# Foundations of Reinforcement Learning with Applications in Finance

Ashwin Rao, Tikhon Jelvis

# 1 The Hamilton-Jacobi-Bellman (HJB) Equation

In this Appendix, we provide a quick coverage of the Hamilton-Jacobi-Bellman (HJB) Equation, which is the continuous-time version of the Bellman Optimality Equation. Although much of this book covers Markov Decision Processes in a discrete-time setting, we do cover some classical Mathematical Finance Stochastic Control formulations in continuous-time. To understand these formulations, one must first understand the HJB Equation, which is the purpose of this Appendix. As is the norm in the Appendices in this book, we will compromise on some of the rigor and emphasize the intuition to develop basic familiarity with HJB.

## 1.1 HJB as a continuous-time version of Bellman Optimality Equation

In order to develop the continuous-time setting, we shall consider a (not necessarily time-homogeneous) process where the set of states at time $t$ are denoted as $\mathcal{S}_t$ and the set of allowable actions for each state at time $t$ are denoted as $\mathcal{A}_t$. Since time is continuous, Rewards are represented as a *Reward Rate* function $\mathcal{R}$ such that for any state $s_t \in \mathcal{S}_t$ and for any action $a_t \in \mathcal{A}_t$, $\mathcal{R}(t, s_t, a_t) \cdot dt$ is the *Expected Reward* in the time interval $(t, t + dt]$, conditional on state $s_t$ and action $a_t$ (note the functional dependency of $\mathcal{R}$ on $t$ since we will be integrating $\mathcal{R}$ over time). Instead of the discount factor $\gamma$ as in the case of discrete-time MDPs, here we employ a *discount rate* (akin to interest-rate discounting) $\rho \in \mathbb{R}_{\geq 0}$ so that the discount factor over any time interval $(t, t + dt]$ is $e^{-\rho \cdot dt}$.

We denote the Optimal Value Function as $V^*$ such that the Optimal Value for state $s_t \in \mathcal{S}_t$ at time $t$ is $V^*(t, s_t)$. Note that unlike Section **??** in Chapter **??** where we denoted the Optimal Value Function as a time-indexed sequence $V_t^*(s_t)$, here we make $t$ an explicit functional argument of $V^*$. This is because in the continuous-time setting, we are interested in the time-differential of the Optimal Value Function.

Now let us write the Bellman Optimality Equation in it's continuous-time version, i.e, let us consider the process $V^*$ over the time interval $(t, t + dt]$ as follows:

$$V^*(t, s_t) = \max_{a_t \in \mathcal{A}_t} \{\mathcal{R}(t, s_t, a_t) \cdot dt + \mathbb{E}_{(t,s_t,a_t)}[e^{-\rho \cdot dt} \cdot V^*(t + dt, s_{t+dt})]\}$$

Multiplying throughout by $e^{-\rho t}$ and re-arranging, we get:

$$\max_{a_t \in \mathcal{A}_t} \{e^{-\rho t} \cdot \mathcal{R}(t, s_t, a_t) \cdot dt + \mathbb{E}_{(t,s_t,a_t)}[e^{-\rho(t+dt)} \cdot V^*(t + dt, s_{t+dt}) - e^{-\rho t} \cdot V^*(t, s_t)]\} = 0$$

$$\Rightarrow \max_{a_t \in \mathcal{A}_t} \{e^{-\rho t} \cdot \mathcal{R}(t, s_t, a_t) \cdot dt + \mathbb{E}_{(t,s_t,a_t)}[d\{e^{-\rho t} \cdot V^*(t, s_t)\}]\} = 0$$

$$\Rightarrow \max_{a_t \in \mathcal{A}_t} \{e^{-\rho t} \cdot \mathcal{R}(t, s_t, a_t) \cdot dt + \mathbb{E}_{(t,s_t,a_t)}[e^{-\rho t} \cdot (dV^*(t, s_t) - \rho \cdot V^*(t, s_t) \cdot dt)]\} = 0$$

Multiplying throughout by $e^{\rho t}$ and re-arranging, we get:

$$\rho \cdot V^*(t, s_t) \cdot dt = \max_{a_t \in \mathcal{A}_t} \{\mathbb{E}_{(t,s_t,a_t)}[dV^*(t, s_t)] + \mathcal{R}(t, s_t, a_t) \cdot dt\} \tag{1.1}$$

For a finite-horizon problem terminating at time $T$, the above equation is subject to terminal condition:

$$V^*(T, s_T) = \mathcal{T}(s_T)$$

for some terminal reward function $\mathcal{T}(\cdot)$.

Equation (1.1) is known as the Hamilton-Jacobi-Bellman Equation - the continuous-time analog of the Bellman Optimality Equation. In the literature, it is often written in a more compact form that essentially takes the above form and "divides throughout by dt." This requires a few technical details involving the stochastic differentiation operator. To keep things simple, we shall stick to the HJB formulation of Equation (1.1).

## 1.2 HJB with State Transitions as an Ito Process

Although we have expressed the HJB Equation for $V^*$, we cannot do anything useful with it unless we know the state transition probabilities (all of which are buried inside the calculation of $\mathbb{E}_{(t,s_t,a_t)}[\cdot]$ in the HJB Equation). In continuous-time, the state transition probabilities are modeled as a stochastic process for states (or of it's features). Let us assume that states are real-valued vectors, i.e, state $\boldsymbol{s}_t \in \mathbb{R}^n$ at any time $t \geq 0$ and that the transitions for $\boldsymbol{s}$ are given by an Ito process, as follows:

$$d\boldsymbol{s}_t = \boldsymbol{\mu}(t, \boldsymbol{s}_t, a_t) \cdot dt + \boldsymbol{\sigma}(t, \boldsymbol{s}_t, a_t) \cdot d\boldsymbol{z}_t$$

where the function $\boldsymbol{\mu}$ (drift function) gives an $\mathbb{R}^n$ valued process, the function $\boldsymbol{\sigma}$ (dispersion function) gives an $\mathbb{R}^{n \times m}$-valued process and $\boldsymbol{z}$ is an $m$-dimensional process consisting of $m$ independent standard brownian motions.

Now we can apply multivariate Ito's Lemma (Equation (**??**) from Appendix **??**) for $V^*$ as a function of $t$ and $s_t$ (we lighten notation by writing $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ instead of $\boldsymbol{\mu}(t, \boldsymbol{s}_t, a_t)$ and $\boldsymbol{\sigma}(t, \boldsymbol{s}_t, a_t)$):

$$dV^*(t, \boldsymbol{s}_t) = (\frac{\partial V^*}{\partial t} + (\nabla_{\boldsymbol{s}} V^*)^T \cdot \boldsymbol{\mu}_t + \frac{1}{2} Tr[\boldsymbol{\sigma}_t^T \cdot (\Delta_{\boldsymbol{s}} V^*) \cdot \boldsymbol{\sigma}_t]) \cdot dt + (\nabla_{\boldsymbol{s}} V^*)^T \cdot \boldsymbol{\sigma}_t \cdot d\boldsymbol{z}_t$$

Substituting this expression for $dV^*(t, \boldsymbol{s}_t)$ in Equation (1.1), noting that

$$\mathbb{E}_{(t,\boldsymbol{s}_t,a_t)}[(\nabla_{\boldsymbol{s}} V^*)^T \cdot \boldsymbol{\sigma}_t \cdot d\boldsymbol{z}_t] = 0$$

and dividing throughout by $dt$, we get:

$$\rho \cdot V^*(t, \boldsymbol{s}_t) = \max_{a_t \in \mathcal{A}_t} \{\frac{\partial V^*}{\partial t} + (\nabla_{\boldsymbol{s}} V^*)^T \cdot \boldsymbol{\mu}_t + \frac{1}{2} Tr[\boldsymbol{\sigma}_t^T \cdot (\Delta_{\boldsymbol{s}} V^*) \cdot \boldsymbol{\sigma}_t] + \mathcal{R}(t, \boldsymbol{s}_t, a_t)\} \tag{1.2}$$

For a finite-horizon problem terminating at time $T$, the above equation is subject to terminal condition:

$$V^*(T, \boldsymbol{s}_T) = \mathcal{T}(\boldsymbol{s}_T)$$

for some terminal reward function $\mathcal{T}(\cdot)$.