Approximately Optimal Control of MDPs over a Long Operation-Dependent Time Horizon and Application to Battery Energy Storage Systems

Ioannis Kordonis, Alexandros C. Charalampidis, Pierre Haessig

***Abstract*—This paper considers the optimal control problem for Markov Decision Processes (MDPs), with a time horizon which is both long and operation-dependent. The motivation for this study comes from the optimal management problem for grid-connected Battery Energy Storage Systems (BESSs). As the investment cost for BESSs represents a large portion of the total cost, and the BESSs can be used only for a limited number of cycles, it is important to maximize the benefits of the use of a BESS over its lifetime, which, however, depends on its actual operation. First, we prove that the generic optimal control problem can be approximated by the minimization of the ratio of two infinite-horizon average-cost problems. We then characterize the optimal policy in terms of appropriate Bellman-type equations, for infinite-horizon average-cost problems, propose a Relative Value Iteration type algorithm and prove its convergence. Finally, we present some numerical results illustrating the efficiency of the proposed methods for the BESS application.**

***Index Terms*— Stochastic optimal control, Energy systems, Optimization algorithms, Battery energy storage systems**

I. Introduction

This paper studies the optimal control of stochastic systems over their lifetime, in the case where the lifetime depends on their actual operation. This study's motivation comes from the study of the optimal operation of grid-connected Battery Energy Storage Systems (BESSs). Batteries are expected to play an essential role in the future power system. However, despite the continuing battery cost reduction, the initial investment costs for BESSs remain high and constitute one of the highest components of BESSs cost. Furthermore, batteries are degrading due to calendar aging and usage (cycling aging), and their lifetime depends on their actual operation. Thus, to optimize the profits from a BESS operation, it is essential to take into account the effects of BESS usage on its lifetime.

The authors are with CentraleSupélec, Automatic Control Group - IETR, Avenue de la Boulaie, 35576 Cesson-Sévigné, France. A. Charalampidis is additionally affiliated with the Department of Electrical Engineering and Computer Science, Technische Universität Berlin, Control Systems Group, Einsteinufer 17, Berlin D-10587, Germany.
Author emails: I. Kordonis: jkordonis1920@yahoo.com, A.C. Charalampidis: alexandros.charalampidis@centralesupelec.fr, P. Haessig: pierre.haessig@centralesupelec.fr.

This work has been supported by the Regional Council of Brittany (SAD 2016 - REINMASE, 9693) and People Programme (Marie Curie Actions) of the European Unions Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n. PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France. The work of A. C. Charalampidis has been also supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 705982.

Battery aging is a complex nonlinear phenomenon, and various battery aging models have been proposed [1]. Still, most of the works on optimal management of grid-connected BESSs use simplified models. The articles [2]–[6] assume that there is a certain cost for using the battery, proportional to the initial investment cost. Other papers put certain constraints on battery usage [7]–[9]. Another relevant idea is to use separation of time scales [10], [11] (see also [12]). However, most of the works do not consider the impact of BESS usage on the problem's time horizon. Notable exceptions are [13], [14], which formulate the problem as a Stochastic Shortest Path (SSP) problem. To achieve this, however, the state space should be extended to include the total degradation up to the current time step. This leads to huge state space cardinalities and makes the analysis and optimization of detailed models computationally difficult.

In this work, we study the general MDP optimal control problem with long, operation-dependent time horizon, and approximately reformulate it into the problem of minimizing the ratio of two long-time average-cost criteria. This latter problem dates back to the 1960s, where problems with ratio objectives were studied in the context of maintenance/replacement scheduling [15], [16]. Derman in [15] proves the existence of optimal strategies within the class of memory-less, stationary policies, and reformulates the problem initially as a fractional linear program and then as a linear program. Ratio objectives also appear in the optimal control of Semi-Markov Decision Processes (SMDPs) [17], [18] (see also [19], [20]). There is also a large body of works applying Reinforcement Learning schemes to SMDPs (e.g. [21]–[24]). Recently, similar problems involving a ratio of two long-time average-cost functions were studied in the context of synthesizing systems which achieve an optimal trade-off between a given cost and reward model [25]–[27]. The difference of this work with the above-mentioned ones is that in all these works the cost function in the denominator can take only positive (or in some cases non-negative) values. There are, however, interesting problems involving a long, operation-dependent time horizons, for which the corresponding ratio objective problem has a denominator that could take both positive and negative values (see for example Section VII).

The technical contribution of the current work is threefold. First, we show that the solution of the optimal control problems with long, operation dependent time horizon can be approximated by ratio cost problems. Second, we extend theory for the optimal control of cost functions with ratio objectives to

the case of indefinite denominators (namely, they could take both positive and negative values). Particularly, the optimal policy for a ratio of two long-time average-cost problems is first characterized by a pair of coupled Bellman equations (fixed-point characterization). Based on this formulation, we propose a simple relative value iteration scheme. We then provide a simpler characterization using a single Bellman-type equation and propose a value iteration type algorithm to solve it. The idea for constructing and proving the convergence of this algorithm comes from the proof of convergence of the value iteration scheme proposed in [28] for long-time average-cost problems. The third contribution is the application of the proposed methods to a BESS example.

The rest of this paper is organized as follows: Section II introduces some notation and presents some results of Stochastic Optimal Control theory. In Section III, the optimal control problem with long operation-dependent time horizon is stated and the approximate reduction to a ratio of two long-time average-cost problems is proved. In Section IV, the optimal policy is characterized in terms of a pair of coupled Bellman-type equations. Section V derives a simpler characterization of the optimal solution in terms of a single Bellman equation, presents a relative value iteration algorithm and proves its convergence. Section VI presents some numerical results, for the BESS application. Finally, Section VIII summarizes the main contributions of the current work.

II. Background, Notation and Preliminary Results

We consider stochastic dynamics of the form:

$$x_{k+1} = f(x_k, u_k, w_k), \tag{1}$$

where x_k represents the state variable, u_k the control action and w_k a stochastic disturbance with known distribution depending only on x_k. We assume that x_k has a finite state space $X = \{1, \dots, n\}$ and the admissible set of control actions u_k is a finite set $U = \{1, \dots, n_U\}$. The controlled Markov chain notation $p_{ij}(u) = P[x_{k+1} = j | x_k = i, u_k = u]$, will be also used. We assume that the distribution of the initial condition x_0 is known.

A general policy is a function of the past states and actions, and randomizes among the several possible actions. Particularly, we define the history at time k as $\eta_k = (x_0, u_0, \dots, x_{k-1}, u_{k-1}, x_k)$. A policy is a sequence $\mu = (\mu_0, \mu_1, \mu_2, \dots)$, where μ_k is a function that maps the history η_k to a probability distribution over the feasible actions. This is the most general class of policies considered in this work, and it will be called 'behavioral policies'. A policy μ is called stationary if each μ_k depends only on the current sate x_k and $\mu_k = \mu_0$. A subset of stationary policies is the set of deterministic stationary policies i.e., policies where $\mu_k(x)$ assigns the entire probability to a single element of U, for all states $x \in X$. We denote the class of behavioral, stationary, and deterministic stationary policies by $\mathcal{U}_b, \mathcal{U}_s, \mathcal{U}_d$, respectively.

Let us state a couple of assumptions which are very common in the literature (see for example Bertsekas's textbook [19]).

Assumption 1: For the controlled Markov chain (1) it holds:

(a) **(Weak Accessibility (WA)):** The state space X can be partitioned into two subsets X_c and X_t such that:
- The states in X_t are transient for all stationary policies.
- For every pair of states $i, j \in X_c$, there is a stationary policy μ and an integer k such that, under μ, it holds:
$$P(x_k = j | x_0 = i) > 0.$$

(b) **(unichain assumption):** For any stationary deterministic policy, the closed loop Markov chain has a single irreducible aperiodic ergodic class.

Let us note that the unichain assumption implies the weak accessibility assumption.

There are several possible formulations for the cost functions. This section, deals with the Long-Time Average (LTA) cost (known also as average-cost per stage or ergodic cost). We consider two different costs per stage $g_1(x_k, u_k)$ and $g_2(x_k, u_k)$, where the first represents an actual cost per stage and the second a degradation rate. The LTA cost is given by:

$$\lambda_s(\mu, i) = \lim_{N\to\infty} \frac{1}{N} E\left[\sum_{k=1}^{N} g_s(x_k, u_k) \middle| x_0 = i, u_k = \mu_k(x_k) \right], \tag{2}$$

for $s = 1, 2$. Note that under Assumption 1(b), for every $\mu \in \mathcal{U}_s$, the cost $\lambda_s(\mu, i)$ does not depend on i. In this case we write $\lambda_s(\mu)$ in the place of $\lambda_s(\mu, i)$. Furthermore, under Assumption 1.a, the optimal cost for the LTA problem is the same for all initial states.

The optimal policy $\mu^\star$ is stationary, and under Assumption 1.a, is characterized by the corresponding Bellman equation:

$$\lambda + h_s(i) = \min_u \left\{ g_s(i, u) + \sum_{j=1}^{n} p_{ij}(u) h_s(j) \right\}, \tag{3}$$

for all $i \in X$, where h_s is a vector in $\mathbb{R}^n$ called bias (e.g. [19], [20]).

Following Bertsekas ([19]), let us introduce the 'shorthand' notation for this problem. For a stationary policy μ, denote by $\mathcal{T}_{s,\mu}$ the Bellman operator associated with the LTA cost λ_s (with $s = 1, 2$), mapping a vector $h_s \in \mathbb{R}^n$ to a vector $\mathcal{T}_{s,\mu} h_s \in \mathbb{R}^n$, given by:

$$(\mathcal{T}_{s,\mu} h_s)(i) = g_s(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) h_s(j). \tag{4}$$

If the LTA cost under a stationary policy μ (not necessarily optimal) is the same for all the initial conditions i.e., $\lambda_s(\mu, i) = \lambda_{s,\mu}$, for some real number $\lambda_{s,\mu}$, then it holds (Bellman equation for policy evaluation):

$$h_s + \lambda_{s,\mu} \mathbf{1} = \mathcal{T}_{s,\mu} h_s, \tag{5}$$

where $\mathbf{1}$ is the n-vector of ones. Conversely if a vector $h_s \in \mathbb{R}^n$ and a scalar $\lambda_{s,\mu}$ satisfy (5), then the LTA cost under μ is equal to $\lambda_{s,\mu}$, for every initial state.

We are also interested in optimization problems, considering the ratio of two LTA costs, in the form:

$$\begin{array}{ll} \underset{\mu}{\text{minimize}} & \dfrac{\lambda_1(\mu, i)}{\lambda_2(\mu, i)} \\ \text{subject to} & \lambda_2(\mu, i) > 0 \end{array}. \tag{6}$$

The following proposition describes the set of possible values of $\lambda_1(\mu,i), \lambda_2(\mu,i)$.

***Proposition** 1:* Under the weak accessibility assumption (Assumption 1.a), there is a convex compact polygon D such that:

(i) Assume that for a policy $\mu \in \mathcal{U}_b$ the limits in (2) exist for $s = 1, 2$. Then $(\lambda_1(\mu,i), \lambda_2(\mu,i)) \in D$.
(ii) For every vertex of D, with coordinates $(\bar{\lambda}_1, \bar{\lambda}_2)$, there is a deterministic stationary policy $\mu \in \mathcal{U}_d$ such that, for all $i \in X$, $\lambda_1(\mu,i) = \bar{\lambda}_1$ and $\lambda_2(\mu,i) = \bar{\lambda}_2$.
(iii) For every point of $(\bar{\lambda}_1, \bar{\lambda}_2) \in D$ there is a behavioral policy $\mu \in \mathcal{U}_b$ such that, for all $i \in X$, $\lambda_1(\mu,i) = \bar{\lambda}_1$ and $\lambda_2(\mu,i) = \bar{\lambda}_2$.

Proof: See Appendix A.

We call D the 'feasible region'. If for a policy $\mu \in \mathcal{U}_d$ there is a vertex of D with coordinates $(\bar{\lambda}_1, \bar{\lambda}_2)$ such that $\lambda_1(\mu,i) = \bar{\lambda}_1$ and $\lambda_2(\mu,i) = \bar{\lambda}_2$, for all $i \in X$, we call μ a 'corner policy'.

***Remark** 1:* Results similar to Proposition 1 appear in the literature of constrained MDPs, under the unichain assumption (e.g. [29]).

III. Problem Formulation and Approximate Infinite-Horizon Reformulation

In this section we present the general form of the optimal control problem with operation-dependent time horizon, as well as an infinite-horizon approximate reformulation. Consider a dynamics in the form (1). Assume that the system has initially a remaining life denoted by R i.e., R denotes the initial total degradation capacity of the system. During time period k, the remaining life is reduced by $g_2(x_k, u_k)$ (note that g_2 could be positive or negative). Thus, the time horizon of the problem is random and it is given by the stopping time:

$$T = \inf\left\{ t : \sum_{k=1}^{t} g_2(x_k, u_k) \geq R \right\}. \tag{7}$$

We are interested in solving the problem:

$$\underset{\mu}{\text{minimize}}\ J_a(\mu) = \lim_{N\to\infty} E\left[\sum_{k=0}^{T\wedge N-1} g_1(x_k, u_k) \right] / R, \tag{8}$$

where $T \wedge N = \min(T, N)$, $\mu = (\mu_1, \mu_2, \dots)$ and u_k is computed according to $\mu_k(x_k)$. In the rest of the paper we refer to (8) as the 'original problem'. The cost function J_a represents the total cost over the operation-dependent time horizon, scaled by the initial remaining life. Indeed, we expect the total cost $\lim_{N\to\infty} E\left[\sum_{k=1}^{T\wedge N} g_1(x_k, u_k)\right]$ to be proportional to R, for large R, and thus we use the scaling by R to make the cost approximately independent of it.

***Remark** 2:* The original Problem (8) can be written as a Stochastic Shortest Path problem. To this end, the system state space should be extended to include the cumulative degradation $y_k = \sum_{k'=1}^{k} g_2(x_k, u_k)$. Let Y denote the possible values of y_k. Then, a natural state space for (8) is $\bar{X} = X \times Y$. For large R the cardinality of the state space of the problem becomes huge. In the case where $g_2(x,u)$ can take negative values, Y becomes countably infinite. In the next subsection, we propose an approximate reformulation of Problem (8) as a ratio of two LTA costs, which allows an efficient solution.

A. Problem Reformulation

We then perform a series of approximate transformations to the original Problem (8), which eventually leads to a problem with ratio objectives. Assuming that R is very large, we approximate the original problem (8) by its 'deterministic horizon' counterpart:

$$\underset{T_d\in\mathbb{N}}{\text{minimize}} \left\{ \begin{array}{ll} \min\limits_{\mu} & E\left[\sum\limits_{k=1}^{T_d} g_1(x_k,u_k)\right]/R \\ \text{subject to} & \left|\sum\limits_{k=1}^{T_d} E[g_2(u_k)] - R\right| \leq \bar{G} \end{array} \right\}, \tag{9}$$

where T_d is a deterministic decision variable and $\bar{G} \ll R$. The constraint represents the fact that the expected discrete degradation is approximately equal to R. The internal minimization of (9) i.e., the corresponding problem with a large fixed T_d, can be further approximated by its infinite-horizon counterpart. To this end, we use $\lambda_1(\mu,i)$ and $\lambda_2(\mu,i)$ defined in (2).

The random variable T could take finite or infinite values. However, we are interested in cases where, under the optimal policy, the expected time horizon is finite and appropriate assumptions will be made in the following (Assumption 2) to ensure that this happens. There are two qualitatively distinct cases. In the first case, there is a policy μ under which the system is producing value i.e., $\lambda_1(\mu,i) < 0$. In this case we are interested in delaying the stopping time T (keeping the system alive), while enjoying negative costs. In the second case every policy produces losses i.e, for every policy μ it holds $\lambda_1(\mu,i) > 0$. In this case, we are interested in making the time horizon T as short as possible.

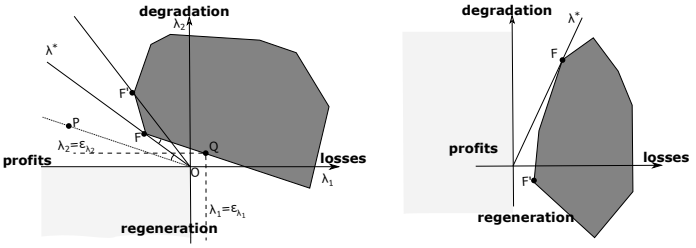
In the case where there exists a value producing policy, we assume that if for a policy μ we have $\lambda_1(\mu,i) < 0$, then it also holds $\lambda_2(\mu,i) > 0$. Otherwise, many of the system sample paths would produce $-\infty$ cost. Conversely, if all policies produce losses we assume that there is a policy μ such that $\lambda_2(\mu,i) > 0$. Otherwise, all the policies will lead to infinite total cost. We will make thus the following assumption.

Assumption 2: For every initial condition $i \in X$ one of the following is true:

(a) **(There exists a value producing policy)** There is a policy μ such that $\lambda_1(\mu,i) < 0$, and for every policy μ' it holds: either $\lambda_1(\mu',i) > 0$ or $\lambda_2(\mu',i) > 0$ (see Figure 1.a).
(b) **(All policies produce losses)** For every policy μ, it holds $\lambda_1(\mu,i) > 0$. Furthermore, there is a policy μ' such that $\lambda_2(\mu',i) > 0$ (see Figure 1.b).

The 'deterministic horizon' Problem (9) can be approximately restated as:

$$\begin{array}{ll} \underset{T_d\in\mathbb{R}_+,\ \mu}{\text{minimize}} & J_b(\mu,i) = T_d\lambda_1(\mu,i)/R \\ \text{subject to} & T_d\lambda_2(\mu,i) = R \\ & \lambda_2(\mu) > 0 \end{array}, \tag{10}$$



(a) There exists a value producing policy **(b)** All policies produce losses

Fig. 1: The dark gray areas in parts (a) and (b) of this figure correspond to the feasible regions D, under Case (a) and Case (b) of Assumption 2 respectively. The light gray areas represent areas forbidden by Assumption 2. In both cases we denote by F the point which is optimal for (10) (or equivalently for (11)). If the manager of the system ignores lifetime effects (degradation/regeneration) he/she will choose the point F', which is the one minimizing λ_1. In Case (a), it turns out that F' gives larger benefits per unit time, but much less time horizon compared with F, and thus it turns out to be inefficient. In Case (b), F' produces the smallest possible loss per unit time, but it leads to a negative λ_2 and thus on average these losses last forever and the total expected cost is infinite. Let us note that the part (a) of this figure is also used in the Appendix, where the points Q and P are also defined.

for large R. Let us call this problem the 'infinite-horizon approximate problem'. Substituting the constraint back to the objective, (10) can be equivalently stated as a 'ratio cost problem':

$$\begin{array}{ll} \underset{\mu}{\text{minimize}} & J_c(\mu,i)=\lambda_1(\mu,i)/\lambda_2(\mu,i) \\ \text{subject to} & \lambda_2(\mu)>0 \end{array}. \tag{11}$$

We then show that if an optimal policy for the 'ratio cost problem' is applied to the original Problem (8) it will be ε-optimal.

***Theorem** 1 (Problem Approximation):* Suppose that Assumption 1.a and Assumption 2 hold. Then:

(i) Problem (11) is feasible and there is a corner optimal policy $\mu^\star \in \mathcal{U}_d$ minimizing J_c for all i. Furthermore, the optimal value is the same for all $i \in X$.

(ii) For every $\varepsilon > 0$, there is an R_0 such that any corner optimal policy for 'ratio cost problem' (11), is ε-optimal for the original Problem (8), for any $R > R_0$. Furthermore, if for a policy μ, the original cost J_a is finite, then the expected time horizon $E[T]$ is finite and the original cost may be written equivalently as:

$$J_a = E\left[\sum_{k=0}^{T} g_1(x_k,u_k)\right]/R. \tag{12}$$

Proof: See Appendix B. □

***Remark** 3:* This is our last reformulation and (11) is an important topic of the rest of this paper. While (11) has a much smaller state space than (8), dynamic programming is not directly applicable to it. In Sections IV and V we propose techniques to solve (11) efficiently.

***Remark** 4:* Problems involving the ratio of two LTA costs have been already studied in the literature (e.g. [15] - [27]). In these works the authors assume that the cost function in the denominator g_2 can take only positive (or non-negative) values. Here we generalize the theory to the case where g_2 can also take negative values.

***Remark** 5:* Case (b) of Assumption 2 can be reduced to case (a). To this end, suppose that Assumption 2.b holds and observe that:

$$\begin{aligned}\arg\min_{\bar\mu}\{\lambda_1(\bar\mu)/\lambda_2(\bar\mu):\lambda_2(\bar\mu)>0\} = \\ = \arg\min_{\bar\mu}\{-\lambda_2(\bar\mu)/\lambda_1(\bar\mu):\lambda_2(\bar\mu)>0\}.\end{aligned}$$

But, since λ_1 is positive under any policy, the minimum in the right-hand side is attained for a $\bar\mu^\star$, such that $\lambda_2(\bar\mu^\star)$ is positive (see Assumption 2.b). Thus, using $\lambda_1' = -\lambda_2$ and $\lambda_2' = \lambda_1$ we end up with a problem of the same form. This transformation corresponds to a 90° counterclockwise rotation of Figure 1.(b).

IV. Fixed Point Characterization and the FPRVI Algorithm

In this section, we characterize the corner optimal policies of (11) in terms of a pair of coupled Bellman-type equations (fixed point characterization). This characterization is then used to propose a value iteration type algorithm.

A. Fixed Point Characterization of the Optimal Policy

***Proposition** 2 (Fixed Point Characterization):* Suppose that Assumption 1.a and Assumption 2.a hold.

(i) Assume that for a policy $\mu^\star$ there exist vectors $h_1^\star, h_2^\star \in \mathbb{R}^n$ and scalars $\lambda_1^\star, \lambda_2^\star \in \mathbb{R}$ satisfying the fixed point equations:

$$\mathcal{T}_{1,\mu^\star} h_1^\star = \lambda_1^\star \mathbf{1} + h_1^\star, \tag{13}$$

$$\mathcal{T}_{2,\mu^\star} h_2^\star = \lambda_2^\star \mathbf{1} + h_2^\star, \tag{14}$$

$$\mu^\star \in \arg\min_{\mu}[\lambda_2^\star \mathcal{T}_{1,\mu} h_1^\star - \lambda_1^\star \mathcal{T}_{2,\mu} h_2^\star], \tag{15}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the vector consisting of ones, in (15) the $\arg\min$ is considered for each component separately, and $T_{s,\mu}$ is given in (4). Then, $\mu^\star$ is optimal for (11), and the optimal value satisfies $\lambda^\star = \lambda_1^\star/\lambda_2^\star$.

(ii) Conversely, if $\mu^\star$ is a corner optimal policy, then it satisfies (13)-(15).

Proof: Let $\mu^\star$ be the optimal policy for (11) i.e., $\lambda_2(\mu^\star) > 0$ and $\lambda_1(\mu,i)/\lambda_2(\mu,i) \geq \lambda_1(\mu^\star)/\lambda_2(\mu^\star)$, for every policy μ with $\lambda_2(\mu,i) > 0$. By Assumption 2.a, the point $(\lambda_1(\mu^\star), \lambda_2(\mu^\star))$ lies in the interior of the second quadrant (see Figure 2.a) and thus $\lambda_2(\mu^\star) > 0$. Hence, if $\mu^\star$ is optimal then:

$$\lambda_2(\mu^\star)\lambda_1(\mu,i) - \lambda_1(\mu^\star)\lambda_2(\mu,i) \geq 0, \tag{16}$$

for every μ and $i \in X$ with $\lambda_2(\mu,i) > 0$. Furthermore, if there is an $i \in X$ and a policy μ such that $\lambda_2(\mu,i) \leq 0$ and $\lambda_2(\mu^\star)\lambda_1(\mu,i) - \lambda_1(\mu^\star)\lambda_2(\mu,i) < 0$, then the feasible set D intersects the third quadrant which contradicts Assumption 2.

Thus, if $\mu^\star$ is optimal then (16) holds true for all the policies μ.

Conversely assume that a policy $\mu^\star$ satisfies (16). Then from, and Assumption 2.a we conclude that $(\lambda_1(\mu^\star), \lambda_2(\mu^\star))$ cannot be in the first quadrant, because there are points of D in the second quadrant, and thus (16) cannot hold. We then show that, $(\lambda_1(\mu^\star), \lambda_2(\mu^\star))$ cannot be in the fourth quadrant. Assume the contrary i.e., $\lambda_1(\mu^\star) > 0$, $\lambda_2(\mu^\star) < 0$. Using Assumption 2.a we consider a point $(\lambda_1(\mu, i), \lambda_2(\mu, i))$ in the second quadrant. Then, take the convex combination $[A(\lambda_1(\mu, i), \lambda_2(\mu, i)) + (1-A)(\lambda_1(\mu^\star), \lambda_2(\mu^\star))] \in D$, where $A = \lambda_1(\mu^\star)/(\lambda_1(\mu^\star) - \lambda_1(\mu, i))$. Observe that $A\lambda_1(\mu, i) + (1-A)\lambda_1(\mu^\star) = 0$ and $A\lambda_2(\mu, i) + (1-A)\lambda_2(\mu^\star) \leq 0$, which violates assumption 2.a. Finally, Assumption 2.a implies that $(\lambda_1(\mu^\star), \lambda_2(\mu^\star))$ cannot belong to the third quadrant. Hence, $(\lambda_1(\mu^\star), \lambda_2(\mu^\star))$ should belong to the second quadrant.

Dividing (16) by $\lambda_2(\mu, i)\lambda_2(\mu^\star)$ we conclude that $\lambda_1(\mu, i)/\lambda_2(\mu, i) \geq \lambda_1(\mu^\star)/\lambda_2(\mu^\star)$, for every policy μ with $\lambda_2(\mu, i) > 0$.

Now (16) holds true for every μ if and only if:

$$\mu^\star \in \arg\min_{\mu} \{\lambda_2(\mu^\star)\lambda_1(\mu) - \lambda_1(\mu^\star)\lambda_2(\mu)\}. \tag{17}$$

(i) Let $\mu^\star$ a policy satisfying (13)-(15). Multiplying (13) by $\lambda_2^\star$ and subtracting (14) times $\lambda_1^\star$, we get:

$$\lambda_2^\star h_1^\star(i) - \lambda_1^\star h_2^\star(i) = \lambda_2^\star g_1(i, u) - \lambda_1^\star g_2(i, u) + \\ + \sum_{j=1}^{n} p_{ij}(u)[\lambda_2^\star h_1^\star(j) - \lambda_1^\star h_2^\star(j)], \tag{18}$$

for $u = \mu^\star(i)$. Thus, (18) and (15) imply (17). Hence, $\mu^\star$ is optimal.

(ii) Using Proposition 1.(ii) and Theorem 1 we conclude that there is an optimal policy $\mu^\star$ for (11) such that $\lambda_1(\mu^\star, i)$ and $\lambda_2(\mu^\star, i)$ do not depend on i. Then, using (5) and (17) we get (15). □

***Remark** 6:* Proposition 2 characterizes corner optimal policies. It is possible that there exist deterministic stationary optimal policies which are not corner. The following proposition shows that this does not happen in the generic case.

***Proposition** 3 (Non-Corner Optimal Policies are Non-Generic):* Suppose that Assumption 1.a holds. Then, the set of vectors of the form $[g_s(x, u) : x \in X, u \in U, s = 1, 2] \in \mathbb{R}^{2nn_U}$ such that there is a non-corner stationary deterministic optimal policy for (11) has Lebesgue measure zero.

Proof: See Appendix C. □

B. The FPRVI Algorithm

The fixed point characterization (13)-(15) can be used to derive a relative value iteration algorithm for Problem (11).

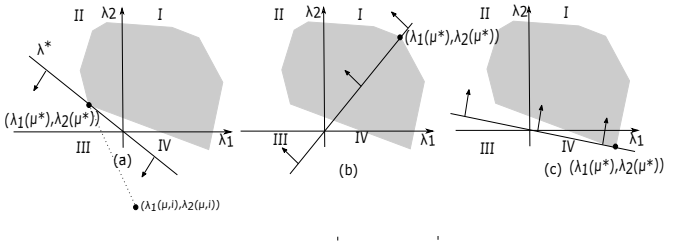


Fig. 2: The possible positions for $(\lambda_1(\mu^\star), \lambda_2(\mu^\star))$. The gray area in these figures correspond to the feasible region D. The three different pictures illustrate three possible positions of $(\lambda_1(\mu^\star), \lambda_2(\mu^\star))$. The arrows correspond to the minus gradient of the function $f(\lambda_1(\mu), \lambda_2(\mu)) = \lambda_2(\mu^\star)\lambda_1(\mu) - \lambda_1(\mu^\star)\lambda_2(\mu)$. Part (a) of the figure corresponds to the correct position of the optimum. Parts (b) and (c) correspond to other candidate solutions in the first and fourth quadrant, which however cannot satisfy (16).

Starting from some initial guesses $\lambda_1^0, \lambda_2^0 \in \mathbb{R}$, $h_1^0, h_2^0 \in \mathbb{R}^n$ and an initial policy μ^0, the iteration is given by:

$$\lambda_1^{k+1} = (\mathcal{T}_{1,\mu^k} h_1^k)(n), \tag{19}$$
$$\lambda_2^{k+1} = (\mathcal{T}_{2,\mu^k} h_2^k)(n), \tag{20}$$
$$h_1^{k+1}(i) = (\mathcal{T}_{1,\mu^k} h_1^k)(i) - \lambda_1^{k+1}, \tag{21}$$
$$h_2^{k+1}(i) = (\mathcal{T}_{2,\mu^k} h_2^k)(i) - \lambda_2^{k+1}, \tag{22}$$
$$\mu^{k+1}(i) = \arg\min_u \Big[\lambda_2^k g_1(i, u) - \lambda_1^k g_2(i, u) + \\ + \sum_{j=1}^{n} p_{ij}(u)(\lambda_2^k h_1^k(j) - \lambda_1^k h_2^k(j))\Big], \tag{23}$$

where $i = 1, \ldots, n$ and n is the last element of the state space X. We will refer to this algorithm as the Fixed Point Relative Value Iteration (FPRVI) algorithm. It is not difficult to see that a fixed point of (19)-(23) satisfies (13)-(15).

The algorithm is intuitively easy to derive, does not have any tuning parameters and works well for the numerical example analyzed (Section VI). However, the analysis of this algorithm is difficult. In the following section, we propose an alternative characterization in terms of a single Bellman equation, which leads to an algorithm with simpler convergence analysis. A comparison of the two algorithms is presented in Remark 8.

V. Bellman Equation Characterization and the RVI Algorithm

A. Bellman Equation Characterization of the Optimal Policy

In this section, we characterize the optimal policy and optimal value in terms of a single Bellman-type equation. This characterization leads to a simple value iteration type algorithm, described in the next subsection.

***Proposition** 4 (Bellman Equation Characterization):* Suppose that Assumption 1.a and Assumption 2.a hold. Then:

(i) If $\lambda^\star$ is the optimal value for the ratio cost function (11) then:

$$\min_{\mu} [\lambda_1(\mu) - \lambda^\star \lambda_2(\mu)] = 0. \tag{24}$$

(ii) There are at most two values of $\lambda^\star$ satisfying (24).

(iii) A policy $\mu^\star$ is optimal for (11) if and only if:

$$\mu^\star \in \arg\min_{\mu}[\lambda_1(\mu) - \lambda^\star\lambda_2(\mu)], \tag{25}$$

where $\lambda^\star$ is the optimal value.

(iv) If $\lambda^\star$ is the optimal value for (11), then there exists a vector $h^\star \in \mathbb{R}^n$ satisfying the Bellman equation:

$$h^\star(i) = \min_u \left[g_1(i,u) - \lambda^\star g_2(i,u) + \sum_{j=1}^{n} h^\star(j) p_{ij}(u) \right]. \tag{26}$$

Furthermore, among all the vectors h satisfying (26), there is a unique one with $h(n) = 0$. Conversely, assume that a vector $h^\star$ and a scalar $\lambda^\star$ satisfy the Bellman equation (26), along with $h(n) = 0$ and denote by $\mu^\star$ the policy attaining the minimum in (26). If additionally $\lambda_2(\mu^\star) > 0$, then $\lambda^\star$ is the optimal value and $\mu^\star$ is the optimal policy.

Proof: (i) Assume that $\lambda^\star$, $\mu^\star$ is the optimal value-optimal policy pair for (11). Then $\lambda^\star = \lambda_1(\mu^\star)/\lambda_2(\mu^\star)$. Equation (16) and the fact that $\lambda_2(\mu^\star) > 0$ imply that $\min_{\mu}[\lambda_1(\mu) - \lambda^\star\lambda_2(\mu)] \geq 0$. Observe that substituting $\mu^\star$ in the place of μ in the left-hand side of (16) we get 0. Thus, (24) holds true.

(ii) Consider the feasible region, illustrated in Figure 1.a. Due to the fact that the feasible region D is a two-dimensional compact convex set not containing the origin, there are at most two lines passing through the origin which support the feasible region (particularly the lines in Figure 2, (a) and (c)).

(iii) Observe that (17) and (25) are equivalent.

(iv) The direct part follows from (i). To prove the converse, consider a scalar $\lambda^\star$ and a vector $h^\star$ satisfying (26). Consider the problem of minimizing the LTA cost criterion:

$$\lambda_1(\mu) - \lambda^\star \cdot \lambda_2(\mu) = \\ = E\left[\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \left(g_1(x_k, \mu(x_k)) - \lambda^\star g_2(x_k, \mu(x_k)) \right) \right].$$

For this criterion there is a unique pair of a scalar θ and a vector $h(1), \ldots, h(n)$ with $h(n) = 0$ such that:

$$\theta + h(i) = \min_u \left[g_1(i,u) - \lambda^\star g_2(i,u) + \sum_{j=1}^{n} h(j) p_{ij}(u) \right],$$

for $i = 1, \ldots, n$. Uniqueness and (26) imply that $\theta = 0$ and thus (24) holds true. Inequality $\lambda_2(\mu^\star) > 0$ and (ii) imply that $\lambda^\star$ is the optimal value. The characterization of the optimal policy follows from (iii). □

***Corollary** 1:* Suppose that Assumption 1.a and Assumption 2.b hold. Then (i)-(iv) of Proposition 4 hold true.

B. A Relative Value Iteration Algorithm

In this section, we propose a Relative Value Iteration (RVI) algorithm for the problem, based on the single Bellman equation characterization (26) and prove its convergence. The iteration is given by:

$$\theta^{k+1} = \min_u \left[g_1(n,u) - \lambda^k g_2(n,u) + \tau \sum_{j=1}^{n} h^k(j) p_{nj}(u) \right], \tag{27}$$

$$h^{k+1}(i) = \min_u \left[g_1(i,u) - \lambda^k g_2(i,u) + \tau \sum_{j=1}^{n} h^k(j) p_{ij}(u) \right] - \\ - \theta^{k+1}, \tag{28}$$

$$\lambda^{k+1} = \lambda^k + \gamma\theta^{k+1}, \tag{29}$$

where the initial conditions are $\lambda^0 = 0$ and $h^0 = 0$, γ is a positive parameter (step size), $0 < \tau \leq 1$ is a damping factor, $i = 1, \ldots, n$, and n is the last element of the state space X. The idea of the algorithm is the following. Let us assume that λ is 'frozen' (is held fixed). Then (27)–(28) corresponds to a modified version of the relative value iteration scheme for the cost function:

$$E[\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} [g_1(x_k, u_k) - \lambda g_2(x_k, u_k)]]. \tag{30}$$

The parameter τ is used to avoid cycling (periodic behavior) in (27)–(28) and ensure convergence (for $\tau = 1$ we retrieve the usual relative value iteration scheme). Under (27)–(28), the value of θ^k converges to the minimum value of (30). On the other hand, it will be shown that the optimal value of (30) is strictly decreasing in λ, for λ in an appropriate interval. Thus, allowing λ to vary according to (29), if θ^k is positive, then λ will increase and will eventually lead to a lower value of θ. The following proposition shows the convergence of the algorithm.

***Theorem** 2 (Convergence of the Algorithm):* Suppose that the Assumptions 1.b and 2.a hold and that $\tau < 1$. Then, there is a value $\bar{\gamma}$ such that if $\gamma \leq \bar{\gamma}$, then the (h^k, λ^k) converges to $(h^\star/\tau, \lambda^\star)$ at the rate of a geometric progression.

Proof: See Appendix D. The proof is based on ideas of singular perturbations (e.g. [30]) □

***Remark** 7:* There are two basic differences of the RVI algorithm described in this section and the Reinforcement Learning algorithms for Semi-Markov Decision Processes (e.g. [21]– [24]). First this algorithm can handle denominators taking both positive and negative values. Second, since it does not use stochastic approximation it is able to achieve a faster (geometric) convergence rate. Despite the fact that this change of the convergence rate is intuitively trivial, the proof of convergence is quite different. With vanishing step-size the ODE method applies. On the other hand, here we develop a form of discrete time singular perturbation stability analysis.

***Remark** 8:* Each iteration of the RVI algorithm ((27)-(29)) takes slightly less time and about half of the memory compared to the FPRVI algorithm (Section IV-B). Furthermore, RVI has guaranteed convergence. On the other hand, FPRVI computes both λ_1 and λ_2 and does not need any tuning.

VI. Battery Energy Storage System Application

Consider a BESS connected to the grid[1]. The discrete time dynamics for the State of Energy (SoE) of the battery is given by:

$$x_{k+1}^1 = x_k^1 + u_k, \tag{31}$$

where $x_k^1 \in [0,1]$ is the SoE i.e., the fraction of the energy stored in the BESS over the maximum possible stored energy, $u_k \in [-u^M, u^M]$ is the charging or discharging power.

At the beginning of each time period, before the decision u_k is made, the grid operator sends a vector signal $x_k^2 = [l_k\ p_k\ d_k]^T$ to the BESS. In this vector, l_k represents the power which the system requires from the battery to absorb or release (hereafter requested power), p_k the price at which the battery owner will be paid for the energy exchange, and d_k the penalty for deviating. The cost (which corresponds to the minus profit) for time period k is given by:

$$g_1(x_k^2, u_k) = d_k|u_k - l_k| - p_k u_k.$$

We assume that p_k has the same sign with l_k and that $d_k > 0$. Thus the battery owner always has the motivation to use a u_k such that $u_k l_k \geq 0$ i.e., to follow the sign of the grid operator signal. The cost $g_1(x_k^2, u_k)$ corresponds to the deviation penalty minus the revenues the BESS owner receives from the energy exchange. The vectors x_k^2 are modeled as i.i.d random variables (that is x_k^2 and $x_{k'}^2$ are independent for $k \neq k'$ and have the same distribution). For the consistency of the notation we write the evolution of the state variable x^2 as $x_{k+1}^2 = w_k$. The system at time step k has state variable $x_k = [x_k^1\ (x_k^2)^T]^T$ and the dynamics is written in compact form as $x_{k+1} = f(x_k, u_k, w_k)$.

In this example, we use a very simple Ah throughput model (e.g. [34]) to predict the battery lifetime. Particularly, we assume that after time interval k, the remaining useful life of the BESS is reduced by an amount of $g_2(x_k, u_k) = c_1 + c_2|u_k|$. The first term corresponds to calendar aging and is independent of the battery use, while the second term corresponds to cycling aging. The lifetime of the BESS is thus given by:

$$T = \min_t \left\{ t : \sum_{k=1}^{t} g_2(x_k, u_k) \geq R \right\},$$

where R is a large positive constant denoting the initial remaining useful life of the battery. The problem of maximizing the revenues of the BESS over its lifetime can be stated in a form similar to (8):

$$\underset{\mu}{\text{minimize}} \quad J_a = \lim_{N\to\infty} E\left[\sum_{k=1}^{T\wedge N-1} g_1(x_k, u_k) \right] / R, \tag{32}$$

where $\mu = (\mu_1, \dots)$ and $u_k = \mu_k(x_k)$.

Let us observe that, if the SoE of the battery permits, the battery manager can always choose an action u_k making the instant cost g_1 non-positive. Thus, it is in the best interest of the battery manager to try to make the lifetime T long, while continuing profiting from its use. Since (31) is controllable and w_k is i.i.d., this model satisfies Assumption 1.a. Furthermore, since $g_2(u) \geq 0$ for all u, the model satisfies also assumption 2.a.

***Remark** 9:* More complex degradation models, depending non-linearly on the power, on the state of charge of the battery, or its temperature can be considered. Furthermore, the requested energy, and the price may have some more complicated stochastic dynamics, depending probably on the time of the day. In this section, however, we use the simplest possible models to illustrate the application of the methods developed in the previous sections.

We then apply both FPRVI and RVI algorithms to this problem, compare their performance, compute the optimal policy and then study its sensitivity to the aging parameters.

A. Optimal Control Computation

Let us first specify some parameters used in the numerical computations. The SoE x_k^1 is bounded in $[0,1]$, the maximum absolute power u^M is 0.1, the calendar aging constant c_1 is 0.01 and the cycling aging constant c_2 is 1 (the cycling aging is dominant). For simplicity, we assume that the requested power l_k is distributed uniformly in $[-0.1, 0.1]$ and that the price per unit energy p_k and the deviation penalty are given by $p_k = 100 l_k$ and $d_k = 1.2 p_k + 0.01$ respectively. This choice reduces the dimensionality of the state variable x^2 to one. We choose to discretize the SoE part of the state variable into 101 points and the x^2 part into 21 points. Thus, the state space has 2121 points.

We implemented both algorithms in Julia 0.6.2. Figure 3 compares their speed of convergence. Particularly, the vertical axis of the figure corresponds to the logarithm of the relative difference of the consecutive iterates of the algorithms. For RVI this corresponds to the quantity:

$$\log_{10} \frac{\|h^{k+1} - h^k\|}{\|h^k\|},$$

and for FPRVI to:

$$\log_{10} \frac{\|[h_1^{k+1}\ h_2^{k+1}] - [h_1^k\ h_2^k]\|}{\|[h_1^k\ h_2^k]\|}.$$

Both algorithms converge linearly (geometrically). The runs were performed in an Intel Core i5 CPU 750 @2.67GHz, 4GB RAM desktop PC. The run time[2] for 200 iterations of RVI algorithm is 1.8s, while 200 iterations of the FPRVI algorithm (Section IV-B) take 1.9s. The parameters for the RVI algorithm are $\tau = 1$ (there is no need to use $\tau < 1$, because there is no periodicity) and $\gamma = 2$ (experimentally tuned).

Figure 4 illustrates the optimal control law, which is piecewise linear with respect to the SoE. It is interesting that at some combinations of SoE with requested power the optimal control is to absorb exactly the power requested i.e., to use $u_k = l_k$. In other combinations it is optimal to use zero control, in anticipation of possible future losses and battery aging. For example, for $l = 0.05$ and for a SoE 80% the

[1]Usual applications are frequency regulation, peak shaving, energy arbitrage, microgrid operation in island mode, et.c. (e.g. [31], [32], [33]) .

[2]In the implementation of both algorithms, we used a monotonicity property of the optimal control law, which is specific to this problem. This reduces the running time, but all the other results are not affected.

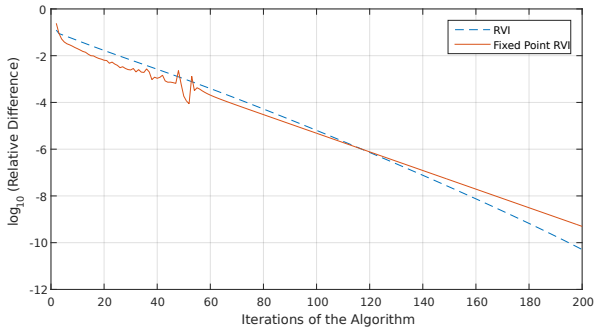


Fig. 3: The convergence of the suggested methods. The vertical axis represents the logarithm of the relative difference of two consecutive iterations. The solid line corresponds to the FPRVI algorithm (Section IV) and the dashed line to RVI algorithm. Numerical results suggest that both algorithms have a linear (geometric) convergence rate.

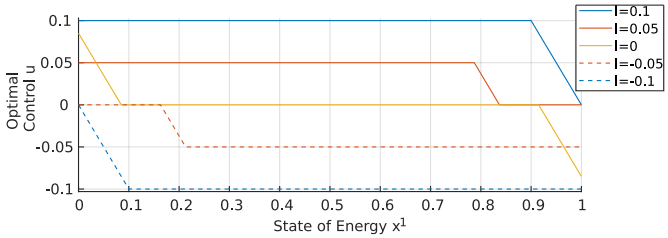


Fig. 4: The optimal control law for the battery Problem (32). The horizontal axis is the State of Energy x^1 and the different lines correspond to different values of the requested power l. We may observe that there is a symmetry around the point $(0.5, 0)$. Particularly, for $l' = -l$ and $x^{1'} = 1 - x^1$ we have $\mu^\star(x^{1'}, l') = -\mu^\star(x^1, l)$.

optimal action is to absorb no power. If the battery absorbed some energy then it would increase the probability that, at a subsequent time step, when the deviation penalty would probably be higher, it would not have the ability to follow the signal l, while at the same time it would certainly (with probability 1) contribute to the battery aging at the current time step.

We then show experimentally that, for a reasonable value of R, the horizon is approximately deterministic. Assume that $R = 6000$, which corresponds to a battery capable of 3000 full cycles (a reasonable number of cycles for a stationary lithium ion battery storage system). For the optimal policy $\mu^\star$, the expected time horizon is $R/\lambda_2(\mu^\star) = 103294$ time steps. On the other hand, running 100 Monte-Carlo simulations using the optimal policy and uniform initial conditions, we get a mean time horizon of 103298 steps and maximum absolute deviation of 554 steps or 0.53%. The root mean square deviation is 192 steps or 0.18%. These results agree with the intuition of the approximately deterministic time horizon.

B. Quantifying Approximate Optimality

We then compare the value of the cost J_a under the approximately optimal policy, $J_a(\mu^\star, x^1, x^2)$, computed in the previous subsection, to the minimum value of J_a.

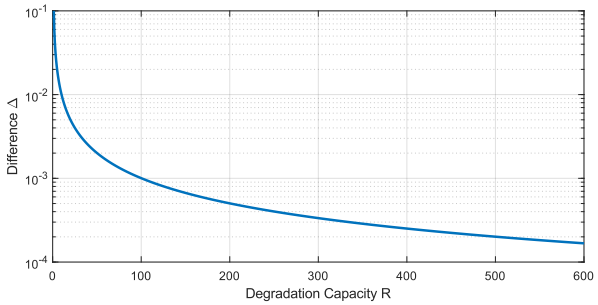


Fig. 5: The difference between the cost of the approximately optimal policy $\mu^\star$ and the optimal policy $\bar{\mu}$, computed using the SSP reformulation.

Using Remark 2, we first transform the Problem (32) into a Stochastic Shortest Path (SSP) problem. Assume that all the parameters are as in the previous subsection. Consider an auxiliary state variable $y_k = \sum_{t=0}^{k} g_2(u_k)$. Since all the possible values of $g_2(u_k)$ are integer multiples of 0.01, the state space for y_k is $Y = \{0, 0.01, 0.02 \ldots, R\}$.

To simplify the numerical computations, we downscale the problem to $R = 600$. Thus, Y has 60001 elements and the extended state space $\bar{X} = X \times Y$ has $127,262,121$ elements. The Bellman equation for this problem is:

$$J^\star_{\text{SSP}}(y, x^1, x^2) = \min_u \Big\{ \frac{g_1(x,u)}{R} I_{y+g_2(u)<R} + \\ + \frac{1}{21} \sum_{x'^2=1}^{21} J^\star_{\text{SSP}}(y + g_2(u), x^1 + u_k, x'^2) \Big\}, \quad (33)$$

where $I_\cdot$ is the indicator (characteristic) function. Furthermore, $J^\star_{\text{SSP}}(y, x^1, x^2) = 0$, for $y \geq R$. The Bellman equation has a special structure. Particularly, since always $g_2(u) > 0$, the value of $J^\star_{\text{SSP}}(y, x^1, x^2)$ depends only on values of $J^\star_{\text{SSP}}(y', x'^1, x'^2)$ with $y' > y$. This allows us to solve (33) from $y = R$ down to $y = 0$. Let us note that a similar property was used in [13].

Denote by $\bar{\mu}$ the optimal policy computed for the SSP problem. The optimal value $J^\star_a(x^1, x^2)$ of J_a is equal to $J^\star_{\text{SSP}}(0, x^1, x^2)$. Figure 5 illustrates the difference:

$$\Delta = \max_{x^1, x^2} \left\{ J_a(\mu^\star, x^1, x^2) - J_a(\bar{\mu}, x^1, x^2) \right\},$$

for several values of R.

For $R = 600$ the value of the difference Δ is 1.673×10^{-4}. The relative difference $\Delta / \min_{x^1, x^2} |J_a(\bar{\mu}, x^1, x^2)|$ is equal to 2.99×10^{-5}. Therefore, if instead of the approximately optimal policy $\mu^\star$ we used the actual optimal policy $\bar{\mu}$ the increase in our profits would be approximately 0.003%.

Now regarding the computation cost, each iteration of (33), requires almost the same number of operations as the algorithm (27)–(29). For (27)–(29) we need in this example 200 iterations, whereas for (33) we need 60000 iterations (for $R = 6000$, as in the previous subsection, we would need 600,000 iterations). Thus, algorithm (27)–(29) needs almost 300 times less operations than (27)–(29) (3000 times, for $R = 6000$).

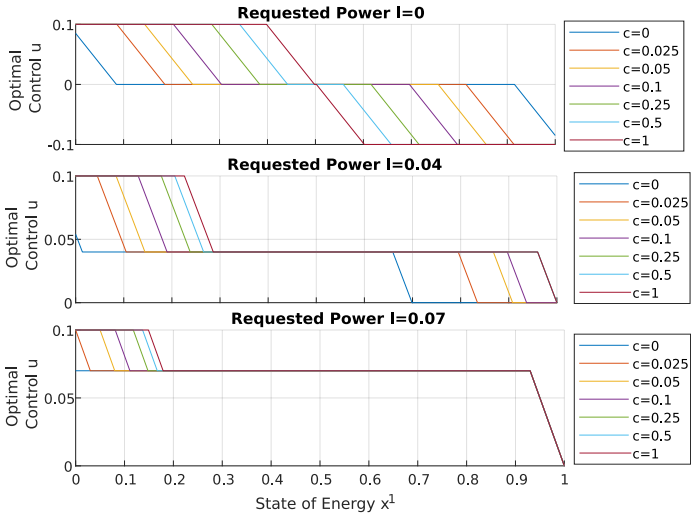


Fig. 6: The optimal control law for different values of c. The value of the optimal controls, for negative l can be deduced from the symmetry observation in Figure 4.

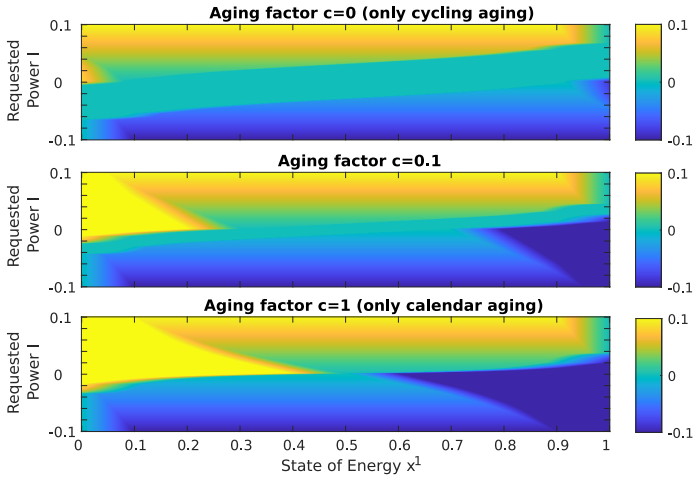


Fig. 7: The optimal control law for different values of c. Particularly, the first heat map corresponds to the case where cycling aging is dominant, the third plot to the case where the calendar aging is dominant (equivalently we do not care about aging) and the second to a case in between.

C. Sensitivity to the Aging Parameters

Another interesting question is how the optimal control law changes when the proportion of the calendar aging over the cycling aging varies. We run the same optimization problem with parameters $c_1 = c$ and $c_2 = 1 - c$, for various values of c. The results are illustrated in Figure 6. For $c = 1$, the optimal control law is identical to the one we would obtain by ignoring the aging and optimizing the long-term average cost λ_1. Indeed, calendar aging is independent of the control actions. As c becomes smaller the cycling aging starts to dominate. We may observe from the numerical results that for smaller c the optimal control action has always smaller absolute value.

Figure 7 illustrates the optimal control law in the form of a heat map, for three different values of c. For a small (in absolute value) requested power and for an almost full or an almost empty battery, we observe an 'opportunistic' charging or discharging. This behavior becomes more important, as the weight of the calendar aging becomes larger (or equivalently as we care less about aging). Furthermore, for small c, we observe that there is a region where the BESS uses $u = 0$ i.e., does not charge or discharge, despite the existence of a non-zero signal l. This turquoise ($u = 0$) region shrinks and eventually disappears as the calendar aging becomes dominant. Another region is the upper right and lower left part of the heat maps, where the BESS is not able to follow the signal l and thus we have saturation. Finally, there is a linear region where the BESS follows exactly the requested power signal.

For this sensitivity analysis we used a finer discretization with 201201 points to produce smoothly varying heat maps. Particularly we discretized x^1 variable in 1001 points and x^2 in 201 points.

D. Comparing of the Proposed Solution to MPC

We then compare the control law computed in the previous subsections to a Stochastic Model Predictive Control (SMPC). The control action u_k at time step k solves the problem:

$$\underset{\mu}{\text{minimize}} \left[\sum_{k'=k}^{k+N_0} g_1(x_k, u_k) \middle| u_{k'} = \mu_{k'}(x_{k'}) \right], \qquad (34)$$

where the optimization horizon N_0 may take several possible values. The parameters are as in subsection VI-A.

It is interesting that, if $N_0 = 1$, the control law becomes:

$$u_k = \begin{cases} l_k & \text{if } \ 0 \le x_k + l_k \le 1 \\ -x_k & \text{if } \ x_k + l_k < 0 \\ 1 - x_k & \text{if } \ x_k + l_k > 1 \end{cases}, \qquad (35)$$

that is, u_k follows the requested signal l_k as close as possible, without violating the battery constraints. We call this the 'myopic' control law.

Figure 8 illustrates the cost J_a for the SMPC control laws for several values of N_0. Interestingly, the cost J_a is not decreasing as the horizon increases. This phenomenon occurs because (34) does not take into account the effect of the control law to the problem horizon.

The use of the approximately optimal control law increases the profits by 6.21% when compared with the 'myopic' control law (35), by 4.02% when compared to the SMPC with $N_0 = 3$, which happens to result the optimal cost among SMPC controllers, and by 11.78% when compared with the SMPC with long horizon N_0.

VII. Examples with Indefinite Denominator

1) Maximum Range of an Electric Vehicle: Consider an electric vehicle with regenerative breaking and let us study the problem of designing a controller to achieve the longest possible distance, before recharging the battery. In contrast with the BESS example, here the system manager (driver) can choose actions that extend the remaining life of the system, for example he/she can use regenerative breaking in a downhill

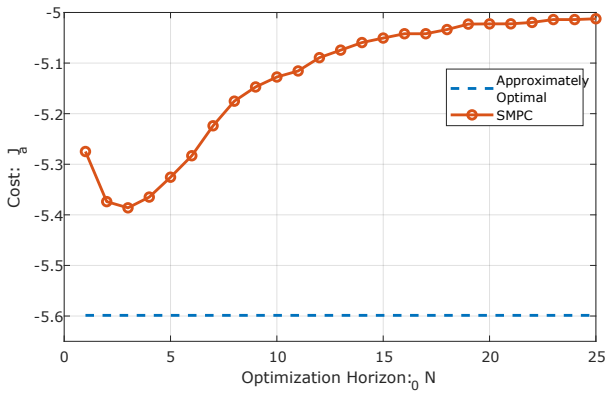


Fig. 8: The cost of SMPC law for several values of the optimization horizon N_0, compared to the cost of the approximately optimal control law obtained in Subsection VI-A.

road. The velocity of the vehicle x_k^1 evolves according to Newton's second law: $x_{k+1}^1 = x_k^1 + u_k - F_{\text{Slope}}(x_k^2) - F_{\text{Friction}}(x_k^1)$, where we assume that the vehicle has unit mass, $F_{\text{Slope}}(x_k^2)$ denotes the force to the vehicle due to the road inclination, x_k^2 is the slope of the road, $F_{\text{Friction}}(x_k^1)$ is the force due to friction and u_k is the force (from the road to the vehicle) due to the use of the accelerator or the break. We model the slope of the road x_k^2 as a sequence of dependent random variables evolving according to: $x_{k+1}^2 = (1-\alpha x_k^1)x_k^2 + \alpha x_k^1 w_k$, where w_k are i.i.d. random variables (not necessarily zero-mean) and α is a small positive constant. Let us denote the energy consumption per time step by $g_2(x_k, u_k)$.

If the mean slope of the road is non-negative, then we expect that the vehicle will consume eventually all the energy stored in its battery. The time, at which this will happen, is given by: $T = \inf_t \left\{ t : \sum_{k=1}^t g_2(x_k, u_k) \geq R \right\}$, where R is the amount of energy initially stored in the battery of the vehicle. The maximum range problem corresponds to the minimization of the following quantity: $\underset{u_k = \mu_k(x_k)}{\text{minimize}} -E\left[\sum_{k=1}^T x_k^1\right]$.

2) Selling the Stock of a Non-Viable Retail Firm: Consider a small retail business, which does not have any profits for a long time, and decides to exit the market. This example studies the problem of optimizing the revenue from selling the remaining stock R of products. Assume that if the business sets a price $u_k^1 \in [0,1]$ at time k, then the number of products it sells per time step is given by $d(u_k^1)$. Furthermore, assume that at time step k, the business is able to buy products at a price x_k which is random. However, as far as the firm is working, it has to pay a constant amount of money c per unit time (e.g. rent). Denoting by $u_k^2 \in [0,1]$ the number of products the business buys at time step k, the cost per stage at time step k is given by $g_1(x_k, u_k^1, u_k^2) = x_k u_k^2 - d(u_k^1)u_k^1 + c$ and the stock after step k is reduced by $g_2(x_k, u_k^1, u_k^2) = d(u_k^1) - u_k^2$. Therefore, the problem of optimizing firm's revenues, until the stock is over has the same structure of the previous problems.

***Remark** 10:* In this problem, we have assumed that the business is not viable. However, it is probably profitable to buy new products, if the price is low. There are policies which make the remaining stock increase in the long run. However, these policies produce long-term losses.

VIII. Conclusion

The paper considered the problem of the optimal control of MDPs over a long and operation-dependent time horizon and reduced it approximately to a problem involving the minimization of the ratio of two long-time average-costs. The optimal control law was characterized by appropriate sets of Bellman-type equations. Particularly, two characterizations were given. In the first one (fixed point characterization), a control law is optimal if and only if there exist two vectors satisfying a pair of coupled Bellman equations. In the second characterization, a necessary condition is given in terms of a single Bellman equation. This condition also becomes sufficient if an additional inequality holds true. Based on each characterization, we proposed a value iteration-based algorithm (i.e. FPRVI and RVI algorithms). For the RVI algorithm we also proved its convergence.

The proposed techniques were then applied to the problem of the optimal management of a Battery Energy Storage System. The results show that the optimal control law consists of different behaviors, whose boundaries move according to the aging parameters. In the future it would be interesting to analyze more complex models for battery degradation or more accurate stochastic models to describe the grid operator signal.

Appendix

A. Proof of Proposition 1

Consider the set of points in $\mathbb{R}^2$:

$$D_0 = \{(\lambda_1(\mu, i), \lambda_2(\mu, i)) : \mu \in \mathcal{U}_d, i \in X\},$$

and denote by D its convex hull. Since the set of deterministic policies $\mathcal{U}_d$ is finite, D_0 is finite, and D is a compact and convex.

(i) We use contradiction. Assume that there is a behavioral policy μ, such that for some i, it holds $(\lambda_1(\mu, i), \lambda_2(\mu, i)) \notin D$. Then, there is a line strictly separating D and $(\lambda_1(\mu, i), \lambda_2(\mu, i))$. Thus, there are constants c_1, c_2 such that:

$$c_1\lambda_1(\mu, i) + c_2\lambda_2(\mu, i) < c_1\tilde{\lambda}_1 + c_2\tilde{\lambda}_2, \tag{36}$$

for all $(\tilde{\lambda}_1, \tilde{\lambda}_2) \in D$.

Consider the cost function:

$$\lim_{N\to\infty} \frac{1}{N} E\left[\sum_{k=1}^N (c_1 g_1(x_k, u_k) + c_2 g_2(x_k, u_k)) \middle| x_0 = i\right],$$

and denote by $\tilde{\mu} \in \mathcal{U}_d$, its optimal policy. Then:

$$c_1\lambda_1(\tilde{\mu}, i) + c_2\lambda_2(\tilde{\mu}, i) \leq c_1\lambda_1(\mu, i) + c_2\lambda_2(\mu, i). \tag{37}$$

On the other hand, $\tilde{\mu}$ is deterministic and thus $(\lambda_1(\tilde{\mu}, i), \lambda_2(\tilde{\mu}, i)) \in D_0 \subset D$. Hence, (37) contradicts (36).

(ii) Denote by $D_v = \{A_1, \ldots, A_m\}$ the extremal points (vertices) of D (see figure 9). Of course $D_v \subset D_0$. Observe that D is the convex hull of D_v. Consider an $A_j \in D_v$. Let A_jB be a segment bisecting the angle $\widehat{A_{j-1}A_jA_{j+1}}$, l a line perpendicular to A_jB and θ the angle of A_jB from

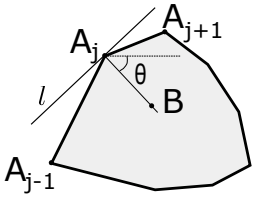


Fig. 9: The set D.

the horizontal axis. Since A_j is an extremal point, the angle $\widehat{A_{j-1}A_jA_{j+1}}$ is strictly less than $180°$. Thus, $A_j = (\bar{\lambda}_1, \bar{\lambda}_2)$ is the unique common point of line l and set D, and the unique minimizer of the problem:

$$\underset{(\tilde{\lambda}_1,\tilde{\lambda}_2)\in D}{\text{minimize}} \ \{\cos\theta\tilde{\lambda}_1 + \sin\theta\tilde{\lambda}_2\}.$$

Consider the cost function:

$$\lim_{N\to\infty}\frac{1}{N}E\left[\sum_{k=1}^{N}(\cos\theta g_1(x_k,u_k) + \sin\theta g_2(x_k,u_k))\right],$$

and denote by $\mu \in \mathcal{U}_d$, its optimal policy. Weak accessibility implies that the optimal value is the same for all the initial conditions and thus the point $(\lambda_1(\mu,i), \lambda_2(\mu,i))$ lies on line l for all $i \in X$. Thus, since $(\lambda_1(\mu,i), \lambda_2(\mu,i)) \in D$, and A_j is the unique common point of l and D, $\lambda_1(\mu,i) = \bar{\lambda}_1$ and $\lambda_2(\mu,i) = \bar{\lambda}_2$, for all $i \in X$. This completes the proof of (ii).

(iii) For every vertex A_j there is deterministic stationary policy μ^j such that if $A_j = (\bar{\lambda}_1^j, \bar{\lambda}_2^j)$, then $\lambda_1(\mu^j,i) = \bar{\lambda}_1^j$, $\lambda_2(\mu^j,i) = \bar{\lambda}_2^j$, for all $i \in X$.

Consider a point $(\bar{\lambda}_1, \bar{\lambda}_2) = A \in D$. Then, since D is the convex hull of D_v, there are scalars $\alpha_1, \dots, \alpha_m$ such that $\bar{\lambda}_1 = \alpha_1\bar{\lambda}_1^1 + \dots + \alpha_m\bar{\lambda}_1^m$ and $\bar{\lambda}_2 = \alpha_1\bar{\lambda}_2^1 + \dots + \alpha_m\bar{\lambda}_2^m$. Assuming that $m \geq 2$, let us describe a policy $\bar{\mu} = (\bar{\mu}_1, \bar{\mu}_2, \dots)$ such that the limits in definition of $\lambda_1(\bar{\mu},i)$, $\lambda_2(\bar{\mu},i)$ exist and $\lambda_1(\bar{\mu},i) = \bar{\lambda}_1$, $\lambda_2(\bar{\mu},i) = \bar{\lambda}_2$. Policy $\bar{\mu}$ is separated in different epochs. The ν-th epoch lasts for ν time steps. That is, for $\nu(\nu-1)/2 < k \leq \nu(\nu+1)/2$ we are in the $\nu-$th epoch. During the $\nu-$th epoch $\bar{\mu}_k = \mu^1$ for the first $\beta_\nu^1 = [\alpha_1\nu]$ time steps, then $\bar{\mu}_k = \mu^2$ for the next $\beta_\nu^2 = [\alpha_2\nu]$ time steps and so on. Here $[\cdot]$ denotes the integer part. Finally, $\bar{\mu}_k = \mu^m$, for the last $\beta_m = \nu - \beta_1 - \dots - \beta_{m-1}$ time steps. Denote by $I_{\nu,j}$ the set of time steps during the $\nu-$th period, where $\bar{\mu}_k = \mu^j$. That is, $I_{\nu,j} = \{\nu(\nu-1)/2 + \sum_{j'=1}^{j-1}\beta_\nu^{j'} + 1, \dots, \nu(\nu-1)/2 + \sum_{j'=1}^{j}\beta_\nu^{j'}\}$.

Let us derive some inequalities for the duration of $I_{\nu,j}$. For $j = 1, \dots, m-1$ it holds:

$$0 \leq \alpha_j\nu - \beta_\nu^j < 1.$$

Adding these inequalities and substituting $\sum_{j=1}^{m-1}\alpha_j = 1-\alpha_m$ and $\sum_{j=1}^{m-1}\beta_\nu^j = \nu - \beta_\nu^m$ we get:

$$0 \leq \beta_\nu^m - \nu\alpha_m < m-1.$$

Therefore, recalling that $m \geq 2$, for all $j = 1, \dots, m$ it holds:

$$|\beta_\nu^j - \nu\alpha_j| < m-1. \tag{38}$$

Consider the sequences:

$$S_{\nu,s} = E\left[\frac{1}{\nu}\sum_{k=\nu(\nu-1)/2+1}^{\nu(\nu+1)/2} g_s(x_k,u_k)\,\middle|\, u_k = \mu_k(x_k)\right],$$

for $s = 1, 2$. Then, $S_{\nu,s}$ can be written as:

$$S_{\nu,s} = \frac{1}{\nu}\sum_{j=1}^{m}\beta_\nu^j\bar{S}_{\nu,s}^j, \tag{39}$$

where:

$$\bar{S}_{\nu,s}^j = E\left[\frac{1}{\beta_\nu^j}\sum_{k\in I_{\nu,j}} g_s(x_k,u_k)\,\middle|\, u_k = \mu^j(x_k)\right].$$

We then prove the following claim:

Claim: The sequence $S_{\nu,s}$ converges to $\bar{\lambda}_s$ for $s = 1, 2$.

Combining this with (39) and (38), and using the fact that $\sum_{j=1}^{m}\beta_\nu^j = \nu$, we get:

$$\begin{aligned}|S_{\nu,s} - \bar{\lambda}_s| \leq& \sum_{j=1}^{m}|\frac{\beta_\nu^j}{\nu}\bar{S}_{\nu,s}^j - \alpha_j\bar{\lambda}_s^j| = \frac{1}{\nu}\sum_{j=1}^{m}|\beta_\nu^j\bar{S}_{\nu,s}^j - \nu\alpha_j\bar{\lambda}_s^j| \\ \leq& \frac{1}{\nu}\sum_{j=1}^{m}[\beta_\nu^j|\bar{S}_{\nu,s}^j - \bar{\lambda}_s^j| + |\bar{\lambda}_s^j||\beta_\nu^j - \nu\alpha_j|] \\ \leq& \max_j|\bar{S}_{\nu,s}^j - \bar{\lambda}_s^j| + \frac{m(m-1)}{\nu}\max_j|\bar{\lambda}_s^j| \end{aligned} \tag{40}$$

Inequality (38) implies that $\beta_\nu^j \geq \nu\alpha_j - m + 1$. Thus, $\bar{S}_{\nu,s}^j \to \bar{\lambda}_s^j$, for all $j = 1, \dots, m$, $s = 1, 2$ and all x_0. Hence, $S_{\nu,s} \to \bar{\lambda}_s$, which completes the proof of the claim.

Now, consider the sequence:

$$\Xi_{s,i,N} = \frac{1}{N}E\left[\sum_{k=1}^{N} g_s(x_k,u_k)\,\middle|\, x_0 = i\right]. \tag{41}$$

To prove (iii) it is sufficient to show that $\Xi_{s,i,N} \to \bar{\lambda}_s$, as $N \to \infty$, for all $i = 1, \dots, n$ and for $s = 1, 2$. First, observe that the subsequence $\Xi_{s,i,\bar{\nu}(\bar{\nu}+1)/2}$ converges to the desired limit, as $\bar{\nu} \to \infty$. Indeed since:

$$\Xi_{s,i,\bar{\nu}(\bar{\nu}+1)/2} = \frac{2}{\bar{\nu}(\bar{\nu}+1)}\sum_{\nu=1}^{\bar{\nu}}\nu S_{j,\nu}, \tag{42}$$

The claim implies that $\Xi_{s,i,\bar{\nu}(\bar{\nu}+1)/2} \to \bar{\lambda}_s$.

Let us then examine $\Xi_{s,i,N}$ for N between two successive evaluations of the subsequence. If $\bar{\nu}(\bar{\nu}+1)/2 \leq N < (\bar{\nu}+1)(\bar{\nu}+2)/2$ then:

$$\begin{aligned}\Xi_{s,i,N} - \Xi_{s,i,\bar{\nu}(\bar{\nu}+1)/2} =& \frac{1}{N}\frac{\bar{\nu}(\bar{\nu}+1)}{2}\Xi_{s,i,\bar{\nu}(\bar{\nu}+1)/2} + \\ &+ \frac{1}{N}E\left[\sum_{k=\frac{\bar{\nu}(\bar{\nu}+1)}{2}+1}^{N} g_s(x_k,u_k)\,\middle|\, x_0 = i\right] - \Xi_{s,i,\bar{\nu}(\bar{\nu}+1)/2}.\end{aligned}$$

Thus:

$$\begin{aligned}|\Xi_{s,i,N} - \Xi_{s,i,\bar{\nu}(\bar{\nu}+1)/2}| \leq& \left|\frac{\bar{\nu}(\bar{\nu}+1)}{2N} - 1\right|\Xi_{s,i,\bar{\nu}(\bar{\nu}+1)/2} + \\ &+ \frac{\bar{\nu}\max_{i',u} g_s(i',u)}{N}\end{aligned}$$

Therefore, the variation of $\Xi_{s,i,N}$ for N between $\bar{\nu}(\bar{\nu}+1)/2$ and $(\bar{\nu}+1)(\bar{\nu}+2)/2$ tends to zero as $\bar{\nu} \to \infty$.

B. Proof of Theorem 1

(i) Problem (11) is feasible, because for any $i \in X$ there is a policy $\bar{\mu}$ with $\lambda_2(\bar{\mu}, i) > 0$. Due to Proposition 1, for all i Problem (11) has the same value with:

$$\underset{(\bar{\lambda}_1, \bar{\lambda}_2) \in D}{\text{minimize}} \ \{\bar{\lambda}_1/\bar{\lambda}_2 : \bar{\lambda}_2 > 0\}. \tag{43}$$

Assumption 2 and the compactness of D imply that this problem has a finite value (exists and it is greater than $-\infty$). Furthermore, since D is a compact polygon and $\bar{\lambda}_1/\bar{\lambda}_2$ is quasi-linear function (e.g. [35]), $\bar{\lambda}_1/\bar{\lambda}_2$ is minimized on a vertex of D. Therefore, there is a deterministic optimal policy, according to Proposition 1.

(ii) Denote by $\lambda^\star$ the optimal value of (11). The proof proceeds in two discrete steps.

First, we show that J_a cannot be much smaller than $\lambda^\star$, for a large value of R (Lemma 3). Conversely, it is shown that a corner optimal policy for (10) has at most a slightly higher cost when applied to (8) (Lemma 4).

Throughout the proof we use repeatedly Doob's (sub-) martingale optional stopping theorem (e.g. [36]).

***Theorem** 3 (Doob's optional stopping theorem):* Let v_t be an $\mathcal{F}_t$ submartingale and T a stopping time such that $E[T] < \infty$. Assume further that there is a constant C such that $|v_{t+1} - v_t| < C$, for all t. Then:

$$E[v_T] \geq E[v_0].$$

The last relation holds as an equality if v_t is a martingale.

We use also the following lemma the proof of which is almost identical to that of the optional stopping theorem. For the shake of completeness we present its proof.

***Lemma** 1:* Let Z_k be a sequence of random variables and assume that there is a constant C such that $|Z_k| < C$. Assume that T is a random time with $E[T] < \infty$. Then:

$$\lim_{N \to \infty} E\left[\sum_{k=1}^{T \wedge N} Z_k\right] = E\left[\sum_{k=1}^{T} Z_k\right]. \tag{44}$$

Equation (44) asserts also the existence of the limit.

Proof: Since $E[T] < \infty$, the random variable $Y_N = \sum_{k=1}^{T \wedge N} Z_k$ converges almost surely to $Y = \sum_{k=1}^{T} Z_k$ as $N \to \infty$. Furthermore, for all N the random variable Y_N is bounded above by:

$$|Y_N| \leq C \sum_{k=1}^{\infty} I_{T \geq k} = CT,$$

where $I_\cdot$ is the indicator (characteristic) function. But, since $E[T] < \infty$, the dominated convergence theorem applies and it holds:

$$\lim_{N \to \infty} E[Y_N] = E[Y],$$

which is exactly (44). Additionally, dominated convergence theorem proves also the existence of the limit. □

For an n-vector h denote by let us introduce the span seminorm:

$$\text{sp}(h) = \max_i h(i) - \min_i h(i). \tag{45}$$

***Lemma** 2:* Suppose that assumptions 1 and 2 hold. Consider any behavioral policy $\mu = (\mu_0, \mu_1, \dots)$. If under μ:

$$\underline{\text{L}} = \liminf_{N \to \infty} E\left[\sum_{k=0}^{T \wedge N - 1} g_1(x_k, u_k)\right] < \infty,$$

then:

(i) The expected horizon is finite i.e., $E[T] < \infty$.

(ii) The limit in J_a exists and satisfies:

$$J_a(x_0, \mu) = \lim_{N \to \infty} E\left[\sum_{k=0}^{T \wedge N - 1} g_1(x_k, u_k)/R\right] \tag{46}$$

$$= E\left[\sum_{k=0}^{T-1} g_1(x_k, u_k)/R\right]. \tag{47}$$

Therefore, the limit in the definition of J_a exists always (either finite or infinite).

Proof: Consider closed third quadrant:

$$Q_3 = \{(x, y) \in \mathbb{R}^2 : x \leq 0, y \leq 0\},$$

and the feasible set D. The set D is compact and convex and Q_3 is closed and convex. According to Assumption 2 $D \cap Q_3 = \emptyset$. Thus, there is a strictly separating line between D and Q_3. This implies that there are constants $c_1, c_2, c_3 > 0$ such that:

$$c_1 \lambda_1(\bar{\mu}, i) + c_2 \lambda_2(\bar{\mu}, i) \geq c_3, \tag{48}$$

for all the stationary policies $\bar{\mu}$.

Consider the cost per stage $\bar{g}(i, u) = c_1 g_1(i, u) + c_2 g_2(i, u) - c_3$. Then, due to (48), the minimal LTA cost for $\bar{g}$ is non-negative. Thus, there is a pair $\bar{\lambda} \geq 0$, $\bar{h} \in \mathbb{R}^n$ such that:

$$\bar{\lambda} + \bar{h}(i) = \min_u \left[\bar{g}(i, u) + \sum_{j=1}^{n} p_{ij}(u) \bar{h}(j)\right]. \tag{49}$$

Denote by $\mathcal{F}_t$ the σ algebra generated by $(x_0, u_0, \dots, x_t, u_t)$, and consider the stochastic process:

$$v_t = \bar{h}(x_t) + \sum_{k=0}^{t-1} \bar{g}(x_k, u_k).$$

Observe that v_t is a $\mathcal{F}_t$-submartingale. Indeed:

$$E[v_{t+1} - v_t | \mathcal{F}_t] = -\bar{h}(x_t) + \bar{g}(x_t, u_t) + \sum_{j=1}^{n} p_{ij}(u_k) \bar{h}(j)$$

$$\geq \bar{\lambda} \geq 0,$$

where for the first inequality we used the Bellman equation (49). Let us consider the stopped process v_{T_t}, where $T_t = T \wedge t$. Since, $E[T_t] \leq t$, the optional stopping theorem implies that $E[v_{T_t}] \geq E[v_0]$, or equivalently:

$$E\left[c_1 \sum_{k=0}^{T_t - 1} g_1(x_k, u_k) + c_2 \sum_{k=0}^{T_t - 1} g_2(x_k, u_k) - c_3(T_t - 1)\right] \geq$$

$$\geq E[h(x_0) - h(x_{T_t})].$$

Furthermore:

$$E[h(x_0) - h(x_{T_t})] \geq -\text{sp}(\bar{h}),$$

where $\text{sp}(h)$ is defined in (45). Thus:

$$E\left[T_t\right] \leq \frac{c_1}{c_3} E\left[\sum_{k=0}^{T_t-1} g_1(x_k,u_k)\right] + \frac{c_2}{c_3} E\left[\sum_{k=0}^{T_t-1} g_2(x_k,u_k)\right] + \\ +1 + \text{sp}(\bar{h})/c_3.$$

Observe that $\sum_{k=0}^{T_t-1} g_2(x_k,u_k) \leq R$. Hence:

$$E\left[T_t\right] \leq \frac{c_1}{c_3} S_t + \Delta,$$

where $\Delta = \frac{c_2}{c_3} R + 1 + \text{sp}(\bar{h})/c_3$ and $S_t = E\left[\sum_{k=0}^{T_t-1} g_1(x_k,u_k)\right]$.

There is a subsequence S_{t_ν} which attains the limit inferior i.e. $S_{t_\nu} \to \underline{\text{L}} < \infty$. Thus, there is a ν_0 such that $S_{t_\nu} \leq \underline{\text{L}} + 1$, for all $\nu \geq \nu_0$. Observe that, since T_t is non-decreasing in t, it holds:

$$E\left[T_t\right] \leq \frac{c_1}{c_3}(\underline{\text{L}} + 1) + \Delta,$$

for all t. Thus, due to monotone convergence theorem we conclude that:

$$E\left[T\right] \leq \frac{c_1}{c_3}(\underline{\text{L}} + 1) + \Delta < \infty.$$

This completes the proof of (i) of the lemma. Then, (ii) is a consequence of Lemma 1. □

Lemma 3: For every $\varepsilon > 0$ there is an R_0 such that for any $R \geq R_0$ and any behavioral policy $\mu = (\mu_0, \mu_1, \dots)$ it holds:

$$\lambda^\star \leq J_a(x_0,\mu) + \varepsilon,$$

where $\lambda^\star$ is the optimal value of (11).

Proof: Throughout this proof we assume that u_k is chosen according to $\mu_k(x_k)$. In the case where $J_a(x_0,\mu) = \infty$, the lemma is trivially true. Thus, we assume that $J_a(x_0,\mu) < \infty$. According to Proposition 4 and Corollary 1[3], there is a vector $h^\star$ satisfying:

$$h^\star(i) = \min_u \left[g_1(i,u) - \lambda^\star g_2(i,u) + \sum_{j=1}^n p_{ij}(u) h^\star(j) \right]. \tag{50}$$

Consider the stochastic process:

$$v_t = h^\star(x_t) + \sum_{k=0}^{t-1} [g_1(x_k,u_k) - \lambda^\star g_2(x_k,u_k)].$$

Observe that v_t is an $\mathcal{F}_t$-submartingale. Indeed, due to (50) it holds:

$$E[v_{t+1} - v_t | \mathcal{F}_t] = -h^\star(x_t) + \\ + \left[g_1(x_t,u_t) - \lambda^\star g_2(x_t,u_t) + \sum_{j=1}^n p_{ij}(u_k) h^\star(j) \right] \geq 0.$$

Using optional stopping theorem we get $E[v_{T_t}] \geq E[v_0] = E[h(x_0)]$. Therefore:

$$E\left[h^\star(x_t) + \sum_{k=0}^{T_t-1} [g_1(x_k,u_k) - \lambda^\star g_2(x_k,u_k)] \right] \geq E[h(x_0)].$$

[3]Let us first note that the proof of Proposition 2 and Proposition 4 do not depend on the proof of (ii) of this proposition and thus this argument is not circular.

Thus:

$$E\left[\sum_{k=0}^{T_t-1} g_1(x_k,u_k)\right] \geq \lambda^\star E\left[\sum_{k=0}^{T_t-1} g_2(x_k,u_k)\right] - \text{sp}(h). \tag{51}$$

We examine the two cases of Assumption 2 separately.

Case 1: Assumption 2.a holds true.

In this case $\lambda^\star < 0$ and thus $\sum_{k=0}^{T_t-1} g_2(x_k,u_k) < R$ implies:

$$\lambda^\star E\left[\sum_{k=0}^{T_t-1} g_2(x_k,u_k)\right] > \lambda^\star R.$$

Combining this with (51) we get:

$$\lambda^\star \leq E\left[\sum_{k=1}^{T_t-1} g_1(x_k,u_k)/R\right] + \text{sp}(h)/R.$$

Taking the limit of the right-hand side we complete the proof of the lemma for Case 1.

Case 2: Assumption 2.b holds true.

Taking the limits in (51) we get:

$$J_a(\mu,x_0) \geq \lambda^\star \lim_{t\to\infty} E\left[\sum_{k=0}^{T_t-1} g_2(x_k,u_k)]\right] - \text{sp}(h). \tag{52}$$

Using Lemma 1 we get:

$$\lim_{t\to\infty} E\left[\sum_{k=0}^{T_t-1} g_2(x_k,u_k)]\right] = E\left[\sum_{k=0}^{T-1} g_2(x_k,u_k)]\right] \geq \\ \geq R - \max_{i,u} g_2(i,u).$$

Combining this inequality with (52) and observing that $\lambda^\star > 0$ we get:

$$\lambda^\star \leq J_a(x_0,\mu) + \frac{\max_{i,u} g_2(i,u)\lambda^\star}{R} + \frac{\text{sp}(h)}{R}$$

This completes the proof. □

We then prove the converse:

Lemma 4: Let $\mu^\star$, $\lambda^\star$ be a corner optimal policy and the optimal value for (11). Then, for any $\varepsilon > 0$, there is an R_0 such that:

$$J_a(\mu^\star,x_0) \leq \lambda^\star + \varepsilon,$$

for all x_0 and all $R \geq R_0$.

Proof: (i) Since $\mu^\star$ is a corner policy, there is a pair $h_2^\star \in \mathbb{R}^n$, $\lambda_2^\star > 0$ satisfying:

$$\lambda_2^\star + h_2^\star(i) = g_2(i,\mu^\star(i)) + \sum_{j=1}^n p_{ij}(\mu^\star(i)) h_2^\star(j).$$

Then, stochastic process:

$$v_t = h_2^\star(x_t) + \sum_{k=0}^{t-1} [g_2(x_k,u_k) - \lambda_2^\star],$$

is a martingale. Similarly to the previous lemma we have $E[v_{T_t}] = E[v_0] = E[h_2^\star(x_0)]$. Hence:

$$E\left[\sum_{k=0}^{T_t-1} g_2(x_k,u_k) - \lambda_2^\star(T_t - 1)\right] = E[h_2^\star(x_0)] - E[h_2^\star(x_t)].$$

Thus,

$$E[T_t] \leq \frac{1}{\lambda_2^\star}(R + \text{sp}(h_2^\star)) + 1,$$

and monotone convergence theorem shows that $E[T] < \infty$.

For the optimal policy optimal value pair $\mu^\star, \lambda^\star$, there is a vector $h^\star \in \mathbb{R}^n$ such that:

$$h^\star(i) = g_1(i, \mu^\star(i)) - \lambda^\star g_2(i, \mu^\star(i)) + \sum_{j=1}^{n} p_{ij}(\mu^\star(i)) h^\star(j).$$

Then, stochastic process:

$$v_t = h^\star(x_t) + \sum_{k=0}^{t-1} [g_1(x_k, u_k) - \lambda^\star g_2(x_k, u_k)],$$

is a martingale. Similarly to the previous lemma we have $E[v_{T_t}] = E[v_0] = E[h^\star(x_0)]$. Hence:

$$E\left[\sum_{k=0}^{T_t-1} [g_1(x_k, u_k) - \lambda^\star g_2(x_k, u_k)]\right] = \\ = E[h^\star(x_0)] - E[h^\star(x_t)].$$

Therefore:

$$E\left[\sum_{k=0}^{T_t-1} g_1(x_k, u_k)\right] \leq \lambda^\star E\left[\sum_{k=0}^{T_t-1} g_2(x_k, u_k)\right] + \text{sp}(h^\star). \tag{53}$$

Case 1: Assumption 2.b holds.
In this case, using that $\sum_{k=0}^{T_t-1} g_2(x_k, u_k) \leq R$, and that $\lambda^\star > 0$, (53) becomes:

$$E\left[\sum_{k=0}^{T_t-1} g_1(x_k, u_k)\right] \leq \lambda^\star R + \text{sp}(h^\star). \tag{54}$$

and dividing by R and taking the limit we arrive to the desired conclusion.
Case 2: Assumption 2.a holds.
Similarly with Case 2 of Lemma 3, we have:

$$\lim_{t\to\infty} E\left[\sum_{k=0}^{T_t-1} g_2(x_k, u_k)]\right] = E\left[\sum_{k=0}^{T-1} g_2(x_k, u_k)]\right] \geq \\ \geq R - \max_{i,u} g_2(i, u).$$

Therefore:

$$J_a(x_0, \mu^\star) = \lim_{t\to\infty} E\left[\sum_{k=0}^{T_t-1} g_1(x_k, u_k)/R\right] \leq \\ \leq \lambda^\star + \frac{\text{sp}(h^\star) - \lambda^\star \max_{i,u} g_2(i, u)}{R}.$$

This completes the proof. □

Proof of Theorem 1: Combining lemmas 3 and 4, for large R and any policy μ, we get:

$$J_a(\mu^\star, x_0) \leq \lambda^\star + \varepsilon \leq J_a(\mu, x_0) + 2\varepsilon.$$

Thus, $\mu^\star$ is ε-optimal for large R. □

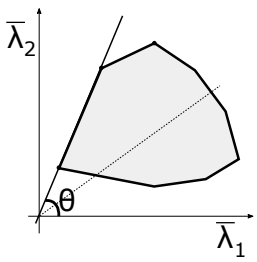


Fig. 10: The set D and the .

C. Proof of Proposition 3

We show that in the generic case any deterministic static optimal policy is a corner policy. We first show that for the most of the values of the costs g_1, g_2 there is a unique optimum in the Problem (43).

Lemma 5: Suppose that the controlled Markov chain is weakly accessible (Assumption 1.a holds). Then, the set of vectors of the form $[g_s(x, u) : x \in X, u \in U, s = 1, 2] \in \mathbb{R}^{2nn_U}$ such that Problem (43) has multiple minima has Lebesgue measure zero.

Proof: Observe that the value of $\bar{\lambda}_1/\bar{\lambda}_2$ is constant, in the $\bar{\lambda}_1 - \bar{\lambda}_2$ plane, on half-lines starting from the origin. Therefore, if (43) has multiple minima then they correspond to a line segment i.e., the intersection of the half-line with D. Furthermore, observe that $\bar{\lambda}_1/\bar{\lambda}_2 = \cot\theta$, where θ is the angle of the half line with the horizontal axis (see Figure 10). Thus, since $\cot(\cdot)$ is strictly decreasing in $(0, \pi)$, if (43) has multiple minima they correspond to an edge of D. Therefore, there are two vertices of D co-linear with the origin.

Using Proposition 1, if (43) has multiple minima then there is a pair of policies $\mu, \mu' \in \mathcal{U}_d$ and a pair of states $i, i' \in X$ such that:

$$\frac{\lambda_1(\mu, i)}{\lambda_2(\mu, i)} = \frac{\lambda_1(\mu', i')}{\lambda_2(\mu', i')}, \\ \lambda_1(\mu, i) \neq \lambda_1(\mu', i'), \quad \lambda_2(\mu, i), \lambda_2(\mu', i') > 0 \tag{55}$$

The set of states X and the set of stationary deterministic strategies $\mathcal{U}_d$ are finite. Thus, it is sufficient to show that for any pair of states i, i' and any pair of strategies $\mu, \mu' \in \mathcal{U}_d$, (55) does not hold in the generic case.

Denote by $\mathfrak{g}_s = [g_s(1,1), \ldots, g_s(1, n_U), \ldots, g_s(n,1), \ldots, g_s(n, n_U)]$, where $s = 1, 2$, and n, n_U the cardinalities of the state space and the set of actions. Then, using [20] (Theorem 8.1.1), for a state $\bar{i} \in X$ and a policy $\mu \in \mathcal{U}_d$ we have:

$$\lambda_s(\bar{\mu}, \bar{i}) = e_{\bar{i}}^T P_{\bar{\mu}}^\star S_{\bar{\mu}} \mathfrak{g}_s,$$

where $e_{\bar{i}}$ an n-vector with $\bar{i}$-th entry equal to one and all the other entries equal to zero, $P_{\bar{\mu}}^\star$ is the limiting matrix of the Markov chain under the policy $\bar{\mu}$ and $S_{\bar{\mu}}$ is an $n \times n|U|$ matrix such that $S_{\bar{\mu}} \mathfrak{g}_s = [g_s(1, \bar{\mu}(1)), \ldots, g_s(n, \bar{\mu}(1))]^T$. Since the limiting matrix $P_{\bar{\mu}}^\star$ is a stochastic matrix and $S_{\bar{\mu}}$ has all its entries equal to zero except of a single entry on each row which is equal to one, the $1 \times nn_U$ row vector:

$$d_{\bar{\mu}, \bar{i}} = e_{\bar{i}}^T P_{\bar{\mu}}^\star S_{\bar{\mu}},$$

has non-negative entries summing to one.

Assume a pair of policies μ, μ' and a pair of states i, i'. If (55) holds then:

$$\frac{d_{\mu,i}\mathfrak{g}_1}{d_{\mu',i'}\mathfrak{g}_1} = \frac{d_{\mu,i}\mathfrak{g}_2}{d_{\mu',i'}\mathfrak{g}_2} \neq 1. \tag{56}$$

If $d_{\mu,i} = d_{\mu',i'}$ then (55) does not holds. Thus, assume $d_{\mu,i} \neq d_{\mu',i'}$.

Denote by by $\mathfrak{g}_1^1$ the first entry of $\mathfrak{g}_1$ and by $\mathfrak{g}_1^{\text{Rest}}$ the rest of the entries. Similarly, denote by $d^1_{\mu,i}, d^1_{\mu',i'}$ the first entries $d_{\mu,i}, d_{\mu',i'}$ and $d^{\text{Rest}}_{\mu,i}, d^{\text{Rest}}_{\mu',i'}$ the rest of the entries. Observe that for all the values of $\mathfrak{g}_1^{\text{Rest}}$ and $\mathfrak{g}_2$ there is at most one value of $\mathfrak{g}_1^1$ such that (56) holds which, whenever is defined, is given by:

$$f(\mathfrak{g}_1^{\text{Rest}}, \mathfrak{g}_2) = \frac{(d_{\mu,i}\mathfrak{g}_2)d^{\text{Rest}}_{\mu',i'}\mathfrak{g}_1^{\text{Rest}} - (d_{\mu',i'}\mathfrak{g}_2)d^{\text{Rest}}_{\mu,i}\mathfrak{g}_1^{\text{Rest}}}{(d_{\mu',i'}\mathfrak{g}_2)d^1_{\mu,i} - (d_{\mu,i}\mathfrak{g}_2)d^1_{\mu',i'}}.$$

Thus, the set of points in $\mathbb{R}^{2nn_U}$, such that (55) holds is a subset of the graph of f. Therefore, it has Lebesgue measure zero ([37], exercise 3.10.52). □

To prove the proposition, assume that $\mu^\star$ is such a policy. Proposition A implies that $\lambda_1(\mu^\star, i)/\lambda_2(\mu^\star, i)$ has the same (optimal) value for all i. But, since, $\mu^\star$ is not a corner policy, there is a pair $i, i' \in X$ such that $\lambda_1(\mu, i) \neq \lambda_1(\mu, i')$. This corresponds to a case where Problem (43) has multiple minima. □

D. Proof of Theorem 2

Assumption 1.a and the compactness of the feasible region (see Figure 1) imply that there are positive constants $\varepsilon_{\lambda_1}, \varepsilon_{\lambda_2}$ such that for each policy μ either $\lambda_1(\mu) \geq \varepsilon_{\lambda_1}$ or $\lambda_2(\mu) \geq \varepsilon_{\lambda_2}$. Denote by Q the point $(\varepsilon_{\lambda_1}, \varepsilon_{\lambda_2})$. We state first the following lemma.

Lemma 6: There is a positive scalar δ_λ such that if $\lambda > \lambda^\star - \delta_\lambda$ then the solution μ^λ of (30) satisfies $\lambda_2(\mu^\lambda) \geq \varepsilon_{\lambda_2}$ and $\lambda_1(\mu^\lambda) \leq 0$.

Proof : Consider the optimal solution $\mu^\star$, the costs $\lambda_1^\star, \lambda_2^\star$, and the associated value $\lambda^\star = \lambda_1^\star/\lambda_2^\star$. Denote by F the point $(\lambda_1^\star, \lambda_2^\star)$. Draw from the origin the line OP which is parallel to FQ. It is easy to see that for any $\lambda < 0$ such that the line $x - \lambda y$ is above OP in the second quadrant, the optimal solution μ^λ of (30) satisfies $\lambda_2(\mu^\lambda) \geq \varepsilon_{\lambda_2}$ and $\lambda_1(\mu^\lambda) \leq 0$. The angle $\widehat{FOP}$ is equal to the angle $\widehat{OFQ}$. There is a minimum value of the angle $\widehat{OFQ}$ for the various points F in the intersection of the rectangle with the second quadrant. Furthermore, this minimum value is positive and depends only on ε_{λ_2}, $\max|g_1|$ and $\max|g_2|$. Thus, using the fact that the function $\arctan(\cdot)$ has a derivative less than or equal to one and some simple trigonometric manipulations, we conclude that there is a positive constant δ_λ such that if $\lambda > \lambda^\star - \delta_\lambda$ then the solution μ^λ of (30) satisfies $\lambda_2(\mu^\lambda) \geq \varepsilon_{\lambda_2}$ and $\lambda_1(\mu^\lambda) \leq 0$. □

Let us introduce some notation. The norm symbol $\|\cdot\|$ denotes the infinity norm. For a scalar λ, denote by $\mathcal{T}_\lambda$ the Bellman operator:

$$(\mathcal{T}_\lambda h)(i) = \min_u \left[g_1(i,u) - \lambda g_2(i,u) + \tau \sum_{j=1}^n h^k(j) p_{ij}(u) \right],$$

and by F_λ the corresponding relative value iteration operator:

$$F_\lambda h = \mathcal{T}_\lambda h - (\mathcal{T}_\lambda h)(n) \cdot \mathbf{1}.$$

Let h^λ be the unique fixed point of F_λ. It is not difficult to see that $(\lambda, h^\lambda/\tau)$ satisfy the Bellman equation for (30). Denote also by $\bar{G}$ the maximum value of $|g_2(i,u)|$.

For a pair of scalars λ and λ' and a pair of n-vectors h and h', operator T_λ satisfies the following pair of inequalities:

$$\|T_\lambda h - T_{\lambda'} h\| \leq \bar{G}|\lambda - \lambda'|, \quad \|T_\lambda h - T_\lambda h'\| \leq \|h - h'\|. \tag{57}$$

Similarly:

$$\|F_\lambda h - F_{\lambda'} h\| \leq 2\bar{G}|\lambda - \lambda'|, \quad \|F_\lambda h - F_\lambda h'\| \leq 2\|h - h'\|. \tag{58}$$

Inequalities (57), (58) will be repeatedly used throughout the proof.

Denote by $\phi^k_{\lambda,h}$ the vector defined by:

$$\begin{aligned} \phi^{k+1}_{\lambda,h} &= F_\lambda \phi^k_{\lambda,h}, \\ \phi^0_{\lambda,h} &= h. \end{aligned} \tag{59}$$

Let $q^k_{\lambda,h}$ be the difference of two successive iterates of $\phi_{\lambda,h}$ i.e., $q^k_{\lambda,h} = \phi^{k+1}_{\lambda,h} - \phi^k_{\lambda,h}$. Finally, denote by $\text{sp}(\cdot)$ the span semi-norm of an n-vector defined as $\text{sp}(q) = \max_i q(i) - \min_i q(i)$. As an intermediate step for proving the convergence of the usual value iteration method it has been shown that there is a positive integer m and an $\varepsilon > 0$ such that:

$$\text{sp}(q^k_{\lambda,h}) \leq (1 - \varepsilon)\text{sp}(q^{k-m}_{\lambda,h}), \tag{60}$$

for all $k \geq m$ (see for example [19], [20]). More importantly, m and ε depend only on the properties of the controlled Markov chain and not on the cost function. Thus, (60) is satisfied for all λ, for the same values of m and ε. Using (60), we can bound the evolution of $\text{sp}(q^k_{\lambda,h})$, in the form:

$$\text{sp}(q^k_{\lambda,h}) \leq B(h,\lambda)\xi^k,$$

where $\xi = (1 - \varepsilon)^{1/m}$ and:

$$B(h,\lambda) \geq \max_{k=0,\ldots,m} \text{sp}(q^k_{\lambda,h})/(1 - \varepsilon). \tag{61}$$

Based on this bound, a convergence rate (59) can be derived. It holds $q^k_{\lambda,h}(n) = 0$ and thus $\|q^k_{\lambda,h}\| \leq \text{sp}(q^k_{\lambda,h})$. Therefore,

$$\|\phi^k_{\lambda,h} - h^\lambda\| \leq \sum_{t=k}^\infty \|q^t_{\lambda,h}\| \leq \frac{B(h,\lambda)\xi^k}{1 - \xi}. \tag{62}$$

We then derive a formula for $B(h,\lambda)$ which does not depend on λ. To this end, observe that for a vector h' it holds:

$$\|F_\lambda h' - h'\| \leq \|F_\lambda h' - F_\lambda h^\lambda\| + \|F_\lambda h^\lambda - h'\|.$$

Applying the second inequality of (58) to the first term of the right-hand side and recalling that $F_\lambda h^\lambda = h^\lambda$ we get:

$$\|F_\lambda h' - h'\| \leq 3\|h' - h^\lambda\|.$$

Thus:

$$\|\phi^{k+1}_{\lambda,h} - h^\lambda\| \leq \|\phi^{k+1}_{\lambda,h} - \phi^k_{\lambda,h}\| + \|\phi^{k+1}_{\lambda,h} - h^\lambda\| \leq 4\|\phi^k_{\lambda,h} - h^\lambda\|.$$

Using iteratively this inequality, we get:

$$\|\phi^{k+1}_{\lambda,h}-h^\lambda\| \leq 4^k\|h-h^\lambda\|,\ \ \|\phi^{k+1}_{\lambda,h}-\phi^k_{\lambda,h}\| \leq 3\cdot 4^{k-1}\|h-h^\lambda\|. \tag{63}$$

Furthermore, $\text{sp}(q^k_{\lambda,h}) \leq 2\|q^k_{\lambda,h}\|$. Thus, (61) is satisfied with:

$$B(h,\lambda) = 6\cdot 4^{m-1}\|h-h^\lambda\| = B_0\|h-h^\lambda\|. \tag{64}$$

Define K as the minimum positive integer such that $\rho = B_0\xi^K/(1-\xi) < 1$. Then the function:

$$V(h,\lambda) = \sum_{k=0}^{K-1}\|\phi^k_{\lambda,h}-h^\lambda\|^2, \tag{65}$$

is a Lyapunov function for the dynamics (59). Indeed:

$$V(F_\lambda h,\lambda)-V(h,\lambda) = \|\phi^k_{\lambda,h}-h^\lambda\|^2-\|h-h^\lambda\|^2 \leq \\ \leq -(1-\rho)\|h-h^\lambda\|^2, \tag{66}$$

where the last inequality holds true due to (62).

The proof of the proposition depends on the following four lemmas, the proof of which is presented after the proof of the proposition.

***Lemma** 7:* There is a positive constant L_1 such that:

$$\|h^\lambda - h^{\lambda'}\| \leq L_1\|\lambda-\lambda'\|,$$

for any λ, λ'

Proof: The limit of the dynamics $\phi^{k+1}_{\lambda,h'} = F_\lambda\phi^k_{\lambda,h'}$ is h^λ irrespectively of the initial condition h'. Using $h^{\lambda'}$ as the initial condition, we get $h^\lambda = \lim_{k\to\infty}\phi^k_{h^{\lambda'},\lambda}$.

Claim: If for some constant Γ it holds $\|h-h^{\lambda'}\| \leq \Gamma$ then:

$$\|F_\lambda h - h\| \leq 2\bar{G}|\lambda-\lambda'| + 4\Gamma, \tag{67}$$

$$\|F_\lambda h - h^{\lambda'}\| \leq 2\bar{G}|\lambda-\lambda'| + 5\Gamma. \tag{68}$$

To prove the claim, observe that:

$$\|F_\lambda h - h\| \leq \|F_\lambda h - F_{\lambda'}h\| + \|F_{\lambda'}h - F_{\lambda'}h^{\lambda'}\| + \\ + \|F_{\lambda'}h^{\lambda'} - h\|.$$

Using both inequalities in (58) we get (67). Inequality (68) is proved using (67) and triangle inequality, which completes the proof of the claim.

We then apply (67), (68) in $\phi^k_{h^{\lambda'},\lambda}$. For $k=0$, we have $\|\phi^0_{h^{\lambda'},\lambda} - h^{\lambda'}\| = 0$. Thus, applying (68) recursively we get:

$$\|\phi^k_{h^{\lambda'},\lambda} - h^{\lambda'}\| \leq 2\bar{G}|\lambda-\lambda'|(5^k-1) < 2\bar{G}|\lambda-\lambda'|5^k.$$

Hence, applying (67) it holds:

$$\text{sp}(q^k_{\lambda,h^{\lambda'}}) \leq 2\|F_\lambda\phi^k_{h^{\lambda'},\lambda} - \phi^k_{h^{\lambda'},\lambda}\| \leq \\ \leq 4\bar{G}|\lambda-\lambda'| + 16\Gamma\bar{G}|\lambda-\lambda'|5^k < \Gamma\bar{G}|\lambda-\lambda'|5^{k+1}.$$

Then:

$$\max_{k=0,\ldots,m}\text{sp}(q^k_{\lambda,h})/(1-\varepsilon) \leq \bar{G}|\lambda-\lambda'|5^{m+1}/(1-\varepsilon),$$

Therefore, using (62) with $k=0$ and the fact that $h^\lambda = \lim_{k\to\infty}\phi^k_{h^{\lambda'},\lambda}$ we get:

$$\|h^\lambda - h^{\lambda'}\| \leq \frac{\bar{G}5^{m+1}}{(1-\varepsilon)(1-\xi)}|\lambda-\lambda'| = L_1|\lambda-\lambda'|,$$

which completes the proof of the lemma. □

***Lemma** 8:* Denote by $v(h,\lambda) = (T_\lambda h)(n)$. Then, if $\lambda, \lambda' > \lambda^\star - \delta_\lambda$ and $h, h' \in \mathbb{R}^n$, it holds:

$$|v(h,\lambda) - v(h',\lambda)| \leq \|h-h'\|, \tag{69}$$

$$A_2|\lambda-\lambda'| \leq |v(h^\lambda,\lambda) - v(h^{\lambda'},\lambda')| \leq \bar{G}|\lambda-\lambda'|, \tag{70}$$

$$A_2|\lambda-\lambda^\star| \leq |v(h^\lambda,\lambda)| \leq \bar{G}|\lambda-\lambda^\star|, \tag{71}$$

$$(\lambda-\lambda^\star)v(h^\lambda,\lambda) \leq -A_2(\lambda-\lambda^\star)^2. \tag{72}$$

Proof: Inequality (69) is a direct consequence of (57). To prove (70) consider a $\lambda' > \lambda$. Recall that $v(h^\lambda,\lambda)$ and $v(h^{\lambda'},\lambda')$ are the minimum values for problems in the form (30). Denote by $\mathcal{M}$ the class of policies μ that satisfy $\lambda_2(\mu) \geq A_2$. Under any policy $\mu \in \mathcal{M}$ it holds:

$$\lambda_1(\mu) - \lambda'\cdot\lambda_2(\mu) + (\lambda'-\lambda)\min_{\mu'\in\mathcal{M}}[\lambda_2(\mu')] \leq \lambda_1(\mu) - \\ -\lambda\cdot\lambda_2(\mu) \leq \lambda_1(\mu) - \lambda'\cdot\lambda_2(\mu) + (\lambda'-\lambda)\max_{\mu'\in\mathcal{M}}[\lambda_2(\mu')].$$

Taking the minimum with respect to μ, we get:

$$v(h^{\lambda'},\lambda') + (\lambda'-\lambda)\min_{\mu'\in\mathcal{M}}[\lambda_2(\mu')] \leq v(h^\lambda,\lambda) \leq \\ \leq v(h^{\lambda'},\lambda') + (\lambda'-\lambda)\max_{\mu'}[\lambda_2(\mu')].$$

Furthermore, $A_2 \leq \min_{\mu'}[\lambda_2(\mu')] \leq \min_{\mu'}[\lambda_2(\mu')] \leq \bar{G}$. Thus,

$$A_2(\lambda-\lambda') \leq v(h^\lambda,\lambda) - v(h^{\lambda'},\lambda') \leq \bar{G}(\lambda-\lambda'),$$

if $\lambda' > \lambda$. Interchanging the roles of λ and λ' we prove (70).

To prove (71) recall that $v(h^{\lambda^\star},\lambda^\star) = 0$. Inequality (72) follows similarly. □

***Lemma** 9:* There are positive constants L_2, L_3 such that:

$$|V(h,\lambda) - V(h,\lambda')| \leq L_2|\lambda-\lambda'|^2 + L_3|\lambda-\lambda'|\|h-h^\lambda\| \tag{73}$$

Proof: It holds:

$$|V(\lambda,h) - V(\lambda',h)| \leq \sum_{k=0}^{K-1}\left|\|\phi^k_{\lambda,h}-h^\lambda\|^2 - \|\phi^k_{\lambda',h}-h^{\lambda'}\|^2\right|.$$

Let us write each term of the summation as:

$$\left|\|\phi^k_{\lambda,h}-h^\lambda\| - \|\phi^k_{\lambda',h}-h^{\lambda'}\|\right|(\|\phi^k_{\lambda,h}-h^\lambda\| + \|\phi^k_{\lambda',h}-h^{\lambda'}\|). \tag{74}$$

Applying twice the triangle inequality, the first factor of (74) can be bounded by:

$$\left|\|\phi^k_{\lambda,h}-h^\lambda\| - \|\phi^k_{\lambda',h}-h^{\lambda'}\|\right| \leq \|\phi^k_{\lambda,h}-\phi^k_{\lambda',h}\| + \|h^\lambda-h^{\lambda'}\|.$$

Claim: There is a positive constant L_5 such that:

$$\|\phi^k_{\lambda,h} - \phi^k_{\lambda',h}\| \leq L_4|\lambda-\lambda'|,$$

for $k=0,\ldots,K$.

To prove the claim, observe that $\|\phi^0_{\lambda,h} - \phi^0_{\lambda',h}\| = 0$ and that for any vectors h', h'', it holds:

$$\|F_\lambda h' - F_{\lambda'}h''\| \leq \|F_\lambda h' - F_{\lambda'}h'\| + \|F_{\lambda'}h' - F_{\lambda'}h''\| \\ \leq 2\bar{G}|\lambda-\lambda'| + 2\|h'-h''\|.$$

To prove the claim it is sufficient to apply recursively the last inequality for $k=0,\ldots,K$.

Using the claim, we conclude that:

$$\left| \|\phi^k_{\lambda,h} - h^\lambda\| - \|\phi^k_{\lambda',h} - h^{\lambda'}\| \right| \leq (L_1 + L_4)|\lambda - \lambda'|,$$

for $k = 0, \ldots, K$.

Using (62), (64) to the second factor of (74) we get:

$$\begin{aligned}\|\phi^k_{\lambda,h} - h^\lambda\| + \|\phi^k_{\lambda',h} - h^{\lambda'}\| &\leq \frac{B_0}{1-\xi}(\|h - h^\lambda\| + \|h - h^{\lambda'}\|) \\ &\leq \frac{B_0}{1-\xi}(2\|h - h^\lambda\| + L_1|\lambda - \lambda'|).\end{aligned}$$

Taking $L_2 = \frac{KL_4L_1B_0}{1-\xi}$ and $L_2 = \frac{2KL_4B_0}{1-\xi}$ we conclude to the desired result. □

Lemma 10: If γ^k is small enough, then for a $\delta < \delta_\lambda/2$, there is a k_0 such that the state (λ^k, h^k) satisfies $\|h^k - h^{\lambda_k}\| \leq \delta$ and $\lambda^k > \lambda^\star - \delta$, for all $k > k_0$.

Proof: Refer to Figure 11. For some k, denote by $h = h^k$, $h^+ = h^{k+1}$, $\lambda = \lambda^k$, $\lambda^+ = \lambda^{k+1}$ and assume that $0 > \lambda > \lambda^\star - \delta_\lambda$. The Lyapunov function (65) satisfies:

$$\begin{aligned}&V(h^+, \lambda^+) - V(h, \lambda) = \\ &= V(h^+, \lambda^+) - V(h^+, \lambda) + V(h^+, \lambda) - V(h, \lambda) \\ &\leq L_2\gamma^2 v^2(h, \lambda) + L_3\gamma v(h, \lambda)\|h - h^\lambda\| - (1-\rho)\|h - h^\lambda\|^2,\end{aligned} \tag{75}$$

where $v(h, \lambda) = (T_\lambda h)(n)$ and (73) was used. Furthermore, using (69), (71) we get:

$$v(h, \lambda) \leq \|h - h^\lambda\| + v(h^\lambda, \lambda) \leq \|h - h^\lambda\| + \bar{G}|\lambda - \lambda^\star|.$$

Substituting back to (75), we get:

$$\begin{aligned}V(h^+, \lambda^+) - V(h, \lambda) \leq &-[(1-\rho) - L_2\gamma^2 + L_3\gamma]\|h - h^\lambda\|^2 + \\ &+ (L_2\gamma + L_3)\gamma\bar{G}|\lambda^\star - \delta_\lambda|\|h - h^\lambda\| + L_2\gamma^2\bar{G}|\lambda^\star - \delta_\lambda|^2.\end{aligned}$$

The last inequality implies that if γ is small enough, the dynamics (27)-(29) will enter the gray area in Figure 11 and remain there as soon as $\lambda > \lambda_\star$. However, in this region, (76) is a Laypunov function and thus, the dynamics will enter into the circle in Figure 11. This is true, due to the fact that for γ small there is no possibility of moving from the left to the right of the cycle within the gray area in a single step. The circle is positively invariant and thus, λ will remain greater than or equal to $\lambda^\star - \delta_\lambda$. □

Now consider the dynamics (27)-(29) and the Lyapunov function candidate:

$$V_c(h, \lambda) = V(h, \lambda) + (\lambda - \lambda^\star)^2. \tag{76}$$

For $h = h^k$, $\lambda = \lambda^k$ and $h^+ = h^{k+1}, \lambda^+ = \lambda^{k+1}$ and assuming that $\lambda > \lambda^\star - \delta$ we have:

$$\begin{aligned}&V_c(h^+, \lambda^+) - V_c(h, \lambda) = (V(h^+, \lambda^+) - V(h^+, \lambda)) \\ &+ (V(h^+, \lambda) - V(h, \lambda)) + ((\lambda^+ - \lambda^\star)^2 - (\lambda - \lambda^\star)^2)\end{aligned} \tag{77}$$

The first term of (77) is bounded above using Lemma 3 and (63):

$$\begin{aligned}&V(h^+, \lambda^+) - V(h^+, \lambda) \leq \\ &\leq L_2(\gamma v(h, \lambda))^2 + L_3\gamma|v(h, \lambda)|\|h^+ - h^\lambda\| \\ &\leq L_2(\gamma v(h, \lambda))^2 + 3L_3\gamma|v(h, \lambda)|\|h - h^\lambda\|.\end{aligned}$$

The second term of the right-hand side of (77) is bounded above by $-(1-\rho)\|h - h^\lambda\|^2$, due to (66). The last term may be rewritten as $[(\gamma v(h, \lambda))^2 + 2\gamma v(h, \lambda)(\lambda - \lambda^\star)]$. Thus:

$$\begin{aligned}&V_c(h^+, \lambda^+) - V_c(h, \lambda) \leq \\ &\leq (L_2 + 1)(\gamma v(h, \lambda))^2 + 3L_3\gamma|v(h, \lambda)|\|h - h^\lambda\| - \\ &- (1-\rho)\|h - h^\lambda\|^2 + 2\gamma v(h, \lambda)(\lambda - \lambda^\star)\end{aligned} \tag{78}$$

Using $v(h, \lambda) = v(h^\lambda, \lambda) + (v(h, \lambda) - v(h^\lambda, \lambda))$, along with (69), (71) and (72) we get:

$$\begin{aligned}(v(h, \lambda))^2 \leq &\bar{G}^2|\lambda - \lambda^\star|^2 + \\ &+ \|h - h^\lambda\|^2 + 2\bar{G}|\lambda - \lambda^\star|\|h - h^\lambda\|,\end{aligned} \tag{79}$$

$$\begin{aligned}&v(h, \lambda)(\lambda - \lambda^\star) \leq \\ &\leq -\bar{G}'(\lambda - \lambda^\star)^2 + \|h - h^\lambda\||\lambda - \lambda^\star|.\end{aligned} \tag{80}$$

Substituting (79), (80) and the right-hand side inequality of (71) into (78) we get:

$$\begin{aligned}&V_c(h^+, \lambda^+) - V_c(h, \lambda) \leq \\ &\leq -\|h - h^\lambda\|^2[(1-\rho) + \bar{G}^2\gamma^2(L_2 + 1)] + \\ &+ \|h - h^\lambda\||\lambda - \lambda^\star|[2\gamma + 2\bar{G}\gamma^2(L_2 + 1) + 3L_3\gamma\bar{G}] - \\ &- |\lambda - \lambda^\star|^2[2\bar{G}'\gamma - \bar{G}^2\gamma^2(L_2 + 1)^2],\end{aligned}$$

which for appropriate positive constants $C_1, \ldots, C_5$, can be written as:

$$\begin{aligned}V_c(h^+, \lambda^+) - V_c(h, \lambda) \leq &-\begin{bmatrix}\|h - h^\lambda\| & |\lambda - \lambda^\star|\end{bmatrix} \cdot \\ &\cdot \begin{bmatrix}(1-\rho) + \gamma^2 C_1 & -C_2\gamma - \gamma^2 C_3 \\ -C_2\gamma - \gamma^2 C_3 & C_4\gamma - \gamma^2 C_5\end{bmatrix} \begin{bmatrix}\|h - h^\lambda\| \\ |\lambda - \lambda^\star|\end{bmatrix}.\end{aligned}$$

Let us denote by $A(\gamma)$ the matrix in the right-hand side of the last equation. It is not difficult to see that there is a positive constant $\bar{\gamma}$ such that for $\gamma < \bar{\gamma}$ the matrix $A(\gamma)$ is positive definite. Furthermore, it is not difficult to see that fixing a γ, such that $0 \leq \gamma \leq \bar{\gamma}$, there is a positive constant ε_0 such that $A(\gamma) \succeq \varepsilon_0 I$, for all $\gamma \leq \gamma \leq \bar{\gamma}$. Therefore, $\lambda^k \to \lambda^\star$ and $h^k \to h^{\lambda^\star} = h^\star/\tau$ at a rate of a geometric progression.

Now due to Lemma 10, the state (λ^k, h^k) will enter the gray region in Figure 11. In this region, $\lambda > \lambda^\star - \delta_\lambda$ and the Lyapunov analysis above holds true.

REFERENCES

[1] A. Barré, B. Deguilhem, S. Grolleau, M. Gérard, F. Suard, and D. Riu, "A review on lithium-ion battery ageing mechanisms and estimations for automotive applications," *Journal of Power Sources*, vol. 241, pp. 680–689, 2013.

[2] C. D. Barley and C. B. Winn, "Optimal dispatch strategy in remote hybrid power systems," *Solar Energy*, vol. 58, no. 4-6, pp. 165–179, 1996.

[3] M. Koller, T. Borsche, A. Ulbig, and G. Andersson, "Defining a degradation cost function for optimal control of a battery energy storage system," in *PowerTech (POWERTECH), 2013 IEEE Grenoble*. IEEE, 2013, pp. 1–6.

[4] W. Ying, Z. Zhi, A. Botterud, K. Zhang, and D. Qia, "Stochastic coordinated operation of wind and battery energy storage system considering battery degradation," *Journal of Modern Power Systems and Clean Energy*, vol. 4, no. 4, pp. 581–592, 2016.

[5] Y. Shi, B. Xu, D. Wang, and B. Zhang, "Using battery storage for peak shaving and frequency regulation: Joint optimization for superlinear gains," *IEEE Transactions on Power Systems*, 2017.

[6] B. Xu, Y. Shi, D. S. Kirschen, and B. Zhang, "Optimal regulation response of batteries under cycle aging mechanisms," *arXiv preprint arXiv:1703.07824*, 2017.

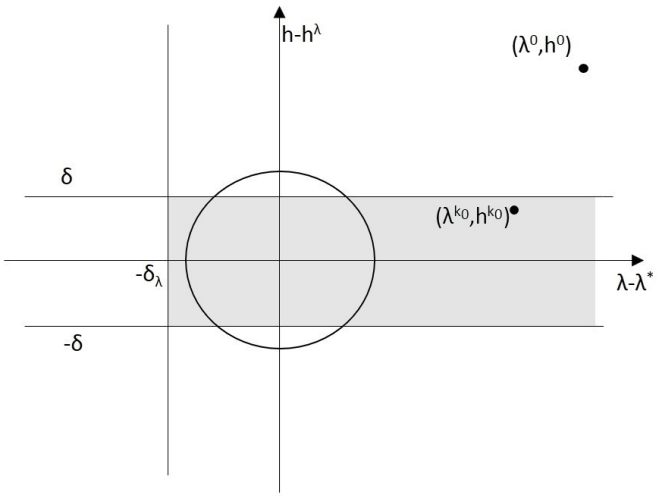


Fig. 11: The Lyapunov argument for Theorem 2

[7] P. Haessig, H. B. Ahmed, and B. Multon, "Energy storage control with aging limitation," in *PowerTech*, 2015.

[8] D.-I. Stroe, V. Knap, M. Swierczynski, A.-I. Stroe, and R. Teodorescu, "Suggested operation of grid-connected lithium-ion battery energy storage system for primary frequency regulation: Lifetime perspective," in *Energy Conversion Congress and Exposition (ECCE), 2015 IEEE*. IEEE, 2015, pp. 1105–1111.

[9] M. Kazemi and H. Zareipour, "Long-term scheduling of battery storage systems in energy and regulation markets considering batterys lifespan," *IEEE Transactions on Smart Grid*, 2017.

[10] P. Carpentier, J.-P. Chancelier, M. De Lara, and T. Rigaut, "Algorithms for two-time scales stochastic optimization with applications to long term management of energy storage," 2019.

[11] B. Heymann, P. Martinon, and F. Bonnans, "Long term aging: an adaptative weights dynamic programming algorithm," 2016.

[12] P. Carpentier, J.-P. Chancelier, M. De Lara, and T. Rigaut, "Time blocks decomposition of multistage stochastic optimization problems," *arXiv preprint arXiv:1804.01711*, 2018.

[13] X. Tan, Y. Wu, and D. H. Tsang, "A stochastic shortest path framework for quantifying the value and lifetime of battery energy storage under dynamic pricing," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 769–778, 2017.

[14] ——, "Pareto optimal operation of distributed battery energy storage systems for energy arbitrage under dynamic pricing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 7, pp. 2103–2115, 2016.

[15] C. Derman, "On sequential decisions and Markov chains," *Management Science*, vol. 9, no. 1, pp. 16–24, 1962.

[16] M. Klein, "Inspection maintenance replacement schedules under markovian deterioration," *Management Science*, vol. 9, no. 1, pp. 25–32, 1962.

[17] W. S. Jewell, "Markov-renewal programming II: Infinite return models, example," *Operations Research*, vol. 11, no. 6, pp. 949–971, 1963.

[18] S. M. Ross, "Average cost semi-Markov decision processes," *Journal of Applied Probability*, vol. 7, no. 3, pp. 649–656, 1970.

[19] D. Bertsekas, *Dynamic programming and optimal control*. Athena scientific Belmont, MA, 1995.

[20] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[21] T. K. Das, A. Gosavi, S. Mahadevan, and N. Marchalleck, "Solving semi-Markov decision problems using average reward reinforcement learning," *Management Science*, vol. 45, no. 4, pp. 560–574, 1999.

[22] A. Gosavi, "Reinforcement learning for long-run average cost," *European Journal of Operational Research*, vol. 155, no. 3, pp. 654–674, 2004.

[23] Z. Ren and B. H. Krogh, "Markov decision processes with fractional costs," *IEEE transactions on automatic control*, vol. 50, no. 5, pp. 646–650, 2005.

[24] J. Yang, Y. Li, H. Chen, and J. Li, "Average reward reinforcement learning for semi-Markov decision processes," in *International Conference on Neural Information Processing*. Springer, 2017, pp. 768–777.

[25] C. Von Essen and B. Jobstmann, "Synthesizing systems with optimal average-case behavior for ratio objectives," *arXiv preprint arXiv:1102.4118*, 2011.

[26] ——, "Synthesizing efficient controllers," in *International Workshop on Verification, Model Checking, and Abstract Interpretation*. Springer, 2012, pp. 428–444.

[27] C. Von Essen, B. Jobstmann, D. Parker, and R. Varshneya, "Synthesizing efficient systems in probabilistic environments," *Acta Informatica*, vol. 53, no. 4, pp. 425–457, 2016.

[28] D. P. Bertsekas, "A new value iteration method for the average cost dynamic programming problem," *SIAM journal on control and optimization*, vol. 36, no. 2, pp. 742–759, 1998.

[29] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[30] H. K. Khalil, "Nonlinear systems, 3rd," *New Jewsey, Prentice Hall*, vol. 9, 2002.

[31] B. Xu, Y. Dvorkin, D. S. Kirschen, C. A. Silva-Monroy, and J.-P. Watson, "A comparison of policies on the participation of storage in US frequency regulation markets," in *Power and Energy Society General Meeting (PESGM), 2016*. IEEE, 2016, pp. 1–5.

[32] M. Koller, T. Borsche, A. Ulbig, and G. Andersson, "Review of grid applications with the zurich 1 mw battery energy storage system," *Electric Power Systems Research*, vol. 120, pp. 128–135, 2015.

[33] S. Chen, H. B. Gooi, and M. Wang, "Sizing of energy storage for microgrids," *IEEE Transactions on Smart Grid*, vol. 3, no. 1, pp. 142–151, 2011.

[34] D. U. Sauer and H. Wenzl, "Comparison of different approaches for lifetime prediction of electrochemical systemsusing lead-acid batteries as example," *Journal of Power sources*, vol. 176, no. 2, pp. 534–546, 2008.

[35] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[36] R. Van Handel, "Stochastic calculus, filtering, and stochastic control," *Course notes., URL http://www. princeton. edu/rvan/acm217/ACM217. pdf*, vol. 14, 2007.

[37] V. I. Bogachev, *Measure theory*. Springer Science & Business Media, 2007, vol. 1.

Ioannis Kordonis received both his Diploma in Electrical and Computer Engineering and his Ph.D. degree from the National Technical University of Athens, Greece, in 2009 and 2015 respectively. During 2016-2017 he was a post-doctoral researcher in the University of Southern California and he is currently employed at temporary teaching and research position (ATER) at CentraleSupélec, on the Rennes campus, in the Automatic Control Group - IETR.

His research interests include Game Theory with an emphasis on Dynamic Games and Stochastic Control theory. He is also interested in applications in the areas of Energy - Power Systems, Transportation Systems and in Bioengineering.

PLACE PHOTO HERE

Alexandros C. Charalampidis was born in Athens, Greece, in 1984. He received the Diploma from the National Technical University of Athens (NTUA), School of Electrical and Computer Engineering, in 2007 and the PhD from the same school in 2011. In 2012-2014 he was a postdoctoral researcher at the Swiss Federal Institute of Technology in Lausanne (EPFL) and in 2014-2015 he was a research associate at NTUA. Since 2016 he is an Associate Professor ("maître de conférences") at CentraleSupélec, on the Rennes campus, in the Automatic Control Group - IETR. Temporarily (2017-2019), he is also with TU Berlin, in the Control Systems Group. His research interests are in the field of Systems and Control. Specifically, his theoretic interests mainly concern nonlinear stochastic systems (Stochastic Estimation and Optimal Control) and he has experience with applications in Power Systems and Biomedical Engineering as well as with autonomous road vehicles.

Pierre Haessig received the Master degree in electrical engineering from École Normale Suprieure (ENS) Paris-Saclay, France, in 2011, and the PhD degree in 2014.

He is an assistant professor at CentraleSupélec, in Rennes, France, where he conducts his research in the Automatic Control team of the IETR lab. His research interests include the sizing and the management of Energy Storage Systems, for mitigating the fluctuations of renewable energies (wind and solar), and more generally the optimization of energy systems in the face of uncertainty.