An Optimal Approximate Dynamic Programming Algorithm for the Economic Dispatch Problem with Grid-Level Storage

Juliana M. Nascimento and Warren B. Powell [1]

January 12, 2012

[1]Department of Operations Research and Financial Engineering, Princeton University

Abstract

We prove convergence of an approximate dynamic programming algorithm for a class of high-dimensional stochastic control problems linked by a scalar storage device. Our problem is motivated by the problem of optimizing hourly dispatch and energy allocation decisions in the presence of grid-level storage. The model makes it possible to capture hourly, daily and seasonal variations in wind, solar and demand, while also modeling the presence of hydro-electric storage to smooth energy production over both daily and seasonal cycles. The problem is formulated as a stochastic, dynamic program, where we approximate the value of stored energy using a piecewise linear value function approximation. We provide a rigorous convergence proof for an approximate dynamic programming algorithm, which can capture the presence of both the amount of energy held in storage as well as other exogenous variables. Our algorithm exploits the natural concavity of the problem to avoid any need for explicit exploration policies.

We propose an approximate dynamic programming algorithm that is characterized by multi-dimensional (and potentially high-dimensional) controls, but where each time period is linked by a single, scalar storage device. Our problem is motivated by the problem of allocating energy resources over both the electric power grid (generally referred to as the economic dispatch problem) as well as other forms of energy allocation (conversion of biomass to liquid fuels, conversion of petroleum to gasoline or electricity, and the use of natural gas in home heating or electric power generation). Determining how much energy can be converted from each source to serve each type of demand can be modeled as a series of single-period linear programs linked by a single, scalar storage variable, as would be the case with grid-level storage. The problem is described in detail in Powell et al. (2011); here, we present the convergence proof for the algorithm used in Powell et al. (2011).

The problem is described as follows. Let R_t be the scalar quantity of stored energy on hand, and let W_t be a discrete, vector-valued (but low dimensional) stochastic process describing exogenously varying parameters such as available energy from wind and solar, demand of different types (with hourly, daily and weekly patterns), energy prices and rainfall. We assume that W_t is Markovian to be able to model processes such as wind, prices and demand where the future (say, the price P_{t+1} at time $t+1$), depends on the current price P_t plus an exogenous change $\hat{P}_{t+1}$, allowing us to write $P_{t+1} = P_t + \hat{P}_{t+1}$. If E_t is the wind at time t, P_t is a price (or vector of prices), and D_t is the demand, we would let $W_t = (E_t, P_t, D_t)$, where the future of each stochastic process depends on the current value. The state of our system, then, is given by $S_t = (R_t, W_t)$.

In our problem, x_t is a vector-valued control giving the allocation of different sources of energy (oil, coal, wind, solar, nuclear, etc.) over different energy pathways (conversion to electricity, transmission over the electric power grid, conversion of biomass to liquid fuel) to satisfy different types of demands. The decision x_t also includes the amount of energy stored in an aggregate, grid-level storage device such as a reservoir for hydroelectric power.

An optimal policy is described by Bellman's equation

$$V_t(S_t) = \max_{x_t \in \mathcal{X}_t} \left(C(S_t, x_t) + \gamma \mathbb{E} \left\{ V_{t+1}(S_{t+1}) | S_t \right\} \right) \tag{1}$$

where $S_{t+1} = f(S_t, x_t, \xi_{t+1})$ and $C(S_t, x_t)$ is a contribution function that is linear in x_t, but where the feasible region $\mathcal{X}_t$ is determined by a set of linear inequalities.

We are going to assume that W_t is a set of discretized random variables, which includes exogenous

energy supply (such as wind and solar), demand, prices, rainfall and any other uncertain parameters. It is possible to show that $V_t(S_t) = V_t(R_t, W_t)$ can be written as a piecewise linear function in R_t for any given value of W_t. We treat x_t as a continuous variable, but we will show that the optimal policy involves solving a linear program which returns an optimal x_t that is also defined on a discrete set of outcomes (the extreme point solutions of the linear program). In this case, we can ensure that the solution x_t also takes on discrete values, but we solve the optimization problem as a continuous problem using a linear programming solver. Thus, we can view x_t and R_t as continuous, while ensuring that they always take on discrete values.

Although S_t may have only five or ten dimensions, this is often enough to render equation (1) computationally intractable. There is an extensive literature in approximate dynamic programming and reinforcement learning which assumes either discrete action spaces, or low dimensional, continuous controls (see Bertsekas and Tsitsiklis (1996), Sutton and Barto (1998), Szepesvari (2010) and Bertsekas (2012*a*)). For our problem, x_t can easily have hundreds or thousands of dimensions, depending on the level of aggregation. In addition, we are unable to compute the expectation analytically, introducing the additional complication of maximizing an expectation that is itself computationally intractable. The problem of vector-valued state, outcome and action spaces is known as the three curses of dimensionality (Powell (2011)).

Current proofs of convergence for approximate dynamic programming algorithms (see, for example, Tsitsiklis (1994), Jaakkola et al. (1994), Borkar and Meyn (2000), Gordon (2001), Precup and Perkins (2003), Bertsekas, Abounadi and Borkar (2003), Szita (2007), Antos et al. (2007), Munos and Szepesvari (2008), Antos et al. (2008*a*), Antos et al. (2008*b*)) assume discrete action spaces and as a result require some form of explicit exploration of the action space, strategies that would never scale to our problems (see Szepesvari (2010) for a concise but modern review). Bertsekas (2011) provides a nice discussion of the challenges of exploration, and highlights problems when this is not handled properly. A convergence proof for a Real Time Dynamic Programming algorithm (Barto et al., 1995) that considers a pure exploitation scheme is provided in Bertsekas and Tsitsiklis (1996) [Prop. 5.3 and 5.4], but it assumes that expected values can be computed and the initial approximations are optimistic, which produces a type of forced exploration. We make no such assumptions, but it is important to emphasize that our result depends on the concavity of the optimal value functions.

Our strategy represents a form of approximate value iteration where we focus on finding the slopes of a function rather than the value (similar to the concept of "dual heuristic dynamic pro-

gramming" first proposed by Werbos (1989)). Approximate value iteration is well known to have poor convergence (Bertsekas (2012*b*)), and can diverge, but we exploit concavity of the value function, and this allows us to design an algorithm that works very well in practice (Godfrey and Powell (2002), Topaloglu and Powell (2006), Enders et al. (2010), He et al. (2010)) and has worked quite well on the specific problem considered in this paper (Powell et al. (2011)). However, this work has not been supported by any formal convergence results. This paper presents the first formal convergence proof.

Others have taken advantage of concavity (convexity for minimization) in the stochastic control literature. Hernandez-Lerma and Runggladier (1994) presents a number of theoretical results for general stochastic control problems for the case with i.i.d. noise (an assumption we do not make), but does not present a specific algorithm that could be applied to our problem class. There is an extensive literature on stochastic control which assumes linear, additive noise (see, for example, Bertsekas (2005) and the references cited there). For our problem, noise arises in costs and, in particular, the constraints.

Other approaches to deal with our problem class would be different flavors of Benders decomposition (Van Slyke and Wets (1969), Higle and Sen (1991), Chen and Powell (1999)) and sample average approximation (SAA) given in Shapiro (2003). A thorough and modern view of these methods is given in Shapiro et al. (2009). However, techniques based on stochastic programming such as Benders require that we create scenario trees to capture the trajectory of information. We are interested in modeling our problem over a full year to capture seasonal variations, producing a problem with 8,760 time periods. 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 and Shapiro (2002).

This paper provides a convergence proof for an algorithm that is applied to the economic dispatch problem in Powell et al. (2011). Our proof technique builds on the convergence proof in Nascimento and Powell (2009) (henceforth referred to as N&P in the remainder of this paper), which in turn combines the convergence proof for a monotone mapping in Bertsekas and Tsitsiklis (1996) with the proof of the SPAR algorithm in Powell et al. (2004). The proof in N&P, however, is developed for a much simpler problem. Specifically, it assumes a) that x_t is a scalar in the set $\{0, 1, \ldots, M\}$, b) the state variable $S_t = (R_t, P_t)$ where P_t is a price, c) a contribution function $C(S_t, x_t) = -P_t x_t$ that

is linear over the entire feasible region, and d) a transition function $R_{t+1} = R_t + x_t$. The proof in N&P exploits these properties throughout the proof.

The experimental work in Powell et al. (2011) showed a) the algorithm produces results that closely match the optimal solution from a linear programming solver when applied to a deterministic model, b) it appears to converge with 100 iterations, even on a stochastic model, 3) the algorithm allows the resource variable R_t to be discretized arbitrarily finely since breakpoints are generated on the fly, and 4) the algorithm is tested on problems where x_t has 200 and 20,000 dimensions, on a problem with 8,760 time periods (hourly increments over a year). The algorithm as presented is limited to problems where W_t is discretized into a relatively small number of outcomes, but we report on recent research in machine learning that has the potential for dramatically expanding the complexity of W_t. Thus, we have shown that the algorithm has fast convergence and scales to very large-scale problems; in this paper we show that it also works in theory.

The paper is organized as follows. Section 1 opens by presenting the basic model. Section 2 gives the optimality equations and describes some properties of the value function. Section 3 describes the algorithm in detail. Section 4 gives the convergence proof, which is the heart of the paper. Section 5 concludes the paper.

1 The model

Our problem is depicted in figure 1, which is described in more detail in Powell et al. (2011). We assume there are two types of flows. The first are intra-time period flows, representing the flow of power from generator to consumer, which are assumed to happen at time period t. Included are flows into and out of a storage unit. The second type of flow is the energy held in storage from one time period to the next. Also linking time periods is information about prices, demands, and exogenous energy generation sources such as wind and solar.

The problem can be modeled as a generic stochastic, dynamic program with state $S_t = (R_t, W_t) \in \mathcal{S}$, where R_t is a scalar describing the amount of energy in storage at time t, and $W_t \in \mathcal{W}$ is the state of our information process at time t which, in our energy application, consists of the price of electricity P_t, the demand for electricity D_t and the current level of exogenous energy generation E_t (say, the speed of wind). We assume that the support $\mathcal{W}$ is discrete, and while x_t is continuous, we are going to show that we can ensure that R_t will only take on discrete values, allowing us to model

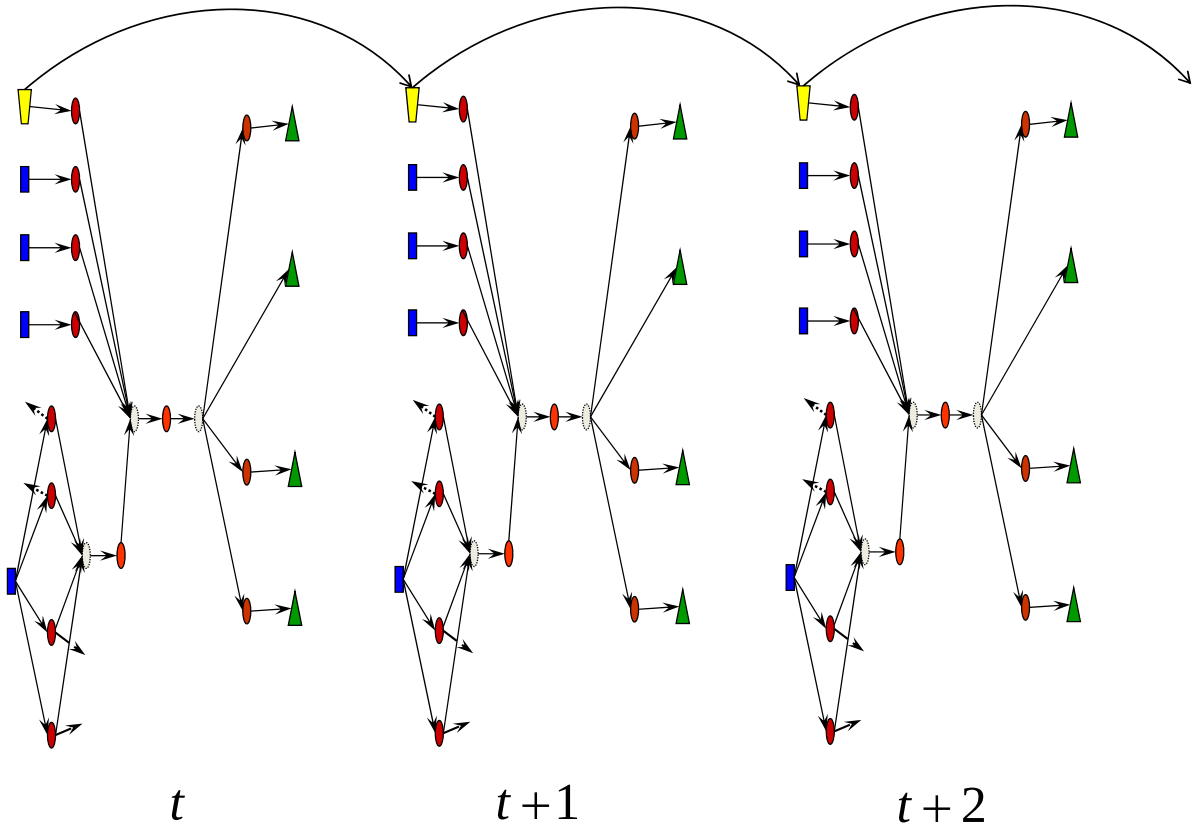


Figure 1: Illustration of time-staged networks linked by storage.

$\mathcal{S}$ as being discrete.

The state evolves under the influence of a decision vector x_t and exogenous information ξ_t. The vector x_t includes both intra-time period flows, as well as the holding of energy in storage from one time period to the next. We represent the feasible region using $x_t \in \mathcal{X}_t(R_t, W_t)$ defined by

$$\begin{aligned} Ax_t &= b_t(R_t, W_t), \\ x_t &\leq u_t(W_t), \\ x_t &\geq 0, \end{aligned}$$

where $b_t(R_t, W_t)$ is a vector of right hand side constraints, where one reflects the amount held in storage (we allow the right hand sides to depend on W_t). We describe transitions using the state transition model, represented by

$$S_{t+1} = S^M(S_t, x_t, \xi_{t+1}).$$

Let ω be a sample realization of $\xi = (\xi_1, \xi_2, \ldots, \xi_T)$, and let Ω be the (finite) set of all possible sample realizations. Let $\mathcal{F}$ be the sigma-algebra on Ω), with filtrations

$$\mathcal{F}_t = \sigma(\xi_1, \ldots, \xi_t).$$

Finally, we define our probability space $(\Omega, \mathcal{F}, \mathcal{P})$ where $\mathcal{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Throughout we use the convention that any variable indexed by t is $\mathcal{F}_t$-measurable.

The state transition model is defined as follows. The evolution of the energy in storage is given by

$$R_{t+1} = R_t + A^S x_t + \hat{R}_{t+1},$$

where $\hat{R}_{t+1}$ represents exogenous changes in storage occurring between t and $t+1$ (this might represent rainfall into a water reservoir). The storage incidence (row) vector A^S captures the elements of x_t that represent moving energy into or out of storage. Variables such as P_t, D_t and E_t evolve according to

$$\begin{aligned} P_{t+1} &= P_t + \hat{P}_{t+1}, \\ D_{t+1} &= D_t + \hat{D}_{t+1}, \\ E_{t+1} &= E_t + \hat{E}_{t+1}. \end{aligned}$$

Our exogenous information, then, is given by $\xi_t = (\hat{R}_t, \hat{P}_t, \hat{D}_t, \hat{E}_t)$.

Throughout our presentation, we use the concept of the post-decision state denoted S_t^x, which is the state of the system at time t, immediately after we have made a decision. S_t^x is given by

$$S_t^x = (R_t^x, W_t),$$

where R_t^x is the post-decision resource state given by

$$R_t^x = f^x(R_t, x_t) = R_t + A^S x_t.$$

Throughout our presentation, we primarily represent the state as (R_t, W_t) rather than the more compact S_t because of the need to represent the effect of x_t, which only impacts R_t.

We seek to maximize a contribution $C(S_t, x_t)$ which we assume is a linear function of the form

$$C(S_t, x_t) = c(W_t)x_t.$$

We need to emphasize, however, that while our cost function is linear in x_t, the effect of the constraints allows us to easily model piecewise linear functions.

We let $X_t^\pi(S_t)$ be a decision function (or a policy) that returns a feasible decision vector x_t given the information in S_t. Our policy is time dependent (rather than being a single function that depends on S_t) because we are solving a nonstationary, finite-horizon problem. We denote the family of policies by $(X^\pi(S_t)),\ \pi \in \Pi$. Our goal is to find the optimal policy defined by

$$\max_{\pi\in\Pi} F^\pi(S_0) = \mathbb{E}\sum_{t=0}^{T} \gamma^t C(S_t, X_t^\pi(S_t)), \tag{2}$$

where the discount factor γ may be equal to 1.

If the matrix A is unimodular, and if $b(R_t, W_t)$ and $u_t(W_t)$ are integer, then it is well known that the solution to the corresponding single-period linear program produce an optimal x_t^* that is integer (Ahuja et al., 1993). If $b(R_t, W_t)$ and $u_t(W_t)$ can be written as an integer times a scaling factor, then the same is true of x_t^*. In this case, the value function can be discretized on the integers. A backward induction proof shows that this is true for all time periods. If the matrix A is not unimodular, but if $b(R_t, W_t)$ and $u_t(W_t)$ can be written as an integer times a scaling factor, then the optimal solution of the single-period linear program can still be represented by the extreme points, which means the optimal value of the LP is piecewise linear (but not necessarily on integer breakpoints). Since $\mathcal{W}$ is discrete and finite, then there is a finite set of extreme points. If we can find the greatest common divisor of the breakpoints corresponding to each value of $\mathcal{W}$, then we can discretize the value function into a set of breakpoints $\{1, \ldots, B^R\}$ which can be represented by an integer times a scaling factor that we denote by ρ. While the scaling factor ρ may be very small, these can be generated on the fly (as is done in Powell et al. (2011)), making the algorithmic implementation independent of the level of discretization. For notational simplicity, we assume that the breakpoints are evenly spaced, but this is not required by the algorithm or the proof, where we could replace the discretization parameter ρ with $\rho(r)$ which gives the length of the rth interval.

Our convergence proof builds on, but is a significant generalization of, the lagged asset acquisition problem in N&P, so it is useful to compare the two models. The lagged asset acquisition problem is

	N&P	This paper
Decision x_t	Discrete, scalar	Continuous vector
Constraints	Simple bounds	Linear polytope
Exogenous process	Scalar price P_t	Low dimensional vector W_t
Contribution function	Linear	Constrained linear
Resource transition	$R_t + x_t$	$R_t + A^S x_t + \hat{R}_{t+1}$

Table 1: Comparison of the lagged asset acquisition model in Nascimento and Powell (2009) and the model in this paper.

described by a scalar resource variable R_t (similar to our paper), a scalar price variable P_t (analogous to W_t in our model), and a scalar decision x_t that determines how many assets to purchase at time t to be used at the end of horizon T, where $x_t \in \{0, 1, \ldots, M\}$. In the lagged asset acquisition problem, the contribution function was given by

$$C(S_t, x_t) = -P_t x_t,$$

which is linear over the entire feasible region. We note that the contribution function $C(S_t, x_t)$ in this paper is also linear, but subject to $x_t \in \mathcal{X}_t$, which allows us to represent nonlinear (piecewise linear) behavior. In N&P, the resource variable R_t evolved according to $R_{t+1} = R_t + x_t$ making it possible to use a fairly coarse discretization of R_t. A comparison of the two models is given in table 1.

Our problem is motivated by the grid-level storage application that arises in economic dispatch, but the problem is more general. Other applications arise in physical distribution, where we pick up or deliver goods each day, and the days are connected only by the inventory of product. Financial companies have to collect and distribute cash for different investments, while managing the cash on hand (the inventory) from one time period to the next. The Red Cross has to manage inventories of blood, collecting donations and distributing supplies, where the activities over a time period (a day or a week) are linked only by the inventory of blood. Even the energy applications can come in different flavors, such as storing chemical energy in a battery or potential energy in a pumped-hydro facility.

2 The Optimal Value Functions

We define, recursively, the optimal value functions associated. We denote by $V_t^*(R_t, W_t)$ the optimal value function around the pre-decision state (R_t, W_t) and by $V_t^x(R_t^x, W_t)$ the optimal value function around the post-decision state (R_t^x, W_t). Using this notation, Bellman's equation can be broken into two steps as follows:

$$V_t^*(R_t, W_t) = \max_{x_t \in \mathcal{X}_t(R_t, W_t)} \big(C_t(R_t, W_t, x_t) + \gamma V_t^x(R_t^x, W_t)\big). \quad (3)$$

$$V_t^x(R_t^x, W_t) = \mathbb{E}\left[V_{t+1}^*(R_{t+1}, W_{t+1})|(R_t^x, W_t)\right], \quad (4)$$

We note that the optimization problem in (3) is deterministic. When we replace $V_t^x(R_t^x, W_t)$ with a piecewise linear approximation $\bar{V}_t(R_t^x, W_t)$, equation (3) becomes a deterministic linear program. This is the essential step that allows us to handle high-dimensional decision variables x_t, with hundreds to tens of thousands of dimensions. The goal of this paper is to design a sequence of piecewise linear approximations that converge to an optimal policy.

In the remainder of the paper, we only use the value function $V_t^x(S_t^x)$ defined around the post-decision state variable, since it allows us to make decisions by solving a deterministic problem as in (4). We show that $V_t^x(R, W_t)$ is concave and piecewise linear in R with B^R discrete breakpoints. This structural property combined with the optimization/expectation inversion is the foundation of our algorithmic strategy and its proof of convergence.

For $W_t \in \mathcal{W}_t$, let $v_t(W_t) = (v_t(1, W_t), \ldots, v_t(B^R, W_t))$ be a vector representing the slopes of a function $V_t(R, W_t) : [0, \infty) \to \Re$ that is concave and piecewise linear with breakpoints $R = 1, \ldots, B^R$. Using the piecewise linear value function approximation, our decision problem at time t is given by

$$F_t^*\left(v_t(W_t), R_t, W_t\right) = \max_{x_t, y_t} C_t(R_t, W_t, x_t) + \gamma \sum_{r=1}^{B^R} v_t(r, W_t) y_{tr}, \quad (5)$$

subject to

$$A_t x_t = b_t(R_t, W_t), \tag{6}$$

$$x_t \leq u_t(W_t), \tag{7}$$

$$x_t \geq 0, \tag{8}$$

$$\sum_{r=1}^{B^R} y_{tr}\rho = f^x(R_t, x_t), \tag{9}$$

$$0 \leq y_{tr} \leq \rho, \quad r \in \{0, 1, \ldots, B^R.\} \tag{10}$$

In this formulation, we are representing the value function explicitly as a piecewise linear function with slopes $v_t(r, W_t)$ where r denotes the segment, and y_{tr} captures the amount of flow (bounded by the scaling factor ρ) allocated to a segment. Constraint (9) ensures that the sum over all y_{tr} adds up to the post-decision resource value. Concavity ensures that we will always allocate flow to segments with the highest values of $v_t(r, W_t)$. Our model assumes that the breakpoints are evenly spaced, but this is purely for notational simplicity. We could have unequal intervals, in which case we simply replace ρ with $\rho(r)$.

It is easy to see that the function $F_t^*(v_t(W_t), R_t, W_t)$ is concave and piecewise linear with breakpoints $R = 1, \ldots, B^R$, representing the extreme points of the linear program. Moreover, the optimal solution (x_t^*, y_t^*) to the linear programming problem that defines $F_t^*(v_t(W_t), R_t, W_t)$ does not depend on $V_t(0, W_t)$ (that is, adding a constant to the value function does not change the optimal solution). We also have that $F_t^*(v_t(W_t), R_t, W_t)$ is bounded for all $W_t \in \mathcal{W}_t$.

The linear program represented by (5)-(10) scales to very large scale problems. In Powell et al. (2011), one problem has a 200-dimensional decision vector x_t with 175,000 time periods; a second problem has a vector x_t with 20,000 dimensions (this problem was solved with 8,760 time periods - hourly intervals over a year). Also, if the matrix A_t is unimodular (which would occur if it corresponds to network flows) and the right hand sides of the constraints are all integer (or an integer times a scaling factor), then the optimal solution (x_t, y_t) will also be integer (possibly times a scaling factor). In this case, we can guarantee that there exists a set of breakpoints for the value function approximation where we are assured that the optimal solution will fall on one of the breakpoints.

We use F_t^* to prove the following proposition about the optimal value function.

Proposition 1 *For $t = 0, \ldots, T$ and information vector $W_t \in \mathcal{W}_t$, the optimal value function*

$V_t^x(R, W_t)$ is concave and piecewise linear with breakpoints $R = 1, \ldots, B^R$. We denote its slopes by $v_t^*(W_t) = \big(v_t^*(1, W_t), \ldots, v_t^*(B^R, W_t)\big)$, where, for $R = 1, \ldots, B^R$ and $t < T$, $v_t^*(R, W_t)$ is given by

$$\begin{aligned} v_t^*(R, W_t) &= V_t^x(R, W_t) - V_t^x(R-1, W_t) \\ &= \mathbb{E}\big[F_{t+1}^*(v_{t+1}^*(W_{t+1}), R + \hat{R}_{t+1}(W_{t+1}), W_{t+1}) \\ &\quad - F_{t+1}^*(v_{t+1}^*(W_{t+1}), R - 1 + \hat{R}_{t+1}(W_{t+1}), W_{t+1}) | W_t\big]. \end{aligned} \tag{11}$$

Proof: The proof is by backward induction on t. The base case $t = T$ holds as $V_T^x(R_T, W_T)$ is equal to zero for all $W_T \in \mathcal{W}_T$. For $t < T$ the proof is obtained noting that

$$V_t^x(R_t^x, W_t) = \mathbb{E}\big[\gamma V_{t+1}^x(0, W_{t+1}) + F_t^*(v_{t+1}^*(W_{t+1}), R_t^x + \hat{R}_{t+1}(W_{t+1}), W_{t+1}) | W_t\big].$$

Due to the concavity of $V_t^x(\cdot, W_t)$, the slope vector $v_t^*(W_t)$ is monotone decreasing, that is, $v_t^*(R, W_t) \geq v_t^*(R + \rho, W_t)$ (throughout, we use $v_t^*(W_t)$ to refer to the slope of the value function $V_t^x(R_t^x, W_t)$ defined around the post-decision state variable). Moreover, throughout the paper, we work with the translated version $V_t^*(\cdot, W_t)$ of $V_t^x(\cdot, W_t)$ given by $V_t^*(R + y, W_t) = \sum_{r=1}^{R} v_t^*(r, W_t) + y v_t^*(R + \rho, W_t)$, where R is nonnegative and $0 \leq y \leq \rho$, since the optimal solution (x_t^*, y_t^*) associated with $F^*(v_{t+1}^*(W_{t+1}), R, W_{t+1})$ does not depend on $V_t^x(0, W_t)$.

Following Bertsekas and Tsitsiklis (1996), we next introduce the dynamic programming operator H associated with the storage class. We define H using the slopes of piecewise linear functions instead of the functions themselves.

Let $v = \{v_t(W_t) \text{ for } t = 0, \ldots, T, \quad W_t \in \mathcal{W}_t\}$ be a set of slope vectors, where $v_t(W_t) = (v_t(1, W_t), \ldots, v_t(B^R), W_t)$. The dynamic programming operator H associated with the storage class maps a set of slope vectors v into a new set Hv as follows. For $t = 0, \ldots, T-1$, $W_t \in \mathcal{W}_t$ and $R = 1, \ldots, B^R$,

$$\begin{aligned} (Hv)_t(R, W_t) &= \mathbb{E}\big[F_t^*(v_{t+1}(W_{t+1}), R + \hat{R}_{t+1}(W_{t+1}), W_{t+1}) \\ &\quad - F_t^*(v_{t+1}(W_{t+1}), R - 1 + \hat{R}_{t+1}(W_{t+1}), W_{t+1}) | W_t\big]. \end{aligned} \tag{12}$$

It is well known that the optimal value function in a dynamic program is unique. This is a trivial result for the finite horizon problems that we consider in this paper. Therefore, the set of slopes v^*

corresponding to the optimal value functions $V_t^*(R, W_t)$ for $t = 0, \ldots, T$, $W_t \in \mathcal{W}_t$ and $R = 1, \ldots, B_t$ are unique.

Let $\tilde{v} = \{\tilde{v}_t(W_t) \text{ for } t = 0, \ldots, T, \quad W_t \in \mathcal{W}_t\}$ and $\tilde{\tilde{v}} = \{\tilde{\tilde{v}}_t(W_t) \text{ for } t = 0, \ldots, T, W_t \in \mathcal{W}_t\}$ be sets of slope vectors such that

$$\tilde{v}_t(W_t) \quad = \quad (\tilde{v}_t(1, W_t), \ldots, \tilde{v}_t(B_t, W_t))$$

and

$$\tilde{\tilde{v}}_t(W_t) \quad = \quad (\tilde{\tilde{v}}_t(1, W_t), \ldots, \tilde{\tilde{v}}_t(B_t, W_t))$$

are monotone decreasing and $\tilde{v}_t(W_t) \leq \tilde{\tilde{v}}_t(W_t)$. Our proof builds on the theory in Bertsekas and Tsitsiklis (1996), which makes the assumption that the dynamic programming operator H defined by (12) is assumed to satisfy the following conditions for $W_t \in \mathcal{W}_t$ and $R = 1, \ldots, B_t$:

$$(H\tilde{v})_t(W_t) \qquad \text{is monotone decreasing in } \tilde{v}, \tag{13}$$

$$(H\tilde{v})_t(R, W_t) \quad \leq \quad (H\tilde{\tilde{v}})_t(R, W_t), \tag{14}$$

$$(H\tilde{v})_t(R, W_t) - \eta e \quad \leq \quad (H(\tilde{v} - \eta e))_t(R, W) \leq (H(\tilde{v} + \eta e))_t(R, W) \leq (H\tilde{v})_t(R, W) + \eta e, \tag{15}$$

where η is a positive integer and e is a vector of ones. Conditions (13) and (15) imply that the mapping H is continuous (see the discussion in Bertsekas and Tsitsiklis (1996), page 158). The dynamic programming operator H and the associated conditions (13)-(15) are used later on to construct deterministic sequences that are provably convergent to the optimal slopes. These assumptions are used in proposition 2 below. The proof (given in the appendix) establishes that they are satisfied.

3 The SPAR-Storage Algorithm

We propose a pure exploitation algorithm, namely the SPAR-Storage Algorithm, that provably learns the optimal decisions to be taken at parts of the state space that can be reached by an optimal policy, which are determined by the algorithm itself. This is accomplished by learning the slopes of the optimal value functions at important parts of the state space, through the construction of value function approximations $\bar{V}_t^n(\cdot, W_t)$ that are concave, piecewise linear with breakpoints $R = 1, \ldots, B^R$. The approximation is represented by its slopes $\bar{v}_t^n(W_t) = (\bar{v}_t^n(1, W_t), \ldots,$

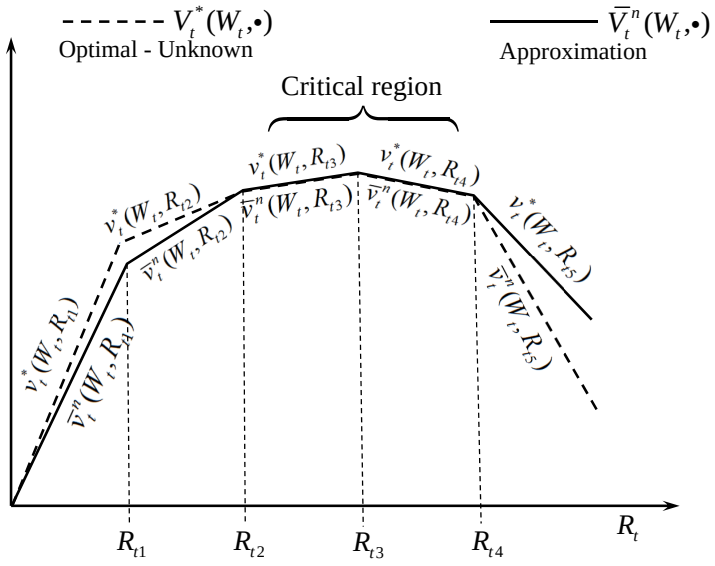


Figure 2: Optimal value function and the constructed approximation

$\bar{v}_t^n(B^R, W_t)\big)$. Figure 2 illustrates the exact (but unknown) and approximate value function approximation, where the two will match (in the limit) only in the vicinity of the optimum. The algorithm combines Monte Carlo simulation in a pure exploitation scheme and stochastic approximation integrated with a projection operation.

Figure 3 describes the SPAR-Storage algorithm. The algorithm requires an initial concave piecewise linear value function approximation $\bar{V}^0(W_t, \cdot)$, represented by its slopes $\bar{v}_t^0(W_t) = \big(\bar{v}_t^0(1, W_t), \ldots,$ $\bar{v}_t^0(B^R, W_t)\big)$, for each information vector $W_t \in \mathcal{W}_t$. Therefore the initial slope vector $\bar{v}_t^0(W_t)$ has to be monotone decreasing. For example, it is valid to set all the initial slopes equal to zero. For completeness, since we know that the optimal value function at the end of the horizon is equal to zero, we set $\bar{v}_T^n(R, W_T) = 0$ for all iterations n, information vectors $W_T \in \mathcal{W}_T$ and asset levels $R = 1, \ldots, B^R$. The algorithm also requires an initial asset level to be used in all iterations. Thus, for all $n \geq 0$, $R_{-1}^{x,n}$ is set to be a nonnegative value, as in STEP 0b.

At the beginning of each iteration n, the algorithm observes a sample realization of the information sequence $W_0^n, \ldots, W_T^n$, as in STEP 1. The sample can be obtained from a sample generator or actual data. After that, the algorithm goes over time periods $t = 0, \ldots, T$.

First, the pre-decision asset level R_t^n is computed, as in STEP 2. Then, the decision x_t^n, which is optimal with respect to the current pre-decision state (R_t^n, W_t^n) and value function approximation

---

STEP 0: Algorithm Initialization:

STEP 0a: Initialize $\bar{v}_t^0(W_t)$ for $t = 1, \ldots, T-1$ and $W_t \in \mathcal{W}_t$ monotone decreasing.

STEP 0b: Set $R_{-1}^{x,n} = \bar{r} = k\rho$ for some $k \geq 0$, for all $n \geq 0$

STEP 0c: Set $n = 1$.

STEP 1: Sample/Observe the information sequence $W_0^n, \ldots, W_T^n$.

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

STEP 2: Compute the pre-decision asset level: $R_t^n = R_{t-1}^{x,n} + \hat{R}_t(W_t^n)$.

STEP 3: Find the optimal solution x_t^n by solving the linear program

$$\max_{x_t \in \mathcal{X}_t(R_t^n, W_t^n)} C_t(R_t^n, W_t^n, x_t) + \gamma \bar{V}_t^{n-1}\left(f^x(R_t^n, x_t)\right).$$

STEP 4: Compute the post-decision asset level:

$$R_t^{x,n} = f^x(R_t^n, x_t^n).$$

STEP 5: Update slopes:

If $t < T$ **then**

STEP 5a: Observe $\hat{v}_{t+1}^n(R_t^{x,n})$ and $\hat{v}_{t+1}^n(R_t^{x,n} + \rho)$. See (16).

STEP 5b: For $W_t \in \mathcal{W}_t$ and $R = 1, \ldots, B^R$:

$$z_t^n(R, W_t) = (1 - \bar{\alpha}_t^n(R, W_t))\bar{v}_t^{n-1}(R, W_t) + \bar{\alpha}_t^n(R, W_t)\hat{v}_{t+1}^n(R).$$

STEP 5c: Perform the projection operation $\bar{v}_t^n = \Pi_{\mathcal{C}}(z_t^n)$. See (20).

STEP 6: Increase n by one and go to step 1.

Figure 3: SPAR-Storage Algorithm

$\bar{V}_t^{n-1}(\cdot, W_t^n)$ is taken, as stated in STEP 3. We have that $\bar{V}_t^n(R + y, W_t) = \sum_{r=1}^{R} \bar{v}_t^n(r, W_t) + y\bar{v}_t^n(R + \rho, W_t)$, where R is a nonnegative value and $0 \leq y \leq \rho$. Next, taking into account the decision, the algorithm computes the post-decision asset level $R_t^{x,n}$, as in STEP 4.

Time period t is concluded by updating the slopes of the value function approximation. Steps 5a-5c describes the procedure. Sample slopes relative to the post-decision states $(R_t^{x,n}, W_t^n)$ and $(R_t^{x,n} + \rho, W_t^n)$ are observed in STEP 5a. After that, these samples are used to update the approximation slopes $\bar{v}_t^{n-1}(W_t^n)$, through a temporary slope vector $z_t^n(W_t^n)$. This procedure requires the use of a stepsize rule that is state dependent, denoted by $\bar{\alpha}_t^n(R, W_t)$, and it may lead to a violation of the property that the slopes are monotonically decreasing, see STEP 5b. Thus, a projection operation

$\Pi_{\mathcal{C}}$ is performed to restore the property and updated slopes $\bar{v}_t^n(W_t^n)$ are obtained in STEP 5c.

After the end of the planning horizon T is reached, the iteration counter is incremented, as in STEP 6, and a new iteration is started from STEP 1.

We note that x_t^n is easily computed by solving the linear program (5)-(10). Moreover, given our assumptions and the properties of F_t^*, it is clear that R_t^n, x_t^n and $R_t^{x,n}$ are all limited to an extreme point solution of the linear program. We also know that they are bounded. Therefore, the sequences of decisions, and the pre- and post- decision states generated by the algorithm, given by $\{x_t^n\}_{n\geq 0}$, $\{S_t^n\}_{n\geq 0} = \{(R_t^n, W_t^n)\}_{n\geq 0}$ and $\{S_t^{x,n}\} = \{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$, respectively, have at least one accumulation point. Since these are sequences of random variables, their accumulation points, denoted by x_t^*, S_t^* and $S_t^{x,*}$, respectively, are also random variables.

The sample slopes used to update the approximation slopes are obtained by replacing the expectation and the slopes $v_{t+1}^*(W_{t+1})$ of the optimal value function in (11) by a sample realization of the information W_{t+1}^n and the current slope approximation $\bar{v}_{t+1}^{n-1}(W_{t+1}^n)$, respectively. Thus, for $t = 1\ldots, T$, the sample slope is given by

$$\begin{aligned}\hat{v}_{t+1}^n(R) = {} & F_t^*(\bar{v}_{t+1}^{n-1}(W_{t+1}^n), R + \hat{R}_{t+1}(W_{t+1}^n), W_{t+1}^n) \\ & - F_t^*(\bar{v}_{t+1}^{n-1}(W_{t+1}^n), R - 1 + \hat{R}_{t+1}(W_{t+1}^n), W_{t+1}^n). \end{aligned} \tag{16}$$

The update procedure is then divided into two parts. First, a temporary set of slope vectors $z_t^n = \{z_t^n(R, W_t) : W_t \in \mathcal{W}_t, R = 1, \ldots, B^R\}$ is produced combining the current approximation and the sample slopes using the stepsize rule $\bar{\alpha}_t^n(R, W_t)$. We have that

$$\bar{\alpha}_t^n(R, W_t) = \alpha_t^n 1_{\{W_t = W_t^n\}}(1_{\{R = R_t^{x,n}\}} + 1_{\{R = R_t^{x,n} + \rho\}}),$$

where α_t^n is a scalar between 0 and 1 and can depend only on information that became available up until iteration n and time t. Moreover, on the event that $(R_t^{x,*}, W_t^*)$ is an accumulation point of $\{(R_t^{x,n}), W_t^n\}_{n\geq 0}$, we make the standard assumptions that

$$\sum_{n=1}^{\infty} \bar{\alpha}_t^n(R_t^{x,*}, W_t^*) = \infty \text{ a.s.}, \tag{17}$$

$$\sum_{n=1}^{\infty} (\bar{\alpha}_t^n(R_t^{x,*}, W_t^*))^2 \leq B^\alpha < \infty \text{ a.s.}, \tag{18}$$

where B^α is a constant. Clearly, the rule $\alpha_t^n = \frac{1}{N^n(R_t^{x,*}, W_t^*)}$ satisfies all the conditions, where $N^n(R_t^{x,*}, W_t^*)$ is the number of visits to state $(R_t^{x,*}, W_t^*)$ up until iteration n. Furthermore, for all

positive integers N,

$$\prod_{n=N}^{\infty} \left(1 - \bar{\alpha}_t^n(R_t^{x,*}, W_t^*)\right) = 0 \text{ a.s.} \tag{19}$$

The proof for (19) follows directly from the fact that $\log(1+x) \leq x$.

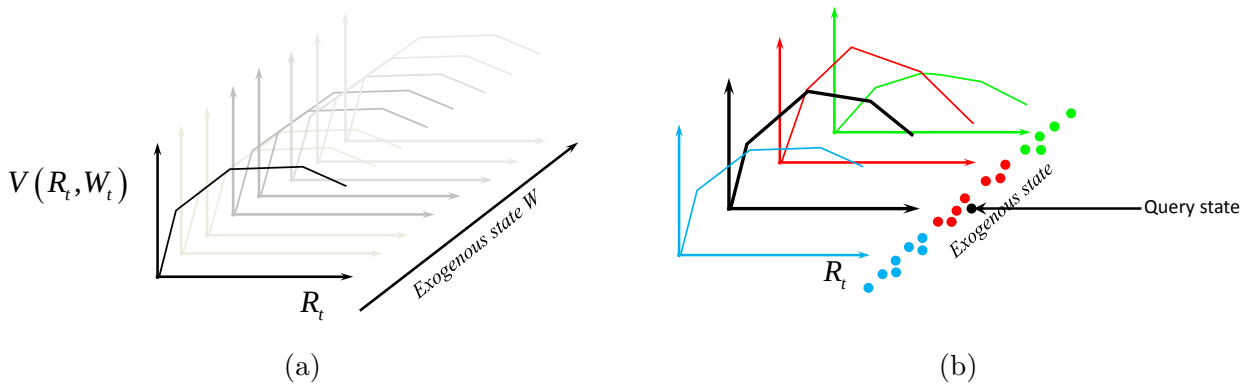
The second part is the projection operation, where the temporary slope vector $z_t^n(W_t)$, that may not be monotone decreasing, is transformed into another slope vector $\bar{v}_t^n(W_t)$ that has this structural property. The projection operator imposes the desired property by simply forcing the violating slopes to be equal to the newly updated ones. For $W_t \in \mathcal{W}_t$ and $R = 1, \ldots, B^R$, the projection is given by

$$\Pi_{\mathcal{C}}(z_t^n)(R, W_t) = \begin{cases} z_t^n(R_t^{x,n}, W_t^n), & \text{if } W_t = W_t^n, \quad R < R_t^{x,n}, \quad z_t^n(R, W_t) \leq z_t^n(R_t^{x,n}, W_t^n) \\ z_t^n(R_t^{x,n} + \rho, W_t^n), & \text{if } W_t = W_t^n, \quad R > R_t^{x,n} + \rho, \\ & \quad z_t^n(R, W_t) \ \geq \ z_t^n(R_t^{x,n} + \rho, W_t^n) \\ z_t^n(R, W_t), & \text{otherwise.} \end{cases} \tag{20}$$

Let the sequence of slopes of the value function approximation generated by the algorithm be denoted by $\{\bar{v}_t^n(R, W_t)\}_{n \geq 0}$. Moreover, as the function $F_t^*(\bar{v}_t^{n-1}(W_t^n), R_t^n, W_t^n)$ is bounded and the stepsizes α_t^n are between 0 and 1, we can easily see that the sample slopes $\hat{v}_t^n(R)$, the temporary slopes $z_t^n(R, W_t)$ and, consequently, the approximated slopes $\bar{v}_t^n(R, W_t)$ are all bounded. Therefore, the slope sequence $\{\bar{v}_t^n(R, W_t)\}_{n \geq 0}$ has at least one accumulation point, as the projection operation guarantees that the updated vector of slopes are elements of a compact set. The accumulation points are random variables and are denoted by $\bar{v}_t^*(R, W_t)$, as opposed to the deterministic optimal slopes $v_t^*(R, W_t)$.

The ability of the SPAR-storage algorithm to avoid visiting all possible values of R_t was significant in our energy application. In our energy model in Powell et al. (2011), R_t was the energy stored in a hydroelectric reservoir which served as a source of storage for the entire grid. The algorithm required that we discretize R_t into approximately 10,000 elements. However, the power of concavity requires that we visit only a small number of these a large number of times for each value of W_t. We found that even when R_t was discretized into tens of thousands of increments, we obtained very high quality solutions in approximately 100 iterations.

An important practical issue is that we effectively have to visit the entire support of W_t infinitely often. If W_t contains a vector of as few as five or ten dimensions, this can be computationally expensive. In a practical application, we would use statistical methods to produce more robust



(a) (b)

Figure 4: (a) Set of piecewise linear, concave functions for each value of W_t. (b) Piecewise linear value functions built around clusters, using interpolation to estimate the value function at a general query state.

estimates using smaller numbers of observations. For example, we could use the hierarchical aggregation strategy suggested by George et al. (2008) to produce estimates of $V(R, W)$ at different levels of aggregation of W. These estimates can then be combined using a simple weighting formula.

Recently, Hannah et al. (2010) proposed the idea of estimating piecewise linear value functions around clusters, using a mixture based on Dirichlet processes. The idea works as follows. Imagine that if we enumerate the entire support $\mathcal{W}$ of W_t, we would have to estimate an extremely large number of piecewise linear functions depicted in figure 4(a). An alternative is to organize subsets of $\mathcal{W}$ into clusters $W^1, \ldots, W^k$ where k is the number of clusters, producing a much smaller number of functions as depicted in figure 4(b). If we are in a state W (the query state), we can use a weighted sum of piecewise linear value functions estimated around the clustered points. Hannah et al. (2010) proves that such an approach can produce an asymptotically unbiased estimate of the value function for any W using this strategy.

It is beyond the scope of this paper to analyze the convergence properties of these algorithms, but it is important to realize that these strategies are available to overcome the discretization of W_t.

4 Convergence Analysis

We start this section by presenting the convergence results we want to prove. The major result is the almost sure convergence of the approximation slopes corresponding to states that are visited

infinitely often. Substantial portions of the proof follow the reasoning first presented in N&P, but modified to reflect the more general setting of our problem class. In this section, we provide only the detailed proof of convergence to the optimal solution, which is fundamentally different because of our use of continuous, vector-valued decisions.

On the event that $(R_t^{x,*}, W_t^*)$ is an accumulation point of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$, we obtain

$$\bar{v}_t^n(R_t^{x,*}, W_t^*) \to v_t^*(R_t^{x,*}, W_t^*) \quad \text{and} \quad \bar{v}_t^n(R_t^{x,*} + \rho, W_t^*) \to v_t^*(R_t^{x,*} + \rho, W_t^*) \quad \text{a.s.}$$

As a byproduct of the previous result, we show that, for $t = 0, \ldots, T$, on the event that (R_t^*, W_t^*, x_t^*) is an accumulation point of $\{(R_t^n, W_t^n, x_t^n)\}_{n\geq 0}$,

$$x_t^* = \arg \max_{x_t \in \mathcal{X}_t(R_t^*), W_t^*} C_t(R_t^*, W_t^*, x_t) + \gamma V_t^* \left(f^x(R_t^*, x_t), W_t^*\right) \quad \text{a.s.} \tag{21}$$

where $V_t^*(\cdot, W_t^*)$ is the translated optimal value function.

Equation (21) implies that the algorithm has learned almost surely an optimal decision for all states that can be reached by an optimal policy. This implication can be easily justified as follows. Pick ω in the sample space. We omit the dependence of the random variables on ω for the sake of clarity. For $t = 0$, since $R_{-1}^{x,n} = \bar{r}$, a given constant, for all iterations of the algorithm, we have that $R_{-1}^{x,*} = \bar{r}$. Moreover, all the elements in $\mathcal{W}_0$ are accumulation points of $\{W_0^n\}_{n\geq 0}$, as $\mathcal{W}_0$ has finite support. Thus, (21) tells us that the accumulation points x_0^* of the sequence $\{x_0^n\}_{n\geq 0}$ along the iterations with pre-decision state $(R_{-1}^{x,*} + \hat{R}_0(W_0^*), W_0^*)$ are in fact an optimal policy for period 0 when the information is W_0^*. This implies that all accumulation points $R_0^{x,*} = f^x(R_{-1}^{x,*} + \hat{R}_0(W_0^*), x_0)$ of $\{R_0^{x,n}\}_{n\geq 0}$ are post-decision resource 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 $\{W_1^n\}_{n\geq 0}$. Hence, (21) tells us that the accumulation points x_1^* of the sequence $\{x_1^n\}$ along iterations with $(R_1^n, W_1^n) = (R_{-0}^{x,*} + \hat{R}_0(W_1^*), W_1^*)$ are indeed an optimal policy for period 1 when the information is W_1^* and the pre-decision resource level is $R_1^* = R_{-0}^{x,*} + \hat{R}_0(W_1^*)$. As before, the accumulation points $R_1^{x,*} = f^x(R_1^*, x_1)$ of $\{R_1^{*,n}\}_{n\geq 0}$ are post-decision resource levels that can be reached by an optimal policy. The same reasoning can be applied for $t = 2, \ldots, T$.

4.1 Outline of the Convergence Proofs

Our proof follows the style of N&P, which builds on the ideas presented in Bertsekas and Tsitsiklis (1996) and in Powell et al. (2004). Bertsekas and Tsitsiklis (1996) proves convergence assuming that

all states are visited infinitely often. The authors do not consider a concavity-preserving step, which is the key element that has allowed us to obtain a convergence proof when a pure exploitation scheme is considered. As a result, their algorithmic strategy would never scale to vector-valued decisions. Although the framework in Powell et al. (2004) also considers the concavity of the optimal value functions in the resource dimension, the use of a projection operation to restore concavity and a pure exploitation routine, their proof is restricted to two-stage problems. The difference is significant. In Powell et al. (2004), it was possible to assume that Monte Carlo estimates of the true value function were unbiased, a critical assumption in that paper. In our paper, estimates of the marginal value of additional resource at time t depends on a value function approximation at $t+1$. Since this is an approximation, the estimates of marginal values at time t are biased.

The main concept to achieve the convergence of the approximation slopes to the optimal ones is to construct deterministic sequences of slopes, namely, $\{L^k(R,W_t)\}_{k\geq 0}$ and $\{U_t^k(R,W_t)\}_{k\geq 0}$, that are provably convergent to the slopes of the optimal value functions. These sequences are based on the dynamic programming operator H, as introduced in (12). We then use these sequences to prove almost surely that for all $k \geq 0$,

$$L_t^k(R_t^{x,*},W_t^*) \leq \bar{v}_t^n(R_t^{x,*},W_t^*) \leq U_t^k(R_t^{x,*},W_t^*), \tag{22}$$

$$L_t^k(R_t^{x,*}+1,W_t^*) \leq \bar{v}_t^n(R_t^{x,*}+1,W_t^*) \leq U_t^k(R_t^{x,*}+\rho,W_t^*), \tag{23}$$

on the event that the iteration n is sufficiently large and $(R_t^{x,*},W_t^*)$ is an accumulation point of $\{(R_t^{x,n},W_t^n)\}_{n\geq 0}$, which implies the convergence of the approximation slopes to the optimal ones.

Establishing (22) and (23) requires several intermediate steps that need to take into consideration the pure exploitation nature of our algorithm and the concavity preserving operation. We give all the details in the proof of $L_t^k(R_t^{x,*},W_t^*) \leq \bar{v}_t^n(R_t^{x,*},W_t^*)$ and $L_t^k(R_t^{x,*}+\rho,W_t^*) \leq \bar{v}_t^n(R_t^{x,*}+\rho,W_t^*)$. The upper bound inequalities are obtained using a symmetrical argument.

First, we define two auxiliary stochastic sequences of slopes, namely, the noise and the bounding sequences, denoted by $\{\bar{s}_t^n(R,W_t)\}_{n\geq 0}$, and $\{\bar{l}_t^n(R,W_t)\}_{n\geq 0}$, respectively. The first sequence represents the noise introduced by the observation of the sample slopes, which replaces the observation of true expectations and the optimal slopes. The second one is a convex combination of the deterministic sequence $L_t^k(R,W_t)$ and the transformed sequence $(HL^k)_t(R,W_t)$.

We then define the set $\tilde{\mathcal{S}}_t$ to contain the states $(R_t^{x,*},W_t^*)$ and $(R_t^{x,*}+\rho,W_t^*)$, such that $(R_t^{x,*},W_t^*)$ is an accumulation point of $\{(R_t^{x,n},W_t^n)\}_{n\geq 0}$ and the projection operation decreased or kept the same

the corresponding unprojected slopes infinitely often. This is not the set of all accumulation points, since there are some points where the slope may have increased infinitely often.

The stochastic sequences $\{\bar{s}_t^n(R,W_t)\}$, and $\{\bar{l}_t^n(R,W_t)\}$ are used to show that on the event that the iteration n is big enough and $(\tilde{R}_t^{x,*},W_t^*)$ is an element of the random set $\tilde{\mathcal{S}}_t$,

$$\bar{v}_t^{n-1}(\tilde{R}_t^{x,*},W_t^*) \geq \bar{l}_t^{n-1}(\tilde{R}_t^{x,*},W_t^*) - \bar{s}_t^{n-1}(\tilde{R}_t^{x,*},W_t^*) \quad \text{a.s.}$$

Then, on $\{(\tilde{R}_t^{x,*},W_t^*) \in \tilde{\mathcal{S}}_t\}$, convergence to zero of the noise sequence, the convex combination property of the bounding sequence and the monotone decreasing property of the approximate slopes, give us

$$L_t^k(\tilde{R}_t^{x,*},W_t^*) \leq \bar{v}_t^{n-1}(\tilde{R}_t^{x,*},W_t^*) \text{ a.s.}$$

Note that this inequality does not cover all the accumulation points of the sequence $\{(R_t^{x,n},W_t^n)\}_{n\geq 0}$, since they are restricted to states in the set $\tilde{\mathcal{S}}_t$. Nevertheless, this inequality and some properties of the projection operation are used to fulfill the requirements of a bounding technical lemma, which is used repeatedly to obtain the desired lower bound inequalities for all accumulation points.

In order to prove (21) when x_t is a vector (an issue that did not arise in N&P), we note that

$$F_t(\bar{v}_t^{n-1}(W_t^n),R_t^n,W_t^n,x_t) = C_t(R_t^n,W_t^n,x_t) + \gamma \bar{V}_t^{n-1}\left(f^x(R_t^n,x_t),W_t^n\right)$$

is a concave function of X_t and $\mathcal{X}_t(R_t^n,W_t^n)$ is a convex set. Let $\mathcal{X}_t^N(R_t^n,W_t^n,x_t^n)$ be the normal cone of $\mathcal{X}_t(R_t^n,W_t^n)$ at x_t^n, and let $\partial F_t(v,R,W,x)$ be the set of subdifferentials of $F(v,R,W,x)$ at (v,R,W,x). Then,

$$0 \in \partial F_t(\bar{v}_t^{n-1}(W_t^n),R_t^n,W_t^n,x_t^n) - \mathcal{X}_t^N(R_t^n,W_t^n,x_t^n),$$

where x_t^n is the optimal decision of the optimization problem in *STEP 3a* of the algorithm. This inclusion and the first convergence result are then combined to show that

$$0 \in \partial F_t(v_t^*(W_t^*),R_t^*,W_t^*,x_t^*) - \mathcal{X}_t^N(R_t^*,W_t^*,x_t^*).$$

We provide the full proof that this condition is satisfied in section 4.4.

4.2 Technical Elements

In this section, we set the stage to the convergence proofs by defining some technical elements. We start with the definition of the deterministic sequence $\{L_t^k(R,W_t)\}_{k\geq 0}$. For this, we let B^v be a

deterministic integer that bounds $|\hat{v}_t^n(R)|$, $|z_t^n(R,W_t)|$, $|\bar{v}_t^n(R,W_t)|$ and $|v_t^*(R,W_t)|$ for all $W_t \in \mathcal{W}_t$ and $R = 1,\ldots,B^R$. Then, for $t = 0,\ldots,T-1$, $W_t \in \mathcal{W}_t$ and $R = 1,\ldots,B^R$, we have that

$$\begin{aligned} L_t^0(R,W_t) &= v_t^*(R,W_T) - 2B^v, \\ L_t^{k+1}(R,W_t) &= \frac{L_t^k(R,W_T) + (HL^k)_t(R,W_t)}{2}. \end{aligned} \tag{24}$$

At the end of the planning horizon T, $L_T^k(R,W_T) = 0$ for all $k \geq 0$. The proposition below introduces the required properties of the deterministic sequence $\{L^k(R,W_t)\}$ for $k \geq 0$. Its proof is deferred to the appendix.

Proposition 2 *Given assumptions* (13)-(15)*, for* $t = 0,\ldots,T-1$*, information vector* $W_t \in \mathcal{W}_t$ *and resource levels* $R = 1,\ldots,B^R$*,*

$$L_t^k(R,W_t) \text{ is monotone decreasing,} \tag{25}$$

$$(HL^k)_t(R,W_t) \geq L_t^{k+1}(R,W_t) \geq L_t^k(R,W_t), \tag{26}$$

$$L_T^k(R,W_t) < v_t^*(R,W_t), \quad \text{and} \quad \lim_{k\to\infty} L_t^k(R,W_t) = v_t^*(R,W_t). \tag{27}$$

In the proof, we demonstrate that assumptions (13)-(15) are satisfied. The deterministic sequence $\{U^k(R,W_t)\}_{k\geq 0}$ is defined in a symmetrical way. It also has the properties stated in proposition 2, with the reversed inequality signs.

We move on to define the random index $\bar{N}$ that 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 states (actions) visited (taken) by the algorithm after iteration $\bar{N}$ are accumulation points of the sequence of states (actions) generated by the algorithm. In fact, $\bar{N}$ can be required to satisfy other constraints of the type: if an event did not happen infinitely often, then it did not happen after $\bar{N}$. Since we need $\bar{N}$ to be finite almost surely, the additional number of constraints have to be finite.

We introduce the set of iterations, namely $\mathcal{N}_t(R,W_t)$, that keeps track of the effects produced by the projection operation. For $W_t \in \mathcal{W}_t$ and $R = 1,\ldots,B^R$, let $\mathcal{N}_t(R,W_t)$ be the set of iterations in which the unprojected slope corresponding to state (R,W_t), that is, $z_t^n(R,W_t)$ was too large and had to be decreased by the projection operation. Formally,

$$\mathcal{N}_t(R,W_t) = \{n \in \mathbb{N} : z_t^n(R,W_t) > \bar{v}_t^n(R,W_t)\}.$$

A related set is the set of states $\tilde{\mathcal{S}}_t$. A state (R, W_t) is an element of $\tilde{\mathcal{S}}_t$ if (R, W_t) is equal to an accumulation point $(R_t^{x,*}, W_t^*)$ of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$ or is equal to $(R_t^{x,*} + \rho, W_t^*)$. Its corresponding approximate slope also has to satisfy the condition $z_t^n(R, W_t) \geq \bar{v}_t^n(R, W_t)$ for all $n \geq \bar{N}$, that is, the projection operation decreased or kept the same the corresponding unprojected slopes infinitely often.

We close this section dealing with measurability issues. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space under consideration. The sigma-algebra $\mathcal{F}$ is defined by $\mathcal{F} = \sigma\{(W_t^n, x_t^n), \quad n \geq 1, \quad t = 0, \ldots, T\}$. Moreover, for $n \geq 1$ and $t = 0, \ldots, T$,

$$\mathcal{F}_t^n = \sigma\left(\{(W_{t'}^m, x_{t'}^m), \quad 0 < m < n, \quad t' = 0, \ldots, T\} \bigcup \{(W_{t'}^n, x_{t'}^n), \quad t' = 0, \ldots, t\}\right).$$

Clearly, $\mathcal{F}_t^n \subset \mathcal{F}_{t+1}^n$ and $\mathcal{F}_T^n \subset \mathcal{F}_0^{n+1}$. Furthermore, given the initial slopes $\bar{v}_t^0(W_t)$ and the initial resource level $\bar{r}$, we have that R_t^n, $R_t^{x,n}$ and α_t^n are in $\mathcal{F}_t^n$, while $\hat{v}_{t+1}^n(R_t^x)$, $z_t^n(W_t)$, $\bar{v}_t^n(W_t)$ are in $\mathcal{F}_{t+1}^n$. A pointwise argument is used in all the proofs of almost sure convergence presented in this paper. Thus, zero-measure events are discarded on an as-needed basis.

4.3 Almost sure convergence of the slopes

We prove that the approximation slopes produced by the SPAR-Storage algorithm converge almost surely to the slopes of the optimal value functions of the storage class for states that can be reached by an optimal policy. This result is stated in theorem 1 below. Along with the proof of the theorem, we present the noise and the bounding stochastic sequences and introduce three technical lemmas. Their proofs are given in the appendix so that the main reasoning is not disrupted.

Before we present the theorem establishing the convergence of the value function, we introduce the following technical condition. Given $k \geq 0$ and $t = 0, \ldots, T-1$, we assume there exists a positive random variable N_t^k such that on $\{n \geq N_t^k\}$,

$$(HL^k)_t(R_t^n, W_t^n) \leq (H\bar{v}^{n-1})_t(R_t^n, W_t^n) \leq (HU^k)_t(R_t^n, W_t^n) \quad \text{a.s.}, \tag{28}$$

$$(HL^k)_t(R_t^n \rho, W_t^n) \leq (H\bar{v}^{n-1})_t(R_t^n + \rho, W_t^n) \leq (HU^k)_t(R_t^n + \rho, W_t^n) \quad \text{a.s.} \tag{29}$$

Proving this condition is nontrivial. It draws on a proof technique in Bertsekas and Tsitsiklis (1996), Section 4.3.6, although this technique requires an exploration policy that ensures that each state is visited infinitely often. N&P (section 6.1) shows how this proof can be adapted to a scalar problem that is linear over the feasible region, but this proof is presented specifically in the context of the

lagged asset acquisition problem, which has very specific structure (as described in section 1. We show that it is satisfied when the decisions are vector-valued subject to a set of linear constraints.

The proof requires the following lemmas, all of which are proven in the appendix.

Lemma 1 *On the event that* $(R_t^{x,*}, W_t^*)$ *is an accumulation point of the sequence* $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$, *we have that*

$$\{\bar{s}_t^n(R_t^{x,*}, W_t^*)\}_{n\geq 0} \to 0 \quad \textit{and} \quad \{\bar{s}_t^n(R_t^{x,*}+\rho, W_t^*)\}_{n\geq 0} \to 0 \ \textit{a.s.} \tag{30}$$

This lemma shows that the stochastic noise sequences asymptotically vanish.

Lemma 2 *On the event that* $\{n \geq N_t^k, \quad (\tilde{R}_t^{x,*}, W_t^*) \in \tilde{\mathcal{S}}_t\}$,

$$\bar{v}_t^{n-1}(\tilde{R}_t^{x,*}, W_t^*) \geq \bar{l}_t^{n-1}(\tilde{R}_t^{x,*}, W_t^*) - \bar{s}_t^{n-1}(\tilde{R}_t^{x,*}, W_t^*) \ \textit{a.s.} \tag{31}$$

This lemma bounds slopes occurring at resource points $\tilde{R}_t^{x,*}, W_t^*)$ that are accumulation points that approach the limit from above.

Lemma 3 *Given an information vector* $W_t \in \mathcal{W}_t$ *and a resource level* $R = 1, \ldots, B^R - 1$, *if for all* $k \geq 0$, *if there exists an integer random variable* $N^k(R, W_t)$ *such that* $L_t^k(R, W_t) \leq \bar{v}_t^{n-1}(R, W_t)$ *almost surely on* $\{n \geq N^k(R, W_t), \quad \mathcal{N}_t(R+\rho, W_t) = \infty\}$, *then for all* $k \geq 0$, *there exists another integer random variable* $N^k(R+\rho, W_t)$ *such that* $L_t^k(R+\rho, W_t) \leq \bar{v}_t^{n-1}(R+\rho, W_t)$ *almost surely on* $\{n \geq N^k(R+\rho, W_t)\}$.

This lemma shows that if $L_t^k(R, W_t) \leq \bar{v}_t^{n-1}(R, W_t)$ for a finite n, then the same is true for the next slope for $R+\rho$, which allows us to prove convergence of adjacent slopes of R is visited infinitely often.

These lemmas are used in the proof of the main result.

Theorem 1 *Assume the stepsize conditions* (17)–(18). *Also assume* (28) *and* (29). *Then, for all* $k \geq 0$ *and* $t = 0, \ldots, T$, *on the event that* $(R_t^{x,*}, W_t^*)$ *is an accumulation point of* $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$, *the sequences of slopes* $\{\bar{v}_t^n(R_t^{x,*}, W_t^*)\}_{n\geq 0}$ *and* $\{\bar{v}_t^n(R_t^{x,*}+\rho, W_t^*)\}_{n\geq 0}$ *generated by the SPAR-Storage algorithm for the storage class converge almost surely to the optimal slopes* $v_t^*(R_t^{x,*}, W_t^*)$ *and* $v_t^*(R_t^{x,*}+\rho, W_t^*)$, *respectively.*

The proof largely parallels the proof in N&P, since both algorithms are estimating scalar, piecewise linear value functions. However, the proof in N&P assumed throughout that: a) the decision x_t is a discrete scalar $x_t \in \{0, 1, 2, \ldots, M\}$, b) the contribution function $C(R, W, x) = -Px$ is linear over this entire region (here, P was a price), and c) the scalar quantity R evolved according to $R_{t+1} = R_t + x_t$, which made it trivial to assume that R_t was always an integer. Our proof is adapted for the more general setting of our problem, but otherwise follows the same structure, and for this reason the proof (which is quite lengthy) has been moved to the appendix. There is, however, one final step that is different as a result of our use of a linear programming subproblem to handle the fact that x_t is a vector. We handle this issue in the next section.

4.4 Optimality of the Decisions

The most important difference between our work and that of N&P is that x_t is a vector, and our decision problem has to be solved as a linear program. 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. This result is not immediate from theorem 1 because we do not guarantee that we find the optimal value function; instead, theorem 1 shows only that we find the correct slopes for points that are visited infinitely often.

Proposition 4.1 *Assume the conditions of Theorem 1 are satisfied. For $t = 0, \ldots, T$, on the event that $(R_t^*, W_t^*, \bar{v}_t^*, x_t^*)$ is an accumulation point of the sequence $\{(R_t^n, W_t^n, \bar{v}_t^{n-1}, x_t^n)\}_{n\geq 1}$ generated by the SPAR-Storage algorithm, x_t^*, is almost surely an optimal solution of*

$$\max_{x_t \in \mathcal{X}_t(R_t^*, W_t^*)} F_t(v_t^*(W_t^*), R_t^*, W_t^*, x_t), \tag{32}$$

where

$$F_t(v_t^*(W_t^*), R_t^*, W_t^*, x_t) = C_t(R_t^*, W_t^*, x_t) + \gamma V_t^*(f^x(R_t^*, x_t), W_t^*).$$

Proof: Fix $\omega \in \Omega$. As before, the dependence on ω is omitted. At each iteration n and time t of the algorithm, the decision x_t^n in *STEP 3* of the algorithm is an optimal solution to the optimization problem $\max_{x_t \in \mathcal{X}_t(R_t^n, W_t^n)} F_t(\bar{v}_t^{n-1}(W_t^n), R_t^n, W_t^n, x_t)$. Since $F_t(\bar{v}_t^{n-1}(W_t^n), R_t^n, W_t^n, \cdot)$ is concave and $\mathcal{X}_t(R_t^n, W_t^n)$ is convex, we have that

$$0 \in \partial F_t(\bar{v}_t^{n-1}(W_t^n), R_t^n, W_t^n, x_t^n) - \mathcal{X}_t^N(R_t^n, W_t^n, x_t^n),$$

where $\partial F_t(\bar{v}_t^{n-1}(W_t^n), R_t^n, W_t^n, x_t^n)$ is the subdifferentials of $F_t(\bar{v}_t^{n-1}(W_t^n), R_t^n, W_t^n, \cdot)$ at x_t^n and $\mathcal{X}_t^N(R_t^n, W_t^n, x_t^n)$ is the normal cone of $\mathcal{X}_t(R_t^n, W_t^n)$ at x_t^n.

Then, by passing to the limit, we can conclude that each accumulation point $(R_t^*, W_t^*, \bar{v}_t^*, x_t^*)$ of the sequence $\{(R_t^n, W_t^n, \bar{v}_t^{n-1}, x_t^n)\}_{n\geq 1}$ satisfies the condition

$$0 \in \partial F_t(\bar{v}_t^*(W_t^*), R_t^*, W_t^*, x_t^*) - \mathcal{X}_t^N(R_t^*, W_t^*, x_t^*).$$

We now derive an expression for the subdifferential. We have that

$$\partial F_t(\bar{v}_t^*(W_t^*), R_t^*, W_t^*, x_t) = \nabla_x C_t(R_t^*, W_t^*, x_t) + \gamma \partial \bar{V}_t^*(R_t^* + A^S x_t, W_t^*),$$

where we substitute $R_t^* + A^S x_t$ for $f^x(R_t^*, x_t)$ because we now need to take advantage of the specific structure of the transition function. From (Bertsekas, Nedic and Ozdaglar, 2003, Proposition 4.2.5),

$$\partial \bar{V}_t^*(R_t^* + A^S x_t, W_t^*) = \{(A^S z)^T : \\ z \in [\bar{v}_t^*(R_t^* + A^S x_t + \rho, W_t^*), \bar{v}_t^*(R_t^* + A^S x_t, W_t^*)]\}.$$

This captures the property that the value of a marginal increase in the post-decision resource state is in the interval bounded by the slopes of the value function adjacent to the corresponding resource level. Therefore, as x_t^* falls on one of the break points,

$$\partial F_t(\bar{v}_t^*(W_t^*), R_t^*, W_t^*, x_t^*) = \{\nabla C_t(R_t^*, W_t^*, x_t^*) + \gamma A^S z : \\ z \in [\bar{v}_t^*(R_t^* + A^S x_t^* + \rho, W_t^*), \bar{v}_t^*(R_t^* + A^S x_t^*, W_t^*)]\}.$$

Since $(R_t^* + A^S x_t^*, W_t^*)$ is an accumulation point of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$, it follows from theorem 1 that

$$\bar{v}_t^*(R_t^* + A^S x_t^*, W_t^*) = v_t^*(R_t^* + A^S x_t^*, W_t^*)$$

and

$$\bar{v}_t^*(R_t^* + A^S x_t^* + \rho, W_t^*) = v_t^*(R_t^* + A^S x_t^* + \rho, W_t^*).$$

which means that the left and right derivatives of our approximate slopes are equal to the optimal slopes.

Hence, $\partial F_t(\bar{v}_t^*(W_t^*), R_t^*, W_t^*, x_t^*) = \partial F_t(v_t^*(W_t^*), R_t^*, W_t^*, x_t)$ and

$$0 \in \partial F_t(v_t^*(W_t^*), R_t^*, W_t^*, x_t^*) - \mathcal{X}_t^N(R_t^*, W_t^*, x_t^*),$$

which proves that x_t^* is the optimal solution of (32).

5 Summary

We propose an approximate dynamic programming algorithm using pure exploitation for the problem of planning energy flows over a year, in hourly increments, in the presence of a single storage device, taking advantage of our ability to represent the value function as a piecewise linear function. We are able to prove almost sure convergence of the algorithm using a pure exploitation strategy. This ability is critical for this application, since we discretized each value function (for 8,760 time periods) into thousands of breakpoints. Rather than needing to visit all the possible values of the storage level many times, we only need to visit a small number of these points (for each value of our information vector W_t) many times.

A key feature of our algorithm is that it is able to handle high-dimensional decisions in each time period. For the energy application, this means that we can solve the relatively small (but with hundreds of dimensions) decision problem that determines how much energy is coming from each energy source and being converted to serve different types of energy demands. We are able to solve sequences of deterministic linear programs very quickly by formulating the value function around the post-decision state variable.

Acknowledgements

This research was supported in part by grant AFOSR contract FA9550-08-1-0195 and the National Science Foundation, grant CMMI-0856153

Appendix

The appendix consists of a brief summary of notation to assist with reading the proofs and the proofs themselves.

Notation

For each random element, we provide its measurability.

- Filtrations
  $\mathcal{F} = \sigma\{(W_t^n, x_t^n),\ n \geq 1,\ t = 0, \ldots, T\}$.
  $\mathcal{F}_t^n = \sigma\left(\{(W_{t'}^m, x_{t'}^m),\ 0 < m < n,\ t' = 0, \ldots, T\} \cup \{(W_{t'}^n, x_{t'}^n),\ t' = 0, \ldots, t\}\right)$.

- Post-decision state $(R_t^{x,n}, W_t^n)$
  $R_t^{x,n} \in \mathcal{F}_t^n$: resource level after the decision taking a value $k\rho$, $0 \leq k \leq B^R$.
  $W_t^n \in \mathcal{F}_t^n$: Markovian information vector. Independent of the resource level.
  $D_t^n \in \mathcal{F}_t^n$: demand vector. Nonnegative and integer valued.
  $\mathcal{W}_t$: finite support set of W_t^n.

- Slopes (monotone decreasing in R and bounded)
  $v_t^*(R, W_t)$: slope of the optimal value function at (R, W_t).
  $z_t^n(R, W_t) \in \mathcal{F}_{t+1}^n$: unprojected slope of the value function approximation at $(R, W_)$.
  $\bar{v}_t^n(R, W_t) \in \mathcal{F}_{t+1}^n$: slope of the value function approximation at (R, W_t).
  $\bar{v}_t^*(R, W_t) \in \mathcal{F}$: accumulation point of $\{\bar{v}_t^n(R, W_t)\}_{n \geq 0}$.
  $\hat{v}_t^n(R) \in \mathcal{F}_t^n$: sample slope at R.

- Stepsizes (bounded by 0 and 1, sum is $+\infty$, sum of the squares is $< +\infty$)
  $\alpha_t^n \in \mathcal{F}_t^n$ and $\bar{\alpha}_t^n(R, W_t) = \alpha_t^n \left(1_{\{R = R_t^{x,n}, W_t = W_t^n\}} + 1_{\{R = R_t^{x,n} + \rho, W_t = W_t^n\}}\right)$.

- Finite random variable $\bar{N}$. Iteration big enough for convergence analysis.

- Set of iterations and states (due to the projection operation)
  $\mathcal{N}_t(R, W_t) \in \mathcal{F}$: iterations in which the unprojected slope at (R, W_t) was decreased.
  $\tilde{\mathcal{S}}_t \in \mathcal{F}$: states in which the projection had not decreased the unprojected slopes i.o.

- Dynamic programming operator H.

- Deterministic slope sequence $\{L_t^k(R, W_t)\}_{k\geq 0}$. Converges to $v_t^*(R, W_t)$.
- Error variable $\hat{s}_{t+1}^n(R) \in \mathcal{F}_{t+1}^n$
- Stochastic noise sequence $\{\bar{s}_t^n(R, W_t)\}_{n\geq 0}$
- Stochastic bounding sequence $\{\bar{l}_t^n(R, W_t)\}_{n\geq 0}$

Each lemma assumes all the conditions imposed and all the results obtained before its statement in the proof of Theorem 1. To improve the presentation of each proof, all the assumptions are presented beforehand.

We start with the proof of proposition 2, then we present the proofs for three technical lemmas needed in the proof of theorem 1. We close with the proof of theorem 1, adapted from the proof in N&P.

Proof of proposition 2

Proof: In the proof of proposition 2, we establish that assumptions (13)-(15) are satisfied. We use the notational convention that L^k is the entire set of slopes $L^k = \{L_t^k(W_t) \text{ for } t = 0, \ldots, T, \quad W_t \in \mathcal{W}_t\}$, where $L_t^k(W_t) = (L_t^k(1, W_t), \ldots, L_t^k(B^R, W_t))$.

We start by showing (25). Clearly, $L_T^k(W_T)$ is monotone decreasing (MD) for all $k \geq 0$ and $W_T \in \mathcal{W}_T$, since $L_T^k(W_T)$ is a vector of zeros. Thus, using condition (13), we have that $(HL^k)_{T-1}(W_{T-1})$ is MD. We keep in mind that, for $t = 0, \ldots, T-1$, $L_t^0(W_t)$ is MD as $L^0 = v^* - 2B^v e$. By definition, we have that $L_{T-1}^1(W_{T-1})$ is MD. A simple induction argument shows that $L_{T-1}^k(W_{T-1})$ is MD. Using the same argument for $t = T-2, \ldots, 0$, we show that $(HL^k)_t(W_t)$ and $L_t^k(W_t)$ are MD.

We now show (26). Since v^* is the unique fixed point of H, then $L^0 = Hv^* - 2B^v e$. Moreover, from condition (15), we have that $(Hv^*)_t(R, W_t) - 2B^v \leq (H(v^* - 2B^v e))_t(R, W_t) = (HL^0)_t(R, W_t)$ for all $W_t \in \mathcal{W}_t$ and $R = 1, \ldots, B_t^R$. Hence, $L^0 \leq HL^0$ and $L^0 \leq L^1 = \frac{L^0 + HL^0}{2} \leq HL^0$. Suppose that (26) holds true for some $k > 0$. We shall prove for $k+1$. The induction hypothesis and (25) tell us that condition (13) 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 (27). 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 $L_t(W_t)$ is monotone decreasing. Therefore, conditions (13)-(15) are also true when applied to L.

Hence, as shown in Bertsekas and 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 and Tsitsiklis, 1996, Lemma 3.4)

$$L = \frac{L + HL}{2}.$$

It follows that $L = v^*$, as v^* is the unique fixed point of H.

Proof of Lemma 1

Proof: Assume stepsize conditions (17)-(18).

Fix $\omega \in \Omega$. Omitting the dependence on ω, let $(R_t^{x,*}, W_t^*)$ be an accumulation point of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$. In order to simplify notation, let $\bar{s}_t^n(R_t^{x,*}, W_t^*)$ be denoted by $\bar{s}_t^{*,n}$ and $\bar{\alpha}_t^n(R_t^{x,*}, W_t^*)$ be denoted by $\bar{\alpha}_t^{*,n}$. Furthermore, let

$$\hat{\theta}_{t+1}^n = \hat{s}_{t+1}^n(R_t^n 1_{\{R_t^{x,*} \leq R_t^{x,n}\}} + (R_t^n + \rho)1_{\{R_t^{x,*} > R_t^{x,n}\}}).$$

We have, for $n \geq 1$,

$$(\bar{s}_t^{*,n})^2 \leq \left[\left(1 - \bar{\alpha}_t^{*,n}\right)\bar{s}_t^{*,n-1} + \bar{\alpha}_t^{*,n}\hat{\theta}_{t+1}^n\right]^2 = (\bar{s}_t^{*,n-1})^2 - 2\bar{\alpha}_t^{*,n}(\bar{s}_t^{*,n-1})^2 + A_t^n, \tag{33}$$

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

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

It is trivial to see that both $\bar{s}_t^{*,n-1}$ and $\hat{\theta}_{t+1}^n$ are bounded. Thus, $(\hat{\theta}_{t+1}^n - \bar{s}_t^{*,n-1})^2$ is bounded and (18) tells us that

$$\sum_{n=1}^{\infty}(\bar{\alpha}_t^{*,n})^2(\hat{\theta}_{t+1}^n - \bar{s}_t^{*,n-1})^2 < \infty. \tag{34}$$

Define a new sequence $\{g_{t+1}^n\}_{n\geq 0}$, where $g_{t+1}^0 = 0$ and

$$g_{t+1}^n = \sum_{m=1}^{n} \bar{\alpha}_t^{*,m} \bar{s}_t^{*,m-1} \hat{\theta}_{t+1}^m.$$

We can easily check that $\{g_{t+1}^n\}_{n\geq 0}$ is a $\mathcal{F}_T^n$-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 consequentially the integrability can be obtained by noticing that $(g_{t+1}^n)^2 = (g_{t+1}^{n-1})^2 + 2g_{t+1}^{n-1}\bar{\alpha}_t^{*,n}\bar{s}_t^{*,n-1}\hat{\theta}_{t+1}^n + (\bar{\alpha}_t^{*,n})^2(\bar{s}_t^{*,n-1}\hat{\theta}_{t+1}^n)^2$. From the martingale equality and boundedness of $\bar{s}_t^{*,n-1}$ and $\hat{\theta}_{t+1}^n$, we get

$$\mathbb{E}\left[(g_{t+1}^n)^2 | \mathcal{F}_T^{n-1}\right] \leq (g_{t+1}^{n-1})^2 + C\mathbb{E}\left[(\bar{\alpha}_t^{*,n})^2 | \mathcal{F}_T^{n-1}\right],$$

where C is a constant. Hence, taking expectations and repeating the process, we obtain, from the stepsize assumption (18) and $\mathbb{E}[(g_{t+1}^0)^2] = 0$,

$$\mathbb{E}\left[(g_{t+1}^n)^2\right] \leq \mathbb{E}\left[(g_{t+1}^{n-1})^2\right] + C\mathbb{E}\left[(\bar{\alpha}_t^{*,n})^2\right] \leq \mathbb{E}\left[(g_{t+1}^0)^2\right] + C\sum_{m=1}^{n}\mathbb{E}\left[(\bar{\alpha}_t^{*,m})^2\right] < \infty.$$

Therefore, the L^2-Bounded Martingale Convergence Theorem (Shiryaev, 1996, page 510) tells us that

$$\sum_{n=1}^{\infty} \bar{\alpha}_t^{*,n} \bar{s}_t^{*,n-1} \hat{\theta}_{t+1}^n < \infty. \tag{35}$$

Inequalities (34) and (35) show us that $\sum_{n=1}^{\infty} A_t^n < \infty$, and so, it is valid to write

$$A_t^n = \sum_{m=n}^{\infty} A_t^m - \sum_{m=n+1}^{\infty} A_t^m.$$

Therefore, as $-2\bar{\alpha}_t^{*,n}(\bar{s}_t^{*,n-1})^2 < 0$, inequality (33) can be rewritten as

$$(\bar{s}_t^{*,n})^2 + \sum_{m=n+1}^{\infty} A_t^m \leq (\bar{s}_t^{*,n-1})^2 + \sum_{m=n}^{\infty} A_t^m. \tag{36}$$

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

Finally, as inequality (33) holds for all $n \geq 1$, it yields

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

Passing to the limits we obtain:

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

This implies, together with the convergence of $\{\bar{s}_t^{*,n}\}_{n\geq 0}$, that

$$\sum_{m=1}^{\infty} \bar{\alpha}_t^{*,m}(\bar{s}_t^{*,m-1})^2 < \infty.$$

On the other hand, stepsize assumption (17) tells us that $\sum_{m=1}^{\infty} \bar{\alpha}_t^{*,m} = \infty$. Hence, there must exist a subsequence of $\{\bar{s}_t^{*,n}\}_{n\geq 0}$ that converges to zero. Therefore, as every subsequence of a convergent sequence converges to its limit, it follows that $\{\bar{s}_t^{*,n}\}_{n\geq 0}$ converges to zero.

Proof of Lemma 2

Proof: We make the following assumptions: Given $t = 0, \ldots, T-1$, $k \geq 0$ and integer N_t^k, assume on $\{n \geq N_t^k\}$, that (42), (28) and (29) hold true.

Again, fix $\omega \in \Omega$ and omit the dependence on ω. The proof is by induction on n. Let $(\tilde{R}_t^{x,*}, W_t^*)$ be in $\tilde{\mathcal{S}}_t$. The proof for the base case $n = N_t^k$ is immediate from the fact that $\bar{s}_t^{N_t^k-1}(\tilde{R}_t^{x,*}, W_t^*) = 0$, $\bar{l}_t^{N_t^k-1}(\tilde{R}_t^{x,*}, W_t^*) = L_t^k(\tilde{R}_t^{x,*}, W_t^*)$ and by the assumption that (42) holds true for $n \geq N_t^k$. Now suppose (31) holds for $n > N_t^k \geq \bar{N}$ and we need to prove for $n+1$.

To simplify the notation, let $\bar{\alpha}_t^n(\tilde{R}_t^{x,*}, W_t^*)$ be denoted by $\tilde{\alpha}_t^n$ and $\bar{v}_t^n(\tilde{R}_t^{x,*}, W_t^*)$ be denoted by $\tilde{v}_t^n$. We use the same shorthand notation for $z_t^n(\tilde{R}_t^{x,*}, W_t^*)$, $\bar{l}_t^n(\tilde{R}_t^{x,*}, W_t^*)$ and $\bar{s}_t^n(\tilde{R}_t^{x,*}, W_t^*)$ as well. Keeping in mind that the set of iterations $\mathcal{N}_t(\tilde{R}_t^{x,*}, W_t^*)$ is finite and for all $n \geq N_t^k \geq \bar{N}$, $\bar{v}_t^n(\tilde{R}_t^{x,*}, W_t^*) \geq z_t^n(\tilde{R}_t^{x,*}, W_t^*)$. We consider three different cases:

Case 1: $W_t^* = W_t^n$ and $\tilde{R}_t^{x,*} = R_t^{x,n}$.

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

Thus,

$$\begin{aligned}
\tilde{v}_t^n \geq \tilde{z}_t^n &= (1-\tilde{\alpha}_t^n)\tilde{v}_t^{n-1} + \tilde{\alpha}_t^n \hat{v}_{t+1}^n(R_t^{x,n}) \\
&\geq (1-\tilde{\alpha}_t^n)(\tilde{l}_t^{n-1} - \tilde{s}_t^{n-1}) + \tilde{\alpha}_t^n \hat{v}_{t+1}^n(R_t^{x,n}) \\
&\quad - \tilde{\alpha}_t^n (H\bar{v}^{n-1})_t(R_t^{x,n}, W_t^n) + \tilde{\alpha}_t^n (H\bar{v}^{n-1})_t(R_t^{x,n}, W_t^n) \qquad (37) \\
&\geq (1-\tilde{\alpha}_t^n)(\tilde{l}_t^{n-1} - \tilde{s}_t^{n-1}) - \tilde{\alpha}_t^n \hat{s}_{t+1}^n(R_t^{x,n}) + \tilde{\alpha}_t^n (HL^k)_t(R_t^{x,n}, W_t^n) \qquad (38) \\
&= \tilde{l}_t^n - ((1-\tilde{\alpha}_t^n)\tilde{s}_t^{n-1} + \tilde{\alpha}_t^n \hat{s}_{t+1}^n(R_t^{x,n})) \qquad (39) \\
&\geq \tilde{l}_t^n - \left(\max\left(0, (1-\tilde{\alpha}_t^n)\tilde{s}_t^{n-1} + \tilde{\alpha}_t^n \hat{s}_{t+1}^n(R_t^{x,n})\right)\right) \\
&= \tilde{l}_t^n - \tilde{s}_t^n. \qquad (40)
\end{aligned}$$

The first inequality is due to the construction of set $\tilde{\mathcal{S}}_t$, while (37) is due to the induction hypothesis. Inequalities (28) and (29) for $n \geq N_t^k$ explains (38). Finally, (39) and (40) come from the definition of the stochastic sequences $\tilde{l}_t^n$ and $\tilde{s}_t^n$, respectively.

Case 2: $W_t^* = W_t^n$ and $\tilde{R}_t^{x,*} = R_t^n + \rho$.

This case is analogous to the previous one, except that we use the sample slope $\hat{v}_{t+1}^n(R_t^{x,n} + \rho)$ instead of $\hat{v}_{t+1}^n(R_t^{x,n})$. We also consider the $(R_t^{x,n} + \rho, W_t^n)$ component, instead of $(R_t^{x,n}, W_t^n)$.

Case 3: Else.

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

$$\tilde{l}_t^n = \tilde{l}_t^{n-1} \quad \text{and} \quad \tilde{s}_t^n = \tilde{s}_t^{n-1}.$$

Therefore, from the construction of set $\tilde{\mathcal{S}}_t$ and the induction hypothesis

$$\tilde{v}_t^n \geq \tilde{z}_t^n = \tilde{v}_t^{n-1} \geq \tilde{l}_t^{n-1} - \tilde{s}_t^{n-1} = \tilde{l}_t^n - \tilde{s}_t^n.$$

Proof of Lemma 3

Assumptions: Assume stepsize conditions (17)-(18). Moreover, assume for all $k \geq 0$ and integer N_t^k that inequalities (28) and (29) hold true on $\{n \geq N_t^k\}$. Finally, assume there exists an integer random variable $N^k(R, W_t)$ such that $\bar{v}_t^{n-1}(R, W_t) \geq L_t^k(R, W_t)$ on $\{n \geq N^k(R, W_t), \quad \mathcal{N}_t(R+\rho, W_t) = \infty\}$.

Pick $\omega \in \Omega$ and omit the dependence of the random elements on w. We start by showing, for each $k \geq 0$, there exists an integer $N_t^{k,s}$ such that $\bar{v}_t^{n-1}(R+\rho, W_t) \geq L_t^k(R+\rho, W_t) - \bar{s}_t^{n-1}(R+\rho, W_t)$

for all $n \geq N_t^{k,s}$. Then, we show, for all $\epsilon > 0$, there is an integer $N_t^{k,\epsilon}$ such that $\bar{v}_t^{n-1}(R+\rho, W_t) \geq L_t^k(R+\rho, W_t) - \epsilon$ for all $n \geq N_t^{k,\epsilon}$. Finally, using these results, we prove existence of an integer $N^k(R+\rho, W_t)$ such that $\bar{v}_t^{n-1}(R+\rho, W_t) \geq L_t^k(R+\rho, W_t)$ for all $n \geq N^k(R+\rho, W_t)$.

Pick $k \geq 0$. Let $N_t^{k,s} = \min\left\{n \in \mathcal{N}_t(R+\rho, W_t) : n \geq N^k(R, W_t)\right\} + 1$, where $N^k(R, W_t)$ is the integer such that $\bar{v}_t^{n-1}(R, W_t) \geq L_t^k(R, W_t)$ for all $n \geq N^k(R, W_t)$. Given the projection properties discussed in the proof of Theorem 1, $\mathcal{N}_t(R+\rho, W_t)$ is infinite and $\bar{v}_t^{N_t^{k,s}-1}(R, W_t) = \bar{v}_t^{N_t^{k,s}-1}(R+\rho, W_t)$. Therefore, $N_t^{k,s}$ is well defined. Redefine the noise sequence $\{\bar{s}_t^n(R+\rho, W_t)\}_{n\geq 0}$ introduced in the proof of Theorem 1 using $N_t^{k,s}$ instead of N_t^k. We prove that $\bar{v}_t^{n-1}(R+\rho, W_t) \geq L_t^k(R+\rho, W_t) - \bar{s}_t^{n-1}(R+\rho, W_t)$ for all $n \geq N_t^{k,s}$ by induction on n.

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

$$\begin{aligned}
\bar{v}_t^{n-1}(R+\rho, W_t) &= \bar{v}_t^{n-1}(R, W_t) \\
&\geq L_t^k(R, W_t) \\
&\geq L_t^k(R+\rho, W_t) \\
&= L_t^k(R+\rho, W_t) - \bar{s}_t^{n-1}(R+\rho, W_t).
\end{aligned}$$

Now, we suppose $\bar{v}_t^{n-1}(R+\rho, W_t) \geq L_t^k(R+\rho, W_t) - \bar{s}_t^{n-1}(R+\rho, W_t)$ is true for $n-1 > N_t^{k,s}$ and prove for n. We have to consider two cases:

Case 1: $n \in \mathcal{N}_t(R+\rho, W_t)$

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

$$\begin{aligned}
\bar{v}_t^{n-1}(R+\rho, W_t) &= \bar{v}_t^{n-1}(R, W_t) \\
&\geq L_t^k(R, W_t) \\
&\geq L_t^k(R+\rho, W_t) \\
&\geq L_t^k(R+\rho, W_t) - \bar{s}_t^{n-1}(R+\rho, W_t).
\end{aligned}$$

Case 2: $n \notin \mathcal{N}_t(R+\rho, W_t)$

The analysis of this case is analogous to the proof of inequality (31) of lemma 2 for $(R+\rho, W_t) \in \tilde{\mathcal{S}}_t$. The difference is that we consider L_t^k instead of the stochastic bounding sequence

$\{\bar{l}_t^n(R+\rho, W_t)\}_{n\geq 0}$ and we note that

$$(1-\bar{\alpha}_t^n(R+\rho,W_t))L_t^k(R+\rho,W_t)+\bar{\alpha}_t^n(R+\rho,W_t)(HL^k)_t(R+\rho,W_t) \geq L_t^k(R+\rho,W_t).$$

Hence, we have proved that for all $k \geq 0$ there exists an integer $N_t^{k,s}$ such that

$$\bar{v}_t^{n-1}(R+\rho,W_t) \quad \geq \quad L_t^k(R+\rho,W_t)-\bar{s}_{t+}^{n-1}(R+\rho,W_t) \quad \text{for all } n \geq N_t^{k,s}.$$

We move on to show, for all $k \geq 0$ and $\epsilon > 0$, there is an integer $N_t^{k,\epsilon}$ such that $\bar{v}_t^{n-1}(R+\rho,W_t) \geq L_t^k(R+\rho,W_t)-\epsilon$ for all $n \geq N_t^{k,\epsilon}$. We have to consider two cases: (i) $(R+\rho,W_t)$ is either equal to an accumulation point $(R_t^{x,*},W_t^*)$ or $(R_t^{x,*}+\rho,W_t^*)$ and (ii) $(R+\rho,W_t)$ is neither equal to an accumulation point $(R_t^{x,*},W_t^*)$ nor $(R_t^{x,*}+\rho,W_t^*)$. For the first case, lemma 1 tells us that $\bar{s}_{t+}^n(R+\rho,W_t)$ goes to zero. Then, there exists $N^\epsilon > 0$ such that $\bar{s}_{t+}^n(R+\rho,W_t) < \epsilon$ for all $n \geq N^\epsilon$. Therefore, we just need to choose $N_t^{k,\epsilon} = \max(N_t^{k,s}, N^\epsilon)$. For the second case, $\bar{\alpha}_t^n(R+\rho,W_t) = 0$ for all $n \geq N_t^{k,s}$ and $\bar{s}_{t+}^{N_t^{k,s}-1}(R+\rho,W_t) = 0$. Thus, $\bar{s}_{t+}^n(R+\rho,W_t) = \bar{s}_{t+}^{N_t^{k,s}-1}(R+\rho,W_t) = 0$ for all $n \geq N_t^{k,s}$ and we just have to choose $N_t^{k,\epsilon} = N_t^{k,s}$.

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

Proof of Theorem 1

As discussed in section 4.1, since the deterministic sequences $\{L_t^k(R_t^x,W_t)\}_{k\geq 0}$ and $\{U_t^k(R_t^x,W_t)\}_{k\geq 0}$ do converge to the optimal slopes, the convergence of the approximation sequences is obtained by showing that for each $k \geq 0$ there exists a nonnegative

random variable $N_t^{*,k}$ such that on the event that $n \geq N_t^{*,k}$ and $(R_t^{x,*}, W_t^*)$ is an accumulation point of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$, we have

$$L_t^k(R_t^{x,*}, W_t^*) \leq \bar{v}_t^n(R_t^{x,*}, W_t^*) \leq U_t^k(R_t^{x,*}, W_t^*) \text{ a.s.}$$
$$L_t^k(R_t^{x,*} + \rho, W_t^*) \leq \bar{v}_t^n(R_t^{x,*} + \rho, W_t^*) \leq U_t^k(R_t^{x,*} + \rho, W_t^*) \text{ a.s.}$$

We concentrate on the inequalities $L_t^k(R_t^{x,*}, W_t^*) \leq \bar{v}_t^n(R_t^{x,*}, W_t^*)$ and $L_t^k(R_t^{x,*} + \rho, W_t^*) \leq \bar{v}_t^n(R_t^{x,*} + \rho, W_t^*)$. The upper bounds are obtained using a symmetrical argument.

The proof is by backward induction on t. The base case $t = T$ is trivial as $L_T^k(R, W_T) = \bar{v}_T^n(R, W_T) = 0$ for all $W_T \in \mathcal{W}_T$, $R = 1, \ldots, B^R$, $k \geq 0$ and iterations $n \geq 0$. Thus, we can pick, for example, $N_T^{*,k} = \bar{N}$, where $\bar{N}$, as defined in section 4.2, is a random variable that denotes when an iteration of the algorithm is large enough for convergence analysis purposes. The backward induction proof is completed when we prove, for $t = T-1, \ldots, 0$ and $k \geq 0$, that there exists $N_t^{*,k}$ such that on the event that $n \geq N_t^{*,k}$ and $(R_t^{x,*}, W_t^*)$ is an accumulation point of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$,

$$L_t^k(R_t^{x,*}, W_t^*) \leq \bar{v}_t^n(R_t^{x,*}, W_t^*) \quad \text{and} \quad L_t^k(R_t^{x,*} + \rho, W_t^*) \leq \bar{v}_t^n(R_t^{x,*} + \rho, W_t^*) \text{ a.s.} \tag{41}$$

Given the induction hypothesis for $t+1$, the proof for time period t is divided into two parts. In the first part, we prove for all $k \geq 0$ that there exists a nonnegative random variable N_t^k such that

$$L_t^k(\tilde{R}_t^{x,*}, W_t^*) \leq \bar{v}_t^{n-1}(\tilde{R}_t^{x,*}, W_t^*), \quad \text{a.s. on } \{n \geq N_t^k, \quad (\tilde{R}_t^{x,*}, W_t^*) \in \tilde{\mathcal{S}}_t\}. \tag{42}$$

Its proof is by induction on k. Note that it only applies to states in the random set $\tilde{\mathcal{S}}_t$. Then, again for t, we take on the second part, which takes care of the states not covered by the first part, proving the existence of a nonnegative random variable $N_t^{*,k}$ such that the lower bound inequalities are true on $\{n \geq N_t^{*,k}\}$ for all accumulation points of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$.

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

Induction Hypothesis: Given $t = T-1, \ldots, 0$, assume, for $t+1$, and all $k \geq 0$ the existence of integers N_{t+1}^k and $N_{t+1}^{*,k}$ such that, for all $n \geq N_{t+1}^k$, (42) is true, and, for all $n \geq N_{t+1}^{*,k}$, inequalities in (41) hold true for all accumulation points $(R_t^{x,*}, W_t^*)$.

Part 1:

For our fixed time period t, we prove for any k, the existence of an integer N_t^k such that for $n \geq N_t^k$, inequality(42) is true. The proof is by forward induction on k.

We start with $k = 0$. For every state (R, W_t), we have that $-B^v \leq v_t^*(R, W_t) \leq B^v$, implying, by definition, that $L_t^0(R, W_t) \leq -B^v$. Therefore, (42) is satisfied for all $n \geq 1$, since we know that $\bar{v}_t^{n-1}(R, W_t)$ is bounded by $-B^v$ and $+B^v$ for all iterations. Thus, $N_t^0 = \max\left(1, N_{t+1}^{*,0}\right) = N_{t+1}^{*,0}$.

The induction hypothesis on k assumes that there exists N_t^k such that, for all $n \geq N_t^k$ (42) is 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 depart from our pointwise argument in order to define the stochastic noise sequence using Lemma 1. We start by defining, for $R = 1, \ldots, B^R$, the random variable $\hat{s}_{t+1}^n(R) = (H\bar{v}^{n-1})_t(R, W_t^n) - \hat{v}_{t+1}^n(R)$ that measures the error incurred by observing a sample slope. Using $\hat{s}_{t+1}^n(R)$ we define for each $W_t \in \mathcal{W}_t$ the stochastic noise sequence $\{\bar{s}_t^n(R, W_t)\}_{n \geq 0}$. We have that $\bar{s}_t^n(R, W_t) = 0$ on $\{n < N_t^k\}$ and, on $\{n \geq N_t^k\}$, we have that is equal to

$$\begin{aligned}\bar{s}_t^n(R, W_t) \quad = \quad & \max\left(0, (1 - \bar{\alpha}_t^n(R, W_t))\, \bar{s}_t^{n-1}(R, W_t)\right. \\ & \left. + \bar{\alpha}_t^n(R, W_t) \hat{s}_{t+1}^n(R_t^{x,n} 1_{\{R \leq R_t^{x,n}\}} + (R_t^n + \rho) 1_{\{R > R_t^n\}})\right).\end{aligned}$$

The sample slopes are defined in a way such that

$$\mathbb{E}\left[\hat{s}_{t+1}^n(R) | \mathcal{F}_t^n\right] = 0. \tag{43}$$

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 lemma 1.

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

$$\delta^k = \min\left\{ \frac{(HL^k)_t(\tilde{R}_t^{x,*}, W_t^*) - L_t^k(\tilde{R}_t^{x,*}, W_t^*)}{4} : \begin{array}{c} (\tilde{R}_t^{x,*}, W_t^*) \in \tilde{\mathcal{S}}_t \\ (HL^k)_t(\tilde{R}_t^{x,*}, W_t^*) > L_t^k(\tilde{R}_t^{x,*}, W_t^*) \end{array} \right\}.$$

If $\delta^k < +\infty$ we define an integer $N_L \geq N_t^k$ to be such that

$$\prod_{m=N_t^k}^{N_L - 1} \left(1 - \bar{\alpha}_t^m(\tilde{R}_t^{x,*}, W_t^*)\right) \leq 1/4 \quad \text{and} \quad \bar{s}_t^{n-1}(\tilde{R}_t^{x,*}, W_t^*) \leq \delta^k, \tag{44}$$

for all $n \geq N_L$ and states $(\tilde{R}_t^{x,*}, W_t^*) \in \tilde{\mathcal{S}}_t$. Such an N_L exists because both (19) and (30) are true. If $\delta^k = +\infty$ then, for all states $(\tilde{R}_t^{x,*}, W_t^*) \in \tilde{\mathcal{S}}_t$, $(HL^k)_t(\tilde{R}_t^{x,*}, W_t^*) = L_t^k(\tilde{R}_t^{x,*}, W_t^*)$ since (26) tells us that $(HL^k)_t(W_t^*) \geq L_t^k(W_t^*)$. Thus, $L_t^{k+1}(\tilde{R}_t^{x,*}, W_t^*) = L_t^k(\tilde{R}_t^{x,*}, W_t^*)$ and we define the integer N_L to be equal to N_t^k. We let $N_t^{k+1} = \max\left(N_L, N_{t+1}^{*,k+1}\right)$ and show that (42) holds for $n \geq N_t^{k+1}$.

We pick a state $(\tilde{R}_t^{x,*}, W_t^*) \in \tilde{\mathcal{S}}_t$. If $L_t^{k+1}(\tilde{R}_t^{x,*}, W_t^*) = L_t^k(\tilde{R}_t^{x,*}, W_t^*)$, then inequality (42) follows from the induction hypothesis. We therefore concentrate on the case where $L_t^{k+1}(\tilde{R}_t^{x,*}, W_t^*) > L_t^k(\tilde{R}_t^{x,*}, W_t^*)$.

First, we depart one more time from the pointwise argument to introduce the stochastic bounding sequence. We use lemma 2, combining this sequence with the stochastic noise sequence. For $W_t \in \mathcal{W}_t$ and $R = 1, \ldots, B^R$, we have on $\{n < N_t^k\}$ that $\bar{l}_t^n(R, W_t) = L_t^k(R, W_t)$ and, on $\{n \geq N_t^k\}$,

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

A simple inductive argument proves that $\bar{l}_t^n(R, W_t)$ is a convex combination of $L_t^k(R, W_t)$ and $(HL^k)_t(R, W_t)$. Therefore we can write,

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

where $\tilde{b}_t^{n-1} = \prod_{m=N_t^k}^{n-1} \left(1 - \bar{\alpha}_t^m(\tilde{R}_t^{x,*}, W_t^*)\right)$. For $n \geq N_t^{k+1} \geq N_L$, we have $\tilde{b}_t^{n-1} \leq 1/4$. Moreover, $L_t^k(\tilde{R}_t^{x,*}, W_t^*) \leq (HL^k)_t(\tilde{R}_t^{x,*}, W_t^*)$. Thus, using (24) and the definition of δ^k, we obtain

$$\begin{aligned}
\bar{l}_t^{n-1}(\tilde{R}_t^{x,*}, W_t^*) &\geq \frac{1}{4} L_t^k(\tilde{R}_t^{x,*}, W_t^*) + \frac{3}{4}(HL^k)_t(\tilde{R}_t^{x,*}, W_t^*) \\
&= \frac{1}{2} L_t^k(\tilde{R}_t^{x,*}, W_t^*) + \frac{1}{2}(HL^k)_t(\tilde{R}_t^{x,*}, W_t^*) \\
&\quad + \frac{1}{4}((HL^k)_t(\tilde{R}_t^{x,*}, W_t^*) - L_t^k(\tilde{R}_t^{x,*}, W_t^*)) \\
&\geq L_t^{k+1}(\tilde{R}_t^{x,*}, W_t^*) + \delta^k. \qquad (45)
\end{aligned}$$

Combining (31) and (45), we obtain, for all $n \geq N_t^{k+1} \geq N_L$,

$$\begin{aligned}
\bar{v}_t^{n-1}(\tilde{R}_t^{x,*}, W_t^*) &\geq L_t^{k+1}(\tilde{R}_t^{x,*}, W_t^*) + \delta^k - \bar{s}_t^{n-1}(\tilde{R}_t^{x,*}, W_t^*) \\
&\geq L_t^{k+1}(\tilde{R}_t^{x,*}, W_t^*) + \delta^k - \delta^k \\
&= L_t^{k+1}(\tilde{R}_t^{x,*}, W_t^*),
\end{aligned}$$

where the last inequality follows from (44).

Part 2:

We continue to consider ω picked in the beginning of the proof of the theorem. In this part, we take care of the states $(R_t^{x,*}, W_t^*)$ that are accumulation points but are not in $\tilde{\mathcal{S}}_t$. In contrast to part 1, the proof technique here is not by forward induction on k. We rely entirely on the definition of the projection operation and on the elements defined in section 4.2, as this part of the proof is all about states for which the projection operation decreased the corresponding approximate slopes infinitely often, which might happen when some of the optimal slopes are equal. Of course this fact is not verifiable in advance, as the optimal slopes are unknown.

Remember that at iteration n time period t, we observe sample realizations of the slopes $\hat{v}_{t+1}^n(R_t^{x,n})$ and $\hat{v}_{t+1}^n(R_t^{x,n}+\rho)$ and it is always the case that $\hat{v}_{t+1}^n(R_t^{x,n}) \geq \hat{v}_{t+1}^n(R_t^{x,n}+\rho)$, implying that the resulting temporary slope $z_t^n(R_t^{x,n}, W_t^n)$ is bigger than $z_t^n(R_t^{x,n}+\rho, W_t^n)$. Therefore, according to our projection operator, the updated slopes $\bar{v}_t^n(R_t^{x,n}, W_t^n)$ and $\bar{v}_t^n(R_t^{x,n}+\rho, W_t^n)$ are always equal to $z_t^n(R_t^{x,n}, W_t^n)$ and $z_t^n(R_t^{x,n}+\rho, W_t^n)$, respectively. Due to our stepsize rule, as described in section 3, the slopes corresponding to $(R_t^{x,n}, W_t^n)$ and $(R_t^{x,n}+\rho, W_t^n)$ are the only ones updated due to a direct observation of sample slopes at iteration n, time period t. All the other slopes are modified only if a violation of the monotone decreasing property occurs. Therefore, the slopes corresponding to states with information vector $W_t \in \mathcal{W}$ different than W_t^n, no matter the resource level $R = 1, \ldots, B^R$, remain the same at iteration n time period t, that is, $\bar{v}_t^{n-1}(R, W_t) = z_t^n(R, W_t) = \bar{v}_t^n(R, W_t)$. On the other hand, it is always the case that the temporary slopes corresponding to states with information vector W_t^n and resource levels smaller than $R_t^{x,n}$ can only be increased by the projection operation. If necessary, they are increased to be equal to $\bar{v}_t^n(R_t^{x,n}, W_t^n)$. Similarly, the temporary slopes corresponding to states with information vector W_t^n and resource levels greater than $R_t^{x,n}+\rho$ can only be decreased by the projection operation. If necessary, they are decreased to be equal to $\bar{v}_t^n(R_t^{x,n}+\rho, W_t^n)$.

Keeping the previous discussion in mind, it is easy to see that for each $W_t^* \in \mathcal{W}_t$, if R_t^{Min} is the minimum resource level such that (R_t^{Min}, W_t^*) is an accumulation point of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$, then the slope corresponding to (R_t^{Min}, W_t^*) could only be decreased by the projection operation a finite number of iterations, as a decreasing requirement could only be originated from a resource level smaller than R_t^{Min}. However, no state with information vector W_t^* and resource level smaller than R_t^{Min} is visited by the algorithm after iteration $\bar{N}$ (as defined in section 4.2), since only accumulation points are visited after $\bar{N}$. We thus have that (R_t^{Min}, W_t^*) is an element of the set $\tilde{\mathcal{S}}_t$, showing that

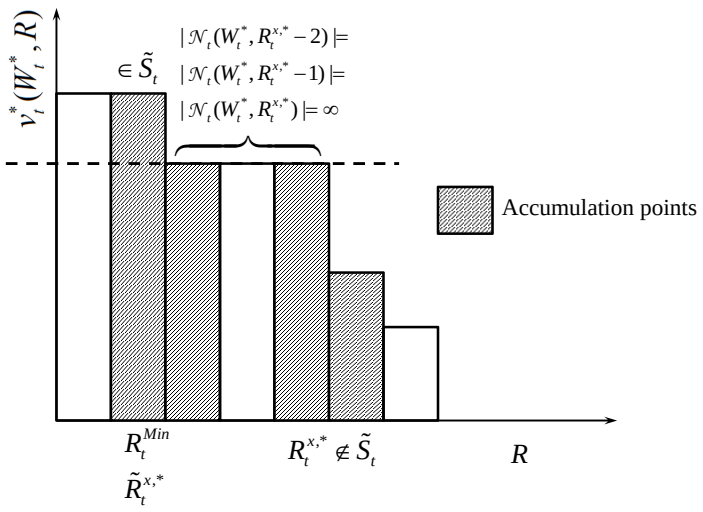


Figure 5: Illustration of technical elements related to the projection operation

$\tilde{\mathcal{S}}_t$ is a proper set.

Hence, for all states $(R_t^{x,*}, W_t^*)$ that are accumulation points of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$ and are not elements of $\tilde{\mathcal{S}}_t$, there exists another state $(\tilde{R}_t^{x,*}, W_t^*)$, where $\tilde{R}_t^{x,*}$ is the maximum resource level smaller than $R_t^{x,*}$ such that $(\tilde{R}_t^{x,*}, W_t^*) \in \tilde{\mathcal{S}}_t$. We argue that for all resource levels R between $\tilde{R}_t^{x,*}$ and $R_t^{x,*}$ (inclusive), we have that $|\mathcal{N}_t(R, W_t^*)| = \infty$. Figure 5 illustrates the situation.

As introduced in section 4.2, we have that $\mathcal{N}_t(R, W_t^*) = \{n \in \mathbb{N} : z_t^n(R, W_t^*) > \bar{v}_t^n(R, W_t^*)\}$. By definition, the sets $\tilde{\mathcal{S}}_t$ and $\mathcal{N}_t(R, W_t^*)$ share the following relationship. Given that (R, W_t^*) is an accumulation point of $\{(R_t^{x,n}, W_t^n)\}_{n\geq 0}$, then $|\mathcal{N}_t(R, W_t^*)| = \infty$ if and only if the state (R, W_t^*) is not an element of $\tilde{\mathcal{S}}_t$. Therefore, $|\mathcal{N}_t(R_t^{x,*}, W_t^*)| = \infty$, otherwise $(R_t^{x,*}, W_t^*)$ would be an element of $\tilde{\mathcal{S}}_t$. If $\tilde{R}_t^{x,*} = R_t^{x,*} - 1$ we are done. If $\tilde{R}_t^{x,*} < R_t^{x,*} - 1$, we have to consider two cases, namely $(R_t^{x,*} - 1, W_t^*)$ is an accumulation point and $(R_t^{x,*} - 1, W_t^*)$ is not an accumulation point. For the first case, we have that $|\mathcal{N}_t(R_t^{x,*} - 1, W_t^*)| = \infty$ from the fact that this state is not an element of $\tilde{\mathcal{S}}_t$. For the second case, since $(R_t^{x,*} - 1, W_t^*)$ is not an accumulation point, its corresponding slope is never updated due to a direct observation of sample slopes for $n \geq \bar{N}$, by the definition of $\bar{N}$. Moreover, every time the slope of $(R_t^{x,*}, W_t^*)$ is decreased due to a projection (which is coming from the left), the slope of $(R_t^{x,*} - 1, W_t^*)$ has to be decreased as well. Therefore, $\mathcal{N}_t(R_t^{x,*}, W_t^*) \bigcap \{n \geq \bar{N}\} \subseteq \mathcal{N}_t(R_t^{x,*} - 1, W_t^*) \bigcap \{n \geq \bar{N}\}$, implying that $|\mathcal{N}_t(R_t^{x,*} - 1, W_t^*)| = \infty$. We then apply the same reasoning for states $(R_t^{x,*} - 2, W_t^*), \ldots, (\tilde{R}_t^{x,*} + \rho, W_t^*)$, obtaining that the corresponding sets of iterations have an infinite number of elements. The same reasoning applies to states $(R_t^{x,*} + \rho, W_t^*)$

that are not in $\tilde{\mathcal{S}}_t$.

Now, pick $k \geq 0$ and a state $(R_t^{x,*}, W_t^*)$ that is an accumulation point but is not in $\tilde{\mathcal{S}}_t$. The same applies if $(R_t^{x,*} + \rho, W_t^*) \notin \tilde{\mathcal{S}}_t$. Consider the state $(\tilde{R}_t^{x,*}, W_t^*)$ where $\tilde{R}_t^{x,*}$ is the maximum resource level smaller than $R_t^{x,*}$ such that $(\tilde{R}_t^{x,*}, W_t^*) \in \tilde{\mathcal{S}}_t$. This state satisfies the condition of lemma 3 with $N^k(\tilde{R}_t^{x,*}, W_t^*) = N_t^k$ (from part 1 of the proof). Thus, we can apply this lemma in order to obtain, for all $k \geq 0$, an integer $N^k(\tilde{R}_t^{x,*} + \rho, W_t^*)$ such that $L_t^k(\tilde{R}_t^{x,*} + \rho, W_t^*) \leq \bar{v}_t^{n-1}(\tilde{R}_t^{x,*} + \rho, W_t^*)$, for all $n \geq N^k(\tilde{R}_t^{x,*} + \rho, W_t^*)$.

After that, we use lemma 3 again, this time considering state $(\tilde{R}_t^{x,*} + \rho, W_t^*)$. Note that the first application of lemma 3 gave us the integer $N^k(\tilde{R}_t^{x,*} + \rho, W_t^*)$, necessary to fulfill the conditions of this second usage of the lemma. We repeat the same reasoning, applying lemma 3 successively to the states $(\tilde{R}_t^{x,*} + 2, W_t^*), \ldots, (R_t^{x,*} - 1, W_t^*)$. In the end, we obtain, for each $k \geq 0$, an integer $N^k(R_t^{x,*}, W_t^*)$, such that $L_t^k(R_t^{x,*}, W_t^*) \leq \bar{v}_t^{n-1}(R_t^{x,*}, W_t^*)$, for all $n \geq N^k(R_t^{x,*}, W_t^*)$. For additional discussion and illustration of this logic, we refer to N&P.

Finally, if we pick $N_t^{*,k}$ to be greater than N_t^k of part 1 and greater than $N^k(R_t^{x,*}, W_t^*)$ and $N^k(R_t^{x,*} + \rho, W_t^*)$ for all accumulation points $(R_t^{x,*}, W_t^*)$ that are not in $\tilde{\mathcal{S}}_t$, then (41) is true for all accumulation points and $n \geq N_t^{*,k}$.

References

Ahmed, S. and Shapiro, A. (2002), The sample average approximation method for stochastic programs with integer recourse. E-print available at http://www.optimization-online.org.

Ahuja, R. K., Magnanti, T. L. and Orlin, J. (1993), 'Network Flows:', *Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs, Ahuja, R. K., J. B. Orlin, D. Sharma* **91**, 71–97.

Antos, A., Szepesvari, C. and Munos, R. (2007), 'Value-Iteration Based Fitted Policy Iteration: Learning with a Single Trajectory', *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning* pp. 330–337.

Antos, A., Szepesvári, C. and Munos, R. (2008*a*), 'Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path', *Machine Learning* **71**(1), 89–129.

Antos, A., Szepesvari, C. and Munos, R. (2008*b*), 'Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path', *Machine Learning* **71**(1), 89–129.

Barto, A. G., Bradtke, S. J. and Singh, S. P. (1995), 'Learning to act using real-time dynamic programming', *Artificial Intelligence, Special Volume on Computational Research on Interaction and Agency* **72**, 81–138.

Bertsekas, D., Nedic, A. and Ozdaglar, E. (2003), *Convex Analysis and Optimization*, Athena Scientific, Belmont, Massachusetts.

Bertsekas, D. P. (2005), *Dynamic Programming and Stochastic Control Vol. I*, Athena Scientific.

Bertsekas, D. P. (2011), 'Approximate policy iteration: a survey and some new methods', *Journal of Control Theory and Applications* **9**(3), 310–335.

Bertsekas, D. P. (2012*a*), Approximate Dynamic Programming, *in* 'Dynamic Programming and Optimal Control Volume II', 3rd edn, Vol. II, Cambridge, MA, chapter Chapter 6.

Bertsekas, D. P. (2012*b*), 'Dynamic Programming and Optimal Control 3rd Edition , Volume II by Chapter 6 Approximate Dynamic Programming Approximate Dynamic Programming', *Control* **II**.

Bertsekas, D. P., Abounadi, J. and Borkar, V. (2003), 'Stochastic approximation for nonexpansive maps: Application to Q-learning algorithms', *SIAM Journal on Control and Optimization* **41**(1), 1–22.

Bertsekas, D. and Tsitsiklis, J. (1996), *Neuro-Dynamic Programming*, Athena Scientific, Belmont, MA.

Borkar, V. S. and Meyn, S. P. (2000), 'The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning', *SIAM Journal on Control and Optimization* **38**(2), 447.

Chen, Z.-L. and Powell, W. B. (1999), 'A convergent cutting-plane and partial-sampling algorithm for multistage stochastic linear programs with recourse', *Journal of Optimization Theory and Applications* **102**(3), 497–524.

Enders, J., Powell, W. B. and Egan, D. M. (2010), 'Robust policies for the transformer acquisition and allocation problem', *Energy Systems* **1**(3), 245–272.

George, A., Powell, W. B. and Kulkarni, S. (2008), 'Value Function Approximation using Multiple Aggregation for Multiattribute Resource Management', *J. Machine Learning Research* **9**, 2079–2111.

Godfrey, G. and Powell, W. B. (2002), 'An adaptive, dynamic programming algorithm for stochastic resource allocation problems I: Single period travel times', *Transportation Science* **36**(1), 21–39.

Gordon, G. J. (2001), 'Reinforcement learning with function approximation converges to a region', *Advances in Neural Information Processing Systems* **13**, 1040–1046.

Hannah, L. A., Powell, W. B. and Blei, D. M. (2010), Nonparametric Density Estimation for Stochastic Optimization with an Observable State Variable, *in* 'Neural Information Processing Society', Vancouver, BC, pp. 1–9.

He, M., Zhao, L. and Powell, W. B. (2010), 'Optimal control of dosage decisions in controlled ovarian hyperstimulation', *Annals of Operations Research* **178**, 223–245.

Hernandez-Lerma, O. and Runggladier, W. (1994), 'Monotone approximations for convex stochastic control problems', *J. Math. Syst., Estim. and Control* **4**, 99–140.

Higle, J. and Sen, S. (1991), 'Stochastic decomposition: An algorithm for two stage linear programs with recourse', *Mathematics of Operations Research* **16**(3), 650–669.

Jaakkola, T., Jordan, M. and Singh, S. P. (1994), 'On the convergence of stochastic iterative dynamic programming algorithms', *Neural Computation* **1201**(1988), 1185–1201.

Munos, R. and Szepesvari, C. (2008), 'Finite-Time Bounds for Fitted Value Iteration', *Journal of Machine Learning Research* **1**, 815–857.

Nascimento, J. M. and Powell, W. B. (2009), 'An Optimal Approximate Dynamic Programming Algorithm for the Lagged Asset Acquisition Problem', *Mathematics of Operations Research* **34**(1), 210–237.

Powell, W. B. (2011), *Approximate Dynamic Programming: Solving the curses of dimensionality*, 2nd. edn, John Wiley & Sons, Hoboken, NJ.

Powell, W. B., George, A., Lamont, A. and Stewart, J. (2011), 'SMART: A Stochastic Multiscale Model for the Analysis of Energy Resources, Technology and Policy', *Informs J. on Computing* .

Powell, W. B., Ruszczyński, A. and Topaloglu, H. (2004), 'Learning algorithms for separable approximations of stochastic optimization problems', *Mathematics of Operations Research* **29**(4), 814–836.

Precup, D. and Perkins, T. (2003), 'A convergent form of approximate policy iteration', *Advances in neural information processing systems* pp. 1627–1634.

Shapiro, A. (2003), Monte carlo sampling methods, *in* A. Ruszczyński and A. Shapiro, eds, '*Handbooks in Operations Research and Management Science: Stochastic Programming*', Vol. 10, Elsevier, Amsterdam, pp. 353–425.

Shapiro, A., Dentcheva, D. and Ruszczynski, A. (2009), *Lectures on stochastic programming: modeling and theory*, SIAM, Philadelphia.

Shiryaev, A. (1996), *Probability Theory*, Vol. 95 of *Graduate Texts in Mathematics*, Springer-Verlag, New York.

Sutton, R. and Barto, A. (1998), *Reinforcement Learning*, The MIT Press, Cambridge, Massachusetts.

Szepesvari, C. (2010), 'Algorithms for Reinforcement Learning', *Synthesis Lectures on Artificial Intelligence and Machine Learning* **4**(1), 1–103.

Szita, I. (2007), Rewarding excursions: extending reinforcement learning to complex domains, PhD thesis.

Topaloglu, H. and Powell, W. B. (2006), 'Dynamic programming approximations for stochastic, time-staged integer multicommodity flow problems', *Informs Journal on Computing* **18**(1), 31–42.

Tsitsiklis, J. N. (1994), 'Asynchronous stochastic approximation and Q-learning', *Machine Learning* **16**, 185–202.

Van Slyke, R. and Wets, R. (1969), 'L-shaped linear programs with applications to optimal control and stochastic programming', *SIAM Journal of Applied Mathematics* **17**(4), 638–663.

Werbos, P. J. (1989), 'Backpropagation and neurocontrol: A review and prospectus', *Neural Networks* pp. 209—-216.