Information Theoretic views of Path Integral Control.

Evangelos A. Theodorou and Emannuel Todorov

Abstract

We derive the connections of Path Integral(PI) and Kulback-Liebler(KL) control as presented in machine learning and robotics communities [1–4] with earlier work in controls on the fundamental dualities between relative entropy and free energy and the logarithmic transformations of diffusions processes [5–10]. Our analysis offers an information theoretic view of PI stochastic optimal control based on the duality between Free Energy and Relative Entropy as expressed by the Legendre-Fenchel transformation in Statistical Mechanics [11]. Finally we overview the cases of partial observability and min-max control.

1 Introduction

Recent developments in nonlinear stochastic optimal control and machine learning suggest new ways to solve optimization problems for dynamical systems in continuous state action spaces. The mathematical analysis is based on the quantum-mechanical concept of Path Integral. The path integral formalism provides an unified view of Newtonian and Quantum mechanics since it generalizes the action principle from Newtonian to Quantum mechanics. In stochastic optimal control theory, path integrals are used to represent value functions and solutions of partial differential equations. Here we provide an information theoretic view of path integral control and show its connections to earlier findings in controls. To do so, in section 2 we derive the dualities between free energy and relative entropy. In section 3 we derive PI control based on these dualities and discuss the cases of min-max control and partial observability. In section 4 we derive PI based on the Bellman principle for continuous and discrete time. In the last section 5 we compare the different approaches to PI control and conclude.

2 Free Energy and Relative Entropy Dualities

In this section we derive the fundamental duality relationships between free energy and relative entropy [9]. This relationship is important for the derivation of stochastic optimal control. Let $(\mathcal{Z}, \boldsymbol{\mathcal{Z}})$ denote a measurable space and $\boldsymbol{\mathcal{P}}(\mathcal{Z})$ the corresponding probability measure defined on the measurable space. For our analysis we consider the following definitions.

Definition 1: Let $\mathbb{P} \in \boldsymbol{\mathcal{P}}(\mathcal{Z})$ and the function $\mathcal{J}(\mathbf{x}) : \mathcal{Z} \rightarrow \Re$ be a measurable function. Then the term: $\mathbb{E}\Big(\mathcal{J}(\mathbf{x})\Big) = \log \int \exp\left(\rho \mathcal{J}(\mathbf{x})\right) d\mathbb{P}$ 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:

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

We will also consider the objective function: $\xi(\mathbf{x}) = \frac{1}{\rho}\mathbb{E}\Big(\mathcal{J}(\mathbf{x})\Big) = \frac{1}{\rho}\log \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}\Big[\exp\left(\rho \mathcal{J}(\mathbf{x})\right)\Big]$ with $\mathcal{J}(\mathbf{x}) = \phi(\mathbf{x}_{t_N}) + \int_{t_i}^{t_N} q(\mathbf{x}) dt$ is the state dependent cost. The objective function above takes

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

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

By taking the logarithm of both sides of the equations above and making use of the Jensen's inequality we will have: $\log \mathcal{E}^{(0)}_{\boldsymbol{\tau}_i}\left[\exp\left(\rho\mathcal{J}(\mathbf{x})\right)\right] = \log \int \exp\left(\rho\mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}} d\mathbb{Q} \geq \int \log\left(\exp\left(\rho\mathcal{J}(\mathbf{x})\right) \frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{Q}$. We multiply the inequality above with $\frac{1}{\rho}$ for case of $\rho < 0$ or $\rho = -|\rho|$ and thus we have:

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

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:

$$-\frac{\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right)}{|\rho|} = \inf_{\mathbb{Q}}\left[\mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right)\right] \quad \text{and} \quad -\frac{\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right)}{|\rho|} = \inf_{\mathbb{Q}}\left[\mathcal{E}^{(1)}\left(\mathcal{J}(\mathbf{x})\right) + \frac{1}{\rho}\mathbb{E}\left(\mathcal{J}(\mathbf{x})\right)\right] \tag{2}$$

The infimum in (2) is attained at $\mathbb{Q}^*$ given by:

$$d\mathbb{Q}^* = \frac{\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right) d\mathbb{P}}{\int \exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right) d\mathbb{P}} \tag{3}$$

When $\rho > 0$ the inequality in (1) becomes from $\leq$ to $\geq$ and the $\inf$ in (2) and becomes $\sup$. Therefore we will have that:

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

In the next section we show how inequality (2) is transformed to a stochastic optimal control problem for the case of markov diffusion processes.

3 Information theoretic view of stochastic optimal control.

We consider the uncontrolled and controlled stochastic dynamics of the form: $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}(\mathbf{x})d\mathbf{w}^{(0)}(t)$ and $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}dt + \frac{1}{\sqrt{|\rho|}}d\mathbf{w}^{(1)}(t)\right)$ with $\mathbf{x}_t \in \Re^{n\times 1}$ denoting the state of the system, $\boldsymbol{\mathcal{B}}(\mathbf{x},t) : \Re^n \times \Re \to \Re^{n\times n}$ is the control and diffusions matrix, $\mathbf{f}(\mathbf{x},t) : \Re^n \times \Re \to \Re^{n\times 1}$ 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. 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 $\mathcal{E}^{(0)}$ and $\mathcal{E}^{(1)}$ respectively. The corresponding probability measures of the aforementioned expectations are $\mathbb{P}$ and $\mathbb{Q}$. We continue our analysis with the main result in (1) and the definition of the Radon-Nikodým derivative: $\frac{d\mathbb{Q}}{d\mathbb{P}} = \exp\left(\zeta(\mathbf{u})\right)$ and $\frac{d\mathbb{P}}{d\mathbb{Q}} = \exp\left(-\zeta(\mathbf{u})\right)$ where according to Girsanov's theorem [12] adapted to the diffusion processes considered here, the

term $\zeta(\mathbf{u})$ is expressed as: $\zeta(\mathbf{u}) = \frac{1}{2}|\rho| \int_{t_i}^{t_N} \mathbf{u}^T \mathbf{u} dt + \sqrt{|\rho|} \int_{t_i}^{t_N} \mathbf{u}^T d\mathbf{w}^{(1)}(t)$. Substitution of the Radon-Nikodým derivative into (2) gives the following result:

$$\xi(\mathbf{x}) = -\frac{1}{|\rho|} \log \mathcal{E}^{(0)} \left[\exp\left(-|\rho| \mathcal{J}(\mathbf{x})\right) \right] \leq \mathcal{E}^{(1)} \left[\mathcal{J}(\mathbf{x}) + \frac{1}{|\rho|} \zeta(\mathbf{u}) \right] = \mathcal{E}^{(1)} \left[\mathcal{J}(\mathbf{x}) + \frac{1}{2} \int_{t_i}^{t_N} \mathbf{u}^T \mathbf{u} dt \right] \tag{5}$$

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 (5) 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 (5) 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 (5). Essentially inequality (5) defines a minimization problem 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 (5) reaches its optimal $\xi(\mathbf{x})$.

An important question to is the link between (5) 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 $\Phi(\mathbf{x}, t) = \mathcal{E}^{(0)}(\exp(\rho \mathcal{J}(\mathbf{x})))$. The Feynman-Kac lemma [13] tells us that this function satisfies the backward Chapman Kolmogorov PDE. Therefore the following equation holds:

$$-\partial_t \Phi = \rho q_0 \Phi + \mathbf{f}^T (\nabla_{\mathbf{x}} \Phi) + \frac{1}{2|\rho|} tr\left((\nabla_{\mathbf{xx}} \Phi) \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{B}}^T \right) \tag{6}$$

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:

$$-\partial_t \xi = q_0 + (\nabla_{\mathbf{x}} \xi)^T \mathbf{f} - \underline{\frac{1}{2} (\nabla_{\mathbf{x}} \xi)^T \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{B}}^T (\nabla_{\mathbf{x}} \xi)} + \frac{1}{2|\rho|} tr\left((\nabla_{\mathbf{xx}} \xi) \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{B}}^T \right) \tag{7}$$

The nonlinear PDEs above corresponds to the HJB equation [14] for the case of the minimizing optimal control problem with control weight $\mathbf{R} = I$ and therefore, $\xi(\mathbf{x})$ is the corresponding minimizing value function. Note that in order to derive the PDEs above we did not use any principle of optimality. Similar results can be derived for $\rho = |\rho| > 0$ as shown in [6]. 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 uncontrolled and controlled dynamics respectively. The function $\xi(\mathbf{x}, t)$ specified as: $\xi(\mathbf{x}, t) = \frac{sign(\rho)}{|\rho|} \log \mathcal{E}^{(0)}(\exp(sign(\rho)|\rho| \mathcal{J}(\mathbf{x})))$ is the value function of the stochastic optimal control problems: $\xi(\mathbf{x}, t_i) = \min_{\mathbf{u}} \mathcal{E}^{(1)} \int_{t_i}^{t_N} \left(q(\mathbf{x}) - \frac{1}{2} \mathbf{u}^T \mathbf{u} \right) dt, \ \ \forall \rho > 0$ and $\xi(\mathbf{x}, t_i) = \max_{\mathbf{u}} \mathcal{E}^{(1)} \int_{t_i}^{t_N} \left(q(\mathbf{x}) + \frac{1}{2} \mathbf{u}^T \mathbf{u} \right) dt, \ \ \forall \rho < 0$.*

3.1 Information theoretic view of stochastic optimal control: The case of partial observability.

In the partial observable case [9], besides the stochastic dynamics there are also the observation diffusions:

$$d\mathbf{y} = \mathbf{h}(\mathbf{x})dt + \frac{1}{\sqrt{|\rho|}}\mathbf{C}(\mathbf{x})d\mathbf{v}^{(0)}(t) \text{ and } d\mathbf{y} = \mathbf{h}(\mathbf{x})dt + \mathbf{C}(\mathbf{x})\left(\mathbf{b}dt + \frac{1}{\sqrt{|\rho|}}\mathbf{C}(\mathbf{x})d\mathbf{v}^{(1)}(t)\right) \quad (8)$$

The term $\mathbf{y}$ denotes the observations, $\mathbf{b}$ plays the role of control that act on observations, and $d\mathbf{v}^0$ and $d\mathbf{v}^0$ is the observation noise under the uncontrolled and controlled measurement dynamics. In this case the Radon-Nikodým derivative is expressed as $\frac{d\mathbb{P}}{d\mathbb{Q}} = \exp(-\zeta(\mathbf{u}))$ with $\zeta(\mathbf{u}) = \frac{1}{2}|\rho|\int_{t_i}^{t_N}\left(\mathbf{u}^T\mathbf{u} + \mathbf{b}^T\mathbf{b}\right)dt + \sqrt{|\rho|}\int_{t_i}^{t_N}\left(\mathbf{u}^T d\mathbf{w}^{(1)}(t) + \mathbf{b}^T d\mathbf{v}^{(1)}(t)\right)$. Under the observation dynamics above the free energy is the lower bound of the following expression:

$$\xi(\mathbf{x}) = -\frac{1}{|\rho|}\log\mathcal{E}^{(0)}\left[\exp\left(-|\rho|\mathcal{J}(\mathbf{x})\right)\right] \leq \mathcal{E}^{(1)}\left[\mathcal{J}(\mathbf{x}) + \frac{1}{2}\int_{t_i}^{t_N}\left(\mathbf{u}^T\mathbf{u} + \mathbf{b}^T\mathbf{b}\right)dt\right] \quad (9)$$

The expectation $\mathcal{E}^{(0)}$ is with respect to the process and observations noise $d\mathbf{w}^{(0)}(t), d\mathbf{v}^{(0)}(t)$ while the expectation $\mathcal{E}^{(1)}$ is with respect to $d\mathbf{w}^{(1)}(t), d\mathbf{v}^{(1)}(t)$. Again the free energy is the lower bound on a cost that is typically found in stochastic optimal control. The partial observable case includes controls in the observations.

3.2 Information theoretic view of stochastic optimal control: The case of min-max optimal control.

We consider stochastic dynamics : $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}dt + \frac{1}{\sqrt{|\rho|}}d\mathbf{w}^{(0)}(t)\right)$ and cost function $S(\mathbf{x}) = \mathcal{E}^{(0)}\left(\exp\left(|\rho|\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u})dt\right)\right) = \mathcal{E}^{(0)}\left(\exp\left(|\rho|\int_{t_i}^{t_N}L(\mathbf{x},\mathbf{u})dt\right)\right)$. Next we define free energy as follows $\mathbb{E}\left(\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u})\right) = \log\int\exp\left(\rho\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u})\right)d\mathbb{P}$ and use the Legendre transformation that leads to maximization problem: $\frac{\mathbb{E}(\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u}))}{|\rho|} = \sup\left[\mathcal{E}^{(1)}\left(\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u})\right) - \frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right)\right]$. By taking into account then stochastic dynamics and the Radon-Nikodým derivative we will have:

$$\begin{aligned}
\inf_{\mathbf{u}}\left[\frac{1}{|\rho|}\log\int\exp\left(\rho\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u})\right)d\mathbb{P}\right] &= \inf_{\mathbf{u}}\sup_{\mathbb{Q}}\left[\mathcal{E}^{(1)}\left(\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u})\right) - \frac{1}{|\rho|}\mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right)\right] \\
\inf_{\mathbf{u}}\left[\frac{1}{|\rho|}\log\int\exp\left(\rho\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u})\right)d\mathbb{P}\right] &= \inf_{\mathbf{u}}\sup_{\boldsymbol{\pi}}\left[\mathcal{E}^{(1)}\left(\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u}) - \frac{1}{2}\int_{t_i}^{t_N}\boldsymbol{\pi}^T\boldsymbol{\pi}dt\right)\right] \quad (10)
\end{aligned}$$

where $\mathcal{E}^{(1)}$ is the expectation on trajectories generated based on the diffusion $d\mathbf{x} = \mathbf{f}(\mathbf{x})dt + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}dt + \boldsymbol{\pi}dt + \frac{1}{\sqrt{|\rho|}}d\mathbf{w}^{(1)}(t)\right)$. The term $\boldsymbol{\pi}$ can be thought as a destabilizing controller. Equation (10) is a short but elegant way to show the equivalence of min-max control and differential game theory with risk sensitivity. Essentially all that it takes is an appropriate definition of free energy and the use of the Legendre transformation. Notice that no dynamic programming arguments were used in this analysis.

4 Derivation based on Bellman Principle: The continuous case.

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:

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

subject to the nonlinear stochastic dynamics: $\mathbf{dx} = \mathbf{F}(\mathbf{x},\mathbf{u})dt + \mathbf{B}(\mathbf{x})d\mathbf{w}$ 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: $\boldsymbol{\mathcal{L}}(\mathbf{x},\mathbf{u},t) = q_0(\mathbf{x},t) + q_1(\mathbf{x},t)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbf{R}\mathbf{u}$. 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 [5, 14] associated with this stochastic optimal control problem is expressed as follows: $-\partial_t V = \min_{\mathbf{u}}\left(\boldsymbol{\mathcal{L}} + (\nabla_{\mathbf{x}}V)^T\mathbf{F} + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}V)\mathbf{B}\mathbf{B}^T\right)\right)$. The corresponding optimal control is given by the equation: $\mathbf{u}(\mathbf{x}_t) = -\mathbf{R}^{-1}\left(q_1(\mathbf{x},t) + \mathbf{G}(\mathbf{x})^T\nabla_{\mathbf{x}}V(\mathbf{x},t)\right)$. 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:

$$-\partial_t V = \tilde{q} + (\nabla_{\mathbf{x}}V)^T\tilde{\mathbf{f}} - \underline{\frac{1}{2}(\nabla_{\mathbf{x}}V)^T\mathbf{G}\mathbf{R}^{-1}\mathbf{G}^T(\nabla_{\mathbf{x}}V)} + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}V)\mathbf{B}\mathbf{B}^T\right) \tag{11}$$

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 $\lambda\mathbf{G}(\mathbf{x})\mathbf{R}^{-1}\mathbf{G}(\mathbf{x})^T = \mathbf{B}(\mathbf{x})\mathbf{B}(\mathbf{x})^T = \boldsymbol{\Sigma}(\mathbf{x}_t) = \boldsymbol{\Sigma}$ the resulting PDE is formulated as follows:

$$-\partial_t\Psi = -\frac{1}{\lambda}\tilde{q}\Psi + \tilde{\mathbf{f}}^T(\nabla_{\mathbf{x}}\Psi) + \frac{1}{2}tr\left((\nabla_{\mathbf{xx}}\Psi)\boldsymbol{\Sigma}\right) \tag{12}$$

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

$$\Psi\left(\mathbf{x}_{t_i}\right) = E^{(0)}\left[\exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x})dt\right)\Psi(\mathbf{x}_{t_N})\right] \tag{13}$$

The expectation $E^{(0)}$ 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 the function $V(\mathbf{x},t)$ is the minimum of the expectation of the objective function $J(\mathbf{x},\mathbf{u})$, it can be trivially shown that:

$$V(\mathbf{x},t_i) = -\lambda\log E^{(0)}\left[\exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x})dt\right)\Psi(\mathbf{x}_{t_N})\right] \leq E^{(1)}\left(J(\mathbf{x},\mathbf{u})\right) \tag{14}$$

Note that the inequality above in similar to (5) when the following equations hold: $q_1(\mathbf{x}) = 0,\ \mathbf{R} = I,\ \lambda = \frac{1}{|\rho|}, \mathbf{G} = \boldsymbol{\mathcal{B}}, \mathbf{B} = \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}$. 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 $\mathcal{E}^{(0)} \equiv E^{(0)}$ and $\mathcal{E}^{(1)} \equiv E^{(1)}$. Under the conditions above the Kolmogorov PDEs (6) and (12) and the HJB equations (11) and (7) are identical.

4.1 Derivation based on Bellman Principle: The discrete case.

In the KL control framework [3,4,15] the analysis starts with the application of the Bellman principle of optimality on Markov Decision Processes (MDP) and under the running cost specified as a sum of a state depended term and the Kullback Leibler Divergence between the transition densities of the controlled and uncontrolled dynamics. In particular, the running cost is specified as $L\left(\mathbf{x},\mathbf{u}\right) = q(\mathbf{x}) + \mathcal{H}\left(\mathbb{Q}||\mathbb{P}\right) = q(\mathbf{x}) + \mathcal{E}^{(1)}\left(\log\frac{p(\mathbf{x}'|\mathbf{x},\mathbf{u})}{p(\mathbf{x}'|\mathbf{x})}\right)$. The transition probabilities under the controlled and uncontrolled dynamics are represented as $p(\mathbf{x}'|\mathbf{x},\mathbf{u})$ and $p(\mathbf{x}'|\mathbf{x})$. Application of the Bellman principle of optimality results in the minimization of the quantity:

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

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

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

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

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

Substitution of the expression above into the Bellman minimization equation results in: $\min_{\mathbf{u}\in\mathcal{U}} \left(q(\mathbf{x}) - \log \mathcal{G}[\Phi](\mathbf{x}) + \mathcal{H}\left(\mathbf{u}(\mathbf{x}|\mathbf{x}) \middle|\middle| \frac{\mathbf{p}(\mathbf{x}'|\mathbf{x})\Phi(\mathbf{x}')}{\mathcal{G}[\Phi](\mathbf{x})} \right) \right)$. The minimum of the Bellman equation is attained by: $p^*(\mathbf{x}'|\mathbf{x},\mathbf{u}) = \frac{\mathbf{p}(\mathbf{x}'|\mathbf{x})\Phi(\mathbf{x})}{\mathcal{G}[\Phi](\mathbf{x})}$. The last equation provides the transition probability under the optimal control law and in that sense it the optimal transition probability. Clearly the optimal distribution above is identical to equations (3). Substitution of the optimal distribution above will result in the Bellman equation: $\Phi(\mathbf{x}) = \exp\left(-q(\mathbf{x})\right)\mathcal{G}[\Phi](\mathbf{x}')$. The link with the continuous case is established by writing the Bellman equation for an MDP with continuous state space $\Phi_{(\delta t)}(\mathbf{x}) = \exp\left(-q(\mathbf{x})\delta t\right)\mathcal{G}[\Phi_{(\delta t)}](\mathbf{x}')$. Rearrangement of the terms results in: $\left(\exp\left(q(\mathbf{x})\delta t\right) - 1\right)\Phi_{(\delta t)}(\mathbf{x}) = \mathcal{E}^{(0)}(\Phi_{(\delta t)}(\mathbf{x}') - \Phi_{(\delta t)}(\mathbf{x}))$. Under the limit $\delta \to 0$ the equation results the backward Chaplman Kolmogorov PDE in (12) for $\rho = 1$.

5 Discussion

We show the connection of path integral control framework as presented in the machine learning and robotic communities [1–3, 16–18] with work in the control theoretic community on risk sensitivity [5, 6, 8, 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 transformation of the HJB in to the backward Chapman Kolmogorov. The Feynman-Kac lemma is applied and the solution of the Chapman Kolmogorov PDE together with the lower bound on the objective function are provided. The second methodological approach starts with the duality between free energy and relative entropy and the resulting optimization problem as expressed in (2). For diffusion processes affine in control and noise and under the use of Girsanov's theorem, the aforementioned optimization results in formulating the bound $\xi(\mathbf{x})$ of the objective function $\mathcal{E}^{(1)}(J(\mathbf{x},\mathbf{u}))$ which is typically found in stochastic optimal control. The link to Bellman optimality is established by showing that, this bound $\xi(\mathbf{x})$ satisfies the HJB equation and therefore it is a value function. Inside the class of the stochastic dynamics of markov diffusion processes affine in control and noise, Dynamic Programming is more general since it incorporates general cost functions and stochastic dynamics. This generalization however, is reduced by the assumption regarding control cost and the variance of the noise $\lambda\mathbf{G}(\mathbf{x})\mathbf{R}^{-1}\mathbf{G}(\mathbf{x})^T = \mathbf{B}(\mathbf{x})\mathbf{B}(\mathbf{x})^T$.

In the second approach the lower bound $\xi(\mathbf{x})$ of the accumulated trajectory cost $\mathcal{E}^{(1)}(J(\mathbf{x},\mathbf{u}))$ is derived without relying on the Bellman Principle. In fact, this lower bound defines a new form of optimality which, as it is shown in [5,6] as well as in this work, for the case of diffusion processes is equivalent to the Bellman principle of optimality. In the KL stochastic optimal control framework the derivation relies on the Bellman Principle of Optimality in discrete time. The resulting distribution

$p_k^*(\mathbf{x}'|\mathbf{x},\mathbf{u})$ is optimal since it is the distribution that results when actions are optimal. In that sense the KL framework, in its initial formulation [3] does not explicitly provide an optimal control law but instead it provides the optimal distribution or optimal transition probability under the use of optimal control law. For the case of control affine diffusions, KL control framework incorporates control-only and state-only depended terms in contrast to PI derived based on the Bellman principle in which cross terms between controls and states may be considered. The compositionally of optimal controls includes control laws and thus it can incorporate any analytically derived optimal control as well as PI control.

The information theoretic view and in particular the use of the Legendre transformation and the duality between free energy and relative is an elegant way to derive many important results in control theory and machine learning. Furthermore, it allows for generalizations in the sense that inequality (5) holds for more general models of stochasticity such as the case of jump diffusions.

References

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

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

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

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

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

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

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

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

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

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

[11] H. Touchette, "The large deviation approach to statistical mechanics," *Physics Reports*, vol. 478, pp. 1–69, 2009.

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

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

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

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

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

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

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