Relative Entropy and Free Energy Dualities: Connections to Path Integral Control and Applications to Tendon Driven Systems

Evangelos A. Theodorou[1], Mark Malhotra[1], Eric Rombokas[2] , Emo Todorov[1,3],

Abstract

This work integrates recent mathematical developments 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. Our analysis suggests an information theoretic view of nonlinear stochastic optimal control theory that does not rely on the Bellman principle of optimality and is applicable to general models of stochasticity.

We derive path integral optimal control and its iterative version by using the relationship between free energy and relative entropy for the special class of Markov diffusion processes. In contrast to previous work, the resulting formulation is valid for feedback policies without

1. Department of Computer Science and Engineering, University of Washington, Seattle, WA.
2. Department of Electrical Engineering, University of Washington, Seattle, WA.
3. Department of Applied Mathematics, University of Washington, Seattle, WA.
Correspondence Author: Evangelos Theodorou, E-mail: etheodor@cs.washington.edu,
see http://www.cs.washington.edu/homes/etheodor/

pre-specified policy parameterizations. The mathematical analysis is based on successive applications of Girsanov's theorem and the use of the Radon-Nikodým derivative for the case of markov diffusion processes. We compare optimal control policies derived based on the Dynamic Programming with control policies based on the duality between free energy and relative entropy. 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. We complete our analysis with an application to control of tendon driven robotic systems.

Index Terms

Stochastic Optimal Control, Information Theory, Path Integrals, Tendon Driven Robots

I. Introduction

Steering a dynamical system from an initial state to a target state under the minimization of a performance criterion is the topic of stochastic optimal control theory. Among the different fields of science and engineering control theory, machine learning and robotics played key role for the mathematical development of optimal control theory and its applicability to robotic systems. The challenges in applying stochastic optimal control to robotic systems are due to the nonlinear nature of dynamics, the dimensionality of the state space and very often the lack of accurate models for the dynamics and the environment. Recent developments on stochastic optimal control for nonlinear markov diffusions processes based on path integrals demonstrated remarkable applicability to robotic control and planning problems.

The framework of path integral (PI) stochastic optimal control was introduced in [14], [15]. In this work new insights regarding symmetry breaking phenomena and their connection to optimal control is presented. In [1], PI control framework was extended to stochastic optimal control problems for multi-agents systems. With the goal to build scalable algorithms applicable to robotic systems, the work [22], [23] derives PI control for the case of markov 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 desired trajectory and/or control gain representation. 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 [2], [3], [18], [21].

Parallel to the mathematical developments in continuous time, in [24], [26] 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 of KL control is general due to its applicability to a large class of stochastic optimal control problems such as finite, infinite horizon, exponentially discounted and first exit (see the supplementary material of [26]). In this work we present the connection of nonlinear stochastic optimal control theory with the duality between free energy and relative entropy [4] while demonstrating its applicability to robotics. More precisely the contributions of this work are summarized as follows:

- We derive the mathematical links of PI and KL control as presented in machine learning and robotics communities [14], [15], [24], [26] with earlier work in controls theory, by using the fundamental dualities between relative entropy and free energy and the logarithmic transformations of diffusions processes [4], [7]–[10]. The aforementioned connections provide an alternative view of stochastic optimal control theory that does not rely on the Bellman principle of optimality. We shown that it takes only the application of Girsanov's theorem and Jensen's inequality to derive a mathematical expression which results in computing the optimal cost-to-go. This computation involves the forward sampling of the uncontrolled dynamics and the evaluation of a *risk seeking* state dependent cost function on the resulting trajectories.

- We present PI control and its iterative version based on the fundamental dualities between free energy and relative entropy as applied to nonlinear markov diffusion processes. In contrast to previous work [23], 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. These 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 generizability in terms of types of assumptions, cost functions and forms of stochastic dynamics.
- We extend our analysis to stochastic optimal control for markov jump diffusion processes of one dimension based on the fundamental relationship between free energy and relative entropy and derive the corresponding bound on the cost function. The analysis relies again on Girsanov's theorem and the use of Radon-Nikodym derivative when jump and diffusion terms appear in the stochastic dynamics.
- We apply the iterative path integral control to a tendon driven robotic finger. The robotic finger is an anatomical correct testbed of the human finger that mimics its active and passive dynamics. We are inspired by biological motor control that

is capable of learning even movements containing contact transitions and unknown force requirements while adapting the impedance of the system. We seek to achieve robotic mimicry of this compliance, employing stiffness only when it is necessary for task completion. We demonstrate the simultaneous learning of feedback gains and desired tendon trajectories with iterative path integral control in a dynamically complex sliding-switch task for a tendon-driven robot hand. The learned controls look noisy but nonetheless result in smooth and expert task performance.

The paper is organized as follows: in Section III we provide the basic dualities between free energy and relative entropy. In Section IV we discuss how these dualities are linked to maximizing or minimizing stochastic optimal control problems for the case of diffusions processes. In Section V we derive the iterative case based on successive applications of Girsanov's theorem. In section VI 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 IV. In VI-A we derive the iterative path integral control based on the bellman principle. We expand our analysis on path integral control for the case of markov jump diffusions in Section VII. In section VIII we apply the iterative path integral control to control of a biomimetic tendon driven robotic finger. Finally in section IX we conclude by discussing the generalizability of the aforementioned approaches in terms of application to types of dynamical systems and cost functions.

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 [4]. This relationship is important for the derivation of stochastic optimal control. Let $(\mathcal{Z}, \boldsymbol{\mathcal{Z}})$ 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 call 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:

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

with $\mathcal{J}(\mathbf{x}) = \phi(\mathbf{x}_{t_N}) + \int_{t_i}^{t_N} q(\mathbf{x})dt$ is the state depended cost. The objective function above takes the form $\xi(\mathbf{x}) = \mathcal{E}^{(0)}_{\boldsymbol{\mathcal{T}}_i}\left(\mathcal{J}\right) + \frac{\rho}{2}Var\left(\mathcal{J}\right)$ 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 and the variance of $\mathcal{J}(\mathbf{x})$. 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 $\mathcal{E}^{(0)}_{\boldsymbol{\mathcal{T}}_i}$ taken under the measure $\mathbb{P}$ as a function of the expectation $\mathcal{E}^{(1)}_{\boldsymbol{\mathcal{T}}_i}$ taken under the probability measure $d\mathbb{Q}$. More precisely will have:

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

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

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

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 (10) 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 (9) becomes from $\leq$ to $\geq$ and the $\inf$ in (10) and (11) becomes $\sup$. In the next section we show how inequality (10) is transformed to a stochastic optimal control problem for the case of markov diffusion processes.

III. Basic Duality Relationships of Free Energy and relative entropy

In this section we derive the fundamental duality relationships between free energy and relative entropy [4]. 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:

$$\mathbb{E}\bigg(\mathcal{J}(\mathbf{x})\bigg) = \log\int \exp\left(\rho\mathcal{J}(\mathbf{x})\right)d\mathbb{P} \tag{7}$$

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:

$$\xi(\mathbf{x}) = \frac{1}{\rho}\mathbb{E}\bigg(\mathcal{J}(\mathbf{x})\bigg) = \frac{1}{\rho}\log \mathcal{E}^{(0)}_{\boldsymbol{\mathcal{T}}_i}\bigg[\exp\left(\rho\mathcal{J}(\mathbf{x})\right)\bigg] \tag{8}$$

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 $\xi(\mathbf{x}) = \mathcal{E}^{(0)}_{\boldsymbol{\mathcal{T}}_i}\left(\mathcal{J}\right) + \frac{\rho}{2}Var\left(\mathcal{J}\right)$ 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 $\mathcal{E}^{(0)}_{\boldsymbol{\mathcal{T}}_i}$ taken under the measure $\mathbb{P}$ as a function of the expectation $\mathcal{E}^{(1)}_{\boldsymbol{\mathcal{T}}_i}$ taken under the probability measure $d\mathbb{Q}$. More precisely will have:

$$\mathcal{E}^{(0)}_{\boldsymbol{\mathcal{T}}_i}\bigg[\exp\left(\rho\mathcal{J}(\mathbf{x})\right)\bigg] = \int \exp\left(\rho\mathcal{J}(\mathbf{x})\right)d\mathbb{P} = \int \exp\left(\rho\mathcal{J}(\mathbf{x})\right)\frac{d\mathbb{P}}{d\mathbb{Q}}d\mathbb{Q}$$

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

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

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

$$\xi(\mathbf{x}) = -\frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right) \leq \mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right) \tag{9}$$

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:

$$-\frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right) = \inf\left[\mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right)\right] \tag{10}$$

and the dual minimization:

$$-\frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right) = \inf\left[\mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right)\right] \tag{11}$$

The infimum in (10) is attained at $\mathbb{Q}^*$ given by:

$$d\mathbb{Q}^* = \frac{\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)d\mathbb{P}}{\int \exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)d\mathbb{P}} \tag{12}$$

When $\rho > 0$ the inequality in (9) becomes from $\leq$ to $\geq$ and the $\inf$ in (10) and (11) becomes $\sup$. In the next section we show how inequality (10) is transformed to a stochastic optimal control problem for the case of markov diffusion processes.

IV. 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 [4], [8]. We consider the uncontrolled and controlled stochastic dynamics of the form:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t) \tag{13}$$

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}dt + \frac{1}{\sqrt{|\rho|}}\mathbf{L}d\mathbf{w}^{(1)}(t)\right) \tag{14}$$

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 (14). 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 (9) and the definition of the Radon-Nikodým derivative:

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\zeta(\mathbf{u})\right) \quad \text{and} \quad \frac{d\mathbb{P}}{d\mathbb{Q}} = \exp\left(-\zeta(\mathbf{u})\right) \tag{15}$$

where according to Girsanov's theorem [16] (see also section X) adapted to the diffusion processes (13) and (14) the term $\zeta(\mathbf{u})$ is expressed as follows:

$$\zeta(\mathbf{u}) = \frac{1}{2}|\rho| \int_{t_i}^{t_N} \mathbf{u}^T\mathbf{u}dt + \sqrt{|\rho|} \int_{t_i}^{t_N} \mathbf{u}^T d\mathbf{w}^{(1)}(t) \tag{16}$$

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 (15) and (27) into inequality (9) gives the following result:

$$\xi(\mathbf{x}) = -\frac{1}{|\rho|} \log \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}\left[\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)\right] \leq \mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{|\rho|}\zeta(\mathbf{u})\right] \tag{17}$$

The expectation on the right side of the inequality in (17) is further simplified as follows:

$$\xi(\mathbf{x}) \leq \mathcal{E}^{(1)}_{\boldsymbol{\tau}_i}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N} \mathbf{u}^T\mathbf{u}dt\right] \tag{18}$$

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 (18) 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 (18) 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 (18). Essentially inequality (18) 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 (18) reaches its optimal $\xi(\mathbf{x})$. Under the optimal control $\mathbf{u}^*$ and according to (19) the optimal distribution takes the from:

$$d\mathbb{Q}^*(\mathbf{x}) = \frac{\exp\Bigg(-|\rho| \int q(\mathbf{x})dt \Bigg) d\mathbb{P}(\mathbf{x})}{\int \exp\Bigg(-|\rho| \int q(\mathbf{x})dt \Bigg) d\mathbb{P}(\mathbf{x})} \tag{19}$$

An important question to ask is what is the link between (18) 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 $\Phi(\mathbf{x}, t) = \mathcal{E}^{(0)}_{\boldsymbol{\mathcal{T}}_i}(\exp(\rho \mathcal{J}(\mathbf{x})))$. The Feynman-Kac lemma [11] tells us that this function satisfies the backward Chapman Kolmogorov PDE. Therefore the following equation holds:

$$-\partial_t \Phi = \rho q_0 \Phi + \mathbf{f}^T (\nabla_{\mathbf{x}} \Phi) + \frac{1}{2|\rho|} tr \left((\nabla_{\mathbf{xx}} \Phi) \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{B}}^T \right) \tag{20}$$

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 shown that $\xi(\mathbf{x})$ satisfies the nonlinear PDE:

$$-\partial_t \xi = q_0 + (\nabla_{\mathbf{x}} \xi)^T \mathbf{f} \underline{- \frac{1}{2} (\nabla_{\mathbf{x}} \xi)^T \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{B}}^T (\nabla_{\mathbf{x}} \xi)} + \frac{1}{2|\rho|} tr \left((\nabla_{\mathbf{xx}} \xi) \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{B}}^T \right) \tag{21}$$

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

$$-\partial_t \xi = q_0 + (\nabla_{\mathbf{x}} \xi)^T \mathbf{f} \underline{+ \frac{1}{2} (\nabla_{\mathbf{x}} \xi)^T \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{B}}^T (\nabla_{\mathbf{x}} \xi)} + \frac{1}{2|\rho|} tr \left((\nabla_{\mathbf{xx}} \xi) \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{B}}^T \right) \tag{22}$$

The nonlinear PDEs above corresponds to the HJB equation [20] for the case of the minimizing and maximizing optimal control problem with control weight $\mathbf{R} = I$ 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 $sign(x) = -1 \;\; \forall x < 0$ and $sign(x) = 1 \;\; \forall x > 0$. More precisely we will have:

Corollary 1: *Consider the expectation operators $\mathcal{E}^{(0)}$, $\mathcal{E}^{(1)}$ evaluated on state trajectories sampled according to (13) and (14) respectively. The function $\xi(\mathbf{x}, t)$ specified as:*

$$\xi(\mathbf{x}, t) = \frac{sign(\rho)}{|\rho|} \log \mathcal{E}^{(0)} \Big[\exp\left(sign(\rho) |\rho| \mathcal{J}(\mathbf{x}) \right) \Big] \tag{23}$$

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 (14).

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 (13) and evaluation of the expectation in (23) on the resulting state trajectories. To derive (23) it takes only the application of Girsanov's theorem and Jensen's inequality.

V. Feedback Control For Markov Diffusion Processes

There are different ways to make use of the fundamental inequality in (18) 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 (23) 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:*

$$S(\mathbf{x},\mathbf{u}_k) = \mathcal{J}(\mathbf{x}) + \frac{1}{2}\left(\eta(\mathbf{u}) + \int_{t_i}^{t_N}||\mu(\mathbf{x})||^2_{\boldsymbol{\Sigma}^{-1}}\delta t\right) \tag{24}$$

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:

$$\begin{aligned}\xi(\mathbf{x}) &= -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)d\mathbb{P} = -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)\frac{d\mathbb{P}}{d\mathbb{Q}}d\mathbb{Q}\\ &= -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x}) - \zeta(\mathbf{u})\right)d\mathbb{Q}\end{aligned} \tag{25}$$

The measure $d\mathbb{Q}$ takes the form of a path integral [19] and thus it is expressed as:

$$\mathbb{Q}\Bigg(\mathbf{x}_N, t_N|\mathbf{x}_i, t_i\Bigg) = \frac{\exp\left(-\frac{|\rho|}{2}\left(\int_{t_i}^{t_N}\mu(\mathbf{x})^T\boldsymbol{\Sigma}^{-1}\mu(\mathbf{x})dt\right)\right)}{(2\pi dt)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \tag{26}$$

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 [12], [17] will become equal to:

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

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

$$\eta(\mathbf{u}) = \int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{u}_k dt + \int_{t_i}^{t_N}2\mathbf{u}_k^T\boldsymbol{\mathcal{B}}^{-T}\mu(\mathbf{x})dt = \int_{t_i}^{t_N}\mathbf{u}^T\mathbf{u}dt + \frac{1}{\sqrt{|\rho|}}\int_{t_i}^{t_N}2\mathbf{u}^T d\mathbf{w}^{(1)}(t) \tag{27}$$

Substitution of the function above $\zeta(\mathbf{u})$ and the path integral into (41) results in the expression:

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

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:

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

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

■

After deriving lemma 1 we proceed wit the final result 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:

$$\boxed{\mathbf{u}_{k+1}(\mathbf{x},t)dt = \mathbf{u}_k(\mathbf{x},t)dt + \frac{1}{\sqrt{\rho}}\mathcal{E}_{p_k}\Big(d\mathbf{w}_k(t)\Big)} \tag{28}$$

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 (40).

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:

$$\nabla_{\mathbf{x}_{t_i}}\xi(\mathbf{x}_{t_i}) = -\frac{1}{|\rho|}\nabla_{\mathbf{x}_{t_i}}\Bigg(\log\int\exp\Bigg[-|\rho|S(\mathbf{x},\mathbf{u}_k)\Bigg]\mathbf{dx}\Bigg) = -\frac{1}{|\rho|}\frac{\nabla_{\mathbf{x}_{t_i}}\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))}\mathbf{dx}}{\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t)}\mathbf{dx}}$$

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:

$$\nabla_{\mathbf{x}_{t_i}} \xi(\mathbf{x}) = \frac{\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))} \nabla_{\mathbf{x}_{t_i}} S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t))\mathbf{dx}}{\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t)}\mathbf{dx}} = \mathcal{E}_{P_k}\Bigg(\nabla_{\mathbf{x}_{t_i}} S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t))\Bigg)$$
$$= \mathcal{E}_{P_k}\Bigg(\nabla_{\mathbf{x}_{t_i}} q(\mathbf{x})\delta t + \nabla_{\mathbf{x}_{t_i}} \mu(\mathbf{x})^T \mathbf{\Sigma}^{-1}(\mu(\mathbf{x}) + \boldsymbol{\mathcal{B}}\mathbf{u}_k(\mathbf{x}, t))dt\Bigg)$$

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:

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

The optimal controls are given by:

$$\mathbf{u}_{k+1}(\mathbf{x}, t)dt = -\mathbf{R}^{-1}\boldsymbol{\mathcal{B}}^T \nabla_{\mathbf{x}_{t_i}} \xi(\mathbf{x})dt = \mathbf{R}^{-1}\boldsymbol{\mathcal{B}}^T \mathcal{E}_{P_k}\Bigg(\mathbf{\Sigma}^{-1}\boldsymbol{\mathcal{B}}\mathbf{u}_k(\mathbf{x}, t)dt + \mathbf{\Sigma}^{-1}\mu(\mathbf{x})dt\Bigg) + O(dt^2)$$
$$= \mathbf{R}^{-1}\boldsymbol{\mathcal{B}}^T \mathbf{\Sigma}^{-1}\boldsymbol{\mathcal{B}}\mathcal{E}_{P_k}\Bigg(\mathbf{u}_k(\mathbf{x}, t)dt + \frac{1}{\sqrt{\rho}} d\mathbf{w}_k(t)\Bigg) = \mathcal{E}_{P_k}\Bigg(\mathbf{u}_k(\mathbf{x}, t)dt + \frac{1}{\sqrt{\rho}} d\mathbf{w}_k(t)\Bigg)$$

For the last two lines we make use of the fact that $\lim_{dt\to 0} O(dt^2) = 0$, $\mathbf{\Sigma} = \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T$. In addition from section (IV) we know that $\mathbf{R} = I$ and $\boldsymbol{\mathcal{B}}$ is invertible. The feedback policy $\mathbf{u}_k(\mathbf{x}, t)$ is evaluated at the current state $\mathbf{x}_{t_i}$ and so we have (28). ■

There are stochastic dynamical systems in which the control and diffusion matrices are partitioned such that $\boldsymbol{\mathcal{B}} = [0^T, \quad \boldsymbol{\mathcal{B}}_c^T]^T$ with $\boldsymbol{\mathcal{B}}_c$ invertible, while the drift term can also be partitioned accordingly $\mathbf{f} = [\mathbf{f}_m^T, \quad \mathbf{f}_c^T]^T$. In [22] it has been shown that the path integral formulation is expressed as in (26) 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 $\mathbf{\Sigma} = \boldsymbol{\mathcal{B}}_c \boldsymbol{\mathcal{B}}_c^T$. Our analysis in theorem 1 holds for the aforementioned types of systems as well

1) Initially the controls $\mathbf{u}_{k+1}(\mathbf{x}, t)$ are computed for any state $\mathbf{x}$ and time t and therefore the iterative path integral control policy is closed loop. The closed loop formulation requires 1) full observability of the state $\mathbf{x}$ at every time of the time horizon and 2) the necessary actuation capabilities to steer the dynamics towards each state $\mathbf{x}$ and time t so that the state space is fully sampled and the feedback policy is computed as a lookup table.
2) In many control and robotic applications there is no direct access to the full state vector due to sensing limitations. Moreover, measurements which are functions of the state are available and they can be used for building a cost function. In these cases (28) could be used in an open loop form in which the the dependence of $\mathbf{u}$ with respect to the state is dropped and $\mathbf{u}$ is only a function of time. More precisely, state trajectories are sampled from a starting state $\mathbf{x}_{t_0}$ towards the target state $\mathbf{x}^*_{t_N}$ and (28) is applied for every time $t_i \in [t_0, t_N]$ from $t_0 < t_1 < ... < t_N$. Table (I) illustrates the iterative path integral control algorithm derived based on the relationship between free energy and relative entropy in its open loop formulation.
3) The path integral control policy (28) only partially depends on the stochastic dynamics in (14). In particular, the optimal policy depends only on the control(=diffusion) matrices of the initial dynamics and not on the drift. This characteristic of the path integral control is important for robotic applications in which an accurate state space representation of the stochastic dynamics is not available. The algorithm becomes completely model free by augmenting the dynamics and instead of controlling the $\mathbf{u}$, the control variable is the change $\delta\mathbf{u}$ of the initial controls per time unit.
4) In many stochastic dynamical systems the control and diffusion matrices are partitioned such that $\boldsymbol{\mathcal{B}} = [0, \ \ \boldsymbol{\mathcal{B}}_c]$ while the drift term can also be partitioned accordingly $\mathbf{f} = [\mathbf{f}_m, \ \ \mathbf{f}_c]$. In [22] it has been show that the path integral formulation

is expressed as in (26) with $\boldsymbol{\mathcal{B}}_c d\mathbf{w}_k = \sqrt{\rho}\mu(\mathbf{x})\delta t$, $\mu(\mathbf{x}) = \left(\frac{\delta \mathbf{x}_c}{\delta t} - \mathbf{f}_c(\mathbf{x}) - \boldsymbol{\mathcal{B}}_c \mathbf{u}_k\right)$ and $\Sigma_c = \boldsymbol{\mathcal{B}}_c \boldsymbol{\mathcal{B}}_c^T$. This change results in substituting $\boldsymbol{\mathcal{B}}$ with $\boldsymbol{\mathcal{B}}_c$ in table I.

Overall the iterative path integral control is easy to implement while it provides the practitioner with the flexibility not to rely on models when these are not available. In the next section we derive the iterative path integral control based on the dynamic programming principle.

TABLE I: Iterative Path Integral Control Based on the Duality Between Free Energy and Relative Entropy.

- **Given**:
  - The cost term $q(\mathbf{x}_t)$, variance $\rho > 0$, initials controls $\mathbf{u}_0$
- **Repeat** until convergence of the trajectory cost R:
  - Create M roll-outs of the system by forward sampling of the diffusion $\mathbf{dx} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}_k dt + \frac{1}{\sqrt{\rho}}\mathbf{dw}^{(k)}(t)\right)$.
  - **For** $k = 1...M$, compute:
    - $\eta(\mathbf{u}, t_i, t_N)$ as in (27).
    - $S(\vec{\mathbf{x}}_{i,k}) = \phi_{t_N} + \sum_{j=i}^{N-1}\left(q(t_j)dt + \eta(\mathbf{u}, t_i, t_N)\right)$
    - $P\left(\vec{\mathbf{x}}_{i,k}\right) = \frac{e^{-\frac{1}{\lambda}S(\vec{\mathbf{x}}_{i,k})}}{\sum_{k=1}^{K}[e^{-\frac{1}{\lambda}S(\vec{\mathbf{x}}_{i,k})}]}$
  - **For** $i = 1...(N-1)$, compute:
    - $\delta\mathbf{u}_{\tilde{\mathbb{P}}} = E_P\left(d\mathbf{w}^{(k)}(t_i)\right)$
    - $\mathbf{u}_{k+1}(t_i)dt = \mathbf{u}_k(t_i)dt + \frac{1}{\sqrt{\rho}}\delta\mathbf{u}_{\tilde{\mathbb{P}}}$

VI. 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:

$$V(\mathbf{x}) = \min_{\mathbf{u}} E\left[J(\mathbf{x}, \mathbf{u})\right] = \min_{\mathbf{u}} E\left[\int_{to}^{t_N} \boldsymbol{\mathcal{L}}(\mathbf{x}, \mathbf{u}, t)dt\right] \tag{29}$$

subject to nonlinear stochastic dynamics as specified by the markov diffusion process:

$$\mathbf{dx} = \mathbf{F}(\mathbf{x},\mathbf{u})dt + \mathbf{B}(\mathbf{x})d\mathbf{w} \tag{30}$$

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 $\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},\mathrm{t})$ is the immediate cost and it is expressed as:

$$\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},t) = q_0(\mathbf{x},t) + q_1(\mathbf{x},t)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbf{R}\mathbf{u} \tag{31}$$

Essentially, the immediate cost has three terms, the first $q_0(\mathbf{x}_t,t)$ is an arbitrary state-dependent cost, the second term depends on states and controls and the third is the control cost with $\mathbf{R} > 0$ the corresponding weight. The stochastic HJB equation [9], [20] associated with this stochastic optimal control problem is expressed as follows:

$$-\partial_t V = \min_{\mathbf{u}} \left(\boldsymbol{\mathcal{L}} + (\nabla_{\mathbf{x}} V)^T \mathbf{F} + \frac{1}{2} tr\left((\nabla_{\mathbf{xx}} V)\mathbf{G}\mathbf{G}^T \right) \right) \tag{32}$$

To find the minimum, the cost function (31) is inserted into (32) 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:

$$\mathbf{u}(\mathbf{x}_t) = -\mathbf{R}^{-1}\left(q_1(\mathbf{x},t) + \mathbf{G}(\mathbf{x})^T \nabla_{\mathbf{x}} V(\mathbf{x},t) \right) \tag{33}$$

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:

$$-\partial_t V = \tilde{q} + (\nabla_{\mathbf{x}} V)^T \tilde{\mathbf{f}} - \underline{\frac{1}{2}(\nabla_{\mathbf{x}} V)^T \mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T(\nabla_{\mathbf{x}} V)} + \frac{1}{2} tr\left((\nabla_{\mathbf{xx}} V)\mathbf{B}\mathbf{B}^T \right) \tag{34}$$

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 $V(\mathbf{x}_{t_N}) = \phi(\mathbf{x}_{t_N})$. To transform the PDE above into a linear one, we use a exponential transformation of the value function $V(\mathbf{x},t) = -\lambda\log\Psi(\mathbf{x},t)$. Given this exponential transformation and by considering 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:

$$-\partial_t\Psi = -\frac{1}{\lambda}\tilde{q}\Psi + \tilde{\mathbf{f}}^T(\nabla_{\mathbf{x}}\Psi) + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}\Psi)\mathbf{\Sigma}\right) \tag{35}$$

with boundary condition: $\Psi(\mathbf{x}(t_N)) = \exp\left(-\frac{1}{\lambda}\phi(\mathbf{x}(t_N))\right)$. By applying the Feynman-Kac lemma to the Chapman-Kolmogorov PDE (35) yields its solution in form of an expectation over system trajectories. This solution is mathematically expressed as:

$$\Psi\left(\mathbf{x}_{t_i}\right) = E^{(0)}\left[\exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x})dt\right)\Psi(\mathbf{x}_{t_N})\right] \tag{36}$$

The expectation $E^{(0)}$ 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 $E^{(1)}$ above, is evaluated on trajectories generated with forward sampling of the controlled diffusion in (30). The optimal controls are specified as:

$$\mathbf{u}_{PI}(\mathbf{x}) = -\mathbf{R}^{-1}\left(q_1(\mathbf{x},t) - \lambda\mathbf{G}(\mathbf{x})^T\frac{\nabla_{\mathbf{x}}\Psi(\mathbf{x},t)}{\Psi(\mathbf{x},t)}\right) \tag{37}$$

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 (30), it can be trivially shown that:

$$V(\mathbf{x},t_i) = -\lambda\log E^{(0)}_{\boldsymbol{\tau}_i}\left[\exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x})dt\right)\Psi(\mathbf{x}_{t_N})\right] \leq E^{(1)}_{\boldsymbol{\tau}_i}\Big(J(\mathbf{x},\mathbf{u})\Big) \tag{38}$$

Note that the inequality above in similar to (18).

$$q_1(\mathbf{x}) = 0, \quad \mathbf{R} = I, \quad \lambda = \frac{1}{|\rho|}, \quad \mathbf{G}(\mathbf{x},t) = \boldsymbol{\mathcal{B}}(\mathbf{x},t) \;\text{ and }\; \mathbf{B}(\mathbf{x},t) = \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}(\mathbf{x},t) \tag{39}$$

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 $\mathcal{E}^{(0)}_{\boldsymbol{\tau}_i} \equiv E^{(0)}$ and $\mathcal{E}^{(1)} \equiv E^{(1)}$. Under the conditions above the Kolmogorov PDEs (20) and (35) and the HJB equations (34) and (21) are identical. All of the analysis in this section is summarized by the corollary that follows:

Corollary 2: *The path integral stochastic optimal control derived based on dynamic programming results in the same value function with the one derived based on the fundamental relationship between free energy and relative entropy under the conditions (39). Under the conditions in (39) $V(\mathbf{x}, t) = \xi(\mathbf{x}, t)$ and $\Psi(\mathbf{x}, t) = \Phi(\mathbf{x}, t)$.*

A. Optimal Control Derivation

We will follow the same arguments of the analysis in section V and derive the new versions of lemma 1 and theorem 1. We show that with respect to iterative path integral update rule given in (1) the resulting algorithm incorporates more general costs functions but also has additional assumptions between noise and control cost. For simplicity in our analysis we will assume that $\mathbf{G}(\mathbf{x}) = \mathbf{G}$ and $\mathbf{B}(\mathbf{x}) = \mathbf{B}$. More precisely we will have:

Lemma 2: *Consider the stochastic dynamics $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \mathbf{G}(\mathbf{x})\mathbf{u}_k dt + \mathbf{B}(\mathbf{x})d\mathbf{w}^{(1)}(t)$ with the control policy $\mathbf{u}_k(\mathbf{x}, t)$ at iteration k. When sampling from these dynamics, the value function $V(\mathbf{x}, t)$ in (38) takes the form:*

$$V(\mathbf{x}, t) = -\lambda \log \int \exp\left[-\frac{1}{\lambda} S(\mathbf{x}, \mathbf{u}_k(\mathbf{x}, t)) \right] \mathbf{dx}$$

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

$$S(\mathbf{x}, \mathbf{u}_k) = \mathcal{J}(\mathbf{x}) + \frac{1}{2} \left(\varpi(\mathbf{u}) + \int_{t_i}^{t_N} ||\mu(\mathbf{x})||^2_{\boldsymbol{\Sigma}^{-1}} \delta t \right) \tag{40}$$

The term $\zeta(\mathbf{u})$ in the path cost above is defined as $\varpi(\mathbf{u}) = \int_{t_i}^{t_N} \mathbf{u}^T \mathbf{G}^T \left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T \right)^{-1} \mathbf{G}\mathbf{u}\delta t + \int_{t_i}^{t_N} 2\mathbf{u}^T \mathbf{G}^T \left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T \right)^{-1} \mathbf{B} d\mathbf{w}^{(1)}(t)$ and terms $\mu(\mathbf{x}) = \left(\frac{\delta \mathbf{x}}{\delta t} - \mathbf{f}(\mathbf{x}) - \boldsymbol{\mathcal{B}}\mathbf{u}_k \right)$.

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:

$$\begin{aligned} V(\mathbf{x},t) &= -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)d\mathbb{P} = -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)\frac{d\mathbb{P}}{d\mathbb{Q}}d\mathbb{Q} \\ &= -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x})-\zeta(\mathbf{u})\right)d\mathbb{Q} \end{aligned} \tag{41}$$

The measure $d\mathbb{Q}$ takes the form of a path integral [19] and thus it is expressed as: $\mathbb{Q}\Bigg(\mathbf{x}_N, t_N|\mathbf{x}_i, t_i\Bigg) = \frac{\exp\left(-\frac{|\rho|}{2}\left(\int_{t_i}^{t_N}\mu(\mathbf{x})^T\boldsymbol{\Sigma}^{-1}\mu(\mathbf{x})\delta t\right)\right)}{(2\pi\delta t)^{m/2}|\boldsymbol{\Sigma}|^{1/2}}$ where we use the fact that $\mathbf{B}d\mathbf{w}_k = \sqrt{\rho}\mu(\mathbf{x})\delta t$ and $\mu(\mathbf{x}) = \left(\frac{\delta\mathbf{x}}{\delta t} - \mathbf{f}(\mathbf{x}) - \mathbf{G}\mathbf{u}_k\right)$ and $\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^{-1}$. Based on the aforementioned inequalities the term $\zeta(\mathbf{u})$ in the Girsanov's theorem [12], [17] will become equal to:

$$\zeta(\mathbf{u}, t_i, t_N) = \frac{1}{2}\int_{t_i}^{t_N}\mathbf{u}^T\mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{u}\delta t + \int_{t_i}^{t_N}\mathbf{u}^T\mathbf{G}^T\boldsymbol{\Sigma}^{-1}\mathbf{B}d\mathbf{w}^{(1)}(t)$$

Since $\lambda\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T = \mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$ we will have that

$$\begin{aligned} \zeta(\mathbf{u}, t_i, t_N) &= \frac{1}{2\lambda}\int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{G}^T\left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T\right)^{-1}\mathbf{G}\mathbf{u}_k\delta t + \frac{1}{\lambda}\int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{G}^T\left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T\right)^{-1}\mu(\mathbf{x})\delta t \\ &= \frac{1}{2\lambda}\varpi(\mathbf{u}, t_i, t_N) \end{aligned}$$

with $\varpi(\mathbf{u}, t_i, t_N)$ defined as follows:

$$\begin{aligned} \varpi(\mathbf{u}, t_i, t_N) &= \int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{G}^T\left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T\right)^{-1}\mathbf{G}\mathbf{u}_k\delta t + \frac{1}{\lambda}\int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{G}^T\left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T\right)^{-1}\mu(\mathbf{x})\delta t \\ &= \int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{G}^T\left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T\right)^{-1}\mathbf{G}\mathbf{u}_k\delta t + \int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{G}^T\left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T\right)^{-1}\mathbf{B}d\mathbf{w}^{(1)}(t)\delta t \end{aligned} \tag{42}$$

Substitution of the function above $\zeta(\mathbf{u})$ and the path integral into 41 results in the expression:

$$V(\mathbf{x}) = -\lambda\log\int\exp\left(-\frac{1}{\lambda}\mathcal{J}(\mathbf{x}) - \zeta(\mathbf{u}_k)\right)d\mathbb{Q} = -\lambda\log\int\exp\left[-\frac{1}{\lambda}\left(\mathcal{J}(\mathbf{x}) + \frac{\varpi(\mathbf{u}) + \int_{t_i}^{t_N}||\mu(\mathbf{x})||^2_{\boldsymbol{\Sigma}^{-1}}\delta t}{2}\right]\right]\mathbf{dx}$$

with $\mathbf{dx}$ defined as $\mathbf{dx} = d\mathbf{x}_{t_{i+1}}, ..., d\mathbf{x}_{t_N}$.

■

Having derived lemma 2 we can now provide the optimal control. More precisely we have that:

Theorem 2: *Consider the stochastic optimal control problem:*

$$V(\mathbf{x}) = \min_{\mathbf{u}} E^{(1)} \left[\int_{to}^{t_N} q(\mathbf{x}) + \frac{1}{2}\mathbf{u}^T \mathbf{R}\mathbf{u} dt \right] \tag{43}$$

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:*

$$\mathbf{u}_{k+1} dt = -\mathbf{R}^{-1} q_1 dt + \boldsymbol{\Omega} \mathcal{E}_{P_k} \Bigg(\mathbf{G}\mathbf{u}_k dt + \mathbf{B} d\mathbf{w}(t) \Bigg) \tag{44}$$

with $\boldsymbol{\Omega} = \mathbf{R}^{-1}\mathbf{G}^T \left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T \right)^{-1}$ *and the term*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 (40).*

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:

$$\nabla_{\mathbf{x}_{t_i}} V(\mathbf{x}_{t_i}) = -\lambda \nabla_{\mathbf{x}_{t_i}} \Bigg(\log \int \exp \Bigg[-\frac{1}{\lambda} S(\mathbf{x}, \mathbf{u}_k) \Bigg] \mathbf{dx} \Bigg) = -\lambda \frac{\nabla_{\mathbf{x}_{t_i}} \int e^{-\frac{1}{\lambda} S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))} \mathbf{dx}}{\int e^{-|\rho| S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t)} \mathbf{dx}}$$

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 $e^{-\frac{1}{\lambda} S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))}$ and $\nabla_{\mathbf{x}} e^{-\frac{1}{\lambda} S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))}$ are jointly continuous we will have that:

$$\nabla_{\mathbf{x}_{t_i}} V(\mathbf{x}) = \mathcal{E}_{P_k} \Bigg(\nabla_{\mathbf{x}} q(\mathbf{x}) \delta t + \lambda \nabla_{\mathbf{x}} \mu(\mathbf{x})^T \boldsymbol{\Sigma}^{-1} (\mu(\mathbf{x}) + \mathbf{G}\mathbf{u}_k(\mathbf{x}, t)) \delta t \Bigg)$$

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

$$\begin{aligned}\nabla_{\mathbf{x}} V(\mathbf{x}) &= \mathcal{E}_{P_k} \Bigg(\nabla_{\mathbf{x}} q(\mathbf{x}) \delta t \Bigg) + \lambda \mathcal{E}_{P_k} \Bigg(\left(-I + \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}) \delta t + \mathbf{G} \nabla_{\mathbf{x}} \mathbf{u}(\mathbf{x}) \delta t \right) \boldsymbol{\Sigma}^{-1} \mu(\mathbf{x}) \Bigg) \\ &+ \lambda \mathcal{E}_{P_k} \Bigg(\left(-I + \nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}) \delta t + \mathbf{G} \nabla_{\mathbf{x}} \mathbf{u}(\mathbf{x}) \delta t \right) \boldsymbol{\Sigma}^{-1} \mathbf{G}\mathbf{u}_k(\mathbf{x}, t) \Bigg)\end{aligned}$$

The optimal controls are given by:

$$\mathbf{u}_{k+1}(\mathbf{x},t)dt = -\mathbf{R}^{-1}q_1 dt - \mathbf{R}^{-1}\mathbf{G}^T\nabla_{\mathbf{x}}V(\mathbf{x})\delta t = -\mathbf{R}^{-1}q_1 dt + \lambda\mathbf{R}^{-1}\mathbf{G}^T\mathbf{\Sigma}^{-1}\mathcal{E}_{P_k}\Bigg(\mathbf{G}\mathbf{u}_k(\mathbf{x},t) + \mu(\mathbf{x})\delta t\Bigg)$$

$$= -\mathbf{R}^{-1}q_1 dt + \mathbf{\Omega}\mathcal{E}_{P_k}\Bigg(\mathbf{G}\mathbf{u}_k(\mathbf{x},t) + \mathbf{B}d\mathbf{w}(t)\delta t\Bigg)$$

where the term $\mathbf{\Omega} = \mathbf{R}^{-1}\mathbf{G}^T\left(\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T\right)^{-1}$. Since the feedback policy $\mathbf{u}_k(\mathbf{x},t)$ is evaluated at the current state $\mathbf{x}$ we will have the final result of this theorem. ■

Table II illustrates the iterative path integral optimal control derived based on the dynamic programming principle. All the points made regarding the iterative path integral control in table I regarding the different ways of using this approach are valid also for the algorithm in II. However there are few differences since algorithm in table II provides the optimal control for cost functions with cross terms of the form $q_1(\mathbf{x},t)^T\mathbf{u}$. Note also that the addition of the aforementioned cross terms results in sampling with a diffusion that is different from the initial diffusion in (30). Furthermore, in case of II the assumption $\lambda\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T = \mathbf{B}\mathbf{B}^T$ should always be valid. The algorithm in I is simpler to implement and it has less parameters to consider.

VII. Stochastic optimal control for Markov Jump Diffusions Processes based on the basic inequalities

In this section we consider Markov jump diffusion processes to show the generizabilty of the information theoretic approach to stochastic optimal control. In particular we derive the lower bound on cost functions that typically appear in the cases of stochastic optimal control of markov jump diffusion processes. The analysis relies again on Girsanov's theorem the use of Radon-Nikodym derivative when poisson-jump and diffusion terms appear in the stochastic dynamics. To keep the analysis simple we consider Markov Jump Diffusions in 1D, the analysis for multidimensional case is similar. More precisely we have the controlled $d\mathbf{x} = \mathbf{f}(\mathbf{x})\delta t + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}\delta t + \frac{1}{\sqrt{|\rho|}}d\mathbf{w}^{(1)}(t)\right) + \mathbf{h}(\mathbf{x})d\mathbf{P}^{(1)}(t)$ as

TABLE II: Iterative Path Integral Control Based on the Dynamic Programming Principle.

- **Given**:
  - The time horizon t_N. The cost terms $q_0(\mathbf{x}_t), q_1(\mathbf{x}_t), \mathbf{R}, \lambda,\ \mathbf{G}, \mathbf{B}$ and the total state depended cost $\tilde{q}(\mathbf{x}, t) = q_0(\mathbf{x}, t) - \frac{1}{2} q_1(\mathbf{x}, t)^T \mathbf{R}^{-1} q_1(\mathbf{x}, t)$
  - The quantities $\mathbf{\Omega} = \mathbf{R}^{-1}\mathbf{G}^T\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma} = \mathbf{B}\mathbf{B}^T$, initials controls $\mathbf{u}_0$ and $\lambda \mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T = \mathbf{B}\mathbf{B}^T$
- **Repeat** until convergence of the trajectory cost R:
  - Create M roll-outs of the system by forward sampling of the diffusion $\mathbf{dx} = \Big(\mathbf{f}(\mathbf{x}) - \mathbf{G}(\mathbf{x})\mathbf{R}^{-1} q_1(\mathbf{x})\Big) dt + \mathbf{G}(\mathbf{x})\mathbf{u}_k dt + \mathbf{B}(\mathbf{x}) d\mathbf{w}(t)$.
  - **For** $k = 1 ... M$, compute:
    - $*$ $\zeta(\mathbf{u}, t_i, t_N)$ as in (42).
    - $*$ $S(\vec{\mathbf{x}}_{i,k}) = \phi_{t_N} + \sum_{j=i}^{N-1} \left(q(t_j)dt + \zeta(\mathbf{u}, t_i, t_N)\right)$
    - $*$ $P\left(\vec{\mathbf{x}}_{i,k}\right) = \frac{e^{-\frac{1}{\lambda} S(\vec{\mathbf{x}}_{i,k})}}{\sum_{k=1}^{K} [e^{-\frac{1}{\lambda} S(\vec{\mathbf{x}}_{i,k})}]}$
  - **For** $i = 1 ... (N-1)$, compute:
    - $*$ $\delta \mathbf{u}_{\tilde{\mathbb{P}}} = E_P \left(d\mathbf{w}^{(k)}(t_i) \right)$
    - $*$ $\mathbf{u}_{k+1}(t_i) dt = \Big(- \mathbf{R}^{-1} q_1(t_i) + \mathbf{\Omega}\mathbf{G}\mathbf{u}_k(t_i) \Big) dt + \mathbf{\Omega}\mathbf{B}\delta\mathbf{u}_{\tilde{\mathbb{P}}}$

well as its uncontrolled versions for $\mathbf{u} = 0$ 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 $i = 1, ..., m$. 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. Thus: $\text{Cov}\left[d\mathbf{P}(t_j) d\mathbf{P}(t_k)\right] = \text{Var}\left[dP(t_j\right] \delta_{k,j} = \nu(t_j) dt \delta_{k,j}$ Based on Girsanov's theorem [13] 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:

$$\zeta(\mathbf{u}) = \int_{t_i}^{t_N} \frac{1}{2}|\rho|\mathbf{u}(t)^2\delta t + \sqrt{|\rho|}\int_{t_i}^{t_N} \mathbf{u}(t)d\mathbf{w}^{(1)}(t) + \int_{t_i}^{t_N} \left(\left(1-\gamma^{(J)}(t)\right)\nu_0(t)\right)dt + \sum_{j=1}^{\mathbf{P}^{(1)}(t)} \log\gamma^{(J)}(t)$$

with $\gamma^{(J)}(t) = \frac{\nu^{(1)}(t)}{\nu^{(0)}(t)}$. The lower bound on the value function is now derived by incorporating the Radon-Nikodým derivative into (18).

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

where $\mathcal{V}(\gamma^{(J)}(t)) = \rho\int_{t_i}^{t_N}\left(\left(\gamma^{(J)}(t)-1\right)\nu_0(t)\right)\delta t + \mathcal{E}^{(1)}_{\boldsymbol{\mathcal{T}}_i}\Bigg(\sum_{j=1}^{\mathbf{P}^{(1)}(t)}\log\gamma^{(J)}(t)^{\rho}\Bigg)$ Thus we will have:

$$\boxed{\xi_J(\mathbf{x}) \leq \mathcal{E}^{(1)}_{\boldsymbol{\mathcal{T}}_i}\Big[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N}\mathbf{u}(t)^2\delta t\Big]}$$

The new bound under sampling based on markov jump diffusion processes is defined as: $\xi_J(\mathbf{x}) = \xi(\mathbf{x}) - \mathcal{V}(\gamma^{(J)}(t))$. 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: $\xi(\mathbf{x}) = \xi_J(\mathbf{x}) \leq \mathcal{E}^{(1)}_{\boldsymbol{\mathcal{T}}_i}\Big[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N}\mathbf{u}(t)^2\delta t\Big]$. 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.

VIII. Application of Iterative Path Integral Control to Tendon driven robotic finger.

Tendon driven robotic systems are difficult to control because of the nonlinearities of their dynamics. The aforementioned nonlinearities are due to 1) the antagonistic relationship between the tendons. A rather *naive* control policy of just pulling all tendons simultaneously does not result in a desired movement or force production. Thus, tendons have to work together and synchronize their tensions such that the desired movement is achieved. 2) Tendons can only pull and not push. This idiosyncrasy of the reduced actuation per tendon is one of the reasons why tendon driven system require a large number of tendons to generate movement and force control 3) Tasks that involve contact with objects and surfaces impose further nonlinear phenomena. Besides the nonlinearities, tendon driven systems are hard to model. System identification is usually the step towards building dynamical models. However for the case of tendon driven systems system id is difficult because of the large dimensionality of the state as well as the requirement for expensive sensors for force, joint position and velocity measurements.

The experiments presented here use the Anatomically Correct Testbed (ACT) index finger [5]. The ACT index finger has the full 4 degree-of-freedom joint mobility and is controlled by six motor-driven tendons acting through a crocheted tendon-hood. Two tendons, the Flexor Digitorum Profundus (FDP) and Flexor Digitorum Superficialis (FDS) act as flexors; the EI(Extensor Indicis) , Radial Interosseous(RI) , and Proximal Interesseous (PI) act as extensors and ab/aductors; the Lumbrical (LUM) is an abductor but switches from extensor to flexor depending on finger posture. There are also 3 joints starting from the Metacarpophalangeal joint (MCP), the Proximal Interphalangeal (PIP) and the Distal Interhalangeal (DIP). By sharing the redundancies and nonlinearities of human hands [6], the system constitutes a challenging testbed for model identification, control, and task learning, while also providing a unique perspective for the study of biomechanics and human motor control. The 6 tendons are torque-controlled by 6 DC motors at 200 Hz and measure tendon displacements at a resolution 2.30 μm; the tendon displacements alone are used for feedback control as there is no direct measurement of joint kinematics. Successfully performing manipulation tasks requires a control policy that can handle the nonlinear dynamics and high dimensionality of the robot as well as the dynamics of the task itself.

A. Sliding Switch Task

We examine the task of contacting a sliding switch and pushing it down (see fig. 1(a)). The switch in our apparatus is coupled to a belt and motor which allow the imposition of synthetic dynamics. In particular the switch is made springy such that it can return back to resting position if contact is lost. The position of the switch x is measured with a linear potentiometer. Importantly, the finger may loose contact with the switch at x_{reach} before reaching the bottom of the possible range, denoted x_{min}.

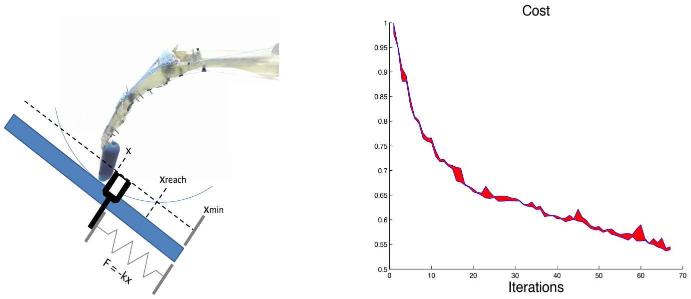
We begin with a single demonstration of the desired task in which a human holds the finger and moves it through a motion of pushing the switch down. The tendon excursions produced by this externally-powered example grossly resemble those required for the robot to complete the task, but simply replaying them using a general-purpose PID controller would not result in successful task completion for two main reasons (see figures 3(a) and 3(b)). Firstly, during demonstration the tendons are not loaded, which changes the configuration of the tendon network in comparison to when it is actively moving. Secondly, and more importantly, the tendon trajectories encountered during a demonstration do not impart any information about the necessary forces required to accommodate the dynamics of the task. For instance, at the beginning of the task, the finger must transition from moving through air freely, to contacting and pushing the switch. A feedback controller following a reference trajectory has no way of anticipating this contact transition, and therefore will fail to initially strike the switch with enough force to produce the desired motion.

B. Application of iterative path integral control.

We apply Iterative path integral control as table I in its open loop formulation. The torque sent to every motor is specified as a P controller:$\tau_i = K_{p_i}(t)\left(l^{(i)}_{actual}(t) - l^{(i)}_{desired}(t)\right), \forall i = 1, ..., 6$ where $l^{(i)}_{actual}(t)$ is the actual tendon excursion and $l^{(i)}_{desired}(t)$ is the desired one. The gain $K_{p_i}(t)$ is time varying. To perform simultaneous control of trajectories and gains, we augment the dynamics with stochastic differential equations. The state of these stochastic differential equations correspond to the 6 tendon excursions and 6 control gains. Thus we have: $dl_{desired}^{(i)}(t) = al_{desired}^{(i)}(t)dt + u_{li}(t)dt + \sigma dw(t), \quad \forall i = 1, ..., 6$, and $dK_{pi}(t) = aK_{pi}(t)dt + u_{Ki}(t)dt + \sigma dw(t), \quad \forall i = 1, ..., 6$ with $a < 0$.

The *new* control variables are u_{Ki} and u_{li} which correspond to the change in control gain and the change in tendon excursion per time unit. Therefore the dimensionality of the control space is $\mathbf{u} \in \Re^{12\times1}$. Note also that the state space formulation of the augmented stochastic dynamics has the form which allows iterative path integral control to be applied. More precisely the control and diffusion matrices are partitioned as follows $[0, \quad \boldsymbol{\mathcal{B}}_c]$ and $[0, \quad \sigma\boldsymbol{\mathcal{B}}_c]$ with $\boldsymbol{\mathcal{B}}_c = I_{12\times12}$. Additionally, the fact that that path integral control does not rely on the drift of the stochastic dynamics is very important for this application as no dynamical model of the tendon driven finger has been available so far. For the experiment the cost function has the form of $L = q(x_{t_N}) + \sum \left(q(x_{t_i}) + \mathbf{u}_{t_i}^T \mathbf{u}_{t_i} \right) dt$. In this cost the x_t is the position in the switch at time t, while $q(x_t)$ is the state dependent cost. In the experiment presented here we use $q(x_t) = 2\times10^4 x_t$ and $q(x_{t_N}) = 300\times q(x_t)$. The position x_t is measured by the potentiometer and it is positive.

Figure 1(b) illustrates the normalized cost during learning for the task of sliding and holding the slider switch, together with the one sigma standard deviations. The number of learning iterations is 201. However, in order to speedup learning we store the cost at every 3rd iteration which correspond to 67 cost-checking iterations from the total of 201 as shown in 1(b). Note that the cost drops to 50% of its initial value. At this level of performance, the finger has learned to push the slider until the point which it does not looses contact. Figure 2 illustrates the torque(=force) in mA applied to the tendons after learning. We have split the tendons activations into 3 subplots to emphasize the role they play during the task. The task starts with an initial burst of activity in the main extensor EC, the abductor-adductor tendons PI and RI as well as the Lumbrical. We speculate that this initial burst of activity is for the purposes of



(a) Slider with spring dynamics.

(b) Normalized cost during learning session.

Fig. 1

stabilizing the finger by rejecting abduction-adduction movements. The movement is initiated with a burst of activity in the PI and RI which generate the rotation around the MCP joint. This rotation, in turn, generates the downward movement. During the phases of *contact* and *moving in contact*, the tendons FDS, FDP and LUM are activated to generate the torque such that the finger overcomes the force applied by the slider. At the end of the *moving in contact*-phase (see the blue arrow in figure 2) Lumbrical acts in a opposite fashion than FDP and FDS. In particular, its activation increases when FDS and FDP activation decreases and vice versa until there is no movement. This observation is in contrast to activation profiles during the *contact* phase. During this phase Lumbrical and FDS are simultaneously activated and therefore it Lumbrical plays the role of flexor. This experiments demonstrate that Lumbrical can act either as a flexor or as an extensor depending on the task under consideration as well as phase of the task. The videos from the initial and learned behaviors can be found on the website: http://www.cs.washington.edu/homes/etheodor/videos.html.

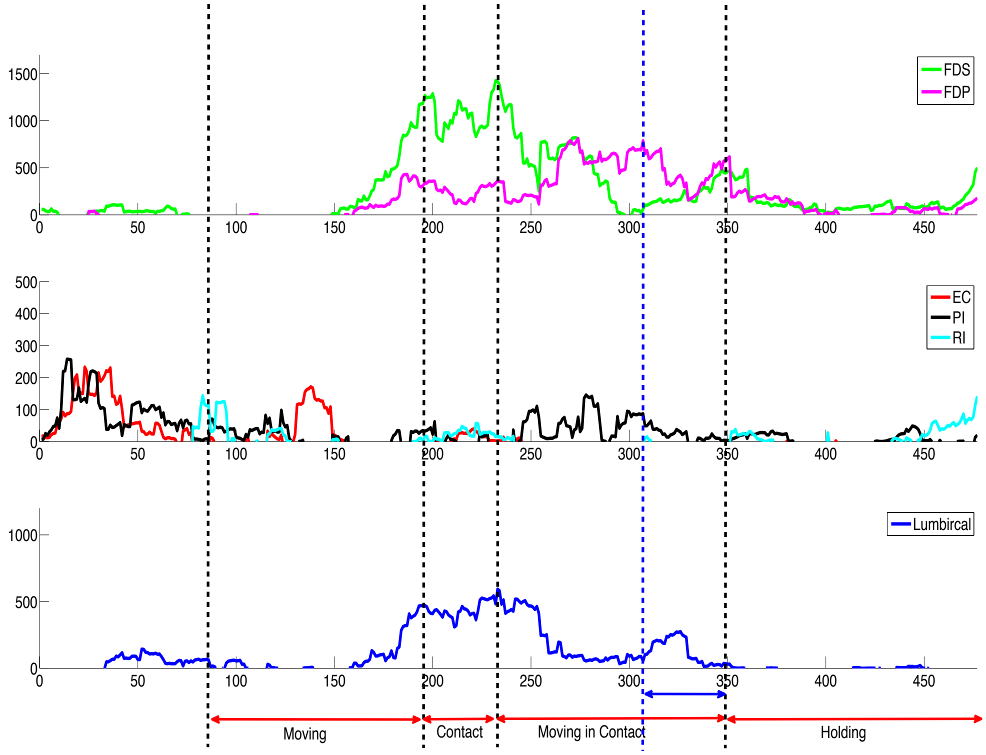
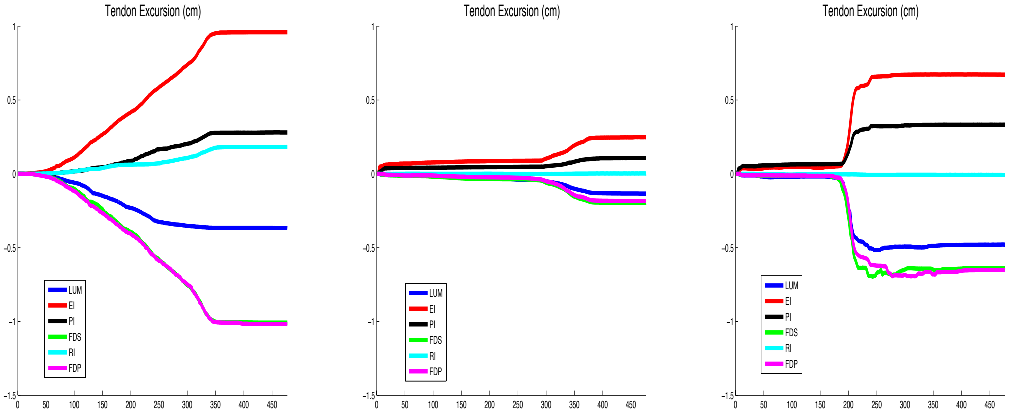


Fig. 2: Forces applied to tendons after learning.

IX. Discussion

Our work in this paper demonstrates the connection of path integral control framework as presented in the machine learning and robotic communities [3], [14], [15], [22], [23], [26] with work in the control theoretic community on risk sensitivity [4], [7]–[9]. 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 $E^{(1)}(J(\mathbf{x},\mathbf{u}))$ 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



(a) Demostrated excursions.

(b) Replayed excursions with a conventional PID controller.

(c) Learned excursions.

Fig. 3

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 approach is developed in the opposite direction. The approach starts with the risk sensitive version of the state dependent part $\mathcal{J}(\mathbf{x})$of the cost function $E^{(1)}(J(\mathbf{x},\mathbf{u}))$. With the application of Girsanov's theorem between controlled and uncontrolled dynamics and the use of Jensen inequality the upper bound $\xi(\mathbf{x})$ of the objective function $\mathcal{E}^{(1)}(J(\mathbf{x},\mathbf{u}))$ is derived. As a last step, the link to Bellman optimality is established by showing that $\xi(\mathbf{x})$ satisfies the HJB equation and therefore it is a value function.

Risk sensitivity and the relationship between free energy and relative entropy offers an alternative formalism of optimality which, for the case of diffusions processes, and under the conditions specified in this work, turns out to be identical to the bellman principle of optimality. In this work we have derived the iterative version of path integral control with the use of successive application of Girsanov's theorem as applied to markov diffusion processes. Previous work in the area of policy improvement with path integrals [23], derived iterative versions of path integral control but for a restricting class of policies parameterized as Dynamic Movement Primitives. Here, the derivation and formulation of iterative path integral control is general and therefore it is valid for generalized feedback policies with no pre-specified parameterization.

Inside the class of the stochastic dynamics described by markov diffusion processes, the path integral control approach derived based on Dynamic Programming may be more general since based on the conditions (39) it can incorporate general cost functions and stochastic dynamics. In particular, it incorporates cost functions which besides the state-only and control-only depended terms they can include terms as functions of both state and control. Another level of generalization is that the control transition and diffusion matrices may be different. This generalization however, is reduced by the assumption regarding control cost and the variance of the noise $\lambda \mathbf{G}(\mathbf{x})\mathbf{R}^{-1}\mathbf{G}(\mathbf{x})^T = \mathbf{B}(\mathbf{x})\mathbf{B}(\mathbf{x})^T$.

The aforementioned assumption holds by construction in the second approach as the control transition and diffusion matrices are almost identical. In addition, in the second approach the lower bound of the accumulated trajectory cost 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 [8], [9] as well as in this work, for the case of diffusion processes turns out to be equivalent to the Bellman principle of optimality. Here we have done on step forward by deriving the lower bound of the cost of stochastic optimal control problem for a special class of nonlinear markov jump diffusion processes. As it turns out the form of the lower bound remains similar with the case of diffusion processes for as long the change in the probability measure of the markov jump diffusion is only due to the changes in the in the drift of the stochastic dynamics.

On the application side we have applied iterative optimal control in its open loop formulation for controlling the ACT tendon driven finger. Our experimental results demonstrate the efficiency of the method in terms of dealing with complex nonlinear systems with unknown dynamics in a tasks that involves contact. Our analysis of the biomechanical properties of the ACT index finger is not conclusive as more experiments will be required in order to 1) further investigate the bio-mechanical properties of the ACT 2) suggests new designs with better actuation and sensing capabilities that will improve control and speedup learning of new tasks 3) scale the analysis to full ACT hand which involves the control of 24 tendons.

X. Appendix

Girsanov's theorem

We will consider Girsanov's theorem for the case of stochastic 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 $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$ and $\mathbf{B}(\mathbf{x})\boldsymbol{\Sigma}\mathbf{B}(\mathbf{x})^T = \boldsymbol{\Sigma}_{\mathbf{w}}$. The corresponding probability measures:

$$d\mathbb{P} = \frac{\exp\left(-\frac{1}{2}\left(\int_{t_i}^{t_N}||\mu_k(\mathbf{x})||^2_{\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}}\delta t\right)\right)}{(2\pi\delta t)^{m/2}|\boldsymbol{\Sigma}_{\mathbf{w}}|^{1/2}}d\mathbf{x} \quad \textit{and} \quad d\mathbb{Q} = \frac{\exp\left(-\frac{1}{2}\left(\int_{t_i}^{t_N}||\lambda_k(\mathbf{x})||^2_{\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}}\delta t\right)\right)}{(2\pi\delta t)^{m/2}|\boldsymbol{\Sigma}_{\mathbf{w}}|^{1/2}}d\mathbf{x}$$

with $\mu_k(\mathbf{x}) = \left(\frac{\delta\mathbf{x}}{\delta t} - \mathbf{f}(\mathbf{x}, t)\right)$ thus $\mu_k(\mathbf{x})\delta t = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t)$ 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 $\lambda_k(\mathbf{x})\delta t = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t)$. From all these equations we can also get a relationship between the noise terms $d\mathbf{w}^{(0)}(t)$ and $d\mathbf{w}^{(1)}(t)$. More precisely $\lambda_k(\mathbf{x})\delta t = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) = \mu_k(\mathbf{x})\delta t - \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)\delta t$ and thus $\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t) - \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)\delta t$. Now we would like to find the expression:

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

$$= \exp\left[-\frac{1}{2}\int_{t_i}^{t_N}\left(-\mathbf{u}_k(t)^T\mathbf{G}(\mathbf{x})^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\mathbf{G}(\mathbf{x})\mathbf{u}_k(t)\delta - \int_{t_i}^{t_N}\mathbf{u}_k(t)^T\mathbf{G}(\mathbf{x})^T\boldsymbol{\Sigma}_{\mathbf{w}}^{-1}\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t)\right)\right]$$

Since $\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) = \boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(0)}(t) - \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)\delta t$ 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 $\boldsymbol{\mathcal{B}}(\mathbf{x})\mathbf{L}d\mathbf{w}^{(1)}(t) + \mathbf{G}(\mathbf{x})\mathbf{u}_k(t)\delta t$. Thus the ratio of the probability measures is:

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

REFERENCES

[1] 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*, 32(1):95–122, 2008.

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

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

[4] Paolo Dai Pra, Lorenzo Meneghini, and Wolfgang Runggaldier. Connections between stochastic control and dynamic games. *Mathematics of Control, Signals, and Systems (MCSS)*, 9(4):303–326, 1996-12-08.

[5] A. D. Deshpande, Z. Xu, M. J. V. Weghe, L. Y. Chang, B. H. Brown, D. D. Wilkinson, S. M. Bidic, and Y. Matsuoka. Mechanisms of anatomically correct testbed (ACT) hand. *Trans. Mechatronics*, 2011.

[6] A.D. Deshpande, J. Ko, D. Fox, and Y. Matsuoka. Anatomically correct testbed hand control: muscle and joint control strategies. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 4416–4422. IEEE, 2009.

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

[8] W. H. Fleming and H. Mete Soner. *Controlled Markov processes and viscosity solutions*. Applications of mathematics. Springer, New York, 1nd edition, 1993.

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

[10] W.H. Fleming. Exit probabilities and optimal stochastic control. *Applied Math. Optim*, 9:329–346, 1971.

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

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

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

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

[15] 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.

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

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

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

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

[21] 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.

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

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

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

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

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

[27] F.J. Valero-Cuevas, F.E. Zajac, C.G. Burgar, et al. Large index-fingertip forces are produced by subject-independent patterns of muscle excitation. *Journal of Biomechanics*, 31(8):693–704, 1998.

[28] M Venkadesan and F.J. Valero-Cuevas. *The journal of Neuroscience*, 28(6):1366–1373, 2008.