Entropy **2015**, *17*, 3352-3375; doi:10.3390/e17053352

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Nonlinear Stochastic Control and Information Theoretic Dualities: Connections, Interdependencies and Thermodynamic Interpretations

Evangelos A. Theodorou [1,2]

[1] The Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0150, USA; E-Mail: evangelos.theodorou@ae.gatech.edu

[2] Institute of Robotics and intelligent Machines, Georgia Institute of Technology, Atlanta, GA 30332-0150, USA

Academic Editor: Kevin H. Knuth

Received: 2 February 2015 / Accepted: 29 April 2015 / Published: 15 May 2015

Abstract: In this paper, we present connections between recent developments on the linearly-solvable stochastic optimal control framework with early work in control theory based on the fundamental dualities between free energy and relative entropy. We extend these connections to nonlinear stochastic systems with non-affine controls by using the generalized version of the Feynman–Kac lemma. We present alternative formulations of the linearly-solvable stochastic optimal control framework and discuss information theoretic and thermodynamic interpretations. On the algorithmic side, we present iterative stochastic optimal control algorithms and applications to nonlinear stochastic systems. We conclude with an overview of the frameworks presented and discuss limitations, differences and future directions.

Keywords: Stochastic Optimal Control; Information Theory; Thermodynamics

1. Introduction

While the topic of nonlinear stochastic control has been traditionally studied within control and applied mathematics, over the past 10–15 years, there has been an increasing interest by researchers in machine learning and robotics communities to expand nonlinear stochastic optimal control in terms of theoretical generalizations and algorithms. The main motivation for this increasing interest is the ability to solve stochastic optimal control problems with forward sampling of stochastic differential equations

(SDEs). There have been a few approaches in the literature on this topic, called path integral (PI) control [1–3], Kullback–Leibler (KL) control or linearly-solvable control [4,5].

The PI control framework is derived for continuous time stochastic systems affine in controls and noise and for finite horizon optimal control problems. In the KL control, the derivation is in discrete time Markov decision processes (MDPs) and includes finite horizon, infinite horizon, exponentially-discounted and first exit optimal control problems. The continuous time equivalent of the KL control is recovered when transition probabilities are defined based on the corresponding SDEs. Due the central role that linear partial differential equations (PDEs) play in the analysis of the aforementioned approaches, we will refer to them as linearly-solvable optimal control (LSOC), and when necessary, we will use the explicit names of PI or KL control. Moreover, we will restrict our analysis to the finite horizon case. Similar connections have been identified for the infinite horizon case [6]. The analysis for the infinite horizon case will be presented in a follow-up manuscript.

One of the important findings in the LSOC framework is the observation that under certain conditions related to the process noise and control authority, stochastic optimal control problems can be solved with forward sampling of SDEs and the evaluation of expectations. The fundamental theorem that made this observation possible, especially for the continuous case, is the Feynman–Kac lemma [7–9]. The Feynman–Kac lemma connects SDEs and linear backward PDEs by providing a probabilistic representation of solutions of backwards PDEs. Alternative computational algorithms to the sampling-based LSOC framework incorporate methods on low rank tensor approximation to find solution of linear PDEs on a domain of interest [10].

With the goal to unify different views on stochastic optimal control as developed within different disciplines in sciences and engineering, this work aims to present recent developments and to discuss their connections with previous work using information theoretic concepts. In particular, we expand upon our previous work on this topic [11] and present connections between the LSOC framework as presented within the machine learning and statistical physics communities with the information theoretic view of nonlinear stochastic optimal control theory using the free energy-relative entropy relationship [12–15].

Below, we summarize the main points of our analysis:

(i) The PI and KL control framework can be derived using the relative entropy-free energy relationship, and therefore, there are direct connections of the LSOC framework to previous work in control theory. These connections were recently shown in [11]. From the epistemological stand point, the aforementioned connections provide a deeper understanding of optimality principles and identify the conditions under which these optimality principles emerge from information theoretic postulates. Essentially, there are alternative views/methodological approaches of looking into nonlinear stochastic optimal control that are illustrated in Figure 1.

(ii) The derivation of nonlinear stochastic optimal control using the free energy and relative entropy relationship does not rely on the Bellman principle. In other words, one can derive the Hamilton–Jacobi–Bellman (HJB) equation without using dynamic programming. When the form of the optimal control policy has to be found, then the connection with stochastic optimal control based on dynamic programming is necessary. In this paper, we generalize the connection between free energy-relative entropy dualities and stochastic optimal control to systems that are non-affine in controls. The analysis leverages the generalized version of the Feynman–Kac lemma and identifies the necessary and sufficient conditions under which the aforementioned connections are valid. This generalization creates future research directions towards the development of optimal control algorithms for stochastic systems nonlinear in the state and control. In addition, it shows that there is a deeper relation between the Legendre transformation and stochastic control that goes beyond the class of control affine systems.

(iii) While typically in stochastic optimal control theory, the cost function is pre-specified, this is not the case when the stochastic optimal control framework is derived using the free energy-relative entropy relationship. In the latter case, the form of the cost function related to control effort emerges from the structure of the underlying stochastic dynamics. This observation indicates that there are strong interdependencies between cost functions and dynamics and that the choice of the control cost function is not arbitrary. Another way to understand the importance of the aforementioned interdependencies is that, while in the traditional approach, the cost function is imposed to the problem, in the information theoretic view of stochastic optimal control, the cost function partially emerges from the formulation of the problem (see Figure 1).

(iv) We illustrate connections between stochastic control and the maximum entropy principle. The analysis relies on the generalized Boltzmann, Gibbs and Shannon entropy [16]. We show that the stochastic control framework is recovered as the maximization of the generalized Boltzmann, Gibbs and Shannon entropy subject to energy and probability measure normalization constraints.

(v) For the class of stochastic systems that are affine in control and noise, there are cost function formulations that cannot be represented within the information theoretic approach. Thus, although the information theoretic formulation of stochastic optimal control provides a general framework, there are cases in which the *a priori* specification of the cost function and the use of dynamic programming provide more flexibility.

(vi) Besides the analysis on the connections between different formulations of stochastic optimal control theory, we also present iterative algorithms designed for stochastic systems and demonstrate some examples.

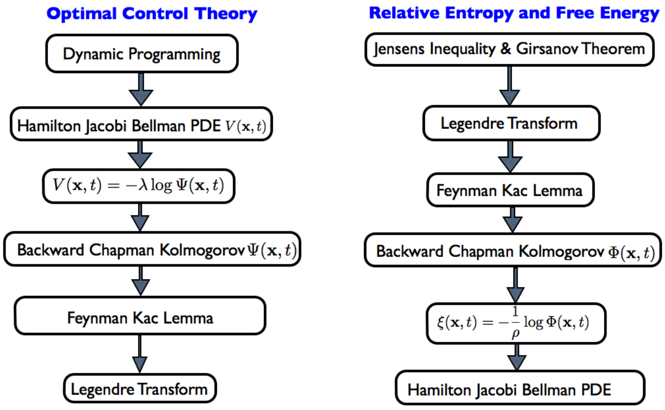

Optimal Control Theory
Dynamic Programming
Hamilton Jacobi Bellman PDE $V(\mathbf{x},t)$
$V(\mathbf{x},t) = -\lambda \log \Psi(\mathbf{x},t)$
Backward Chapman Kolmogorov $\Psi(\mathbf{x},t)$
Feynman Kac Lemma
Legendre Transform
Relative Entropy and Free Energy
Jensens Inequality & Girsanov Theorem
Legendre Transform
Feynman Kac Lemma
Backward Chapman Kolmogorov $\Phi(\mathbf{x},t)$
$\xi(\mathbf{x},t) = -\frac{1}{\rho} \log \Phi(\mathbf{x},t)$
Hamilton Jacobi Bellman PDE

Figure 1. An unified view of nonlinear stochastic optimal control theory based on dynamic programming and free energy-relative entropy information theoretic dualities.

The paper is organized as follows. In Section 2, we provide the definitions of free energy and relative entropy and derive their mathematical connection. In Section 3 we present the connection of the relative entropy and free energy relationship with the theory of stochastic control. In particular, in Section 3.1, we apply the relative entropy and free energy relationship to nonlinear stochastic dynamical systems affine in noise. In Section 3.2, the analysis on the application of the aforementioned relationship to nonlinear stochastic dynamics affine in controls and noise is presented together with connections to dynamic programming. In Section 4, we discuss thermodynamic interpretations and connections to the maximum entropy principle. In Section 5, we provide the derivation of the PI control as presented within the machine learning and statistical physics. In Section 6, the discrete time formulation is derived, and in Subsection 6.1, the connections to continuous time are shown. Finally in Section 7, we present algorithms, and in Section 8, we conclude with a discussion on the equivalencies and differences between the different views of stochastic optimal control.

2. Fundamental Relationship between Free Energy and Relative Entropy

In this section, we discuss the fundamental relationship between free energy and relative entropy [13]. This relationship is key for deriving the stochastic optimal control problem. Let $(\Omega, \mathcal{F})$ be a measurable space, where Ω denotes the sample space and $\mathcal{F}$ denotes a σ-algebra, and let $\boldsymbol{\mathcal{P}}(\Omega)$ define a probability measure on the σ-algebra $\mathcal{F}$. For the concepts that we shall propose, we need the following definitions.

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

$$\mathcal{E} = \log_e \int \exp(\rho \mathcal{J}(\mathbf{x})) \mathrm{d}\mathbb{P}, \tag{1}$$

is called the free energy (the function $\log_e$ denotes the natural logarithm) of $\mathcal{J}(\mathbf{x})$ with respect to $\mathbb{P}$ and $\rho \in \Re$.

Definition 2. [13]: Let $\mathbb{P} \in \boldsymbol{\mathcal{P}}(\Omega)$ and $\mathbb{Q} \in \boldsymbol{\mathcal{P}}(\Omega)$; then, the relative entropy of $\mathbb{P}$ with respect to $\mathbb{Q}$ is defined as:

$$\mathbb{KL}\left(\mathbb{Q}||\mathbb{P}\right) = \begin{cases} \int \log_e \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \mathrm{d}\mathbb{Q}, & \text{if } \mathbb{Q} << \mathbb{P} \text{ and } \log_e \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \in \mathcal{L}_1 \ , \\ +\infty, & \text{otherwise,} \end{cases}$$

where "$<<$" denotes the absolute continuity of $\mathbb{Q}$ with respect to $\mathbb{P}$ and $\mathcal{L}_1$ denotes the space of Lebesgue measurable functions on $[0, \infty)$. We say that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, and we write $\mathbb{Q} << \mathbb{P}$ if $\mathbb{P}(H) = 0 \Rightarrow \mathbb{Q}(H) = 0, \ \forall H \in \mathcal{F}$.

The free energy and relative entropy relationship is expressed by the theorem that follows.

Theorem 1. *Let $(\Omega, \mathcal{F})$ be a measurable space, where Ω denotes the sample space and $\mathcal{F}$ denotes a σ-algebra, and let $\boldsymbol{\mathcal{P}}(\Omega)$ define a probability measure on the σ-algebra $\mathcal{F}$. Consider $\mathbb{P}, \mathbb{Q} \in \boldsymbol{\mathcal{P}}(\Omega)$ and the definitions of free energy and relative entropy as expressed in Definitions (1) and (2). Under the assumption that $\mathbb{Q} << \mathbb{P}$, the following inequality holds:*

$$-\frac{1}{|\rho|} \log_e \mathbb{E}_{\mathbb{P}} \bigg[\exp(-|\rho| \mathcal{J}) \bigg] \leq \bigg[\mathbb{E}_{\mathbb{Q}} \left(\mathcal{J}\right) + |\rho|^{-1} \mathbb{KL}\left(\mathbb{Q}||\mathbb{P}\right) \bigg], \tag{2}$$

where $\mathbb{E}_{\mathbb{P}}, \mathbb{E}_{\mathbb{Q}}$ is the expectation under the probability measure $\mathbb{P}, \mathbb{Q}$, respectively, and $\rho \in \Re^-$ and $\mathcal{J} : \Re^M \rightarrow \Re$ and $M \in \mathcal{Z}^+$. The inequality in (2) is the so-called Legendre transform.

Proof. We express the expectation $\mathbb{E}_{\mathbb{P}}$ as a function of the expectation $\mathbb{E}_{\mathbb{Q}}$. In particular,

$$\mathbb{E}_{\mathbb{P}}\bigg[\exp(\rho\mathcal{J})\bigg] = \int \exp(\rho\mathcal{J})\mathrm{d}\mathbb{P} = \int \exp(\rho\mathcal{J})\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\mathrm{d}\mathbb{Q}. \tag{3}$$

Taking the logarithm of both sides of Equation (3) and using Jensen's inequality yields:

$$\log_e \mathbb{E}_{\mathbb{P}}\bigg[\exp\left(\rho\mathcal{J}\right)\bigg] = \log_e \int \exp\left(\rho\mathcal{J}\right)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\mathrm{d}\mathbb{Q} \geq \int \log_e\bigg(\exp\left(\rho\mathcal{J}\right)\frac{d\mathbb{P}}{d\mathbb{Q}}\bigg)\mathrm{d}\mathbb{Q} \tag{4}$$

The inequality (4) can be written as:

$$\log_e \mathbb{E}_{\mathbb{P}}\bigg[\exp(\rho\mathcal{J}(\mathbf{x},t))\bigg] \geq \int \bigg(\rho\mathcal{J} + \log\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\bigg)\mathrm{d}\mathbb{Q} = \int \rho\mathcal{J}\mathrm{d}\mathbb{Q} - \mathbb{KL}\left(\mathbb{Q}||\mathbb{P}\right). \tag{5}$$

Multiplying Equation (5) with $\frac{1}{\rho}$, where $\rho < 0$ or $\rho = -|\rho|$, it follows Equation (2) with $\mathbb{E}_{\mathbb{Q}}\left(\mathcal{J}\right) = \int \mathcal{J}\mathrm{d}\mathbb{Q}$. □

Inequality (2) gives a dual relationship between relative entropy and free energy, which leads to the minimization problem:

$$-\frac{1}{|\rho|}\log_e \mathbb{E}_{\mathbb{P}}\bigg[\exp(-|\rho|\mathcal{J})\bigg] = \inf_{\mathrm{d}\mathbb{Q}}\bigg[\mathbb{E}_{\mathbb{Q}}\left(\mathcal{J}\right) + |\rho|^{-1}\mathbb{KL}\left(\mathbb{Q}||\mathbb{P}\right)\bigg], \tag{6}$$

The infimum in Equation (6) attained at $\mathbb{Q}^*$ is given by:

$$\mathrm{d}\mathbb{Q}^* = \frac{\exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}}{\int \exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}}. \tag{7}$$

To verify that the infimum is attained by Equation (2), we have the following lemma.

Lemma 1. *Given the definitions of free energy and relative entropy and the assumption of absolutely continuous measures $\mathbb{Q} << \mathbb{P}$, the LHS of the Legendre transformation in Equation (2) is attained by the optimal measure in Equation (7).*

Proof. The proof is rather, simple and it is based on the substitution of Equation (7) into Equation (2). More precisely:

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}^*}\left[\mathcal{J}\right] + \frac{1}{|\rho|}\mathbb{KL}\left(\mathbb{Q}^*||\mathbb{P}\right) &= \mathbb{E}_{\mathbb{Q}^*}\left[\mathcal{J}\right] + \frac{1}{|\rho|}\int \log_e \frac{\mathrm{d}\mathbb{Q}^*}{d\mathbb{P}}\mathrm{d}\mathbb{Q}^* \\
&= \mathbb{E}_{\mathbb{Q}^*}\left[\mathcal{J}\right] + \frac{1}{|\rho|}\int \log_e \frac{\frac{\exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}}{\int \exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}}}{\mathrm{d}\mathbb{P}}\mathrm{d}\mathbb{Q}^* \\
&= \mathbb{E}_{\mathbb{Q}^*}\left[\mathcal{J}\right] + \frac{1}{|\rho|}\int \log_e \frac{\exp(-|\rho|\mathcal{J})}{\int \exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}}\mathrm{d}\mathbb{Q}^* \\
&= \mathbb{E}_{\mathbb{Q}^*}\left[\mathcal{J}\right] + \frac{1}{|\rho|}\int \bigg[-|\rho|\mathcal{J}(\mathbf{x}))\bigg]\mathrm{d}\mathbb{Q}^* \\
&- \frac{1}{|\rho|}\int \log_e \int \exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}\bigg]\mathrm{d}\mathbb{Q}^* \\
&= \frac{1}{|\rho|}\int \bigg[-\log_e \int \exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}\bigg]\mathrm{d}\mathbb{Q}^* \\
&= -\frac{1}{|\rho|}\log_e \int \exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}\int \mathrm{d}\mathbb{Q}^* \\
&= -\frac{1}{|\rho|}\log_e \int \exp(-|\rho|\mathcal{J})\mathrm{d}\mathbb{P}
\end{aligned}$$

□

In the case where $\rho > 0$, the inequality in (2) is flipped, and hence, the infimum in Equation (6) reverts to a supremum.

3. The Legendre Transformation and Stochastic Optimal Control

With the goal to use the Legendre transformation and to show its connection to optimal control, we define $\mathcal{J}$ as a state- and time-dependent cost function evaluated on trajectories starting at $\mathbf{x}(t) \in \Re^n$ at time t and with a time horizon $t_N \geq t$. More precisely, we have the mathematical form:

$$\mathcal{J} = \mathcal{J}(\mathbf{x}(\cdot), t) = \phi(\mathbf{x}(t_N), t_N) + \int_t^{t_N} q(\mathbf{x}, \tau)\mathrm{d}\tau. \tag{8}$$

where $\phi : \Re^n \times \Re \to \Re$ is a state-dependent terminal cost and $q : \Re^n \times \Re \to \Re$ is state- and time-dependent running cost. We also define the function $\xi : \Re^n \times \Re \to \Re$ as follows:

$$\xi(\mathbf{x}, t) = \frac{1}{\nu} \log_e \mathbb{E}_{\mathbb{P}}\Big[\exp(\nu \mathcal{J}(\mathbf{x}, t))\Big], \tag{9}$$

where $\nu \in \Re$. Depending on the sign of ν, the function $\xi(\mathbf{x}, t)$ has different interpretations. For small ν, Equation (9) is a function of the mean and the variance $\xi(\mathbf{x}, t) = \mathbb{E}_{\mathbb{P}}(\mathcal{J}(\mathbf{x}, t)) + \frac{\nu}{2}\mathbb{VAR}(\mathcal{J}(\mathbf{x}, t))$. For $\nu = |\rho|$, Equation (9) is risk sensitive, whereas for $\nu = -|\rho|$, it is risk seeking. For our analysis, $\nu = -|\rho|$. Next, we incorporate the state and time dependencies in the Legendre transformation in Equation (6), and we have:

$$\xi(\mathbf{x}, t) = \underbrace{-\frac{1}{|\rho|} \log_e \mathbb{E}_{\mathbb{P}}\overbrace{\Big[\exp(-|\rho|\mathcal{J}(\mathbf{x}, t))\Big]}^{\textbf{Desirability}}}_{\textbf{Helmholtz Free Energy}} = \inf_{\mathrm{d}\mathbb{Q}}\Big[\underbrace{\mathbb{E}_{\mathbb{Q}}(\mathcal{J}(\mathbf{x}, t))}_{\textbf{State Cost}} + \underbrace{|\rho|^{-1}\mathbb{KL}(\mathbb{Q}||\mathbb{P})}_{\textbf{Information Cost}}\Big]. \tag{10}$$

The RHS of the minimization problem in Equation (10) is the sum of a state-dependent cost and the relative entropy between the two measures $\mathbb{P}, \mathbb{Q}$; moreover, the minimization is w.r.t the probability measures $\mathbb{Q}$. We assign the probability measures $\mathbb{P}$ and $\mathbb{Q}$ to passive, in the sense of uncontrolled dynamics, and to controlled dynamics and consider the task of steering a dynamical system from an initial to a target state. The goal in Equation (10) is to find the optimal probability measure/control that steers the system from the initial to the terminal state by minimizing the state cost at the expense of the information cost. The information cost is an implicit measure of control effort, and its final formulation depends on the structure of the underlying dynamics. The LHS in Equation (10) corresponds to Helmholtz free energy, while there is also a term that corresponds to the concept of the desirability function. This concept was introduced in [4] and plays a key role in our derivations and analysis that follow. In the next two sections, we apply the Legendre transformation to stochastic systems and identify the cases where there is a direct relationship with dynamic programming and the LSOC.

3.1. Application to Nonlinear Stochastic Dynamics with Affine Stochastic Disturbances

In this section, we consider stochastic dynamics of the form:

$$\mathrm{d}\mathbf{x} = \mathbf{F}(\mathbf{x}, \mathbf{u})\mathrm{d}t + \mathbf{B}(\mathbf{x})\mathrm{d}\mathbf{w}^{(1)}, \quad \mathbf{x}(0) = \mathbf{x}_0, \quad t \geq 0, \tag{11}$$

where $\mathbf{x} \in \Re^n$ denotes the state of the system, $\mathbf{u} \in \Re^m$ denotes the control vector, $\mathbf{B}(\mathbf{x}) : \Re_n \to \Re^{n\times p}$ is the diffusion matrix function, $\mathbf{F}(\mathbf{x},\mathbf{u}) : \Re^n \times \Re^m \to \Re^n$ are the drift dynamics and $\mathrm{d}\mathbf{w} \in \Re^p$ is a Gaussian white noise disturbance. The diffusion matrix is partitioned as $\mathbf{B}(\mathbf{x}) = [0^{\mathrm{T}}_{(n-p)\times p},\ \ \mathbf{B}^{\mathrm{T}}_c(\mathbf{x})]^{\mathrm{T}}$, where $\mathbf{B}_c(\mathbf{x}) : \Re^n \to \Re^{p\times p}$ is invertible and $\mathbf{\Sigma}_{\mathbf{B}_c}(\mathbf{x}) = \mathbf{B}(\mathbf{x})\mathbf{B}^{\mathrm{T}}(\mathbf{x}) : \Re^n \to \Re^{p\times p}$. Similarly, the drift term is partitioned as $\mathbf{F}(\mathbf{x},\mathbf{u}) = [\mathbf{F}^{\mathrm{T}}_m(\mathbf{x},\mathbf{u}),\ \ \mathbf{F}^{\mathrm{T}}_c(\mathbf{x},\mathbf{u})]^{\mathrm{T}}$, where $\mathbf{F}_m(\mathbf{x},\mathbf{u}) : \Re^n \times \Re^m \to \Re^{(n-p)}$ and $\mathbf{F}_c(\mathbf{x},\mathbf{u}) : \Re^n \times \Re^m \to \Re^p$. Next, define the stochastic differential equation:

$$\mathrm{d}\mathbf{x} = \mathbf{A}(\mathbf{x})\mathrm{d}t + \mathbf{B}(\mathbf{x})\mathrm{d}\mathbf{w}^{(0)}, \quad \mathbf{x}(0) = \mathbf{x}_0, \quad t \geq 0, \tag{12}$$

where the drift term $\mathbf{A}(\mathbf{x}) : \Re^n \to \Re^n$ is defined as $\mathbf{A}(\mathbf{x}) \triangleq \mathbf{F}(\mathbf{x}, 0)$ and corresponds to the uncontrolled dynamics in Equation (11). Here, we denote the expectations evaluated on the system trajectories generated by the controlled dynamics and uncontrolled dynamics by $\mathbb{E}_{\mathbb{Q}}$ and $\mathbb{E}_{\mathbb{P}}$, respectively. In addition, $\Delta\mathbf{F}_m(\mathbf{x},\mathbf{u}) \triangleq \mathbf{F}_c(\mathbf{x},\mathbf{u}) - \mathbf{A}_c(\mathbf{x}) = \mathbf{F}_c(\mathbf{x},\mathbf{u}) - \mathbf{F}_c(\mathbf{x}, 0)$. The definition of the Radon–Nikodym [17] derivative for the stochastic differential Equations (11) and (12) has the form:

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp(\zeta(\mathbf{u}, t)), \tag{13}$$

where the term $\zeta(\mathbf{u}, t)$ is now defined as:

$$\zeta(\mathbf{u}, t) = \int_t^{t_N} \Delta\mathbf{F}^{\mathrm{T}}_c(\mathbf{x},\mathbf{u})\mathbf{B}_c(\mathbf{x})^{-1}(\mathbf{x})\mathrm{d}\mathbf{w}^{(1)}\mathrm{d}\tau + \int_t^{t_N} \frac{1}{2}\Delta\mathbf{F}^{\mathrm{T}}_c(\mathbf{x},\mathbf{u})\mathbf{\Sigma}^{-1}_{\mathbf{B}_c}(\mathbf{x})\Delta\mathbf{F}_c(\mathbf{x},\mathbf{u})\mathrm{d}\tau. \tag{14}$$

Now, substituting Equations (13) and (14) into Equation (2), we obtain:

$$\xi(\mathbf{x}, t) \leq \underbrace{\mathbb{E}_{\mathbb{Q}}\left[\mathcal{J}(\mathbf{x}, t)\right]}_{\textbf{State Cost}} + \underbrace{\mathbb{E}_{\mathbb{Q}}\left[\frac{1}{2|\rho|}\int_t^{t_N} \Delta\mathbf{F}^{\mathrm{T}}_c(\mathbf{x},\mathbf{u})\mathbf{\Sigma}^{-1}_{\mathbf{B}_c}(\mathbf{x})\Delta\mathbf{F}_c(\mathbf{x},\mathbf{u})\mathrm{d}\tau\right]}_{\textbf{Information Cost}}. \tag{15}$$

When the minimum is attained for $\mathbb{Q}^*$ given by Equation (7), we have:

$$\xi(\mathbf{x}, t) = \mathbb{E}_{\mathbb{Q}^*}\left[\phi(\mathbf{x}(t_N), t_N) + \int_t^{t_N} q(\mathbf{x}, \tau)\mathrm{d}\tau + \frac{1}{2|\rho|}\int_t^{t_N} \Delta\mathbf{F}^{\mathrm{T}}_c(\mathbf{x},\mathbf{u}^*)\mathbf{\Sigma}^{-1}_{\mathbf{B}_c}(\mathbf{x})\Delta\mathbf{F}_c(\mathbf{x},\mathbf{u}^*)\mathrm{d}\tau\right], \tag{16}$$

where the $\Delta\mathbf{F}_c(\mathbf{x},\mathbf{u}^*) \triangleq \mathbf{F}_c(\mathbf{x},\mathbf{u}^*) - \mathbf{F}_c(\mathbf{x}, 0)$ corresponds to the difference between the drift of the optimally-controlled (*i.e.*, $\mathbf{u} = \mathbf{u}^*$) and the drift of the uncontrolled (*i.e.*, $\mathbf{u} = 0$) dynamics. Equations (15) and (16) demonstrate how the structure of the dynamics appears in the information cost under minimization in the Legendre transformation. Therefore, there is a straight-forward relationship between the structure of the stochastic dynamics under consideration and the form of the control cost function under minimization. While this observation is not surprising when the Legendre transformation is used, it suggests ways to design control cost functions in stochastic optimal control theory based on the form of the stochastic dynamics. Another interesting observation is that the LHS of Equation (16) is the minimum attained under the optimal, in the sense of the Legendre transformation, probability measure $\mathbb{Q}^*$.

A question that arises here is related to the connection between the two forms of optimality, namely the optimality in the Legendre sense and the optimality in the dynamic programming sense. To further investigate this connection, we will leverage the Feynman–Kac Lemma in its more general form for the case of backward PDEs [8]. We also consider the stochastic dynamics in Equation (11) under the optimal control law $\mathbf{u}^*(\mathbf{x},t)$:

$$\mathrm{d}\mathbf{x} = \mathbf{F}(\mathbf{x},\mathbf{u}^*(\mathbf{x},t))\mathrm{d}t + \mathbf{B}(\mathbf{x})\mathrm{d}\mathbf{w}^{(1)}, \quad \mathbf{x}(0) = \mathbf{x}_0, \quad t \geq 0, \tag{17}$$

Under the assumptions of continuity and linear growth for $\mathbf{F}(\mathbf{x},\mathbf{u}^*(\mathbf{x},t)), \mathbf{B}(\mathbf{x})$ and the existence and uniqueness of the weak solution of Equation (17) (see pages 364 and 366 of [8]), we have the following theorem:

Theorem 2. *Feynman–Kac: Let $\Psi(\mathbf{x},t) : [0,t_N] \times \Re^n \to \Re$ be continuous and $\Psi(\mathbf{x},t) \in C^{1,2}$, and it satisfies the Cauchy problem:*

$$-\partial_t\Psi = -\frac{1}{\lambda}\ell\Psi + \mathbf{F}^T(\nabla_{\mathbf{x}}\Psi) + \frac{1}{2}\mathrm{tr}\left((\nabla_{\mathbf{xx}}\Psi)\mathbf{B}\mathbf{B}^T\right) + \mathcal{L}(\mathbf{x},t) \tag{18}$$

in $[0,t_N) \times \Re^{n\times 1}$ with the boundary condition:

$$\Psi(\mathbf{x},t_N) = \beta(\mathbf{x}) \tag{19}$$

then $\Psi(\mathbf{x},t)$ admits the stochastic representation:

$$\Psi(\mathbf{x},t) = \mathbb{E}\left[\beta(\mathbf{x}_{t_N})\exp\left(-\frac{1}{\lambda}\int_t^{t_N}\ell(\mathbf{x}_s,\tau)\mathrm{d}\tau\right) + \int_t^{t_N}\mathcal{L}(\mathbf{x},t)\exp\left(-\frac{1}{\lambda}\int_t^{s}\ell(\mathbf{x},\tau)\mathrm{d}\tau\right)\mathrm{d}s\right] \tag{20}$$

$\forall t_0 \in [0,t_N]$. *In particular, such a solution is unique. The expectation above is taken with respect to sampled trajectories generated using (17).*

Given the form of the Feynman–Kac lemma and the expectation in Equation (16), we set the terms $\mathcal{L}(\mathbf{x},t)$, ℓ and $\beta(\mathbf{x}_{t_N})$ in Equation (20) as follows:

$$\mathcal{L}(\mathbf{x},t) = q(\mathbf{x}) + \frac{1}{2|\rho|}\Delta\mathbf{F}_c^{\mathrm{T}}(\mathbf{x},\mathbf{u}^*)\boldsymbol{\Sigma}_{\mathbf{B}_c}^{-1}(\mathbf{x})\Delta\mathbf{F}_c(\mathbf{x},\mathbf{u}^*)\mathrm{d}\tau, \quad \ell(\mathbf{x},t) = 0, \quad \beta(\mathbf{x}_{t_N}) = \phi(\mathbf{x}(t_N),t_N) \tag{21}$$

Based on the Feynman–Kac lemma, the free energy term $\boldsymbol{\xi}(\mathbf{x},t)$ can be interpreted as the unique solution of the backward PDE:

$$\begin{aligned} -\frac{\partial\xi(\mathbf{x},t)}{\partial t} = \underbrace{q(\mathbf{x},t)}_{\textbf{State Cost}} + \underbrace{\frac{1}{2|\rho|}\Delta\mathbf{F}_c^{\mathrm{T}}(\mathbf{x},\mathbf{u}^*)\boldsymbol{\Sigma}_{\mathbf{B}_c}^{-1}(\mathbf{x})\Delta\mathbf{F}_c(\mathbf{x},\mathbf{u}^*)}_{\textbf{Optimal Control Cost}} + \xi_{\mathbf{x}}^{\mathrm{T}}(\mathbf{x},t)\mathbf{F}(\mathbf{x},\mathbf{u}^*) \\ + \frac{1}{2}\mathrm{tr}\left(\xi_{\mathbf{xx}}(\mathbf{x},t)\mathbf{B}(\mathbf{x})\mathbf{B}^{\mathrm{T}}(\mathbf{x})\right) \end{aligned} \tag{22}$$

with the boundary condition $\xi(\mathbf{x}(t_N),t_N) = \phi(\mathbf{x}(t_N),t_N)$. The interesting observation here is that the PDE in Equation (22) is the optimal HJB PDE for a stochastic optimal control problem with state cost $q(\mathbf{x},t)$ and control cost term $\frac{1}{2|\rho|}\Delta\mathbf{F}_c^{\mathrm{T}}(\mathbf{x},\mathbf{u}^*)\boldsymbol{\Sigma}_{\mathbf{B}_c}^{-1}(\mathbf{x})\Delta\mathbf{F}_c(\mathbf{x},\mathbf{u}^*)$ subject to the dynamics in Equation (11). It is clear therefore that there is a fundamental connection between the Legendre transformation and dynamic programing for the general class of stochastic systems that are affine only in the stochastic disturbances and nonlinear in controls and states. This observation generalizes our previous work on identifying the connections between the relative entropy and free energy dualities and the PI and KL controls [11]. Essentially, the two methodologies result in the same HJB PDE when the state and the control cost function are defined as:

$$\text{State Cost} = q(\mathbf{x}, t), \quad \text{Control Cost} = \frac{1}{2|\rho|}\Delta\mathbf{F}_c^{\mathrm{T}}(\mathbf{x},\mathbf{u})\boldsymbol{\Sigma}_{\mathbf{B}_c}^{-1}(\mathbf{x})\Delta\mathbf{F}_c(\mathbf{x},\mathbf{u}) \tag{23}$$

The implications of this finding can be summarized as follows

(i) The Helmholtz free energy satisfies the HJB PDE for the case of systems that are non-affine in controls and affine in stochastic disturbances. This observation has direct consequences to the development of algorithms that can compute the value function for a stochastic optimal control problem with forward sampling of SDEs. While this connection was known within the LSOC framework for dynamics affine in controls and noise, it is the first time that this connection has been derived for general classes of stochastic systems with dynamics nonlinear in state and control and affine only in noise.

(ii) The optimal measure $\mathrm{d}\mathbb{Q}^*$ for the stochastic control problem with state and control cost as specified in Equation (23) is given by Equation (7). Note that this is the probability measure that corresponds to trajectories generated under the optimal control policy $\mathbf{u}^*(\mathbf{x}, t)$. A fundamental question at this point is related to how this optimal control can be numerically computed, such that $\mathrm{d}\mathbb{Q} = \mathrm{d}\mathbb{Q}^*$. The difficulty arises from the fact that for the case of dynamic systems that are nonlinear in controls and cost functions that are non quadratic, there is no explicit form for the optimal control policy $\mathbf{u}^*(\mathbf{x}, t)$. This difficulty could be addressed by an *a priori* specification of the structure of the optimal control policy $\mathbf{u}(\mathbf{x}, t)$ and then optimization of this structure, such that for any state $\mathbf{x}$ and time t, the optimal probability measure $\mathrm{d}\mathbb{Q}^* = \mathrm{d}\mathbb{Q}^*(\mathbf{x}, t)$ is reached.

Next, we discuss the connection between the free energy-relative entropy dualities and stochastic optimal control of systems with dynamics affine in control and noise. Again, the Feynman–Kac lemma plays a key role, but as will be shown, the way that it is applied differs from our analysis in this section, since it is directly applied to the desirability function.

3.2. Application to Nonlinear Stochastic Dynamics with Affine Controls and Disturbances

In this section, we apply the Legendre transformation for probability measures that correspond to stochastic dynamics affine in control and stochastic disturbances. In particular, we consider the stochastic dynamics [13,18]:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x})\mathrm{d}t + \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}(\mathbf{x})\mathrm{d}\mathbf{w}^{(0)}, \mathbf{x}(0) = \mathbf{x}_0, t \geq 0, \tag{24}$$

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x})\mathrm{d}t + \boldsymbol{\mathcal{B}}(\mathbf{x})\left(\mathbf{u}\mathrm{d}t + \frac{1}{\sqrt{|\rho|}}\mathrm{d}\mathbf{w}^{(1)}\right), \mathbf{x}(0) = \mathbf{x}_0, \quad t \geq 0 \tag{25}$$

where $\mathbf{x} \in \Re^n$ denotes the state of the system, $\boldsymbol{\mathcal{B}}(\mathbf{x}) : \Re \rightarrow \Re^{n\times p}$ is the control and diffusion matrix function partitioned as $\boldsymbol{\mathcal{B}}(\mathbf{x}) = [0^{\mathrm{T}}_{(n-p)\times p}, \ \ \boldsymbol{\mathcal{B}}_c^{\mathrm{T}}(\mathbf{x})]^{\mathrm{T}}$, where $\boldsymbol{\mathcal{B}}_c(\mathbf{x}) : \Re^n \rightarrow \Re^{p\times p}$ is invertible, $\mathbf{f}(\mathbf{x}) : \Re^n \rightarrow \Re^n$ denotes the passive dynamics, $\mathbf{u} \in \Re^p$ is the control vector and $\mathrm{d}\mathbf{w} \in \Re^p$ is a Gaussian white noise disturbance. Note that the difference between the two diffusion terms in Equations (24) and (25) is the fact that the control appears in Equation (25). This control, together with the passive dynamics, defines a new drift term. Expectations evaluated on the system trajectories generated by the uncontrolled and controlled dynamics are represented by $\mathbb{E}_{\mathbb{P}}$ and $\mathbb{E}_{\mathbb{Q}}$, respectively. The corresponding probability measures of the aforementioned expectations are $\mathbb{P}$ and $\mathbb{Q}$. Next, we use Equation (10) and the Radon–Nikodym derivative given by Equations (13) and (14), which now takes the form $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(\zeta(\mathbf{u}, t)\right)$ [8], where the term $\zeta(\mathbf{u}, t)$ is given by:

$$\zeta(\mathbf{u}, t) = \frac{1}{2}|\rho| \int_t^{t_N} \mathbf{u}^{\mathrm{T}}\mathbf{u}\mathrm{d}\tau + \sqrt{|\rho|} \int_t^{t_N} \mathbf{u}^{\mathrm{T}}\mathrm{d}\mathbf{w}^{(1)}, \tag{26}$$

Substituting $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}$ into inequality (10) gives:

$$\xi(\mathbf{x}, t) = -\frac{1}{|\rho|}\log_e \mathbb{E}_{\mathbb{P}}\Big[\exp\left(-|\rho|\mathcal{J}(\mathbf{x}, t)\right)\Big] \leq \mathbb{E}_{\mathbb{Q}}\Big[\mathcal{J}(\mathbf{x}, t) + \frac{1}{|\rho|}\zeta(\mathbf{u}, t)\Big] \tag{27}$$

Substitution of $\zeta(\mathbf{u})$ in the last equation results in:

$$\xi(\mathbf{x}, t) = -\frac{1}{|\rho|}\log_e \mathbb{E}_{\mathbb{P}}\Big[\exp\left(-|\rho|\mathcal{J}(\mathbf{x}, t)\right)\Big] \leq \mathbb{E}_{\mathbb{Q}}\Big[\mathcal{J}(\mathbf{x}, t) + \frac{1}{2}\int_t^{t_N} \mathbf{u}^{\mathrm{T}}\mathbf{u}\mathrm{d}\tau\Big]. \tag{28}$$

The last term in Inequality (28) corresponds to the cost function of a stochastic optimal control problem and is bounded from below by the free energy. In addition to providing a lower-bound on the objective function for the stochastic optimal control problem, Inequality (28) provides an explicit construction on how this lower bound can be computed. This computation involves forward sampling of the uncontrolled dynamics, evaluation of the expectation of the exponentiated state-dependent part, $\phi(\mathbf{x}(t_N))$ and $q(\mathbf{x}(t))$, and the logarithmic transformation of this expectation. Note that Inequality (28) is derived without relying on any principle of optimality and involves the application of Girsanov's theorem between controlled and uncontrolled stochastic dynamics, as well as the use of the dual relationship between the free energy and the relative entropy needed to compute the lower bound in (28). Inequality (28) defines a minimization problem where the RHS of the inequality is minimized with respect $\zeta(\mathbf{u}, t)$ and, hence, with respect to the control $\mathbf{u}$. At the minimum $\mathbf{u} = \mathbf{u}^*$, the right part of the inequality in (28) attains its optimal $\xi(\mathbf{x}, t)$. Under the optimal control policy $\mathbf{u}^*$, the optimal distribution takes the from:

$$\mathrm{d}\mathbb{Q}^*(\mathbf{x}, t) = \frac{\exp\left[-|\rho|\left(\phi(\mathbf{x}(t_N)) + \int_t^{t_N} q(\mathbf{x}, \tau)\mathrm{d}\tau\right)\right]\mathrm{d}\mathbb{P}}{\int \exp(-|\rho|\left(\phi(\mathbf{x}(t_N)) + \int_t^{t_N} q(\mathbf{x}, \tau)\mathrm{d}\tau\right)\Big]\mathrm{d}\mathbb{P}}. \tag{29}$$

An important question that arises is: What is the link between (28) and the principle of optimality in dynamic programming? To address this question, one needs to show that $\xi(\mathbf{x}, t)$ satisfies the HJB equation, and hence, $\xi(\mathbf{x}, t)$ is the corresponding value function [18]. More precisely, we introduce a new variable $\Phi(\mathbf{x}, t)$ defined as $\Phi(\mathbf{x}, t) \triangleq \mathbb{E}_{\mathbb{P}}(\exp\left(\rho\mathcal{J}(\mathbf{x}, t)\right))$ and apply the Feynman–Kac lemma [7] to arrive at the backward Chapman–Kolmogorov PDE:

$$-\partial_t\Phi(\mathbf{x}, t) = -|\rho|q(\mathbf{x}, t)\Phi(\mathbf{x}, t) + \mathbf{f}^{\mathrm{T}}(\mathbf{x})\Phi_{\mathbf{x}}(\mathbf{x}, t) + \frac{1}{2|\rho|}\mathrm{tr}\left(\Phi_{\mathbf{xx}}(\mathbf{x}, t)\boldsymbol{\mathcal{B}}(\mathbf{x})\boldsymbol{\mathcal{B}}^{\mathrm{T}}(\mathbf{x})\right). \tag{30}$$

Since $\xi(\mathbf{x},t) = \frac{1}{\rho}\log\Phi(\mathbf{x},t) = -\frac{1}{|\rho|}\log\Phi(\mathbf{x},t)$, it follows that $\partial_t\Phi(\mathbf{x},t) = -|\rho|\Phi(\mathbf{x},t)\partial_t\xi(\mathbf{x},t)$, $\Phi_{\mathbf{x}}(\mathbf{x},t) = -|\rho|\Phi(\mathbf{x},t)\xi_{\mathbf{x}}$ and $\Phi_{\mathbf{xx}}(\mathbf{x},t) = |\rho|\Phi(\mathbf{x},t)\xi_{\mathbf{xx}}(\mathbf{x},t) - |\rho|^2\Phi(\mathbf{x},t)\xi_{\mathbf{x}}(\mathbf{x},t)\xi_{\mathbf{x}}^{\mathrm{T}}(\mathbf{x},t)$. In this case, it can be shown that $\xi(\mathbf{x},t)$ satisfies the nonlinear PDE:

$$-\partial_t\xi(\mathbf{x},t) = q(\mathbf{x},t) + \xi_{\mathbf{x}}^{\mathrm{T}}(\mathbf{x},t)\mathbf{f}(\mathbf{x}) - \frac{1}{2}\xi_{\mathbf{x}}^{\mathrm{T}}(\mathbf{x},t)\boldsymbol{\mathcal{B}}(\mathbf{x})\boldsymbol{\mathcal{B}}^{\mathrm{T}}(\mathbf{x})\xi_{\mathbf{x}}(\mathbf{x},t) + \frac{1}{2|\rho|}\mathrm{tr}\left(\xi_{\mathbf{xx}}(\mathbf{x},t)\boldsymbol{\mathcal{B}}(\mathbf{x})\boldsymbol{\mathcal{B}}^{\mathrm{T}}(\mathbf{x})\right). \quad (31)$$

The nonlinear PDE in Equation (31) corresponds to the HJB equation [19] for the case of the minimizing optimal control problem, and hence, $\xi(\mathbf{x},t)$ is the corresponding minimizing value function. It is important to note that the principle of optimality was not used to derive Equation (31). Furthermore, while the mathematical analysis results in the HJB PDE, it does not explicitly provide the form of the optimal control policy. This means that to derive Equation (31), it is not required to have an expression for the optimal control policy. This observation is in stark contrast with the classical treatment of stochastic optimal control theory, based on dynamic programming, where first the optimal control is specified and then the final form of the HJB Equation (31) is derived.

The optimal control policy associated with Equation (31) is expressed as:

$$\mathbf{u}(\mathbf{x},t) = -\boldsymbol{\mathcal{B}}^{\mathrm{T}}(\mathbf{x})\xi_{\mathbf{x}}(\mathbf{x},t). \quad (32)$$

To recover the optimal control policy Equation (32), one needs to be aware of the optimal control derivation that is based on dynamic programming.

4. Thermodynamic Interpretations and Connections to the Maximum Entropy Principle

In this section, we discuss thermodynamic interpretations of nonlinear stochastic optimal control theory using the relative entropy-free energy relationship. More precisely, we consider the Baroh–Jaunch entropy or generalized Boltzmann–Gibbs–Shannon entropy [16] defined as:

$$\mathcal{S}\left(\mathbb{Q}||\mathbb{P}\right) \triangleq -\mathbb{KL}\left(\mathbb{Q}||\mathbb{P}\right) = -\int \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\log_e\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}\mathrm{d}\mathbb{P}, \quad (33)$$

then Equation (2) takes the form:

$$\underbrace{-\frac{1}{|\rho|}\log_e\mathbb{E}_{\mathbb{P}}\left[\exp(-|\rho|\mathcal{J}(\mathbf{x},t))\right]}_{\textbf{F: Helmholtz Free Energy}} = \inf_{\mathrm{d}\mathbb{Q}}\left[\underbrace{\mathbb{E}_{\mathbb{Q}}\left(\mathcal{J}(\mathbf{x},t)\right)}_{\textbf{U: State Cost}} - \underbrace{T\mathcal{S}\left(\mathbb{Q}||\mathbb{P}\right)}_{\textbf{S: Generalized Entropy}}\right]. \quad (34)$$

At $\mathbb{Q} = \mathbb{Q}^*$, we have that:

$$\underbrace{-\frac{1}{|\rho|}\log_e\mathbb{E}_{\mathbb{P}}\left[\exp(-|\rho|\mathcal{J}(\mathbf{x},t))\right]}_{\textbf{F: Helmholtz Free Energy}} = \underbrace{\mathbb{E}_{\mathbb{Q}^*}\left(\mathcal{J}(\mathbf{x},t)\right)}_{\textbf{U: State Cost}} - \underbrace{T\mathcal{S}\left(\mathbb{Q}^*||\mathbb{P}\right)}_{\textbf{S: Generalized Entropy}}, \quad (35)$$

The last equation has the form $F = U - T\mathcal{S}$, where F is the free energy, $T = |\rho|^{-1}$ is the temperature and $\mathcal{S}$ is the generalized Boltzmann–Gibbs–Shannon entropy. Note that Baroh–Jaunch entropy is a concave function, and it is a generalized form of entropy, since it incorporates the Boltzmann, Gibbs and Shannon entropy [16]. In addition, it is negative, and its maximum is reached for $\mathbb{P} = \mathbb{Q}$. Minimization of the $\mathbb{KL}(\mathbb{P}||\mathbb{Q})$ is equivalent to maximization of the generalized Boltzmann–Gibbs–Shannon entropy

$\mathcal{S}(\mathbb{P}||\mathbb{Q})$. In the absence of the state cost, the optimal measure is the one that maximizes the Boltzmann–Gibbs–Shannon entropy, and therefore, $\mathbb{P} = \mathbb{Q}$. However, as it is shown next, in the presence of the state-dependent cost constraint, the optimal measure $\mathbb{Q}^*$ should be "far" from the baseline probability measure $\mathbb{P}$. When the probability measures $\mathbb{Q}$ and $\mathbb{P}$ are assigned to state distributions of controlled and uncontrolled stochastic dynamical systems, the Kullback–Leibler divergence between $\mathbb{Q}$ and $\mathbb{P}$ is an implicit measure of control effort. In this sense, minimization of the Kullback–Leibler divergence or maximization of the generalized Boltzmann–Gibbs–Shannon entropy is equivalent to minimization of the control effort.

Another interesting connection with thermodynamics emerges from the fact that the optimal policy can be derived using the maximum entropy principle. The form of entropy under maximization is the generalized Boltzmann–Gibbs–Shannon entropy. To makes things concrete, lets consider the following maximum entropy constrained optimization problem specified as follows:

$$\begin{aligned} &\max_{\mathrm{d}\mathbb{Q}^*} \mathcal{S}\left(\mathbb{Q}||\mathbb{P}\right) \\ &\text{Subject to: } \mathbb{E}_{\mathbb{Q}}[\mathcal{J}(\mathbf{x},t)] = c \text{ and } \int \mathrm{d}\mathbb{Q} = 1. \end{aligned} \tag{36}$$

where c is positive constant. To find the solution, we form the augmented objective function by incorporating the constraints with proper Lagrange multipliers:

$$\begin{aligned} \mathcal{L}(\mathbb{Q},\lambda,\mu,c) &= \mathcal{S}\left(\mathbb{Q}||\mathbb{P}\right) + \lambda\Big(c - \mathbb{E}_{\mathbb{Q}}[\mathcal{J}(\mathbf{x},t)]\Big) + \mu\Big(1 - \int \mathrm{d}\mathbb{Q}\Big) \\ &= -\int \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \log_e \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} \mathrm{d}\mathbb{P} + \lambda\Big(c - \int \mathcal{J}(\mathbf{x},t)\mathrm{d}\mathbb{Q}\Big) + \mu\Big(1 - \int \mathrm{d}\mathbb{Q}\Big) \\ &= -\int \Big(\log_e \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} + \lambda\mathcal{J}(\mathbf{x},t) + \mu\Big)\mathrm{d}\mathbb{Q} + \lambda c + \mu. \end{aligned} \tag{37}$$

Next, we define the term:

$$\mathbb{L} = \int \underbrace{\Big(\log_e \frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} + \lambda\mathcal{J}(\mathbf{x},t) + \mu\Big)}_{f} \mathrm{d}\mathbb{Q} \tag{38}$$

Given the assumptions that f is $\mathbb{Q}$-integrable and the signed measure $\mathbb{L}$ is absolute continuous with respect to $\mathbb{Q}$, ($\mathbb{L} << \mathbb{Q}$), the signed measure $\mathbb{L}$ is finite (see the general form of Radon–Nikodym Theorem 5.5.3 in [20]). Later, it will be shown that $\mathbb{L}$ is a measure (positive-signed measure), but for now, we consider the more general case of the signed measure. Under the aforementioned assumptions, the Radon–Nikodym derivative $\frac{\mathrm{d}\mathcal{L}(\mathbb{Q})}{\mathrm{d}\mathbb{Q}}$ is a well-defined operation. To find the optimal measure $\mathbb{Q}$, we apply the Radon–Nikodym derivative in Equation (38) and set it to zero. In mathematical terms, this operation results in:

$$\log_e \frac{\mathrm{d}\mathbb{Q}^*}{\mathrm{d}\mathbb{P}} + \lambda\mathcal{J}(\mathbf{x},t) + \mu = 0 \Rightarrow \mathrm{d}\mathbb{Q}^* = \exp(-\lambda\mathcal{J}(\mathbf{x},t) - \mu)\mathrm{d}\mathbb{P}.$$

Integration of the optimal measure $\mathrm{d}\mathbb{Q}^*$ to one gives an expression for μ:

$$\mu = \log_e \int \exp(-\lambda\mathcal{J}(\mathbf{x},t))\mathrm{d}\mathbb{P}. \tag{39}$$

Substitution of μ back in Equation (39) gives the optimal probability measure in Equation (7). There are few interesting observations:

(i) Substitution of the optimal measure $\mathrm{d}\mathbb{Q}^*$ in Lagrangian (37) results in:

$$\mathcal{L}(\mathbb{Q}^*, \lambda, \mu, c) = \lambda c + \log_e \int \exp(-\lambda \mathcal{J}(\mathbf{x}, t))\mathrm{d}\mathbb{P}.$$

Moreover, given a certain performance level c, the Lagrange multiplier λ can be found by using the equation:

$$\int \mathcal{J}(\mathbf{x}, t)\mathrm{d}\mathbb{Q}^* - c = 0 \Rightarrow \int \Big(\mathcal{J}(\mathbf{x}, t) - c\Big) \exp(-\lambda \mathcal{J}(\mathbf{x}, t)\mathrm{d}\mathbb{P} = 0. \tag{40}$$

(ii) The term $-\frac{1}{\lambda}\mu = -\frac{1}{\lambda}\mu(\mathbf{x}, t)$ corresponds to the Helmholtz free energy, since:

$$-\frac{1}{\lambda}\mu(\mathbf{x}, t) = -\frac{1}{\lambda} \log_e \int \exp(-\lambda \mathcal{J}(\mathbf{x}, t))\mathrm{d}\mathbb{P}. \tag{41}$$

Therefore, for the case stochastic dynamics affine in control and noise, the term $-\frac{1}{\lambda}\mu$ is a value function and satisfies the HJB equation.

(iii) Initially, we considered $\mathbb{L}$ as a singed measure. However, given the optimal measure $\mathbb{Q}^*$ and the form of the Lagrange multiplier μ, the signed measure $\mathbb{L}$ is positive, and therefore, it is a measure. To show this, one can use the Legendre transformation between free energy and relative entropy.

The thermodynamic equilibrium of maximum entropy corresponds to maximization of the generalized Boltzmann–Gibbs–Shannon entropy that is equivalent to minimization of control effort subject to the performance and normalization constrains as expressed in the optimization problem in Equation (36). Moreover, the equilibrium measure is the optimal measure as specified in Equation (7). For the case of stochastic dynamics affine in controls and noise, this measure corresponds to trajectories sampled from the stochastic dynamics under the optimal, in the sense of dynamic programming, control policy.

5. Bellman Principle of Optimality

In this section, we consider the classical stochastic optimal control problem as a constrained optimization problem and derive the LSOC framework in continuous time. The analysis in this section is more known under the name of PI control, and it has been presented mostly in the machine learning and statistical physics communities [2,21].

Here, we present a generalized version, which was also derived in [11], which allows terms in the cost function to be both state and control dependent. This formulation is important for our later discussion on the generalizability of every approach (information theoretic, path integral, KL control) in the LSOC framework. In particular, we start with the cost functional:

$$V(\mathbf{x}(t), t) = \min_{\mathbf{u}} J(\mathbf{x}, \mathbf{u}) = \min_{\mathbf{u}} \mathbb{E}_{\mathbb{Q}} \left[\phi(\mathbf{x}, t_N) + \int_t^{t_N} \mathcal{L}(\mathbf{x}, \mathbf{u}, \tau)\mathrm{d}\tau\right]. \tag{42}$$

The expectation $\mathbb{E}_{\mathbb{Q}}$ in Equation (42) is evaluated on system trajectories generated by forward sampling of the controlled diffusion process Equation (11). We assume that the function $\mathbf{F}(\mathbf{x}, \mathbf{u})$ is a nonlinear function of the state $\mathbf{x} \in \Re^n$ and affine in the control $\mathbf{u} \in \Re^m$, and hence,

$\mathbf{F}(\mathbf{x},\mathbf{u}) = \mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}$. The matrix function $\mathbf{G}(\mathbf{x}) : \Re^n \to \Re^{n\times m}$ is the control transition matrix, and $\mathbf{f}(\mathbf{x}) : \Re^n \to \Re^n$ denotes the passive dynamics. Under the optimal control $\mathbf{u} = \mathbf{u}^*$, the cost function $J(\mathbf{x},\mathbf{u})$ is equal to the value function $V(\mathbf{x},t)$. Next, let $\mathcal{L}(\mathbf{x},\mathbf{u},t)$ denote the running cost defined as $\mathcal{L}(\mathbf{x},\mathbf{u},t) \triangleq q_0(\mathbf{x},t) + q_1^{\mathrm{T}}(\mathbf{x},t)\mathbf{u} + \frac{1}{2}\mathbf{u}^T\mathbf{R}\mathbf{u}$, where $q_0(\mathbf{x},t)$ is a nonlinear, nonquadratic state-dependent cost, $q_1^{\mathrm{T}}(\mathbf{x},t)\mathbf{u}$ is a cross-term depending on the state and control and $\frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{R}\mathbf{u}$ is a quadratic control cost with $\mathbf{R} > 0$. The stochastic HJB equation [15,19] associated with this stochastic optimal control problem can be expressed as follows:

$$-\partial_t V(\mathbf{x},t) = \min_{\mathbf{u}}(\mathcal{L}(\mathbf{x},\mathbf{u},t) + V_{\mathbf{x}}^{\mathrm{T}}(\mathbf{x},t)\mathbf{F}(\mathbf{x},\mathbf{u}) + \frac{1}{2}\mathrm{tr}\left(V_{\mathbf{xx}}(\mathbf{x},t)\mathbf{B}(\mathbf{x})\mathbf{B}^{\mathrm{T}}(\mathbf{x})\right)). \tag{43}$$

The corresponding optimal control is given by:

$$\mathbf{u}(\mathbf{x},t) = -\mathbf{R}^{-1}\left(q_1(\mathbf{x},t) + \mathbf{G}^{\mathrm{T}}(\mathbf{x})V_{\mathbf{x}}(\mathbf{x},t)\right). \tag{44}$$

The optimal control drives the system dynamics in the direction opposite that of the gradient of the value function $V_{\mathbf{x}}(\mathbf{x},t)$. Furthermore, the value function satisfies the nonlinear, second-order PDE:

$$\begin{aligned}-\partial_t V(\mathbf{x},t) = \tilde{q}(\mathbf{x},t) + V_{\mathbf{x}}^{\mathrm{T}}(\mathbf{x},t)\tilde{\mathbf{f}}(\mathbf{x},t) - \frac{1}{2}V_{\mathbf{x}}^{\mathrm{T}}(\mathbf{x},t)\mathbf{G}(\mathbf{x})\mathbf{R}^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{x})V_{\mathbf{x}}(\mathbf{x},t)\\ + \frac{1}{2}\mathrm{tr}\left(V_{\mathbf{xx}}(\mathbf{x},t)\mathbf{B}(\mathbf{x})\mathbf{B}^{\mathrm{T}}(\mathbf{x})\right),\end{aligned} \tag{45}$$

where $\tilde{q}(\mathbf{x},t) \triangleq q_0(\mathbf{x},t) - \frac{1}{2}q_1(\mathbf{x},t)^{\mathrm{T}}\mathbf{R}^{-1}q_1(\mathbf{x},t)$ and $\tilde{\mathbf{f}}(\mathbf{x},t) \triangleq \mathbf{f}(\mathbf{x}) - \mathbf{G}(\mathbf{x})\mathbf{R}^{-1}q_1(\mathbf{x},t)$, and the boundary condition is $V(\mathbf{x}(t_N),t_N) = \phi(\mathbf{x}(t_N),t_N)$. Given the exponential transformation and the relationship between control authority and noise:

$$\begin{aligned}V(\mathbf{x},t) = -\lambda\log\psi(\mathbf{x},t),\\ \lambda\mathbf{G}(\mathbf{x})\mathbf{R}^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{x}) = \mathbf{B}(\mathbf{x})\mathbf{B}^{\mathrm{T}}(\mathbf{x}) = \mathbf{\Sigma}(\mathbf{x}),\end{aligned} \tag{46}$$

the PDE in Equation (45) yields:

$$-\partial_t\psi(\mathbf{x},t) = -\frac{1}{\lambda}\tilde{q}(\mathbf{x},t)\psi(\mathbf{x},t) + \tilde{\mathbf{f}}^{\mathrm{T}}(\mathbf{x})\psi_{\mathbf{x}}(\mathbf{x},t) + \frac{1}{2}\mathrm{tr}\left(\psi_{\mathbf{xx}}(\mathbf{x},t)\mathbf{\Sigma}(\mathbf{x})\right), \tag{47}$$

with boundary condition $\psi(\mathbf{x},t_N) = \exp\left(-\frac{1}{\lambda}\phi(\mathbf{x},t_N)\right)$. Now, applying the Feynman–Kac lemma to the Chapman–Kolmogorov PDE Equation (47) yields a solution in the form of an expectation over system trajectories; namely:

$$\psi\left(\mathbf{x}(t),t\right) = \mathbb{E}_{\tilde{\mathbb{P}}}\left[\exp\left(-\int_t^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x},\tau)\mathrm{d}\tau\right)\psi(\mathbf{x}(t_N),t_N)\right]. \tag{48}$$

The expectation $\mathbb{E}_{\tilde{\mathbb{P}}}$ in Equation (48) is taken on sample paths generated with the forward sampling of the uncontrolled diffusion equation $\mathrm{d}\mathbf{x} = \tilde{f}(\mathbf{x})\mathrm{d}t + \mathbf{B}(\mathbf{x})\mathrm{d}\mathbf{w}$, and the optimal control is given by:

$$\mathbf{u}(\mathbf{x}(t),t) = -\mathbf{R}^{-1}\left(q_1(\mathbf{x},t) - \lambda\mathbf{G}^{\mathrm{T}}(\mathbf{x})\frac{\psi_{\mathbf{x}}(\mathbf{x},t)}{\psi(\mathbf{x},t)}\right). \tag{49}$$

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, it can be shown that:

$$V(\mathbf{x},t) = \underbrace{-\lambda\log_e\mathbb{E}_{\mathbb{P}}\left[\exp\left(-\int_t^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x},t)\mathrm{d}\tau\right)\Psi(\mathbf{x}(t_N),t_N)\right]}_{\textbf{Helmholtz Free Energy}} \leq \underbrace{\mathbb{E}_{\mathbb{Q}}\left[J(\mathbf{x},\mathbf{u})\right]}_{\textbf{Total Cost}}. \tag{50}$$

Note that Equation (50) is a form of the Legendre transformation, and in fact, it is identical to Equation (28) for the case where $q_1(\mathbf{x}, t) = 0$, $\mathbf{R} = I$, $\lambda = \frac{1}{|\rho|}$, $\mathbf{G}(\mathbf{x}) = \boldsymbol{\mathcal{B}}(\mathbf{x})$ and $\mathbf{B}(\mathbf{x}) = \frac{1}{\sqrt{|\rho|}}\boldsymbol{\mathcal{B}}(\mathbf{x})$. With the derivation of the PI stochastic control starting with dynamic programming in continuous time, it is obvious that the mathematical steps follow the opposite direction as in the section where the the same framework is derived based on the relative entropy-free energy dualities; see Figure 1. Furthermore, within the class of stochastic systems affine in controls and stochastic disturbances, the approach that is discussed in this section provides more general formulations, since it allows cost functions with terms that are both state and control dependent, such as the term $q_1^{\mathrm{T}}(\mathbf{x}, t)\mathbf{u}$. These terms cannot be recovered when the information theoretic approach is used for the class of stochastic systems with affine controls and disturbances. Therefore, under these certain conditions, the dynamic programming approach provides more flexibility in designing cost functions for optimal control problems.

6. Kullback–Leibler Control in Discrete Formulations

The KL control was presented in its most generalized form in [4]. In this section, we will review the KL control for the finite horizon case. A preliminary analysis on the information theoretic connection of the KL control for the infinite horizon case can be found in [6]. Within the KL control framework, the stochastic optimal control problem is formalized as a Markov decision process (MDP) with a stage-wise cost described as:

$$\ell(\mathbf{x}, \mathbf{u}) = q(\mathbf{x}) + \mathbb{KL}\Big(\mathcal{U}\left(\cdot|\mathbf{x}\right)\mathcal{P}\left(\cdot|\mathbf{x}\right)\Big) = q(\mathbf{x}) + \mathbb{E}_{\mathbf{x}'\sim\mathcal{U}(\cdot|\mathbf{x})}\left[\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)}\right)\right].$$

The KL divergence in the last expression is applied to the one step ahead transition probabilities of the control $\mathcal{U}\left(\mathbf{x}'|\mathbf{x}\right)$ and uncontrolled dynamics $\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)$. Application of the Bellman principle of optimality in the finite horizon case results in:

$$V(\mathbf{x}, t_k) = \min_{\mathcal{P}(\cdot|\mathbf{x})}\left(q(\mathbf{x}) + \mathbb{E}_{\mathbf{x}'\sim\mathcal{U}(\cdot|\mathbf{x})}\left[\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)}\right) + V(\mathbf{x}, t_k + 1)\right]\right), \tag{51}$$

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

$$\begin{aligned}\mathbb{E}_{\mathbf{x}'\sim\mathcal{U}(\cdot|\mathbf{x})}\left[\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)}\right) + V(\mathbf{x}', t_{k+1})\right] &= \mathbb{E}_{\mathbf{x}'\sim\mathcal{U}(\cdot|\mathbf{x})}\left[\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)}\right) + \log\left(\frac{1}{\exp\left(-V(\mathbf{x}', t_{k+1}\right)}\right)\right]\\ &= \mathbb{E}_{\mathbf{x}'\sim\mathcal{U}(\cdot|\mathbf{x})}\left[\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)\exp\left(-V(\mathbf{x}', t_{k+1})\right)}\right)\right].\end{aligned}$$

For these purposes, the normalization term $\mathcal{G}_{t_k}[\Phi](\mathbf{x})$ is introduced with $\Phi(\mathbf{x}, t_k) = \exp\left(-V(\mathbf{x}, t_k)\right)$ being the desirability function. More precisely, we will have:

$$\mathcal{G}_{t_k}[\Phi](\mathbf{x}) = \sum_{\mathbf{x}'}\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)\Phi(\mathbf{x}', t_{k+1}) = \mathbb{E}_{\mathbf{x}'\sim\mathcal{P}(\cdot|\mathbf{x})}\Big[\Phi(\mathbf{x}', t_{k+1})\Big]. \tag{52}$$

Therefore, we have:

$$\mathbb{E}_{\mathbf{x}'\sim\mathcal{U}(\cdot|\mathbf{x})}\log\left(\frac{\mathcal{U}\left(\mathbf{x}'|\mathbf{x}\right)}{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)}\right) + V(\mathbf{x}', t_{k+1}) = -\log\left(\mathcal{G}_{t_k}[\Phi](\mathbf{x})\right) + \mathbb{KL}\left(\mathcal{U}\left(\cdot|\mathbf{x}\right)||\frac{\mathcal{P}\left(\mathbf{x}'|\mathbf{x}\right)\Phi(\mathbf{x}', t_{k+1})}{\mathcal{G}_{t_k}[\Phi](\mathbf{x})}\right).$$

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

$$V(\mathbf{x},t_k) = \min_{\mathbf{u}\in\mathcal{U}}\left[q(\mathbf{x}) - \log\left(\mathcal{G}_t[\Phi](\mathbf{x})\right) + \mathbb{KL}\left(\mathcal{U}(\cdot|\mathbf{x})||\frac{\mathcal{P}(\mathbf{x}'|\mathbf{x})\,\Phi(\mathbf{x}',t_{k+1})}{\mathcal{G}_{t_k}[\Phi](\mathbf{x})}\right)\right].$$

The minimum of the Bellman equation is attained by:

$$\mathcal{U}^*(\mathbf{x}'|\mathbf{x}) = \frac{\mathcal{P}(\mathbf{x}'|\mathbf{x})\,\Phi(\mathbf{x}',t_{k+1})}{\mathcal{G}_{t_k}[\Phi](\mathbf{x})}.$$

The equation above provides the transition probability under the optimal control law and, in that sense, the optimal transition probability. Substitution of the optimal distribution above will result in the linear Bellman equation:

$$\Phi(\mathbf{x},t_k) = \exp\left(-q(\mathbf{x})\right)\mathcal{G}_{t_k}[\Phi](\mathbf{x}). \tag{53}$$

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

$$\Phi(\mathbf{x},t_k) = \mathbb{E}_{\mathbf{x}_{\tau+1}\sim\mathcal{P}(\cdot|\mathbf{x}_\tau)}\left[\exp\left(-\sum_{\tau=t_k}^{T} q(\mathbf{x}_\tau)\right)\right]. \tag{54}$$

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 $\mathcal{U}(\vec{\mathbf{x}})$ for the trajectory $\vec{\mathbf{x}} = \{\mathbf{x}_{t_i}, ..., \mathbf{x}_{t_k}, ..\mathbf{x}_{t_N}\}$ that is specified as follows:

$$\begin{aligned}
\mathcal{U}(\vec{\mathbf{x}}) &= \prod_{k=i}^{N}\mathcal{U}^*(\mathbf{x}'|\mathbf{x}) = \prod_{k=i}^{N}\frac{\mathcal{P}(\mathbf{x}'|\mathbf{x})\,\Phi(\mathbf{x}_{t_{k+1}},t_{k+1})}{\mathcal{G}_\tau[\Phi](\mathbf{x})}\\
&= \left(\frac{\mathcal{P}\left(\mathbf{x}_{t_{k+1}}|\mathbf{x}_{t_k}\right)\Phi(\mathbf{x}_{t_{k+1}},t_{k+1})}{\mathcal{G}_{t_k}[\Phi](\mathbf{x})}\right)\left(\frac{\mathcal{P}\left(\mathbf{x}_{t_{k+2}}|\mathbf{x}_{t_{k+1}}\right)\Phi(\mathbf{x}_{t_{k+2}},t_{k+2})}{\mathcal{G}_{t_{k+1}}[\Phi](\mathbf{x})}\right)\cdots\\
&= \left(\frac{\mathcal{P}\left(\mathbf{x}_{t_{k+1}}|\mathbf{x}_{t_k}\right)\exp\left(-q(\mathbf{x}_{t_{k+1}})\right)\mathcal{G}_{t_{k+1}}[\Phi](\mathbf{x})}{\mathcal{G}_{t_k}[\Phi](\mathbf{x})}\right)\times\left(\frac{\mathcal{P}\left(\mathbf{x}_{t_{k+2}}|\mathbf{x}_{t_{k+1}}\right)\exp\left(-q(\mathbf{x}_{t_{k+2}})\right)\mathcal{G}_{t_{k+2}}[\Phi](\mathbf{x})}{\mathcal{G}_{t_{k+1}}[\Phi](\mathbf{x})}\right)\cdots\\
&= \left(\prod_{k=i}^{N}\mathcal{P}\left(\mathbf{x}_{t_{k+1}}|\mathbf{x}_{t_k}\right)\right)\frac{\exp\left(-\mathcal{J}(\vec{\mathbf{x}})\right)}{\mathcal{G}_{t_k}[\Phi](\mathbf{x})}.
\end{aligned}$$

Therefore, the optimal trajectory probability has the form:

$$\mathcal{U}(\vec{\mathbf{x}}) = \frac{\mathcal{P}(\vec{\mathbf{x}})\exp\left(-\mathcal{J}(\vec{\mathbf{x}})\right)}{\mathbb{E}_{\vec{\mathbf{x}}'\sim\mathcal{P}(\cdot)}\left[\exp\left(-\mathcal{J}(\vec{\mathbf{x}}')\right)\right]}. \tag{55}$$

The optimal trajectory probability in the last expression is identical to Equations (7) and (29).

6.1. Connections to Continuous Time

The link of the discrete Bellman Equation in (53) to the corresponding HJB PDE is achieved when expectations $\mathbb{E}_{\mathbf{x}'\sim\mathcal{P}(\cdot|\mathbf{x})}$ and $\mathbb{E}_{\mathbf{x}'\sim\mathcal{U}(\cdot)\mathbf{x}}$ are computed using one step ahead states sampled from the uncontrolled and controlled dynamics:

$$\begin{aligned}
\mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x})\mathrm{d}t + \mathbf{C}(\mathbf{x})\mathrm{d}\mathbf{w}\\
\mathrm{d}\mathbf{x} &= \mathbf{f}(\mathbf{x})\mathrm{d}t + \mathbf{C}(\mathbf{x})(\mathbf{u}\mathrm{d}t + \mathrm{d}\mathbf{w})
\end{aligned} \tag{56}$$

Due to space limitations, we summarize the derivation of the continuous time LSOC with the following lemma. The derivation can be found in the Supplementary Material of [4].

Lemma 2. *Lets consider the dynamics in Equation (56) and the function $V(\mathbf{x}, t) : \Re^n \times \Re \to \Re$ and $\Phi(\mathbf{x}, t) : \Re^n \times \Re \to \Re$ with $\Phi(\mathbf{x}, t)$ satisfying the linear Bellman equation:*

$$\Phi_{(\mathrm{d}t)}(\mathbf{x}, t_k) = \exp\left(-q(\mathbf{x})\mathrm{d}t\right)\mathcal{G}_{t_k}[\Phi_{(\mathrm{d}t)}](\mathbf{x}). \tag{57}$$

where $q(\mathbf{x})$ is state-dependent cost and the operator $\mathcal{G}_{t_k}[\Phi_{(\mathrm{d}t)}](\mathbf{x})$ is defined as in Equation (52). If $V(\mathbf{x}, t) = -\log_e \Phi(\mathbf{x}, t)$, then $V(\mathbf{x}, t)$ satisfies the HJB PDE of an optimal control problem.

$$V(\mathbf{x}(t_0), t_0) = \min_{\mathbf{u}} \mathbb{E}\left[\int_{t_0}^{t_N}\left(q(\mathbf{x}) + \frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{u}\right)\mathrm{d}t\right] \tag{58}$$

subject to controlled dynamics in Equation (56).

This lemma can be seen as an alternative derivation of the Feynman–Kac lemma [9].

7. Algorithms

In this section, we review the derivation of iterative PI control as shown in our previous work [3,11] and also discuss applications and algorithms. In particular, we will start our analysis with the expectation as expressed in Equation (48). Note that this expectation is evaluated over trajectories sampled via forward propagation of uncontrolled diffusion $\mathrm{d}\mathbf{x} = \tilde{\mathbf{f}}(\mathbf{x})\mathrm{d}t + \mathbf{B}(\mathbf{x})\mathrm{d}\mathbf{w}^{(0)}(t)$ in which $\tilde{\mathbf{f}}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}) - \mathbf{G}(\mathbf{x})\mathbf{R}^{-1}q_1(\mathbf{x}, t)$. In this paper, we assume that the state of the stochastic dynamics is partitioned as $\mathbf{x} = [\mathbf{x}_m \;\; \mathbf{x}_c]^{\mathrm{T}}$, and the drift and control transition terms are partitioned as follows $\tilde{\mathbf{f}}(\mathbf{x}) = [\tilde{\mathbf{f}}_m^{\mathrm{T}}(\mathbf{x}) \;\; \tilde{\mathbf{f}}_c^{\mathrm{T}}(\mathbf{x})]^{\mathrm{T}}$ $\mathbf{G}(\mathbf{x}) = [\mathbf{0}_{(n-m)\times m}^{\mathrm{T}} \;\; \mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m)]^{\mathrm{T}}$, with $\tilde{\mathbf{f}}_m(\mathbf{x}) : \Re^n \to \Re^{(n-p)}$, $\tilde{\mathbf{f}}_c(\mathbf{x}) : \Re^n \to \Re^p$, $\mathbf{G}_c(\mathbf{x}_m) : \Re^p \to \Re^{m\times m}$ and diffusion term $\mathbf{B}(\mathbf{x}) = [\mathbf{0}_{(n-p)\times p}^{\mathrm{T}} \;\; \mathbf{B}_c^{\mathrm{T}}(\mathbf{x}_m)]^{\mathrm{T}}$ with $\mathbf{B}_c(\mathbf{x}_m) : \Re^p \to \Re^p$. Note that systems such as multi-body systems have this form. We also assume that:

$$\lambda\mathbf{G}_c(\mathbf{x}_m)\mathbf{R}^{-1}\mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m) = \mathbf{B}_c(\mathbf{x}_m)\mathbf{B}_c^{\mathrm{T}}(\mathbf{x}_m) \tag{59}$$

To derive the iterative path integral control, we will start our analysis with the stochastic representation of the solution of backward Chapman–Kolmogorov PDE:

$$\begin{aligned}\Psi\left(\mathbf{x}(t_i), t_i\right) &= \mathbb{E}_{\tilde{\mathbb{P}}}\left[\exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x}, t)\mathrm{d}t\right)\Psi(\mathbf{x}_{t_N})\right] \\ &= \int \exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x}, t)\mathrm{d}t\right)\Psi(\mathbf{x}_{t_N})\mathrm{d}\tilde{\mathbb{P}}\end{aligned} \tag{60}$$

Since at every iteration k, the sampling process takes place with the use of the control policy $\mathbf{u}_k(\mathbf{x}, t)$, the expression above is formulated as:

$$\Psi\left(\mathbf{x}(t_i), t_i\right) = \int \exp\left(-\int_{t_i}^{t_N}\frac{1}{\lambda}\tilde{q}(\mathbf{x}, , t)\mathrm{d}t\right)\Psi(\mathbf{x}_{t_N})\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\tilde{\mathbb{Q}}}\mathrm{d}\tilde{\mathbb{Q}} \tag{61}$$

where the $\tilde{\mathbb{Q}}$ is the probability measure that corresponds to the diffusion process $\mathrm{d}\mathbf{x} = \tilde{\mathbf{f}}(\mathbf{x})\mathrm{d}t + \mathbf{G}(\mathbf{x})\mathbf{u}_k(\mathbf{x}, t)\mathrm{d}t + \mathbf{B}(\mathbf{x})\mathrm{d}\mathbf{w}^{(1)}$. The terms $\mathbf{u}_k(\mathbf{x}, t)$ and $\mathrm{d}\mathbf{w}^{(1)}$ are the control and noise used at iteration k. The ratio of the two probability measures $\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\tilde{\mathbb{Q}}}$ is the Radon–Nikodým. The aforementioned ratio for stochastic dynamics is formulated as follows:

$$\frac{\mathrm{d}\tilde{\mathbb{P}}}{\mathrm{d}\tilde{\mathbb{Q}}} = \exp\left[-\frac{1}{2\lambda}\int_{t_i}^{t_N}\left(\mathbf{u}_k^{\mathrm{T}}(t)\mathbf{\Upsilon}_{\mathbf{uu}}\mathbf{u}_k(t)\delta t\right)\right] \times \exp\left[-\frac{1}{\lambda}\int_{t_i}^{t_N}\left(\mathbf{u}_k^{\mathrm{T}}(t)\mathbf{\Upsilon}_{\mathbf{uw}}\mathrm{d}\mathbf{w}^{(1)}(t)\right)\right]$$

where the terms $\boldsymbol{\Upsilon}_{\mathbf{uu}}(\mathbf{x}), \boldsymbol{\Upsilon}_{\mathbf{uw}}(\mathbf{x})$ and $\boldsymbol{\Upsilon}$ are defined as $\boldsymbol{\Upsilon}_{\mathbf{uu}}(\mathbf{x}_m) = \mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m)\boldsymbol{\Upsilon}^{-1}\mathbf{G}_c(\mathbf{x}_m)$ and $\boldsymbol{\Upsilon}_{\mathbf{uw}}(\mathbf{x}_m) = \mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m)\boldsymbol{\Upsilon}^{-1}\mathbf{B}_c(\mathbf{x}_m)$ and $\boldsymbol{\Upsilon} = \mathbf{G}_c(\mathbf{x}_m)\mathbf{R}^{-1}\mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m) = \mathbf{B}_c(\mathbf{x}_m)\mathbf{B}_c^{\mathrm{T}}(\mathbf{x}_m)$. After formulating the probability measure $\tilde{\mathbb{P}}$ and using the equation above, Equation (61) will take the form:

$$\Psi\left(\mathbf{x}(t_i), t_i\right) = \lim_{\mathrm{d}t \to 0} \int \frac{1}{\mathfrak{D}(\vec{\mathbf{x}}_i)} \exp\left(-\frac{1}{2\lambda}\mathfrak{L}_k(\vec{\mathbf{x}}_i, \vec{\mathbf{u}}_i^{(k)})\right)\mathrm{d}\vec{\mathbf{x}} \tag{62}$$

where $\mathfrak{L}_k(\vec{\mathbf{x}}_i, \vec{\mathbf{u}}_i^{(k)})$ plays the role of the Lagrangian at iteration k that is specified as follows:

$$\begin{aligned}\mathfrak{L}_k(\vec{\mathbf{x}}_k(t_i), \vec{\mathbf{u}}_k(t_i)) = {} & \phi(\mathbf{x}(t_N)) + \frac{1}{2}\sum_{j=i}^{N-1} \tilde{q}(\mathbf{x}(t_j), t_j)\mathrm{d}t \\ & + \frac{1}{2}\sum_{j=i}^{N-1}\left[\left\| \frac{\mathbf{x}_c(t_j + \mathrm{d}t) - \mathbf{x}_c(t_j)}{\mathrm{d}t} - \boldsymbol{\alpha}(\mathbf{x}(t_j), \mathbf{u}_k(t_j))\right\|^2_{\boldsymbol{\Upsilon}^{-1}_{t_j}}\right]\mathrm{d}t \\ & + \frac{1}{2}\sum_{j=i}^{N-1} \mathbf{u}_k^{\mathrm{T}}(t_j)\Big[\boldsymbol{\Upsilon}_{\mathbf{uu}}\mathbf{u}_k(t_j) \ + 2\mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m(t_j))\boldsymbol{\Upsilon}^{-1}\mu(\mathbf{x}_j)\Big]\mathrm{d}t\end{aligned} \tag{63}$$

where the term $\boldsymbol{\alpha}(\mathbf{x}(t_j), \mathbf{u}_k(t_j))$ is defined as $\boldsymbol{\alpha}(\mathbf{x}(t_j), \mathbf{u}_k(t_j)) = \tilde{\mathbf{f}}_c(\mathbf{x}(t_j)) - \mathbf{G}_c(\mathbf{x}_m(t_j))\mathbf{u}_k(t_j)$ and $\tilde{\mathbf{f}}_c(\mathbf{x}(t_j))$ is the drift term defined as $\tilde{\mathbf{f}}_c(\mathbf{x}(t_j)) = \mathbf{f}_c(\mathbf{x}(t_j)) - \mathbf{G}_c(\mathbf{x}(t_j))\mathbf{R}^{-1}q_1(\mathbf{x}(t_j), t_j)$. The terms $\vec{\mathbf{u}}_k(t_i) = \{\mathbf{u}_k(t_i), ..., \mathbf{u}_k(t_{N-1})\}$ and $\vec{\mathbf{x}}_k(t_i) = \{\mathbf{x}_k(t_i), ..., \mathbf{x}_k(t_N)\}$ are the state and control trajectories at iteration k. In addition, the term $\mu(\mathbf{x}_j) = \frac{\mathbf{x}_c(t_j+\mathrm{d}t) - \mathbf{x}_c(t_j)}{\mathrm{d}t} - \mathbf{f}_c(\mathbf{x}(t_j)) - \mathbf{G}_c(\mathbf{x}_m(t_j))\mathbf{u}_k(t_j)$, and thus, $\mu(\mathbf{x}_j)\mathrm{d}t = \mathbf{B}_c(\mathbf{x}_m(t_j))\mathrm{d}\mathbf{w}^{(1)}(t)$. In a more compact form, Equation (62) can be written as:

$$\Psi\left(\mathbf{x}(t_i), t_i\right) = \lim_{dt \to 0} \int \exp\left(-\frac{1}{2\lambda}\tilde{\mathfrak{L}}_k(\vec{\mathbf{x}}(t_i), \vec{\mathbf{u}}_k(t_i))\right)\mathrm{d}\vec{\mathbf{x}} \tag{64}$$

where $\tilde{\mathfrak{L}}_k = \mathfrak{L}_k + 2\lambda \log \mathfrak{D}$.

Lemma 3. *(Iterative path integral optimal control:) Given the form of the Lagrangian in Equation (63) and the desirability function in Equation (64), the iterative optimal path integral control is specified as:*

$$\begin{aligned}\mathbf{u}_{k+1}(\mathbf{x}(t_i), t_i)\mathrm{d}t = {} & -\underbrace{\mathbf{R}^{-1}q_1(\mathbf{x}(t_i), t_i)\mathrm{d}t}_{\textit{Cost Function}} + \underbrace{\boldsymbol{\Omega}(\mathbf{x}_m(t_i))\mathbf{G}_c(\mathbf{x}_m(t_i))\mathbf{u}_k(\mathbf{x}(t_i), t_i)\mathrm{d}t}_{\textit{Previous Control}} \\ & + \underbrace{\boldsymbol{\Omega}(\mathbf{x}_m(t_i))\mathbf{B}_c(\mathbf{x}(t_i))\delta\mathbf{u}_{PI}(\mathbf{x}(t_i), t_i)}_{\textit{Path Integral Correction}}\end{aligned} \tag{65}$$

The path integral correction term $\delta\mathbf{u}_{PI}$ is given by:

$$\delta\mathbf{u}_{PI}(\mathbf{x}(t_i), t_i) = \mathbb{E}_{P(\vec{\mathbf{x}})}\left(\mathrm{d}\mathbf{w}^{(1)}(t_i)|\mathbf{x}(t_i)\right) \tag{66}$$

where $P(\vec{\mathbf{x}}) = \frac{e^{-\frac{1}{\lambda}\tilde{\mathfrak{L}}(\vec{\mathbf{x}}_i)}}{\int e^{-\frac{1}{\lambda}\tilde{\mathfrak{L}}(\vec{\mathbf{x}}_i)}\mathrm{d}\vec{\mathbf{x}}_i^{(c)}}$, while the term $\boldsymbol{\Omega}(\mathbf{x}_m(t_i))$ is defined as:

$$\boldsymbol{\Omega}(\mathbf{x}_m(t_i)) = \mathbf{R}^{-1}\mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m(t_i))\boldsymbol{\Upsilon}^{-1}$$

Proof. The optimal controls based on Relation (49) is specified as:

$$\begin{aligned}\mathbf{u}(\mathbf{x}(t), t) &= -\mathbf{R}^{-1}\left(q_1(\mathbf{x}, t) - \lambda[\mathbf{0}_{k\times p} \ \ \mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m)]\frac{\psi_{\mathbf{x}}(\mathbf{x}, t)}{\psi(\mathbf{x}, t)}\right) \\ &= -\mathbf{R}^{-1}q_1(\mathbf{x}, t) + \lambda\mathbf{R}^{-1}\mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m)\frac{\psi_{\mathbf{x}_c(t)}(\mathbf{x}, t)}{\psi(\mathbf{x}, t)}\end{aligned} \tag{67}$$

Next, the term $\frac{\psi_{\mathbf{x}_c(t)}(\mathbf{x},t)}{\psi(\mathbf{x},t)}$ is computed where $\psi_{\mathbf{x}_c(t)}(\mathbf{x},t) = \nabla_{\mathbf{x}_c(t)}\psi(\mathbf{x},t)$. In particular, by pushing the gradient inside the expectation in the definition of the desirability function, we have that:

$$\frac{\nabla_{\mathbf{x}_c(t)}\psi(\mathbf{x},t)}{\psi(\mathbf{x},t)} = \mathbb{E}_{P(\vec{\mathbf{x}})}\left(\nabla_{\mathbf{x}_c(t_i)}\tilde{\mathfrak{L}}(\vec{\mathbf{x}}_i)\right)$$

The term $\mathbb{E}_{P(\vec{\mathbf{x}})}$ is the expectation under the probability $P(\vec{\mathbf{x}})$, which is defined as $P(\vec{\mathbf{x}}) = \frac{e^{-\frac{1}{\lambda}\tilde{\mathfrak{L}}(\vec{\mathbf{x}}_i)}}{\int e^{-\frac{1}{\lambda}\tilde{\mathfrak{L}}(\vec{\mathbf{x}}_i)}\mathrm{d}\vec{\mathbf{x}}_i^{(c)}}$. Based on the form of the Lagrangian in Equation (63), the term $\left(\nabla_{\mathbf{x}_c(t_i)}\tilde{\mathfrak{L}}(\vec{\mathbf{x}}_i)\right)$ takes the form:

$$\left(\nabla_{\mathbf{x}_c(t_i)}\tilde{\mathfrak{L}}(\vec{\mathbf{x}}_i)\right) = -\boldsymbol{\Upsilon}^{-1}\bigg(\mathbf{G}_c(\mathbf{x}_m(t_i))\mathbf{u}_k(t_i) + \mu(\mathbf{x}_{t_i})\bigg) + \mathcal{O}(\mathrm{d}t) \tag{68}$$

The notation $\mathcal{O}(\mathrm{d}t)$ is used for terms of order $\mathrm{d}t$. We will keep this notation, as we will see that these terms will cancel. The optimal control is expressed as:

$$\begin{aligned}\mathbf{u}_{k+1}(\mathbf{x}(t_i),t_i)\mathrm{d}t &= -\mathbf{R}^{-1}q_1(\mathbf{x}_{t_i},t_i)\mathrm{d}t + \mathbf{R}^{-1}\mathbf{G}_c^{\mathrm{T}}(\mathbf{x}_m(t_i))\mathbb{E}_{P(\vec{\mathbf{x}})}\left(\nabla_{\mathbf{x}_{t_i}^{(c)}}\tilde{\mathfrak{L}}(\vec{\mathbf{x}}_i)\right)\mathrm{d}t \\ &\approx -\mathbf{R}^{-1}q_1(\mathbf{x}_{t_i},t_i)\mathrm{d}t + \mathbb{E}_{P(\vec{\mathbf{x}})}\bigg(\mathbf{u}_L\bigg) + \mathcal{O}(\mathrm{d}t^2)\end{aligned} \tag{69}$$

The term $\mathbf{u}_L$ in the expression above takes the form:

$$\mathbf{u}_L = \mathbf{R}^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{x}_m(t_i))\boldsymbol{\Upsilon}(\mathbf{x}_m(t_i))^{-1}\bigg(\mathbf{G}_c(\mathbf{x}_m(t_i)\mathbf{u}_k(t_i)\mathrm{d}t + \mu(\mathbf{x}_{t_i})\mathrm{d}t\bigg) \tag{70}$$

The multiplication of the optimal controls with $\mathrm{d}t$ is done in terms of quadratic order with respect to $\mathrm{d}t$. These terms cancel out as $\mathrm{d}t \rightarrow 0$ or for very small $\mathrm{d}t$. Finally, since $\mu(\mathbf{x})\mathrm{d}t = \mathbf{B}(\mathbf{x})\mathrm{d}\mathbf{w}^{(1)}(t)$, we will have that the final result is:

$$\mathbf{u}_L = \mathbf{R}^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{x}_m(t_i))\boldsymbol{\Upsilon}(\mathbf{x}_m(t_i))^{-1}\bigg(\mathbf{G}_c(\mathbf{x}_m(t_i))\mathbf{u}_k(t_i)\mathrm{d}t + \mathbf{B}_c(\mathbf{x}_m(t_i))\mathrm{d}\mathbf{w}^{(1)}(t)\bigg) \tag{71}$$

By combining Equations (69), (68) and (71), the final form for the iterative optimal control is expressed in Equation (67). □

7.1. Open Loop Formulations and Application to an Inverted Pendulum

One of the characteristics of the iterative optimal control in Equation (65) is that the control $\mathbf{u}_{k+1}(\mathbf{x},t)$ at iteration $k+1$ requires the knowledge of the control $\mathbf{u}_k(\mathbf{x},t)$ for every pair $(\mathbf{x},t)$. While the iterative characteristic of the proposed scheme improves scalability, the requirement for computing $\mathbf{u}_k(\mathbf{x},t)$ for any state and time $(\mathbf{x},t)$ prohibits the application of this scheme to high dimensional systems. An alternative approach to address this is to use a parametric or non-parametric approximation method to represent $\mathbf{u}_k(\mathbf{x},t)$ and apply iterative path integral control in its initial feedback form (65).

Here, we suggest a receding horizon open loop formulation and restrict our analysis to stochastic systems with $\mathbf{B}_c(\mathbf{x}(t)) = \mathbf{B}_c$, $\mathbf{G}_c(\mathbf{x}) = \mathbf{G}_c$ and $q_1(\mathbf{x}(t),t) = 0$. The algorithm is provided in Tables Algorithm 1 and Algorithm 2 and consists of three procedures, namely $FnSample\_Trajectories$, $FnUpdate\_Controls$ and $FnApply\_Control\_Dynamics$. In particular, the functionality for the procedure $FnSample\_Trajectories$ is to sample trajectories starting from state $\mathbf{x}_k$ by using an initial control trajectory $\vec{\mathbf{u}}(:,k \rightarrow T) = (\mathbf{u}(k),\mathbf{u}(k+1),...,\mathbf{u}(T))$ and to return these sample trajectories and noise profiles used for sampling $\mathrm{d}\mathbf{w}(:, k \to T)$. The next procedure is $FnSample\_Update$ and has as input the control trajectory $\vec{\mathbf{u}}(:, k \ \to \ T)$, the sampled state trajectories $Sampled\_Trajectories$ and the noise profiles $\mathrm{d}\mathbf{w}(:, k \to T)$. Its functionality, illustrated in Algorithm 2, is to apply the iterative path integral control in its open loop formulation and to compute the new control trajectory $\vec{\mathbf{u}}_{\text{updated}}(:, k \ \to \ T)$. In the open loop formulation, the state dependence of the correction term in Equation (66) is dropped, and therefore, the term $\delta\mathbf{u}_{\text{PI}}(t)$ becomes only time varying. In $FnApply\_Control\_Dynamics$, the control is applied for one time step, and the overall algorithm repeats again.

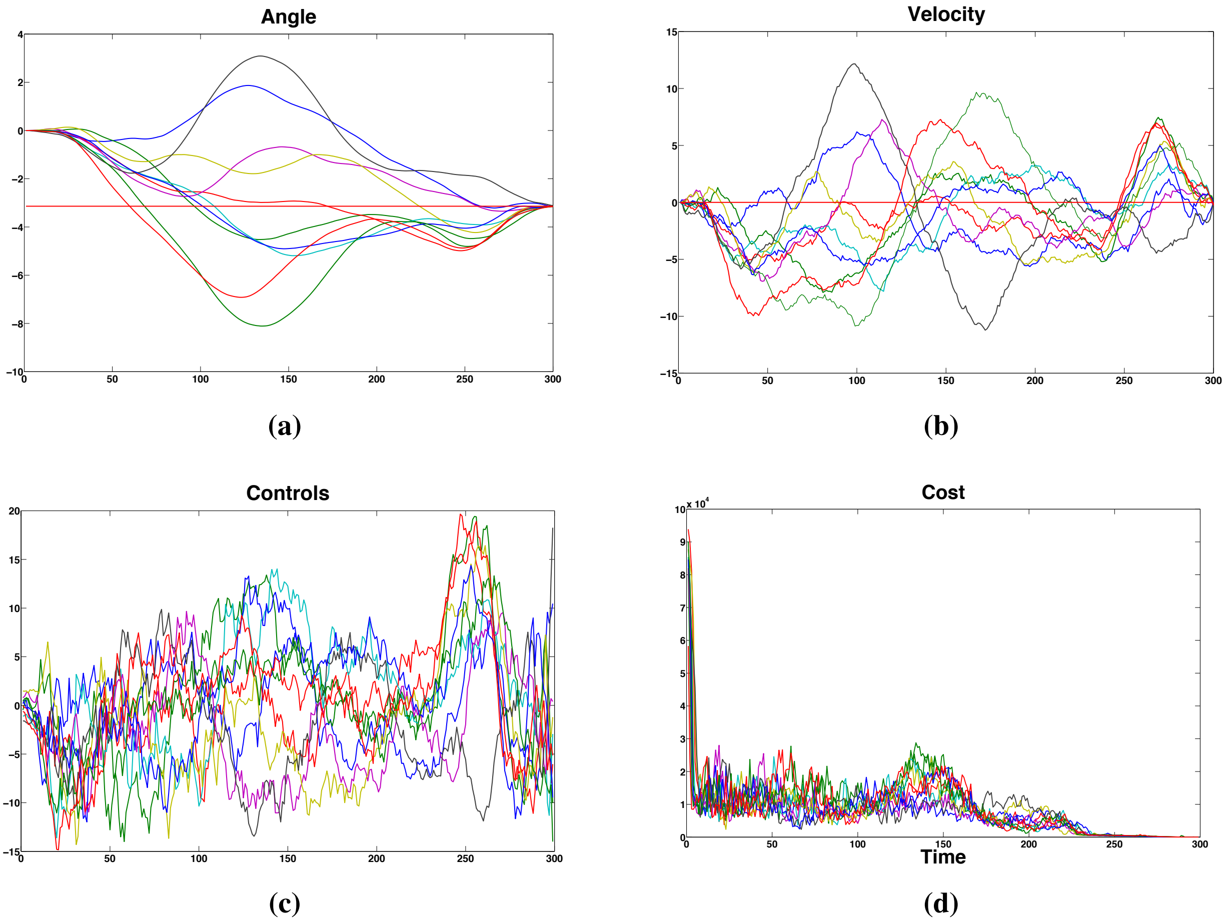
Algorithm 1: Iterative stochastic optimal control.

```
Given the start state $\mathbf{x}_0$, initial control trajectory $\vec{\mathbf{u}}_k$;
for $k = 0$ to $T$ do
    $[Sampled\_Trajectories, \mathrm{d}\mathbf{w}(k \to T)] = FnSample\_trajectories(\vec{\mathbf{u}}_{\text{current}}(:, k \ \to \ T), \mathbf{x}_k, k)$;
    $[\vec{\mathbf{u}}_{\text{next}}]$= $FnUpdate\_Controls(\vec{\mathbf{u}}_k, Sampled\_Trajectories, \mathrm{d}\mathbf{w}(k \to T))$;
    $[\mathbf{x}_{k+1}] = FnApply\_Control\_Dynamic(\mathbf{x}_k, \vec{\mathbf{u}}_{\text{next}}(:, 1))$;
    $\vec{\mathbf{u}}_{\text{current}} = \vec{\mathbf{u}}_{\text{next}}$;
end
```

Algorithm 2: Update_Controls.

```
FnUpdate_Controls($\vec{\mathbf{u}}_k, Sampled\_Trajectories, \mathrm{d}\mathbf{w}(k \to T)$);
Given $Sampled\_Trajectories$, and controls $\vec{\mathbf{u}}_k(:, k \ \to \ T)$;
for $i = 1$ to $Number\_of\_Iterations$ do
    for $t = k$ to $T$ do
        Compute the path integral correction term $\delta\mathbf{u}_{\text{PI}}(t) = \mathbb{E}_{P(\vec{\mathbf{x}})}(\mathrm{d}\mathbf{w}(t))$ in Equation (66) with $\vec{\mathbf{x}}_{\text{sampled}}$ ;
        Compute $\mathbf{u}_{i+1}(t)$ based on Equation (65);
    end
end
$\vec{\mathbf{u}}_{\text{updated}}(:, k \ \to \ T) = \vec{\mathbf{u}}_{Number\_of\_Iterations}(:, k \ \to \ T)$;
return $\vec{\mathbf{u}}_{\text{updated}}(:, k \ \to \ T)$
```

Here, we apply the proposed algorithm to a swing up task of an inverted pendulum. The task is to bring the pendulum from initial state $x = [x_1, x_2] = [0, 0]$ to target state $p^* = [p_1^*, p_2^*] = [-\pi, 0]$. The pendulum has mass $m = 1$ kg and link length $l = 0.5$ m. The number of sampled trajectories returned by the function $FnSample\_Trajectories$ is 200. The terminal cost is $\phi(x_{t_N}, t_N) = 1,000 * (x_1(t_N) - p_1^*)^2 + 100 * (x_2(t_N) - p_2^*)^2$, and the state cost $q(x) = 0$ and control cost $\frac{1}{2\sigma^2}u^2$. The variance of noise is $\sigma = 0.5$, and the time horizon used is $t_N = 300 * 0.01 = 3$ s. The state, control trajectories and the cost are illustrated in Figure 2a–d. In particular, Figure 2a illustrates a set of 10 angular trajectories that reach the desired state(red horizontal line), and Figure 2b illustrates the corresponding angular velocities that also reach the desired state (red horizontal line). Figure 2c illustrates the stochastic iterative path integral control trajectories. Finally, Figure 2d illustrates the cost for the 10 trials as the system moves towards the target state p^* under the application of the iterative optimal path integral control.



(a) **(b)** **(c)** **(d)**

Figure 2. (**a**) Angle; (**b**) rotational velocity; (**c**) controls; and (**d**) cost trajectories.

8. Discussion

In this paper we present four different approaches to LSOC. In the first approach, which is also the most traditional one, stochastic optimal control is formulated as the minimization of an objective function $J(\mathbf{x}, \mathbf{u})$ in Equation (42) 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 Equation (46) transforms the HJB into 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 starts with the risk-seeking version of the cost $\mathcal{J}(\mathbf{x})$. This quantity has also the form of the Helmholtz free energy. With the application of Girsanov's theorem between controlled and uncontrolled dynamics and the use of Jensen inequality, the Helmholtz Free energy is the lower bound of the objective function that consists of a state-dependent cost and an information cost, which is a measure of control effort. The link to Bellman optimality is established by showing that the Helmholtz free energy satisfies the HJB equation, and therefore, it is a value function. It should be clear by now that steps to information theoretic representation are in the opposite direction, as shown in Figure 1. While in the information theoretic approach, the analysis starts with the derivation of the Legendre transformation and ends with the HJB PDE, in the traditional approach, the analysis starts from dynamic programming and ends in a special case of the Legendre transformation.

In the third approach, the stochastic optimal control problem is derived using the maximum entropy principle. The optimization problem is formulated as the maximization of the generalized Boltzmann–Gibbs–Shannon entropy subject to performance constraints. The optimization is with respect to a probability measure that corresponds to the controlled dynamics. At the thermodynamic equilibrium, we have the maximization of the generalized Boltzmann–Gibbs–Shannon entropy, which is equivalent to the minimization of the control effort subject to the performance and normalization constrains as expressed in the optimization problem in (29).

In the KL stochastic optimal control framework [4], the treatment is for MDP. The analysis starts with the construction of a cost function that consists of state cost and an information cost defined as the KL divergence between the one step ahead transition probabilities of the control and uncontrolled dynamics. Next dynamic programming is used to derive the Bellman equation in discrete time. The connection to continuous time stochastic optimal control is performed when one step ahead transition probabilities of the control and uncontrolled dynamics correspond to controlled and uncontrolled diffusion processes with the same drift.

In this work, we present different views of nonlinear stochastic control and provide connections, new generalizations and algorithms. Given all of the aforementioned approaches, it is clear that the idea of exponential transformation of the value function existed already in the early work of control theory. However, it was recently conceptualized as desirability and further explored in terms of algorithms, quantum mechanical interpretations and discrete time formulations. While significant progress has been made in both theory and algorithms, there are fundamental assumptions in the frameworks presented in this work that restrict their applicability to systems where the uncertainty is only of a stochastic nature. This means that there are assumptions on the structure of the dynamics that allow uncertainty only due to noise. Given the progress on non-parametric regression methods in statistical machine learning and the different ways to represent uncertainty, future work on stochastic control will focus on the development of theory and algorithms for the stochastic control of systems with unknown and stochastic dynamics. In these cases, uncertainty will not only incorporate stochasticity due to the existence of noise, but it will also include probabilistic representations of the unknown dynamics. Finally, the generalizations and thermodynamic interpretations presented in this work create new research directions towards the development of stochastic control algorithms for general classes of stochastic systems and for information theoretic measures, such as non-extensive entropies that go beyond the entropy measures used in Boltzmann Gibbs statistical mechanics.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Kappen, H.J. An introduction to stochastic control theory, path integrals and reinforcement learning. In *Cooperative Behavior in Neural Systems*; Marro, J., Garrido, P.L., Torres, J.J., Eds.; American Institute of Physics: College Park, MD, USA, 2007; Volume 887, pp. 149–181.
2. Kappen, H.J. Path integrals and symmetry breaking for optimal control theory. *J. Stat. Mech. Theory Exp.* **2005**, *11*, P11011.
3. Theodorou, E.; Buchli, J.; Schaal, S. A Generalized Path Integral Approach to Reinforcement Learning. *J. Mach. Learn. Res.* **2010**, 3137–3181.
4. Todorov, E. Efficient computation of optimal actions. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 11478–11483.
5. Todorov, E. Linearly-solvable markov decision problems. In *Advances in Neural Information Processing Systems 19*; Scholkopf, B., Platt, J., Hoffman, T., Eds.; MIT Press: Cambridge, MA, USA, 2007.
6. Pan, Y.; Theodorou, E. Nonparametric infinite horizon Kullback-Leibler stochastic control. In Proceedings of the 2014 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), Orlando, USA, 9–12 December 2014; pp. 1–8.
7. Friedman, A. *Stochastic Differential Equations And Applications*; Academic Press: Waltham, MA, USA, 1975.
8. Karatzas, I.; Shreve, S.E. *Brownian Motion and Stochastic Calculus (Graduate Texts in Mathematics)*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1991.
9. Øksendal, B.K. *Stochastic differential equations: An Introduction with Applications*, 6th ed.; Springer: Berlin, Germany, 2003.
10. Horowitz, M.B.; Damle, A.; Burdick, J.W. Linear Hamilton Jacobi Bellman Equations in High Dimensions. **2014**, arXiv:1404.1089.
11. Theodorou, E.; Todorov, E. Relative entropy and free energy dualities: Connections to Path Integral and KL control. In Proceedings of 51st IEEE Conference on Decision and Control, Maui, HI, USA, 10–13 December 2012; pp. 1466–1473.
12. Fleming, W. Exit probabilities and optimal stochastic control. *Appl. Math. Optim.* **1971**, *9*, 329–346.
13. Dai Pra, P.; Meneghini, L.; Runggaldier, W. Connections between stochastic control and dynamic games. *Math. Control Signals Syst. (MCSS)* **1996**, *9*, 303–326.
14. Mitter, S.K.; Newton, N.J. A Variational Approach to Nonlinear Estimation. *SIAM J. Control Optim.* **2003**, *42*, 1813–1833.
15. Fleming, W.H.; Soner, H.M. *Controlled Markov Processes and Viscosity Solutions*, 2nd ed.; Springer: New York, NY, USA, 2006.
16. Wehrl, A. The many facets of entropy. *Rep. Math. Phys.* **1991**, *30*, 119–129.
17. Yang, J.; Kushner, J.H. A monte carlo method for sensitivity analysis and parametric optimization of nonlinear stochastic systems. *SIAM J. Control Optim.* **1991**, *29*, 1216–1249.
18. Fleming, W.H.; Soner, H.M. *Controlled Markov Processes and Viscosity Solutions*, 1st ed.; Springer: New York, NY, USA, 1993.

19. Stengel, R.F. *Optimal Control and Estimation*; Dover Publications: New York, NY, USA, 1994.
20. Leadbetter, R.; Cambanis, S.; Pipiras, P. *Basic Course in Measure and Probabilty*; Cambridge University Press: Cambridge, UK, 2014.
21. Kappen, H.J. Linear theory for control of nonlinear stochastic systems. *Phys. Rev. Lett.* **2005**, *95*, 200201.