The final version of this manuscript will appear in IEEE Wireless Communication Letters, 2013.
DOI: 10.1109/WCL.2013.13.130213

Markov Decision Processes with Threshold Based Piecewise Linear Optimal Policies

Tomaso Erseghe, Andrea Zanella, Claudio G. Codemo

***Abstract*—This letter investigates the structure of the optimal policy for a class of Markov decision processes (MDPs), having convex and piecewise linear cost function. The optimal policy is proved to have a piecewise linear structure that alternates flat and constant-slope pieces, resembling a staircase with tilted rises and as many steps (thresholds) as the breakpoints of the cost function. This particular structure makes it possible to express the policy in a very compact manner, particularly suitable to be stored in low-end devices. More importantly, the threshold-based form of the optimal policy can be exploited to reduce the computational complexity of the iterative dynamic programming (DP) algorithm used to solve the problem. These results apply to a rather wide set of optimization problems, typically involving the management of a resource buffer such as the energy stored in a battery, or the packets queued in a wireless node.**

***Index Terms*—dynamic programming, low complexity, Markov decision processes, optimization, piecewise linear policy**

I. Introduction

A Markov decision process (MDP) is a mathematical framework for modeling decision making problems in stochastic systems whose state evolution is partially random and partially under the control of a decision maker. In this context, a *policy* π is a function that maps each system state x into an action a that takes values in the admissible region $\mathcal{A}(x)$. Given the policy π, the system state evolves as a discrete-time Markov process, so that the next state only depends upon the current state x and the associated action $a = \pi(x)$. Furthermore, each transition incurs in penalty γ that only depends upon the current state x and the corresponding action a. The aim is to find the policy π^* that minimizes some cumulative function of γ, e.g., the average cost per stage defined as

$$\lambda = \min_{\pi} \left\{ \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{E}\left[\gamma(\pi(x(t)), x(t))\right] \right\} \qquad (1)$$

where the expectation is with respect to all possible transitions from state $x(t)$, with action $\pi(x(t))$. A common method to solve MDP problems is dynamic programming (DP) [1], which often requires the iterative evaluation of the cost function for all possible states and admissible actions, until convergence to the optimal policy π^*. This method, however, may be computational demanding for large state space and action regions. Furthermore, the optimal policy π^* is typically expressed as an action vector with the same size of the state space.

The authors are with Dipartimento di Ingegneria dell'Informazione, Università di Padova, Via G. Gradenigo 6/B, 35131 Padova, Italy. tel: +39 049 827 7656, fax: +39 049 827 7699, mailto: {erseghe, zanella, codemocl}@dei.unipd.it

In this letter we prove that a certain class of MDPs characterized, among other properties, by a convex and piecewise linear cost function γ, admits an optimal policy with extremely simple piecewise linear structure, which alternates flat pieces (thresholds) and linear segments with constant negative slope, thus resembling the shape of a staircase with tilted rises and irregular steps as exemplified in later Fig. 1. A similar finding was presented in [2], where a two-threshold optimal policy was identified under the assumption of *linear* and *non-decreasing* cost functions. Here we generalize the result to *piecewise linear* cost function, relaxing the monotonicity assumption with *convexity*. The particular structure of π^* makes it possible to code the policy in a very compact form, particularly suitable to be stored in resource-limited devices. Furthermore, we propose a value iteration method that searches for the breakpoints and intercepts of the optimal policy rather than for a policy with generic shape, thus greatly reducing the computational complexity.

II. Problem model and examples of application

Formally, the class of MDPs considered in this letter is characterized by the following properties.

A1 The system state can be expressed as $x = [x_s, x_v]$, where the evolution of (scalar) x_s is under the control of the (scalar) action $a \in \mathcal{A}(x)$ through a deterministic function $f(a, x)$, while x_v evolves as a (possibly vectorial) Markov process independent of a. Furthermore, x_s can take values in a convex set $\mathcal{X}_s$. Denoting by $y = [y_s, y_v]$ the next state of the system, the update transition probability can hence be expressed as

$$p(y|a,x) = \delta(y_s - f(a,x))\, p(y_v|x_v) \ , \qquad (2)$$

with $\delta(\cdot)$ the Kronecker delta function, and $p(y_v|x_v)$ the transition probability of the Markov process x_v.

A2 The deterministic function $f(a, x)$ is linear in both the action a and the previous state x_s, i.e.,

$$f(a,x) = c_1(x_v) \cdot x_s + c_2(x_v) \cdot a + c_3(x_v) \ ,$$

where $c_i(\cdot)$ are coefficients that may depend on x_v, with $c_1(\cdot)$ and $c_2(\cdot)$ always positive. This assumption basically requires the action to affect directly and linearly on x_s, ruling out all the cases where the action has (even partially) random effects on x_s.

A3 The cost function γ only depends on x_v and a, i.e.,

$$\gamma(a,x) = \gamma(a,x_v) \ . \qquad (3)$$

Moreover, $\gamma(a, x_v)$ is *convex* and *piecewise linear* in a. Hence, for any given x_v, function $\gamma(\cdot, x_v)$ is identified

by a ordered set of breakpoints, $b_1(x_v) < \ldots < b_B(x_v)$, and an ordered set of derivatives, $d_1(x_v) < \ldots < d_{B+1}(x_v)$. For later use, we also set $b_0(x_v) = -\infty$ and $b_{B+1}(x_v) = +\infty$. Note that, while x_s does not impact directly on the instantaneous cost γ, it does play a role in determining the long-term average cost λ because it defines the range $\mathcal{A}(x)$ of admissible actions.

A4 Given x_v, the action space

$$\mathcal{R}(x_v) = \left\{ (a, x_s) \,\middle|\, a \in \mathcal{A}(x_s, x_v),\, x_s \in \mathcal{X}_s \right\} \quad (4)$$

is closed and convex.

The model defined by the assumptions A1-A4 broadly applies to systems that meet the Markovian resource request x_v by possibly using part of the buffered resources x_s. The decision problem consists in determining the optimal buffering action $a = \pi(x_s, x_v)$, subject to some state-dependent constraints $\mathcal{A}(x_s, x_v)$, in order to minimize the long term average of cost γ (see, e.g., [3], [4]). We discuss an application example.

Power control under buffer constraints [5]: Assume that, at every slot of length T, a sensor generates L bits to be delivered to a controller. Let x_s be the number of queued bits at the beginning of a given slot, and x_v the channel gain which is supposed to be constant over one slot, but evolving in time as a Markov process. Assume that the number of bits transmitted per slot is $d = WT \log_2(1 + P\,x_v/(W\,N_0))$ where W is the channel bandwidth, T the slot length, P the transmit power, and N_0 the noise power density. The buffer size at the next slot is $y_s = f(a, x) = x_s + a + L$, and the action is $a = -d$. The objective is to minimize the long-term average power, $\overline{P}$, given that the queue length x_s cannot exceed threshold Q. The action space (4) is given by $\mathcal{X}_s = [0, Q]$ and $\mathcal{A}(x_s, x_v) = [-x_s - L, Q - x_s - L] \cap [-\infty, 0]$. The cost function is

$$\gamma(a, x_v) = \frac{P}{N_0 W} = \frac{2^{-a/(WT)} - 1}{x_v}\,, \quad (5)$$

which is convex in a for any given x_v, but needs to be approximated as piecewise linear to apply the proposed method.

III. Optimal Threshold-Based Policy

DP theory [1] assures that an optimal stationary policy π^* for problem (1) can always be identified by solving the following Belmann's equation

$$\lambda + g(x) = \min_{a \in \mathcal{A}(x)} \gamma(a, x) + \int p(y|a, x) g(y)\, dy\,, \quad (6)$$

where $g(x)$ is some function satisfying $g(0) = 0$. Then, for a given state x, the optimal stationary policy is identified by $\pi^*(x) = a^*$, with a^* the point of minimum in (6). A classical method to solve the optimization problem (6) is by means of value iteration, i.e., via the recursion

$$\begin{aligned} U_t(a, x) &= \int p(y|a, x) g_t(y)\, dy \\ \pi_t^*(x) &= \operatorname*{argmin}_{a \in \mathcal{A}(x)} \gamma(a, x) + U_t(a, x) \\ \tilde{g}_{t+1}(x) &= \gamma(\pi_t^*(x), x) + U_t(\pi_t^*(x), x) \\ g_{t+1}(x) &= \tilde{g}_{t+1}(x) - \tilde{g}_{t+1}(0) \end{aligned} \quad (7)$$

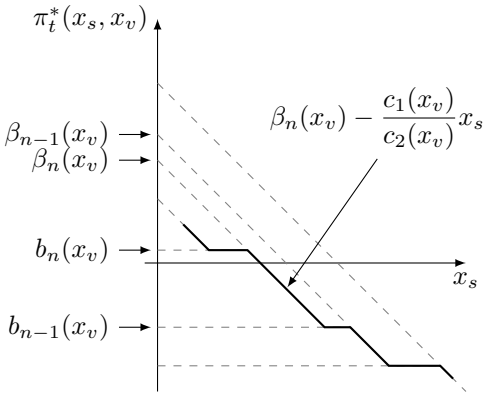

Fig. 1. Staircase structure of the optimal policy $\pi_t^*(x)$ as a function of x_s and for a given x_v.

which starts from $g_0(x) = 0$. It can be proved that, as the time horizon grows to infinity, g_t converges to g in (6) and π_t^* converges to an optimal strategy π^*.

Note that from assumption A1 it is

$$\begin{aligned} U_t(a, x) &= G_t(f(a, x), x_v) \\ G_t(x) &= \int p(y_v|x_v) g_t(x_s, y_v)\, dy_v\,. \end{aligned} \quad (8)$$

The main consequence of assumptions A1-A4 is, however, that the optimal policy $\pi_t^*(x_s, x_v)$ exhibits, for any given x_v, the regular threshold-based structure illustrated in Fig. 1, which can be very compactly expressed through the ordered sequence of intercepts on the π_t^*-axis. The formal result is provided by the following theorem whose proof is given in Section IV.

Theorem 1. *Under A1-A4, an optimal strategy $\pi_t^*(x)$ is*

$$\pi_t^*(x) = \max\Big(\min \mathcal{A}(x), \min\Big(\tilde{\pi}_t(x), \max \mathcal{A}(x) \Big) \Big) \quad (9)$$

where $\tilde{\pi}_t(x_s, x_v)$ is, for any given x_v, a non-increasing piecewise linear function of x_s with staircase shape as sketched in Fig. 1. This function is uniquely identified by the points where the pieces prolongations intercept the π_t^*-axis, which are equal to the breakpoints $b_n(x_v)$ of γ (see A3), and the points

$$\beta_n(x_v) = \frac{w_n(x_v) - c_3(x_v)}{c_2(x_v)} \quad (10)$$

where $c_i(x_v)$ are the linear coefficients of $f(a, x)$ (see A2), while constants $w_n(x_v)$ are defined via

$$w_n(x_v) = \operatorname*{argmin}_{x_s \in \mathcal{X}_s} \Big| d_n(x_v) + c_2(x_v)\, \partial_{x_s} G_t(x_s, x_v) \Big|\,, \quad (11)$$

where $|\cdot|$ denotes the absolute value operator applied to each of the points in the target set, $d_n(x_v)$ are the derivatives of $\gamma(a, x_v)$ (see A3), and $\partial_{x_s} G_t$ is the sub-differential[1] with

[1]The sub-differential generalizes the derivative to convex functions which are not differentiable. In our context it equals the derivative of the function where it exists, while it equals the *interval* between the left and right derivatives of the function in correspondence of breakpoints. The sub-differential is a monotone (non-decreasing) operator.

respect to x_s of the convex function $G_t(x)$ defined in (8). Formally, we hence have

$$\tilde{\pi}_t(x) = \max_{n=1,\dots,B+1} \min\left(\beta_n(x_v) - \frac{c_1(x_v)}{c_2(x_v)} x_s, b_n(x_v)\right) . \quad (12)$$

Being Theorem 1 valid for any t, it is also valid in the limit, and therefore the staircase structure holds also for the optimal policy $\pi^* = \lim_{t\to\infty} \pi_t^*$. Despite the simple formulation of (9)-(12), the derivation of constants $w_n(x_v)$ according to (11) may be troublesome because of the necessity of differentiating $G_t(x)$. It is therefore of interest to identify some form of policy iteration based upon the sub-differential $\partial_{x_s} G_t$ in order to fully exploit the results of Theorem 1. This is given by the following result whose proof is given in Section IV.

Theorem 2. *The optimal policy $\pi^*(x)$ can be obtained iteratively by tracking the sub-differential $\partial_{x_s} G_t(x)$, namely:*

1: *Set $\partial_{x_s} G_0(x) = 0$*
2: ***for** $t = 0, 1, \dots$ **do***
3: *Evaluate constants $w_n(x_v)$ using (11)*
4: *Synthesize the optimal policy $\pi_t^*(x)$ using (9), (10), (12)*
5: *Define $\tilde{\partial}_{x_s} \pi_t^*(x)$ as the derivative of $\pi_t^*(x)$ with respect to x_s where it exists, and as the* interval *between the left and right derivatives of the function in correspondence of breakpoints*[2]
6: *Evaluate the sub-differential of $g_{t+1}(x)$ using*

$$\partial_{x_s} g_{t+1}(x) = \partial_a \gamma\Big(\pi_t^*(x), x_v\Big)\, \tilde{\partial}_{x_s} \pi_t^*(x) + \partial_{x_s} G_t\Big(f(\pi_t^*(x), x), x_v\Big)\Big[c_1(x_v) + c_2(x_v)\, \tilde{\partial}_{x_s} \pi_t^*(x)\Big] \quad (13)$$

7: *Update the sub-differential $\partial_{x_s} G_{t+1}(x)$ using*

$$\partial_{x_s} G_{t+1}(x) = \int p(y_v|x_v) \partial_{x_s} g_{t+1}(x_s, y_v)\, dy_v . \quad (14)$$

8: ***end for***

In general Theorem 2 is much more efficient than standard policy iteration (7) because the search for the minimum is greatly simplified. In fact, for a given x_v, evaluating π_t^* only requires identifying constants $w_n(x_v)$ using (11), which is a search on $x_s \in \mathcal{X}_s$ of the points where the (monotone) sub-differential reaches levels $-d_n(x_v)/c_2(x_v)$. Instead, the search for the optimal policy in the second of (7) is performed on the much wider two-dimensional space $\mathcal{R}(x_v)$ defined in A4.

Proof-of-concept example: We apply the proposed method to the power-control example of Section II. We assume constant bit rate data transmission at 64 kbit/s in a 802.15.4 scenario, and set $T = 10$ ms, $L = 640$ bit, $Q = 10\,L$, $W = 156$ kHz (equivalent bandwidth). The wireless channel gain is modeled as Rayleigh fading and simulated using the Jake's model, with maximum Doppler frequency $f_d = 11$ Hz (roughy 5 km/h speed) and average gain of -25 dB. Performance was evaluated for: a) **CP (continuous policy)**, that is, by numerically solving (7) with the continuous cost function

[2]Note that this is not a proper sub-differential since π_t^* is not convex in x_s, and it is hence denoted with a *tilda* as $\tilde{\partial}_{x_s}$.

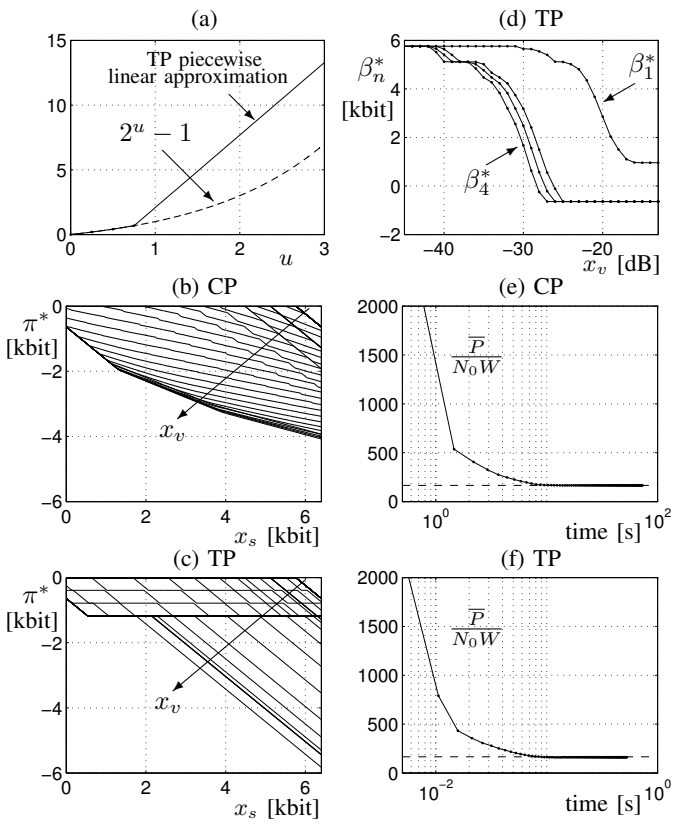

Fig. 2. Power control example: (a) continuous and piecewise linear approximation of the cost function with $B + 1 = 4$ pieces ($u = -a/WT$); (b) optimal CP; (c) optimal TP; (d) parameters $\beta_n(x_v)$ that compactly describe TP; (e)-(f) average normalized power $\overline{P}/(N_0 W)$ for CP and TP, respectively, computed as the value iteration t evolves, and shown as a function of processing time.

γ in (5); b) **TP (threshold based policy)**, that is, by using Theorem 2 with a piecewise linear approximation of γ having four pieces (solid line in Fig. 2a). Transition probabilities $p(y_v|x_v)$ are estimated from channel samples.

Fig. 2b-c report the shape of CP and TP optimal actions, respectively, as a function of x_s (buffer size), and for increasing of x_v (channel gain). Although policies are quite different, with TP exhibiting the threshold-based regular structure predicted by Theorem 1, their performance in terms of $\bar{P}$ is identical. However, TP outperforms CP in two ways. First, Theorem 1 allows for a *compact representation* of TP through coefficients $\beta_n(x_v)$ shown in Fig. 2d. These coefficients are very correlated, so that they can be further compressed yielding an even more compacted representation of TP. Second, thanks to the sub-gradient formulation of Theorem 2, the processing time to compute TP is much lower, as clearly shown in Fig. 2e-f .

IV. Proofs

A. Proof of Theorem 1

The first step towards the identification of a threshold-based optimal policy is the recognition of the second of (7) as a convex minimization problem. Since $\gamma(a, x)$ is assumed

convex by A3, and $\mathcal{A}(x)$ is assumed convex by A4, it remains to prove the convexity of $U_t(a,x)$ in a.

Lemma 3. *Under A1-A4 the function $U_t(a,x)$ is convex in a.*

Proof of Lemma 3: We first by induction on t that g_t is convex in x_s, then show that this implies that U_t is convex in a. For $t=0$ we have $g_0(x)=0$, hence convexity is verified. Assuming g_t convex in x_s for some $t\geq 0$, then also G_t in (8) will be convex in x_s, being a linear combination of g_t. In turn, the linear relation A2 and (8) guarantee that U_t is convex in the couple $[a,x_s]$. We now prove that this result implies the convexity of $\tilde{g}_{t+1}$ in x_s. To this end, for a given x_v, consider two values $x_{s,1}$ and $x_{s,2}$ of x_s, with associated optimal actions

$$a_i^* = \underset{a\in\mathcal{A}(x_{s,i},x_v)}{\operatorname{argmin}} \gamma(a,x_v) + U_t(a,x_{s,i},x_v)\ , \qquad (15)$$

and define $\overline{a}=\alpha a_1+(1-\alpha)a_2$ and $\overline{x}_s=\alpha x_{s,1}+(1-\alpha)x_{s,2}$, with $0\leq\alpha\leq 1$. Note that $\overline{a}\in\mathcal{A}(\overline{x}_s,x_v)$ because of A4. From (7) we can write $\tilde{g}_{t+1}(\overline{x}_s,x_v)\leq\gamma(\overline{a},x_v)+U_t(\overline{a},\overline{x}_s,x_v)$. Now, by exploiting the convexity of both γ (A3 assumption) and U_t we obtain

$$\begin{aligned}\tilde{g}_{t+1}(\overline{x}_s,x_v) &\leq \alpha\gamma(a_1,x_v)+(1-\alpha)\gamma(a_2,x_v)\\ &\quad+\alpha U_t(a_1,x_{s,1},x_v)+(1-\alpha)U_t(a_2,x_{s,2},x_v)\\ &=\alpha\tilde{g}_{t+1}(x_{s,1},x_v)+(1-\alpha)\tilde{g}_{t+1}(x_{s,2},x_v)\end{aligned} \qquad (16)$$

which proves convexity in x_s of $\tilde{g}_{t+1}$. According to the equivalence up to a constant factor stated in the last of (7), convexity of $\tilde{g}_{t+1}$ implies convexity in x_s of g_{t+1}. Hence g_t is proved to be convex in x_s for any $t\geq 0$. Convexity in a of $U_t(a,x_s,x_v)$ follows from convexity in $[a,x_s]$. ∎

Note that, in comparison with [2], we proved that convexity is independent of any non-decreasing property of γ, which is only required to be convex.

Lemma 3 assures that the search for the minimum in (7) is a convex problem, which can be solved by investigating the derivative, or, more correctly, the sub-differential with respect to a of the function $\gamma(a,x)+U_t(a,x)$, which we denote by $\partial_a\gamma(a,x)+\partial_a U_t(a,x)$. Application of this rationale provides the following proof of Theorem 1.

Proof of Theorem 1: Lemma 3 guarantees that $\partial_a\gamma(a,x)+\partial_a U_t(a,x)$ is a continuous monotone (non-decreasing) operator and, hence, the optimal action $\pi_t^*(x)$ is

$$\pi_t^*(x) = \underset{a\in\mathcal{A}(x)}{\operatorname{argmin}} \left|\partial_a\gamma(a,x_v)+\partial_a U_t(a,x)\right| . \qquad (17)$$

The monotonicity of $\partial_a\gamma(a,x_v)+\partial_a U_t(a,x)$ makes it possible to search for the minima in any superset $\mathcal{B}(x)\supseteq\mathcal{A}(x)$, thus obtaining the solution

$$\tilde{\pi}_t(x) = \underset{a\in\mathcal{B}(x)}{\operatorname{argmin}} \left|\partial_a\gamma(a,x_v)+\partial_a U_t(a,x)\right| , \qquad (18)$$

that shall hence be projected onto $\mathcal{A}(x)$. This projection is in fact operated by (9), as assured by A4 which guarantees that $\mathcal{A}(x)$ is a compact interval. We hence focus on the expression of $\tilde{\pi}_t(x)$ by considering $\mathcal{B}(x)=\{a|f(a,x)\in\mathcal{X}_s\}$. Using (8) and A3, we can rewrite (18) as

$$\tilde{\pi}_t(x) = \underset{a\in\mathcal{B}(x)}{\operatorname{argmin}} \left|\partial_a\gamma(a,x_v)+c_2(x_v)\,\partial_{x_s}G_t(f(a,x),x_v)\right| . \qquad (19)$$

Now, for any given x, the optimal policy $\tilde{\pi}_t(x)$ can either fall in-between two consecutive breakpoints of γ, or correspond to one of such breakpoints.

In the first case, there exists an $n\in\{1,\dots,B+1\}$ such that $\tilde{\pi}_t(x)\in(b_{n-1}(x_v),b_n(x_v))\cap\mathcal{B}(x)$. In this interval the sub-gradient of γ is equal to $d_n(x_v)$, and (19) takes the form

$$\tilde{\pi}_t(x) = \underset{a\in\mathcal{B}(x)}{\operatorname{argmin}} \left|d_n(x_v)+c_2(x_v)\,\partial_{x_s}G_t(f(a,x),x_v)\right| . \qquad (20)$$

Because of the definition of $\mathcal{B}(x)$ and of $c_2(x_v)\neq 0$, this result can be equivalently expressed as a function of $w_n(x_v)$ defined in (11) by setting $f(\tilde{\pi}_t(x),x)=w_n(x_v)$, which gives

$$\tilde{\pi}_t(x) = \beta_n(x_v) - \frac{c_1(x_v)}{c_2(x_v)}\,x_s\ , \qquad (21)$$

where $\beta_n(x_v)$ was defined in (10). The range of $\tilde{\pi}_t(x)$ implies that (21) is valid for x_s belonging to the open interval

$$\mathcal{I}_n(x_v)\in\frac{c_2(x_v)}{c_1(x_v)}\Big(\beta_n(x_v)-b_n(x_v),\beta_n(x_v)-b_{n-1}(x_v)\Big)\ , \qquad (22)$$

eventually limited to $\mathcal{X}_s$. Note also that gradient monotonicity and $d_n(x_v)>d_{n-1}(x_v)$ assure $w_n(x_v)\leq w_{n-1}(x_v)$. This guarantees that (22) identifies non-overlapping intervals $\mathcal{I}_n$, satisfying $\mathcal{I}_n<\mathcal{I}_{n-1}$, with inequality intended for any couple of elements belonging to the two sets.

In the second case, we have $\tilde{\pi}_t(x)=b_n(x_v)$ for some $n\in\{1,\dots,B\}$. In this point the sub-gradient of γ is given by the interval $[d_n(x_v),d_{n+1}(x_v)]$. Hence, from (19) we have

$$\begin{aligned}&b_n(x_v)\\ &=\underset{a\in\mathcal{B}(x)}{\operatorname{argmin}} \left|[d_n(x_v),d_{n+1}(x_v)]+c_2(x_v)\partial_{x_s}G_t(f(a,x),x_v)\right|.\end{aligned} \qquad (23)$$

Monotonicity, increasing ordering of $d_n(x_v)$, and (11), make (23) equivalent to requiring $f(b_n(x_v),x)\in[w_{n+1}(x_v),w_n(x_v)]$, that is, x_s belongs to the interval

$$\mathcal{J}_n(x_v)=\frac{c_2(x_v)}{c_1(x_v)}\Big[\beta_{n+1}(x_v)-b_n(x_v),\beta_n(x_v)-b_n(x_v)\Big]\ , \qquad (24)$$

eventually limited to $\mathcal{X}_s$. These are again disjoint intervals satisfying $\mathcal{J}_n>\mathcal{J}_{n-1}$. The union of (24) and (22) covers $\mathcal{X}_s$.

To conclude, (12) is just a compact form to express policy $\tilde{\pi}_t(x)$ given by (22) for $x_s\in\mathcal{I}_n(x_v)$, and $\tilde{\pi}_t(x)=b_n(x_v)$ for $x_s\in\mathcal{J}_n(x_v)$. This result can be inferred from Fig. 1. ∎

B. Proof of Theorem 2

The result follows from (11), the equivalence of g_{t+1} and $\tilde{g}_{t+1}$ up to a constant factor, and the fact that $\tilde{g}_{t+1}(x)=G_t(f(\pi_t^*(x),x_s),x_v)+\gamma(\pi_t^*(x),x_v)$. Note that $\tilde{\partial}_{x_s}\pi_t^*(x)$ can only assume three values, namely, 0, $-c_1(x_v)/c_2(x_v)$, and the interval $[-c_1(x_v)/c_2(x_v),0]$ for breakpoints.

References

[1] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Athena Scientific, 2007.

[2] P. Van de Ven, N. Hegde, L. Massoulie, and T. Salonidis, "Optimal control of residential energy storage under price fluctuations," in *IARIA ENERGY 2011*, Venice (I), May 2011, pp. 159–162.

[3] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1732–1743, Sept. 2011.
[4] C. Codemo, T. Erseghe, and A. Zanella, "Energy storage optimization strategies for smart grids," in *IEEE ICC 2013*, June 2013.
[5] M. Goyal, A. Kumar, and V. Sharma, "Power constrained and delay optimal policies for scheduling transmission over a fading channel," in *IEEE INFOCOM 2003*, Apr. 2003, pp. 311–320.