Proyecciones (Antofagasta. On line) | vol. 40, n. 1 (2021) | pp. 117-137.

10.22199/issn.0717-6279-2021-01-0008

PROYECCIONES
Journal of Mathematics

ISSN 0717-6279 (On line)

A Moreau-Yosida regularization for Markov decision processes

R. Israel Ortega-Gutiérrez[1] orcid.org/0000-0001-8247-9673
Hugo Cruz-Suárez[2] orcid.org/0000-0002-0732-4943

Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas, Puebla, México.
[1] rei_israel@yahoo.com.mx ; [2] hcs@fcfm.buap.mx

Received: February 2020 | Accepted: July 2020

Abstract:

This paper addresses a class of sequential optimization problems known as Markov decision processes. These kinds of processes are considered on Euclidean state and action spaces with the total expected discounted cost as the objective function. The main goal of the paper is to provide conditions to guarantee an adequate Moreau-Yosida regularization for Markov decision processes (named the original process). In this way, a new Markov decision process that conforms to the Markov control model of the original process except for the cost function induced via the Moreau-Yosida regularization is established. Compared to the original process, this new discounted Markov decision process has richer properties, such as the differentiability of its optimal value function, strictly convexity of the value function, uniqueness of optimal policy, and the optimal value function and the optimal policy of both processes, are the same. To complement the theory presented, an example is provided.

Keywords: Discounted Markov decision processes; Uniqueness of optimal policies, Moreau-Yosida regularization.

MSC (2020): 90C40, 49M20.

Cite this article as (IEEE citation style):

R. I. Ortega-Gutiérrez and H. Cruz-Suárez, "A Moreau-Yosida regularization for Markov decision processes", *Proyecciones (Antofagasta, On line)*, vol. 40, no. 1, pp. 117-137, 2021, doi: 10.22199/issn.0717-6279-2021-01-0008

1. Introduction

This article addresses discounted Markov decision processes (MDPs) defined on Euclidean state and action spaces (see [12] and [13]). For this kind of MDPs, this document proposes conditions on the components of the Markov control model, which guarantee the application of Moreau-Yosida regularization. This proposal is the first step to apply numerical methods based on Moreau-Yosida regularization, for example, gradient method (see [14] and [16]) or bundle methods (see [2]), in the context of MDPs. These numerical methods should be applied to approximate the optimal policy of MDPs.

The Moreau-Yosida regularization (see [4], [15] and [16]) is a way to smooth a nonsmooth convex function with a minimum such that the smoothing achieves the same minimum as the original function. It is also important to observe that the Moreau-Yosida regularization is a powerful technique to provide differentiability properties to the corresponding perturbed functions (see [4], [8] and [15]). Moreover, it is worth mentioning that this type of regularization has not yet been applied to MDPs.

The main idea of the methodology of the paper is the following: considering a discounted Markov decision process (named the original process or the original model) that satisfies certain assumptions, the Moreau-Yosida regularization is applied to its cost function, allowing the establishment of a new MDP, designated the perturbed MDP. The perturbed MDP has components that are identical to the original Markov control model; the only difference lies in the cost function. Specifically, its cost function is the cost function of the original model with an added quadratic function (in fact, this kind of addition procedure is the core of the Moreau-Yosida regularization).

Therefore, it should be noted that with this type of perturbation, although the optimal value function in the original model is not necessarily differentiable, the differentiability of the optimal value function in the perturbed MDP is guaranteed. Additionally, it is ensured that the optimal value function and the optimal policy of both models are exactly alike.

This work is organized as follows. In Section 2, the Moreau-Yosida regularization is presented. In Section 3, the way to perturb discounted MDPs using the regularization of Moreau-Yosida and the main result are given. Finally, Sections 4 and 5 provide an example and some final remarks.

Notation. Let W be a Euclidean space. For any set $B \subseteq W$, a point $x \in B$ is called an interior point of B if there exists an open set U such that $x \in U \subseteq B$. The interior of B is the set of all interior points of B and is denoted by $int(B)$. If $x \in W$, then $\| x \|$ represents its Euclidean norm, and x^T represents its transpose (considering x as a column vector). For $x, y \in W$, $\langle x, y \rangle$ denotes their (usual) inner product. When $W = \mathbf{R}$, the absolute value of $x \in W$ is denoted by $| x |$.

2. Moreau-Yosida Regularization

The theory presented in this section extends to two variables (denoted as x and a), and some ideas on the Moreau-Yosida regularization approach are given in [16].

Let X and A be nonempty Borel spaces of $\mathbf{R}^l$ and $\mathbf{R}^m$ (l and m are positive integers), respectively. Suppose that $A(x) \subset A$ is a (nonempty) measurable set for all $x \in X$. Define $\mathbf{K} := \{(x, a) \mid x \in X, a \in A(x)\}$. It is assumed that $\mathbf{K}$ is a measurable subset of the product space $X \times A$.

Let $G : \mathbf{K} \to \mathbf{R}$ be a Borel measurable function, and define
$\psi(x) := \inf_{a \in A(x)} G(x, a), \quad x \in X.$

Assumption 2.1. a) *For each $x \in X$, $A(x)$ is a compact and convex set.*

b) *$xA(x)$ is a continuous multifunction.*

c) *G is a continuous function on $\mathbf{K}$.*

d) *For each $x \in X$, $G(x, \cdot)$ is a convex function on $A(x)$.*

e) *There exists a unique measurable selector $f^* : X \to A$ with $f^*(x) \in A(x)$ such that $\psi(x) = G(x, f^*(x))$ for each $x \in X$.*

Remark 2.2. *Observe that under Assumptions 2.1 a) and c), there exists a measurable selector f^* such that $\psi(x) = G(x, f^*(x))$ for each $x \in X$ (see Proposition D5, p. 182 in [12]); in fact, in Assumption 2.1 e), the uniqueness of this selector is required.*

In all the sequel, it will be assumed that $\lambda > 0$ is fixed. Consider, for each $(x, a) \in \mathbf{K}$ and $b \in A(x)$, the following function: $\overline{H}(x, a, b) := G(x, b) + \frac{\lambda}{2} \| b - a \|^2$.

Definition 2.3. *For $x \in X$ and $a \in A(x)$, define the following functions:*
$H_x(a) := \min_{b\in A(x)} \left\{\overline{H}(x,a,b)\right\}$,
$\widehat{P}(x,a) := arg\min_{b\in A(x)} \left\{\overline{H}(x,a,b)\right\}$, *where H_x and $\widehat{P}$ are called the regularization of Moreau-Yosida of $\overline{H}$ and the proximal operator associated with $\overline{H}$, respectively.*

Remark 2.4. a) *Note that $H_x(a) \leq G(x,a) < +\infty$ for all $(x,a) \in \mathbf{K}$.*

b) *Observe that under Assumption 2.1, since for each $(x,a) \in \mathbf{K}$, $\overline{H}(x,a,\cdot)$ is the sum of a convex function and a strictly convex one, $\overline{H}(x,a,\cdot)$ is strictly convex. Consequently, the proximal operator $\widehat{P}(x,a)$ is nonempty for each $(x,a) \in \mathbf{K}$. In addition, observe that $\overline{H}(x,a,\cdot)$ is a continuous function for each $(x,a) \in \mathbf{K}$.*

Lemma 2.5. *Suppose that Assumption 2.1 holds. Then, $\widehat{P}(x,a)$ is a unit set, that is, $\widehat{P}(x,a) = \{P(x,a)\}$, with $P : \mathbf{K} \to A$, and the function $\overline{H}(x,a,\cdot)$ has a unique minimum in $A(x)$ for each $(x,a) \in \mathbf{K}$. In particular, $\widehat{P}(x,f^*(x)) = \{P(x,f^*(x))\} = \{f^*(x)\}$ for each $x \in X$; i.e., $f^*(x)$ is the unique minimum for the function $\overline{H}(x,f^*(x),\cdot)$ for each $x \in X$. Furthermore, $G(x,f^*(x)) = H_x(f^*(x))$ for each $x \in X$.*

Proof. Let $x \in X$ be fixed. $\overline{H}(x,a,\cdot)$ is a strictly convex function (see Remark 2.4 a)); thus, for each $a \in A(x)$, using Theorem 2.6, p. 41 in [16], $\overline{H}(x,a,\cdot)$ has a unique minimum in $A(x)$ for each $a \in A(x)$. Hence, $\widehat{P}(x,a)$ is a unit set, let us say, $\widehat{P}(x,a) = \{P(x,a)\}$. In particular, $\overline{H}(x,f^*(x),\cdot)$ has a unique minimum in $A(x)$. Now, observe that

$$\begin{aligned} G(x,f^*(x)) &\leq G(x,P(x,f^*(x))) \\ &\leq G(x,P(x,f^*(x))) + \frac{\lambda}{2} \parallel P(x,f^*(x)) - f^*(x) \parallel^2 \\ &= \min_{b\in A(x)} \left\{\overline{H}(x,f^*(x),b)\right\} = H_x(f^*(x)). \end{aligned} \tag{2.1}$$

Using Remark 2.4 a), it is obtained that

$$H_x(f^*(x)) \leq G(x,f^*(x)). \tag{2.2}$$

Then, (1) and (2) imply that $G(x,f^*(x)) = H_x(f^*(x))$ and that $\widehat{P}(x,f^*(x)) = \{f^*(x)\}$. Since x is arbitrary, the proof of Lemma 2.5 is finished. □

Lemma 2.6. *Suppose that Assumption 2.1 holds. Then, for each $x \in X$, both functions $H_x(\cdot)$ and $P(x,\cdot)$ are continuous.*

Proof. Let $x \in X$ be fixed. Under Assumption 2.1, the hypothesis of Theorem 17.31, p. 570 in [1] holds. Consequently, the following statements are valid: (i) H_x is a continuous function; (ii) $\widehat{P}(x,a)$ is a nonempty compact set, in fact, $\widehat{P}(x,a) = \{P(x,a)\}$ for each $a \in A(x)$; and (iii) the multifunction $a\widehat{P}(x,a)$ is upper semicontinuous because A is trivially Hausdorff. Then, for each closed set $F \subset A$, $\{a \in A(x) \mid P(x,a) \in F\} = \left\{a \in A(x) \mid \widehat{P}(x,a) \cap F \neq \emptyset\right\}$ is closed in A. Hence, $P(x,\cdot)$ is a continuous function (see Theorem 8.3, p. 79 in [7]). Since x is arbitrary, the proof of Lemma 2.5 is concluded. □

Lemma 2.7. *Suppose that Assumption 2.1 is fulfilled. Then, the gradient of the function $H_x(\cdot)$, denoted by $\nabla H_x(\cdot)$, is given by $\nabla H_x(a) = \lambda(a - P(x,a))$ for all $a \in int(A(x))$ and $x \in X$.*

Proof. Let $x \in X$, $a \in int(A(x))$ and a direction $d \in \mathbf{R}^m$ with $\| d \| = 1$ be fixed. Consider $t \in \mathbf{R}^+ - \{0\}$; then, using the definition of H_x (see Definition 2.3), it follows that

$$\begin{aligned}\frac{H_x(a+td) - H_x(a)}{t} &= \frac{1}{t}\left[\min_{b\in A(x)}\left\{G(x,b) + \frac{\lambda}{2} \| b - a - td \|^2\right\}\right. \\ &- \left.\min_{\omega\in A(x)}\left\{G(x,\omega) + \frac{\lambda}{2} \| \omega - a \|^2\right\}\right].\end{aligned}$$

Now, since $H_x(a) \leq G(x, P(x,a+td)) + \frac{\lambda}{2} \| P(x,a+td) - a \|^2$, it is obtained that

$$\begin{aligned}\frac{H_x(a+td) - H_x(a)}{t} &\geq \frac{1}{t}\left[G(x,P(x,a+td)) + \frac{\lambda}{2} \| P(x,a+td) - a - td \|^2\right. \\ &- \left. G(x,P(x,a+td)) - \frac{\lambda}{2} \| P(x,a+td) - a \|^2\right].\end{aligned}$$

Equivalently,

$$\begin{aligned}\frac{H_x(a+td)-H_x(a)}{t} &\geq \frac{\lambda}{2t}\left[\| P(x,a+td)-a-td\|^2 - \| P(x,a+td)-a\|^2\right]\\ &= \frac{\lambda}{2t}\left[\| P(x,a+td)-P(x,a)+P(x,a)-a-td\|^2\right.\\ &- \left.\| P(x,a+td)-P(x,a)+P(x,a)-a\|^2\right]. \qquad (2.3)\end{aligned}$$

By simple computations and the polarization identity (see [9]), it follows that

$$\frac{\lambda}{2t}\left[\| \rho+\eta\|^2 - \| \rho+\varphi\|^2\right] = \frac{\lambda}{2t}\left[\| \eta\|^2 - \| \varphi\|^2\right] - \lambda\langle \rho, d\rangle, \qquad (2.4)$$

where $\rho := P(x,a+td)-P(x,a)$, $\eta := P(x,a)-a-td$ and $\varphi := P(x,a)-a$. Substituting (4) in (3), it follows that

$$\begin{aligned}\frac{H_x(a+td)-H_x(a)}{t} &\geq \frac{\lambda}{2t}\left[\| P(x,a)-a-td\|^2 - \| P(x,a)-a\|^2\right]\\ &- \lambda\langle P(x,a+td)-P(x,a), d\rangle\\ &= \frac{\lambda}{2t}\left[\| P(x,a)-a\|^2 - 2t\langle P(x,a)-a, d\rangle + t^2\| d\|^2\right.\\ &- \left.\| P(x,a)-a\|^2\right]\\ &- \lambda\langle P(x,a+td)-P(x,a), d\rangle\\ &= \lambda\langle a-P(x,a), d\rangle + \frac{\lambda}{2}t - \lambda\langle P(x,a+td)-P(x,a), d\rangle. \qquad (2.5)\end{aligned}$$

Now, since $P(x,\cdot)$ is a continuous function (see Lemma 2.5), taking $\liminf$ when t goes to 0 yields that $P(x,a+td)\to P(x,a)$. Then, from inequality (5), it follows that

$$\liminf_{t\to 0}\frac{H_x(a+td)-H_x(a)}{t} \geq \lambda\langle a-P(x,a), d\rangle. \qquad (2.6)$$

On the other hand, since $H_x(a+td)\leq G(x,P(x,a)) + \frac{\lambda}{2}\| P(x,a)-a-td\|^2$, it follows that

$$\begin{aligned}\frac{H_x(a+td)-H_x(a)}{t} &= \frac{1}{t}\left[\min_{b\in A(x)}\left\{G(x,b)+\frac{\lambda}{2}\| b-a-td\|^2\right\}\right.\\ &- \left.\min_{\omega\in A(x)}\left\{G(x,\omega)+\frac{\lambda}{2}\| \omega-a\|^2\right\}\right]\end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{t}\left[G(x,P(x,a)) + \frac{\lambda}{2} \parallel P(x,a) - a - td \parallel^2 \right. \\ &- \left. G(x,P(x,a)) - \frac{\lambda}{2} \parallel P(x,a) - a \parallel^2 \right] \\ &= \frac{\lambda}{2t}\left[\parallel P(x,a) - a - td \parallel^2 - \parallel P(x,a) - a \parallel^2 \right] \\ &= \lambda\langle a - P(x,a), d\rangle + \frac{\lambda}{2}t - \lambda\langle P(x,a+td) - P(x,a), d\rangle. \end{aligned}$$

Then, in the last inequality, taking lim sup as $t \to 0$ implies that

$$\limsup_{t\to 0} \frac{H_x(a+td) - H_x(a)}{t} \leq \lambda\langle a - P(x,a), d\rangle. \tag{2.7}$$

Then, from inequalities (6) and (7), it is obtained that

$$\begin{aligned} \lambda\langle a - P(x,a), d\rangle &\leq \liminf_{t\to 0} \frac{H_x(a+td) - H_x(a)}{t} \\ &\leq \limsup_{t\to 0} \frac{H_x(a+td) - H_x(a)}{t} \\ &\leq \lambda\langle a - P(x,a), d\rangle. \end{aligned}$$

Therefore, $\nabla H_x(a) = \lambda(a - P(x,a))$. Since a, d and x are arbitrary, result follows. □

A direct consequence of Lemma 2.7 is the following result.

Corollary 2.8. *Suppose that Assumption 2.1 holds and the partial derivative with respect to a of the function $P(x,a)$, denoted by $D(P(x,a))$, exists for $x \in X$ and $a \in int(A)$. Then, the Hessian matrix $\mathcal{H}(H_x(\cdot))$ is given by*

$H\,(H_x(a)) = \lambda(I_{m\times m} - D(P(x,a)))$, for each $a \in int(A(x))$ with $x \in X$. Here, $I_{m\times m}$ is the identity matrix of order m, and $D(P(x,a))$ is the matrix with entries $\dfrac{\partial P^i(x,a)}{\partial a^j}$, i.e., $D(P(x,a)) := \left[\dfrac{\partial P^i(x,a)}{\partial a^j}\right]_{i,j}^{n,m}$ with $P(x,a) = (P^1(x,a), P^2(x,a), \ldots, P^m(x,a))$ and $a = (a^1, a^2, \ldots, a^m)$.

3. The Moreau-Yosida Regularization Applied to MDPs

3.1. Discounted Markov Decision Processes

Now, the preliminary Markov decision process context is briefly presented (see [12]). Let $(X, A, \{A(x) \mid x \in X\}, Q, c)$ be a *Markov control model* that

consists of the *state space* X, the *control* or *action space* A, the *admissible sets* $A(x)$, $x \in X$, the *transition law* Q, and the *cost-per-stage function* c. The sets X and A are assumed to be Borel spaces of $\mathbf{R}^l$ and $\mathbf{R}^m$, respectively. For each $x \in X$, $A(x)$ is a nonempty measurable subset of A, and $A(x)$ denotes the set of *feasible actions* in the state $x \in X$. Define $\mathbf{K} := \{(x, a) \mid x \in X, a \in A(x)\}$, which is assumed to be a measurable subset of $X \times A$. The transition law Q is a stochastic kernel on X given $\mathbf{K}$, and the cost-per-stage function $c : \mathbf{K} \to \mathbf{R}$ is measurable.

Definition 3.1. *For each $t = 0, 1, \ldots$, define the space $\mathbf{H}_t$ of admissible histories up to time t as $\mathbf{H}_0 = X$, and $\mathbf{H}_t = \mathbf{K} \times \mathbf{H}_{t-1}$, for $t = 1, 2, \ldots$. A policy is defined as a sequence $\pi = \{\pi_t, t = 0, 1, \ldots\}$ of stochastic kernels, defined on A given $\mathbf{H}_t$ (see [13]). In this paper, the set of policies will be denoted by Π. In particular, a stationary policy is of the form $\pi = \{f, f, \ldots\}$, where f is defined as a measurable function $f : X \to A$ such that $f(x) \in A(x)$ for all $x \in X$. The set of all stationary policies will be denoted by $\mathbf{F}$.*

For each $\pi \in \Pi$ and an initial state $x \in X$, let

$\mathrm{V}(\pi, x) = E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right]$,be the *total expected discounted cost*, when the policy π is applied, given the initial state x. The constant $\alpha \in (0, 1)$ is a given *discount factor* fixed. The sequence of consecutive states and corresponding actions will be denoted by $\{x_t\}$ and $\{a_t\}$, respectively, and E_x^π denotes the expected value with respect to the probability measure P_x^π, which is defined on a canonical measurable space (Ω, F) induced by the Ionescu-Tulcea Theorem (see [12]).

Definition 3.2. *A policy π^* is optimal if $V(\pi^*, x) = \upsilon^*(x)$ for all $x \in X$, where $\upsilon^*(x) := \pi \in \Pi \mathrm{inf} V(\pi, x), \quad x \in X$.*

υ^* is called the optimal value function.

Definition 3.3. *Let $\varpi : X \to [1, +\infty)$ be a measurable function that will be referred to as a weight function. If u is a real-valued function on X, define its ϖ-norm as*

$\| u \|_\varpi := \sup_{x \in X} \dfrac{| u(x) |}{\varpi(x)}$. *Let $\mathbf{B}_\varpi(X)$ be the normed linear space of ϖ-bounded measurable functions u on X.*

Assumption 3.4. a) *For each $x \in X$, $A(x)$ is a compact set.*

b) *The transition law Q is strongly continuous, i.e.,*

$$\theta(x,a) := \int_X u(y)Q(dy|x,a), \text{ for } (x,a) \in \mathbf{K},$$

is a continuous function and bounded on $\mathbf{K}$ for each $u \in \mathbf{B}(X)$, which denotes the Banach space of real-valued bounded measurable functions u on X with the supremum norm $\| u \| := \sup_{x\in X} | u(x) |$.

c) *c is a continuous function on $\mathbf{K}$.*

d) *There exist nonnegative constants $\hat{k}$ and δ, with $1 \leq \delta < \frac{1}{\alpha}$ and a weight function $\varpi : X \to [1, +\infty)$, such that*
$\sup_{a\in A(x)} | c(x,a) | \leq \hat{k}\varpi(x)$, *and*
$\sup_{a\in A(x)} \int \varpi(y)Q(dy|x,a) \leq \delta\varpi(x)$, *for all $x \in X$.*

e) *For every state $x \in X$, the function $\varpi'(x,a) := \int_X \varpi(y)Q(dy \mid x,a)$ is a continuous function in $a \in A(x)$.*

Definition 3.5. *The value iteration functions are defined as:*

$$V_n(x) = \min_{a\in A(x)} \left\{ c(x,a) + \alpha \int_X V_{n-1}(y)Q(dy|x,a) \right\}, \tag{3.1}$$

for all $x \in X$ and $n = 1, 2, \ldots$, with $V_0(\cdot) = 0$.

The proof of the following lemma can be found in Theorem 8.3.6, p. 47 in [13].

Lemma 3.6. *Suppose that Assumption 3.4 is fulfilled. Then,*

a) *The optimal value function $v^* \in \mathbf{B}_\varpi(X)$ is a solution of*

$$v^*(x) = \min_{a\in A(x)} \left[c(x,a) + \alpha \int_X v^*(y)Q(\,dy|\,x,a) \right] \tag{3.2}$$

for each $x \in X$.

b) *There exists a selector $f^* \in \mathbf{F}$ such that, in (9), the minimum is attained, i.e.,*

$$v^*(x) = c(x, f^*(x)) + \alpha \int_X v^*(y)Q(\,dy|\,x, f^*(x)), \tag{3.3}$$

for all $x \in X$, and f^* is optimal.

c) $V_n \in \mathbf{B}_\varpi(X)$, $n = 1, 2, \ldots$, *and* $\{V_n\}$ *converges pointwise to* υ^*. *Moreover, for each* $n = 1, 2, \ldots$ *there exists* $f_n \in \mathbf{F}$ *such that*

$V_n(x) = c(x, f_n(x)) + \alpha \int_X V_{n-1}(y) Q(dy|x, f_n(x))$, *for all* $x \in X$, *where* f_n *is called the minimizer of the value iteration functions.*

3.2. Application to MDPs

Let $M = (X, A, \{A(x)|\, x \in X\}, Q, c)$ be a fixed Markov control model. M will be referred to as the *original model*, and it will be assumed that M satisfies Assumption 3.4. Optimal value function, optimal policy, value iteration function and minimizers of the value iteration functions will be denoted for υ^*, f^*, V_n, and f_n, $n = 0, 1, 2, \ldots$, respectively.

Now, let us define the following MDP with the Markov control model given by $M_\lambda := (X, A, \{A(x)|x \in X\}, Q, c_\lambda)$, where $c_\lambda(x, a) := c(x, a) + \lambda/2 \parallel a - f^*(x) \parallel^2$, $(x, a) \in \mathbf{K}$, f^* is a fixed optimal policy and c is the cost function of the original model M. Note that both MDPs; M and M_λ, are equal except for the cost function. Moreover, observe that the set $\mathbf{F}$ of stationary policies is the same for both models (in fact, the set Π of the all admissible policies is the same for both models). The MDP M_λ will be referred to as the *perturbed model*.

In model M_λ, the total expected discounted cost, when the policy $\pi \in \Pi$ is applied and the initial state is $x \in X$, is defined by

$$\mathrm{W}(\pi, x) = E_x^\pi \left[\sum_{t=0}^{\infty} \alpha^t c_\lambda(x_t, a_t) \right],$$

and the corresponding optimal value function is given by

$$\omega^*(x) = \inf_{\pi \in \Pi} W(\pi, x), \quad x \in X.$$

Remark 3.7. *Under Assumption 3.4, observe that for each* $x \in X$, $W(\pi, x) < +\infty$ *when* $\pi = \{f^*, f^*, \ldots\}$, *where* f^* *is the optimal policy from original model* M.

Notation 3.8. a) *Define the functions* $G(x, a) := c(x, a) + \alpha \int_X \upsilon^*(y) Q(dy|x, a)$ *and*

$G_n(x, a) := c(x, a) + \alpha \int_X V_{n-1}(y) Q(dy|x, a)$, $n = 1, 2, \ldots$, *for each* $(x, a) \in \mathbf{K}$. *Consequently,* $\overline{H}(x, f^*(x), b) = G(x, b) + \frac{\lambda}{2} \parallel b - f^*(x) \parallel^2$ *for each* $x \in X$ *and* $b \in A(x)$.

b) *Evaluating* f^* *in* H_x, $\widehat{P}$ *and* H_x^n, *which were given in Definition 2.3, yields the following*

$$H_x(f^*(x)) = \min_{b \in A(x)} \left\{ c(x, b) + \alpha \int_X \upsilon^*(y) Q(dy|x, b) + \tfrac{\lambda}{2} \parallel b - f^*(x) \parallel^2 \right\},$$

$\widehat{P}(x, f^*(x)) = arg\min_{b\in A(x)} \left\{c(x,b) + \alpha \int_X v^*(y)Q(dy|x,b) + \frac{\lambda}{2} \parallel b - f^*(x) \parallel^2\right\}$,
for each $x \in X$. *Moreover, for each* $n = 1, 2, \ldots$ *and* $x \in X$,
$H_x^n(f^*(x)) = \min_{b\in A(x)} \left\{c(x,b) + \alpha \int_X V_{n-1}(y)Q(dy|x,b) + \frac{\lambda}{2} \parallel b - f^*(x) \parallel^2\right\}$.

It is important to mention that in the rest of the paper, we will work in the context of discounted MDPs described in Subsections 3.1 and 3.2, and we only deal with functions G and G_n defined in Notation 3.8 and with the Moreau-Yosida regularization of $\overline{H}$, the proximal operator of $\overline{H}$ and $H_x^n(\cdot)$ defined by them.

3.3. Convexity and Differentiability in the Perturbed Model

In this section, under convexity assumptions on the components of the Markov control model, it is obtained a version of the Moreau-Yosida regularization for MDPs, see Theorem 3.13.

Assumption 3.9. a) *$A(x)$ is a convex set for each $x \in X$, and the multifunction $xA(x)$ is continuous.*

b) *Assume that $G(x,\cdot)$ and $G_n(x,\cdot)$ are convex functions for each $x \in X$ and $n = 1, 2, \ldots$.*

c) *The integrals:*

$$\int V_n(y)Q(dy \mid \cdot, \cdot), \quad n = 1, 2, \ldots, \tag{3.4}$$

and

$$\int v^*(y)Q(dy \mid \cdot, \cdot) \tag{3.5}$$

are finite and continuous functions on $\mathbf{K}$.

Remark 3.10. a) *There are conditions in [6] (see Assumption 1 and either Condition 1 (C1) or Condition 2 (C2) in [6]) that, under suitable and light modifications of conditions C1 or C2, obtain the result that $G_n(x,\cdot)$ and $G(x,\cdot)$ are convex functions for each $x \in X$ (see also the proof of Lemma 6.2 in the same reference [6]).*

b) *For bounded models (i.e., MDPs with bounded cost functions and compact admissible action sets), the continuity of integrals (11) and (12) follows directly from the strong continuity of the transition law Q. If, moreover, the multifunction $xA(x)$ is continuous, then the continuity of V_n, $n = 1, 2, \ldots$, and v^* is an immediate consequence of Proposition D.3(c) p. 130 in [11] using the continuity of the cost function c and of the integrals (11) and (12), and equations (8) and (9).*

The proof of the following theorem is direct.

Theorem 3.11. *Suppose that Assumptions 3.4 and 3.9 hold. Then, $\overline{H}(x, f^*(x), \cdot)$ is a strictly convex function for each $x \in X$.*

Theorem 3.12. *Suppose that Assumption 3.9 holds. Then, $P(x, f^*(x)) = f^*(x)$ for each $x \in X$, i.e., f^* is the unique optimal policy of the perturbed model M_λ. Moreover, $G(x, f^*(x)) = H_x(f^*(x))$ for each $x \in X$.*

Proof. Let $x \in X$ be fixed. Lemma 2.5 implies that $\widehat{P}(x, f^*(x)) = \{f^*(x)\}$, i.e., $P(x, f^*(x)) = f^*(x)$, and that $G(x, f^*(x)) = H_x(f^*(x))$. It only remains to prove that f^* is the optimal policy for model M_λ. Observe that $W(f^*, x) = V(f^*, x) = v^*(x)$. It is easy to verify that for each $\pi \in \Pi$, $W(\pi, x) \geq V(\pi, x) \geq v^*(x)$. Then, it follows that for each $\pi \in \Pi$, $W(\pi, x) \geq W(f^*, x)$. Therefore, $w^*(\cdot) = v^*(\cdot) = W(f^*, \cdot)$, and f^* is optimal for W. Now, uniqueness of f^* will be proven by contradiction. In this way, it is assumed that $h \in \mathbf{F}$ is an optimal policy for W, and it is assumed, without loss of generality, that function h is not equal to f in the fixed state x. From (1), noting that $w^*(\cdot) = v^*(\cdot)$, it is obtained that

$$\begin{aligned} w^*(x) &= G(x, f^*(x)) \\ &< G(x, h(x)) + \frac{\lambda}{2} \parallel f^*(x) - h(x) \parallel^2 \\ &= c(x, h(x)) + \frac{\lambda}{2} \parallel f^*(x) - h(x) \parallel^2 + \alpha \int_X v^*(y) Q(dy|x, h(x)) \\ &= c(x, h(x)) + \frac{\lambda}{2} \parallel f^*(x) - h(x) \parallel^2 + \alpha \int_X w^*(y) Q(dy|x, h(x)). \end{aligned}$$

Consequently, $w^*(x) < c_\lambda(x, h(x)) + \alpha \int_X w^*(y) Q(dy|x, h(x))$, i.e., $h(x)$ does not satisfy equation (10); which is a contradiction of the optimality of h. Therefore, f^* is the unique optimal policy of the perturbed model M_λ. Since x is arbitrary, result follows. □

Theorem 3.13. *Suppose that Assumptions 3.4 and 3.9 are fulfilled. If* $a \in int(A(x))$, *then* $\nabla H_x(a) = \lambda(a - P(x,a))$. *In particular, if* $f^*(x) \in int(A(x))$, *it is obtained that* $\nabla H_x(f^*(x)) = \mathbf{0}$, *for each* $x \in X$, *where* $\mathbf{0}$ *represents the zero vector in* $\mathbf{R}^m$.

Proof. Let $x \in X$ be fixed. From Lemma 2.7, it is obtained that $\nabla H_x(a) = \lambda(a - P(x,a))$ for each $a \in int(A(x))$. In particular, $\nabla H_x(f^*(x)) = \lambda(f^*(x) - f^*(x)) = \mathbf{0}$. Hence, the arbitrariness of x implies Theorem 3.13. □

4. Example

In this section, one example will be presented to illustrate the theory developed in the previous sections. It is also relevant to note that in Example 4.1, validity of Assumptions 3.4 and 3.9 is given in detail.

Example 4.1. *Let* $X = A = \mathbf{R}^l$ *and* $A(\mathbf{x}) = [- \mid x_1 \mid, \mid x_1 \mid] \times [- \mid x_2 \mid, \mid x_2 \mid] \times \cdots \times [- \mid x_l \mid, \mid x_l \mid]$ *for all* $\mathbf{x}^T = (x_1, x_2, \ldots, x_l) \in X$. *The transition law is given by*

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{a}_t + \xi_t, \quad \text{for each} \quad t = 0, 1, \ldots, \quad \text{with} \quad \mathbf{x}_0 = \mathbf{x} \in X,$$

where $\{\xi_t\}$ *is a sequence of i.i.d column random vectors with values in* $S = \mathbf{R}^l$. *Let* ξ *be a generic element of the sequence* $\{\xi_t\}$. *Assume that* ξ *has a standard multivariate normal distribution with density* Δ. *Furthermore, it is assumed that* $\alpha \in (0, 1/4)$. *The cost function is given by*

$$c(\mathbf{x},\mathbf{a}) = \| \mathbf{x} \|^2 + \| \mathbf{a} \|^2, \quad (\mathbf{x}, \mathbf{a}) \in \mathbf{K}.$$

Remark 4.2. *The standard linear-quadratic problem is unconstrained, that is,* $A(\mathbf{x}) \equiv A = \mathbf{R}^l$ *(see [3]). The selection of the compact control sets* $A(\mathbf{x}) = [- \mid x_1 \mid, \mid x_1 \mid] \times [- \mid x_2 \mid, \mid x_2 \mid] \times \cdots \times [- \mid x_l \mid, \mid x_l \mid]$, $\mathbf{x} \in X$, *was made in order to satisfy Assumption 3.4 a) and, in fact, it is directly seen that* $f_n(\mathbf{x})$, $f^*(\mathbf{x}) \in int\,(A(\mathbf{x}))$ *for all* $\mathbf{x} \in X$ *and* $n \geq 1$ *(see Lemma 4.5, below).*

Lemma 4.3. *Example 4.1 satisfies Assumption 3.4.*

Proof. Assumption 3.4 is verified in the next steps:

a) Observe that $A(\mathbf{x})$ is clearly a compact set for all $\mathbf{x} \in X$.

b) Let $u \in \mathbf{B}(X)$, then for each $(\mathbf{x}, \mathbf{a}) \in \mathbf{K}$:

$$\begin{aligned} \theta(\mathbf{x}, \mathbf{a}) &:= \int_X u(\mathbf{y}) Q(dy \mid \mathbf{x}, \mathbf{a}) \\ &= \int_{\mathbf{R}^l} u(\mathbf{x} + \mathbf{a} + \mathbf{s}) \Delta(\mathbf{s}) d\mathbf{s}. \end{aligned}$$

Since density Δ is a continuous and bounded function, then, as a consequence of the dominated convergence theorem (see [17]), transition law Q is strongly continuous.

c) Clearly, cost function c is a continuous function on $\mathbf{K}$.

d) It is easily verified that $\hat{k} = 1$, $\delta = 4$ and $\varpi(\mathbf{x}) = 3 \parallel \mathbf{x} \parallel^2 +1$, $\mathbf{x} \in X$, satisfy Assumption 3.4 d).

e) Taking $\varpi(\cdot)$ as in the previous step d), it is obtained that for each $(\mathbf{x}, \mathbf{a}) \in \mathbf{K}$,

$$\begin{aligned} \varpi'(\mathbf{x}, \mathbf{a}) &= \int_X \varpi(\mathbf{y}) Q(d\mathbf{y} \mid \mathbf{x}, \mathbf{a}) \\ &= \int_{\mathbf{R}^l} \varpi(\mathbf{x} + \mathbf{a} + \mathbf{s}) \Delta(\mathbf{s}) d\mathbf{s} \\ &= \int_{\mathbf{R}^l} \left[3 \parallel \mathbf{x} + \mathbf{a} + \mathbf{s} \parallel^2 +1 \right] \Delta(\mathbf{s}) d\mathbf{s}. \end{aligned}$$

It implies that $\varpi'(\mathbf{x}, \mathbf{a}) = 3 \parallel \mathbf{x} + \mathbf{a} \parallel^2 +4$, $(\mathbf{x}, \mathbf{a}) \in \mathbf{K}$. Therefore, $\varpi'(\mathbf{x}, \cdot)$ is a continuous function on $A(\mathbf{x})$ for each $\mathbf{x} \in X$.

□

Lemma 4.4. *Example 4.1 satisfies Assumption 3.9.*

Proof. Assumption 3.9 will be verified in the following steps.

a) It will be proven that the multifunction $\mathbf{x}A(\mathbf{x})$ is lower semicontinuous (l.s.c.). To this end, assume that $\{\mathbf{x}_n = (x_{1,n}, x_{2,n}, \ldots, x_{l,n})\} \subset X$ is a sequence convergent to $\mathbf{x} = (x_1, x_2, \ldots, x_l) \in X$, and let $\mathbf{a} \in A(\mathbf{x})$. The proof will proceed by cases. For the first case, suppose that $\mathbf{a}$ is in the boundary of $A(\mathbf{x})$, i.e., $\mathbf{a} = (a_1, a_2, \ldots, x_i, \ldots, a_l)$ for some $i \in \{1, 2, \ldots, l\}$ with $a_j \in [- \mid x_j \mid, \mid x_j \mid]$ for all $j \neq i$, and $j = 1, 2, \ldots, l$. In this case, consider $\mathbf{a}_n = (a_1, a_2, \ldots, x_{i,n}, \ldots, a_l)$. Then, observe

that $\mathbf{a}_n \to \mathbf{a}$ as $n \to \infty$. For the second case, consider $\mathbf{a} \in int(A(\mathbf{x}))$. Hence, there exists $\beta_i \in (0,1)$ such that $a_i = -\beta_i \mid x_i \mid +(1-\beta_i) \mid x_i \mid$ for all $i = 1, 2, \ldots, l$. Taking $a_{i,n} = -\beta_{i,n} \mid x_{i,n} \mid +(1-\beta_{i,n}) \mid x_{i,n} \mid$, then $a_{i,n} \to a_i$ as $n \to \infty$ for all $i = 1, 2, \ldots, l$; in other words, $\mathbf{a}_n \to \mathbf{a}$ when n goes to infinity. According to the previous cases, it is concluded that the multifunction $\mathbf{x}A(\mathbf{x})$ is l.s.c. Now, it will be proven that the multifunction is upper semicontinuous (u.s.c.). Let $\mathbf{x}_n \to \mathbf{x} \in X$ and $\mathbf{a}_n \in A(\mathbf{x}_n)$. Then, $- \mid x_{i,n} \mid \leq a_{i,n} \leq \mid x_{i,n} \mid$ for all $i = 1, 2, \ldots, l$. Taking the limit as $n \to \infty$, it is obtained that $- \mid x_i \mid \leq \liminf a_{i,n} \leq \limsup a_{i,n} \leq \mid x_i \mid$ for all $i = 1, 2, \ldots, l$. Then, sequence $\{a_n\}$ has a limit point in $A(\mathbf{x})$ defined by $\liminf_{n\to\infty} a_{i,n}$ or by $\limsup_{n\to\infty} a_{i,n}$. Therefore, multifunction $\mathbf{x}A(\mathbf{x})$ is u.s.c. and hence continuous.

b) Clearly, this example satisfies Assumption 1 and C2 in [6]. Then, G and G_n are convex functions on $A(\mathbf{x})$ for all $\mathbf{x} \in X$, and the optimal policy f^* is unique.

c) Now, the finiteness and the continuity of $\int v^*(\mathbf{y})Q(d\mathbf{y} \mid \cdot, \cdot)$ will be verified. Taking $\pi = \{f, f, \ldots\}$ with $f(\mathbf{x}) = -\mathbf{x}$, $\mathbf{x} \in X$, it is obtained that

$$V(f, \mathbf{x}) = \| \mathbf{x} \|^2 + \frac{2\alpha}{1-\alpha} < +\infty, \qquad \mathbf{x} \in X. \tag{4.1}$$

Consequently, by (13), it follows that

$$0 \leq v^*(\mathbf{x}) \leq \| \mathbf{x} \|^2 + \frac{2\alpha}{1-\alpha} \tag{4.2}$$

$\mathbf{x} \in X$. Then, for each $(\mathbf{x}, \mathbf{a}) \in \mathbf{K}$,

$$\begin{aligned} \int v^*(\mathbf{y})Q(d\mathbf{y} \mid \mathbf{x}, \mathbf{a}) &= \int I_X(\mathbf{s}) v^*(\mathbf{x}+\mathbf{a}+\mathbf{s})\Delta(\mathbf{s})d\mathbf{s} \\ &\leq \int I_X(\mathbf{s}) \left[\| \mathbf{x}+\mathbf{a}+\mathbf{s} \|^2 + \frac{2\alpha}{1-\alpha} \right] \Delta(\mathbf{s})d\mathbf{s} \\ &= \| \mathbf{x}+\mathbf{a} \|^2 + \frac{1+\alpha}{1-\alpha} < +\infty. \end{aligned}$$

Consider $\{(\mathbf{x}_n, \mathbf{a}_n)\}$ a sequence in $\mathbf{K}$ such that $(\mathbf{x}_n, \mathbf{a}_n) \to (\mathbf{x}, \mathbf{a}) \in \mathbf{K}$ when n goes to infinity. Let $\widehat{L}$ be a positive number such that for each $n = 1, 2, \ldots$

$$0 \leq \| \mathbf{x}_n \|^2 \leq \widehat{L} \quad \text{and} \quad 0 \leq \| \mathbf{a}_n \|^2 \leq \widehat{L}. \tag{4.3}$$

From (14), for each $n = 1, 2, \ldots$ and $\mathbf{s} \in S$, it is obtained that $0 \leq \upsilon^*(\mathbf{s})\Delta(\mathbf{s} - \mathbf{x}_n - \mathbf{a}_n) \leq h_n(\mathbf{s})$, where $h_n(\mathbf{s}) = \left[\| \mathbf{s} \|^2 + \frac{2\alpha}{1-\alpha} \right] \Delta(\mathbf{s} - \mathbf{x}_n - \mathbf{a}_n)$, $\mathbf{s} \in S$. Then, from (15), it yields that

$$\begin{aligned} 0 \leq \int_X h_n(\mathbf{s}) d\mathbf{s} &= \int_X \left[\| \mathbf{s} \|^2 + \frac{2\alpha}{1-\alpha} \right] \Delta(\mathbf{s} - \mathbf{x}_n - \mathbf{a}_n) d\mathbf{s} \\ &= \int_X \left[\| \mathbf{x}_n + \mathbf{a}_n + \mathbf{s} \|^2 + \frac{2\alpha}{1-\alpha} \right] \Delta(\mathbf{s}) d\mathbf{s} \\ &= \| \mathbf{x}_n + \mathbf{a}_n \|^2 + \frac{1+\alpha}{1-\alpha} \\ &\leq 4\widehat{L} + \frac{1+\alpha}{1-\alpha} < +\infty, \end{aligned}$$

for each $n = 1, 2, \ldots$. Furthermore, observe that $\{h_n\}$ converge pointwise to the function

$$h(\mathbf{s}) = \left[\| \mathbf{s} \|^2 + \frac{2\alpha}{1-\alpha} \right] \Delta(\mathbf{s} - \mathbf{x} - \mathbf{a}),$$

and

$$0 \leq h_n(\mathbf{s}) \leq e^{-2\widehat{L}} \left[\| \mathbf{s} \|^2 + \frac{2\alpha}{1-\alpha} \right] \Delta(\mathbf{s}), \quad \mathbf{s} \in S, \quad n = 1, 2, \ldots.$$

Now, by using the dominated convergence theorem (see [17]), it follows that $\int h_n(\mathbf{s}) d\mathbf{s} \to \int h(\mathbf{s}) d\mathbf{s}$, when n goes to infinity. On the other hand, due to the continuity of Δ, it follows that $\upsilon^*(\mathbf{s})\Delta(\mathbf{s} - \mathbf{x}_n - \mathbf{a}_n) \to \upsilon^*(\mathbf{s})\Delta(\mathbf{s} - \mathbf{x} - \mathbf{a})$, as $n \to \infty$. Applying the dominated convergence theorem again, it is obtained that

$$\begin{aligned} \lim_{n \to \infty} \int \upsilon^*(\mathbf{y}) Q(d\mathbf{y} \mid \mathbf{x}_n, \mathbf{a}_n) &= \lim_{n \to \infty} \int I_X(\mathbf{s}) \upsilon^*(\mathbf{s}) \Delta(\mathbf{s} - \mathbf{x}_n - \mathbf{a}_n) d\mathbf{s} \\ &= \int I_X(\mathbf{s}) \upsilon^*(\mathbf{s}) \Delta(\mathbf{s} - \mathbf{x} - \mathbf{a}) d\mathbf{s} \\ &= \int \upsilon^*(\mathbf{y}) Q(d\mathbf{y} \mid \mathbf{x}, \mathbf{a}), \end{aligned}$$

i.e., $\int v^*(\mathbf{y})Q(d\mathbf{y} \mid \mathbf{x}, \mathbf{a})$ is a continuous function on $\mathbf{K}$. In an analogous way, $\int V_n(\mathbf{y})Q(d\mathbf{y} \mid \cdot, \cdot)$ is finite and continuous on $\mathbf{K}$ for each $n = 1, 2, \ldots$.

□

Lemma 4.5. *In Example 4.1, $f_n(\mathbf{x})$ and $f^*(\mathbf{x})$ are in the $int(A(\mathbf{x}))$ for all $\mathbf{x} \in X$ and $n \geq 1$.*

Proof. Let $\mathbf{x} \in X$ be fixed. First, it will be proven that $f_n(\mathbf{x}) \in int(A(\mathbf{x}))$, $n \geq 1$. Applying the value iteration algorithm (see Lemma 3.6 c)), it is obtained that for $n \geq 0$,

$$V_{n+1}(\mathbf{x}) = \min_{\mathbf{a} \in A(\mathbf{x})} \left\{ \mathbf{a}^T (I + \alpha K_n) \mathbf{a} + \mathbf{x}^T (I + \alpha K_n) \mathbf{x} + 2\alpha \mathbf{x}^T K_n \mathbf{a} + \alpha\theta_n + \alpha\beta_n \right\}, \tag{4.4}$$

with $V_0(\mathbf{x}) = 0$, where K_n is a positive semidefinite and symmetric matrix and $\{\theta_n\}$, $\{\beta_n\}$ are sequences of nonnegative real numbers. Define $g(\mathbf{a}) = \mathbf{a}^T (I + \alpha K_n) \mathbf{a} + 2\alpha \mathbf{x}^T K_n \mathbf{a}$ for each $\mathbf{a} \in A(\mathbf{x})$ (see (16)). Then, observe that g is a quadratic form with a minimum $\mathbf{a}^* = -\alpha \mathbf{x}^T K_n (I + \alpha K_n)^{-1}$ and $\| \mathbf{a}^* \| \leq \| \mathbf{x} \|$. The last assertion is verified via the following inequalities:

$$\begin{aligned} \| \mathbf{a}^* \| &= \| -\alpha K_n (I + \alpha K_n)^{-1} \mathbf{x} \| \\ &\leq \| \alpha K_n (I + \alpha K_n)^{-1} \| \| \mathbf{x} \| \\ &= \| ((I + \alpha K_n) - I) (I + \alpha K_n)^{-1} \| \| \mathbf{x} \| \\ &= \| I - (I + \alpha K_n)^{-1} \| \| \mathbf{x} \| \leq \| \mathbf{x} \| . \end{aligned}$$

The second inequality is obtained because $(I + \alpha K_n)^{-1}$ and I are nonnegative, definite and symmetric matrices with an application of Exercise 2.2-10 in [5] p. 57. Consequently, since the Hessian matrix of the quadratic form g is positive semidefinite and symmetric, it follows that $\mathbf{a}^* = f_n(\mathbf{x}) \in int(A(\mathbf{x}))$ is the unique minimum for each $n \geq 1$.

Then, substituting $\mathbf{a}^* = -\alpha \mathbf{x}^T K_n (I + \alpha K_n)^{-1}$ in (16), it is obtained that

$$\begin{aligned} V_{n+1}(\mathbf{x}) &= \mathbf{x}^T (I + \alpha K_n) \mathbf{x} + \alpha\theta_n + \alpha\beta_n + \min_{\mathbf{a} \in A(\mathbf{x})} \left\{ \mathbf{a}^T (I + \alpha K_n) \mathbf{a} + 2\alpha \mathbf{x}^T K_n \mathbf{a} \right\} \\ &= \mathbf{x}^T (I + \alpha K_n) \mathbf{x} + \alpha\theta_n + \alpha\beta_n \end{aligned}$$

$$
\begin{aligned}
&+ \left(-\alpha \mathbf{x}^T K_n \left(I+\alpha K_n\right)^{-1}\right)\left(I+\alpha K_n\right)\left(-\alpha\left(I+\alpha K_n\right)^{-1} K_n \mathbf{x}\right) \\
&+ 2\alpha \mathbf{x}^T K_n\left(-\alpha\left(I+\alpha K_n\right)^{-1} K_n \mathbf{x}\right) \\
&= \mathbf{x}^T\left(I+\alpha K_n\right)\mathbf{x}+\alpha\theta_n+\alpha\beta_n-\alpha^2 \mathbf{x}^T K_n\left(I+\alpha K_n\right)^{-1} K_n \mathbf{x} \\
&= \mathbf{x}^T\left(I+\alpha\left(K_n-\alpha K_n\left(I+\alpha K_n\right)^{-1} K_n\right)\right)\mathbf{x}+\alpha\theta_n+\alpha\beta_n.
\end{aligned}
$$

Equivalently, for $n \geq 0$,

$$V_{n+1}(\mathbf{x}) = \mathbf{x}^T K_{n+1}\mathbf{x} + D_{n+1},$$

where $K_{n+1} = I+\alpha\left(K_n-\alpha K_n\left(I+\alpha K_n\right)^{-1} K_n\right)$ and $D_{n+1} = \alpha\theta_n+\alpha\beta_n$.

On the other hand, from Lemma 3.6 c), it follows that $\{V_n\}$ converges pointwise to v^*. Then, there exist a positive semidefinite and symmetric matrix K and real numbers θ and α, such that $K_n \to K$, $\theta_n \to \theta$, $\alpha_n \to \alpha$. Consequently,

$$
\begin{aligned}
v^*(\mathbf{x}) &= \min_{\mathbf{a}\in A(\mathbf{x})}\left\{\mathbf{a}^T\left(I+\alpha K\right)\mathbf{a}+\mathbf{x}^T\left(I+\alpha K\right)\mathbf{x}+2\alpha\mathbf{x}^T K\mathbf{a}+\alpha\theta+\alpha\beta\right\} \\
&= \mathbf{x}^T\left(I+\alpha\left(K-\alpha K\left(I+\alpha K\right)^{-1} K\right)\right)\mathbf{x}+\alpha\theta+\alpha\beta \\
(4.5) \quad &= \mathbf{x}^T K\mathbf{x}+D,
\end{aligned}
$$

where K satisfies the Ricatti's equation $K = I+\alpha\left(K-\alpha K\left(I+\alpha K\right)^{-1} K\right)$ and $D = \alpha\theta+\alpha\beta$ is a positive number.

Finally, using (17) and proceeding in a similar way to the previous case, it is possible to prove that $f^*(\mathbf{x}) \in int(A(\mathbf{x}))$. Since $\mathbf{x}$ is arbitrary, result follows. □

Theorem 4.6. *In Example 4.1, the gradient of $H_{\mathbf{x}}$ is given by*

$$\nabla H_{\mathbf{x}}(\mathbf{a}) = \lambda\left(\mathbf{a}-\left(\lambda I\mathbf{a}-2\alpha K\mathbf{x}\right)\left(2I+2\alpha K+\lambda I\right)^{-1}\right)$$

for each $(\mathbf{x},\mathbf{a}) \in \mathbf{K}$, where K is a matrix positive semidefinite and symmetric that satisfies the Ricatti's equation $K = I+\alpha\left(K-\alpha K\left(I+\alpha K\right)^{-1} K\right)$.

Furthermore, $f^*(\mathbf{x}) = -\alpha\mathbf{x}^T K\left(I+\alpha K\right)^{-1}$, $\mathbf{x} \in X$.

Proof. Let $(\mathbf{x},\mathbf{a}) \in \mathbf{K}$ be fixed. Then, substituting (17) in $H_{\mathbf{x}}(\mathbf{a})$ (see Definition 2.3 and Remark 3.8 a)), it is obtained that

$$H_{\mathbf{x}}(\mathbf{a}) = \min_{\mathbf{b}\in A(\mathbf{x})}\left\{\mathbf{x}^T\mathbf{x}+\mathbf{b}^T\mathbf{b}+\alpha E\left[v^*\left(\mathbf{x}+\mathbf{b}+\xi\right)\right]+\frac{\lambda}{2}(\mathbf{b}-\mathbf{a})^T(\mathbf{b}-\mathbf{a})\right\}$$

$$
\begin{aligned}
&= \min_{\mathbf{b}\in A(\mathbf{x})} \Big\{ \mathbf{x}^T\mathbf{x} + \mathbf{b}^T\mathbf{b} + \alpha E\left[(\mathbf{x}+\mathbf{b}+\xi)^T K (\mathbf{x}+\mathbf{b}+\xi) + D \right] \\
&+ \frac{\lambda}{2}(\mathbf{b}-\mathbf{a})^T(\mathbf{b}-\mathbf{a}) \Big\} \\
&= \min_{\mathbf{b}\in A(\mathbf{x})} \Big\{ \mathbf{x}^T\mathbf{x} + \mathbf{b}^T\mathbf{b} + \alpha (\mathbf{x}+\mathbf{b})^T K (\mathbf{x}+\mathbf{b}) + \alpha\sigma^2 + \alpha D \\
&+ \frac{\lambda}{2}(\mathbf{b}-\mathbf{a})^T(\mathbf{b}-\mathbf{a}) \Big\}, \qquad (4.6)
\end{aligned}
$$

where K is a matrix positive semidefinite and symmetric that satisfies the Ricatti's equation $K = I + \alpha\left(K - \alpha K (I+\alpha K)^{-1} K\right)$ and $D = \alpha\theta + \alpha\beta$ is a positive number. Then, minimizer $\mathbf{b}^*$ of (18), satisfies the following equation:

$$2I\mathbf{b}^* + 2\alpha K(\mathbf{x}+\mathbf{b}^*) + \lambda I(\mathbf{b}^* - \mathbf{a}) = \mathbf{0}.$$

This implies that

$$\mathbf{b}^* = (\lambda I\mathbf{a} - 2\alpha K\mathbf{x})(2I + 2\alpha K + \lambda I)^{-1}.$$

Therefore,

$$P(\mathbf{x},\mathbf{a}) = (\lambda I\mathbf{a} - 2\alpha K\mathbf{x})(2I + 2\alpha K + \lambda I)^{-1}.$$

Finally, by Theorem 3.13, it results that

$$\nabla H_{\mathbf{x}}(\mathbf{a}) = \lambda\left(\mathbf{a} - (\lambda I\mathbf{a} - 2\alpha K\mathbf{x})(2I + 2\alpha K + \lambda I)^{-1}\right)$$

and

$$\nabla H_{\mathbf{x}}(f^*(\mathbf{x})) = \lambda\left(f^*(\mathbf{x}) - (\lambda I f^*(\mathbf{x}) - 2\alpha K\mathbf{x})(2I + 2\alpha K + \lambda I)^{-1}\right) = \mathbf{0}.$$

Then, the optimal policy is given by

$$f^*(\mathbf{x}) = -\alpha\mathbf{x}^T K (I + \alpha K)^{-1}.$$

□

5. Final Remarks

For suitable discounted MDPs, a procedure to perturb the corresponding cost function by adding a convenient quadratic function was found. It is also significant to mention that if the cost function is not differentiable, then for the perturbed cost function, the answer regarding this property is

positive; that is, the perturbed cost function is differentiable. Moreover, the optimal policies for the original model and the perturbed one are the same. For future work, the authors will seek to develop an algorithm to calculate the optimal policy and the optimal value function for the perturbed model by applying Lipschitzian gradient methods in the context of Moreau-Yosida regularization (see [14]). In addition to considering alternative methods as regularized Bellman operators (see [10]) to complement the methodology of Moreau-Yosida regularization.

Acknowledgments

The authors are deeply grateful to the reviewers and the associate editor for their careful reading of the original manuscript, and for their helpful suggestions to improve the paper.

References

[1] C. D. Aliprantis and K. C. Border, *Infinite dimensional analysis: a hitchhikers guide*. Berlin: Springer, 2006, doi: 10.1007/3-540-29587-9

[2] A. Belloni, *Lecture notes for IAP 2005 course introduction to Bundle methods*, 2005. [On line]. Available: https://bit.ly/3qbJfmb

[3] D. P. Bertsekas, *Dynamic programming and stochastic control*. New York, NY: Academic Press, 1976.

[4] J. M. Borwein and Q. J. Zhu, *Techniques of variational analysis*. New York: Springer, 2005.

[5] P. G. Ciarlet, B. Miara, and J. M. Thomas, *Introduction to numerical linear algebra and optimisation*. New York, NY: Cambridge University Press, 1989.

[6] D. Cruz-Suárez, R. Montes-de-Oca, and F. Salem-Silva, "Conditions for the uniqueness of optimal policies of discounted Markov decision processes", *Mathematical methods of operational research*, vol. 60, no. 3, pp. 415–436, Dec. 2004, doi: 10.1007/s001860400372

[7] J. Dugundji, *Topology*. Boston, MA: Allyn and bacon, 1966. [On line]. Av Available: https://bit.ly/398XKQA

[8] I. Đuranović-Miličić and M. Gardašević-Filipović, "On an Algorithm in nondifferential convex optimization", *Yugoslav journal of operations research,* vol. 23, no.1, pp. 59-71, 2013, doi: 10.2298/YJOR110501024D

[9] S. H. Friedberg, A. Insel, and L. Spence, *Linear algebra.* Upper Saddle River, NJ: Pearson, 2003.

[10] M. Geist, B. Scherrer, O. Pietquin, "Theory of regularized Markov decision processes", *Proceedings of machine learning research* , vol. 97, pp. 2160-2169, 2019. [On line]. Available: https://bit.ly/3bdfpcC

[11] O. Hernández-Lerma, *Adaptive Markov control processes.* New York, NY: Springer, 1989.

[12] O. Hernández-Lerma and J. B. Lasserre, *Markov control processes: basic optimality criteria: discrete-time.* New York, NY: Springer, 1996.

[13] O. Hernández-Lerma and J. B. Lasserre, *Discrete-time Markov control processes.* New York, NY: Springer, 1999.

[14] C. Lemaréchal and C. Sagastizábal, "Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries", *SIAM journal on optimization,* vol. 7, no. 2, pp. 367-385, 1997, doi: 10.1137/S1052623494267127

[15] R. T. Rockafellar, *Convex analysis*. Princeton, NJ: Princeton University Press, 1970.

[16] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Heidelberg: Springer, 2004.

[17] H. L. Royden, *Real analysis*. New York, NY: Macmillan, 1988.