An Optimal Approximate Dynamic Programming Algorithm for Concave Single Asset Management Problems

Juliana M. Nascimento
Warren B. Powell
Department of Operations Research and Financial Engineering
Princeton University

February 22, 2007
Abstract
1 Introduction

We propose an approximate dynamic programming algorithm that provably finds an optimal policy to multistage stochastic problems that are part of a problem class here referred to as the concave single asset management class. In this class of problems, the objective is to manage a single asset over a finite and discrete time horizon in order to maximize the expected profits. At each time period, we must act on the asset taking into consideration the current asset level and the new exogenous information that became available since the previous action. The decisions are integer valued vectors and the exogenous information process is Markov and independent of the decisions. Most importantly, its probability distribution is unknown. Not even a parametric form is known. The state of the system at time $t$ is the available information and the current asset level, denoted by $W_t$ and $R_t$, respectively. The key fact that makes a problem part of this problem class is that its optimal value functions must be concave and piecewise linear with breakpoints on the integers in the asset dimension.

A fair amount of applications falls under the framework. In section 5, we describe a single asset inventory system where demand, selling and replenishing prices fluctuates over time and the inventory manager must decide at each period the portion of the random demand that should be satisfied and the amount of assets that should be purchased for replenishment, given that the replenishing quantity must be lower than a finite bound. In section 5, we also prove that this problem is part of our problem class. Another example is a mutual fund application where the fund manager must decide the cash level of the fund, in order to meet customer redemptions and not to lose investment opportunities. An application involving the purchasing of forward contracts over time to satisfy a random demand for a commodity in the future, as described in Nascimento & Powell (2006), also falls in our framework.

Since we do not know the probability distribution of the stochastic process, standard methods that requires the computation of expected values, like classical backwards dynamic programming and multistage stochastic programming, can not be applied. In fact, solving stochastic optimization problems using a distribution free, nonparametric approach has been attracting a lot of attention. A distribution free revenue management and multiproduct pric-
ing applications can be found in van Ryzin & McGill (2000) and Rusmevichientong et al. (2006), respectively. A single-period newsvendor problem and its multi-period extension, when the demand distribution is unknown, are considered in Levi et al. (2006). The authors 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.

Moreover, even though the asset level state variable is a scalar, the state space can still be quite large as the curse of dimensionality may arise even with one dimensional state spaces. Therefore, Q-learning (Watkins & Dayan (1992)), which is considered the standard technique for nonparametric problems, is not applicable, since the state space is even bigger, as it is enlarged with all possible actions. Furthermore, even if the distribution was known, standard dynamic programming may not be viable. For example, in Halman et al. (2006), the authors prove that finding an optimal policy to a single-item stochastic lot-sizing problem with known distribution is NP-hard. They also develop approximation algorithms to deal with it.

We propose a pure exploitation algorithm that explores explicitly the structural properties of the optimal value functions of this problem class. It combines Monte Carlo simulation, stochastic approximation and a projection operation. It is a generalization of the SPAR (Powell et al. (2004)) algorithm, which is presented in the context of a two-stage problem. At each iteration \( n \) and time period \( t \), the algorithm proceeds by solving problems of the form:

\[
x_t^n = \arg \max_{x \in \mathcal{A}(R_{t-1}^n, W_t^n)} C_t(R_{t-1}^n, W_t^n, x) + \gamma \bar{V}_{t-1}^n(W_t^n, g(R_{t-1}^n) + A \cdot x),
\]

where \( R_{t-1}^n \) is the asset level in iteration \( n \) after the decision at time \( t - 1 \) has been made and \( W_t^n \) is a sample realization of the exogenous information at iteration \( n \) and period \( t \). Our convergence proof requires \( W_t^n \) to have a finite support discrete, as would occur in any practical application. However, our algorithm allows it to be continuous even though we discretize the value function approximation. Moreover, \( \mathcal{A}(R_{t-1}^n, W_t^n) \) is a convex set constraint set, \( \bar{V}_{t-1}^n(W_t^n, R_t^n) \) is an approximation to the dynamic programming optimal value function, \( g \) is a bounded scalar function, \( A \) is an integer valued input-output vector
and $A \cdot x$ is an inner product operation that translates the effect the decisions have on the asset.

Let $F_t(R_{t-1}^n, W_t^n, \bar{V}_{t-1}^{n-1}, \cdot) : \mathbb{R}^l \rightarrow \mathbb{R}$ be given by

$$F_t(R_{t-1}^n, W_t^n, \bar{V}_{t-1}^{n-1}, x) = C_t(R_{t-1}^n, W_t^n, x) + \bar{V}_{t-1}^{n-1}(W_t^n, g(R_{t-1}^n) + A \cdot x),$$

The slopes of $F_t(R_{t-1}^n, W_t^n, \bar{V}_{t-1}^{n-1}, x)$ to the left and right of $x^n_t$ (which is an integer breakpoint of $F_t(R_{t-1}^n, W_t^n, \bar{V}_{t-1}^{n-1}, x)$) are used to update $\bar{V}_{t-1}^{n-1}$ obtaining $\bar{V}_{t-1}^{n}$. These slopes depend both on the sample information and on $\bar{V}_{t-1}^{n-1}(W_t^n, \cdot)$, 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 dependence on sample information and on the approximation of the value function in the future is common in approximate dynamic programming algorithms (see Bertsekas & Tsitsiklis (1996), Sutton & Barto (1998)), 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 & Powell (2002) and Topaloglu & Powell (2006)). To the best of our knowledge, as of this writing, the only convergence result for multistage problems that uses a pure exploitation scheme can be found in Nascimento & Powell (2006), which is just a special case of our problem class.

Our proof techniques come from the field of approximate dynamic programming (notably Bertsekas & Tsitsiklis (1996)) as well as the proof of the SPAR algorithm in Powell et al. (2004), generalizing the proof of the ADP-Lagged algorithm in Nascimento & Powell (2006). Current proofs of convergence for approximate dynamic programming algorithms such as Q-learning (Tsitsiklis (1994), Jaakkola et al. (1994)) and optimistic policy iteration (Tsitsiklis (2002)) require that we visit states (and possibly actions) infinitely often. A convergence proof for a Real Time Dynamic Programming (Barto et al. (1995)) algorithm that considers a pure exploitation scheme is provided in Bertsekas & Tsitsiklis (1996) [Prop. 5.3 and 5.4], but it 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 optimal value functions.
Competing approaches to deal with our problem class would be different flavors of Benders decomposition (Van Slyke & Wets (1969), Higle & Sen (1991), Chen & Powell (1999)) and the sample average approximation (Shapiro (2003)) (SAA). However, Benders decomposition will not handle an arbitrary type of exogenous information. On the other hand, SAA 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 & Shapiro (2002).

The contributions of the paper are: a) We propose an approximate dynamic programming algorithm for a whole class of multistage nonparametric stochastic problems using pure exploitation; b) we prove convergence of the algorithm and c) we show that this algorithm can be applied to find an optimal policy to realistic single asset inventory problem, where demand, selling and replenishing prices are random and possibly correlated.

This paper is organized as follows. Section 2 introduces the problem class and the corresponding dynamic programming model. Section 3 describes the algorithm. Section 4 proves that the algorithm does converge to an optimal policy. Section 5 describes the inventory system and shows that indeed this problem is part of the problem class. Finally, section 6 presents the conclusions.

2 The Storage Class And Its Model

A single asset is storaged and as new information becomes available at each time period $t$, a management decision must be taken. The objective is to maximize the expected profits over the horizon $t = 0, \ldots, T$, when a discount factor $0 < \gamma \leq 1$ is considered.

The information process is denoted by $W_t$. It is exogenous to the system and we assume it is Markov, even though its underlying distribution might be unknown. We also assume it has finite support. In our problem class, $W_t$ is a vector in $R^n$. Demand for the storaged asset is always part of the vector. The demand can be heterogenous and do not need to be satisfied, however, no backlogging is allowed. We assume it is integer-valued. We emphasize
that all sorts of information can also be contained in $W_t$. Good examples are: buying/selling prices, interest rates, temperatures and so on.

We denote by $R_t$ the storage level right before a decision is taken. It is a scalar quantity. We define $S_t = (W_t, R_t)$ as the pre-decision state. Taking into consideration $S_t$, the decision $x_t$ is made. It is a vector in $\mathbb{R}^l$ satisfying a set of constraints that may depend on $S_t$. We denote by $\mathcal{X}(S_t)$ the constraint set. We assume that it is convex and is contained in $\{x \in \mathbb{R}^l : 0 \leq x \leq M_t\}$, where $M_t$ is a deterministic bounding vector. Moreover, the decision may be chosen to be integer.

The storage level right after the decision is taken is denoted by $R^x_t$. We define $S^x_t = (W_t, R^x_t)$ as the post-decision state. We have that the asset level evolves according to

$$R_t = g_1(R^x_{t-1}, W_t) \quad \text{and} \quad R^x_t = g_2(W_t, R_t) + A \cdot x_t,$$

where $g_1$ and $g_2$ are an integer bounded transition functions and $A \in \mathbb{Z}^l$ is an input-output vector. Note that the decision $x_t$ impacts the asset level in a linear way. Moreover, it does not influence in any way the information process. Given our assumptions, we can infer that both $R_t$ and $R^x_t$ are integer, bounded and nonnegative.

Finally, the contribution in each period, which is linear is $x_t$, is given by

$$C_t(S_t, x_t) = c_{t0}(S_t) + \sum_{i=1}^l c_{ti}(S_t)x_{ti},$$

where, for $i = 0, \ldots, l$, $c_{ti}(S_t)$ is a bounded scalar function.

We now define, recursively, the optimal value functions associated with our problem class. At time $t = T$, since it is the end of the planning horizon, the value of being in any state $(W_T, R^x_T)$ is zero. Hence, $V^*_{T}(W_T, R^x_T) = 0$. At time $t - 1$, for $t = T, \ldots, 1$, the value of being in any pre-decision state $(W_{t-1}, R_{t-1})$ does not involve expectations, since the next post-decision state $(W_{t-1}, R^x_{t-1})$ is a deterministic function of $W_{t-1}$, it only requires the solution of an optimization problem. On the other hand, the value of being in any post-decision state $(W_{t-1}, R^x_{t-1})$ does not involve the solution of an optimization problem, it only involves an expectation, since the next pre-decision state $(W_t, R_t)$ only depends on the
information that first becomes available at \( t \). Therefore,

\[
V_t^x(W_{t-1}, R_{t-1}) = \max_{x_{t-1} \in \mathcal{X}(W_{t-1}, R_{t-1})} C_{t-1}(W_{t-1}, R_{t-1}, x_{t-1}) + \gamma V_{t-1}^x(W_{t-1}, R_{t-1}) \tag{1}
\]

\[
V_{t-1}^x(W_{t-1}, R_{t-1}) = \mathbb{E} \left[ V_t^x(W_t, R_t)|(W_{t-1}, R_{t-1}) \right]. \tag{2}
\]

If we plug-in (1) on (2), we get

\[
V_{t-1}^x(W_{t-1}, R_{t-1}) = \mathbb{E} \left[ \max_{x_t \in \mathcal{X}(W_t, R_t)} C_t(W_t, R_t, x) + \gamma V_t^x(W_t, R_t)|(W_{t-1}, R_{t-1}) \right]. \tag{3}
\]

Throughout the paper, we only use (3), instead of (1) and (2), i.e., we only consider the value function around the post-decision state. In order to simplify notation, we will just drop the superscript \( x \) in the value function notation. A discussion and an application of value functions around post-decision state variables can be found in Powell & Van Roy (2004) and Van Roy et al. (1997), respectively. We can see that its main feature is the inversion of the optimization/expectation order in the value function formula, allowing for more effective computational strategies.

Given our assumptions, it is not hard to see that the optimal value functions are piecewise linear with integer break points and concave in the asset dimension. Moreover, the optimal decision \( x_t^* = \arg\max_{x_t \in \mathcal{X}(S_t)} C_t(S_t, x_t) + \gamma V_t^x(W_t, g_2(W_t, R_t) + A \cdot x_t) \) is independent of the value of \( V_t^x(W_t, 0) \), for all possible information vectors \( W_t \). We can thus work from now on with a translated version of \( V_t^x \). The version is obtained, for each \( W_t \), subtracting \( V_t^x(W_t, 0) \) from \( V_t^x(W_t, 1) \), \( V_t^x(W_t, 2) \), \( \ldots \), \( V_t^x(W_t, B_t) \), where \( B_t \) is an upper bound for \( R_t^x \). With the translation, the optimal value functions \( V_{t-1}^x(W_{t-1}, \cdot) \) can be uniquely identified by its decreasing slopes \( (v_{t-1}^x(W, 1), \ldots, v_{t-1}^x(W, B_{t-1})) \), where \( v_{t-1}^x(W_{t-1}, R_{t-1}^x) = V_{t-1}^x(W_{t-1}, R_{t-1}^x) - V_{t-1}^x(W_{t-1}, R_{t-1}^x - 1) \). Note that the slopes are not affect by the translation.

Let \( \hat{G}_t(R_{t-1}^x, W_t, V_t^*) \) be a random variable defined as:

\[
\max_{x_t \in \mathcal{X}(W_t, g_1(R_{t-1}^x, W_t))} C_t(W_t, g_1(R_{t-1}^x, W_t), x_t) + \gamma V_t^x(W_t, g_2(W_t, g_1(R_{t-1}^x, W_t)) + A \cdot x_t)
\]

\[
- \max_{y_t \in \mathcal{X}(W_t, g_1(R_{t-1}^x - 1, W_t))} C_t(W_t, g_1(R_{t-1}^x - 1, W_t), y_t) + \gamma V_t^x(W_t, g_2(W_t, g_1(R_{t-1}^x - 1, W_t)) + A \cdot y_t).
\]

We have that

\[
v_{t-1}^x(W_{t-1}, R_{t-1}^x) = \mathbb{E} \left[ \hat{G}_t(R_{t-1}^x, W_t, V_t^*)|W_{t-1}, R_{t-1}^x \right]. \tag{4}
\]
3 The SPAR-MultiPeriod Algorithm

We propose an algorithm, namely the SPAR-MultiPeriod Algorithm, that provably learns the optimal decisions to be taken at important parts of the state space, which are determined by the algorithm itself. This is accomplished constructing concave and piecewise linear function approximations $\bar{V}_n^t(W_t, \cdot)$, learning its slopes $\bar{v}_n^t(W_t, R_{t1}), \ldots, \bar{v}_n^t(W_t, R_{tN})$ over the iterations. Figure 1 illustrates the idea. The algorithm combines Monte Carlo simulation in a pure exploitation scheme, stochastic approximation integrated with a projection operation.

Figure 1: Optimal value function and the constructed approximation

We use the notation convention that an arbitrary pre-decision state at time $t$ is denoted by $S_t = (W_t, R_t)$, while $S^n_t = (W^n_t, R^n_t)$ denotes the actual state visited by the algorithm at iteration $n$ and time $t$. Furthermore, $\{S^n_t\}_{n \geq 0} = \{(W^n_t, R^n_t)\}_{n \geq 0}$ is the sequence of pre-decision states visited by the algorithm. We use the same notation convention for the post-decision state $(S^x_t = (W_t, R^x_t), S^{x,n}_t = (W^n_t, R^{x,n}_t), \{S^{x,n}_t\}_{n \geq 0} = \{(W^n_t, R^{x,n}_t)\}_{n \geq 0})$ and for the decision $(x_t, x^n_t, \{x^n_t\}_{n \geq 0})$. Moreover, the sequences of slopes of the value function approximations are denoted by $\{\bar{v}^n_t\}_{n \geq 0}$. There is one sequence $\{\bar{v}^n_t(W_t, R^x_t)\}_{n \geq 0}$ for each time $t$ and state $(W_t, R_t)$. The notation $\{\bar{v}^n_t\}_{n \geq 0}$ represents the family of all such sequences.

Figure 2 describes the SPAR-MultiPeriod algorithm, a modified version of the SPAR (Powell et al. (2004)) and of the ADP-Lagged (Nascimento & Powell (2006)) algorithms. The algorithm requires initial piecewise linear value function approximations represented by their slopes $\bar{v}^0$. The slopes must be decreasing in the asset dimension and bounded between $-B$ and $B$, where $B$ is a deterministic bound bigger that the upper value of the information.
STEP 0: Algorithm Initialization:

STEP 0a: Initialize $\bar{v}_t^0(W_t, R_t^n)$ for all $t$ and $(W_t, R_t^n)$ monotone decreasing in $R_t^n$.

STEP 0b: Set $n = 1$.

STEP 1: Planning Horizon Initialization: Observe the initial asset level $R^n_{-1}$.

Do for $t = 0, \ldots, T$:

STEP 2: Sample/Observ the information vector $W^n_t$.

STEP 3: Compute the pre-decision asset level: $R^n_t = g_1(R^n_{t-1}, W^n_t)$.

STEP 4: Update Slopes Procedure:

If $t < T$ then

STEP 4a: Observe $\hat{v}_t^n(R^n_{t-1})$ and $\hat{v}_t^n(R^n_{t-1} + 1)$. See [5].

STEP 4b: For all possible states $S^n_t$:

$$z^n_{t-1}(S^n_{t-1}) = (1 - \hat{\alpha}^n_{t-1}(S^n_{t-1}))\bar{v}^{n-1}_{t-1}(S^n_{t-1}) + \hat{\alpha}^n_{t-1}(S^n_{t-1})\hat{v}_t^n(R^n_{t-1})$$

STEP 4c: Perform the projection operation $\bar{v}^n_{t-1} = \Pi_C(z^n_{t-1})$. See [5].

STEP 5: Find the optimal solution $x^n_t$ of

$$\max_{x \in X(W^n_t, R^n_T)} C_t(S^n_t, x) + \gamma \bar{V}^{n-1}_t(W^n_t, R^n_T).$$

STEP 6: Compute the post-decision asset level: $R^{x,n}_t = g_2(W^n_t, R^n_t) + A \cdot x^n_t$.

STEP 7: Increase $n$ by one and go to step 1.

Figure 2: SPAR-MultiPeriod Algorithm

process $W_t$, for $t = 0, \ldots, T$. For example, it is valid to set all the initial slopes equal to zero.

As discussed before, there is a one to one correspondence between the value function and its slopes. Therefore we refer interchangeably to $\bar{V}^n_t(W_t, \cdot)$ and $\bar{v}^n_t(W_t)$. For completeness, we set $\bar{v}^n_T(W_T, R) = 0$ for all iterations $n$ and states $(W_T, R^n_T)$.

At the beginning of each iteration $n$, the algorithm observes the initial asset level $R^{x,n}_{-1}$, as in STEP 1. It has to be a positive integer. After that, the algorithms goes over time periods $t = 0, \ldots, T$. At the beginning of time period $t$, the algorithms observes a sample realization of the information vector, as in STEP 2. The sample can be obtained from a sample generator or actual data. After that, the pre-decision cash level $R^n_t$ is computed, as in STEP 3.
Before the decision at time period $t$ is taken, the algorithm uses the sample information to update the slopes of time period $t - 1$. Steps 4a-4c describes the procedure and figure 3 illustrates it. Sample slopes relative to the post-decision states $(W^{n}_{t-1}, R^{x,n}_{t-1})$ and $(W^{n}_{t-1}, R^{x,n}_{t-1} + 1)$ are observed, see STEP 4a and figure 3a. After that, these sample are used to update the approximation slopes $\hat{v}^{n-1}_{t-1}$, through a temporary slope vector $z^{n}_{t-1}$. This procedure requires the use of a stepsize rule that is state dependent, denoted by $\alpha^{n}_{t}(S^{x}_{t})$, and it may lead to a violation of the slopes monotone decreasing property, see STEP 4a and figure 3b. Thus, a projection operation is performed to restore the property and updated slopes $\bar{v}^{n}_{t-1}$ are obtained, see STEP 4c and figure 3c.

Then, the decision $x^{n}_{t}$, which is optimal with respect to the current pre-decision state $(W^{n}_{t}, R^{n}_{t})$ and value function approximation $\bar{V}^{n-1}_{t}(W^{n}_{t}, \cdot)$ is taken, as stated in STEP 5. Time period $t$ is concluded when the algorithm computes the post-decision state $R^{x,n}_{t}$, as in STEP 6. Same steps are repeated until $t = T$. After that the iteration counter is incremented, as in STEP 7, and a new iteration is started from STEP 1.

We obtain sample slopes by replacing the expectation and the optimal value function $V^{*}_{t}$ in (4) by a sample realization of the information $W^{n}_{t}$ and the current approximation $\bar{V}^{n-1}_{t}$, respectively. Thus, for $t = 1, \ldots, T$, the sample slope is given by

$$\hat{v}^{n}_{t}(R) = \hat{G}_{t}(R, W^{n}_{t}, \bar{V}^{n-1}_{t}).$$

(5)

The projection operator $\Pi_{C}$ maps a vector $z^{n}_{t}$ that may not be monotone decreasing in the cash level dimension, into another vector $\bar{v}^{n}_{t}$ that has this structural property. The operator imposes the property by forcing the newly updated slope at $(W^{n}_{t}, R^{x,n}_{t})$ to be greater or equal to the newly updated slope at $(W^{n}_{t}, R^{x,n}_{t} + 1)$ and then forcing the other violating slopes to be equal to the newly updated ones. For any state $(W^{n}_{t}, R^{x}_{t})$, the projection is given by

$$\Pi_{C}(z^{n}_{t})(W^{n}_{t}, R^{x}_{t}) = \begin{cases} 
\frac{z^{n}(W^{n}_{t}, R^{x,n}_{t}) + z^{n}(W^{n}_{t}, R^{x,n}_{t} + 1)}{2}, & \text{if C1} \\
\Pi_{C}(z^{n}_{t})(W^{n}_{t}, R^{x,n}_{t}), & \text{if C2} \\
\Pi_{C}(z^{n}_{t})(W^{n}_{t}, R^{x,n}_{t} + 1), & \text{if C3} \\
z^{n}_{t}(W^{n}_{t}, R^{x}_{t}), & \text{otherwise},
\end{cases}$$

(6)
where the conditions C1, C2 and C3 are

C1: \( W_t = W^n_t, \quad R^x_t = (R^{x,n}_t \text{ or } R^{x,n}_t + 1), \quad z^n_t(W^n_t, R^{x,n}_t) < z^n_t(W^n_t, R^{x,n}_t + 1) \);

C2: \( W_t = W^n_t, \quad R^x_t < R^{x,n}_t, \quad z^n_t(W_t, R^x_t) \leq \Pi_C(z^n_t)(W^n_t, R^{x,n}_t) \);

C3: \( W_t = W^n_t, \quad R^x_t > R^{x,n}_t + 1, \quad z^n_t(W_t, R^x_t) \geq \Pi_C(z^n_t)(W^n_t, R^{x,n}_t + 1) \).

We close the section with some theoretical remarks. We start with the sequences gener-
ated by the algorithm. The sequences of decisions, pre and post decision states have at least one accumulation point. This result derives from the fact that the information process and the asset level have finite support. Moreover, the decisions are elements of a compact set. Since these are sequences of random variables, their accumulation points are also random variables. We denote them by $x^*_t$, $S^*_t$ and $S^{x,*}_t$, respectively.

We now deal with measurability issues. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be our probability space and $\{\mathcal{F}^n_t; n \geq 0, t = 0, \ldots, T\}$ be the filtration generated by the algorithm defined on $\mathcal{F}$. Clearly, $\mathcal{F}^n_t$ is the sigma-algebra generated by the algorithm up until iteration $n$ and time period $t$. Hence, for $t = 0, \ldots, T - 1$, $\mathcal{F}^n_t \subset \mathcal{F}^n_{t+1}$ and $\mathcal{F}^n_n \subset \mathcal{F}^{n+1}_0$. We have, for all $t = 0, \ldots, T$, that $S^n_t$, $x^n_t$, $S^{x,n}_t$ and $\hat{v}^n_t(R) \in \mathcal{F}^n_t$, while $\bar{v}^n_t(W, R)$ and $z^n_t(W, R) \in \mathcal{F}^n$.

We move on to the stepsizes $\bar{\alpha}^n_{t-1}(S^*_{t-1})$ used in STEP 4b of the algorithm. For each iteration $n$ time period $t$, the stepsize rule is as follow. For the states $(W^n_t, R^{x,n}_t)$ and $(W^n_t, R^{x,n}_t + 1)$, we have that $\bar{\alpha}^n_t(W^n_t, R^{x,n}_t) = \bar{\alpha}^n_t(W^n_t, R^{x,n}_t + 1) > 0$. For all the other states $(W_t, R^x_t)$, we have that $\bar{\alpha}^n_t(W_t, R^x_t) = 0$. Moreover, we assume for each accumulation point $(W^*_t, R^{x,*}_t)$ of the post-decision sequence that:

\begin{align}
\bar{\alpha}^n_t(W_t, R^x_t) \in [0, 1] \text{ and } \bar{\alpha}^n_t(W_t, R^x_t) \in \mathcal{F}^n_t \\
\sum_{n=0}^{\infty} \bar{\alpha}^n_t(W^*_t, R^{x,*}_t)^2 \leq B \text{ a.s.} \\
\sum_{n=0}^{\infty} \bar{\alpha}^n_t(W^*_t, R^{x,*}_t) = \infty \text{ a.s.,}
\end{align}

where $\bar{B}$ is a deterministic constant. Clearly, the stepsize rule

$$
\bar{\alpha}^n_t(W^n_t, R^{x,n}_t) = \frac{1}{NV(W^n_t, R^{x,n}_t)}
$$

satisfies conditions (7)–(9), where $NV(W^n_t, R^{x,n}_t)$ is the number of visits to state $(W^n_t, R^{x,n}_t)$ up until iteration $n$.

Furthermore, for all positive integers $N$,

$$
\prod_{n=N}^{\infty} (1 - \bar{\alpha}^n_t(W^*_t, R^{x,*}_t)) = 0 \text{ a.s.}
$$

The proof for (10) follows directly from the fact that $\log(1 + x) \leq x$. 

11
Finally, we can easily see that $\hat{v}_n^n(R), z^n_t(W_t, R^n_t)$ and $\bar{v}_n^n(W_t, R^n_t)$ are bounded by $-B$ and $B$ for all iterations $n$, as the initial slopes are bounded by $-B$ and $B$ and the stepsizes are bounded by 0 and 1. Therefore, the slope sequences $\{\bar{v}^n\}_{n\geq 0}$ also have at least one accumulation point, as the projection operation guarantees that the updated slopes are elements of a compact set. As before, the accumulation points are random variables and are denoted by $\bar{v}^*$, as opposed to the deterministic optimal slopes $v^*$.

4 Convergence Analysis

We start this section presenting the convergence results we want to prove. The major convergence result is, for all accumulation points $(W^*_t, R^x_t)$

$$\bar{v}_t^n(W^*_t, R^x_t) \to v_t^*(W^*_t, R^x_t) \quad \text{and} \quad \bar{v}_t^n(W^*_t, R^x_t + 1) \to v_t^*(W^*_t, R^x_t + 1) \text{ a.s.} \quad (11)$$

As a byproduct of the previous result, we show, for $t = 0, \ldots, T$, that

$$x^*_t = \arg \max_{x \in \mathcal{X}(R^n_{t-1}, W^n_t)} F_t(R^n_{t-1}, W^n_t, V^*_t, x) \quad \text{a.s.,} \quad (12)$$

where $(R^n_{t-1}, W^n_t, x^*_t)$ is an accumulation point of the sequence $\{(R^n_{t-1}, W^n_t, x^n_t)\}_{n\geq 0}$ generated by the algorithm, $F_t(R^n_{t-1}, W^n_t, V^*_t, x) = C_t(R^n_{t-1}, W^n_t, x) + V^*_t(W^n_t, g(R^n_{t-1}) + A \cdot x)$ and $V^*_t$ is the optimal value function.

Equation (12) implies that the algorithm has learned an optimal decision for all states that can be reached by an optimal policy. This implication can be easily justified as follows. For $t = 0$, since, $R^n_{-1} = 0$ for all iterations of the algorithm, we have that $R^*_0 = 0$. Moreover, all the elements in $\mathcal{W}_0$ are accumulation points of $\{W^n_0\}_{n\geq 0}$, as $\mathcal{W}_0$ has a finite support. Thus, (12) tells us that the accumulation points $x^*_0$ of the sequence $\{x^*_0\}_{n\geq 0}$ along the iterations with initial information $W^*_0$ are in fact an optimal policy for period 0 when the information is $W^*_0$. This implies that all accumulation points $R^*_0 = g(0) + A \cdot x^*_0$ of $\{R^*_0\}_{n\geq 0}$ are asset levels that can be reached by an optimal policy. By the same token, for $t = 1$, every element in $\mathcal{W}_1$ is an accumulation point of $\{W^*_1\}_{n\geq 0}$. Hence, (12) tells us that the accumulation points $x^*_1$ of the sequence $\{x^*_1\}$ along iterations with $(R^*_0, W^*_t) = (R^*_0, W^*_1)$ are indeed an optimal policy for period 1 when the asset level is $R^*_0$ and the information is $W^*_1$. As before,
the accumulation points \( R_1^* = g(R_0^n) + A \cdot x_1^* \) of \( \{R^n\}_{n \geq 0} \) are asset levels that can be reached by an optimal policy. The same reasoning can be applied for \( t = 2, \ldots, T \).

Since we have almost surely convergence proofs, we only consider the sets in the sigma-algebra \( \mathcal{F} \) that have strictly positive measure. The rest of this section is presented as follows. First, we provide an outline of the proofs in section 4.1. After that, we introduce the dynamic programming operator associated with the problem class and deterministic bounding sequences that converge to the optimal slopes. In section 4.3 we introduce technical elements required in the convergence proofs. Finally, complete proofs are given in 4.4 and 4.5.

4.1 Outline of the Convergence Proofs

Our proofs follow the ideas presented in Bertsekas & Tsitsiklis (1996) and in Powell et al. (2004). Moreover, they are a generalization of the proofs in Nascimento & Powell (2006). Bertsekas & Tsitsiklis (1996) obtain their convergence proof by assuming that all states are visited infinitely often. They do not consider a concavity-preserving step which accelerates convergence but complicates the proof. Although the framework in Powell et al. (2004) also considers the concavity of the optimal value functions in the asset dimension, the use of a projection operation to restore concavity and a pure exploitation routine, their proofs are restricted to two stage problems. Finally, the convergence proofs in Nascimento & Powell (2006) just apply to a specific problem within our problem class, namely the purchasing of forward contracts to satisfy demand and maximize the profits.

The main concept to achieve (11) is to consider the dynamic programming operator associated with the problem class and to construct deterministic bounding sequences that are based on this operator and that are provably convergent to the slopes of the optimal value functions. The operator is denoted by \( H \) and the deterministic bounding sequences, for each \( t = 0, \ldots, T \) and \((W, R) \in \mathcal{S}_t\), are be denoted by \( \{L^k_t(W, R)\}_{k \geq 0} \) and \( \{U^k_t(W, R)\}_{k \geq 0} \). A full description of these elements is given in the next section. Then, we use the deterministic bounding sequences to prove, for \((\bar{W}^*, \bar{R}^*) \in \mathcal{S}_t^*\), that

\[
L^k_t(\bar{W}^*, \bar{R}^*) \leq \bar{v}^n_t(\bar{W}^*, \bar{R}^*) \leq U^k_t(\bar{W}^*, \bar{R}^*)
\] (13)
for all $k \geq 0$ and for all sufficiently large $n$.

Establishing (13) requires several intermediate steps that need to take into consideration the pure exploitation nature of our algorithm and the concavity preserving operation. In order to do so, we need to define four stochastic sequences, $\{s^n_0\}_{n \geq 0}$, $\{\bar{s}^n_t\}_{t \geq 0}$, $\{\bar{u}^n_t\}_{t \geq 0}$ and $\{\bar{u}^n_t\}_{t \geq 0}$. The first two sequences are called stochastic noise sequences and the last two sequences are called stochastic bounding sequences.

We now present the main intermediate results that are required to obtain (13). First, for $n$ big enough, we show that

\[ \bar{v}^{n-1}_t(\bar{W}^*, \bar{R}^*) \leq \bar{u}^{n-1}_t(\bar{W}^*, \bar{R}^*) + \bar{s}^{n-1}_t(\bar{W}^*, \bar{R}^*), \quad \text{if } (\bar{W}^*, \bar{R}^*) \in \bar{S}_t^-, \]
\[ \bar{v}^{n-1}_t(\bar{W}^*, \bar{R}^*) \geq \bar{m}^{n-1}_t(\bar{W}^*, \bar{R}^*) - \bar{s}^{n-1}_t(\bar{W}^*, \bar{R}^*), \quad \text{if } (\bar{W}^*, \bar{R}^*) \in \bar{S}_t^+, \]

where both $\bar{S}_t^-$ and $\bar{S}_t^+$ are proper subsets of $\bar{S}_t^*$. The definition of these sets will be given in 4.3. Then, convergence to zero of the noise sequences, a convex combination property of the stochastic bounding sequences and concavity, will give us

\[ \bar{v}^{n-1}_t(\bar{W}^*, \bar{R}^*) \leq U^k_t(\bar{W}^*, \bar{R}^*), \quad \text{if } (\bar{W}^*, \bar{R}^*) \in \bar{S}_t^-, \]
\[ \bar{v}^{n-1}_t(\bar{W}^*, \bar{R}^*) \geq L^k_t(\bar{W}^*, \bar{R}^*), \quad \text{if } (\bar{W}^*, \bar{R}^*) \in \bar{S}_t^+. \]

Note that the last two inequalities do not cover all states in $\bar{S}_t^*$. Nevertheless, these inequalities and some properties of the projection operation are used to fulfill the requirements of a bounding technical lemma, which is used repeatedly to obtain (13) for all states in $\bar{S}_t^*$, which implies (11), the first convergence result.

As $F_t(R^n_{t-1}, W^n_t, \bar{V}^{n-1}_t, x) = C_t(R^n_{t-1}, W^n_t, x) + \gamma \bar{V}^{n-1}_t(W^n_t, g(R^n_{t-1}) + A \cdot x)$ is a concave function and $\mathcal{X}(R^n_{t-1}, W^n_t)$ is a convex set, we have that

\[ 0 \in \partial F_t(R^n_{t-1}, W^n_t, \bar{V}^{n-1}_t, x^n_t) - \mathcal{X}^N(R^n_{t-1}, W^n_t, x^n_t), \]

where $x^n_t$ is the optimal decision of the optimization problem in STEP 3a of the algorithm, $\partial F_t(R^n_{t-1}, W^n_t, \bar{V}^{n-1}_t, x^n_t)$ is the subdifferential of $F_t(R^n_{t-1}, W^n_t, \bar{V}^{n-1}_t, \cdot)$ at $x^n_t$ and $\mathcal{X}^N(R^n_{t-1}, W^n_t, x^n_t)$ is the normal cone of $\mathcal{X}(R^n_{t-1}, W^n_t)$ at $x^n_t$. This inclusion and the first convergence result are then used to obtain (12), as it will be shown in section 4.5 that

\[ 0 \in \partial F_t(R^*_t, W^*_t, V^*_t, x^*_t) - \mathcal{X}^N(R^*_t, W^*_t, x^*_t). \]
4.2 Dynamic Programming Operator and Deterministic Sequences

The dynamic programming operator $H$ maps a vector $v$ into a new vector $Hv$ as follows. For $t = 0, \ldots, T$ and $(W, R) \in \mathcal{S}_t$,

$$
(Hv)_t(W, R) = \begin{cases} 
E \left[ \hat{G}_{t+1}(R, \hat{W}_{t+1}, v_{t+1}) | W_t = W \right], & \text{if } t < T \\
0, & \text{if } t = T.
\end{cases}
$$

The expression for the random variable $\hat{G}_{t+1}(R, \hat{W}_{t+1}, v_{t+1})$ depends on the specific problem within the problem class. Clearly, the vector of slopes of the optimal value functions $v^*$ is the unique fixed point of $H$.

We present the deterministic bounding sequences $\{U_k\}_{k \geq 0}$ and $\{L_k\}_{k \geq 0}$. When we refer to them without mentioning the time index $t$ and the state $(W, R)$, we are referring to the family of sequences $\{U_k^t(W, R)\}_{k \geq 0}$ and $\{L_k^t(W, R)\}_{k \geq 0}$, one for each $t = 0, \ldots, T - 1$ and $(W, R) \in \mathcal{S}_t$. Let

$$
L^0 = v^* - 2Be \quad \text{and} \quad L^{k+1} = \frac{L^k + HL^k}{2}, \quad k \geq 0. \quad (14)
$$

$$
U^0 = v^* + 2Be \quad \text{and} \quad U^{k+1} = \frac{U^k + HU^k}{2}, \quad k \geq 0. \quad (15)
$$

Moreover, $L^k_t(W, R) = U^k_t(W, R) = 0$ for all $k \geq 0$ and states $(W, R) \in \mathcal{S}_T$. We conclude the section presenting a proposition with properties of the deterministic sequences necessary for the slope convergence proof. Its proof is deferred to the appendix.

**Proposition 1.** For $t = 0, \ldots, T-1$, states $(W, R) \in \mathcal{S}_t$ and vectors $\tilde{v}$ such that $\tilde{v}_{t+1}(W_{t+1}) \in C_{t+1}$ for all $W_{t+1} \in W_{t+1}$, assume that

$$
(H\tilde{v})_t(W) \in C_t \quad (16)
$$

$$
(H\tilde{v})_t(W, R) \leq (H\tilde{v})_t(W, R), \ \text{if } \tilde{v}_{t+1} \leq \tilde{v}_{t+1} \text{ and } \tilde{v}_{t+1}(W_{t+1}) \in C_{t+1} \quad (17)
$$

$$
(H\tilde{v})_t(W, R) - \eta \leq (H(\tilde{v} - \eta E))_t(W, R) \leq (H(\tilde{v} + \eta E))_t(W, R) \leq (H\tilde{v})_t(W, R) + \eta, \quad (18)
$$

where $\eta$ is a positive constant and $e$ is a vector with all components equal to 1. Then, for all
\[ k \geq 0, \]
\[ L^k_t(W) \in C_t \quad \text{and} \quad U^k_t(W) \in C_t \]  \hspace{1cm} (19)
\[ HL^k \geq L^{k+1} \geq L^k \quad \text{and} \quad HU^k \leq U^{k+1} \leq U^k \]  \hspace{1cm} (20)
\[ L^k < v^*, \quad \lim_{k \to \infty} L^k = v^* \quad \text{and} \quad U^k > v^*, \quad \lim_{k \to \infty} U^k = v^*. \]  \hspace{1cm} (21)

### 4.3 Technical Elements

We start presenting the integer random variable \( \bar{N} \). This variable 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 all accumulation points \( ((W_0^*, R_0^*, x_0^*), \ldots, (W_T^*, R_T^*, x_T^*)) \) of \( \{(W_0^m, R_0^m, x_0^m), \ldots, (W_T^m, R_T^m, x_T^m)\}_m \geq 0 \) have been observed at least once. Moreover, \( \bar{N} \) is also the smallest integer such that if a mathematical statement regarding the sequences of slopes, states and decisions generated by the algorithm is true only for finitely many iterations, then it is false for all iterations \( n \geq \bar{N} \). It is easy to see that \( \bar{N} \) is finite almost surely. Examples of mathematical statements are:

If \( \sum_{n=1}^{\infty} 1\{(R_{t-1}^n, W_t^n, x_t^n) = (R, W, x)\} < \infty \), then \( \sum_{n=\bar{N}}^{\infty} 1\{(R_{t-1}^n, W_t^n, x_t^n) = (R, W, x)\} = 0; \)

If \( \sum_{n=1}^{\infty} 1\{\bar{v}^n_t(W_t, R) < 0\} 1\{\bar{v}^n_t(W_t, R) > 10\} < \infty \), then \( \sum_{n=\bar{N}}^{\infty} 1\{\bar{v}^n_t(W_t, R) < 0\} 1\{\bar{v}^n_t(W_t, R) > 10\} = 0. \)

For each \( (W, R) \in \tilde{S}_t \), we introduce the sets of iterations \( \mathcal{N}^-_t(W, R) \) and \( \mathcal{N}^+_t(W, R) \). They contain the iterations in which the corresponding slopes were modified by the projection operation. Let \( \mathcal{N}^-_t(W, R) \) (\( \mathcal{N}^+_t(W, R) \)) be the set of iterations in which the unprojected slope corresponding to state \( (W, R) \) was too small (large) and had to be increased (decreased) by the projection operation. Formally,

\[ \mathcal{N}^-_t(W, R) = \{n \in \mathbb{N} : z^n_t(W, R) < \bar{v}^n_t(W_t, R)\} \]
\[ \mathcal{N}^+_t(W, R) = \{n \in \mathbb{N} : z^n_t(W, R) > \bar{v}^n_t(W_t, R)\}. \]

For example, based on figure 3c,

\[ n \in \mathcal{N}^-_t(W^n_t, R^n_t - 1) \quad \text{and} \quad n \in \mathcal{N}^+_t(W^n_t, R^n_t + 2). \]
We close this section defining two proper sets of $\tilde{S}_t^*$, namely $\tilde{S}_t^-$ and $\tilde{S}_t^+$. The states in $\tilde{S}_t^-$ ($\tilde{S}_t^+$) are the ones for which the projection operation decreased (increased) or kept the same the corresponding unprojected slopes infinitely often, that is, for $(W, R) \in \tilde{S}_t^*$, $\mathcal{N}_t^-(W, R)$ ($\mathcal{N}_t^+(W, R)$) is finite if and only if $(W, R) \in \tilde{S}_t^- (\tilde{S}_t^+)$. That is,

\[ \tilde{S}_t^- = \{(W, R) \in \tilde{S}_t^* : z_t^n(W, R) \geq \bar{v}_t^n(W, R) \text{ for all } n \geq \bar{N}\} \]
\[ \tilde{S}_t^+ = \{(W, R) \in \tilde{S}_t^* : z_t^n(W, R) \leq \bar{v}_t^n(W, R) \text{ for all } n \geq \bar{N}\}. \]

Due to the definition of the projection operator, $\tilde{S}_t^+$ is not empty, since $(W, R^{\text{Min}}) \in \tilde{S}_t^+$, where $R^{\text{Min}}$ is the minimum asset level such that $(W, R) \in \tilde{S}_t^*$. We can use a similar argument to show that $\tilde{S}_t^-$ is not empty.

### 4.4 Almost Sure Convergence of the Slopes

In this section we prove that the slopes of the approximate functions converges almost surely to the slopes of the optimal value functions for states in $\tilde{S}_t^*$. This result is stated in theorem 1 below. Along the proof of the theorem, we introduce three technical lemmas and give details about the noise and the bounding stochastic sequences. The proofs of the lemmas are given in the appendix so that the main reasoning is not disrupted.

Before stating the theorem, we give a summary of the problem class properties required for the proof. We also introduce a technical assumption, which fulfillment is highly dependent on the specific problem under consideration. We show in section 5.3 that the inventory system does satisfy this assumption.

We start with the problem class properties. For $t = 0, \ldots, T$, assume that

- The random variable $W_t \in \mathbb{R}^n$ has finite support; \hspace{1cm} (22)
- $R_t \in \{0, \ldots, B_t\}$; \hspace{1cm} (23)
- $x_t \in \mathcal{X}(R_{t-1}, W_t)$ is integer and $\mathcal{X}(R_{t-1}, W_t)$ is convex; \hspace{1cm} (24)
- $C_t(R_{t-1}, W_t, x_t)$ is linear in $x_t$; \hspace{1cm} (25)
- $\bar{G}_t(R_{t-1}, W_t, v_t^*)$ is bounded; \hspace{1cm} (26)
- $V_t^*(W, \cdot)$ has integer breakpoints and its slopes $v_t^*(W)$ is in $C_t$. \hspace{1cm} (27)
We now introduce the technical assumption. For a given \( k \geq 0 \) and \( t = 0, \ldots, T - 1 \), if there exists a positive integer \( N^k_t \) such that \( N^k_t \geq \bar{N} \) and for all \( n \geq N^k_t \) and states \((\bar{W}^*, \bar{R}^*) \in \bar{S}_{t+1}^*\), we have that

\[
L^k_{t+1}(\bar{W}^*, \bar{R}^*) \leq \bar{v}^{n-1}_{t+1}(\bar{W}^*, \bar{R}^*) \leq U^k_{t+1}(\bar{W}^*, \bar{R}^*),
\]  

(28)

then, for \( n \geq N^k_t \),

\[
(HL^k)_t(W^n_t, R^n_t) \leq (H\bar{v}^{n-1})_t(W^n_t, R^n_t) \leq (HU^k)_t(W^n_t, R^n_t)
\]  

(29)

\[
(HL^k)_t(W^n_t, R^n_t + 1) \leq (H\bar{v}^{n-1})_t(W^n_t, R^n_t + 1) \leq (HU^k)_t(W^n_t, R^n_t + 1),
\]  

(30)

where \((W^n_t, R^n_t)\) is the state being visited by the algorithm at iteration \( n \) at time \( t \) and \( \bar{v}^{n-1} \) is the approximate slopes obtained at the end of iteration \( n - 1 \). These inequalities are non-trivial due to the pure exploitation nature of our algorithm and are highly dependent on the dynamics of the problem.

**Theorem 1.** Assume the stepsize conditions (7)–(9) and the problem class properties (22)–(27). Also assume the deterministic sequences conditions (19)–(21). Finally, assume that if there exists an integer \( N^k_t \) such that (28) is true for all \( n \geq N^k_t \) and states \((\bar{W}^*, \bar{R}^*) \in \bar{S}^*_t\), then (29)–(30) hold true for all \( n \geq N^k_t \). Then, for all \( k \geq 0 \) and \( t = 0, \ldots, T \), there exists a positive integer \( N^{*,k}_t \) such that, for all \( n \geq N^{*,k}_t \) and states \((\bar{W}^*, \bar{R}^*) \in \bar{S}^*_t\),

\[
L^k_t(\bar{W}^*, \bar{R}^*) \leq \bar{v}^{n-1}_t(\bar{W}^*, \bar{R}^*) \leq U^k_t(\bar{W}^*, \bar{R}^*) \quad \text{a.s.}
\]  

(31)

Therefore,

\[
\bar{v}^n_t(\bar{W}^*, \bar{R}^*) \to v^*_t(\bar{W}^*, \bar{R}^*) \quad \text{a.s.}
\]  

(32)

**Proof.** The proof of the theorem is by backward induction on \( t \). The base case is \( t = T \). As \( v^*_T(W, R) = U^k_T(W, R) = L^k_T(W, R) = \bar{v}^n_T(W, R) = 0 \) for all states \((W, R) \in \bar{S}_T\), integers \( k \geq 0 \) and iterations \( n \geq 0 \), the inequalities in (31) are trivial for \( t = T \). In particular, we can take \( N^{*,k}_T = \bar{N} \).

The backward induction proof is completed when we prove (31) for a general \( t, t = 0, \ldots, T - 1 \). Given the induction hypothesis for \( t + 1 \), the proof for time period \( t \) is divided
into two parts. We prove for all \( k \geq 0 \) that there exists an integer \( N_t^k \) such that, for all \( n \geq N_t^k \),

\[
\begin{align*}
\bar{v}_{t}^{n-1}(\tilde{W}^*, \tilde{R}^*) &\leq U_t^k(\tilde{W}^*, \tilde{R}^*), & \text{if } (\tilde{W}^*, \tilde{R}^*) \in \tilde{S}_t^-, \\
\bar{v}_{t}^{n-1}(\tilde{W}^*, \tilde{R}^*) &\geq L_t^k(\tilde{W}^*, \tilde{R}^*), & \text{if } (\tilde{W}^*, \tilde{R}^*) \in \tilde{S}_t^+.
\end{align*}
\]  

This is the first part. Its proof is by induction on \( k \). Note that this part only applies to states in the sets \( \tilde{S}_t^- \) and \( \tilde{S}_t^+ \). Then, again for \( t \), we take on the second part, which proves the existence of an integer \( N_t^{s,k} \) such that (31) is true for all states in \( \tilde{S}_t^s \) and iterations \( n \geq N_t^{s,k} \). Note that the second part takes care of the states in \( \tilde{S}_t^s \) not covered by the first part. Consequently, (32) is true for \( t \). Figure 4 shows the relationship between the sets of states.

Figure 4: Relationship between the sets of states.

We start the backward induction on \( t \). Remember that the base case \( t = T \) is trivial and we pick \( N_{T}^{s,k} = \bar{N} \). We also pick, for completeness, \( N_{T}^{k} = \bar{N} \).

**Induction Hypothesis:** Given \( t = 0, \ldots, T-1 \), assume, for \( t+1 \), and all \( k \geq 0 \) the existence of integers \( N_{t+1}^{k} \) and \( N_{t+1}^{s,k} \) such that, for all \( n \geq N_{t+1}^{k} \), (33) and (34) are true, and, for all \( n \geq N_{t+1}^{s,k} \), the inequalities in (31) hold true for all states \( (\tilde{W}^*, \tilde{R}^*) \in \tilde{S}_{t+1}^s \).

**Part 1:**

Now we prove for any \( k \), the existence of an integer \( N_{t}^{k} \) such that for \( n \geq N_{t}^{k} \), inequalities (33) and (34) are true. For a particular time \( t \), the proof is by forward induction on \( k \).

We start with \( k = 0 \). For every \( (W, R) \in \tilde{S}_t \), \( -B \leq v_t^*(W, R) \leq B \) implying that, by
definition, \( U_t^0(W, R) \geq B \) and \( L_t^0(W, R) \leq -B \). Therefore, (33) and (34) are satisfied for all \( n \geq 1 \), since we know that \( v^{n-1} \) is bounded by \(-B\) and \( B\) for all iterations. Thus, we can take \( N_t^0 = \max(1, N_{t+1}^{*,0}) = N_{t+1}^{*,0} \).

The induction hypothesis on \( k \) assumes that there exists \( N_t^k \) such that, for all \( n \geq N_t^k \), (33) and (34) are true. Note that we can always make \( N_t^k \) larger than \( N_{t+1}^{*,k} \), thus we assume that \( N_t^k \geq N_{t+1}^{*,k} \). The next step is the proof for \( k + 1 \).

Before we move on, we define the variables \( \hat{s}_{t+1}^n - (R) \) and \( \hat{s}_{t+1}^n + (R) \) to be the error incurred by observing a sample slope. For \( R = 1, \ldots, B_t \),

\[
\hat{s}_{t+1}^n - (R) = \hat{v}_{t+1}^n(R) - (H \bar{v}^{n-1})_t(W_t^n, R) \quad \text{and} \quad \hat{s}_{t+1}^n + (R) = -\hat{s}_{t+1}^n - (R).
\]

Using \( \hat{s}_{t+1}^- \) and \( \hat{s}_{t+1}^+ \), we define the stochastic noise sequences \( \{\hat{s}_{t-}^n\}_{n \geq 0} \) and \( \{\hat{s}_{t+}^n\}_{n \geq 0} \). For \( (W, R) \in \bar{S}_t \),

\[
\hat{s}_{t-}^n(W, R) = 0 \quad \text{and} \quad \hat{s}_{t+}^n(W, R) = 0, \quad \text{for} \ n < N_t^k,
\]

and, for \( n \geq N_t^k \),

\[
\hat{s}_{t-}^n(W, R) = \max(0, (1 - \bar{\alpha}_t^n(W, R)) \hat{s}_{t-}^{n-1}(W, R) + \bar{\alpha}_t^n(W, R) \hat{s}_{t+1-}^n(R_{t}^{n}1_{\{R_{t} \leq R^*_t\}} + (R_t^{n} + 1)1_{\{R_{t} > R^*_t\}}))
\]

\[
\hat{s}_{t+}^n(W, R) = \max(0, (1 - \bar{\alpha}_t^n(W, R)) \hat{s}_{t+}^{n-1}(W, R) + \bar{\alpha}_t^n(W, R) \hat{s}_{t+1+}^n(R_{t}^{n}1_{\{R_{t} \leq R^*_t\}} + (R_t^{n} + 1)1_{\{R_{t} > R^*_t\}}))
\]

The sample slopes are defined in a way such that

\[
E \left[ \hat{s}_{t+1}^- (R) | F_t^n \right] = 0. \quad (35)
\]

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

**Lemma 1.** For \((\bar{W}^*, \bar{R}^*) \in \bar{S}_t^*\),

\[
\{\hat{s}_{t-}^n(\bar{W}^*, \bar{R}^*)\}_{n \geq 0} \to 0 \quad \text{and} \quad \{\hat{s}_{t+}^n(\bar{W}^*, \bar{R}^*)\}_{n \geq 0} \to 0 \ a.s. \quad (36)
\]
Proof of lemma 1. Given in the appendix. \qed 

Using the convention that the minimum of an empty set is $+\infty$, let

$$\delta^k_L = \min \left\{ \frac{(HL^k)_{t}(\tilde{W}^*, \tilde{R}^*) - L^k_t(\tilde{W}^*, \tilde{R}^*)}{4} : (\tilde{W}^*, \tilde{R}^*) \in \tilde{S}_t^+, (HL^k)_{t}(\tilde{W}^*, \tilde{R}^*) > L^k_t(\tilde{W}^*, \tilde{R}^*) \right\}.$$ 

If $\delta^k_L < +\infty$ we define an integer $N_L \geq N^k_t$ to be such that

$$\prod_{m=N^k_t}^{N_L-1} \left( 1 - \tilde{\alpha}^m_\ast(\tilde{W}^*, \tilde{R}^*) \right) \leq 1/4 \quad \text{and} \quad \tilde{s}^{n-1}_\ast(\tilde{W}^*, \tilde{R}^*) \leq \delta^k_L,$$

for all $n \geq N_L$ and states $(\tilde{W}^*, \tilde{R}^*) \in \tilde{S}_t^+$. Such an $N_L$ exists because both (10) and (36) are true. If $\delta^k_L = +\infty$ then, for all states $(\tilde{W}^*, \tilde{R}^*) \in \tilde{S}_t^+$, $(HL^k)_{t}(\tilde{W}^*, \tilde{R}^*) = L^k_t(\tilde{W}^*, \tilde{R}^*)$ since (20) tells us that $HL^k \geq L^k$. Thus, $L^{k+1}_t(\tilde{W}^*, \tilde{R}^*) = L^k_t(\tilde{W}^*, \tilde{R}^*)$ and we define the integer $N_L$ to be equal to $N^k_t$.

We can apply symmetric reasoning to determine $\delta^U_k$ and $N_U$. We just need to consider the deterministic bounding sequence $\{U^k\}_{k \geq 0}$, the set $\tilde{S}_t^-$ and the noise sequence $\{\tilde{s}^n_t\}_{n \geq 0}$ instead of $\{L^k\}_{k \geq 0}$, $\tilde{S}_t^+$ and $\{\tilde{s}^n_t\}_{n \geq 0}$, respectively.

Finally, let $N^{k+1}_t = \max\left( N_L, N_U, N^k_{t+1} \right)$. First, pick a state $(\tilde{W}^*, \tilde{R}^*) \in \tilde{S}_t^+$. If $L^{k+1}_t(\tilde{W}^*, \tilde{R}^*) = L^k_t(\tilde{W}^*, \tilde{R}^*)$, then inequality $L^{k+1}_t(\tilde{W}^*, \tilde{R}^*) \leq \tilde{v}^{n-1}(\tilde{W}^*, \tilde{R}^*)$ follows from the induction hypothesis. We therefore concentrate on the case where $L^{k+1}_t(\tilde{W}^*, \tilde{R}^*) > L^k_t(\tilde{W}^*, \tilde{R}^*)$.

First, we define the stochastic bounding sequences $\{\tilde{l}^n_t\}_{n \geq 0}$ and $\{\tilde{u}^n_t\}_{n \geq 0}$. For each $(W, R) \in \tilde{S}_t$, we have

$$\tilde{l}^n_t(W, R) = L^k_t(W, R) \quad \text{and} \quad \tilde{u}^n_t(W, R) = U^k_t(W, R), \quad \text{for } n < N^k_t,$$

and, for $n \geq N^k_t$,

$$\tilde{l}^n_t(W, R) = (1 - \tilde{\alpha}^n_t(W, R)) \tilde{l}^{n-1}_t(W, R) + \tilde{\alpha}^n_t(W, R)(HL^k)_{t}(W, R)$$

$$\tilde{u}^n_t(W, R) = (1 - \tilde{\alpha}^n_t(W, R)) \tilde{u}^{n-1}_t(W, R) + \tilde{\alpha}^n_t(W, R)(HU^k)_{t}(W, R).$$

A simple inductive argument proves that $\tilde{u}^n_t(W, R)$ is a convex combination of $U^k_t(W, R)$ and $(HU^k)_{t}(W, R)$, while $\tilde{l}^n_t(W, R)$ is a convex combination of $L^k_t(W, R)$ and $(HL^k)_{t}(W, R)$. 

21
Therefore, we can write, taking $\bar{v}_t^{n-1} = \prod_{m=N_t^k}^{n-1} \left(1 - \bar{\alpha}_t^m(\bar{W}^*, \bar{R}^*)\right)$,

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

For $n \geq N_t^{k+1} \geq N_L$, we have $\bar{b}_t^{n-1} \leq 1/4$. Moreover, $L_t^k(\bar{W}^*, \bar{R}^*) \leq (HL^k)_t(\bar{W}^*, \bar{R}^*)$. Thus, using (14) and the definition of $\delta^k_L$, we obtain

$$
\bar{v}_t^{n-1}(\bar{W}^*, \bar{R}^*) \geq \frac{1}{4} L_t^k(\bar{W}^*, \bar{R}^*) + \frac{3}{4}(HL^k)_t(\bar{W}^*, \bar{R}^*)
= \frac{1}{2} L_t^k(\bar{W}^*, \bar{R}^*) + \frac{1}{2}(HL^k)_t(\bar{W}^*, \bar{R}^*) + \frac{1}{4}((HL^k)_t(\bar{W}^*, \bar{R}^*) - L_t^k(\bar{W}^*, \bar{R}^*))
\geq L_t^{k+1}(\bar{W}^*, \bar{R}^*) + \delta^k_L.
$$

Again for $n \geq N_t^{k+1} \geq N_L$, the following lemma is used to show that

$$
\bar{v}_t^{n-1}(\bar{W}^*, \bar{R}^*) \geq \bar{v}_t^{n-1}(\bar{W}^*, \bar{R}^*) - s_{t+1}^{n-1}(\bar{W}^*, \bar{R}^*).
$$

**Lemma 2.** For $n \geq N_t^k$,

$$
\bar{v}_t^{n-1}(\bar{W}^*, \bar{R}^*) \leq \bar{u}_t^{n-1}(\bar{W}^*, \bar{R}^*) + \bar{s}_{t-1}^{n-1}(\bar{W}^*, \bar{R}^*), \quad \text{if } (\bar{W}^*, \bar{R}^*) \in \bar{S}^-_t,
$$

$$
\bar{v}_t^{n-1}(\bar{W}^*, \bar{R}^*) \geq \bar{v}_t^{n-1}(\bar{W}^*, \bar{R}^*) - s_{t+1}^{n-1}(\bar{W}^*, \bar{R}^*), \quad \text{if } (\bar{W}^*, \bar{R}^*) \in \bar{S}^+_t.
$$

**Proof of lemma 2** Given in the appendix.

Combining (38) and (39), we obtain, for all $n \geq N_t^{k+1} \geq N_L$,

$$
\bar{v}_t^{n-1}(\bar{W}^*, \bar{R}^*) \geq L_t^{k+1}(\bar{W}^*, \bar{R}^*) + \delta^k_L - s_{t+1}^{n-1}(\bar{W}^*, \bar{R}^*)
\geq L_t^{k+1}(\bar{W}^*, \bar{R}^*) + \delta^k_L - \delta^k_L
= L_t^{k+1}(\bar{W}^*, \bar{R}^*),
$$

where the last inequality follows from (37).

To finish the proof of part 1, we pick a state $(\bar{W}^*, \bar{R}^*) \in \bar{S}^-$. The reasoning for $U_t^{k+1}(\bar{W}^*, \bar{R}^*)$ is symmetrical to that for $L_t^{k+1}(\bar{W}^*, \bar{R}^*)$, which completes our induction. Thus, we have proved that, for all $k \geq 0$, there exists $N_t^k$ such that (33) and (34) hold for all $n \geq N_t^k$. This concludes the first part of the proof.
Part 2:

In this part, we take care of the states \((\bar{W}^*, \bar{R}^*) \in \bar{S}^+_t \\setminus (\bar{S}^+_t \cap \bar{S}^-_t)\), because if \((\bar{W}^*, \bar{R}^*) \in \bar{S}^+_t \ (\in \bar{S}^-_t)\), we have already proved in part 1 that, for all \(k \geq 0\), there exists \(N^k_t\) such that if \(n \geq N^k_t\), then \(\bar{v}^{-n}_{t}(\bar{W}^*, \bar{R}^*) \geq L^k_t(\bar{W}^*, \bar{R}^*) \leq U^k_t(\bar{W}^*, \bar{R}^*)\). 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. If for all \((\bar{W}^*, \bar{R}^*) \in \bar{S}^+_t\) the corresponding optimal slopes \(v^*_t(\bar{W}^*, \bar{R}^*)\) are distinct, then \(\bar{S}^*_t = \bar{S}^+_t = \bar{S}^-_t\) and Part 2 is not necessary. However, this fact is not verifiable. Figure 5 illustrates a typical situation where \(\bar{S}^+_t \setminus (\bar{S}^+_t \cap \bar{S}^-_t) \neq \emptyset\).

![Figure 5: Optimal slopes that can lead to \(\bar{S}^+_t \setminus (\bar{S}^+_t \cap \bar{S}^-_t) \neq \emptyset\)](image)

An important property of the projection operator is that all the slopes to the left of \(R^t_n\) changed by the projection operation are increased to be equal to the new slope at \(R^t_n\). Similarly, all the slopes to the right of \(R^t_n + 1\) changed by the projection operation are decreased to be equal to the slope at \(R^t_n + 1\) (see figure 3).

There is another interesting property that is necessary for the proof of Part 2. Let \((\bar{W}^*, \bar{R}^*) \in \bar{S}^+_t \setminus \bar{S}^+_t\). We argued in section 4.3 that the state \((\bar{W}^*, R^{Min})\) is an element of \(\bar{S}^+_t\), where \(R^{Min}\) is the minimum asset level of the set \(\{R : (\bar{W}^*, R) \in \bar{S}^+_t\}\). Therefore, the state \((\bar{W}^*, \bar{R}^*_t)\) where \(\bar{R}^*_t\) is the maximum asset level smaller than \(\bar{R}^*\) such that \((\bar{W}^*, \bar{R}^*_t) \in \bar{S}^+_t\) is well defined. We show next that for all asset levels \(R\) between \(\bar{R}^*_t\) and \(\bar{R}^*(\text{inclusive})\),
\(|N^+_i(W^*, R)|\) is also equal to infinity. By definition of the set \(\tilde{S}^-_i\), \(|N^+_i(W^*, R)| = \infty\). If \((\bar{W}^*, \bar{R}^* - 1) = (\bar{W}^*, \bar{R}^*_+ )\) we are done. Otherwise, we have to consider two cases. First, if \((\bar{W}^*, \bar{R}^* - 1) \in \tilde{S}^+_i\), then \(N^+_i(\bar{W}^*, \bar{R}^* - 1)\) is infinite by the definition of the set \(\tilde{S}^+_i\) and from the fact that, in this case, \((\bar{W}^*, \bar{R}^* - 1) \in S^+_i \setminus \tilde{S}^+_i\). Second, if \((\bar{W}^*, \bar{R}^* - 1) \notin \tilde{S}^+_i\), then the corresponding slope is never updated due to a direct observation of sample slopes for \(n \geq \bar{N}\). Moreover, every time the slope of \((\bar{W}^*, \bar{R}^*)\) is decreased due to a projection (which is coming from the left), the slope of \((\bar{W}^*, \bar{R}^* - 1)\) is decreased as well. Therefore, \(N^+_i(\bar{W}^*, \bar{R}^*) \cap \{n \geq \bar{N}\} \subseteq N^+_i(\bar{W}^*, \bar{R}^* - 1) \cap \{n \geq \bar{N}\}\), implying that \(N^+_i(\bar{W}^*, \bar{R}^* - 1)\) is infinite. We then apply the same reasoning for states \((\bar{W}^*, \bar{R}^* - 2), (\bar{W}^*, \bar{R}^* - 3), \ldots,\) until we reach state \((\bar{W}^*, \bar{R}^*_+ )\).

A symmetrical argument handles the states \((\bar{W}^*, \bar{R}^*) \in \tilde{S}^-_i \setminus \tilde{S}^-_i\).

With these properties in mind we go back to the proof of Part 2. We introduce the lemma that is the key element for the proof.

**Lemma 3.** If for all \(k \geq 0\), there exists an integer \(N^k(W, R)\) such that \(v_i^{n-1}(W, R) \geq L_i^k(W, R)\) for all \(n \geq N^k(W, R)\) and \(N^+_i(W, R + 1)\) is infinite, then for all \(k \geq 0\), there exists an integer \(N^k(W, R + 1)\) such that \(v_i^{n-1}(W, R + 1) \geq L_i^k(W, R + 1)\) for all \(n \geq N^k(W, R + 1)\).

Similarly, if for all \(k \geq 0\), there exists an integer \(N^k(W, R)\) such that \(v_i^{n-1}(W, R) \leq U_i^k(W, R)\) for all \(n \geq N^k(W, R)\) and \(N^-_i(W, R - 1)\) is infinite, then for all \(k \geq 0\), there exists an integer \(N^k(W, R - 1)\) such that \(v_i^{n-1}(W, R - 1) \leq U_i^k(W, R - 1)\) for all \(n \geq N^k(W, R - 1)\).

**Proof of lemma** \(\Box\) Given in the appendix.

Pick \(k \geq 0\), \((\bar{W}^*, \bar{R}^*) \in \tilde{S}_i^+ \setminus \tilde{S}_i^+\) and the state \((\bar{W}^*, \bar{R}^*_+ ) \in \tilde{S}_i^+\) introduced in the projection discussion. Note that we can apply lemma 3 considering states \((\bar{W}^*, \bar{R}^*_+ )\) and \((\bar{W}^*, \bar{R}^*_+ + 1)\) in order to obtain, for all \(k \geq 0\), an integer \(N^k(\bar{W}^*, \bar{R}^*_+ + 1)\) such that \(v_i^{n-1}(\bar{W}^*, \bar{R}^*_+ + 1) \geq L_i^k(\bar{W}^*, \bar{R}^*_+ + 1)\), for all \(n \geq (\bar{W}^*, \bar{R}^*_+ + 1)\).

After that, we make use of lemma 3 again, this time to states \((\bar{W}^*, \bar{R}^*_+ + 1)\) and \((\bar{W}^*, \bar{R}^*_+ + 2)\). Note that the first application of lemma 3 gave us the integer \(N^k(\bar{W}^*, \bar{R}^*_+ + 1)\), necessary to fulfill the conditions of this second usage of the lemma. We repeat the same reasoning,
applying lemma 3 successively to the pairs of states \((\bar{W}, \bar{R}_+^t+2)\) and \((\bar{W}, \bar{R}_+^t+3), (\bar{W}, \bar{R}_+^t+3)\) and \((\bar{W}, \bar{R}_+^t+4), \ldots, (\bar{W}, \bar{R}^* - 1)\) and \((\bar{W}, \bar{R}^*)\). In the end, we obtain, for each \(k \geq 0\), an integer \(N^k(\bar{W}, \bar{R}^*)\), such that \(v_t^{n-1}(\bar{W}, \bar{R}^*) \geq L_t^k(\bar{W}, \bar{R}^*)\), for all \(n \geq N^k(\bar{W}, \bar{R}^*)\). Figure 6 illustrates this process.

Similarly, pick \((\bar{W}, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \setminus \bar{\mathcal{S}}_t^-\). By successive applications of the second part of lemma 3 we obtain for each \(k \geq 0\), an integer \(N^k(\bar{W}, \bar{R}^*)\), such that \(v_t^{n-1}(\bar{W}, \bar{R}^*) \leq U_t^k(\bar{W}, \bar{R}^*)\), for all \(n \geq N^k(\bar{W}, \bar{R}^*)\).

Finally, if we consider \(N_t^{*k} = \max \left\{ N_t^k, \max_{(\bar{W}, \bar{R}^*) \in \bar{\mathcal{S}}_t^* \setminus (\bar{\mathcal{S}}_t^* \cap \bar{\mathcal{S}}_t^-)} N^k(\bar{W}, \bar{R}^*) \right\}\), then (31) is true for all states \((\bar{W}, \bar{R}^*) \in \bar{\mathcal{S}}_t^*\) and \(n \geq N_t^{*k}\). Consequently, (32) is also true for all states \((\bar{W}, \bar{R}^*) \in \bar{\mathcal{S}}_t^*\). □

4.5 Optimality of the Decisions

We finish the convergence analysis proving that, with probability one, the algorithm learns an optimal decision for all states that can be reached by an optimal policy.

**Theorem 2.** For \(t = 0, \ldots, T\), let \((R_{t-1}^t, W_t^*, \bar{v}_t^*, x_t^t)\) be an accumulation point of the sequence \(\{(R_{t-1}^n, W_t^n, \bar{v}_t^{n-1}, x_t^n)\}_{n \geq 1}\) generated by the algorithm. Assume all conditions of theorem 1.
are satisfied. Then, with probability one, \( x_t^* \) is an optimal solution of
\[
\max_{x \in \mathcal{X}(R_{t-1}^n, W_t^n)} F_t(R_{t-1}^*, W_t^*, V_t^*, x),
\]
where \( F_t(R_{t-1}^*, W_t^*, V_t^*, x) = C_t(R_{t-1}^*, W_t^*, x) + \gamma V_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x) \).

Proof. The solution \( x_t^n \) in STEP 3a of the algorithm, as it is optimal, satisfies
\[
0 \in \partial F_t(R_{t-1}^n, W_t^n, \bar{v}_t^n, x_t^n) + \mathcal{X}^N(R_{t-1}^n, W_t^n, x_t^n),
\]
where \( \partial F_t(R_{t-1}^n, W_t^n, \bar{v}_t^n, x_t^n) \) is the subdifferential of \( F_t(R_{t-1}^n, W_t^n, \bar{v}_t^n, x_t^n) \) at \( x_t^n \) and \( \mathcal{X}^N(R_{t-1}^n, W_t^n, x_t^n) \) is the normal cone of \( \mathcal{X}(R_{t-1}^n, W_t^n) \) at \( x_t^n \).

Then, by passing to the limit, we can conclude that each accumulation point \((R_{t-1}^*, W_t^*, \bar{v}_t^*, x_t^*) \) of the sequence \( \{(R_{t-1}^n, W_t^n, \bar{v}_t^n, x_t^n)\}_{n \geq 1} \) satisfies the condition
\[
0 \in \partial F_t(R_{t-1}^*, W_t^*, \bar{v}_t^*, x_t^*) + \mathcal{X}^N(R_{t-1}^*, W_t^*, x_t^*).
\]

We now derive an expression for the subdifferential. We have that \( \partial F_t(R_{t-1}^*, W_t^*, V_t^*, x) = \partial C_t(R_{t-1}^*, W_t^*, x) + \gamma \partial \bar{V}_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x) \). Moreover,
\[
\partial C_t(R_{t-1}^*, W_t^*, x) = \nabla C_t(R_{t-1}^*, W_t^*, x) = (c_{t1}(R_{t-1}^*, W_t^*), \ldots, c_{tl}(R_{t-1}^*, W_t^*)).
\]

From [Bertsekas et al., 2003] Proposition 4.2.5, as \( A \in \mathbb{Z}^l \), for \( x \in \mathbb{N}^l \),
\[
\partial \bar{V}_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x) = \{(A_1 y, \ldots, A_l y)^T : y \in [\bar{v}_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x + 1), \bar{v}_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x)]\}.
\]

Therefore, when \( x \in \mathbb{N}^l \),
\[
\partial F_t(R_{t-1}^*, W_t^*, \bar{v}_t^*, x) = \{(c_{t1}(R_{t-1}^*, W_t^*) + \gamma A_1 y, \ldots, c_{tl}(R_{t-1}^*, W_t^*) + \gamma A_l y)^T : y \in [v_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x + 1), v_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x)]\}.
\]

Since \((W_t^*, g(R_{t-1}^*) + A \cdot x_t^*) \) is an accumulation point of \( \{(W_t^n, R_{t-1}^n)\}_{n \geq 0} \), it follows from theorem \( \square \) that
\[
\bar{v}_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x_t^*) = v_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x_t^*) \quad \text{a.s.}
\]

and
\[
\bar{v}_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x_t^* + 1) = v_t^*(W_t^*, g(R_{t-1}^*) + A \cdot x_t^* + 1) \quad \text{a.s.}
\]
Hence, \( \partial F_t(R_{t-1}^*, W_t^*, \bar{V}_t^*, x_t^*) = \partial F_t(R_{t-1}^*, W_t^*, V_t^*, x_t^*) \) and
\[
0 \in \partial F_t(R_{t-1}^*, W_t^*, V_t^*, x_t^*) - \mathcal{X}^N(R_{t-1}^*, W_t^*, x_t^*),
\]
which proves that \( x_t^* \) is the optimal solution of \([42]\) with probability 1. \( \square \)

5 An Inventory System Application

In this section we describe a single asset inventory system and show that this problem is part of the problem class considered in the previous sections. Moreover, we determine an expression for the random variable \( \hat{G}_t \) and prove that all assumptions of theorem 1 are satisfied. Hence, we can apply our algorithm in order to find an optimal policy to this version of a much studied problem.

5.1 Problem Description

At each time period, the inventory manager must determine the number of assets to sell and the number of assets to be purchased in order to replenish the inventory, given the current inventory level, the demand for the asset, the unit selling and replenishing price. The objective is to maximize the expected profits over the horizon \( t = 0, \ldots, T \), when a discount factor \( \gamma \) is considered. There are no holding costs and unsatisfied demand is lost.

Using the notation introduced in section 2, the exogenous process is given by \( \hat{W}_t = (\hat{P}_s^t, \hat{P}_r^t, \hat{D}_t) \), denoting selling price, replenishing price and demand increments, respectively. The information variable \( W_t \) is thus given by \( (P_s^t, P_r^t, D_t) \), the selling price, replenishing price and demand at \( t \). It is not necessary to know the closed form of the transition function \( f \), as long as we assume that it leads to a positive, discrete and bounded set. Moreover, the demand \( D_t \) is assumed to be integer valued. We also assume that \( P_s^t \geq P_r^t > 0 \).

The decision \( x_t = (x_{t1}, x_{t2}) \) is the amount of sold assets and the amount of purchased assets to replenish the inventory, respectively. Assets purchased at time period \( t \) become available at the next time period \( t + 1 \). Available assets not sold at \( t \) remain in the system for \( t + 1 \). Therefore, the inventory level at \( t \) after the decision is made is given by \( R_t = \)
$R_{t-1} - x_{t1} + x_{t2}$, meaning that the function $g$ is the identity function and the input-output vector is given by $A = (-1, 1)$.

We require, at each time period, $x_{t1}$ to be positive, integer and to satisfy $x_{t1} \leq \min(D_t, R_{t-1})$. We also require $x_{t2}$ to be positive, integer and bounded by a constant $M_t$. Hence, the convex constraint set is given by

$$\mathcal{X}(R_{t-1}, W_t) = \{ x \in \mathbb{R}^2 : 0 \leq x_{t1} \leq \min(D_t, R_{t-1}), 0 \leq x_{t2} \leq M_t \}.$$  

implying that $R_t$ is bounded by 0 and $B_t = \sum_{t'=0}^{t} M_{t'}$.

Clearly, the contribution in each period is given by $C_t(R_{t-1}, W_t, x_t) = P_t^r x_{t1} - P_t^r x_{t2}$ and the optimal value functions for $t = 1, \ldots, T$ and $(W,R) \in \mathcal{S}_{t-1}$ are given by

$$V^*_{t-1}(W,R) = E \left[ \max_{x_t \in \mathcal{X}(R,W_t)} C_t(R,W_t,x_t) + \gamma V^*_t(W_t,R-x_{t1}+x_{t2}) \bigg| W_{t-1} = W \right].$$

We also have that $V^*_t(W,R) = 0$ for all $(W,R) \in \mathcal{S}_T$.

### 5.2 Problem Class Properties

The description of the problem made sure that the inventory problem has properties (22)–(25). In order to show properties (26)–(27), we prove the following proposition.

**Proposition 2.** For $t = 1, \ldots, T$, the optimal value functions have integer breakpoints in the asset dimension and $v^*_{t-1}(W) = (v^*_{t-1}(W,1), \ldots, v^*_{t-1}(W, B_{t-1})) \in \mathcal{C}_{t-1}$ for all $W \in \mathcal{W}_{t-1}$, where

$$v^*_{t-1}(W,R) = V^*_{t-1}(W,R) - V^*_{t-1}(W,R-1) = E \left[ \hat{G}_t(R,W_t,V^*_t) \bigg| W_{t-1} = W \right]$$

and

$$\hat{G}_t(R,W_t,v^*_t) = v^*_t(W_t,R+M_t)1_{\{v^*_t(W_t,R+M_t) > P_t^r\}} + P_t^r 1_{\{v^*_t(W_t,R+M_t) \leq P_t^r\}} \left[ 1_{\{D_t \geq R\}} + 1_{\{D_t < R\}'} \right]$$

$$+ \min \left( P_t^r, v^*_t(W_t,R-D_t) \right) 1_{\{D_t < R\}} 1_{\{v^*_t(W_t,R-D_t+M_t) \leq P_t^r\}}$$

$$+ v^*_t(W_t,R-D_t+M_t) 1_{\{D_t < R\}} 1_{\{P_t^r < v^*_t(W_t,R-D_t+M_t) \leq P_t^r\}}.$$  

}\)
Moreover, if we assume that

$$M_t > \min \{ R : \min_{P_t^r} P_t^r > \max_{v_t^i(W_t, R)} v_t^i(W_t, R) \},$$

i.e., the bound on the replenishment decision is big enough, then the random variable $\hat{G}_t(R, W_t, v_t^s)$ simplifies to

$$\hat{G}_t(R, W_t, v_t^s) = P_t^s 1_{\{D_t \geq R\}} + \min (P_t^r, v_t^s(W_t, R - D_t)) 1_{\{D_t < R\}}.$$  (49)

**Proof.** The proof of the proposition is by backward induction on $t$, analyzing the optimal solution $x_t^*$ of the maximization problem inside the expectation in (43). We keep in mind that prices respect the inequality $P_t^r > P_t^r > 0$ and the demand $D_t$ is integer valued.

For $t = 1, \ldots, T$, $R \in \{1, \ldots, B_{t-1}\}$ and $W_t \in W$, let $x_t^* = \arg \max_{x_t \in X(R, W_t)} F_t(R, W_t, V_t^*, x_t)$, where

$$F_t(R, W_t, V_t^*, x_t) = P_t^s x_{t1} - P_t^r x_{t2} + \gamma V_t^s(W_t, R - x_{t1} + x_{t2}).$$

(50)

Moreover, let $\tilde{x}_t^* = \arg \max_{x_t \in X(R-1, W_t)} F_t(R - 1, W_t, V_t^*, x_t)$. Note that

$$\hat{G}_t(R, W_t, v_t^s) = F_t(R, W_t, V_t^*, x_t^*) - F_t(R - 1, W_t, V_t^*, \tilde{x}_t^*).$$

(51)

We start with the base case. Since $V_{T-1}^*(W, R) = 0$ for all states $(W, R) \in S_T$, it is obvious that $x_T^* = \left( \min(\hat{D}_T, R, 0) \right)$ and $\tilde{x}_T^* = \left( \min(\hat{D}_T, R - 1, 0) \right)$. Therefore, for $(W, R) \in S_{T-1}$,

$$V_{T-1}^*(W, R) = \mathbb{E} [P_T^r \min(D_T, R) | W_{T-1} = W].$$

Since $D_T$ is integer valued, $V_{T-1}^*(W, \cdot)$ has integer breakpoints. Moreover, it is clear that

$$G_T(R, W_T, V_T^*) = P_T^s 1_{\{D_T \geq R\}},$$

implying that $G_T(R, W_T, V_T^*)$ is bounded and monotone decreasing in $R$. Therefore, $v_{T-1}^*(W) \in C_{T-1}$. Note that $G_T(R, W_T, V_T^*)$ is indeed the expression in (44)–(47), as $v_T^s$ is equal to zero and the prices are positive.

For all $W \in W$, suppose that $V_{t}^*(W, \cdot)$ has integer breakpoints and $v_t^s(W) \in C_t$. We shall show, for all $W \in W_{t-1}$, that $V_{t-1}^*(W, \cdot)$ has integer breakpoints and $v_{t-1}^s(W) \in C_{t-1}$. 29
Pick \((W, R) \in S_{t-1}\) and a possible realization of \(\hat{W}_t\). Then, \(W_t = (P^s_t, P^r_t, D_t) = f(W, \hat{W}_t)\). We first determine the optimal number of assets to sell, then we determine the optimal number of assets to purchase in order to replenish the inventory. Even though the selling price is always greater than the replenishing price, sometimes it is not optimal to satisfy the demand as much as possible. This is the case as we have a bound \(M_t\) on how much we can order for the next period and the selling price can be smaller than the value of having one more unit of asset in the future.

If \(\gamma v^*_t(W_t, R + M_t) > P^s_t\), then the optimal decision is not to satisfy the demand at all and to replenish the inventory as much as possible. Thus, \(\tilde{x}^t_{t_1} = x^*_t = 0\) and \(\tilde{x}^t_{t_2} = x^*_t = M_t\). Plugging in \(\tilde{x}^*_t\) and \(x^*_t\) in (51), shows (44).

The selling decision is strictly positive, when \(\gamma v^*_t(W_t, R + M_t) \leq P^s_t\). We consider this case to show (45)–(47).

If either \(D_t \geq R\) or \(v^*_t(W_t, R - D_t + M_t) > P^s_t\), the optimal decision does not satisfy the demand completely. Moreover, \(\tilde{x}^t_{t_1} = x^*_t = 0\), and the asset level after the selling decision is the same for both cases, implying that the replenishment decision must be the same for both cases, i.e., \(\tilde{x}^*_t = x^*_t\). Substituting \(\tilde{x}^*_t\) and \(x^*_t\) in (51) shows (45).

On the other hand, if \(D_t < R\) and \(v^*_t(W_t, R - D_t + M_t) \leq P^s_t\) then the whole demand will be satisfied, i.e., \(\tilde{x}^t_{t_1} = x^*_t = D_t\). We have to consider three cases for the replenishment decision and keep in mind that \(v^*_t(W_t) \in C_t\).

(i): If \(\gamma v_t(W_t, R - D_t) \leq P^r_t\) then we do not replenish the inventory, i.e., \(\tilde{x}^t_{t_2} = x^*_t = 0\), as the value of having one more unit of asset in the future is less than the replenishing price.

(ii): If \(\gamma v_t(W_t, R - D_t) > P^r_t\), it is optimal to replenish the inventory in such a way that \(\gamma v_t(W_t, R - D_t + x^*_t) \geq P^r_t\) and \(\gamma v_t(W_t, R - D_t + x^*_t + 1) \leq P^r_t\). Since the post sale inventory levels were \(R - D_t - 1\) and \(R - D_t\), respectively, we have that \(\tilde{x}^t_{t_2} = x^*_t + 1\). We have proved (46).

(iii): Finally, if \(\gamma v^*_t(W_t, R - D_t + M_t) > P^r_t\), then we buy as much as possible and \(\tilde{x}^t_{t_2} =\)
\[ x^*_{t_2} = M_t, \] showing (47).

As \( x^*_t \) is integer valued and \( V^*_t(W_t, \cdot) \) has integer breakpoints, it is not hard to see that \( V^*_{t-1}(W, \cdot) \) has integer breakpoints as well. Furthermore, if \( \bar{v}_t(W) \in C_t \) for all \( W \in \mathcal{W}_t \), it is not hard to see that \( \hat{G}_t(R, W_t, \bar{v}_t) \geq \hat{G}_t(R + 1, W_t, \bar{v}_t) \), as \( P^*_t > P^*_r > 0 \). As a special case, since \( v^*_t(W) \in C_t \) for all \( W \in \mathcal{W}_t \), it follows that \( v^*_{t-1}(W) \in C_{t-1} \) for all \( W \in \mathcal{W}_{t-1} \).

Imposing (48), \( v^*_t(W_t, R - \min(D_t, R) + M_t) \leq v^*_t(W_t, M_t) \leq P^*_r \), implying (49). \( \square \)

### 5.3 Assumptions Verification

We show that the assumptions of theorem 1 are satisfied. The stepsize assumptions are independent of the problem and can be easily enforced. We need to show that the deterministic sequences properties (19)–(21) hold true and the technical assumptions (29)-(30) are true if (28) is true. In order to simplify the proofs, we assume (48) holds true and the random variable \( \hat{G}_t \) is given by (49). With more work, it can also be shown the results are true when \( \hat{G}_t \) is given by (44)-(47).

**Proposition 3.** For \( t = 0, \ldots, T-1 \), states \( (W, R) \in \mathcal{S}_t \) and vectors \( \tilde{v} \) such that \( \tilde{v}_{t+1}(W_{t+1}) \in C_{t+1} \) for all \( W_{t+1} \in \mathcal{W}_{t+1} \), if the random variable \( \hat{G}_t(R, W_t, \tilde{v}_t) \) is given by (49), then (16)–(18) are true. Therefore, using proposition 1, we have that (19)–(21) are true as well.

**Proof.** We have argued that (16) holds true in the proof of the previous proposition.

Clearly, \( \min \left( P^*_t, \bar{v}_t(W_t, R - D_t) \right) \leq \min \left( P^*_t, \tilde{v}_t(W_t, R - D_t) \right) \), if \( \bar{v}_t \leq \tilde{v}_t \), implying (17). Note that we did not have to assume that \( \tilde{v}_t(W_t) \in C_t \). This assumption would have been necessary we were considering that \( \hat{G}_t(R, W_t, \bar{v}_t) \) was given by (44)-(47).

Moreover, we also have that

\[
\min \left( P^*_t, \bar{v}_t(W_t, R - D_t) \right) - \eta \leq \min \left( P^*_t, \bar{v}_t(W_t, R - D_t) - \eta \right) \\
= \min \left( P^*_t, \bar{v}_t(W_t, R - D_t) + \eta \right) \leq \min \left( P^*_t, \bar{v}_t(W_t, R - D_t) + \eta \right),
\]

which implies (18). \( \square \)
Proposition 4. Given \( k \geq 0 \) and \( t = 0, \ldots, T - 1 \), if \( \exists N_t^k \geq \bar{N} \) such that (28) holds for all \( n \geq N_t^k \) and \( (W, R) \in \bar{S}_{t+1} \), then inequalities (29)-(30) hold for all \( n \geq N_t^k \).

Proof. Pick \( n \geq N_t^k > \bar{N} \) and \( \hat{W}_{t+1} \) given \( W_t^n \). We prove (29). The proof of (30) can be obtained using a similar argument. Consider the random variables

\[
\hat{G}^L_{t+1} = \hat{G}_{t+1}(R_t^n, f(W_t^n, \hat{W}_{t+1}), L_t^k) \\
\hat{G}^v_{t+1} = \hat{G}_{t+1}(R_t^n, f(W_t^n, \hat{W}_{t+1}), \hat{v}_{t+1}^{n-1}) \\
\hat{G}^U_{t+1} = \hat{G}_{t+1}(R_t^n, f(W_t^n, \hat{W}_{t+1}), U_t^k),
\]

where \( \hat{G}^L_{t+1}, \hat{G}^v_{t+1} \) and \( \hat{G}^U_{t+1} \) are defined as in (49) replacing \( v_t^* \) by \( L_t^k, \hat{v}_{t+1}^{n-1} \) and \( U_t^k \), respectively. The result follows if we show that \( \hat{G}^L_{t+1} \leq \hat{G}^v_{t+1} \leq \hat{G}^U_{t+1} \).

If \( t = T - 1 \), then \( \hat{G}^L_{t+1} = \hat{G}^v_{t+1} = \hat{G}^U_{t+1} = \hat{G}_{t+1}^a \), as \( L_T(W, R) = \hat{v}_T^{n-1}(W, R) = U_T(W, R) = 0 \) for all \( (W, R) \in \bar{S}_T \).

We now analyze \( t < T - 1 \). Remember that

\[
W_{t+1} = (P_{t+1}^s, P_{t+1}^r, D_{t+1}) = f(W_t^n, \hat{W}_{t+1}).
\]

If \( D_{t+1} \geq R_t^n \), then \( \hat{G}^L_{t+1} = \hat{G}^v_{t+1} = \hat{G}^U_{t+1} = \hat{G}_{t+1}^a \). On the other hand, if \( D_{t+1} < R_t^n \), we have to consider two different cases:

(i): \( (W_{t+1}, R_t^n - D_{t+1}) \in \bar{S}_{t+1}^* \). This case is trivial, as (28) holds for state \( (W_{t+1}, R_t^n - D_{t+1}) \).

Therefore, it is easy to see that \( \hat{G}^L_{t+1} \leq \hat{G}^v_{t+1} \leq \hat{G}^U_{t+1} \).

(ii): \( (W_{t+1}, R_t^n - D_{t+1}) \notin \bar{S}_{t+1}^* \). For this case, we need to rely on the structure of the optimal solution, that is, given our assumption on the bound \( M_{t+1} \), it is always optimal to satisfy the demand. Moreover, for any iteration \( m \), the replenishing decision \( x_{t2}^m \) is such that

\[
\hat{v}_{t+1}^{m-1}(W_{t+1}, R_t^m - D_{t+1} + x_{t2}^m) > P_{t+1}^r \quad \text{and} \quad \hat{v}_{t+1}^{m-1}(W_{t+1}, R_t^m - D_{t+1} + x_{t2}^m + 1) \leq P_{t+1}^r.
\]

Thus, since \( (W_{t+1}, R_t^n - D_{t+1}) \notin \bar{S}_{t+1}^* \), it means that \( \sum_{m=1}^{\infty} 1_{\{\hat{v}_{t+1}^{m-1}(W_{t+1}, R_t^m - D_{t+1} + x_{t2}^m) \leq P_{t+1}^r\}} < \infty \). Thus, by definition of the random variable \( \bar{N} \), it follows that \( \hat{v}_{t+1}^{m-1}(W_{t+1}, R_t^n + \)}
1 − D_{t+1}) > P_{t+1}^r \text{ for all } m \geq \bar{N}, \text{ including the iteration } n. \text{ Since } \bar{v}_{t+1}^{m-1}(W_{t+1}) \in \mathcal{C}_{t+1}, \text{ we have that } \bar{v}_{t+1}^m(W_{t+1}, R_t^n - D_{t+1}) > P_{t+1}^r \text{ as well, implying that } \hat{G}_{t+1}^m = P_{t+1}^r. \text{ Moreover, let } \tilde{R}^* > R_t^n - D_{t+1} \text{ be the maximum asset level such that } \bar{v}_{t+1}^{m-1}(W_{t+1}, \tilde{R}^*) > P_{t+1}^r, \bar{v}_{t+1}^{m-1}(W_{t+1}, \tilde{R}^* + 1) \leq P_{t+1}^r \text{ for some iteration } m \geq N_k^n \text{ and } (W_{t+1}, \tilde{R}^*) \in \bar{S}_{t+1}^*.

Therefore, (28) holds for state \((W_{t+1}, \tilde{R}^*)\) and we have that

\[ P_{t+1}^r < \bar{v}_{t+1}^{m-1}(W_{t+1}, \tilde{R}^*) \leq U_{t+1}^k(W_{t+1}, \tilde{R}^*) \leq U_{t+1}^k(W_{t+1}, R_t^n - D_{t+1}). \]

Thus, \( \min(P_{t+1}^r, L_{t+1}^k(W_{t+1}, R_t^n - D_{t+1})) = \hat{G}_{t+1}^L \leq \hat{G}_{t+1}^m = \hat{G}_{t+1}^U = P_{t+1}^s. \)

\[ \square \]

6 Conclusions

We proposed a pure exploitation approximate dynamic programming algorithm in order to find an optimal policy for a whole class of multistage stochastic single asset problems. The probability distributions of the stochastic processes are not known (not even parametrically) and the problems may suffer from the curse of dimensionality, preventing the use of standard techniques such as backward dynamic programming and Q-learning. The key property of the problem class is that the optimal value functions associated with its problems are concave and piecewise linear with integer breakpoints in the asset dimension. This feature was used extensively both in the design of our algorithm and on the convergence proofs, allowing for the pure exploitation scheme. The algorithm uses monte carlo samples to learn the optimal value function only in important parts of the state space, which are determined by the algorithm itself.

Future research should investigate formally the rate of convergence of the algorithm, although numerical experiments (Nascimento & Powell (2006)) for purchasing forward contracts, which is a problem within our problem class, showed that our algorithm is a very competitive approach, even when the probability distributions are known.

We also described a single asset inventory problem, where the demand, the selling and the replenishing prices are all stochastic and possibly correlated. Moreover, the demand
observations are censored by the inventory level. We showed that this problem is part of the considered problem class and satisfies all the assumptions for the convergence proof. Hence, our algorithm can be used to provide an optimal policy. Even though the inventory problem is much studied, to the best of our knowledge, competing approaches are able to compute optimal policies only through simplifications or some kind of assumption regarding the probability distributions, which is not our case.

References


**A Appendix**

*Proof of proposition* [1] We prove (19)–(21) for the \( \{L^k\}_{k \geq 0} \) sequence. The proof for \( \{U^k\}_{k \geq 0} \) can be obtained using a similar argument.
We start showing (19). Clearly, \( L^k_T(W) \in \mathcal{C}_T \) for all \( k \geq 0 \) and \( W \in \mathcal{W}_T \), since \( L^k_T \) is equal to zero for all possible states. Thus, using assumption (16), we have that \( (HL^k)_{T-1}(W) \in \mathcal{C}_{T-1} \). We keep in mind that, for \( t = 0, \ldots, T-1 \), \( L^0_t(W) \in \mathcal{C}_t \) as \( L^0 = v^* - Be \). By the definition of the \( \{L^k\}_{k \geq 0} \) sequence, we have that \( L^k_{T-1}(W) \in \mathcal{C}_{T-1} \). A simple induction argument shows that \( L^k_{T-1}(W) \in \mathcal{C}_{T-1} \). Using the same argument for \( t = T-2, \ldots, 0 \), we show that \( (HL^k)_t(W) \in \mathcal{C}_t \) and \( L^k_t(W) \in \mathcal{C}_t \).

We now show (20). Since \( v^* \) is the unique fixed point of \( H \), then \( L^0 = Hv^* - 2Be \). Moreover, from assumption (18), we have, for any \( (W, R) \in \mathcal{S}_t \), that \( (Hv^*)_t(W, R) - 2B \leq (H(v^* - 2Be))_t(W, R) = (HL^0)_t(W, R) \). Hence, \( L^0 \leq HL^0 \) and \( L^0 \leq L^1 = \frac{L^0 + HL^0}{2} \leq HL^0 \). Suppose that (20) holds true for some \( k > 0 \). We shall prove for \( k + 1 \). The induction hypothesis and (19) tell us that (17) holds true. Hence, \( HL^{k+1} \geq HL^k \geq L^{k+1} \) and \( HL^{k+1} \geq L^{k+2} \geq L^{k+1} \).

Finally, we show (21). A simple inductive argument shows that \( L^k > v^* \) for all \( k \geq 0 \). Thus, as the sequence is monotone and bounded, it is convergent. Let \( L \) denote the limit. It is clear that \( L_t(W) \in \mathcal{C}_t \). Therefore, assumptions (16)-(18) are also true when applied to \( L \) instead of \( \tilde{v} \). Hence, as shown in (Bertsekas & Tsitsiklis 1996, pages 158-159), we have that

\[
\|HL^k - HL\|_\infty \leq \|L^k - L\|_\infty.
\]

With this inequality, it is straightforward to see that

\[
\lim_{k \to \infty} HL^k = HL.
\]

Therefore, as in the proof of (Bertsekas & Tsitsiklis 1996, Lemma 3.4)

\[
L = L + HL \over 2.
\]

It follows that \( L = v^* \), as \( v^* \) is the unique fixed point of \( H \). □

Each lemma assumes all the conditions imposed and all the results obtained before its statement in the proof of theorem 1. To improve the comprehension of each proof, all the assumptions are presented beforehand.
Proof of lemma 1. Assumptions: Assume stepsize conditions (7)-(9).

Pick \((\hat{W}^*, \hat{R}^*) \in \hat{S}_t^*\). We prove the almost sure convergence to zero of the sequence \(\{\hat{s}_t^n(\hat{W}^*, \hat{R}^*)\}_{n \geq 0}\). The proof for \(\{\hat{s}_t^n(\hat{W}^*, \hat{R}^*)\}_{n \geq 0}\) is symmetrical. In order to simplify notation, let \(\hat{s}_t^n(\hat{W}^*, \hat{R}^*)\) be denoted by \(\hat{s}_t^n\) and \(\bar{\alpha}_t^n(\hat{W}^*, \hat{R}^*)\) be denoted by \(\bar{\alpha}_t^n\). Furthermore, let

\[
\hat{\theta}_{t+1}^n = \hat{s}_{t+1}^n - (R_t^n 1_{\{\hat{R}^* \leq R_t^n\}} + (R_t^n + 1) 1_{\{\hat{R}^* > R_t^n\}}).
\]

We have, for \(n \geq 1\),

\[
(\hat{s}_t^n)^2 \leq \left[ (1 - \bar{\alpha}_t^n) \hat{s}_t^{n-1} + \bar{\alpha}_t^n \hat{\theta}_{t+1}^n \right]^2 = (\hat{s}_t^{n-1})^2 - 2\bar{\alpha}_t^n (\hat{s}_t^{n-1})^2 + A_t^n,
\]

where \(A_t^n = 2\bar{\alpha}_t^n \hat{s}_t^{n-1} \hat{\theta}_{t+1}^n + (\bar{\alpha}_t^n)^2 (\hat{\theta}_{t+1}^n - \hat{s}_t^{n-1})^2\). We want to show that

\[
\sum_{n=1}^{\infty} A_t^n = 2 \sum_{n=1}^{\infty} \bar{\alpha}_t^n \hat{s}_t^{n-1} \hat{\theta}_{t+1}^n + \sum_{n=1}^{\infty} (\bar{\alpha}_t^n)^2 (\hat{\theta}_{t+1}^n - \hat{s}_t^{n-1})^2 < \infty \quad \text{a.s.}
\]

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

\[
\sum_{n=1}^{\infty} (\bar{\alpha}_t^n)^2 (\hat{\theta}_{t+1}^n - \hat{s}_t^{n-1})^2 < \infty \quad \text{a.s.}
\]

Define a new sequence \(\{g_t^n\}_{n \geq 0}\), where \(g_t^0 = 0\) and \(g_t^n = \sum_{m=1}^{n} \bar{\alpha}_t^m \hat{s}_t^{m-1} \hat{\theta}_{t+1}^m\). We can easily check that \(\{g_t^n\}_{n \geq 0}\) is a \(\mathcal{F}^n\)-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 consequently the integrability can be obtained by noticing that (\(g_t^n\)^2) = (\(g_t^{n-1}\)^2) + 2\(g_t^{n-1}\) \(\bar{\alpha}_t^n \hat{s}_t^{n-1} \hat{\theta}_{t+1}^n\) + (\(\bar{\alpha}_t^n\)^2) (\(\hat{s}_t^{n-1} \hat{\theta}_{t+1}^n\)^2). From the martingale equality and boundedness of \(\hat{s}_t^{n-1}\) and \(\hat{\theta}_{t+1}^n\), we get

\[
\mathbb{E} [(g_t^n)^2 | \mathcal{F}^{n-1}] \leq (g_t^{n-1})^2 + C \mathbb{E} [(\bar{\alpha}_t^n)^2 | \mathcal{F}^{n-1}],
\]

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

\[
\mathbb{E} [(g_t^n)^2] \leq \mathbb{E} [(g_t^{n-1})^2] + C \mathbb{E} [(\bar{\alpha}_t^n)^2] \leq \mathbb{E} [(g_t^0)^2] + C \sum_{m=1}^{n} \mathbb{E} [(\bar{\alpha}_t^m)^2] < \infty.
\]
Therefore, the $L^2$-Bounded Martingale Convergence Theorem (Shiryaev 1996, page 510) tells us that
\[
\sum_{n=1}^{\infty} \bar{\alpha}_{t,n}^* \bar{s}_{t,n-1}^* \bar{\theta}_{t+1,n} < \infty \text{ a.s.} \tag{54}
\]

Inequalities (53) and (54) show us that $\sum_{n=1}^{\infty} A_n^t < \infty$ a.s., 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 $-2\bar{\alpha}_{t,n}^* (s_{t,n-1}^*)^2 < 0$, inequality (52) can be rewritten as
\[
(s_{t,n}^*)^2 + \sum_{m=n+1}^{\infty} A_m^t \leq (s_{t,n-1}^*)^2 + \sum_{m=n}^{\infty} A_m^t.
\tag{55}
\]

Thus, the sequence $\{((s_{t,n-1}^*)^2 + \sum_{m=n}^{\infty} A_m^t)\} \geq 0$ is decreasing and bounded from below, as $\sum_{m=1}^{\infty} A_m^t < \infty$. Hence, it is convergent. Moreover, as $\sum_{m=n}^{\infty} A_m^t \to 0$ when $n \to \infty$, we can conclude that $\{s_{t,n}^*\} \geq 0$ converges almost surely.

Finally, as inequality (52) holds for all $n \geq 1$, it yields
\[
(s_{t,n}^*)^2 \leq (s_{t,n-1}^*)^2 - 2\bar{\alpha}_{t,n}^* (s_{t,n-1}^*)^2 + A_n^t
\]
\[
\leq (s_{t,n-2}^*)^2 - 2\bar{\alpha}_{t,n-1}^* (s_{t,n-2}^*)^2 + A_{n-1}^t - 2\bar{\alpha}_{t,n}^* (s_{t,n-2}^*)^2 + A_n^t
\]
\[
\vdots
\]
\[
\leq -2 \sum_{m=1}^{n} \bar{\alpha}_{t,m}^* (s_{t,m-1}^*)^2 + \sum_{m=1}^{n} A_m^t.
\]

Passing to the limits we obtain:
\[
\limsup_{n \to \infty} (s_{t,n}^*)^2 + 2 \sum_{m=1}^{\infty} \bar{\alpha}_{t,m}^* (s_{t,m-1}^*)^2 \leq \sum_{m=1}^{\infty} A_m^t < \infty \text{ a.s.}
\]

This implies, together with the convergence of $\{s_{t,n}^*\}$, that $\sum_{m=1}^{\infty} \bar{\alpha}_{t,m}^* (s_{t,m-1}^*)^2 < \infty$ almost surely. On the other hand, stepsize assumption (9) tells us that $\sum_{m=1}^{\infty} \bar{\alpha}_{t,m}^* = \infty$ almost surely. Hence, there must exist a subsequence of $\{s_{t,n}^*\}$ that converges to zero almost surely. Therefore, as every subsequence of a convergent sequence converges to its limit, it follows that $\{s_{t,n}^*\} \geq 0$ converges to zero almost surely. \qed
Proof of lemma 2. Assumptions: Given \( t = 0, \ldots, T - 1, k \geq 0 \) and integer \( N_t^k \), assume for all \( n \geq N_t^k \), that (33) and (34) hold true.

The proof is by induction on \( n \). Let \((\bar{W}^*, \bar{R}^*)\) be in \( \bar{S}_t^- \) or in \( \bar{S}_t^+ \). The proof for the base case \( n = N_t^k \) is immediate from the fact that \( \bar{s}_{t-}^{N_t^k-1}(\bar{W}^*, \bar{R}^*) = \bar{s}_{t+}^{N_t^k-1}(\bar{W}^*, \bar{R}^*) = 0 \), \( \bar{u}_{t-}^{N_t^k-1}(\bar{W}^*, \bar{R}^*) = L_t^k(\bar{W}^*, \bar{R}^*) \) and \( \bar{u}_{t+}^{N_t^k-1}(\bar{W}^*, \bar{R}^*) = U_t^k(\bar{W}^*, \bar{R}^*) \) and by the assumption that (33) and (34) hold true for \( n \geq N_t^k \). Now suppose true for \( n \) and we need to prove for \( n + 1 \).

In order to simplify the notation, let \( \bar{\alpha}_t^n(\bar{W}^*, \bar{R}^*) \) be denoted by \( \bar{\alpha}_t^n \) and \( \bar{v}_t^n(\bar{W}^*, \bar{R}^*) \) be denoted by \( \bar{v}_t^n \). We use the same shorthand notation for \( z_t^n(\bar{W}^*, \bar{R}^*), \bar{u}_t^n(\bar{W}^*, \bar{R}^*), \bar{s}_{t-}^n(\bar{W}^*, \bar{R}^*), \bar{s}_{t+}^n(\bar{W}^*, \bar{R}^*) \) as well.

We start by considering the states \((\bar{W}^*, \bar{R}^*)\) that are in \( \bar{S}_t^+ \). By the construction of this set, for all \((\bar{W}^*, \bar{R}^*) \in \bar{S}_t^+ \), the set of iterations \( \mathcal{N}_t^+(\bar{W}^*, \bar{R}^*) \) is finite and for all \( n \geq N_t^k \geq N \), \( \bar{v}_t^n(\bar{W}^*, \bar{R}^*) \geq z_t^n(\bar{W}^*, \bar{R}^*) \). We have to consider three different cases:

(i): \( \bar{W}^* = W_t^n \) and \( \bar{R}^* = R_t^n \).

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

\[
\bar{v}_t^n \geq z_t^n = (1 - \bar{\alpha}_t^n)\bar{v}_t^{n-1} + \bar{\alpha}_t^n \bar{v}_{t+1}^n(R_t^n)
\geq (1 - \bar{\alpha}_t^n)(\bar{\bar{l}}_t^{n-1} - \bar{s}_t^{n-1}) + \bar{\alpha}_t^n \bar{s}_{t+1}^n(R_t^n)
- \bar{\alpha}_t^n (H^n)_{t}(W_t^n, R_t^n)
+ \bar{\alpha}_t^n (H^n)_{t}(W_t^n, R_t^n)
\geq (1 - \bar{\alpha}_t^n)(\bar{\bar{l}}_t^{n-1} - \bar{s}_t^{n-1}) - \bar{\alpha}_t^n \bar{s}_{t+1}^n(R_t^n)
+ \bar{\alpha}_t^n (H \bar{L}^k)_{t}(W_t^n, R_t^n)
= \bar{\bar{l}}_t^n - ((1 - \bar{\alpha}_t^n)\bar{s}_t^{n-1} + \bar{\alpha}_t^n \bar{s}_{t+1}^n(R_t^n))
\geq \bar{l}_t^n - (\max(0, (1 - \bar{\alpha}_t^n)\bar{s}_t^{n-1} + \bar{\alpha}_t^n \bar{s}_{t+1}^n(R_t^n)))
= \bar{l}_t^n - \bar{s}_t^n.
\]

The first inequality is due to the construction of set \( \bar{S}_t^+ \), while (56) is due to the induction hypothesis. Inequalities (29) and (30) for \( n \geq N_t^k \) explains (57). Finally, (58) and (59) come from the definition of the stochastic sequences \( \bar{l}_t^n \) and \( \bar{s}_t^n \), respectively.

(ii): \( \bar{W}^* = W_t^n \) and \( \bar{R}^* = R_t^n + 1 \).

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 \((W_t^n, R_t^n + 1)\) component, instead of \((W_t^n, R_t^n)\).

(iii): Else.

Here the state \((\tilde{W}^*, \tilde{R}^*)\) 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,

\[
\overline{l}_{t+1}^n = \overline{l}_t^n - \tilde{s}_{t+1}^n.
\]

Therefore, from the construction of set \(\tilde{S}_t^+\) and the induction hypothesis

\[
\overline{v}_{t+1}^n \geq \tilde{z}_t^n = \overline{v}_t^n - \overline{l}_t^n = \overline{l}_t^n - \overline{s}_{t+1}^n.
\]

The considerations for states \((\tilde{W}^*, \tilde{R}^*)\) that are in \(\tilde{S}_t^-\) are symmetrical. We just need to use the reversed inequality and the sequences \(\{\overline{u}_m^n\} m \geq 0\) and \(\{\overline{s}_m^n\} m \geq 0\).

\[\square\]

Proof of lemma 3. Assumptions: Assume stepsize conditions (7)-(9). Moreover, assume for all \(k \geq 0\) and integer \(N^{k}_t\) that inequalities (29) and (30) hold true for all \(n \geq N^{k}_t\).

We prove the first statement. The second one is symmetrical. We start by showing, for each \(k \geq 0\), there exists an integer \(N^{k,s}_t\) such that \(\overline{v}_t^{n-1}(W, R + 1) \geq L_k^1(W, R + 1) - \overline{s}_{t+1}^n(W, R + 1)\) for all \(n \geq N^{k,s}_t\). Then, we show, for all \(\epsilon > 0\), there is an integer \(N^{k,\epsilon}_t\) such that \(\overline{v}_t^{n-1}(W, R + 1) \geq L_k^1(W, R + 1) - \epsilon\) for all \(n \geq N^{k,\epsilon}_t\). Finally, using these results, we prove existence of an integer \(N^k(W, R + 1)\) such that \(\overline{v}_t^{n-1}(W, R + 1) \geq L_k^1(W, R + 1)\) for all \(n \geq N^k(W, R + 1)\).

Pick \(k \geq 0\). Let \(N^{k,s}_t = \min \{n \in N^{+}_{t}(W, R + 1) : n \geq N^k(W, R)\} + 1\), where \(N^k(W, R)\) is the integer such that \(\overline{v}_t^{n-1}(W, R) \geq L_k^1(W, R)\) for all \(n \geq N^k(W, R)\). Since \(N^{+}_{t}(W, R + 1)\) is infinite, \(N^{k,s}_t\) is well defined. Note that \(\overline{v}_t^{N^{k,s}_t-1}(W, R) = \tilde{v}_t^{N^{k,s}_t-1}(W, R + 1)\). Redefine the noise sequence \(\{\overline{s}_m^n\} m \geq 0\) introduced in the proof of theorem 1 using \(N^{k,s}_t\) instead of \(N^k_t\). We prove that \(\overline{v}_t^{n-1}(W, R + 1) \geq L_k^1(W, R + 1) - \overline{s}_{t+1}^n(W, R + 1)\) for all \(n \geq N^{k,s}_t\) by induction on \(n\).

For the base case \(n - 1 = N^{k,s}_t - 1\), from our choice of \(N^{k,s}_t\) and the monotone decreasing
property of $L^k$, we have that

$$\bar{v}^{n-1}_t(W, R+1) = \bar{v}^{n-1}_t(W, R) \geq L^k_t(W, R) \geq L^k_t(W, R + 1) = L^k_t(W, R + 1) - \bar{s}^{n-1}_{t+1}(W, R + 1).$$

Now, we suppose $\bar{v}^{n-1}_t(W, R + 1) \geq L^k_t(W, R + 1) - \bar{s}^{n-1}_{t+1}(W, R + 1)$ is true for $n - 1 > N^k,s_t$ and prove for $n$. We have to consider two cases:

(i): $n \in \mathcal{N}_t^+(W, R + 1)$

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

$$\bar{v}^{n-1}_t(W, R + 1) = \bar{v}^{n-1}_t(W, R) \geq L^k_t(W, R) \geq L^k_t(W, R + 1) \geq L^k_t(W, R + 1) - \bar{s}^{n-1}_{t+1}(W, R + 1).$$

(ii): $n \notin \mathcal{N}_t^+(W, R + 1)$

The analysis of this case is analogous to the proof of inequalities (40) and (41) of lemma 2 for $(W, R + 1) \in \tilde{S}^+$. The difference is that we consider $L^k_t$ instead of the stochastic bounding sequence $\{\bar{l}^{n}_t\}_{m \geq 0}$ and we note that

$$(1 - \bar{\alpha}^n_t(W, R + 1))L^k_t(W, R + 1) + \bar{\alpha}^n_t(W, R + 1)(HL^k)_t(W, R + 1) \geq L^k_t(W, R + 1).$$

Hence, we have proved that for all $k \geq 0$ there exists an integer $N^k,s_t$ such that

$$\bar{v}^{n-1}_t(W, R + 1) \geq L^k_t(W, R + 1) - \bar{s}^{n-1}_{t+1}(W, R + 1) \text{ for all } n \geq N^k,s_t.$$

We move on to show, for all $k \geq 0$ and $\epsilon > 0$, there is an integer $N^{k,\epsilon}_t$ such that

$$\bar{v}^{n-1}_t(W, R + 1) \geq L^k_t(W, R + 1) - \epsilon \text{ for all } n \geq N^{k,\epsilon}_t. \quad \text{(1)}$$

We have to consider two cases: (i) $(W, R + 1) \in \tilde{S}^*_t$ and (ii) $(W, R + 1) \notin \tilde{S}^*_t$. For the first case, lemma 1 tells us that $\bar{s}^n_{t+1}(W, R + 1)$ goes to zero. Then, there exists $N^\epsilon > 0$ such that $\bar{s}^n_{t+1}(W, R + 1) < \epsilon$ for all $n \geq N^\epsilon$. Therefore, we just need to choose $N^{k,\epsilon}_t = \max(N^k,s_t, N^\epsilon)$. For the second case, $\bar{\alpha}^n_t(W, R + 1) = 0$ for all $n \geq N^k,s_t$ and $\bar{s}^{N^k,s_t-1}_{t+1}(W, R + 1) = 0$. Thus, $\bar{s}^n_{t+1}(W, R + 1) = \bar{s}^{N^k,s_t-1}_{t+1}(W, R + 1) = 0$ for all $n \geq N^k,s_t$ and we just have to choose $N^{k,\epsilon}_t = N^k,s_t$.

We are ready to conclude the proof. For that matter, we use the result of the previous paragraph. Pick $k > 0$. Let $\epsilon = v^*_t(W, R + 1) - L^k_t(W, R + 1) > 0$. Since $\{L^k\}_{k \geq 0}$ increases to
there exists \( k' > k \) such that \( v^* - L^k_t(W, R+1) < \epsilon/2 \). Thus, \( L^{k'}_t(W, R+1) - L^k_t(W, R+1) > \epsilon/2 \) and the result of the previous paragraph tells us that there exists \( N^k_{t, \epsilon/2} \) such that \( \bar{v}_{t-1}^{n-1}(W, R+1) \geq L^{k'}_t(W, R+1) - \epsilon/2 > L^k_t(W, R+1) + \epsilon/2 - \epsilon/2 = L^k_t(W, R+1) \) for all \( n \geq N^k_{t, \epsilon/2} \). Therefore, we just need to choose \( N^k_t(W, R+1) = N^k_{t, \epsilon/2} \) and we have proved that for all \( k \geq 0 \), there exists \( N^k_t(W, R+1) \) such that \( \bar{v}_{t-1}^{n-1}(W, R+1) \geq L^k_t(W, R+1) \) for all \( n \geq N^k_t(W, R+1) \). 
\( \Box \)