Relative Entropy and Free Energy Dualities: Connections to Path Integral and KL control

Evangelos A. Theodorou[1] and Emanuel Todorov[1,2]

***Abstract*— This paper integrates recent work on Path Integral (PI) and Kullback Leibler (KL) divergence stochastic optimal control theory with earlier work on risk sensitivity and the fundamental dualities between free energy and relative entropy. We derive the path integral optimal control framework and its iterative version based on the aforemetioned dualities. The resulting formulation of iterative path integral control is valid for general feedback policies and in contrast to previous work, it does not rely on pre-specified policy parameterizations. The derivation is based on successive applications of Girsanov's theorem and the use of Radon-Nikodým derivative as applied to diffusion processes due to the change of measure in the stochastic dynamics. We compare the PI control derived based on Dynamic Programming with PI based on the duality between free energy and relative entropy. Moreover we extend our analysis on the applicability of the relationship between free energy and relative entropy to optimal control of markov jump diffusions processes. Furthermore, we present the links between KL stochastic optimal control and the aforementioned dualities and discuss its generalizability.**

I. Introduction

Stochastic optimal control for nonlinear diffusion processes based on path integrals demonstrated remarkable applicability to robotic control and planning problems. For continuous state actions spaces and continuous time, work in [1], [2] provided the Path Integral (PI) representation of stochastic optimal control for a special class of dynamics and presented new insights regarding symmetry breaking phenomena and their connection to optimal control. In [3], the PI control framework was extended to stochastic optimal control problems for multi-agents systems.

In [4], [5] PI control was derived for the case of generalized diffusions processes with state dependent control and diffusions matrices. Additionally, an iterative algorithm was provided for the cases in which desired trajectories and/or control gains are parameterized with the use of Dynamic Movement Primitives (DMPs). DMPs are nonlinear point attractors with adjustable landscape and they have been used in robotics for the purposes of smooth representation of desired trajectories and/or control gains. The resulting algorithm Policy Improvement with Path Integrals (PI^2) has been applied to a variety of robotic systems for tasks such as planning, gain scheduling and variable stiffness control [6]–[9].

[1] Department of Computer Science and Engineering, University of Washington, Seattle. email: etheodor@cs.washington.edu

[2] Department of Applied Mathematics, University of Washington, Seattle. email: todorov@cs.washington.edu

This work was submitted for review on March 7th of 2012.

This work was supported by the US National Science Foundation.

Parallel to the work in continuous time, in [10], [11] the Bellman principle of optimality was applied for discrete time optimal control problems in which the control cost is formulated as the Kullback Leibler (KL) divergence between the controlled and uncontrolled dynamics. The resulting framework is applicable to a large class of control problems which include finite, infinite horizon, exponentially discounted and first exit. In this paper we will derive the connections of PI and KL control as presented in the machine learning and robotics communities [1], [2], [10], [11] with earlier work in controls on the fundamental dualities between relative entropy and free energy and the logarithmic transformations of diffusions processes [12]–[17]. More precisely:

- We present PI control and its iterative version based on the fundamental dualities between free energy and relative entropy as applied to nonlinear diffusion processes. In contrast to previous work [4], the derivation and the resulting formulation of iterative path integral control holds for general feedback policies and it does not rely on specific policy parameterizations. The derivation is based on successive applications of Girsanov's theorem due to the change of measure in the stochastic dynamics. The aforementioned change of measure is the outcome of the change in the drift of the stochastic dynamics which, in turn, results from the updates in controls that take place at every iteration.
- We compare the proposed PI optimal control formulation derived based on the application of Girsanov's theorem and Jensen's inequality with the one derived based on the Bellman principle of optimality. We specify the conditions under which the two approaches lead to the same results and discuss their generazability in terms of types of cost functions and forms of stochastic dynamics.
- We extend our analysis to stochastic optimal control for jump diffusions processes of one dimension based on the fundamental relationship between free energy and relative entropy and derive the corresponding bound on the cost function.
- Finally, we provide the connections of KL stochastic optimal control with earlier work on risk sensitivity and discuss the generalizability with respect to different stochastic optimal control problems.

The paper is organized as follows: in Section II we provide the basic dualities between free energy and relative entropy. In Section III we discuss how these dualities are linked to maximizing or minimizing stochastic optimal control

problems for the case of diffusion processes. In Section IV we derive the iterative case based on successive applications of Girsanov's theorem. In section V we show how the path integral control framework is derived based on the Bellman principle of optimality and contrast this approach with the one in Section III. We expand our analysis on path integral control for the case of markov jump diffusions in Section VI. Finally in Section VII we provide links to KL-control and in Section VIII we conclude by discussing the generalizability of the aforementioned approaches.

II. Basic Duality Relationships of Free Energy and Relative Entropy

In this section we derive the fundamental duality relationships between free energy and relative entropy [16]. This relationship is important for the derivation of stochastic optimal control. Let $(\mathcal{Z}, \boldsymbol{\mathcal{Z}})$ denote a measurable space and $\boldsymbol{\mathcal{P}}(\mathcal{Z})$ the corresponding probability measure defined on the measurable space. For our analysis we consider the following definitions.

Definition 1: Let $\mathbb{P} \in \boldsymbol{\mathcal{P}}(\mathcal{Z})$ and the function $\mathcal{J}(\mathbf{x}) : \mathcal{Z} \rightarrow \Re$ be a measurable function. Then the term:

$$\mathbb{E}\Big(\mathcal{J}(\mathbf{x})\Big) = \log \int \exp\left(\rho \mathcal{J}(\mathbf{x})\right) d\mathbb{P} \tag{1}$$

is called free energy of $\mathcal{J}(\mathbf{x})$ with respect to $\mathbb{P}$.

Definition 2: Let $\mathbb{P} \in \boldsymbol{\mathcal{P}}(\mathcal{Z})$ and $\mathbb{Q} \in \boldsymbol{\mathcal{P}}(\mathcal{Z})$, the relative entropy of $\mathbb{P}$ with respect to $\mathbb{Q}$ is defined as:

$$\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right) = \begin{cases} \int \log \frac{d\mathbb{Q}}{d\mathbb{P}} d\mathbb{Q} & \text{if } \mathbb{Q} << \mathbb{P} \text{ and } \log \frac{d\mathbb{Q}}{d\mathbb{P}} d\mathbb{Q} \in L^1 \\ +\infty & \text{otherwise} \end{cases}$$

We will also consider the objective function:

$$\xi(\mathbf{x}) = \frac{1}{\rho}\mathbb{E}\Big(\mathcal{J}(\mathbf{x})\Big) = \frac{1}{\rho}\log \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}\Big[\exp\left(\rho \mathcal{J}(\mathbf{x})\right)\Big] \tag{2}$$

with $\mathcal{J}(\mathbf{x}) = \phi(\mathbf{x}_{t_N}) + \int_{t_i}^{t_N} q(\mathbf{x}) dt$ is the state dependent cost. The objective function above takes the form $\xi(\mathbf{x}) = \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}\left(\mathcal{J}\right) + \frac{\rho}{2} Var\left(\mathcal{J}\right)$ as $\rho \rightarrow 0$. This form allows us to get the basic intuition for constructing such objective functions. Essentially for small ρ the cost is a function of the mean the variance. When $\rho > 0$ the cost function is risk sensitive while for $\rho < 0$ is risk seeking. To derive the basic relationship between free energy and relative entropy we express the expectation $\mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}$ taken under the measure $\mathbb{P}$ as a function of the expectation $\mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}$ taken under the probability measure $d\mathbb{Q}$. More precisely will have:

$$\begin{aligned} \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}\Big[\exp\left(\rho \mathcal{J}(\mathbf{x})\right)\Big] &= \int \exp\left(\rho \mathcal{J}(\mathbf{x})\right) d\mathbb{P} \\ &= \int \exp\left(\rho \mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} \end{aligned}$$

By taking the logarithm of both sides of the equations above and making use of the Jensen's inequality we will have:

$$\begin{aligned} &\log \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}\Big[\exp\left(\rho \mathcal{J}(\mathbf{x})\right)\Big] = \log \int \exp\left(\rho \mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} \geq \\ &\int \log\left(\exp\left(\rho \mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} = \int \left(\rho \mathcal{J}(\mathbf{x}) + \log \frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \\ &= \int \rho \mathcal{J}(\mathbf{x}) d\mathbb{Q} - \mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right) \end{aligned}$$

We multiply the inequality above with $\frac{1}{\rho}$ for case of $\rho < 0$ or $\rho = -|\rho|$ and thus we have:

$$\xi(\mathbf{x}) = -\frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right) \leq \mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right) \tag{3}$$

where $\mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) = \int \mathcal{J}(\mathbf{x}) d\mathbb{Q}$. The inequality above gives us the duality relationship between relative entropy and free energy. Essentially one could define the following two minimization problems:

$$-\frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right) = \inf\left[\mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right)\right] \tag{4}$$

and the dual minimization:

$$-\frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right) = \inf\left[\mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right)\right] \tag{5}$$

The infimum in (4) is attained at $\mathbb{Q}^*$ given by:

$$d\mathbb{Q}^* = \frac{\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right) d\mathbb{P}}{\int \exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right) d\mathbb{P}} \tag{6}$$

When $\rho > 0$ the inequality in (3) becomes from $\leq$ to $\geq$ and the $\inf$ in (4) and (5) becomes $\sup$. In the next section we show how inequality (4) is transformed to a stochastic optimal control problem for the case of markov diffusion processes.

III. Stochastic optimal control for Markov Diffusions Processes based on the fundamental dualities

For our analysis in this section we use the same notation as in [13], [16]. We consider the uncontrolled and controlled stochastic dynamics of the form:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}(\mathbf{x}) d\mathbf{w}^{(0)}(t) \tag{7}$$

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}dt + \frac{1}{\sqrt{|\rho|}} d\mathbf{w}^{(1)}(t)\right) \tag{8}$$

with $\mathbf{x}_t \in \Re^{n \times 1}$ denoting the state of the system, $\boldsymbol{\mathcal{B}}(\mathbf{x}, t) : \Re^n \times \Re \rightarrow \Re^{n \times n}$ is the control and diffusions matrix, $\mathbf{f}(\mathbf{x}, t) : \Re^n \times \Re \rightarrow \Re^{n \times 1}$ the passive dynamics, $\mathbf{u}_t \in \Re^{n \times 1}$ the control vector and $d\mathbf{w} \in \Re^{p \times 1}$ brownian noise. Notice that the difference between the two diffusions above is on the controls that appear in (8). These controls together with the passive dynamics define a new drift term. For our analysis here we assume $\boldsymbol{\mathcal{B}}^{-1}$ exists. Expectations evaluated

on trajectories generated by the controlled dynamics and uncontrolled dynamics are represented as $\mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}$ and $\mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}$ respectively. The corresponding probability measures of the aforementioned expectations are $\mathbb{P}$ and $\mathbb{Q}$. We continue our analysis with the main result in (3) and the definition of the Radon-Nikodým derivative:

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\zeta(\mathbf{u})\right) \quad \text{and} \quad \frac{d\mathbb{P}}{d\mathbb{Q}} = \exp\left(-\zeta(\mathbf{u})\right) \tag{9}$$

where according to Girsanov's theorem [18] (see also section IX) adapted to the diffusion processes (7) and (8) the term $\zeta(\mathbf{u})$ is expressed as follows:

$$\zeta(\mathbf{u}) = \frac{1}{2}|\rho| \int_{t_i}^{t_N} \mathbf{u}^T \mathbf{u} dt + \sqrt{|\rho|} \int_{t_i}^{t_N} \mathbf{u}^T d\mathbf{w}^{(1)}(t) \tag{10}$$

An informal explanation for the applicability of Girsanov's theorem is that it provides the link between expectations evaluated on trajectories generated from diffusions with different drift terms. Substitution of (9) and (21) into inequality (3) gives the following result:

$$\xi(\mathbf{x}) = -\frac{1}{|\rho|} \log \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}\left[\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)\right] \leq \mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{|\rho|}\zeta(\mathbf{u})\right] \tag{11}$$

The expectation on the right side of the inequality in (11) is further simplified as follows:

$$\xi(\mathbf{x}) \leq \mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N} \mathbf{u}^T \mathbf{u} dt\right] \tag{12}$$

The right term of the inequality above corresponds to the cost function of a stochastic optimal control problem that is bounded from below by the free energy. Besides providing a lower bound on the objective function of the stochastic optimal control problem inequality (12) expresses also how this lower bound should be computed. This computation involves forward sampling of the uncontrolled dynamics, evaluation of the expectation of the exponentiated state depended part $\phi(\mathbf{x}_{t_N})$ and $q(\mathbf{x}_t)$ and the logarithmic transformation of this expectation. Surprisingly, inequality (12) was derived without relying on any principle of optimality. It only takes the application of Girsanov theorem between controlled and uncontrolled stochastic dynamics and the use of dual relationship between free energy and relative entropy to find the lower bound in (12). Essentially inequality (12) defines a minimization process in which the right part of the inequality is minimized with respect $\zeta(\mathbf{u})$ and therefore with respect to control $\mathbf{u}$. At the minimum, when $\mathbf{u} = \mathbf{u}^*$ then the right part of the inequality in (12) reaches its optimal $\xi(\mathbf{x})$. Under the optimal control $\mathbf{u}^*$ and according to (13) the optimal distribution takes the from:

$$d\mathbb{Q}^*(\mathbf{x}) = \frac{\exp\left(-|\rho|\int q(\mathbf{x})dt\right)d\mathbb{P}(\mathbf{x})}{\int \exp\left(-|\rho|\int q(\mathbf{x})dt\right)d\mathbb{P}(\mathbf{x})} \tag{13}$$

An important question to ask is what is the link between (12) and the dynamic programming principle. To find this link the next step is to show that $\xi(\mathbf{x})$ satisfies the HJB equations and therefore it is the corresponding value function. More precisely, we introduce a new variable $\Phi(\mathbf{x}, t)$ defined as $\Phi(\mathbf{x}, t) = \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}(\exp\left(\rho\mathcal{J}(\mathbf{x})\right))$. The Feynman-Kac lemma [19] tells us that this function satisfies the backward Chapman Kolmogorov PDE. Therefore the following equation holds:

$$-\partial_t\Phi = \rho q_0 \Phi + \mathbf{f}^T(\nabla_{\mathbf{x}}\Phi) + \frac{1}{2|\rho|} tr\left((\nabla_{\mathbf{xx}}\Phi)\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T\right) \tag{14}$$

For $\rho = -|\rho| < 0$ and since $\xi(\mathbf{x}) = \frac{1}{\rho}\log\Phi(\mathbf{x}, t) = -\frac{1}{|\rho|}\log\Phi(\mathbf{x}, t)$ we will have that $\partial_t\Phi = -|\rho|\Phi\partial_t\xi$, $\nabla_{\mathbf{x}}\Phi = -|\rho|\Phi\nabla_{\mathbf{x}}\xi$ and $\nabla_{\mathbf{xx}}\Phi = |\rho|\Phi\nabla_{\mathbf{xx}}\xi - |\rho|^2\Phi\nabla_{\mathbf{x}}\xi\nabla_{\mathbf{x}}\xi^T$ it can be trivially shown that $\xi(\mathbf{x})$ satisfies the nonlinear PDE:

$$-\partial_t\xi = q_0 + (\nabla_{\mathbf{x}}\xi)^T\mathbf{f} - \frac{1}{2}(\nabla_{\mathbf{x}}\xi)^T\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T(\nabla_{\mathbf{x}}\xi) + \frac{1}{2|\rho|} tr\left((\nabla_{\mathbf{xx}}\xi)\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T\right) \tag{15}$$

Similarly, for the case of $\rho = |\rho| > 0$ the resulting PDE will be:

$$-\partial_t\xi = q_0 + (\nabla_{\mathbf{x}}\xi)^T\mathbf{f} + \frac{1}{2}(\nabla_{\mathbf{x}}\xi)^T\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T(\nabla_{\mathbf{x}}\xi) + \frac{1}{2|\rho|} tr\left((\nabla_{\mathbf{xx}}\xi)\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T\right) \tag{16}$$

The nonlinear PDEs above corresponds to the HJB equation [20] for the case of the minimizing and maximizing optimal control problem with control weight $\mathbf{R}^{-1} = I$ and therefore, $\xi(\mathbf{x})$ is the corresponding minimizing or maximizing value function. Note that in order to derive the PDEs above we did not use any principle of optimality. The analysis so far is summarized by the following corollary in which we use the function $sign(x) = -1 \quad \forall x < 0$ and $sign(x) = 1 \quad \forall x > 0$. More precisely we will have:

Corollary 1: *Consider the expectation operators $\mathcal{E}^{(0)}$, $\mathcal{E}^{(1)}$ evaluated on state trajectories sampled according to (7) and (8) respectively. The function $\xi(\mathbf{x}, t)$ specified as:*

$$\xi(\mathbf{x}, t) = \frac{sign(\rho)}{|\rho|}\log\mathcal{E}^{(0)}\left[\exp\left(sign(\rho)|\rho|\mathcal{J}(\mathbf{x})\right)\right] \tag{17}$$

is the value function of the stochastic optimal control problems:

$$\xi(\mathbf{x}, t_i) = \min_{\mathbf{u}} \mathcal{E}^{(1)}\left[\int_{t_i}^{t_N}\left(q(\mathbf{x}) - \frac{1}{2}\mathbf{u}^T\mathbf{u}\right)dt\right], \quad \forall\rho > 0$$

$$\xi(\mathbf{x}, t_i) = \max_{\mathbf{u}} \mathcal{E}^{(1)}\left[\int_{t_i}^{t_N}\left(q(\mathbf{x}) + \frac{1}{2}\mathbf{u}^T\mathbf{u}\right)dt\right], \quad \forall\rho < 0$$

subject to the stochastic dynamics in (8).

Corollary 1 shows how to compute the value function $\xi(\mathbf{x}, t)$. More precisely, the computation involves sampling of state trajectories based on the uncontrolled dynamics (7) and evaluation of the expectation in (17) on the resulting state trajectories. To derive (17) it takes only the application of Girsanov's theorem and Jensen's inequality.

IV. Feedback Control For Markov Diffusion Processes

There are different ways to make use of the fundamental inequality in (12) and derive controllers. For lower dimensional stochastic control problems evaluation of the free energy under the uncontrolled dynamics provides a good estimate of the value function. For planning and control problems of dynamical systems in high dimensional state spaces, the evaluation of the expectation may become numerical intractable. Here we show the derivation of the iterative case based on successive application of Girsanov's theorem for the change of measure at iteration k of the iterative algorithm.

Lemma 1: *Consider the stochastic dynamics* $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}_k dt + \frac{1}{\sqrt{|\rho|}} d\mathbf{w}^{(1)}(t)\right)$ *with the control policy* $\mathbf{u}_k(\mathbf{x}, t)$ *at iteration k. When sampling from these dynamics, the risk seeking function* $\xi(\mathbf{x}, t)$ *in (17) takes the form:*

$$\xi(\mathbf{x}, t) = -\frac{1}{|\rho|} \log \int \exp\Big[-|\rho| S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t))\Big] \mathbf{dx}$$

with the path cost $S(\mathbf{x}, \mathbf{u}_k)$ *defined as:*

$$S(\mathbf{x}, \mathbf{u}_k) = \mathcal{J}(\mathbf{x}) + \frac{1}{2}\left(\eta(\mathbf{u}) + \int_{t_i}^{t_N} ||\mu(\mathbf{x})||^2_{\boldsymbol{\Sigma}^{-1}} \delta t\right) \quad (18)$$

The term $\eta(\mathbf{u})$ *in the path cost above is defined as* $\eta(\mathbf{u}) = \int_{t_i}^{t_N} \mathbf{u}_k^T \mathbf{u}_k dt + \int_{t_i}^{t_N} 2\mathbf{u}_k^T \boldsymbol{\mathcal{B}}^{-T} \mu(\mathbf{x}) dt$ *and terms* $\mu(\mathbf{x}) = \left(\frac{\delta \mathbf{x}}{\delta t} - \mathbf{f}(\mathbf{x}) - \boldsymbol{\mathcal{B}}\mathbf{u}_k\right)$, $\boldsymbol{\Sigma} = \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T$.

Proof: The proof relies on the change of measure and use of the Radon Nikodym derivative for markov diffusion processes. More precisely we will have that:

$$\begin{aligned}
\xi(\mathbf{x}) &= -\frac{1}{|\rho|} \log \int \exp\left(-|\rho| \mathcal{J}(\mathbf{x})\right) d\mathbb{P} \\
&= -\frac{1}{|\rho|} \log \int \exp\left(-|\rho| \mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} \\
&= -\frac{1}{|\rho|} \log \int \exp\left(-|\rho| \mathcal{J}(\mathbf{x}) - \zeta(\mathbf{u})\right) d\mathbb{Q} \qquad (19)
\end{aligned}$$

The measure $d\mathbb{Q}$ takes the form of a path integral [21] and thus it is expressed as:

$$\mathbb{Q}\left(\mathbf{x}_N, t_N | \mathbf{x}_i, t_i\right) = \frac{\exp\left(-\frac{|\rho|}{2}\left(\int_{t_i}^{t_N} \mu(\mathbf{x})^T \boldsymbol{\Sigma}^{-1} \mu(\mathbf{x}) dt\right)\right)}{(2\pi dt)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \quad (20)$$

where we use the fact that $\boldsymbol{\mathcal{B}} d\mathbf{w}_k = \sqrt{\rho}\mu(\mathbf{x})\delta t$ and $\mu(\mathbf{x}) = \left(\frac{\delta \mathbf{x}}{\delta t} - \mathbf{f}(\mathbf{x}) - \boldsymbol{\mathcal{B}}\mathbf{u}_k\right)$. Based on the aforementioned inequalities the term $\zeta(\mathbf{u})$ in the Girsanov's theorem [22], [23] will become equal to:

$$\begin{aligned}
\zeta(\mathbf{u}) &= \frac{1}{2}|\rho| \int_{t_i}^{t_N} \mathbf{u}^T \mathbf{u} dt + \sqrt{|\rho|} \int_{t_i}^{t_N} \mathbf{u}^T d\mathbf{w}^{(1)}(t) \\
&= \frac{1}{2}|\rho| \int_{t_i}^{t_N} \mathbf{u}_k^T \mathbf{u}_k dt + |\rho| \int_{t_i}^{t_N} \mathbf{u}_k^T \boldsymbol{\mathcal{B}}^{-T} \mu(\mathbf{x}) dt \\
&= \frac{1}{2}|\rho|\ \eta(\mathbf{u}) \qquad (21)
\end{aligned}$$

with $\eta(\mathbf{u})$ defined as:

$$\begin{aligned}
\eta(\mathbf{u}) &= \int_{t_i}^{t_N} \mathbf{u}_k^T \mathbf{u}_k dt + \int_{t_i}^{t_N} 2\mathbf{u}_k^T \boldsymbol{\mathcal{B}}^{-T} \mu(\mathbf{x}) dt \\
&= \int_{t_i}^{t_N} \mathbf{u}^T \mathbf{u} dt + \frac{1}{\sqrt{|\rho|}} \int_{t_i}^{t_N} 2\mathbf{u}^T d\mathbf{w}^{(1)}(t) \qquad (22)
\end{aligned}$$

Substitution of the function above $\zeta(\mathbf{u})$ and the path integral into (19) results in the expression:

$$\xi(\mathbf{x}) = -\frac{1}{|\rho|} \log \int \exp\left(-|\rho| \mathcal{J}(\mathbf{x}) - \zeta(\mathbf{u}_k)\right) d\mathbb{Q} =$$

$$-\frac{1}{|\rho|} \log \int \exp\left[-|\rho| \left(\mathcal{J}(\mathbf{x}) + \frac{\eta(\mathbf{u}) + \int_{t_i}^{t_N} ||\mu(\mathbf{x})||^2_{\boldsymbol{\Sigma}^{-1}} dt}{2} \right) \right] \mathbf{dx}$$

with $\mathbf{dx}$ defined as $\mathbf{dx} = d\mathbf{x}_{t_{i+1}}, ..., d\mathbf{x}_{t_N}$. Thus in a more compact form we will have that:

$$\xi(\mathbf{x}) = -\frac{1}{|\rho|} \log \int \exp\Big[-|\rho| S(\mathbf{x}, \mathbf{u}_k)\Big] \mathbf{dx}$$

with the term $S(\mathbf{x}, \mathbf{u}_k)$ defined as $S(\mathbf{x}, \mathbf{u}_k) = \mathcal{J}(\mathbf{x}) + \frac{1}{2}\left(\eta(\mathbf{u}) + \int_{t_i}^{t_N} ||\mu(\mathbf{x})||^2_{\boldsymbol{\Sigma}^{-1}} dt\right)$.

■

A. Iterative Path Integral Control

In this section we derive the iterative optimal control based on lemma 1. The final result is given in the form of the theorem that follows:

Theorem 1: *Consider the stochastic optimal control problem:*

$$\xi(\mathbf{x}) = \min_{\mathbf{u}} E^{(1)}\left[\int_{to}^{t_N} \left(q(\mathbf{x}) + \frac{1}{2}\mathbf{u}^T \mathbf{u} \right) dt \right]$$

subject to the stochastic constraints:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u} dt + \frac{1}{\sqrt{|\rho|}} d\mathbf{w}^{(1)}(t)\right)$$

The iterative optimal control solution has the form:

$$\boxed{\mathbf{u}_{k+1}(\mathbf{x}, t) dt = \mathbf{u}_k(\mathbf{x}, t) dt + \frac{1}{\sqrt{\rho}} \mathcal{E}_{p_k}\left(d\mathbf{w}_k(t)\right)} \quad (23)$$

with P_k *having the form of a path integral expressed as:* $P_k = \frac{e^{-|\rho| S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t))}}{\int e^{-|\rho| S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t)} \mathbf{dx}}$ *and the path cost term* $S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t)$ *defined as in (18).*

Proof: To get the control we take the derivative of $S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t))$ with respect to $\mathbf{x}_{t_i}$. More precisely we will have that:

$$\begin{aligned}
\nabla_{\mathbf{x}_{t_i}} \xi(\mathbf{x}_{t_i}) &= -\frac{1}{|\rho|} \nabla_{\mathbf{x}_{t_i}} \left(\log \int \exp\Big[-|\rho| S(\mathbf{x}, \mathbf{u}_k)\Big] \mathbf{dx} \right) \\
&= -\frac{1}{|\rho|} \frac{\nabla_{\mathbf{x}_{t_i}} \int e^{-|\rho| S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t))} \mathbf{dx}}{\int e^{-|\rho| S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t)} \mathbf{dx}}
\end{aligned}$$

The support space of the integral is $\mathbf{dx}$ with $\mathbf{dx} = d\mathbf{x}_{t_{i+1}}, ..., d\mathbf{x}_{t_N}$. Under the assumption that the quantities

$e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))}$ and $\nabla_{\mathbf{x}} e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))}$ are jointly continuous we will have that:

$$\nabla_{\mathbf{x}_{t_i}}\xi(\mathbf{x}) = \frac{\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))}\nabla_{\mathbf{x}_{t_i}} S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))\mathbf{dx}}{\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t)}\mathbf{dx}} =$$

$$\mathcal{E}_{P_k}\Big(\nabla_{\mathbf{x}_{t_i}} S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))\Big) =$$

$$\mathcal{E}_{P_k}\Big(\nabla_{\mathbf{x}_{t_i}} q(\mathbf{x})\delta t + \nabla_{\mathbf{x}_{t_i}}\mu(\mathbf{x})^T\mathbf{\Sigma}^{-1}(\mu(\mathbf{x}) + \boldsymbol{\mathcal{B}}\mathbf{u}_k(\mathbf{x},t))dt\Big)$$

The probability P_k is defined as follows: $P_k = \frac{e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))}}{\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t)}\mathbf{dx}}$. The quantity $\nabla_{\mathbf{x}_{t_i}}\mu(\mathbf{x})$ is equal to $\nabla_{\mathbf{x}_{t_i}}\mu(\mathbf{x}) = \frac{1}{\delta t}I + \nabla_{\mathbf{x}_{t_i}}\mathbf{f}(\mathbf{x}) + \boldsymbol{\mathcal{B}}\nabla_{\mathbf{x}_{t_i}}\mathbf{u}(\mathbf{x})$ after substituting back we will have:

$$\nabla_{\mathbf{x}_{t_i}}\xi(\mathbf{x}) = \mathcal{E}_{P_k}\Big(\nabla_{\mathbf{x}} q(\mathbf{x})dt\Big)$$

$$+ \mathcal{E}_{P_k}\Big(\big(-I + \nabla_{\mathbf{x}_{t_i}}\mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}\nabla_{\mathbf{x}_{t_i}}\mathbf{u}(\mathbf{x})dt\big)\mathbf{\Sigma}^{-1}\mu(\mathbf{x})\Big)$$

$$+ \mathcal{E}_{P_k}\Big(\big(-I + \nabla_{\mathbf{x}_{t_i}}\mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}\nabla_{\mathbf{x}_{t_i}}\mathbf{u}(\mathbf{x})dt\big)\mathbf{\Sigma}^{-1}\boldsymbol{\mathcal{B}}\mathbf{u}_k(\mathbf{x},t)\Big)$$

The optimal controls are given by:

$$\mathbf{u}_{k+1}(\mathbf{x},t)dt = -\mathbf{R}^{-1}\boldsymbol{\mathcal{B}}^T\nabla_{\mathbf{x}}\xi(\mathbf{x})dt$$

$$= \mathbf{R}^{-1}\boldsymbol{\mathcal{B}}^T\mathcal{E}_{P_k}\Big(\mathbf{\Sigma}^{-1}\boldsymbol{\mathcal{B}}\mathbf{u}_k(\mathbf{x},t)dt + \mathbf{\Sigma}^{-1}\mu(\mathbf{x})dt\Big) + O(dt^2)$$

$$= \mathbf{R}^{-1}\boldsymbol{\mathcal{B}}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mathcal{B}}\mathcal{E}_{P_k}\Big(\mathbf{u}_k(\mathbf{x},t)dt + \frac{1}{\sqrt{\rho}}d\mathbf{w}_k(t)\Big)$$

$$= \mathcal{E}_{P_k}\Big(\mathbf{u}_k(\mathbf{x},t)dt + \frac{1}{\sqrt{\rho}}d\mathbf{w}_k(t)\Big)$$

Because $\lim_{dt\to 0} O(dt^2) = 0$. Since $\mathbf{R} = I$ and $\mathbf{\Sigma} = \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T$ and $\boldsymbol{\mathcal{B}}$ is invertible. The feedback policy $\mathbf{u}_k(\mathbf{x},t)$ is evaluated at the current state $\mathbf{x}$ we have (23). ■

There are stochastic dynamical systems in which the control and diffusion matrices are partitioned such that $\boldsymbol{\mathcal{B}} = [0^T,\ \boldsymbol{\mathcal{B}}_c^T]^T$ with $\boldsymbol{\mathcal{B}}_c$ invertible, while the drift term can also be partitioned accordingly $\mathbf{f} = [\mathbf{f}_m^T,\ \mathbf{f}_c^T]^T$. In [5] it has been show that the path integral formulation is expressed as in (20) with $\boldsymbol{\mathcal{B}}_c d\mathbf{w}_k = \sqrt{\rho}\mu(\mathbf{x})dt$, $\mu(\mathbf{x}) = \big(\frac{\delta \mathbf{x}_c}{\delta t} - \mathbf{f}_c(\mathbf{x}) - \boldsymbol{\mathcal{B}}_c\mathbf{u}_k\big)$ and $\Sigma_c = \boldsymbol{\mathcal{B}}_c\boldsymbol{\mathcal{B}}_c^T$. Our analysis in theorem 1 holds also for the aforementioned types of systems.

V. Derivation based on Bellman Principle

We consider stochastic optimal control in the classical sense, as a constrained optimization problem, with the cost function under minimization given by the mathematical expression:

$$V(\mathbf{x}) = \min_{\mathbf{u}} E\Big[J(\mathbf{x},\mathbf{u})\Big] = \min_{\mathbf{u}} E\Big[\int_{to}^{t_N} \boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},t)dt\Big]$$

subject to the nonlinear stochastic dynamics:

$$\mathbf{dx} = \mathbf{F}(\mathbf{x},\mathbf{u})dt + \mathbf{B}(\mathbf{x})d\mathbf{w} \tag{24}$$

with $\mathbf{x} \in \Re^{n\times 1}$ denoting the state of the system, $\mathbf{u} \in \Re^{p\times 1}$ the control vector and $d\mathbf{w} \in \Re^{p\times 1}$ brownian noise. The function $\mathbf{F}(\mathbf{x},\mathbf{u})$ is a nonlinear function of the state $\mathbf{x}$ and affine in controls $\mathbf{u}$ and therefore is defined as $\mathbf{F}(\mathbf{x},\mathbf{u}) = \mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}$. The matrix $\mathbf{G}(\mathbf{x}) \in \Re^{n\times p}$ is the control matrix, $\mathbf{B}(\mathbf{x}) \in \Re^{n\times p}$ is the diffusion matrix and $\mathbf{f}(\mathbf{x}) \in \Re^{n\times 1}$ are the passive dynamics. The cost function $J(\mathbf{x},\mathbf{u})$ is a function of states and controls. Under the optimal controls $\mathbf{u}^*$ the cost function is equal to the value function $V(\mathbf{x})$. The term $\boldsymbol{\mathcal{L}}$(x,u,t) is the running cost and it is expressed as:

$$\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},t) = q_0(\mathbf{x},t) + q_1(\mathbf{x},t)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbf{R}\mathbf{u} \tag{25}$$

Essentially, the running cost has three terms, the first $q_0(\mathbf{x}_t,t)$ is a state-dependent cost, the second term depends on states and controls and the third is the control cost with the term $\mathbf{R} > 0$ the corresponding weight. The stochastic HJB equation [12], [20] associated with this stochastic optimal control problem is expressed as follows:

$$-\partial_t V = \min_{\mathbf{u}}\Big(\boldsymbol{\mathcal{L}} + (\nabla_{\mathbf{x}}V)^T\mathbf{F} + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}V)\mathbf{B}\mathbf{B}^T\right)\Big) \tag{26}$$

To find the minimum, the cost function (25) is inserted into (26) and the gradient of the expression inside the parenthesis is taken with respect to controls $\mathbf{u}$ and set to zero. The corresponding optimal control is given by the equation:

$$\mathbf{u}(\mathbf{x}_t) = -\mathbf{R}^{-1}\Big(q_1(\mathbf{x},t) + \mathbf{G}(\mathbf{x})^T\nabla_{\mathbf{x}}V(\mathbf{x},t)\Big) \tag{27}$$

These optimal controls will push the system dynamics in the direction opposite that of the gradient of the value function $\nabla_{\mathbf{x}}V(\mathbf{x},t)$. The value function satisfies nonlinear, second-order PDE:

$$-\partial_t V = \tilde{q} + (\nabla_{\mathbf{x}}V)^T\tilde{\mathbf{f}} - \underline{\frac{1}{2}(\nabla_{\mathbf{x}}V)^T\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T(\nabla_{\mathbf{x}}V)}$$
$$+ \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}V)\mathbf{B}\mathbf{B}^T\right) \tag{28}$$

with $\tilde{q}(\mathbf{x},t)$ and $\tilde{\mathbf{f}}(\mathbf{x},t)$ defined as $\tilde{q}(\mathbf{x},t) = q_0(\mathbf{x},t) - \frac{1}{2}q_1(\mathbf{x},t)^T\mathbf{R}^{-1}q_1(\mathbf{x},t)$ and $\tilde{\mathbf{f}}(\mathbf{x},t) = \mathbf{f}(\mathbf{x},t) - \mathbf{G}(\mathbf{x},t)\mathbf{R}^{-1}q_1(\mathbf{x},t)$ and the boundary condition $V(\mathbf{x}_{t_N}) = \phi(\mathbf{x}_{t_N})$. Given the exponential transformation $V(\mathbf{x},t) = -\lambda\log\Psi(\mathbf{x},t)$ and the assumption $\lambda\mathbf{G}(\mathbf{x})\mathbf{R}^{-1}\mathbf{G}(\mathbf{x})^T = \mathbf{B}(\mathbf{x})\mathbf{B}(\mathbf{x})^T = \mathbf{\Sigma}(\mathbf{x}_t) = \mathbf{\Sigma}$ the resulting PDE is formulated as follows:

$$-\partial_t\Psi = -\frac{1}{\lambda}\tilde{q}\Psi + \tilde{\mathbf{f}}^T(\nabla_{\mathbf{x}}\Psi) + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}\Psi)\mathbf{\Sigma}\right) \tag{29}$$

with boundary condition: $\Psi(\mathbf{x}(t_N)) = \exp\big(-\frac{1}{\lambda}\phi(\mathbf{x}(t_N))\big)$. By applying the Feynman-Kac lemma to the Chapman-Kolmogorov PDE (29) yields its solution in form of an expectation over system trajectories. This solution is mathematically expressed as:

$$\Psi\left(\mathbf{x}_{t_i}\right) = E^{(0)}_{\boldsymbol{\tau}_i}\Big[\exp\Big(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x})dt\Big)\Psi(\mathbf{x}_{t_N})\Big] \tag{30}$$

The expectation $E^{(0)}_{\boldsymbol{\tau}_i}$ is taken on sample paths $\boldsymbol{\tau}_i = (\mathbf{x}_i,...,\mathbf{x}_{t_N})$ generated with the forward sampling of the

uncontrolled diffusion equation $\mathbf{dx} = \tilde{f}(\mathbf{x}_t)\delta t + \mathbf{B}(\mathbf{x})d\mathbf{w}$. The expectation $E^{(1)}_{\boldsymbol{\tau}_i}$ above, is evaluated on trajectories generated with forward sampling of the controlled diffusion in (24). The optimal controls are specified as:

$$\mathbf{u}_{PI}(\mathbf{x}) = -\mathbf{R}^{-1}\left(q_1(\mathbf{x},t) - \lambda\mathbf{G}(\mathbf{x})^T\frac{\nabla_{\mathbf{x}}\Psi(\mathbf{x},t)}{\Psi(\mathbf{x},t)}\right)$$

Since, the initial value the function $V(\mathbf{x},t)$ is the minimum of the expectation of the objective function $J(\mathbf{x},\mathbf{u})$ subject to controlled stochastic dynamics in (24), it can be trivially shown that:

$$\begin{aligned} V(\mathbf{x},t_i) &= -\lambda\log E^{(0)}_{\boldsymbol{\tau}_i}\left[\exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x})dt\right)\Psi(\mathbf{x}_{t_N})\right] \\ &\leq E^{(1)}_{\boldsymbol{\tau}_i}\left(J(\mathbf{x},\mathbf{u})\right) \end{aligned} \tag{31}$$

Note that the inequality above in similar to (12) when the following equations hold:

$$q_1(\mathbf{x}) = 0, \quad \mathbf{R} = I, \quad \lambda = \frac{1}{|\rho|}, \mathbf{G} = \boldsymbol{\mathcal{B}}, \mathbf{B} = \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}} \tag{32}$$

The first three equalities guarantee that $J(\mathbf{x},\mathbf{u}) = \mathcal{J}(\mathbf{x}) - \frac{|\rho|}{\rho}\int_{t_i}^{t_N}\mathbf{u}^T\mathbf{u}dt$ are identical, and the last two equalities make sure that the expectations are evaluated under the same diffusions and therefore $\mathcal{E}^{(0)}_{\boldsymbol{\tau}_i} \equiv E^{(0)}_{\boldsymbol{\tau}_i}$ and $\mathcal{E}^{(1)}_{\boldsymbol{\tau}_i} \equiv E^{(1)}_{\boldsymbol{\tau}_i}$. Under the conditions above the Kolmogorov PDEs (14) and (29) and the HJB equations (28) and (15) are identical.

VI. Markov Jump Diffusions Processes

We consider the one-dimensional uncontrolled and controlled markov jump diffusions processes specified as $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}(\mathbf{x})d\mathbf{w}^{(0)}(t) + \mathbf{h}(\mathbf{x})d\mathbf{P}^{(0)}(t)$ and $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})(\mathbf{u}dt + \frac{1}{\sqrt{|\rho|}}d\mathbf{w}^{(1)}(t)) + \mathbf{h}(\mathbf{x})d\mathbf{P}^{(1)}(t)$, with $\mathbf{x}_t \in \Re^{1\times 1}$ denoting the state of the system, $\mathbf{B}(\mathbf{x},t) \in \Re^{1\times 1}$ the diffusion-control transition matrix, $\mathbf{f}(\mathbf{x},t) \in \Re^{1\times 1}$ the passive dynamics, $\mathbf{u}_t \in \Re^{1\times 1}$ the control vector and $d\mathbf{w} \in \Re^{1\times 1}$ brownian noise. The term $P(t) \in \Re^{1\times 1}$ is Poisson distributed and $\mathbf{h}(\mathbf{x},t) \in \Re^{1\times 1}$ is the jump-amplitude or the Poisson process coefficient with $E\left(d\mathbf{P}(t)^{(i)}\right) = \nu^{(i)}\delta t$ and $\text{Var}\left(d\mathbf{P}(t)^{(i)}\right) = \nu^{(i)}\delta t$ for $i = 0, 1$. The term $\nu(t) > 0$ is the ith jump rate or jump density and $\nu\delta t$ is the mean count of the Poisson process in the time interval $(t, t+dt]$. Poisson processes obey the Markov property while they also have independent increments and thus $\text{Cov}\left[d\mathbf{P}(t)(t_j)d\mathbf{P}(t)(t_k)\right] = \nu(t_j)dt\delta_{k,j}$. Based on Girsanov's theorem [24] for markov jump diffusion processes, the Radon-Nikodým derivative is now specified as $\frac{d\mathbb{P}}{d\mathbb{Q}} = \exp\left(-\zeta(\mathbf{u})\right)$ with $\zeta(\mathbf{u})$ defined as follows:

$$\zeta(\mathbf{u}) = \int_{t_i}^{t_N}\frac{1}{2}|\rho|\mathbf{u}(t)^2dt + \sqrt{|\rho|}\int_{t_i}^{t_N}\mathbf{u}(t)d\mathbf{w}^{(1)}(t) + \mathcal{V}(\gamma^{(J)})$$

$\mathcal{V}(\gamma^{(J)}(t)) = \int_{t_i}^{t_N}\left(\left(\gamma^{(J)}(t) - 1\right)\nu_0(t)\right)\delta t + \sum_{j=1}^{\mathbf{P}^{(1)}(t)}\log\gamma^{(J)}(t)$ and $\gamma^{(J)}(t) = \frac{\nu^{(1)}(t)}{\nu^{(0)}(t)}$. The lower bound on the value function is now derived by incorporating the Radon-Nikodým derivative into (12).

$$\begin{aligned} \xi(\mathbf{x}) &= \frac{1}{\rho}\log\mathcal{E}^{(0)}\left[\exp\left(\rho\mathcal{J}(\mathbf{x})\right)\right] \leq \mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}\left[\mathcal{J}(\mathbf{x}) - \frac{1}{\rho}\zeta(\mathbf{u})\right] \\ &\leq \mathcal{E}^{(1)}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N}\mathbf{u}(t)^2\delta t + \rho\mathcal{V}(\gamma^{(J)}(t))\right] \end{aligned}$$

Thus we will have:

$$\boxed{\xi_J(\mathbf{x}) \leq \mathcal{E}^{(1)}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N}\mathbf{u}(t)^2dt\right]}$$

The new bound under sampling based on markov jump diffusion processes is defined by the equation $\xi_J(\mathbf{x}) = \xi(\mathbf{x}) - \mathcal{E}^{(1)}\left(\rho\mathcal{V}(\gamma^{(J)}(t))\right)$. For the cases where the change of measure between the control and uncontrolled markov jump diffusion includes only changes in the drift $\gamma^{(J)}(t) = 1$, the bound above simplifies to:

$$\xi(\mathbf{x}) = \xi_J(\mathbf{x}) \leq \mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N}\mathbf{u}(t)^2dt\right]$$

Thus when the change of measure in the markov jump diffusion process is only due to the change in the drift, the corresponding bound of the cost function has the same formulation with the one derived for diffusion processes.

VII. Connections to KL control

In the KL control framework [10], [11], [25] the analysis starts with the application of the Bellman principle of optimality on Markov Decision Processes (MDP) and under the running cost specified as a sum of a state depended term and the Kullback Leibler Divergence between the transition densities of the controlled and uncontrolled dynamics. In particular, the running cost is specified as $L(\mathbf{x},\mathbf{u}) = q(\mathbf{x}) + \mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right) = q(\mathbf{x}) + \mathcal{E}^{(1)}\left(\log\frac{p(\mathbf{x}'|\mathbf{x},\mathbf{u})}{p(\mathbf{x}'|\mathbf{x})}\right)$. The transition probabilities under the controlled and uncontrolled dynamics are represented as $p(\mathbf{x}'|\mathbf{x},\mathbf{u})$ and $p(\mathbf{x}'|\mathbf{x})$. Application of the Bellman principle of optimality results in the minimization of the quantity:

$$V_t(\mathbf{x}) = \min_{\mathbf{u}\in\mathcal{U}}\left(q(\mathbf{x}) + \mathcal{E}^{(1)}\left(\log\frac{p(\mathbf{x}'|\mathbf{x},\mathbf{u})}{p(\mathbf{x}|\mathbf{x})} + V_{t+1}(\mathbf{x}')\right)\right)$$

Depending on the stochastic optimal control problem $w(\mathbf{x}')$ is equal to $V(\mathbf{x}'), \alpha V(\mathbf{x}'), V_{t+1}(\mathbf{x}')$. For our presentation here we choose $w(\mathbf{x}') = V_{t+1}(\mathbf{x}')$ that corresponds to finite horizon case. The $\mathbf{u}$ dependent terms in the functional above are minimized and thus we will have that:

$$\begin{aligned} &\mathcal{E}^{(1)}\left(\log\frac{p(\mathbf{x}'|\mathbf{x},\mathbf{u})}{p(\mathbf{x}'|\mathbf{x})} + V_{t+1}(\mathbf{x}')\right) = \\ &\mathcal{E}^{(1)}\left(\log\frac{p(\mathbf{x}'|\mathbf{x},\mathbf{u})}{p(\mathbf{x}'|\mathbf{x})\exp\left(-V_{t+1}(\mathbf{x}')\right)}\right)\Bigg) \end{aligned}$$

For the purposes the normalization term $\mathcal{G}[\Phi](\mathbf{x})$ is introduced with $\Phi(\mathbf{x}) = \exp(-w(\mathbf{x}'))$ being the *desirability* function defined as $\mathcal{G}[\Phi](\mathbf{x}) = \sum p(\mathbf{x}'|\mathbf{x})\Phi(\mathbf{x}') = \mathcal{E}^{(0)}\Big(\Phi(\mathbf{x}')\Big)$, we will have that:

$$\mathcal{E}^{(1)}\Bigg(\log\frac{\mathbf{u}(\mathbf{x}^{'}|\mathbf{x})}{p(\mathbf{x}'|\mathbf{x})} + V_{t+1}(\mathbf{x}')\Bigg) = \\ -\log\mathcal{G}[\Phi](\mathbf{x}) + \mathcal{H}\Bigg(p(\mathbf{x}'|\mathbf{x},\mathbf{u})\Bigg|\Bigg|\frac{\mathbf{p}(\mathbf{x}'|\mathbf{x})\Phi(\mathbf{x}')}{\mathcal{G}[\Phi](\mathbf{x})}\Bigg)$$

Substitution of the expression above into the Bellman minimization equation results in:

$$\min_{\mathbf{u}\in\mathcal{U}}\Bigg(q(\mathbf{x}) - \log\mathcal{G}[\Phi](\mathbf{x}) + \mathcal{H}\Bigg(\mathbf{u}(\mathbf{x}|\mathbf{x})\Bigg|\Bigg|\frac{\mathbf{p}(\mathbf{x}'|\mathbf{x})\Phi(\mathbf{x}')}{\mathcal{G}[\Phi](\mathbf{x})}\Bigg)\Bigg)$$

The minimum of the Bellman equation is attained by:

$$\boxed{p^*(\mathbf{x}'|\mathbf{x},\mathbf{u}) = \frac{\mathbf{p}(\mathbf{x}'|\mathbf{x})\Phi(\mathbf{x})}{\mathcal{G}[\Phi](\mathbf{x})}}$$

The equation above provides the transition probability under the optimal control law and in that sense it the optimal transition probability. Clearly the optimal distribution above is identical to equations (6) and (13). Substitution of the optimal distribution above will result in the Bellman equation:

$$\Phi(\mathbf{x}) = \exp(-q(\mathbf{x}))\mathcal{G}[\Phi](\mathbf{x}')$$

The link with the continuous case is established by writing the Bellman equation for an MDP with continuous state space.

$$\Phi_{(\delta t)}(\mathbf{x}) = \exp(-q(\mathbf{x})\delta t)\mathcal{G}[\Phi_{(\delta t)}](\mathbf{x}')$$

Rearrangement of the terms results in:

$$\Bigg(\exp(q(\mathbf{x})\delta t) - 1\Bigg)\Phi_{(\delta t)}(\mathbf{x}) = \mathcal{E}^{(0)}\Bigg(\Phi_{(\delta t)}(\mathbf{x}') - \Phi_{(\delta t)}(\mathbf{x})\Bigg)$$

Under the limit $\delta \to 0$ the equation results the backward Chaplman Kolmogorov PDE in (29) for $\rho = 1$.

A. The compositionally of optimal controls

In the KL optimal control framework the optimal control is constructed [25] as the convex combination of the K optimally controlled distributions $p_k^*(\mathbf{x}'|\mathbf{x},\mathbf{u})$. These optimally controlled distributions correspond to K optimal control problems which differ only at the terminal cost. Thus the optimal control takes the form: $\mathbf{u}^*(\mathbf{x}) = \sum_k m_k(\mathbf{x})p_k^*(\mathbf{x}'|\mathbf{x},\mathbf{u})$ with the mixing term $m_k(\mathbf{x})$ defined as: $m_k(\mathbf{x}) = \frac{w_k\mathcal{G}[\Phi](\mathbf{x})}{\sum_s w_s\mathcal{G}[\Phi](\mathbf{x})}$. For continuous time optimal control problems the compositionally control law is expressed as $\mathbf{u}^*(\mathbf{x}) = \sum_k m_k(\mathbf{x})\Bigg(\mathbf{R}^{-1}\mathbf{G}^T\frac{\nabla\Phi_k(\mathbf{x})}{\Phi_k(\mathbf{x})}\Bigg)$.

VIII. Discussion

This work shows the connection of path integral control framework as presented in the machine learning and robotic communities [1], [2], [4], [5], [7], [10] with work in the control theoretic community on risk sensitivity [12], [13], [15], [16]. Essentially there are two methodological approaches to derive the path integral framework. In the first, stochastic optimal control is specified as minimization of the objective $E^{(1)}_{\boldsymbol{\tau}_i}(J(\mathbf{x},\mathbf{u}))$ subject to the controlled dynamics. The HJB PDE is derived based on the Bellman principle of optimality. The exponential transformation of the value function $V(\mathbf{x})$ and the connection between control cost and variance result in the transformation of the HJB in to the backward Chapman Kolmogorov. The Feynman-Kac lemma is applied and the solution of the Chapman Kolmogorov PDE together with the lower bound on the objective function are provided. The second methodological approach starts with the duality between free energy and relative entropy and the resulting optimization problem as expressed in (4). For diffusion processes affine in control and noise and under the use of Girsanov's theorem, the aforementioned optimization results in formulating the bound $\xi(\mathbf{x})$ of the objective function $\mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}(J(\mathbf{x},\mathbf{u}))$ which is typically found in stochastic optimal control. The link to Bellman optimality is established by showing that, this bound $\xi(\mathbf{x})$ satisfies the HJB equation and therefore it is a value function.

Inside the class of the stochastic dynamics of markov diffusion processes affine in control and noise, Dynamic Programming is more general since (see conditions (32)) it incorporates general cost functions and stochastic dynamics. This generalization however, is reduced by the assumption regarding control cost and the variance of the noise $\lambda\mathbf{G}(\mathbf{x})\mathbf{R}^{-1}\mathbf{G}(\mathbf{x})^T = \mathbf{B}(\mathbf{x})\mathbf{B}(\mathbf{x})^T$.

In the second approach the lower bound $\xi(\mathbf{x})$ of the accumulated trajectory cost $\mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}(J(\mathbf{x},\mathbf{u}))$ is derived without relying on the Bellman Principle. In fact, this lower bound defines a new form of optimality which, as it is shown in [12], [13] as well as in this work, for the case of diffusion processes is equivalent to the Bellman principle of optimality. Here we derived the lower bound of the cost of stochastic optimal control for nonlinear markov jump diffusion processes. We show that the form of the lower bound remains similar with the case of diffusion processes for as long the change in the probability measure is only due to the changes in the drift of the dynamics of the markov jump diffusion process.

In the KL stochastic optimal control framework the derivation relies on the Bellman Principle of Optimality in discrete time. The resulting distribution $p_k^*(\mathbf{x}'|\mathbf{x},\mathbf{u})$ is optimal since it is the distribution that results when actions are optimal. In that sense the KL framework, in its initial formulation [10] does not explicitly provide an optimal control law but instead it provides the optimal distribution or optimal transition probability under the use of optimal control law. For the case of control affine diffusions, KL control framework incorporates control-only and state-only depended terms in

contrast to PI derived based on the Bellman principle in which cross terms between controls and states may be considered. The compositionally of optimal controls includes control laws and thus it can incorporate any analytically derived optimal control as well as PI control. The strength of KL control framework is in its generalizability. A shown in [10] the KL control is applicable to different forms of stochastic optimal control problems which include finite, infinite horizon, discounted and first exit optimal control. It remains an open question whether this level of generizability is achieved in continuous time since one can derive the optimal control law for infinite horizon by using risk sensitive cost functions as in [15].

IX. Appendix

We will the nonlinear diffusions $\mathbf{dx} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t)$ and $\mathbf{dx} = \mathbf{f}(\mathbf{x})dt + \mathbf{G}(\mathbf{x})\mathbf{u}dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t)$. We also have that $\mathbf{B}(\mathbf{x})\mathbf{L}\mathbf{L}^T\mathbf{B}(\mathbf{x})^T = \boldsymbol{\Sigma}_{\mathbf{w}}$. We consider the corresponding probability measures:

$$d\mathbb{P}\Big(\mathbf{x}_N, t_N|\mathbf{x}_i, t_i\Big) = \frac{\exp\left(-\frac{1}{2}\left(\int_{t_i}^{t_N} ||\mu_k(\mathbf{x})||^2_{\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}}dt\right)\right)}{(2\pi dt)^{m/2}|\boldsymbol{\Sigma}_{\mathbf{w}}|^{1/2}}d\mathbf{x}$$

$$d\mathbb{Q}\Big(\mathbf{x}_N, t_N|\mathbf{x}_i, t_i; \mathbf{u}_k\Big) = \frac{\exp\left(-\frac{1}{2}\left(\int_{t_i}^{t_N} ||\lambda_k(\mathbf{x})||^2_{\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}}dt\right)\right)}{(2\pi dt)^{m/2}|\boldsymbol{\Sigma}_{\mathbf{w}}|^{1/2}}d\mathbf{x}$$

with $\mu_k(\mathbf{x}) = \left(\frac{\delta\mathbf{x}}{\delta t} - \mathbf{f}(\mathbf{x},t)\right)$ thus $\mu_k(\mathbf{x})dt = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t)$ and $\lambda_k(\mathbf{x}) = \frac{\delta\mathbf{x}}{\delta t} - \mathbf{f}(\mathbf{x},t) - \mathbf{G}(\mathbf{x})\mathbf{u}_k(t) = \mu_k(\mathbf{x}) - \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)$ thus $\lambda_k(\mathbf{x})dt = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t)$. Also we have that $\lambda_k(\mathbf{x})dt = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) = \mu_k(\mathbf{x})dt - \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)dt$ and thus $\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t) - \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)dt$. Now we would like to find the expression:

$$\frac{d\mathbb{P}\Big(\mathbf{x}_N, t_N|\mathbf{x}_i, t_i\Big)}{d\mathbb{Q}\Big(\mathbf{x}_N, t_N|\mathbf{x}_i, t_i; \mathbf{u}_k\Big)} = \frac{\exp\left(-\frac{1}{2}\left(\int_{t_i}^{t_N} ||\mu_k(\mathbf{x})||^2_{\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}}dt\right)\right)}{\exp\left(-\frac{1}{2}\left(\int_{t_i}^{t_N} ||\lambda_k(\mathbf{x})||^2_{\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}}dt\right)\right)}$$

$$= \exp\left[-\frac{1}{2}\int_{t_i}^{t_N}\left(||\mu_k(\mathbf{x})||^2_{\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}} - ||\lambda_k(\mathbf{x})||^2_{\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}}\right)dt\right]$$

$$= \exp\left[-\frac{1}{2}\int_{t_i}^{t_N}\left(-\mathbf{u}_k(t)^T\mathbf{G}(\mathbf{x})^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{G}(\mathbf{x})\mathbf{u}_k(t)dt\right]\right.$$

$$\left.\times\exp\left[-\int_{t_i}^{t_N}\mathbf{u}_k(t)^T\mathbf{G}(\mathbf{x})^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t)\right)\right]$$

Since $\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t) - \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)dt$ then we will have that $\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t) = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) + \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)\delta t$. We are going to substitute the expression $\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t)$ with $\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) + \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)dt$ and thus the ratio of the probability measures is expressed as:

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \exp\left[-\frac{1}{2}\int_{t_i}^{t_N}\left(\mathbf{u}_k(t)^T\mathbf{G}(\mathbf{x})^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{G}(\mathbf{x})\mathbf{u}_k(t)\delta t\right]\right.$$

$$\left.\times\exp\left[\int_{t_i}^{t_N}\mathbf{u}_k(t)^T\mathbf{G}(\mathbf{x})^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t)\right)\right]$$

References

[1] H. J. Kappen, "An introduction to stochastic control theory, path integrals and reinforcement learning," in *Cooperative Behavior in Neural Systems* (J. Marro, P. L. Garrido, and J. J. Torres, eds.), vol. 887 of *American Institute of Physics Conference Series*, pp. 149–181, Feb. 2007.

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

[3] B. Van den Broek, W. Wiegerinck, and H. J. Kappen., "Graphical model inference in optimal control of stochastic multi-agent systems," *Journal of Artificial Intelligence Research*, vol. 32, no. 1, pp. 95–122, 2008.

[4] E. Theodorou, J. Buchli, and S. Schaal, "A generalized path integral approach to reinforcement learning," *Journal of Machine Learning Research*, no. 11, pp. 3137–3181, 2010.

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

[6] J. Buchli, F. Stulp, E. Theodorou, and S. Schaal, "Learning variable impedance control," *nternational journal of robotics research*, pp. 820–833, April 2011.

[7] J. Buchli, E. Theodorou, F. Stulp, and S. Schaal, "Variable impedance control - a reinforcement learning approach," in *Robotics: Science and Systems Conference (RSS)*, 2010.

[8] F. Stulp, J. Buchli, E. Theodorou, and S. Schaal, "Reinforcement learning of full-body humanoid motor skills," in *10th IEEE-RAS International Conference on Humanoid Robots*, 2010.

[9] 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.

[10] E. Todorov, "Efficient computation of optimal actions," *Proc Natl Acad Sci U S A*, vol. 106, no. 28, pp. 11478–83, 2009.

[11] E. Todorov, "Linearly-solvable markov decision problems," in *Advances in Neural Information Processing Systems 19 (NIPS 2007)* (B. Scholkopf, J. Platt, and T. Hoffman, eds.), (Vancouver, BC), Cambridge, MA: MIT Press, 2007.

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

[13] W. H. Fleming and H. M. Soner, *Controlled Markov processes and viscosity solutions*. Applications of mathematics, New York: Springer, 1nd ed., 1993.

[14] W. Fleming, "Exit probabilities and optimal stochastic control," *Applied Math. Optim*, vol. 9, pp. 329–346, 1971.

[15] W. H. Fleming and W. M. McEneaney, "Risk-sensitive control on an infinite time horizon," *SIAM J. Control Optim.*, vol. 33, pp. 1881–1915, November 1995.

[16] P. Dai Pra, L. Meneghini, and W. Runggaldier, "Connections between stochastic control and dynamic games," *Mathematics of Control, Signals, and Systems (MCSS)*, vol. 9, no. 4, pp. 303–326, 1996-12-08.

[17] S. K. Mitter and N. J. Newton, "A variational approach to nonlinear estimation," *SIAM J. Control Optim.*, vol. 42, pp. 1813–1833, May 2003.

[18] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus (Graduate Texts in Mathematics)*. Springer, 2nd ed., August 1991.

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

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

[21] M. Schulz, *Control Theory in Physics and other Fields of Science. Concepts, Tools and Applications*. Spinger, 2006.

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

[23] B. K. Oksendal, *Stochastic differential equations : an introduction with applications*. Berlin ; New York: Springer, 6th ed., 2003.

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

[25] E. Todorov, "Compositionality of optimal control laws," *In Advances in Neural Information Processing Systems*, vol. 22, pp. 1856–1864, 2009.