A Lyapunov Shortest-Path Characterization for Markov Decision Processes

Julio B. Clempner and Jesus Medel

Abstract— **In this paper we introduce a modeling paradigm for developing decision process representation for shortest-path problems. Whereas, in previous work attention was restricted to tracking the net using Bellman's equation as a utility function, this work uses a Lyapunov-like function. In this sense, we are changing the traditional cost function by a trajectory-tracking function which is also an optimal cost-to-target function for tracking the net. The main point of the Markov decision process is its ability to represent the system-dynamic and trajectory-dynamic properties of a decision process. Within the system-dynamic properties framework we prove new notions of equilibrium and stability. In the trajectory-dynamic properties framework, we optimize the value of the trajectory-function used for path planning via a Lyapunov-like function, obtaining as a result new characterizations for final decision points (optimum points) and stability. Moreover, we show that the system-dynamic and Lyapunov trajectory-dynamic properties of equilibrium, stability and final decision points (optimum points) meet under certain restrictions.**

Index Terms— **Lyapunov theory, Bellman's equation, Forward Decision Process, Markov decision process.**

I. Introduction

hereas previous efforts have restricted attention to track the net using Bellman's equation as a utility function, this paper introduces a modeling paradigm for developing decision process representation, including Markov decision processes (MDP), using a trajectory function as a tool for path planning ([1], [2]). The main point of this paper is its ability to represent the system-dynamic and the trajectory-dynamic properties of a decision process application. We will identify the system-dynamic properties as those characteristics related only with the global system behavior, and we will identify the trajectory-dynamic properties as those characteristics related with the trajectory function at each state that depends on a probabilistic routing policy.

Within the system-dynamic properties framework we show notions of stability. In this sense, we call equilibrium point to the state in a MDP that does not change, and it is the last state in the net.

In the trajectory-dynamic properties framework we define the trajectory function as a Lyapunov-like function. By an appropriate selection of the Lyapunov-like function, under certain desired criteria, it is possible to optimize the trajectory. By optimizing the trajectory we understand that it is maximum or minimum reward (in a certain sense). In addition, we use the notions of stability in the sense of Lyapunov to characterize the stability properties of the MDP. The core idea of our approach uses a non-negative trajectory function that converges in decreasing form to a (set of) final decision states. It is important to point out that the value of the trajectory function associated with the MDP implicitly determines a set of policies, not just a single policy (in case of having several decisions states that could be reached). We call "optimum point" the best choice selected from a number of possible final decision states that may be reached (to select the optimum point, the decision process chooses the strategy that optimizes the reward).

As a result, we extend the system-dynamic framework including the trajectory-dynamic properties. We show that the system-dynamic and the trajectory-dynamic properties of equilibrium, stability and optimum-point conditions converge under certain restrictions: if the MDP is finite then we have that a final decision state is an equilibrium point.

The paper is structured in the following manner. Section 2 presents the formulation of the decision model, and all the structural assumptions are introduced there. Section 3 discusses the main results of the paper, giving a detailed analysis of the equilibrium, stability and optimum-point conditions for the MDP. Finally, in section 4 some concluding remarks and future work projects are outlined.

II. Formulation

The aim of this section is to introduce the decision model and all the structural assumptions related with the Markov model ([3], [5], [9]).

Notation 1: As usual let $\mathbb{R}$ be the set of real number and let $\mathbb{N}$ be the set of non-negative integers.

Definition 1: A Markov Decision Process is a 5-tuple

$$MDP = \{S, A, \Upsilon, Q, U\} \tag{1}$$

where:

- S is a countable set of feasible states, $S \subset \mathbb{N}$, endowed with discrete topology[1].
- A is the set of actions, which is a metric space. For each $s \in S$, $A(s) \subset A$ is the non-empty set of admissible actions at state $s \in S$. Without loss of generality we may take $A = \bigcup_{s \in S} A(s)$.

J. B. Clempner is with the Center for Computing Research. National Polytechnic Institute, Av. Juan de Dios Batiz s/n, Edificio CIC, Col. Nueva Industrial Vallejo, 07738, Mexico City, Mexico (e-mail: julio@k-itech.com)

J. J. Medel is with the Center for Computing Research. National Polytechnic Institute, Av. Juan de Dios Batiz s/n, Edificio CIC, Col. Nueva Industrial Vallejo, 07738, Mexico City, Mexico (e-mail: jjmedelj@yahoo.com.mx)

[1]Note that the existence of a topology on S is trivial, since S is countable. We introduce it for definition compatibilities.

- $\Upsilon = \{(s,a)|s \in S, a \in A(s)\}$ is the set of admissible state-action pairs, which is a measurable subset of $S \times A$.
- $Q = \left[q_{ij|k}\right]$ is an array of probabilities, where $q_{ij|k} \equiv P(s_j|s_i,a_k)$ representing the probability associated with the transition from state s_i to state s_j under an action $a_k \in A(s_i)$. Note that for any fixed k, $Q|_k$ is a stochastic matrix.
- $U : S \to \mathbb{R}_+$ is a trajectory function, associating to each state a real value. Note that U is a function bounded from below. (moreover, it is convenient to use $\|U\| = \sup_{s\in S} U(s)$).

Interpretation: The control model (1) represents a discrete-time controlled stochastic system that is observed at time $n \in \mathbb{N}$. Denoting by s_n and a_k the state of the system and action applied at time n, respectively, the interpretation of the MDP dynamics is as follows. At each discrete time $n \in \mathbb{N}$ the state of the system $s_n = s \in S$ is observed. For every action $a_n = a \in A(s)$, the probability of the system to find itself in the next state s_{n+1} at time $n+1$ is $\mathbb{P}(s_{n+1}|s_n = s, a_k = a)$. Considering the previous states of the trajectory (path, orbit) $(s_0, s_1, ..., s_n)$ the value of the trajectory function U is obtained and, then the next state s_{n+1} is selected according to U applying some 'criteria'. This is the Markov property of the decision process in (1).

For each $n \in \mathbb{N}$ the cross product $H_n = \Upsilon^n \times S$ is the set of admissible histories up to time n. The vector $h_n = (s_0, a_0, ..., s_{n-1}, a_{n-1}, s_n) \in H_n$ denotes the history of the process at time n. Considering the previous states of the trajectory $(s_0, s_1, ..., s_n)$, and for every action $a_n \in A(s_i)$, the probability of the system to find itself in state $s_j \in S$ is $q_{ij|k}$. A policy π is a (possibly randomized) measurable rule for choosing actions, which depends on the current state. The policy $\pi_{k|i} \equiv P(a_k|s_i)$ represents the probability measure associated with the occurrence of an action a_n from state s_i. The set of all policies is denoted by Π.

We define a process over S as a finite or infinite sequence of elements of S. If $p = (s_0, s_1 ... s_n)$ is a finite process, we say that s_n is the end state of p, and we denote it $last(p) = s_n$. For completeness, $first(p) = s_0$ denote the state in which p starts. Let us define the sample space $\Omega = (S \times A)^\infty$, i.e. Ω represents the set of infinite processes over S. Let us define the random variables $X_n : \Omega \to S$ for each $n \in \mathbb{N}$, so that we have: $X_n(\omega) = x_n$ for $\omega = (x_0, a_0, x_1, ...)$.

Let $(\Omega, \mathcal{F})$ be a measurable space with $\mathcal{F}$ a σ-algebra of subsets of the previously defined sample space Ω. We define a probabilistic process over S as a pair $(S, \mathbb{P})$, where $\mathbb{P}$ is a probability measure on $\mathcal{F}$. If there is an element $s_0 \in S$ such that $X_0 = s_0$, we say that s_0 is the initial state of the probabilistic process $(S, \mathbb{P})$. Let $p = (s_0, ..., s_n)$ be a finite process.

We define the likelihood of p by $\mathbb{P}(p)$. Intuitively, $\mathbb{P}(p)$ is the probability measure of p to occur in an execution of the system. Be aware however that the likelihood function does not define a probability measure on the set of finite processes, since it does not sum to 1.

Let $(S, \mathbb{P})$ be a probabilistic process, and let $p = (s_0, ..., s_0)$ be a finite process over S with $\mathbb{P}(p) > 0$. Let us consider the mapping $g : p \to \Omega$ defined by: $g(s_0, s_1, ..., s_n, X_{n+1}, X_{n+2}, ...) = (s_n, X_{n+1}, X_{n+2}, ...)$.

The mapping g let us define a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ as follows: $\forall A \in \mathcal{F}$, $\mathbb{P}(A) = \mathbb{P}(g^{-1}(A)|p)$, where $\mathbb{P}(\cdot\ |p)$ is the probability conditional on p. We call the new probabilistic process $(S, \mathbb{P})$ the probabilistic future of process p. We denote by the symbol $\mathbb{E}$ the expectation under probability $\mathbb{P}$. By construction, $s_n = last(p)$ is the initial state of the probabilistic future of p.

Definition 2: Two given processes p and p' represent a Path of the following type:

1) OR if one has associated a better probability $\mathbb{P}$ to occur at the same time,
2) AND if they have associated any probability $\mathbb{P}$ they occur at the same time,
3) Concur if they have associated the same probability $\mathbb{P}$ to occur at the same time.

From the previous definition we have the following remark.

Remark 1: In a Concur-Path, we have $last(p) = last(p')$ and therefore we also have $\mathbb{P}(p) = \mathbb{P}(p')$.

Consider an arbitrary $s_j \in S$ and for each fixed action $a_k \in A$ we look at the previous states s_i of the state s_j, denoted by $s_{\eta_{jk}} = \{s_h : h \in \eta_{jk}\}$ where $\eta_{jk} = \{h : (s_h, a_k, s_j)\}$, that materialize the concurrent state-action pair $(s_h, a_k) \in \Upsilon$ and form the sum

$$\sum_{h\in\eta_{jk}} \pi_{k|h} q_{hj|k}\ U_h^{(\pi_{k|h})} \tag{2}$$

Notation 2: With the intention to facilitate the notation we will represent the trajectory function U as follows:

1) $U_i \equiv U(s_i)$ representations of the value of U at state s_i.
2) $U_i \equiv U_i^{(\pi)}$ for an arbitrary policy π.

Continuing with all the a_k's we form the vector indexed by the sequence k identified by $(k_0, k_1, ..., k_f)$ as follows:

$$\begin{bmatrix} \sum_{h\in\eta_{jk_0}} \pi_{k_0|h} q_{hj|k_0} U_h, & \sum_{h\in\eta_{jk_1}} \pi_{k_1|h} q_{hj|k_1} U_h, \\ & ..., \sum_{h\in\eta_{jk_f}} \pi_{k_f|h} q_{hj|k_f} U_h \end{bmatrix} \tag{3}$$

the index sequence k is the set $\lambda = \{k : a_k \in (s_h, a_k, s_j)$, and s_h running over the set $s_{\eta_{jk}}\}$, and $f = \#(\lambda)$ is the number of actions to state s_j.

Intuitively, the vector (3) represents all the possible trajectories through the actions a_k where $(k_0, k_1, ..., k_f)$ to a state s_j for a fixed j.

Continuing the construction of the definition of the trajectory function U, let us introduce the following definition.

Definition 3: Let $MDP = \{S, A, \Upsilon, Q, U\}$ be a Markov Decision Process, let $(s_0, s_1, ..., s_n)$ be a realized trajectory of the system and let $L : \mathbb{R}^n \to \mathbb{R}_+$ be a continuous map. Then L is a Lyapunov-like function [6] if it satisfies the following properties:

1) $\exists s^*$ such that $L(s^*) = 0$,
2) $L(s) > 0$ for $\forall s \neq s^*$,
3) $L(s) \to \infty$ when $s \to \infty$,
4) $\Delta L = L(s_{i+1}) - L(s_i) < 0\ \forall i\ s_i \neq s^*$.

From the previous definition we have the following remark.

Remark 2: In the previous definition point 3 we state that $L(s) \rightarrow \infty$ when $s \rightarrow \infty$ meaning that there is no s^* reachable from some s.

Then, formally we define the trajectory function U as follows:

Definition 4: For the discrete time $n \in \mathbb{N}$ the trajectory function U with respect a Markov Decision Process $MDP = \{S, A, \Upsilon, Q, U\}$ is represented by

$$U_j = \begin{cases} U_0 & if \quad n = 0 \\ L(\alpha) & if \quad n > 0 \end{cases} \tag{4}$$

where

$$\alpha = \begin{bmatrix} \sum_{h \in \eta_{jk_0}} \pi_{k_0|h} q_{hj|k_0} U_h, \sum_{h \in \eta_{jk_1}} \pi_{k_1|h} q_{hj|k_1} U_h, \\ ..., \sum_{h \in \eta_{jk_f}} \pi_{k_f|h} q_{hj|k_f} U_h \end{bmatrix} \tag{5}$$

the function $L : D \subseteq \mathbb{R}^n_+ \rightarrow \mathbb{R}_+$ is a function that optimizes the reward through all possible transitions (i.e. trough all the possible trajectories defined by the different a_k's), D is the decision set formed by the k's : $0 \leq k_l \leq f$ of all those possible transitions (s_h, a_k, s_j), η_{jk} is the index sequence of the list of previous places to s_j through action a_k and s_h $(h \in \eta_{jk_l})$ is a specific previous place of s_j through action a_k.

From the above definition we have the following remark.

Remark 3:

- Note that the Lyapunov-like function L guarantees that the optimal course of action is followed (taking into account all the possible paths defined). In addition, the function L establishes a preference relation because by definition L is asymptotic; this condition gives to the decision maker the opportunity to select a path that optimizes the reward.
- The iteration over time $n \in \mathbb{N}$ for U is as follows:
  1) for $n = 0$ the trajectory function value is U_0 at state s_0 and for the rest of the states s_i the value is 0,
  2) for $n > 0$ the trajectory function value is U_j at each state s_j, is computed by taking into account the value of the previous states s_i.

Property 1: The function U satisfies the following properties:

1) $\exists s^{\triangle}$ such that
   a) if there exists an infinite sequence $\{s_i\}_{i=1}^{\infty}$ with $s_n \underset{n \rightarrow \infty}{\rightarrow} s^{\triangle}$ (s_n converge at $s^{\triangle}$) such that $0 \leq ... < U_n < U_{n-1} ... < U_1$, then $U(s^{\triangle})$ is the infimum of the infinite sequence, i.e. $U(s^{\triangle}) = 0$,
   b) if there exists a finite sequence $s_1, ..., s_n$ with $s_1, ..., s_n \rightarrow s^{\triangle}$ ($s_1, ..., s_n$ converge at $s^{\triangle}$) and there exists a constant $C \in \mathbb{R}$ such that $C = U_n < U_{n-1} ... < U_1$, then $U(s^{\triangle})$ is the minimum of the finite sequence, i.e. $U(s^{\triangle}) = C$, $(s^{\triangle} = s_n)$.
2) there exists a constant $C \in \mathbb{R}$ such that $U(s_i) > \max\{0, C\}$, $\forall s_i$ such that $s_i \neq s^{\triangle}$.
3) $\forall s_i, s_{i-1}$ such that $s_{i-1} \leq_U s_i$ then $\forall i \ \Delta U_i = U_i - U_{i-1} < 0$ (a trajectory function $U : S \rightarrow \mathbb{R}$ is consistent with the preference relationship of a decision problem $(S, \leq)$ if $\forall w, z \in S : w \leq_u z$ if and only if $U_w \leq U_z$)

Property 2: The trajectory function $U : S \rightarrow \mathbb{R}_+$ is a Lyapunov-like function.

Proof: Straightforward from the previous definitions. ■

Explanation. Intuitively, a Lyapunov-like function can be considered as routing function and optimal cost function. In our case, an optimal discrete problem, the cost-to-target values are calculated using a discrete Lyapunov-like function. Every time a discrete vector field of possible actions is calculated over the decision process. Each applied optimal action (selected via some 'criteria') decreases the optimal value, ensuring that the optimal course of action is followed and establishing a preference relation. In this sense, the criteria change the asymptotic behavior of the Lyapunov-like function by an optimal trajectory tracking value. It is important to note, that the process finishes when the equilibrium point is reached. This point determines a significant difference to the use of Bellman's equation.

Definition 5: A final decision point s_f with respect a Markov Decision Process $MDP = \{S, A, \Upsilon, Q, U\}$ is a state s where the infimum of the trajectory function is asymptotically approached (or the minimum is attained), i.e. $U(s) = 0$ or $U(s) = C$.

Definition 6: An optimum point $s^{\triangle}$ with respect a Markov Decision Process $MDP = \{S, A, \Upsilon, Q, U\}$ is a final decision point s_f where the best choice is selected 'according to some criteria'.

Assumption 1: Every Markov Decision Process $MDP = \{S, A, \Upsilon, Q, U\}$ has a final decision point.

Remark 4: In case that $\exists s_1, ..., s_n$, such that $U(s_1) = ... = U(s_n) = 0$, then $s_1, ..., s_n$ are optimum points.

Remark 5: The monotonicity of U guarantees that it is possible to make the search starting from the decision points.

Proposition 1: Let $MDP = \{S, A, \Upsilon, Q, U\}$ be a Markov Decision Process and let $s^{\triangle}$ an optimum point. Then $U(s^{\triangle}) \leq U(s)$, $\forall s$ such that $s \leq_U s^{\triangle}$.

Proof: We have that $U(s^{\triangle})$ is equal to the minimum or the infimum. Therefore, $U(s^{\triangle}) \leq U(s)$ $\forall s$ such that $s \leq_U s^{\triangle}$. ■

Theorem 1: Let $MDP = \{S, A, \Upsilon, Q, U\}$ be a Markov Decision Process. If s^* is an equilibrium point then it is a final decision point.

Proof: Let us suppose that s^* is an equilibrium point we want to show that its trajectory function value has asymptotically approached an infimum (or reached a minimum). Since s^* is an equilibrium point, by definition, it is the last state of the net. But, this implies that the routing policy attached to the transition(s) that follows s^* is 0, (in case there is such a transition(s) i.e., worst case). Therefore, its value can not be modified and since the trajectory function is a decreasing function of s_i an infimum or a minimum is attained. Then, s^* is a final decision point. ■

Theorem 2: Let $MDP = \{S, A, \Upsilon, Q, U\}$ be a (finite) Markov Decision Process (unless s is an equilibrium point). If s_f is a final decision point then it is an equilibrium point.

Proof: If s_f is a final decision point, since the MDP is finite, there exists some n such that $U(s_f) = C$. Let us

suppose that s_n is not an equilibrium point.

case 1. Then, it is not bounded. So, it is possible to fire some transition of s_f in the MDP. Therefore, it is possible to modify its value. As a result, it is possible to obtain a lower value than C.

case 2. Then, it is not the last state in the net. So, it is possible to fire some transition to s_f. Therefore, it is possible to modify the trajectory function value over s_f. As a result, it is possible to obtain a lower value than C. ■

Corollary 1: Let $MDP = \{S, A, \Upsilon, Q, U\}$ be a finite Markov Decision Process (unless s is an equilibrium point). Then, an optimum point $s^{\triangle}$ is an equilibrium point.

Proof: From the previous theorem we know that a final decision point is an equilibrium point and since in particular $s^{\triangle}$ is final decision point then, it is an equilibrium point. ■

Remark 6: The finite condition over the MDP can not be relaxed. Let us suppose that the MDP is not finite, i.e. s is in a cycle then, the Lyapunov-like function converges when $n \to \infty$, to zero i.e., $L(s) = 0$ but the MDP has no final state therefore, it is not an equilibrium point.

III. Conclusion

A formal framework for decision process has been presented. Stability theory was used to characterize the dynamical behavior of the MDP. In addition, we show that the MDP mark-dynamic and trajectory-dynamic properties of equilibrium, stability and optimum point converge under some mild restrictions. There are a number of questions relating classical planning, that may in the future be addressed satisfactorily within this approach.

References

[1] J. Clempner. Towards Modeling The Shortest-Path Problem and Games with Petri Nets. *Proc. of The Doctoral Consortium at the ICATPN*, 1-12, 2006.

[2] J. Clempner and J. Medel. A Simple Modal Approach to Decision Process. *Proceedings of the 9th WSEAS Int.Conf. on Mathematical and Computational Methods in Science and Engineering*, 34-38, 2007.

[3] O. Hernández-Lerma and J.B. Lasserre. Discrete-Time Markov Control Process: Basic Optimality Criteria. — Berlin, Germany : Springer, 1996.

[4] O. Hernández-Lerma, G. Carrasco and R. Pére-Hernández. Markov Control Processes with the Expected Total Cost Criterion: Optimality, Stability and Transient Model. Acta Applicadae Matematicae, 59, 3 229-269, 1999.

[5] O. Hernández-Lerma and J.B. Lasserre. Futher Topics on Discrete-Time Markov Control Process. Berlin, Germany: Springer-Verlag, 1999.

[6] R. E. Kalman and J. E. Bertram. Control System Analysis and Design Via the "Second Method" of Lyapunov. Journal of Basic Engineering, 82(D), 371-393, 1960.

[7] Lakshmikantham, S. Leela and A.A. Martynyuk, *Practical Stability of Nonlinear Systems,* World Scientific, Singapore, 1990.

[8] V. Lakshmikantham, V.M. Matrosov and S. Sivasundaram, *Vector Lyapunov Functions and Stability Analysis of Nonlinear Systems,* Kluwer Academic Publ., Dordrecht, 1991.

[9] A. S. Poznyak, K. Najim and E. Gomez-Ramirez. Self-learning control of finite markov chains. Marcel Dekker, Inc., New York, 2000.

Dr. Julio Clempner holds a Ph.D. from the Center for Computing Research at the National Polytechnic Institute. His research interests are focused on justifying and introducing the Lyapunov equilibrium point in shortest-path decision processes and shortest-path game theory. This interest has lead to several streams of research. One stream is on the use of Markov decision processes for formalizing the previous ideas, changing the Bellman's equation by a Lyapunov-like function which is a trajectory-tracking function, and also it is an optimal cost-to-target function for tracking the net. A second stream is on the use of Petri nets as a language for modeling decision process and game theory introducing colors, hierarchy, etc. Petri nets are used for process representation taking advantage of the formal semantic and the graphical display. The final stream examines the possibility to meet modal logic, decision processes and game theory. He is a member of the Mexican National System of Researchers (SNI).

Dr. Jesus Medel holds a PhD from the Center for Research and Advanced Studies (Cinvestav), National Polytechnic Institute. He is a member of the Mexican National System of Researchers (SNI) and a member of the Mexican Academy of Sciences. Since 1999 he is a Professor at the National Polytechnic Institute. At present, he has graduated seven master's and two doctoral students. He has published more than sixty papers and one book, and a second book is in press. In 2004, he was awarded the degree of Magister and PhD in postgraduate education management. His research areas of interest include Filter Theory, Real-time Filter Theory, Non-stationary Sample Time, Real-time Fuzzy Logic Filters, and others.