From information theoretic dualities to Path Integral and Kullback Leibler control: Continuous and Discrete Time formulations

Evangelos A. Theodorou[1], Krishnamurthy Dvijotham[2] and Emo Todorov[3]

***Abstract*— This paper presents a unified view of stochastic optimal control theory as developed within the machine learning and control theory communities. In particular we show the mathematical connection between 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 discuss the applications of the relationship between free energy and relative entropy to nonlinear stochastic dynamical systems affine in noise and nonlinear stochastic dynamics affine in control and noise. For this last class of systems, we provide the PI optimal control and its iterative formulation. In addition, we present the connection of PI control derived based on Dynamic Programming with the information theoretic dualities. Finally, we provide links to KL stochastic optimal control and discuss generalizations and future work.**

I. Introduction

Optimal control for nonlinear Markov diffusions processes based on path integrals demonstrated remarkable applicability to robotic control and planning problems. For continuous state actions spaces and continuous time, work in [14], [15] 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 [35], PI control framework was extended to stochastic optimal control problems for multi-agents systems. In [26], [27] the PI control was derived 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). 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], [21], [24].

Parallel to the work in continuous time, in [31], [34] 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 work we derive a unified view of PI and KL control as presented in machine learning [14], [15], [31], [34] and control theory [4], [6]–[9] communities. This unified view relies on the relationship between free energy and relative entropy which is derived in Section II. In Section III we apply this relationship to nonlinear stochastic dynamics affine in noise. In Section IV, we apply this relationship to nonlinear stochastic dynamics affine in noise and control. Furthermore we derive the iterative PI control without policy parameterization. In section V we show how the PI control framework is derived based on the Bellman principle of optimality and contrast this approach with the one in Section IV. In Section VI we provide links to KL-control. Finally in Section VII we conclude and discuss the connections between different 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 [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:

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

We will also consider the objective function:

$$\xi(\mathbf{x}) = \frac{1}{\rho}\mathbb{E}\Big(\mathcal{J}(\mathbf{x})\Big) = \frac{1}{\rho}\log E_{\mathbb{P}}\Big[\exp\left(\rho \mathcal{J}(\mathbf{x})\right)\Big] \tag{2}$$

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 $\xi(\mathbf{x}) = E_{\mathbb{P}}\left(\mathcal{J}\right) + \frac{\rho}{2} Var\left(\mathcal{J}\right)$ 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

[1] Evangelos A. Theodorou is a Postdoctoral Research Associate with the Department of Computer Science and Engineering,University of Washington, Seattle. etheodor@cs.washington.edu

[2] Krishnamurthy Dvijotham is graduate student in Department of Computer Science and Engineering, University of Washington, Seattle. dvij@cs.washington.edu

[3] Emo Todorov is Associate professor with the Departments of Computer Science and Engineering, and Applied Math, University of Washington, Seattle. todorov@cs.washington.edu

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 $E_{\mathbb{P}}$ taken under the measure $\mathbb{P}$ as a function of the expectation $E_{\mathbb{Q}}$ taken under the probability measure $\mathbb{Q}$. More precisely will have:

$$\begin{aligned} E_{\mathbb{P}}\Big[\exp\left(\rho\mathcal{J}(\mathbf{x})\right)\Big] &= \int \exp\left(\rho\mathcal{J}(\mathbf{x})\right) d\mathbb{P} \\ &= \int \exp\left(\rho\mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} \end{aligned}$$

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

$$\begin{aligned} \log E_{\mathbb{P}}\Big[\exp\left(\rho\mathcal{J}(\mathbf{x})\right)\Big] &= \log \int \exp\left(\rho\mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} \\ &\geq \int \log\left(\exp\left(\rho\mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \\ &= \int \left(\rho\mathcal{J}(\mathbf{x}) + \log\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q} \\ &= \int \rho\mathcal{J}(\mathbf{x}) d\mathbb{Q} - \mathcal{I}\left(\mathbb{Q}||\mathbb{P}\right) \end{aligned}$$

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

$$\boxed{\xi(\mathbf{x}) = -\frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right) \leq E_{\mathbb{Q}}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{|\rho|}\mathcal{I}\left(\mathbb{Q}||\mathbb{P}\right)} \quad (3)$$

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

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

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

$$d\mathbb{Q}^* = \frac{\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right) d\mathbb{P}}{\int \exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right) d\mathbb{P}} \quad (5)$$

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

A rather intuitive way of writing (4) is to express it in the form that follows:

$$\underbrace{-|\rho|^{-1}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right)}_{\textbf{Helmholtz Free Energy}} \leq \underbrace{\textbf{State Cost} + |\rho|^{-1}\textbf{Information Cost}}_{\textbf{Non-Equilibrium Free Energy}} \quad (6)$$

The terms **State Cost** and **Information Cost** are defined as:

$$\textbf{State Cost} = E_{\mathbb{Q}}\left(\mathcal{J}(\mathbf{x})\right), \quad \textbf{Information Cost} = \mathcal{I}\left(\mathbb{Q}||\mathbb{P}\right)$$

We can think about steering a dynamical system from an initial to a terminal state by minimizing a cost function at the expense of the information cost. The summation of the state and information cost corresponds to the free energy far from the thermodynamic equilibrium. At the thermodynamic equilibrium in which the minimum is attained for $\mathbb{Q} = \mathbb{Q}^*$, equation (4) takes the form:

$$\underbrace{-|\rho|^{-1}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right)}_{\textbf{Helmholtz Free Energy}} = \underbrace{E_{\mathbb{Q}^*}\left(\mathcal{J}(\mathbf{x})\right)}_{\textbf{Energy}} - |\rho|^{-1} \cdot \underbrace{\mathcal{H}\left(\mathbb{Q}^*\right)}_{\textbf{Generalized Entropy}} \quad (7)$$

The entropy functional $\mathcal{H}\left(\mathbb{Q}^*\right)$ above is knonw as Baroh-Jaunch entropy or generalized Boltzmann-Gibbs-Shannon entropy [36] defined as $\mathcal{H}\left(\mathbb{Q}\right) = -\mathcal{I}\left(\mathbb{Q}||\mathbb{P}\right)$. Equation (7) takes has form $F = U - TS$ where F is the free energy, $T = |\rho|^{-1}$ is the temperature and S is entropy.

Next we apply (7) on nonlinear stochastic dynamics affine in noise, nonlinear stochastic dynamics affine in control and noise, and nonlinear dynamics with diffusion and poisson noise.

III. Nonlinear Stochastic Dynamics Affine in Noise

We consider the stochastic dynamics affine in noise:

$$d\mathbf{x} = \mathbf{F}(\mathbf{x}, \mathbf{u})dt + \mathbf{C}(\mathbf{x})d\mathbf{w}(t) \quad (8)$$

with $\mathbf{x} \in \Re^{n\times 1}$ denoting the state of the system and $\mathbf{u}_t \in \Re^{p\times 1}$ the control vector , $\mathbf{C}(\mathbf{x}) \in \Re^{n\times p}$ is the diffusion matrix, $\mathbf{F}(\mathbf{x}, \mathbf{u}) \in \Re^{n\times 1}$ the drift dynamics, $\mathbf{\Sigma_C} = \mathbf{C}(\mathbf{x})\mathbf{C}(\mathbf{x})^T \in \Re^{p\times p}$ and $d\mathbf{w} \in \Re^{p\times 1}$ brownian noise. We also define the stochastic differential equation:

$$d\mathbf{x} = \mathbf{A}(\mathbf{x})dt + \mathbf{C}(\mathbf{x})d\mathbf{w}(t) \quad (9)$$

where the drift term $\mathbf{A}(\mathbf{x}) \in \Re^{n\times 1}$ is defined as $\mathbf{A}(\mathbf{x}) = \mathbf{F}(\mathbf{x}, 0)$ and therefore it corresponds to the uncontrolled dynamics in (8). Expectations evaluated on trajectories generated by the controlled dynamics and uncontrolled dynamics are represented as $E_{\mathbb{P}}$ and $E_{\mathbb{Q}}$ respectively. We also define the following quantity:

$$\delta\mathbf{F}(\mathbf{x}, \mathbf{u}) = \mathbf{F}(\mathbf{x}, \mathbf{u}) - \mathbf{A}(\mathbf{x}) = \mathbf{F}(\mathbf{x}, \mathbf{u}) - \mathbf{F}(\mathbf{x}, 0), \quad \forall \mathbf{x}, \mathbf{u}$$

We continue our analysis with the result in (3) and the definition of the Radon- Nicodyn derivative for the stochastic differential equations (8) and (9). More precisely we will have:

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left[\int_{t_0}^{t_N}\left(\delta\mathbf{F}^T\mathbf{C}(\mathbf{x})^{-1}d\mathbf{w}(t) + \frac{1}{2}\delta\mathbf{F}^T\mathbf{\Sigma}_{\mathbf{C}}^{-1}\delta\mathbf{F}\delta t\right)\right]$$

By substituting the equation above back into (3) we have that:

$$-\frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right) \leq E_{\mathbb{Q}}\left(\mathcal{J}(\mathbf{x}) + \frac{1}{2|\rho|}\int_{t_0}^{t_N} \delta\mathbf{F}^T\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}\delta\mathbf{F}\delta t\right)$$

The equation above can be written in the form (6) with the state cost term defined as:

$$\textbf{State Cost} = E_{\mathbb{Q}}\left(\mathcal{J}(\mathbf{x})\right) \tag{10}$$

And the information cost is expressed as:

$$\textbf{Information Cost} = E_{\mathbb{Q}}\left(\frac{1}{2}\int_{t_0}^{t_N} \delta\mathbf{F}^T\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}\delta\mathbf{F}\delta t\right) \tag{11}$$

At the thermodynamic equilibrium in which the minimum of non-equilibrium free energy is attained for $\mathbb{Q}^*$ given by (5), we have that:

$$-\frac{1}{|\rho|}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right) = E_{\mathbb{Q}^*}\left(\mathcal{J}(\mathbf{x}) + \frac{1}{2|\rho|}\int_{t_0}^{t_N} \delta\mathbf{F}_*^T\boldsymbol{\Sigma}_{\mathbf{C}}^{-1}\delta\mathbf{F}_*\delta t\right)$$

where the term $\delta\mathbf{F}_*$ is expressed as $\delta\mathbf{F}_*(\mathbf{x},\mathbf{u}) = \mathbf{F}(\mathbf{x},\mathbf{u}^*) - \mathbf{A}(\mathbf{x})$ and $\mathbf{u}^*$ and it corresponds to the different between optimally controlled under $\mathbf{u} = \mathbf{u}^*$ and uncontrolled dynamics $\mathbf{u} = 0$.

IV. Nonlinear Stochastic Dynamics Affine in Control and Noise

For our analysis in this section [4], [7] we consider the uncontrolled and controlled stochastic dynamics of the form:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}(\mathbf{x})d\mathbf{w}^{(0)}(t) \tag{12}$$

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

with $\mathbf{x}_t \in \Re^{n\times 1}$ denoting the state of the system, $\mathbf{B}(\mathbf{x}) \in \Re^{n\times p}$ is the control and diffusions matrix, $\mathbf{f}(\mathbf{x},t) \in \Re^{n\times 1}$ the passive dynamics, $\mathbf{u}_t \in \Re^{p\times 1}$ the control vector and $d\mathbf{w} \in \Re^{p\times 1}$ brownian noise. Notice that the difference between the two diffusions above is on the controls that appear in 13. These controls together with the passive dynamics define a new drift term. Expectations evaluated on trajectories generated by the controlled dynamics and uncontrolled dynamics are represented as $E_{\mathbb{P}}$ and $E_{\mathbb{Q}}$ 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- - Nicodyn derivative.

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\zeta(\mathbf{u})\right) \quad \text{and} \quad \frac{d\mathbb{P}}{d\mathbb{Q}} = \exp\left(-\zeta(\mathbf{u})\right) \tag{14}$$

where according to Girsanov's theorem [17] adapted to the Markov diffusion processes (12) and (13) the term $\zeta(\mathbf{u})$ is defined as follows:

$$\zeta(\mathbf{u}) = \frac{1}{2}|\rho|\int_{t_i}^{t_N}\mathbf{u}^T\mathbf{u}dt + \sqrt{|\rho|}\int_{t_i}^{t_N}\mathbf{u}^T d\mathbf{w}^{(1)}(t) \tag{15}$$

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 (14) and (25) into inequality (3) gives the following result:

$$\begin{aligned}\xi(\mathbf{x}) &= -\frac{1}{|\rho|}\log E_{\mathbb{P}}\left[\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)\right] \\ &\leq E_{\mathbb{Q}}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{|\rho|}\zeta(\mathbf{u})\right]\end{aligned} \tag{16}$$

The expectation on the right side of the inequality in (16) is further simplified as follows:

$$\begin{aligned}\xi(\mathbf{x}) &= -\frac{1}{|\rho|}\log E_{\mathbb{P}}\left[\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)\right] \\ &\leq E_{\mathbb{Q}}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N}\mathbf{u}^T\mathbf{u}dt\right]\end{aligned} \tag{17}$$

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 (17) 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 (17) 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 (17). Essentially inequality (17) 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 (17) reaches its optimal $\xi(\mathbf{x})$. Under the optimal control $\mathbf{u}^*$ and according to (18) the optimal distribution takes the from:

$$d\mathbb{Q}^*(\mathbf{x}) = \frac{\exp\left(-|\rho|\int q(\mathbf{x})dt\right)d\mathbb{P}(\mathbf{x})}{\int\exp\left(-|\rho|\int q(\mathbf{x})dt\right)d\mathbb{P}(\mathbf{x})} \tag{18}$$

An important question to ask is what is the link between (17) 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 $\Phi(\mathbf{x},t) = E_{\mathbb{P}}(\exp\left(\rho\mathcal{J}(\mathbf{x})\right))$. The Feynman-Kac lemma

[10] tells us that this function satisfies the backwad Chapman Kolmogorov PDE. Therefore the following equation is true.

$$-\partial_t \Phi = \rho q_0 \Phi + \mathbf{f}^T(\nabla_{\mathbf{x}}\Phi) + \frac{1}{2|\rho|} tr\left((\nabla_{\mathbf{xx}}\Phi)\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^T\right) \quad (19)$$

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:

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

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

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

The nonlinear PDEs above corresponds to the HJB equation [23] for the case of the minimizing and maximizing optimal control problem and therefore, $\xi(\mathbf{x})$ is the corresponding minimizing or maximizing value function. Note that in oder to derive the PDEs above we did not use any principle of optimality.

A. Iterative Path Integral Control

Here we briefly discuss the derivation of the iterative PI control [30] based on successive application of Girsanov's theorem. The analysis starts with the following lemma.

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:*

$$S(\mathbf{x},\mathbf{u}_k) = \mathcal{J}(\mathbf{x}) + \frac{1}{2}\left(\eta(\mathbf{u}) + \int_{t_i}^{t_N}||\mu(\mathbf{x})||^2_{\boldsymbol{\Sigma}^{-1}}\delta t\right) \quad (22)$$

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:

$$\begin{aligned}\xi(\mathbf{x}) &= -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)d\mathbb{P}\\ &= -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)\frac{d\mathbb{P}}{d\mathbb{Q}}d\mathbb{Q}\\ &= -\frac{1}{|\rho|}\log\int\exp\left(-|\rho|\mathcal{J}(\mathbf{x}) - \zeta(\mathbf{u})\right)d\mathbb{Q}\end{aligned} \quad (23)$$

The measure $d\mathbb{Q}$ is expressed as:

$$d\mathbb{Q}\Big(\mathbf{x}_N, t_N|\mathbf{x}_i,t_i\Big) = \frac{\exp\left(-\frac{|\rho|}{2}\left(\int_{t_i}^{t_N}\mu(\mathbf{x})^T\boldsymbol{\Sigma}^{-1}\mu(\mathbf{x})dt\right)\right)}{(2\pi dt)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}\mathbf{dx} \quad (24)$$

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 [11], [20] will become equal to:

$$\begin{aligned}\zeta(\mathbf{u}) &= \frac{1}{2}|\rho|\int_{t_i}^{t_N}\mathbf{u}^T\mathbf{u}dt + \sqrt{|\rho|}\int_{t_i}^{t_N}\mathbf{u}^T d\mathbf{w}^{(1)}(t)\\ &= \frac{1}{2}|\rho|\int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{u}_k dt + |\rho|\int_{t_i}^{t_N}\mathbf{u}_k^T\boldsymbol{\mathcal{B}}^{-T}\mu(\mathbf{x})dt\\ &= \frac{1}{2}|\rho|\ \eta(\mathbf{u})\end{aligned} \quad (25)$$

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

$$\begin{aligned}\eta(\mathbf{u}) &= \int_{t_i}^{t_N}\mathbf{u}_k^T\mathbf{u}_k dt + \int_{t_i}^{t_N}2\mathbf{u}_k^T\boldsymbol{\mathcal{B}}^{-T}\mu(\mathbf{x})dt\\ &= \int_{t_i}^{t_N}\mathbf{u}^T\mathbf{u}dt + \frac{1}{\sqrt{|\rho|}}\int_{t_i}^{t_N}2\mathbf{u}^T d\mathbf{w}^{(1)}(t)\end{aligned} \quad (26)$$

Substitution of the function above $\zeta(\mathbf{u})$ and the path integral into (23) 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 $S(\mathbf{x},\mathbf{u}_k) = \mathcal{J}(\mathbf{x}) + \frac{1}{2}\left(\eta(\mathbf{u}) + \int_{t_i}^{t_N}||\mu(\mathbf{x})||^2_{\boldsymbol{\Sigma}^{-1}}dt\right)$.

■

The next step is to compute the gradient of $\xi(\mathbf{x})$ with respect to the state $\mathbf{x}$. The final result is given by 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:

$$\boxed{\mathbf{u}_{k+1}(\mathbf{x},t)dt = \mathbf{u}_k(\mathbf{x},t)dt + \frac{1}{\sqrt{\rho}}\mathcal{E}_{p_k}\Big(d\mathbf{w}_k(t)\Big)} \quad (27)$$

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 (22).

Proof: The gradient of $\xi(\mathbf{x})$ with respect to $\mathbf{x}_{t_i}$ is formulated as:

$$\nabla_{\mathbf{x}_{t_i}}\xi(\mathbf{x}_{t_i}) = -\frac{1}{|\rho|}\frac{\nabla_{\mathbf{x}_{t_i}}\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t))}\mathbf{dx}}{\int e^{-|\rho|S(\mathbf{x},\mathbf{u}_k(\mathbf{x},t)}\mathbf{dx}} \tag{28}$$

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 have that:

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

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 the optimal controls takes the form:

$$\begin{aligned}\mathbf{u}_{k+1}(\mathbf{x},t)dt &= -\mathbf{R}^{-1}\boldsymbol{\mathcal{B}}^T\nabla_{\mathbf{x}_{t_i}}\xi(\mathbf{x})dt\\ &= \mathcal{E}_{P_k}\Big(\mathbf{u}_k(\mathbf{x},t)dt + \frac{1}{\sqrt{\rho}}d\mathbf{w}_k(t)\Big)\end{aligned}$$

The policy $\mathbf{u}_k(\mathbf{x},t)$ is evaluated with trajectories starting from $\mathbf{x}_{t_i}$ and so we have (27). ■

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 [26] it has been shown that the path integral formulation is expressed as in (24) with $\boldsymbol{\mathcal{B}}_c d\mathbf{w}_k = \sqrt{\rho}\mu(\mathbf{x})dt$, $\mu(\mathbf{x}) = \left(\frac{\delta \mathbf{x}_c}{\delta t} - \mathbf{f}_c(\mathbf{x}) - \boldsymbol{\mathcal{B}}_c\mathbf{u}_k\right)$ and $\mathbf{\Sigma} = \boldsymbol{\mathcal{B}}_c\boldsymbol{\mathcal{B}}_c^T$. Our analysis in theorem 1 holds for the aforementioned types of systems as well.

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:

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

The expectation $E_{\mathbb{Q}}$ above, is evaluated on trajectories generated with forward sampling of the controlled diffusion:

$$\mathbf{dx} = \mathbf{F}(\mathbf{x},\mathbf{u})dt + \mathbf{B}(\mathbf{x})d\mathbf{w} \tag{29}$$

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:

$$\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},t) = q_0(\mathbf{x},t) + q_1(\mathbf{x},t)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbf{R}\mathbf{u} \tag{30}$$

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 [8], [23] associated with this stochastic optimal control problem is expressed as follows:

$$-\partial_t V = \min_{\mathbf{u}}\left(\boldsymbol{\mathcal{L}} + (\nabla_{\mathbf{x}}V)^T\mathbf{F} + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}V)\mathbf{B}\mathbf{B}^T\right)\right) \tag{31}$$

To find the minimum, the cost function (30) is inserted into (31) 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:

$$\mathbf{u}(\mathbf{x}_t) = -\mathbf{R}^{-1}\left(q_1(\mathbf{x},t) + \mathbf{G}(\mathbf{x})^T\nabla_{\mathbf{x}}V(\mathbf{x},t)\right) \tag{32}$$

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:

$$\begin{aligned}-\partial_t V = \tilde{q} + (\nabla_{\mathbf{x}}V)^T\tilde{\mathbf{f}} - \frac{1}{2}(\nabla_{\mathbf{x}}V)^T\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T(\nabla_{\mathbf{x}}V)\\ + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}V)\mathbf{B}\mathbf{B}^T\right)\end{aligned} \tag{33}$$

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:

$$-\partial_t\Psi = -\frac{1}{\lambda}\tilde{q}\Psi + \tilde{\mathbf{f}}^T(\nabla_{\mathbf{x}}\Psi) + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}\Psi)\mathbf{\Sigma}\right) \tag{34}$$

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 (34) yields its solution in form of an expectation over system trajectories. This solution is mathematically expressed as:

$$\Psi\left(\mathbf{x}_{t_i}\right) = E_{\mathbb{P}}\left[\exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x})dt\right)\Psi(\mathbf{x}_{t_N})\right] \tag{35}$$

The expectation $E_{\mathbb{P}}$ is taken on sample paths 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 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 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 (29), it can be trivially shown that:

$$\begin{aligned} V(\mathbf{x},t_i) &= -\lambda \log E_{\mathbb{P}}\left[\exp\left(-\int_{t_i}^{t_N} \frac{1}{\lambda}\tilde{q}(\mathbf{x})dt\right)\Psi(\mathbf{x}_{t_N})\right] \\ &\leq E_{\mathbb{Q}}\Big(J(\mathbf{x},\mathbf{u})\Big) \end{aligned} \tag{36}$$

Note that the inequality above in similar to (17) when the following equations hold: $q_1(\mathbf{x}) = 0, \quad \mathbf{R} = I, \quad \lambda = \frac{1}{|\rho|}, \mathbf{G} = \boldsymbol{\mathcal{B}}, \mathbf{B} = \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}$

VI. Discrete-time results

In this section, we show that the general framework of section II applies to control problems in discrete time as well. In particular, we will derive the framework of linearly solvable MDPs or KL Control [34] as a special case of the general framework.

To make things concrete, consider a state space $\mathcal{X}$ ($= \mathcal{Z}$ in section II) and a finite-horizon discrete time dynamical system:

$$\mathbf{x}_{t+1} \sim \mathcal{P}\left(\cdot|\mathbf{x}_t\right)$$

where Π is the probability transition density. To simplify the exposition, we will assume that $\mathcal{X}$ is a finite state space. Consider trajectories of length T from this system $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T]$. The probability density function over trajectories is simply

$$\mathcal{P}\left(\mathbf{X}\right) = [\mathbf{x}_0; \mathbf{x}_1; \ldots; \mathbf{x}_T] = \prod_{t=0}^{T-1} \mathcal{P}\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right). \tag{37}$$

Now consider applying control to this dynamical system to change the trasition density to $\mathcal{U}\left(\cdot|\mathbf{x}_t\right)$ and the corresponding trajectory density to $\mathcal{U}\left(\mathbf{X}\right)$.

In the framework of KL-control or Linearly Solvable MDPs [34] [33] [31], the control designer is allowed to pick $\mathcal{U}\left(\cdot|\mathbf{x}_t\right)$ however he wishes, as long as it has the same support as $\mathcal{P}\left(\cdot|\mathbf{x}_t\right)$, but he needs to pay a price equal to the KL divergence (or relative entropy) $\mathcal{I}\left(\mathcal{U}\left(\cdot|\mathbf{x}_t\right) \parallel \mathcal{P}\left(\cdot|\mathbf{x}_t\right)\right)$ (akin to a control cost) in addition to a state cost $\rho\mathcal{J}\left(\mathbf{x}_t\right)$ (where ρ is a scaling factor). The expectation of the total cost under $\mathcal{U}\left(\mathbf{X}\right)$ then becomes

$$\underset{\mathbf{X}\sim\mathcal{U}(\cdot)}{E}\left[\sum_t \rho\mathcal{J}\left(\mathbf{x}_t\right) + \mathcal{I}\left(\mathcal{U}\left(\cdot|\mathbf{x}_t\right) \parallel \mathcal{P}\left(\cdot|\mathbf{x}_t\right)\right)\right].$$

By exploiting the Markovian structure of $\mathcal{U}\left(\mathbf{X}\right)$ (37), the second term can be rewritten as

$$\begin{aligned}
&\sum_{\mathbf{X}} \prod_{\tau=0}^{T-1} \mathcal{U}\left(\mathbf{x}_{\tau+1}|\mathbf{x}_\tau\right)\left(\sum_{t=0}^{T-1}\sum_{\mathbf{x}'} \mathcal{U}\left(\mathbf{x}'|\mathbf{x}_t\right)\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}_t\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}_t\right)}\right)\right) \\
&= \sum_{t=0}^{T-1}\sum_{\mathbf{X},\mathbf{x}'} \prod_{\tau=0}^{T-1} \mathcal{U}\left(\mathbf{x}_{\tau+1}|\mathbf{x}_\tau\right)\mathcal{U}\left(\mathbf{x}'|\mathbf{x}_t\right)\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}_t\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}_t\right)}\right) \\
&= \sum_{t=0}^{T-1}\sum_{\mathbf{x}_0,\ldots,\mathbf{x}_t,\mathbf{x}'} \prod_{\tau=0}^{t-1} \mathcal{U}\left(\mathbf{x}_{\tau+1}|\mathbf{x}_\tau\right)\mathcal{U}\left(\mathbf{x}'|\mathbf{x}_t\right)\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}_t\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}_t\right)}\right) \\
&= \sum_{t=0}^{T-1}\sum_{\mathbf{x}_0,\ldots,\mathbf{x}_{t+1}} \prod_{\tau=0}^{t-1} \mathcal{U}\left(\mathbf{x}_{\tau+1}|\mathbf{x}_\tau\right)\log\left(\frac{\mathcal{U}\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)}{\mathcal{P}\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)}\right) \\
&= \sum_{t=0}^{T-1}\sum_{\mathbf{x}_0,\ldots,\mathbf{x}_T} \prod_{\tau=0}^{T-1} \mathcal{U}\left(\mathbf{x}_{\tau+1}|\mathbf{x}_\tau\right)\log\left(\frac{\mathcal{U}\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)}{\mathcal{P}\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)}\right) \\
&= \sum_{\mathbf{x}_0,\ldots,\mathbf{x}_T} \prod_{\tau=0}^{T-1} \mathcal{U}\left(\mathbf{x}_{\tau+1}|\mathbf{x}_\tau\right)\left(\sum_{t=0}^{T-1}\log\left(\frac{\mathcal{U}\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)}{\mathcal{P}\left(\mathbf{x}_{t+1}|\mathbf{x}_t\right)}\right)\right) \\
&= \sum_{\mathbf{X}} \mathcal{U}\left(\mathbf{X}\right)\log\left(\frac{\mathcal{U}\left(\mathbf{X}\right)}{\mathcal{P}\left(\mathbf{X}\right)}\right) = \mathcal{I}\left(\mathcal{U}\left(\cdot\right) \parallel \mathcal{P}\left(\cdot\right)\right)
\end{aligned}$$

Denoting $\mathcal{J}\left(\mathbf{X}\right) = \sum_t \mathcal{J}\left(\mathbf{x}_t\right)$, the overall control objective becomes

$$\underbrace{\rho \underset{\mathbf{X}\sim\mathcal{U}(\cdot)}{E}\left[\mathcal{J}\left(\mathbf{X}\right)\right]}_{\text{Expected state cost}} \quad + \quad \underbrace{\mathcal{I}\left(\mathcal{U}\left(\cdot\right) \parallel \mathcal{P}\left(\cdot\right)\right)}_{\text{KL Control Cost}}.$$

Thus, the control problem amounts to

$$\min_{\mathcal{U}(\cdot)} \underset{\mathbf{X}\sim\mathcal{U}(\cdot)}{E}\left[\mathcal{J}\left(\mathbf{X}\right)\right] + \frac{1}{\rho}\mathcal{I}\left(\mathcal{U}\left(\cdot\right) \parallel \mathcal{P}\left(\cdot\right)\right)$$

Replacing the densities $\mathcal{U}\left(\mathbf{X}\right)$ and $\mathcal{P}\left(\mathbf{X}\right)$ with the corresponding measures $\mathbb{Q}$ and $\mathbb{P}$ from section II, this matches the RHS of equation (4). The inequality still holds valid:

$$\underset{\mathbf{X}\sim\mathcal{P}(\cdot)}{E}\left[\mathcal{J}\left(\mathbf{X}\right)\right] \leq \min_{\mathcal{U}(\cdot)} \underset{\mathbf{X}\sim\mathcal{U}(\cdot)}{E}\left[\mathcal{J}\left(\mathbf{X}\right)\right] + \frac{1}{\rho}\mathcal{I}\left(\mathcal{U}\left(\cdot\right) \parallel \mathcal{P}\left(\cdot\right)\right).$$

Thus, the framework of KL-control can be derived as a special case of the general measure-theoretic formulation of relative-entropy control presented in section II.

A. Derivation from the Bellman Optimality Principle

As with the continuous time case, we can derive things based on the Bellman optimality principle in the discrete-time setting as well. We formalize the problem as a Markov Decision Process (MDP) with a stagewise cost described as above:

$$\begin{aligned}
&\mathcal{J}\left(\mathbf{x}\right) + \mathcal{I}\left(\mathcal{U}\left(\cdot|\mathbf{x}\right) \parallel \mathcal{P}\left(\cdot|\mathbf{x}\right)\right) \\
&= \mathcal{J}\left(\mathbf{x}\right) + \underset{\mathbf{x}'\sim\mathcal{U}(\cdot|\mathbf{x})}{E}\left[\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)}\right)\right]
\end{aligned}$$

Application of the Bellman principle of optimality in the finite horizon case gives us :

$$V_t(\mathbf{x}) = \min_{\mathcal{U}(\cdot|\mathbf{x})} \left(\mathcal{J}(\mathbf{x}) + \underset{\mathcal{U}(\cdot|\mathbf{x})}{E} \left[\log \left(\frac{\mathcal{U}(\mathbf{x}'|\mathbf{x})}{\mathcal{P}(\mathbf{x}'|\mathbf{x})} \right) + V_{t+1}(\mathbf{x}') \right] \right)$$

where $V_t(\mathbf{x})$ is the time-varying cost-to-go function. The $\mathcal{U}(\cdot|\mathbf{x})$ dependent terms in the functional above are minimized and thus we will have that:

$$\underset{\mathbf{x}' \sim \mathcal{U}(\cdot|\mathbf{x})}{E} \left[\log \left(\frac{\mathcal{U}(\mathbf{x}'|\mathbf{x})}{\mathcal{P}(\mathbf{x}'|\mathbf{x})} \right) + V_{t+1}(\mathbf{x}') \right] =$$
$$\underset{\mathbf{x}' \sim \mathcal{U}(\cdot|\mathbf{x})}{E} \left[\log \left(\frac{\mathcal{U}(\mathbf{x}'|\mathbf{x})}{\mathcal{P}(\mathbf{x}'|\mathbf{x})} \right) + \log \left(\frac{1}{\exp(-V_{t+1}(\mathbf{x}'))} \right) \right] =$$
$$\underset{\mathbf{x}' \sim \mathcal{U}(\cdot|\mathbf{x})}{E} \left[\log \left(\frac{\mathcal{U}(\mathbf{x}'|\mathbf{x})}{\mathcal{P}(\mathbf{x}'|\mathbf{x}) \exp(-V_{t+1}(\mathbf{x}'))} \right) \right]$$

For the purposes the normalization term $\mathcal{G}_t[\Phi](\mathbf{x})$ is introduced with $\Phi_t(\mathbf{x}) = \exp(-V_t(\mathbf{x}))$ being the *desirability* function. More precisely we will have:

$$\mathcal{G}_t[\Phi](\mathbf{x}) = \sum_{\mathbf{x}'} \mathcal{P}(\mathbf{x}'|\mathbf{x}) \Phi_{t+1}(\mathbf{x}') = \underset{\mathbf{x}' \sim \mathcal{P}(\cdot|\mathbf{x})}{E} [\Phi_{t+1}(\mathbf{x}')]$$

Therefore we have

$$\underset{\mathbf{x}' \sim \mathcal{U}(\cdot|\mathbf{x})}{E} \left[\log \left(\frac{\mathcal{U}(\mathbf{x}'|\mathbf{x})}{\mathcal{P}(\mathbf{x}'|\mathbf{x})} \right) + V_{t+1}(\mathbf{x}') \right] =$$
$$-\log(\mathcal{G}_t[\Phi](\mathbf{x})) + \mathcal{I}\left(\mathcal{U}(\cdot|\mathbf{x}) \parallel \frac{\mathcal{P}(\mathbf{x}'|\mathbf{x}) \Phi_{t+1}(\mathbf{x}')}{\mathcal{G}_t[\Phi](\mathbf{x})} \right)$$

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

$$\min_{\mathbf{u} \in \mathcal{U}} \mathcal{J}(\mathbf{x}) - \log(\mathcal{G}_t[\Phi](\mathbf{x})) + \mathcal{I}\left(\mathcal{U}(\cdot|\mathbf{x}) \parallel \frac{\mathcal{P}(\mathbf{x}'|\mathbf{x}) \Phi_{t+1}(\mathbf{x}')}{\mathcal{G}_t[\Phi](\mathbf{x})} \right)$$

The minimum of the Bellman equation is attained by:

$$\boxed{\mathcal{U}^*(\mathbf{x}'|\mathbf{x}) = \frac{\mathcal{P}(\mathbf{x}'|\mathbf{x}) \Phi_{t+1}(\mathbf{x}')}{\mathcal{G}_t[\Phi](\mathbf{x})}} \tag{38}$$

Substitution of the optimal distribution above will result in the linear Bellman equation:

$$\Phi_t(\mathbf{x}) = \exp(-\mathcal{J}(\mathbf{x})) \mathcal{G}_t[\Phi](\mathbf{x}) \tag{39}$$

This can be used to prove the path-integral representation of the desirability function

$$\Phi_t(\mathbf{x}) = \underset{\mathbf{x}_{\tau+1} \sim \mathcal{P}(\cdot|\mathbf{x}_\tau)}{E} \left[\exp\left(-\sum_{\tau=t}^{T} \mathcal{J}(\mathbf{x}_\tau) \right) \right].$$

Thus, the desirability function is just the expectation under the uncontrolled dynamics of the exponentiated path cost starting at state $\mathbf{x}$ at time t. This gives an expression for the optimally controlled trajectory distribution $\mathcal{U}(\mathbf{X})$:

$$\mathcal{U}(\mathbf{X}) = \frac{\mathcal{P}(\mathbf{X}) \exp(-\mathcal{J}(\mathbf{X}))}{E_{\mathbf{X}' \sim \mathcal{P}(\cdot)} [\exp(-\mathcal{J}(\mathbf{X}'))]}$$

which is identical to equation (5).

VII. Discussion

The PI and KL control frameworks presented here constitute a rich mathematical framework that has recently received a lot of attention, following Kappen's work on control-affine diffusions in continuous time [16], and Todorov's work on Markov decision processes in discrete time [32]. The initial studies [16], [32] were done independently, yet they both built upon the same earlier results which we discuss here. For over 30 years these earlier results had remained a curious mathematical fact, that was never actually used to solve control problems.

In continuous time, the trick that makes the HJB equation linear is Applying this exponential (or logarithmic) transformation to 2nd-order PDEs has a long history in Physics [12], [13]. Its first application to control was due to Fleming and Mitter, who showed that non-linear filtering corresponds to a stochastic optimal control problem whose HJB equation can be made linear [5]. Kappen generalized this idea, and noted that the solution to the resulting linear PDE is also a path integral – which yields sampling approximations to the optimal value function [16].

Todorov's work on the KL control framework [32] was motivated by the same earlier results but in a more abstract way: Todorov asked the question, are there classes of linearly-solvable optimal control problems involving arbitrary dynamics? This led to the KL control framework. In discrete time, the trick that makes the Bellman equation linear is

$$\min_q \{\mathcal{I}(q \parallel p) + \underset{q}{E}[V()]\} = -\log \underset{p}{E}[\exp(-V())]$$

where the minimum is achieved at $q^* = \exp(-V()) \times p$. Todorov introduced this trick in [32]. In an earlier paper, Mitter had used very similar ideas to provide a variational interpretation of Bayesian estimation [18]. Indeed if p is a prior and $V()$ is a negative log-likelihood, then the above q^* is a Bayesian posterior.

Theodorou built on Kappen's work [16] in subsequent works [27] and developed the PI^2 algorithm, which has since been successfully applied to many robotic control tasks [2], [3], [28]. He also generalized this work to controlled jump diffusion processes [30] and derived the free-energy interpretation of this work [29] which we presented in this paper. We showed here that this can be viewed as a unifying framework from which all the above works, to the best of our knowledge, can be derived as special cases.

We believe that the free energy interpretation brings together different points of view and allows us to understand these works from an information theoretic perspective. Further, it is mathematically convenient and applies directly to both discrete time and continuous time problems (including jump diffusion processes). There are several interesting and important directions for further work: The probabilistic representation (5) here allows us to explore the use of more sophisticated Monte Carlo algorithms [1] to compute the sampling-based control solution efficiently. Since the control solution we've developed requires sampling-based

exploration, another important area of research is ensuring that exploration is done safely and without damage to the system [19]. Another area of work is to exploit model-based trajectory optimization and control methods [25] (perhaps based on partial/incorrect models) to perform efficient sampling in the model-free methods we developed here.

REFERENCES

[1] S. Brooks, A Gelman, G Jones, and X. L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 1 edition, 2011.

[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] W. Fleming and S. Mitter. Optimal control and nonlinear filtering for nondegenerate diffusion processes. *Stochastics*, 8:226–261, 1982.

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

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

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

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

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

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

[12] C. Holland. A new energy characterization of the smallest eigenvalue of the Schrödinger equation. *Comm Pure Appl Math*, 30:755–765, 1977.

[13] B. Hopf. The partial differential equation $u_t + uu_x = \mu u_{xx}$. *Comm Pure Appl Math*, 3:201–230, 1950.

[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] H.J. Kappen. Linear theory for control of nonlinear stochastic systems. *Physical Review Letters*, 95(20):200201, 2005.

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

[18] S. Mitter and N. Newton. A variational approach to nonlinear estimation. *SIAM J Control Opt*, 42:1813–1833, 2003.

[19] T. M. Moldovan and P. Abbeel. Safe exploration in markov decision processes. In *ICML*, 2012.

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

[21] 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.

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

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

[24] 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.

[25] Y Tassa, T. Erez, and E. Todorov. Synthesis and stabilization of complex behaviors through online trajectory optimization. In *IROS*, pages 4906–4913, 2012.

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

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

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

[29] E.A. Theodorou and E. Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*, pages 1466–1473, 2012.

[30] E.A. Theodorou and E. Todorov. Stochastic optimal control for nonlinear markov jump diffusion processes. In *American Control Conference (ACC), 2012*, pages 1633–1639, 2012.

[31] 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.

[32] E. Todorov. Linearly-solvable Markov decision problems. *Advances in neural information processing systems*, 19:1369, 2007.

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

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

[35] 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.

[36] A. Wehrl. The many facets of entropy. *Reports on Mathematical Physics*, 30(1):119 – 129, 1991.