Stochastic Optimal Control for Nonlinear Markov Jump Diffusion Processes

Evangelos A. Theodorou and Emmanuel Todorov

Abstract— **We consider the problem finite horizon stochastic optimal control for nonlinear markov jump diffusion processes. In particular, by using stochastic calculus for markov jump diffusions processes and the logarithmic transformation of the value function we demonstrate the transformation of the corresponding Hamilton-Jacobi-Bellman (HJB) Partial Differential Equation (PDE) to the backward Chapman Kolmogorov PDE for jump diffusions. Furthermore we derive the Feynman-Kac lemma for nonlinear markov jump diffusions processes and apply it to the transformed HJB equation. Application of the Feynman-Kac lemma yields the solution of the transformed HJB equation. The path integral interpretation is derived. Finally, conclusions and future directions are discussed.**

I. INTRODUCTION

Nonlinear stochastic optimal control theory [1], [2], [3] is one of the most fundamental control theoretic frameworks with a plethora of applications in domains that span from biology [4], [5] and neuroscience [6] to vehicle and mobile robot control [7]. The nonlinear and stochastic nature of most dynamical systems in engineering and biology results in the broad applicability of stochastic nonlinear optimal control framework.

Despite and progress in terms and theory and applications of stochastic optimal control, there are still open theoretical and algorithmic questions as to weather or not stochastic optimal control, under different than brownian noise profiles, is feasible. Stochastic models that incorporate brownian and Poisson distributed noise offer a suitable description of phenomena in which sudden changes in state may occur. The source of these changes varies with the dynamical system under consideration. For example, in case of neuromuscular systems stochasticity comes from noisy neural commands in which the neural firing rate is Poisson distributed. In humanoid and mobile robotics randomness may be caused due to noise in the readings of proprioceptive sensors such as odometers, gyros, accelerometers etc. Very often these readings include sudden changes due to contact phenomena with the environment.

Even when only brownian noise is considered, one of the main issues with stochastic optimal control is that its solution requires the solution of a nonlinear and second order partial differential equation, the so called Hamilton Jacobi Bellman equation [1], [8]. How to solve such partial differential equation especially for high dimensional state space models is still an open research problem. The challenges in solving this PDE have limited the use of stochastic optimal control to low dimensional control problems.

Recently there has been a number of studies [9], [10], [11], [12], [13] and [14] which have shown promising results in terms of efficiency applicability and robustness of the proposed path integral control framework to high dimensional stochastic optimal control problems. In particular, in [10], [11] the path integral control framework was first introduced and its application to symmetry breaking phenomena was investigated. In [9], the path integral approach was applied to the case of multi-agent optimal control problems. An alternative formulation of stochastic optimal control in discrete time was presented in [12]. An advantage of this work is that the control cost is defined as the Kullback-Leibler divergence between the state transition probabilities of the controlled and uncontrolled dynamics. This component allows for the use of, more general than quadratic, control cost functions. In [13], [15] the path integral control framework was generalized for the case of nonlinear diffusions processes with state depended control and diffusion matrices. In addition, an iterative version of the path integral control framework capable of scaling to high dimensional learning control problems, the so called Policy Improvement with Path Integral (PI^2) was presented. In [16], [15], [17], [18], [19] there has been a number of applications of PI^2 to learning robotic control. These applications include planning and gain scheduling for tasks such as grasping, reaching, manipulating objects and jumping with humanoid, quadruped and manipulator robotic systems.

The path integral approach solves nonlinear stochastic optimal control problems with forward sampling of diffusions processes. The main rational of this framework is that under 1) an assumption between the weight in the control cost and the variance of the noise and 2) the logarithmic transformation of the value function, the HJB equation is transformed into a linear and second order PDE. The linear PDE corresponds to the backward Chapman-Kolmogorov PDE. Solutions of the linear PDE can be found via the use of the Feynman-Kac lemma. The Feynman-Kac lemma creates a bridge between PDE and Stochastic Differential Equations(SDEs) and its use can be twofold. On one side, it can be use to find solution of PDE with forward sampling of SDE. On the other side it can be used to find solution of SDE by deterministically solving the corresponding PDE. The approach of using the Feynman- Kac lemma for solving PDEs is very promising especially in cases where an initial policy is considered and successive policy improvements are

E.A. Theodorou is Postdoctoral Research Associate with the Department of Computer Science and Engineering, University of Washington, Seattle, USA etheodor@cs.washington.edu

E. Todorov is Associate Professor with the Department of Computer Science and Engineering and Department of Applied Math, University of Washington, Seattle, USA todorov@cs.washington.edu

performed with successive application of the Feynman-Kac lemma.

In this work, we follow the rational of path integral control for diffusions processes in [13], [15], [14] but this time we consider stochastic optimal control for markov jump diffusion processes. More precisely, in Section II we review important properties of Poisson processes. In Section III we provide the HJB equation for the case of markov jump diffusion processes and demonstrate its transformation to a linear PDE. This linear PDE is the so called backward Chapman Kolmogorov PDE modified for the case of markov jump diffusions. In section IV we derive the Feynman-Kac lemma for markov jump diffusions and apply it to the transformed HJB PDE. In Section V we discuss the path integral formulation of the solution of the transformed PDE. Finally in the last Section VI, we conclude and discuss future directions.

II. Elements of Poisson stochastic calculus

In this section we present [20], [21] fundamental elements of the Poisson stochastic calculus. More precisely, let $\mathbf{P}(t)$ be a m-dimensional Poisson vector process with the differential having the common mean and variance:

$$E\left(dP_i(t)\right) = \mu_i dt, \quad \text{Var}\left(dP_i(t)\right) = \mu_i dt, \quad \text{for } i = 1, ..., m \tag{1}$$

where $\mu_i(t) > 0$ is the ith jump rate or jump density and $\mu_i dt$ is the mean count of the ith Poisson process in the time interval $(t, t + dt]$. Poisson processes obey the Markov property while they also have independent increments. Thus:

$$\text{Cov}\left[dP_i(t_j)dP_i(t_k)\right] = \text{Var}\left[dP_i(t_j)\,\delta_{k,j} = \mu_i(t_j)dt\delta_{k,j}\right.$$

where $\delta_{k,j}$ is the Kronecker delta. If s and t are continuous arguments then:

$$\text{Cov}\left[dP_i(s)dP_i(t)\right] = \mu_i(t_j)dt\delta(t-s)ds \tag{2}$$

For the Poisson differential vector $d\mathbf{P}$ we have that $\text{Var}\left[d\mathbf{P}\right] = \mathbf{Diag}\left(\mu_1, ..., \mu_m\right)$. In case where the Poisson increments are not independent then $\text{Var}\left[d\mathbf{P}\right] = \mathbf{\Sigma_p} dt$. The processes P_i, dP_i are all Poisson distributed and therefore:

$$\text{Prob}\Big(P_i(t) = k\Big) = \exp(-\nu_i)\frac{\nu_i^k}{k!} \tag{3}$$

$$\text{Prob}\Big(dP_i(t) = k\Big) = \exp(-\mu_i)\frac{(\mu_i dt)^k}{k!} \tag{4}$$

Poisson distribution takes a simplified form for the differential $dP_i(t)$ if one use $dt-$precision. In this case the Poisson distribution specifies the *Zero-One Law*(ZOL) for jumps of $dP_i(t)$. In mathematical terms we have:

$$\text{Prob}\Big(dP_i(t) = k\Big) = \Big(1 - \mu_i(t)dt\Big)\delta_{k,0} + \mu_i(t)dt\delta_{k,1} \tag{5}$$

This is a special case in which Poisson distribution reduces to bernoulli distribution since there are only two possible events of zero and one jump. A consequence of ZOL is that:

$$\Big\langle (dP_i(t))^\nu \Big\rangle = \Big(1 - \mu_i(t)dt\Big)\cdot 0^\nu + \mu_i(t)dt \cdot 1^\nu = \mu_i(t)dt \tag{6}$$

The powers of Poisson distribution do not truncate to a finite number, in contradiction to wiener differentials which truncate at the second order and thus contributing derivatives up to second order. The Poisson differential contributes derivatives of all orders, usually represented as functional integrals or delayed arguments that cause global dependence rather than local dependence of partial derivatives of finite order.

III. Stochastic Optimal Control for Markov Jump Diffusion Processes

We consider the stochastic optimal control problem with the objective function under minimization, expressed as follows:

$$V(\mathbf{x}, t_0) = \min_{\mathbf{u}} J(\mathbf{u}, \mathbf{x}) = \min_{\mathbf{u}} \Big\langle \phi(\mathbf{x}_{t_N}) + \int_{t_0}^{t_N} \boldsymbol{\mathcal{L}}(\mathbf{x}, \mathbf{u}) dt \Big\rangle \tag{7}$$

where $\boldsymbol{\mathcal{L}}(\mathbf{x}, \mathbf{u}) = q(\mathbf{x}) + \frac{1}{2}\mathbf{u}^T\mathbf{R}\mathbf{u}$ is the running cost accumulated during the time horizon $\Delta T = t_N - t_0$ and with t_0 being the starting time and t_N the end time. An essential assumption in this work is that the running cost is quadratic with respect to controls. More general formulations of the running cost which include quadratic terms and linear terms in $\mathbf{u}$ could be considered but we do not show these generalizations here. The term $\phi(\mathbf{x}_{t_N})$ in the cost function (7) is the terminal cost with $V(\mathbf{x}, t_N) = \phi(\mathbf{x}_{t_N})$. The minimization is subject to the dynamical constrains:

$$\mathbf{dx} = \left(\mathbf{f}(\mathbf{x}, t) + \mathbf{G}(\mathbf{x}, t)\mathbf{u}\right) dt + \mathbf{B}(\mathbf{x}, t)d\mathbf{w}(t) + \mathbf{h}(\mathbf{x}, t)d\mathbf{P}(t) \tag{8}$$

with $\mathbf{x}_t \in \Re^{n\times 1}$ denoting the state of the system, $\mathbf{G}(\mathbf{x}, t) : \Re^n \times \Re \to \Re^{n\times p}$ the control matrix, $\mathbf{B}(\mathbf{x}, t) : \Re^n \times \Re \to \Re^{n\times p}$ is the diffusions matrix $\mathbf{f}(\mathbf{x}, t) : \Re^n \times \Re \to \Re^n$ the passive dynamics, $\mathbf{u}_t \in \Re^{p\times 1}$ the control vector and $d\mathbf{w} \in \Re^{p\times 1}$ brownian noise. The term $\mathbf{P}(t) \in \Re^{m\times 1}$ is Poisson distributed and $\mathbf{h}(\mathbf{x}, t) : \Re^n \times \Re \to \Re^{n\times m}$ is the jump-amplitude or the Poisson process coefficient. We assume that $\mathbf{G}(\mathbf{x}, t), \mathbf{B}(\mathbf{x}, t), \mathbf{f}(\mathbf{x}, t)$ and $\mathbf{h}(\mathbf{x}, t) \in \boldsymbol{\mathcal{C}}^1$. The conditional expectation of the state differential process is expressed as:

$$E\left(d\mathbf{x}|\mathbf{x}(t)\right) = \Big(\mathbf{f}(\mathbf{x}, t) + \mathbf{G}(\mathbf{x}, t)\mathbf{u} + h(\mathbf{x}, t)\mu(t)\Big)dt \tag{9}$$

where $\mu(t) \in \Re^{m\times 1}$ is the jump-rate vector. The conditional covariance is given as:

$$\text{Cov}\Big(d\mathbf{x}\; d\mathbf{x}^T|\mathbf{x}(t)\Big) = \Big(\mathbf{B}(\mathbf{x}, t)\mathbf{B}(\mathbf{x}, t)^T + \mathbf{h}(\mathbf{x}, t)\mathbf{\Sigma}_p\mathbf{h}(\mathbf{x}, t)^T\Big)dt$$

where $\mathbf{\Sigma}_p = \mathbf{diag}\,(\mu_1(t), ..., \mu_m(t))$ is the Poisson covariance matrix for the case where the Poisson process has undepended increments. The stochastic Bellman principle of optimality yields:

$$V(\mathbf{x},t) = \min_{\mathbf{u}[\mathbf{x},(t,t+\Delta t)]} \left\langle \int_t^{t+\Delta t} \boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},\tau)d\tau + V(\mathbf{x},t+\Delta t) \right\rangle \quad (10)$$

The HJB equation for the case of markov jump diffusions [20] of the form of equation (8) is expressed as follows:

$$\begin{aligned} -\frac{\partial V(\mathbf{x},t)}{\partial t} &= \min_{\mathbf{u}[\mathbf{x},(t,t+dt)]} \Bigg(\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},\tau) \\ &+ \nabla_{\mathbf{x}} V(\mathbf{x},t)^T \Big(\mathbf{f}(\mathbf{x},t) + \mathbf{G}(\mathbf{x},t)\mathbf{u} \Big) \\ &+ \frac{1}{2} tr\Big(\nabla_{\mathbf{xx}} V(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T \Big) \\ &+ \sum_{k=1}^{m} \mu_k(t) \Big[V\Big(\mathbf{x}(t) + \mathbf{h}_k(\mathbf{x}(t),t), t \Big) - V\Big(\mathbf{x}(t), t \Big) \Big] \Bigg) \end{aligned} \quad (11)$$

The optimal control based on the minimization above is given as:

$$\mathbf{u}(\mathbf{x},t) = -\mathbf{R}^{-1}\mathbf{G}(\mathbf{x},t)\ \nabla_{\mathbf{x}} V(\mathbf{x},t) \quad (12)$$

Substitution of the optimal control above into (11) yields the following PDE:

$$\begin{aligned} -\frac{\partial V(\mathbf{x},t)}{\partial t} &= q(\mathbf{x},t) + \nabla_{\mathbf{x}} V(\mathbf{x},t)^T \mathbf{f}(\mathbf{x},t) \\ &- \frac{1}{2} \nabla_{\mathbf{x}} V(\mathbf{x},t)^T \mathbf{G}(\mathbf{x},t)\mathbf{R}^{-1}\mathbf{G}(\mathbf{x},t)^T \nabla_{\mathbf{x}} V(\mathbf{x},t) \\ &+ \frac{1}{2} tr\left((\nabla_{\mathbf{xx}} V(\mathbf{x},t))\mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T \right) \\ &+ \sum_{k=1}^{m} \mu_k(t) \Big[V\Big(\mathbf{x}(t) + \mathbf{h}_k(\mathbf{x}(t),t), t \Big) - V\Big(\mathbf{x}(t), t \Big) \Big] \Big) \end{aligned} \quad (13)$$

The PDE above is second order and nonlinear. Moreover, in comparison to the HJB PDE for diffusions processes, equation (13) incorporates an additional term that corresponds to the jump $\mathbf{h}_k(\mathbf{x}(t),t)$ of the SDE in (8). In fact in case where no Poisson noise is incorporated $\mu(t) = 0$ equation (13) collapses to the HJB for diffusion processes. We follow the rational in [13], [15], [14] and therefore we apply the logarithmic transformation $V(\mathbf{x},t) = -\lambda \log \Psi(\mathbf{x},t)$ to the PDE in (13) and assume that there is a connection between control cost and noise expressed as: $\lambda \mathbf{G}(\mathbf{x},t)\mathbf{R}^{-1}\mathbf{G}(\mathbf{x},t)^T = \mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T$. The intuition for the last assumption is that since the term $\mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T$ corresponds to the variance of the brownian noise high variance means low weight in the control cost the therefore "cheap" controls. Similarly, low variance is equivalent to high weight in the control cost and therefore "expensive" controls. High variability leads to an increased control authority while low variability has the effect of reducing control authority.

It is evident that the strength of the stochastic disturbances determines how much control authority is required for the system such that it can optimally perform the task. Therefore, for the cases where stochastic disturbances have high variability, the control cost is low and thus larger control commands are available. For cases where $\mathbf{G}(\mathbf{x},t) = \mathbf{B}(\mathbf{x},t)$ the control weight is specified as $\mathbf{R} = \frac{1}{\lambda} I$. Next, we find all the partial derivatives and the difference term in (13) as a function of the new value function $\Psi(\mathbf{x},t)$. More precisely we will have that:

$$\begin{aligned} \frac{\partial V(\mathbf{x},t)}{\partial t} &= -\lambda \frac{1}{\Psi(\mathbf{x},t)} \frac{\partial \Psi(\mathbf{x},t)}{\partial t} \\ \nabla_{\mathbf{x}} V(\mathbf{x},t) &= -\lambda \frac{1}{\Psi(\mathbf{x},t)} \nabla_{\mathbf{x}} \Psi(\mathbf{x},t) \\ \nabla_{\mathbf{xx}} V(\mathbf{x},t) &= \lambda \frac{1}{\Psi_t^2} \nabla_{\mathbf{x}} \Psi(\mathbf{x},t)\ \nabla_{\mathbf{x}} \Psi(\mathbf{x},t)^T \\ &\quad - \lambda \frac{1}{\Psi(\mathbf{x},t)} \nabla_{\mathbf{xx}} \Psi(\mathbf{x},t) \\ d_{jump} V(\mathbf{x},t) &= \frac{\partial V(\mathbf{x},t)}{\partial \Psi(\mathbf{x},t)}\ d_{jump} \Psi(\mathbf{x},t) \\ &= -\frac{\lambda}{\Psi(\mathbf{x},t)}\ d_{jump} \Psi(\mathbf{x},t) \end{aligned} \quad (14)$$

where the terms $d_{jump} V(\mathbf{x},t)$ and $d_{jump} \Psi(\mathbf{x},t)$ are defined as:

$$d_{jump} V = \sum_{k=1}^{m} \mu_k \Big[V\Big(\mathbf{x}(t) + \mathbf{h}_k(\mathbf{x}(t),t), t \Big) - V\Big(\mathbf{x}(t), t \Big) \Big] \quad (15)$$

and:

$$d_{jump} \Psi = \sum_{k=1}^{m} \mu_k \Big[\Psi\Big(\mathbf{x}(t) + \mathbf{h}_k(\mathbf{x}(t),t), t \Big) - \Psi\Big(\mathbf{x}(t), t \Big) \Big] \quad (16)$$

By applying the equalities (14) above and under the assumption $\lambda \mathbf{G}(\mathbf{x},t)\mathbf{R}^{-1}\mathbf{G}(\mathbf{x},t)^T = \mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T$ we have the resulting PDE:

$$\begin{aligned} -\frac{\partial_t \Psi(\mathbf{x},t)}{\partial t} &= -\frac{1}{\lambda} q(\mathbf{x},t)\Psi(\mathbf{x},t) + \mathbf{f}(\mathbf{x},t)^T\ \nabla_{\mathbf{x}} \Psi(\mathbf{x},t) \\ &\quad + \frac{1}{2} tr\Big((\nabla_{\mathbf{xx}} \Psi(\mathbf{x},t))\mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T \Big) \\ &\quad + d_{jump} \Psi(\mathbf{x},t) \end{aligned}$$

with the terminal condition $\Psi(\mathbf{x},t_N) = \exp\left(-\frac{1}{\lambda}\phi(\mathbf{x}(t_N))\right)$. By using the differential operator $\boldsymbol{\mathcal{D}}$ defined as:

$$\boldsymbol{\mathcal{D}}\Psi(\mathbf{x},t) = -\frac{1}{\lambda}q(\mathbf{x},t)\Psi(\mathbf{x},t) + \mathbf{f}(\mathbf{x},t)^T\ \nabla_{\mathbf{x}}\Psi(\mathbf{x},t) + \frac{1}{2}tr\Big((\nabla_{\mathbf{xx}}\Psi(\mathbf{x},t))\mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T\Big) + d_{jump}\Psi(\mathbf{x},t) \tag{17}$$

we can write the PDE as follows:

$$-\frac{\partial\Psi(\mathbf{x},t)}{\partial t} = \boldsymbol{\mathcal{D}}\Psi(\mathbf{x},t) \quad \text{or} \quad \frac{\partial\Psi(\mathbf{x},t)}{\partial t} + \boldsymbol{\mathcal{D}}\Psi(\mathbf{x},t) = 0 \tag{18}$$

The initial nonlinear HJB equation in $V(\mathbf{x},t)$ is transformed into the linear PDE in $\Psi(\mathbf{x},t)$ which corresponds to the backward Chapman Kolmogorov PDE for nonlinear jump diffusions. When $\mu(t) = 0$ equation (18) corresponds to the Chapman Kolmogorov PDE for diffusion proccesses. The solution of the linear PDE in (18) is found with the application of the Feynman - Kac lemma extended to markov jump diffusion processes. In particular, application of the Feynman-Kac lemma results in the numerical process of computing $\Psi(\mathbf{x})$ with forward sampling of the uncontrolled jump diffusion dynamics and evaluation of the expectation of the exponential of state dependent $q(\mathbf{x})$ cost function on the trajectories generated by the forward sampling process. This process will become clear in the next section in which Feynman-Kac lemma for markov jump diffusions is derived.

Finally, the optimal control law as function of the value function $V(\mathbf{x},t)$ is given in (12), while as a function of the exponentiated value function $\Psi(\mathbf{x},t)$ is formulated by the equation:

$$\mathbf{u}(\mathbf{x},t) = \lambda\mathbf{R}^{-1}\mathbf{G}(\mathbf{x},t)\ \frac{\nabla_{\mathbf{x}}\Psi(\mathbf{x},t)}{\Psi(\mathbf{x},t)} \tag{19}$$

Essentially under the logarithmic transformation the optimal control $\mathbf{u}(\mathbf{x},t)$ in (19) is acting such that stochastic dynamical systems visits states that maximize $\Psi(\mathbf{x},t)$. In the initial formulation in (12), the optimal controls are acting such that the stochastic dynamical system visits states that minimize the value function $V(\mathbf{x},t)$.

IV. Feynman Kac Lemma for Markov Jump Diffusion Processes.

The Feynman-Kac lemma is one of the most fundamental theoretical tools that bridges the gap between SDEs and PDEs and offers an alternative methodology for solving PDEs with forward sampling of SDEs. There are numerous applications of the Feynman-Kac lemma in financial engineering. Rigorous derivations of different versions of Feynman-Kac lemma are found in classic books for brownian stochastic calculus such as [22], [23]. The version of Feynman Kac Dynkin lemma for one-dimensional jump diffusions is published as an exercise for the reader in chapter 7 of [20], but no derivation is provided. In this work, we provide the proof of the Feynman-Kac lemma for multidimensional markov jump diffusions.

Lemma: *Lets consider the linear parabolic PDE:*

$$\frac{\partial_t\Psi(\mathbf{x},t)}{\partial t} + \boldsymbol{\mathcal{D}}\Psi(\mathbf{x},t) = \Xi(\mathbf{x},t)$$

with the boundary condition: $\Psi(\mathbf{x}(t_N),t_N) = \xi(t_N)$ *and the differential operator* $\boldsymbol{\mathcal{D}}$ *defined in* (17)*. Then its solution takes the form*

$$\Psi(\mathbf{x},t_0) = \left\langle \Psi(\mathbf{x},t_N)\exp\left(-\frac{1}{\lambda}\int_{t_0}^{t_N} q(\mathbf{x})d\tau\right) - \int_{t_0}^{t_N}\ \Xi(\mathbf{x},t)\exp\left(-\int_{t_0}^{t} q(\mathbf{x})d\tau\right)dt\right\rangle \tag{20}$$

with the expectation in (20) *taken under the forward sampling of the markov jump diffusion process:*

$$d\mathbf{x} = \mathbf{f}(\mathbf{x},t)dt + \mathbf{B}(\mathbf{x},t)d\mathbf{w}(t) + \mathbf{h}(\mathbf{x},t)d\mathbf{P}(t)$$

Proof: Let us consider $\boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t) = \Psi(\mathbf{x},t)\ \boldsymbol{\mathcal{Z}}(t_0,t)$ where the term $\boldsymbol{\mathcal{Z}}(t_0,t)$ is defined as follows:

$$\boldsymbol{\mathcal{Z}}(t_0,t) = \exp\left(-\int_{t_0}^{t} q(\mathbf{x})d\tau\right) \tag{21}$$

We apply the multidimensional version of the Itô lemma:

$$d\boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t) = d\Psi(\mathbf{x},t)\ \boldsymbol{\mathcal{Z}}(t_0,t) + \Psi(\mathbf{x},t)\ d\boldsymbol{\mathcal{Z}}(t_0,t) + d\Psi(\mathbf{x},t)\ d\boldsymbol{\mathcal{Z}}(t_0,t) \tag{22}$$

Since $d\Psi(\mathbf{x},t)\ d\boldsymbol{\mathcal{Z}}(t_0,t) = 0$ we will have that: $d\boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t) = d\Psi(\mathbf{x},t)\ \boldsymbol{\mathcal{Z}}(t,t_N) + \Psi(\mathbf{x},t)\ d\boldsymbol{\mathcal{Z}}(t_0,t)$. We calculate the differentials $d\Psi(\mathbf{x},t), d\boldsymbol{\mathcal{Z}}(t,t_N)$ according to the Itô differentiation rule. More precisely for the term $d\boldsymbol{\mathcal{Z}}(t_0,t_N)$ we will have that:

$$d\boldsymbol{\mathcal{Z}}(t_0,t) = -q(\mathbf{x})\boldsymbol{\mathcal{Z}}(t_0,t)\ dt \tag{23}$$

while the term $d\Psi(\mathbf{x},t)$ is specified according to the chain rule of the Markov Jump Diffusion processes. More precisely we will have that:

$$\begin{aligned} d\Psi(\mathbf{x},t) &= \frac{\partial\Psi(\mathbf{x},t)}{\partial t}dt + \nabla_{\mathbf{x}}\Psi(\mathbf{x},t)^T\mathbf{f}(\mathbf{x},t)dt \\ &+ \nabla_{\mathbf{x}}\Psi(\mathbf{x},t)^T\Big(\mathbf{B}(\mathbf{x},t)d\mathbf{w}\Big) \\ &+ \frac{1}{2}tr\Big(\nabla_{\mathbf{xx}}\Psi(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T\Big)dt \\ &+ \sum_{k=1}^{m}\Big[\Psi\Big(\mathbf{x}(t)+\mathbf{h}_k(\mathbf{x}(t),t),t\Big) - \Psi\Big(\mathbf{x}(t),t\Big)\Big]dP_k(t) \end{aligned} \tag{24}$$

By substituting (23) and (24) back to (22) and taking the expectation with respect the diffusion and Poisson differentials given the the state $\mathbf{x}(t_0)$ [1] we will have:

[1]This means that $\mathbf{x}(t_0)$is treated as deterministic variable

$$\left\langle d\boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t)\right\rangle = \left(\frac{\partial\Psi(\mathbf{x},t)}{\partial t}dt + \nabla_{\mathbf{x}}\Psi(\mathbf{x},t)^T\mathbf{f}(\mathbf{x},t)dt + \frac{1}{2}tr\left(\nabla_{\mathbf{xx}}\Psi(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)\mathbf{B}(\mathbf{x},t)^T\right)dt + \sum_{k=1}^{m}\mu_k(t)\left[\Psi\left(\mathbf{x}(t)+\mathbf{h}_k(\mathbf{x}(t),t),t\right)-\Psi\left(\mathbf{x}(t),t\right)\right]\right) \times \boldsymbol{\mathcal{Z}}(t_0,t) - \Psi(\mathbf{x},t)\ q(\mathbf{x})\boldsymbol{\mathcal{Z}}(t_0,t)\ dt \tag{25}$$

By considering the operator $\boldsymbol{\mathcal{D}}$defined in the previous section we will have:

$$\left\langle d\boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t)\right\rangle = \left(\frac{\partial\Psi(\mathbf{x},t)}{\partial t}dt + \boldsymbol{\mathcal{D}}\Psi\right)\ \boldsymbol{\mathcal{Z}}(t_0,t) \tag{26}$$

We integrate the equation above from t_0 to t_N therefore we get:

$$\left\langle \int_{t_0}^{t_N} d\boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t)dt\right\rangle = \int_{t_0}^{t_N}\left(\frac{\partial\Psi(\mathbf{x},t)}{\partial t}dt + \boldsymbol{\mathcal{D}}\Psi\right)\ \boldsymbol{\mathcal{Z}}(t_0,t)dt$$

With substitution of $\frac{\partial\Psi(\mathbf{x},t)}{\partial t}dt + \boldsymbol{\mathcal{D}}\Psi = \Xi(\mathbf{x},t)$ we have:

$$\left\langle \boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t_N) - \boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t_0)\right\rangle = \int_{t_0}^{t_N}\ \Xi(\mathbf{x},t)\boldsymbol{\mathcal{Z}}(t_0,t)dt$$

and then since $\boldsymbol{\mathcal{Z}}(t_0,t) = \exp\left(-\int_{t_0}^{t} q(\mathbf{x})d\tau\right)$ we get the equation:

$$\left\langle \boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t_N) - \boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t_0)\right\rangle = \int_{t_0}^{t_N}\ \Xi(\mathbf{x},t)\exp\left(-\int_{t_0}^{t} q(\mathbf{x})d\tau\right)dt$$

Finally since

$$\begin{aligned}\boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t_0) &= \Psi(\mathbf{x},t_0)\\ \boldsymbol{\mathcal{G}}(\mathbf{x},t_0,t_N) &= \Psi(\mathbf{x},t_N)\exp\left(-\frac{1}{\lambda}\int_{t_0}^{t_N} q(\mathbf{x})d\tau\right)\end{aligned}$$

we reach the final result:

$$\Psi(\mathbf{x},t_0) = \left\langle \Psi(\mathbf{x},t_N)\exp\left(-\frac{1}{\lambda}\int_{t_0}^{t_N} q(\mathbf{x})d\tau\right) - \int_{t_0}^{t_N}\ \Xi(\mathbf{x},t)\exp\left(-\int_{t_0}^{t} q(\mathbf{x})d\tau\right)dt\right\rangle$$

The expectation above is taken based on trajectories generated with forward sampling of the uncontrolled markov jump diffusion:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x},t)dt + \mathbf{B}(\mathbf{x},t)d\mathbf{w}(t) + \mathbf{h}(\mathbf{x},t)d\mathbf{P}(t) \tag{27}$$

This is the end of the proof of the Feynman-Kac lemma. For the linear HJB in (13) the Feynman-Kac Lemma takes the form:

$$\Psi\left(\mathbf{x}(t_0),t_0\right) = \left\langle \Psi\left(\mathbf{x}(t_N),t_N\right)\exp\left(-\frac{1}{\lambda}\int_{t_0}^{t_N} q(\mathbf{x})d\tau\right)\right\rangle$$

The expectation above can also be written as:

$$\Psi\left(\mathbf{x}(t_0),t_0\right) = \int P\left(\mathbf{x}_N,t_N|\mathbf{x}_0,t_0\right)\exp\left(-\frac{1}{\lambda}\int_{t_0}^{t_N} q(\mathbf{x})d\tau\right)\Psi_{t_N}d\mathbf{x}$$

In the next section we compute the probability $P\left(\mathbf{x}_N,t_N|\mathbf{x}_0,t_0\right)$ under the uncontrolled jump diffusion in (27).

V. Path Integral Formulation For Markov Jump Diffusion Processes.

We compute the path integral formulation for the case of markov jump diffusions processes of the form in (27) We follow the approach in [24] and we consider the one-dimensional case. Thus the jump diffusion in (27) will take the form:

$$x(s)-x(t) = \mathbf{f}(x(t),t)\delta t + \mathbf{B}(x(t),t)dw(t) + \mathbf{h}(x(t),t)dP(t) \tag{28}$$

where $s = t+\delta t$. Next we start with the conditional probability density function and marginalize with respect to all possible outcome of the jump differential dP. In particular, we will have:

$$P\left(x(s),s\middle|x(t),t\right) = \int P\left(x(s),dP,s|x(t),t\right)dP \tag{29}$$

By using the bayes rule, the equation above can be further formulated as:

$$P\Big(x(s),s\Big|x(t),t\Big) = \int P\Big(x(s),s|x(t),dP,t\Big)\text{Prob}\Big(dP=k\Big)dP$$

Since the probability $\text{Prob}\Big(dP=k\Big)$ is written as:

$$\begin{aligned}&\text{Prob}\Big(dP=k\Big) = \\ &= \Big(1-\mu(t)\delta t\Big)\delta_{k,0}+\mu(t)\delta t\delta_{k,1} \\ &= \Big(1-\mu(t)\delta t\Big)\delta(dP-0)+\mu(t)\delta t\delta(dP-1)\end{aligned}$$

we will have that:

$$\begin{aligned}&P\Big(x(s),s\Big|x(t),t\Big) = \\ &\Big(1-\mu(t)\delta t\Big)P\Big(x(s),s\Big|x(t),dP=0,t\Big) \\ &+\mu(t)\delta t P\Big(x(s),s\Big|x(t),dP(t)=1,t\Big)\end{aligned}$$

or in a more compact form:

$$P\Big(x(s),s\Big|x(t),t\Big) = P_{nospike}\Big(1-\mu(t)\delta t\Big)+P_{spike}\mu(t)\delta t \tag{30}$$

where $P_{nospike} = P\left(x(s),s\Big|x(t),dP=0,t\right)$ is the conditional probability density function when no spike occurs and $P_{spike} = P\left(x(s),s\Big|x(t),dP=1,t\right)$ is the when a spike occurs. Given the markov jump diffusion in (28) the conditional probability density functions are determined as follows:

$$\begin{aligned}&P\left(x(s),s\Big|x(t),dP=0,t\right) \\ &= \frac{1}{\sqrt{2\pi\boldsymbol{\sigma}\delta t}}\exp\left[-\frac{(x(s)-x(t)-\mathbf{f}(x,t)\delta t)}{2\boldsymbol{\sigma}}\right]\end{aligned}$$

and

$$\begin{aligned}&P\left(x(s),s\Big|x(t),dP=1,t\right) = \frac{1}{\sqrt{2\pi\boldsymbol{\sigma}\delta t}} \\ &\times\exp\left[-\frac{(x(s)-x(t)-(\mathbf{f}(\mathbf{x},t)+\mathbf{h}(x,t)\mu)\delta t)^2}{2\boldsymbol{\sigma}}\right]\end{aligned}$$

with the parameter $\boldsymbol{\sigma}$ defined as $\boldsymbol{\sigma} = \mathbf{B}(x,t)^2$. After defining the CPDF under the markov jump diffusion (28) the path integral formulation is expressed as:

$$P\Big(x_N,t_N|x_0,t_0\Big) = \prod_{i=0}^{N-1}P\Big(x_{i+1},t_{i+1}|x_i,t_i\Big) \tag{31}$$

with $P\Big(x_{i+1},t_{i+1}|x_i,t_i\Big)$ defined as

$$\begin{aligned}P\Big(x_{i+1},t_{i+1}|x_i,t_i\Big) &= P^{(i)}_{nospike}\Big(1-\mu(t_i)\delta t\Big) \\ &+P^{(i)}_{spike}\mu(t_i)\delta t\end{aligned}$$

and $P^{(i)}_{nospike}, P^{(i)}_{spike}$ expressed as:

$$P^{(i)}_{nospike} = P\left(x(t_{i+1}),t_{i+1}\Big|x(t_i),dP(t_i)=0,t_i\right)$$

and

$$P^{(i)}_{spike} = P\left(x(t_{i+1}),t_{i+1}\Big|x(t_i),dP(t_i)=1,t_i\right)$$

VI. Conclusion and Future Work.

In this work we consider stochastic optimal control for nonlinear markov jump diffusion processes. We show that under the logarithmic transformation of the value function, and the assumption that relates the weight in controls and the variance of the brownian noise, the nonlinear and second order HJB is transformed into the backward Chapman Kolmogorov PDE for markov jump diffusions. After deriving the Feynman-Kac lemma for markov jump diffusions we apply it to get the stochastic representation of the solution of the backward Chapman Kolmogorov PDE. Essentially, with the application of the Feynman-Kac lemma we can solve the initial stochastic optimal control problem by evaluating the expectation of the exponential of state dependent part $q(\mathbf{x})$ of the cost function under the uncontrolled stochastic dynamics. Finally, we provide the path integral formulation for one-dimensional markov jump diffusions.

In future work, we will consider more general cases of jump diffusions in which the amplitude of the jump term is not only a function of the state but it is also a function of an additional random variable. This is the class of the "marked" jump diffusions. Furthermore, we plan to work on iterative versions of the proposed framework in which given an initial policy, successive application of the Feynman-Kac will result in policy improvement. The iterative scheme could be derived by using importance sampling based on Girsanov's theorem

and the Radon-Nikodým derivative as applied to markov jump diffusion processes. At each iteration only the drift of the stochastic dynamics will change as the controls are updated. Finally we are currently working on KL control [12] as applied to stochastic optimal control for markov jump diffusions.

VII. Appendix

In this section we show how the nonlinear and second order PDE in (13) is transformed into (18). More precisely by inserting the logarithmic transformation and the derivatives of the value function in (13) we obtain:

$$\begin{aligned}\frac{\lambda}{\Psi}\frac{\partial \Psi}{\partial t} &= q_t - \frac{\lambda}{\Psi}(\nabla_{\mathbf{x}}\Psi)^T\mathbf{f} \\ &\quad - \underline{\frac{\lambda^2}{2\Psi_t^2}(\nabla_{\mathbf{x}}\Psi)^T\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T(\nabla_{\mathbf{x}}\Psi)} \\ &\quad + \frac{1}{2}tr\left(\Gamma\right) - \frac{\lambda}{\Psi}\, d_{jump}\Psi\end{aligned} \tag{32}$$

where the term Γ is expressed as:

$$\Gamma = \left(\lambda\frac{1}{\Psi^2}\nabla_{\mathbf{x}}\Psi\ \nabla_{\mathbf{x}}\Psi^T - \lambda\frac{1}{\Psi}\nabla_{\mathbf{xx}}\Psi\right)\mathbf{B}\mathbf{B}^T$$

The trace of Γ is therefore:

$$\begin{aligned}\frac{1}{2}tr(\Gamma) &= \underline{\lambda\frac{1}{2\Psi^2}tr\left(\nabla_{\mathbf{x}}\Psi^T\mathbf{B}\mathbf{B}^T\nabla_{\mathbf{x}}\Psi\right)} \\ &\quad - \frac{1}{2}\lambda\frac{1}{\Psi}tr\left(\nabla_{\mathbf{xx}}\Psi\mathbf{B}\mathbf{B}^T\right)\end{aligned} \tag{33}$$

Comparing the underlined terms in (32) and (33), one can recognize that these terms will cancel under the assumption $\lambda\mathbf{G}(\mathbf{x})\mathbf{R}^{-1}\mathbf{G}(\mathbf{x})^T = \mathbf{B}(\mathbf{x})\mathbf{B}(\mathbf{x})^T = \mathbf{\Sigma}$. The resulting PDE has the form of (18)

References

[1] Robert F. Stengel. *Optimal control and estimation*. Dover books on advanced mathematics. Dover Publications, New York, 1994.

[2] W. H. Fleming and H. Mete Soner. *Controlled Markov processes and viscosity solutions*. Applications of mathematics. Springer, New York, 2nd edition, 2006.

[3] Peter Dorato, Vito Cerone, and Chaouki Abdallah. *Linear Quadratic Control: An Introduction*. Krieger Publishing Co., Inc., Melbourne, FL, USA, 2000.

[4] E. Todorov. Stochastic optimal control and estimation methods adapted to the noise characteristics of the sensorimotor system. *Neural Computation*, 17(5):1084, 2005.

[5] Weiwei Li, Emanuel Todorov, and Xiuchuan Pan. Hierarchical optimal control of redundant biomechanical systems. In *in 26th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2004.

[6] J. Izawa, T. Rane, O. Donchin, and R. Shadmehr. Motor adaptation as a process of reoptimization. *Journal Of Neuroscience*, 28(11):2883–2891, 2008.

[7] M. Papageorgiou and T. Bauschert. Stochastic optimal control of moving vehicles in a dynamic environment. In *International Journal of Robotic Research*, volume 13, pages 342–354, August 1994.

[8] David H. Jacobson and David Q. Mayne. *Differential dynamic programming*. American Elsevier Pub. Co., New York,, 1970.

[9] H. J. Kappen. An introduction to stochastic control theory, path integrals and reinforcement learning. In J. Marro, P. L. Garrido, and J. J. Torres, editors, *Cooperative Behavior in Neural Systems*, volume 887 of *American Institute of Physics Conference Series*, pages 149–181, February 2007.

[10] H. J. Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 11:P11011, 2005.

[11] H. J. Kappen. Linear theory for control of nonlinear stochastic systems. *Phys Rev Lett*, 95:200201, 2005. Journal Article United States.

[12] E. Todorov. Efficient computation of optimal actions. *Proc Natl Acad Sci U S A*, 106(28):11478–83, 2009.

[13] E. Theodorou, J. Buchli, and S. Schaal. A generalized path integral approach to reinforcement learning. *Journal of Machine Learning Research*, (11):3137–3181, 2010.

[14] E.. Theodorou. *Iterative Path Integral Stochastic Optimal Control: Theory and Applications to Motor Control*. PhD thesis, university of southern California, May 2011.

[15] E. Theodorou, J. Buchli, and S. Schaal. Reinforcement learning of motor skills in high dimensions: A path integral approach. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2010.

[16] Jonas Buchli, Evangelos Theodorou, Freek Stulp, and Stefan Schaal. Variable impedance control - a reinforcement learning approach. In *Robotics: Science and Systems Conference (RSS)*, 2010.

[17] Freek Stulp, Jonas Buchli, Evangelos Theodorou, and Stefan Schaal. Reinforcement learning of full-body humanoid motor skills. In *10th IEEE-RAS International Conference on Humanoid Robots*, 2010.

[18] P. Pastor, M. Kalakrishnan, S. Chitta, E. Theodorou, and S. Schaal. skill learning and task outcome prediction for manipulation. In *robotics and automation (icra), 2011 ieee international conference on*, 2011.

[19] Freek Stulp, Evangelos Theodorou, Jonas Buchli, and Stefan Schaal. Learning to grasp under uncertainty. In *International Conference of Robotics and Automation*, 2010.

[20] Floyd B. Hanson. *Applied Stochastic Processes and Control for Jump-Diffusions*. SIAM, 2007.

[21] C. Gardiner. *Handbook of Stochastic Methods: for Physics, Chemistry and the Natural Sciences*. Spinger, 2004.

[22] A. Friedman. *Stochastic Differential Equations And Applications*. Academic Press, 1975.

[23] Ioannis Karatzas and Steven E. Shreve. *Brownian Motion and Stochastic Calculus (Graduate Texts in Mathematics)*. Springer, 2nd edition, August 1991.

[24] A. Pirrota and R. Santoro. Probabilistic response of nonlinear systems under combined normal and poison white noise via path integral method. *Probabilistic Engineering Mechanics*, 26:26–32, 2011.