Continuous-Time Markov Decision Problems with Binary State

CHIARA BRAMBILLA, LUCA GROSSET, ELENA SARTORI
Dipartimento di Matematica "Tullio Levi-Civita"
Università degli Studi di Padova
Via Trieste, 63 - 35121 Padova
ITALY

Abstract: We analyse a binary state continuous-time Markov decision problem. The standard Hamilton–Jacobi–Bellman equation is introduced and, with suitable assumptions on the probability rate and on the cost function, it can be replaced by a simpler backward differential equation. Through a numerical example, we show how to find an optimal feedback control using the results presented in this paper.

Key-Words: Continuous-time Finite-state Markov Process, Optimal Stochastic Control, Dynamic Programming.

Received: October 24, 2022. Revised: December 20, 2022. Accepted: January 18, 2023. Published: February 16, 2023.

1 Introduction

The theory of continuous-time Markov decision processes is a rapidly growing area of research with numerous practical applications; see, e.g., [4], and [6]. This paper focuses on binary state processes and prioritises an analytical solution over a computational one. The description of the evolution of binary state processes draws inspiration from the Curie-Weiss model [5], but recent models have also incorporated human rationality by allowing decision-making agents to modify their states. Initially, agents were assumed to update their state according to a predetermined probabilistic transition rate based on their surrounding environment. However, this approach does not take into account the rationality of humans. Therefore, to change their states, agents face decision problems and try to minimise or maximise a suitable objective (such as cost, gain, or level of satisfaction). To model this situation, we introduce in the transition rate a control function, which is chosen by the agents once they have solved their optimisation problem. This leads to the definition of a controlled Markov chain. It is well known that the Hamilton–Jacobi–Bellman equation (HJB) associated with a controlled continuous-time finite-state Markov process can be difficult to treat. In this paper, we use techniques developed in [1] and [2] to derive closed or semi-closed expressions for optimal control and state evolution.

2 The model

In this section, we describe the continuous-time Markov decision process that we are going to study, which is defined in the programming interval $[0, T]$ with $T > 0$. Let $S = \{-1, 1\}$ be the state space and let $[0, \upsilon]$, with $\upsilon > 0$, be the control set. For $t \in [0, T]$ denote by X_t the state process and by $U_t \in [0, \upsilon]$ the control process. We formally define the dynamics of the process by

$$\mathbb{P}\left(X_{t+h} = -x \,|\, X_t = x, U_t = u\right) = \\ = \ell(x, u)h + o(h), \quad (1)$$

where $\ell(x, \cdot) : [0, \upsilon] \to [0, +\infty)$ is a continuous function for all $x \in S$. Now, assume that the system is controlled using a feedback function:

$$U_t = u(t, X_{t^-}), \quad (2)$$

where

$$u : [0, T] \times \{-1, 1\} \to [0, \upsilon] \quad (3)$$

is a measurable function. We denote by $\mathcal{U}$ the set of all feasible feedback control functions. We recall that for all feasible feedback functions, there exists a continuous-time Markov process defined by (1).

For the reader's convenience, we sketch how control and state are connected. We observe that the following approach is a bit more general than what we would need because it can be applied not only for the feedback controls, but also for more general non-anticipating controls. Nevertheless, we prefer to keep this approach since it is the most popular in stochastic optimal control problems. Let us denote by $B(S)$ the space of bounded functions with real value in S equipped with the supremum norm.
For all $t \in [0, T]$ and for all measurable control functions u, consider the following operator:

$$\begin{array}{rccc} \Lambda_t^u : & B(S) & \to & B(S) \\ & f & \mapsto & \Lambda_t^u f, \end{array} \quad (4)$$

where

$$\Lambda_t^u f(x) := \ell(x, u(t, x))\left(f(-x) - f(x)\right)$$

for $x \in S$. Let $\mathcal{D} = \mathcal{D}([0,T], S)$ be the space of right-continuous functions with finite left limit defined in $[0,T]$, taking values in the finite binary set $S = \{-1, 1\}$. We endow this space with the Skorokhod topology and denote by $\mathcal{S}$ the Borel σ-algebra on $\mathcal{D}$. In the measurable space $(\mathcal{D}, \mathcal{S})$, we denote by $(X_t)_{t\in[0,T]}$ the canonical process: $X_t(\omega) = \omega(t)$.

Let μ be a probability measure on S. A probability measure $\mathbb{P}^u_{s,\mu}$ on $(\mathcal{D}, \mathcal{S})$ is a solution to the martingale problem characterised by (4) if and only if the following two conditions hold:

1. $\mathbb{P}^u_{s,\mu}(X_s \in A) = \mu(A)$ for $s \in [0,T]$, $A \subset S$;
2. for all functions $f \in B(S)$, the process $f(X_t) - \int_0^t \Lambda^u_r f(X_r)dr$ is a martingale with respect to the natural filtration $\mathcal{F}_t = \sigma(X_r, r \leq t)$.

The process X_t characterised by the previous martingale problem is the unique continuous-time Markov chain with infinitesimal generator given by (4).

This discussion allows us to clarify the connection between a feedback control function and the associated state function. Therefore, we denote by $\mathbb{E}^u_{s,\mu}$ the expectation with respect to the probability measure $\mathbb{P}^u_{s,\mu}$. Furthermore, if μ is the measure that concentrates all probability in the state $x \in S$, we write $\mathbb{P}^u_{s,x}$ and $\mathbb{E}^u_{s,x}$.

3 Stochastic optimal control

In this section, we present the finite-time optimal control problem we are dealing with. We are looking for a feasible feedback control function u that minimises

$$J(u) := \mathbb{E}^u_{0,x}\left\{\int_0^T c(t, X_t, U_t)dt + g(X_T)\right\}, \quad (5)$$

where $x \in S$, $c(\cdot,\cdot,\cdot)$ is a continuous function and $g(\cdot)$ is given.
Let's introduce the optimal value function:

$$V(t,x) = \inf_{u\in\mathcal{U}} \mathbb{E}^u_{t,x}\left\{\int_t^T c(s, X_s, U_s)ds + g(X_T)\right\}.$$

This function is the key point to solving our optimal control problem because it solves a particular equation. More in detail, the HJB equation associated with the previous optimal control problem can be written as

$$\begin{cases} \partial_t V(t,x) + \min\limits_{w\in[0,\upsilon]} \{c(t,x,w) + \Lambda^w_t V(t,x)\} = 0 \\ V(T,x) = g(x), \end{cases} \quad (6)$$

where $x \in \{-1, 1\}$, and $t \in [0,T)$.

In the following Theorem, we explain the connection between the optimal value function, a solution of (6), an optimal feedback control function.

Theorem 1 (HJB Equation)**.** *Let us assume that:*

- $S = \{-1, 1\}$*;*
- $\ell(x,\cdot) : [0,\upsilon] \to [0,+\infty)$ *is a continuous function for all* $x \in S$*;*
- $c : [0,T] \times S \times [0,\upsilon] \to \mathbb{R}$ *is a continuous function.*

Then:

- *There exists a unique function, bounded and continuously differentiable with respect to the first variable, which is a solution of the HJB equation;*
- *An optimal control in the feedback form* $u^* : [0,T] \times \{-1,1\} \to [0,\upsilon]$ *is given by*

$$c(t,x,u^*(t,x)) + \Lambda^{u^*}_t V(t,x) = \\ = \min_{w\in[0,\upsilon]} \{c(t,x,w) + \Lambda^w_t V(t,x)\}. \quad (7)$$

Proof. See [3] Theorem 2.4. □

This standard result is very powerful when we want to solve an optimal control problem driven by a stochastic differential equation. However, when considering a controlled continuous-time Markov chain with binary state, the HJB equation does not seem to be so useful. In the next section, we present an approach that allows us to use all the information of the HJB equation to find an optimal control.

4 From HJB to a backward ODE

This is the main section of this work. At this point, we show how the HJB equation can be replaced by an ordinary backward differential equation. The idea is to use the algebraic difference between the evaluation of the HJB equation in both states of the system to obtain useful information to construct an optimal feedback control. Thus, the HJB equation, which seemed hardly treatable, allows us to obtain an ordinary differential equation that must be solved backwards. Henceforth, we use the discrete gradient notation: for all $x \in S$ we define

$$\nabla_x f(x) := f(-x) - f(x).$$

Now we have all the necessary information to enunciate and prove the following result.

Theorem 2 (Backward ODE)**.** *Assume that the optimal control problem presented in the previous sections has the following formulation:*

- $\ell(x,u) = \alpha(x) + \beta u$*, with* $\alpha(x) \geq 0$ *for all* $x \in S$ *and* $\beta > 0$*;*

- $c(t,x,u) = \gamma(t)x + \kappa(t)u^2/2$ *, with* $\gamma(\cdot), \kappa(\cdot) \in C^0([0,T],\mathbb{R})$*, and* $\kappa(t) > 0$.

Moreover, suppose that we can solve the backward ODE:

$$\begin{cases} \dot{z}(t) = 2\gamma(t) - \nabla_x \alpha(1) + \frac{\beta^2}{2\kappa(t)} |z(t)| z(t) \\ z(T) = \nabla_x g(1) \end{cases}$$

and, finally, assume that the solution $z(t)$ *of this ODE satisfies the inequality:*

$$[z(t)]^- \leq \frac{\upsilon \kappa(t)}{\beta},$$

so that the control constraint is inactive. Therefore, the optimal feedback control is

$$u^*(t,x) = \frac{\beta}{\kappa(t)} [z(t)x]^-.$$

Proof. Under the assumptions, the HJB equation becomes:

$$\partial_t V(t,x) + \min_{w \in [0,\upsilon]} \left\{ \gamma(t)x + \frac{\kappa(t)w^2}{2} + \alpha(x) + \beta w \nabla_x V(t,x) \right\} = 0.$$

Minimising we get

$$w = \min \left\{ \frac{\beta}{\kappa(t)} [(\nabla_x V(t,x))]^-, \upsilon \right\}.$$

Since $\frac{\beta}{\kappa(t)} [(\nabla_x V(t,x))]^- \leq \upsilon$ by the above assumption, we can rewrite the HJB equation as

$$\partial_t V(t,x) + \gamma(t)x + \frac{\beta^2}{\kappa(t)} \left\{ \frac{1}{2} ([(\nabla_x V(t,x))]^-)^2 \right\} + \alpha(x) + \frac{\beta^2}{\kappa(t)} \left\{ [(\nabla_x V(t,x))]^- (\nabla_x V(t,x)) \right\} = 0.$$

Let us introduce a new key variable

$$z(t) := \nabla_x V(t,1).$$

Now we rewrite the previous HJB equation for $x = 1$ and for $x = -1$; we obtain the system:

$$\begin{cases} \partial_t V(t,1) + \gamma(t) + \alpha(1) + \\ \qquad + \frac{\beta^2}{\kappa(t)} \left\{ \frac{1}{2} ([z(t)]^-)^2 + [z(t)]^- z(t) \right\} = 0 \\ \partial_t V(t,-1) - \gamma(t) + \alpha(-1) + \\ \qquad + \frac{\beta^2}{\kappa(t)} \left\{ \frac{1}{2} ([-z(t)]^-)^2 - [-z(t)]^- z(t) \right\} = 0. \end{cases}$$

Recalling that, for all $z \in \mathbb{R}$, we have that $[-z]^- = [z]^+$, and, taking the difference between the two equations, we get

$$\dot{z}(t) - 2\gamma(t) + \nabla_x \alpha(1) + \frac{\beta^2}{\kappa(t)} \left\{ \frac{1}{2} \left(([z(t)]^+)^2 - ([z(t)]^-)^2 \right) \right\} + \frac{\beta^2}{\kappa(t)} \left\{ -z(t) \left([z(t)]^+ + [z(t)]^- \right) \right\} = 0.$$

Finally, for all $z \in \mathbb{R}$ it holds $[z]^+ + [z]^- = |z|$ and $[z]^+ - [z]^- = z$; hence we obtain:

$$\dot{z}(t) = 2\gamma(t) - \nabla_x \alpha(1) + \frac{\beta^2}{2\kappa(t)} |z(t)| z(t).$$

The final condition $V(T,x) = g(x)$ for all $x \in S$ gives us $z(T) = \nabla_x V(T,1) = \nabla_x g(1)$. Therefore, HJB becomes

$$\begin{cases} \dot{z}(t) = 2\gamma(t) - \nabla_x \alpha(1) + \frac{\beta^2}{2\kappa(t)} |z(t)| z(t) \\ z(T) = \nabla_x g(1), \end{cases}$$

which is exactly the backward ODE we are looking for. □

5 A numerical example

We formally define the dynamics of the process with $x \in S$ as follows:

$$\mathbb{P}(X_{t+h} = -x | X_t = x, U_t = u) = = (1 + x + u)h + o(h). \quad (8)$$

We want to choose a feasible feedback control function u which takes values in $[0,2]$, in order to minimise

$$J^u_{0,x} := \mathbb{E}^u_{0,x} \left\{ \int_0^1 \frac{-X_t}{2} + \frac{U_t^2}{4} dt + \frac{X_T}{4} \right\};$$

then, the backward ODE is

$$\begin{cases} \dot{z}(t) = 1 + |z(t)| z(t) \\ z(1) = -1/2. \end{cases}$$

In a left neighborhood of 1, for example in $(1-\varepsilon, 1]$ with $\varepsilon > 0$, the ODE becomes

$$\begin{cases} \dot{z}(t) = 1 - (z(t))^2 \\ z(1) = -1/2\,. \end{cases}$$

The solution to the previous Cauchy problem is

$$z(t) = 1 - \frac{6e^2}{e^{2t} + 3e^2}\,.$$

We observe that, for $t \in [0,1]$, this function is always negative (i.e., $\varepsilon > 1$); moreover, $[z(t)]^- \leq -z(0) = (1-3e^3)/(1+3e^2) < 1$, because the function is strictly negative and strictly increasing. Hence, the constraint on the control $u \in [0,2]$ is not active and the optimal feedback is feasible. Using the optimal feedback control

$$u^*(t,x) = \left[\frac{e^{2t}-3e^2}{e^{2t}+3e^2}\cdot x\right]^- ,$$

we can find the evolution of the expected value of the process. Let us define $m_x^*(t) := \mathbb{E}_{0,x}^{u^*}(X_t)$; using the infinitesimal generator we obtain

$$\dot{m}_x^*(t) = -2\mathbb{E}_{0,x}^{u^*}\left\{X_t\left(1+X_t+\left[\frac{e^{2t}-3e^2}{e^{2t}+3e^2}\cdot X_t\right]^-\right)\right\} ,$$

which becomes

$$\dot{m}_x^*(t) = \left(\frac{e^{2t}-3e^2}{e^{2t}+3e^2}-2\right)(m_x^*(t)+1) \qquad (9)$$

with initial condition:

$$m_x^*(0) = \mathbb{E}_{0,x}^{u^*}(X_0) = x\,.$$

Equation (9) is a linear ODE whose coefficients are continuous functions for all $t \in [0,1]$; therefore, there exists a unique solution $m_x^*(t)$ to the previous Cauchy problem. For $x = -1$ the solution is constant: $m_{-1}^*(t) \equiv -1$. On the other hand, for $x = 1$ we can explicitly find the analytical form of this function, but it is long and inexpressive. We prefer to plot its graph in Figure 1. Moreover, using the evolution of the expected value of the optimal process starting from $x = 1$, we can find the evolution of the probability of each state: $p(t) := \mathbb{P}_{0,1}^{u^*}(X_t = 1) = (1+m_1^*(t))/2$. The probability is displayed in Figure 1.

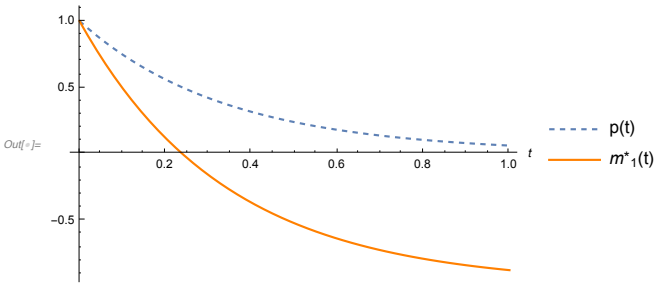

Figure 1: $p(t)$ and $m_1^*(t)$.

We notice that, using the optimal feedback control, the process moves towards the state -1. When the initial position is -1, the process remains in this state; otherwise, when the initial position is 1, the process changes its state with a strictly positive probability rate.

6 Conclusion

In this paper, we analyse the evolution of a continuous-time Markov decision process characterised by a binary state. We introduce the standard Hamilton–Jacobi–Bellman equation and prove that, under suitable analytical formulations of the rate of transition and of the cost function, we can replace the HJB equation with a backward ODE. Then, all information useful to characterise an optimal feedback control is now contained in the solution of the backward ODE. Using a numerical example, we show how to find an optimal feedback control through the results shown in this paper.

This research can be improved in various directions. First, we can try to extend the family of functions that satisfies the hypotheses of Theorem 2. Subsequently, we can investigate whether what is proven in this paper is valid for analogous problems with an infinite-time horizon, too. Finally, it may be interesting to study whether this approach is useful for analysing the interaction between multiple players.

References:

[1] A. Cecchin, P. Dai Pra, M. Fischer & G. Pelino, On the Convergence Problem in Mean Field Games: A Two State Model without Uniqueness, *SIAM Journal on Control and Optimization*, Vol. 57, No. 4, 2019, pp. 2443-2466.

[2] P. Dai Pra, E. Sartori & M. Tolotti, Climb on the Bandwagon: Consensus and Periodicity in a Lifetime Utility Model with Strategic Interactions, *Dynamic Games and Applications*, Vol. 9, No. 4, 2019, pp. 1061-1075.

[3] M.K. Ghosh & S. Saha, Risk-sensitive Control of Continuous-time Markov Chains, *Stochastics*, Vol. 86, No.4, 2014, pp. 655-675.

[4] X. Guo & O. Hernández-Lerma, *Continuous-Time Markov Decision Processes*, Springer Berlin Heidelberg, 2009.

[5] M. Kochmański, T. Paszkiewicz, & S. Wolski, Curie–Weiss magnet — A Simple Model of Phase Transition, *European Journal of Physics*, Vol.34, No.6, 2013, pp. 1555-1578.

[6] W. Wang, X. Su, S. Gan, & L. Qian, Pricing Vulnerable European Options Under a Markov-Modulated Jump Diffusion Process, *WSEAS Transactions on Mathematics*, Vol.16, 2017, pp. 123-132.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)
The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
No funding was received for conducting this study.

Conflict of Interest
The authors have no conflicts of interest to declare that are relevant to the content of this article.