Model-based Path Integral Stochastic Control: A Bayesian Nonparametric Approach

Yunpeng Pan, Evangelos A. Theodorou, and **Michail Kontitsis**
Daniel Guggenheim School of Aerospace Engineering
Institute for Robotics and Intelligent Machines
Georgia Institute of Technology
Atlanta, GA 30332
ypan37@gatech.edu, evangelos.theodorou@ae.gatech.edu

Abstract

Over the last few years, sampling-based stochastic optimal control (SOC) frameworks have shown impressive performances in reinforcement learning (RL) with applications in robotics. However, such approaches require a large amount of samples from many interactions with the physical systems. To improve learning efficiency, we present a novel model-based and data-driven SOC framework based on path integral formulation and Gaussian processes (GPs). The proposed approach learns explicit and time-varying optimal controls autonomously from limited sampled data. Based on this framework, we propose an iterative control scheme with improved applicability in higher-dimensional and more complex control tasks. We demonstrate the effectiveness and efficiency of the proposed framework using two nontrivial examples. Compared to state-of-the-art RL methods, the proposed framework features superior control learning efficiency.

1 Introduction

Stochastic optimal control based on exponential transformation of the value function has demonstrated remarkable applicability in robotic control and planning problems, created new research avenues in terms of theoretical generalizations and scalable optimal control algorithms. Although the exponential transformation of the value function existed already in control theory [1],[2], it was only very recently conceptualized as *desirability* and explored in terms of algorithms [3], path integral interpretations [4] and discrete formulations [5]. The resulting stochastic optimal control frameworks are known under the names of Path Integral (PI) control for continuous time, Kullback Leibler (KL) control for discrete time, or more generally Linearly Solvable Optimal Control [5].

One of the most attractive characteristics of the PI control is that optimal control problems can be solved with forward sampling of Stochastic Differential Equations (SDEs). While the process of sampling with SDEs is more scalable than the process of numerically solving partial differential equations, it still suffers from the curse of dimensionality when performed in a naive fashion. One way to circumvent this problem is to parameterize policies [3] and then perform optimization with sampling. However, in this case one has to impose the structure of the policy a-priori and therefore restrict the possible optimal control solutions within the assumed parameterization.

Motivated by the aforementioned limitations, in this paper we introduce a Bayesian nonparametric model-based approach to PI control. Different from most sampling-based approaches, our method learns a probabilistic model from limited sampled data by taking into account model uncertainties. The optimal controls are given in explicit forms based on analytic expressions of path integrals. Furthermore, we develop an iterative control scheme based on importance sampling. Compared to related works in GP-based RL/control [6][7] and PI controls [3][8][9][10] the proposed framework

features merits from both. Firstly, the proposed method finds optimal controls more efficiently than PI controls thanks to the analytic computations of path integrals. Secondly, the proposed work offers faster learning speed than gradient-based policy search methods, which usually rely on optimization solvers (e.g. CG, BFGS) to find optimal policies. Thirdly, the proposed framework requires significantly less sampled data compared to sampling-based approaches.

2 Problem Formulation

We consider a unknown nonlinear stochastic system described by the following differential equation

$$\mathrm{d}\mathbf{x} = \Big(\mathbf{f}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}\Big)\mathrm{d}t + \mathbf{B}(\mathbf{x})\mathrm{d}\omega, \quad \mathrm{d}\omega \sim \mathcal{N}(0, \mathbf{\Sigma}_\omega), \tag{1}$$

with state $\mathbf{x} \in \mathbb{R}^n$, control $\mathbf{u} \in \mathbb{R}^m$, and standard Brownian motion noise $\omega \in \mathbb{R}^p$. $\mathbf{G}(\mathbf{x}) \in \mathbb{R}^{n\times m}$ is the control matrix and $\mathbf{B}(\mathbf{x}) \in \mathbb{R}^{n\times p}$ is the diffusion matrix. The stochastic optimal control problem is defined as finding the controls $\mathbf{u}_t$ that minimize the expected cost

$$J(\tau_0) = \mathbb{E}\Big[q(\mathbf{x}_T) + \int_{t=0}^{T} \mathcal{L}\Big(\mathbf{x}_t, \mathbf{u}_t, t\Big)\mathrm{d}t\Big], \tag{2}$$

where $q(\mathbf{x}_T)$ is the terminal cost, $\mathcal{L}(\mathbf{x}_t, \mathbf{u}_t, t)$ is the instantaneous cost rate, $\mathbf{u}_t$ is the control input. The cost $J(\tau_0)$ is defined as the expectation of the total cost accumulated from $t = 0$ to T. $\tau(0)$ is a trajectory starting from $\mathbf{x}_0$ to $\mathbf{x}_T$. We use the instantaneous cost $\mathcal{L}(\mathbf{x}_t, \mathbf{u}_t, t) = q(\mathbf{x}_t, t) + \frac{1}{2}\mathbf{u}_t^{\mathrm{T}}\mathbf{R}\mathbf{u}_t$, where $q(\mathbf{x}_t, t)$ is an arbitrary state-dependent cost function, $\mathbf{R}$ is a semi-definite weight matrix of the quadratic control cost. In this paper, we use a quadratic cost function $q(\mathbf{x}_t, t) = (\mathbf{x}_t - \mathbf{x}_t^{goal})^{\mathrm{T}}\mathbf{Q}(\mathbf{x}_t - \mathbf{x}_t^{goal})$, where $\mathbf{x}_t^{goal}$ is the desired states. For numerical implementation we use the discrete-time formulation [1]. For concise formulation we use abbreviated notations $\mathcal{L}_t = \mathcal{L}(\mathbf{x}_t, \mathbf{u}_t, t), \mathbf{G}_t = \mathbf{G}(\mathbf{x}_t), \mathbf{B}_t = \mathbf{B}(\mathbf{x}_t), \mathbf{f}_t = \mathbf{f}(\mathbf{x}_t)$ and $q_t = q(x_t, t)$.

3 Path Integral Control

In this section we briefly review the concept and formulation of Path Integral control. We start with the Hamilton-Jacobi-Bellman (HJB) equation. The HJB equation states the optimality condition for value function. The value function is defined by the Bellman equation

$$V(\mathbf{x}_t) = \min_{\mathbf{u}_{0,\dots,T}} J^{\pi}(\tau_t). \tag{3}$$

And the stochastic HJB equation is defined as

$$-\partial_t V_t = \min_{\mathbf{u}_t}\left(\mathcal{L}_t + (\nabla_{\mathbf{x}} V_t)^{\mathrm{T}}\Big(\mathbf{f}_t + \mathbf{G}_t\mathbf{u}_t\Big) + \frac{1}{2}\mathrm{tr}\Big((\nabla_{\mathbf{xx}} V_t)\mathbf{B}_t\mathbf{B}_t^{\mathrm{T}}\Big)\right). \tag{4}$$

Where ∂_t is the partial derivative w.r.t time. $\nabla_{\mathbf{x}}, \nabla_{\mathbf{xx}}$ refer to the Jacobian and Hessian of the value function w.r.t the state, respectively. Taking the gradient w.r.t $\mathbf{u}_t$ of the expression inside the parenthesis (4), we obtain the corresponding optimal control $\hat{\mathbf{u}}_t = -\mathbf{R}^{-1}\mathbf{G}_t^{\mathrm{T}}(\nabla_{\mathbf{x}} V_t)$. Substitution of the optimal control back into (4) yields the following partial differential equation (PDE)

$$-\partial_t V_t = q_t + (\nabla_{\mathbf{x}} V_t)^{\mathrm{T}}\mathbf{f}_t - \frac{1}{2}(\nabla_{\mathbf{x}} V_t)^{\mathrm{T}}\mathbf{G}_t\mathbf{R}^{-1}\mathbf{G}_t(\nabla_{\mathbf{x}} V_t) + \frac{1}{2}\mathrm{tr}\Big((\nabla_{\mathbf{xx}} V_t)\mathbf{B}_t\mathbf{B}_t^{\mathrm{T}}\Big). \tag{5}$$

In order to solve the above PDE, we apply an exponential transformation of the optimal value function $\Psi(\mathbf{x}_t) = \exp\Big(-\frac{1}{\lambda}V(\mathbf{x}_t)\Big)$, where $\Psi(\mathbf{x}_t)$ is called the *desirability* of $\mathbf{x}_t$. We use abbreviation Ψ_t for the rest of the paper. The HJB equation can be transformed to a linear PDE

$$-\partial_t \Psi_t = \frac{1}{\lambda}q_t\Psi_t + \mathbf{f}_t^{\mathrm{T}}(\nabla_{\mathbf{x}}\Psi_t) + \frac{1}{2}\mathrm{tr}\Big((\nabla_{\mathbf{xx}}\Psi_t)\mathbf{B}_t\mathbf{B}_t^{\mathrm{T}}\Big). \tag{6}$$

By applying the Feynman-Kac formula [8]. Under the assumption that $\mathbf{R} = \lambda\mathbf{\Sigma}_w^{-1}$, the above PDE can be solved as

$$\Psi_t = \lim_{\mathrm{d}t\to 0}\int \mathrm{p}(\tau_t|\mathbf{x}_t)\exp\Big(-\frac{1}{\lambda}\big(\sum_{j=t}^{T} q_j\mathrm{d}t\big)\Big)\mathrm{d}\tau_t. \tag{7}$$

[1] The discrete-time formulation of the dynamics is $\mathrm{d}\mathbf{x}_t = \mathbf{x}_{t+\mathrm{d}t} - \mathbf{x}_t = (\mathbf{f}_t + \mathbf{G}_t\mathbf{u}_t)\mathrm{d}t + \mathbf{B}_t\mathrm{d}\omega\sqrt{\mathrm{d}t}$.

And the optimal control is obtained as

$$\hat{\mathbf{u}}_t = -\mathbf{R}^{-1}\mathbf{G}_t^{\mathrm{T}}(\nabla_{\mathbf{x}} V_t) = \lambda \mathbf{R}^{-1}\mathbf{G}_t^{\mathrm{T}}\Big(\frac{\nabla_{\mathbf{x}}\Psi_t}{\Psi_t}\Big). \tag{8}$$

$\hat{\mathbf{u}}_t$ can be approximated based on path costs of sampled trajectories [8][4][9][3][10][11]. However, these sampling-based approaches require a large amount of data from extensive trials on physical systems. Now we introduce an efficient model-based approach to approximating $\nabla_{\mathbf{x}}\Psi_t$ and Ψ_t.

4 Proposed Approach

4.1 Bayesian nonparametric formulation of path integral control

In this paper, the unknown state transition function $\mathbf{f}(\cdot)$ can be viewed as an inference with the goal of inferring $\mathrm{d}\mathbf{x}$ given $\mathbf{x}$. We view this inference as a nonlinear regression problem, and we assume $\mathbf{f}(\cdot)$ can be represented by Gaussian processes (GP). A GP is defined as a collection of random variables, any finite number subset of which have a joint Gaussian distribution. Given a sequence of state-control pairs $\tilde{\mathbf{X}} = \{(\mathbf{x}_0, \mathbf{u}_0), \dots (\mathbf{x}_T, \mathbf{u}_T)\}$, and the corresponding state transition $\mathrm{d}\mathbf{X} = \{\mathrm{d}\mathbf{x}_0, \dots, \mathrm{d}\mathbf{x}_T\}$, a GP is completely defined by a mean function and a covariance function. The joint distribution of the observed output and the output corresponding to a given test state-control pair $\tilde{\mathbf{x}}^* = (\mathbf{x}^*, \mathbf{u}^*)$ can be written as $\mathrm{p}\begin{pmatrix} \mathrm{d}\mathbf{X} \\ \mathrm{d}\mathbf{x}^* \end{pmatrix} \sim \mathcal{N}\Big(0, \begin{bmatrix} \mathbf{K}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) + \sigma_n \mathbf{I} & \mathbf{K}(\tilde{\mathbf{X}}, \tilde{\mathbf{x}}^*) \\ \mathbf{K}(\tilde{\mathbf{x}}^*, \tilde{\mathbf{X}}) & \mathbf{K}(\tilde{\mathbf{x}}^*, \tilde{\mathbf{x}}^*) \end{bmatrix}\Big)$. The covariance of this multivariate Gaussian distribution is defined via a kernel matrix $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j)$. In particular, in this paper we consider the Gaussian kernel $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_s^2 \exp(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)) + \sigma_n^2$, with $\sigma_s, \sigma_n, \mathbf{W}$ the hyper-parameters. The kernel function can be interpreted as a similarity measure of random variables. More specifically, if the training pairs $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{X}}_j$ are close to each other in the kernel space, their outputs $\mathrm{d}\mathbf{x}_i$ and $\mathrm{d}\mathbf{x}_j$ are highly correlated. The posterior distribution, which is also a Gaussian, can be obtained by constraining the joint distribution to contain the output $\mathrm{d}\mathbf{x}^*$ that is consistent with the observations. Assuming independent outputs (no correlation between each output dimension) and given a test input $\tilde{\mathbf{x}}_t = (\mathbf{x}_t, \mathbf{u}_t)$ at time step t, the one-step predictive mean and variance of the state transition are specified as $\mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_t] = \mathbf{K}(\tilde{\mathbf{x}}_t, \tilde{\mathbf{X}})(\mathbf{K}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) + \sigma_n \mathbf{I})^{-1}\mathrm{d}\mathbf{X}$, $\mathbb{VAR}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_t] = \mathbf{K}(\tilde{\mathbf{x}}_t, \tilde{\mathbf{x}}_t) - \mathbf{K}(\tilde{\mathbf{x}}_t, \tilde{\mathbf{X}})(\mathbf{K}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}) + \sigma_n \mathbf{I})^{-1}\mathbf{K}(\tilde{\mathbf{X}}, \tilde{\mathbf{x}}_t)$. Assume initially $\mathbf{x}_0$ is deterministic, the state distribution at $t = 0 + \mathrm{d}t$ is $\mathrm{p}(\mathbf{x}_t) \sim \mathcal{N}(\mathbf{x}_0 + \mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_0], \mathbb{VAR}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_0])$. When propagating the GP-based dynamics over a trajectory of time horizon T, the input state-control pair $\tilde{\mathbf{x}}_t$ becomes uncertain with a Gaussian distribution. Here we define the joint distribution over state-control pair at t as $\mathrm{p}(\tilde{\mathbf{x}}_t) = \mathrm{p}(\mathbf{x}_t, \mathbf{u}_t) \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$. Thus the distribution over state transition becomes $\mathrm{p}(\mathrm{d}\mathbf{x}_t) = \int \mathrm{p}(\mathbf{f}(\tilde{\mathbf{x}}_t)|\tilde{\mathbf{x}}_t)\mathrm{p}(\tilde{\mathbf{x}}_t)\mathrm{d}\tilde{\mathbf{x}}_t$. Generally, this predictive distribution cannot be computed analytically because the nonlinear mapping of an input Gaussian distribution lead to a non-Gaussian predictive distribution. However, the predictive distribution can be approximated by a Gaussian $\mathrm{p}(\mathrm{d}\mathbf{x}_t) \sim \mathcal{N}(\mathrm{d}\boldsymbol{\mu}_t, \mathrm{d}\boldsymbol{\Sigma}_t)$. Thus the state distribution at $t + \mathrm{d}t$ is also a Gaussian $\mathcal{N}(\boldsymbol{\mu}_{t+\mathrm{d}t}, \boldsymbol{\Sigma}_{t+\mathrm{d}t})$ [7]

$$\boldsymbol{\mu}_{t+\mathrm{d}t} = \boldsymbol{\mu}_t + \mathrm{d}\boldsymbol{\mu}_t, \qquad \boldsymbol{\Sigma}_{t+\mathrm{d}t} = \boldsymbol{\Sigma}_t + \mathrm{d}\boldsymbol{\Sigma}_t + \mathbb{COV}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathbf{x}_t, \mathrm{d}\mathbf{x}_t] + \mathbb{COV}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathrm{d}\mathbf{x}_t, \mathbf{x}_t]. \tag{9}$$

Given an input joint distribution $\mathcal{N}(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t)$, we employ the moment matching approach [12][7] to compute the posterior GP. The predictive mean $\mathrm{d}\boldsymbol{\mu}_t$ is evaluated as

$$\mathrm{d}\boldsymbol{\mu}_t = \mathbb{E}_{\tilde{\mathbf{x}}_t}\big[\mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_t]\big] = \int \mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_t]\mathcal{N}\big(\tilde{\boldsymbol{\mu}}_t, \tilde{\boldsymbol{\Sigma}}_t\big)\mathrm{d}\tilde{\mathbf{x}}_t.$$

Next, we compute the predictive covariance matrix

$$\mathrm{d}\boldsymbol{\Sigma}_t = \begin{bmatrix} \mathbb{VAR}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathrm{d}\mathbf{x}_{t_1}] & \dots & \mathbb{COV}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathrm{d}\mathbf{x}_{t_n}, \mathrm{d}\mathbf{x}_{t_1}] \\ \vdots & \ddots & \vdots \\ \mathbb{COV}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathrm{d}\mathbf{x}_{t_1}, \mathrm{d}\mathbf{x}_{t_n}] & \dots & \mathbb{VAR}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathrm{d}\mathbf{x}_{t_n}] \end{bmatrix},$$

where the variance term on the diagonal for output dimension i is obtained as

$$\mathbb{VAR}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathrm{d}\mathbf{x}_{t_i}] = \mathbb{E}_{\tilde{\mathbf{x}}_t}\big[\mathbb{VAR}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_{t_i}]\big] + \mathbb{E}_{\tilde{\mathbf{x}}_t}\big[\mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_{t_i}]^2\big] - \mathbb{E}_{\tilde{\mathbf{x}}_t}\big[\mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_{t_i}]\big]^2, \tag{10}$$

and the off-diagonal covariance term for output dimension i, j is given by the expression

$$\mathbb{COV}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathrm{d}\mathbf{x}_{t_i}, \mathrm{d}\mathbf{x}_{t_j}] = \mathbb{E}_{\tilde{\mathbf{x}}_t}\big[\mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_{t_i}]\mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_{t_j}]\big] - \mathbb{E}_{\tilde{\mathbf{x}}_t}[\mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_{t_i}]]\mathbb{E}_{\tilde{\mathbf{x}}_t}[\mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_{t_j}]]. \tag{11}$$

The input-output cross-covariance is formulated as

$$\mathbb{COV}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\tilde{\mathbf{x}}_t, \mathrm{d}\mathbf{x}_t] = \mathbb{E}_{\tilde{\mathbf{x}}_t}\left[\tilde{\mathbf{x}}_t \mathbb{E}_{\mathbf{f}}[\mathrm{d}\mathbf{x}_t]^{\mathrm{T}}\right] - \mathbb{E}_{\tilde{\mathbf{x}}_t}[\tilde{\mathbf{x}}_t]\mathbb{E}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathrm{d}\mathbf{x}_t]^{\mathrm{T}}. \tag{12}$$

$\mathbb{COV}_{\mathbf{f},\tilde{\mathbf{x}}_t}[\mathbf{x}_t, \mathrm{d}\mathbf{x}_t]$ can be easily obtained as a sub-matrix of (12). The kernel or hyper-parameters $\Theta = (\sigma_n, \sigma_s, \mathbf{W})$ can be learned by maximizing the log-likelihood of the training outputs given the inputs.

All mean and variance terms can be computed analytically. The hyper-parameters $\sigma_n, \sigma_s, \mathbf{W}$ can be learned by maximizing the log-likelihood of the training outputs given the inputs [13]. Given the transition probability $\mathrm{p}(\mathbf{x}_{t+\mathrm{d}t}|\mathbf{x}_t)$ (9), we now introduce a novel formulation of path integral control based on the GP representation. Firstly we reformulate the desirability (7) as

$$\begin{aligned}
\Psi_t &= \int \mathrm{p}\Big(\tau_t|\mathbf{x}_t\Big) \exp\Big(-\frac{1}{\lambda}(\sum_{j=t}^{T} q_j \mathrm{d}t)\Big)\mathrm{d}\tau_t \\
&= \int ... \underbrace{\int \mathrm{p}\Big(\mathbf{x}_{T-\mathrm{d}t}|\mathbf{x}_{T-2\mathrm{d}t}\Big) \exp\Big(-\frac{1}{\lambda} q_{T-\mathrm{d}t}\mathrm{d}t\Big) \underbrace{\int \mathrm{p}\Big(\mathbf{x}_T|\mathbf{x}_{T-\mathrm{d}t}\Big) \exp\Big(-\frac{1}{\lambda} q_T \mathrm{d}t\Big)\mathrm{d}\mathbf{x}_T}_{\Psi_{T-\mathrm{d}t}} \mathrm{d}\mathbf{x}_{T-\mathrm{d}t}}_{\Psi_{T-2\mathrm{d}t}} ...\mathrm{d}\mathbf{x}_{t+\mathrm{d}t} \\
&= \int \mathrm{p}\Big(\mathbf{x}_{t+\mathrm{d}t}|\mathbf{x}_t\Big) \exp\Big(-\frac{1}{\lambda} q_{t+\mathrm{d}t}\mathrm{d}t\Big) \underbrace{\int \mathrm{p}\Big(\mathbf{x}_{t+2\mathrm{d}t}|\mathbf{x}_{t+\mathrm{d}t}\Big) \exp\Big(-\frac{1}{\lambda} q_{t+2\mathrm{d}t}\mathrm{d}t\Big)\Psi_{t+2\mathrm{d}t}\mathrm{d}\mathbf{x}_{t+2\mathrm{d}t}}_{\Psi_{t+\mathrm{d}t}} \mathrm{d}\mathbf{x}_{t+\mathrm{d}t} \\
&= \mathbb{E}_{\mathrm{p}(\mathbf{x}_{t+\mathrm{d}t}|\mathbf{x}_t)}\Big[\exp(-\frac{1}{\lambda} q_{t+\mathrm{d}t}\mathrm{d}t)\Psi_{t+\mathrm{d}t}\Big].
\end{aligned}$$

The desirability Ψ_t can be evaluated recursively as above. Since the exponential transformation of the cost $\exp(-\frac{1}{\lambda} q_t \mathrm{d}t)$ is an unnormalized Gaussian $\mathcal{N}(\mathbf{x}_t^{goal}, \frac{2\lambda}{\mathrm{d}t}\mathbf{Q}^{-1})$. To obtain Ψ_t, which is an expectation taken with respect to path from t to T, firstly we compute the one-step desirability

$$\begin{aligned}
\Psi_{T-\mathrm{d}t} =& \mathbb{E}_{\mathrm{p}(\mathbf{x}_T|\mathbf{x}_{T-\mathrm{d}t})}\Big[\exp\Big(-\frac{1}{\lambda} q_T \mathrm{d}t\Big)\Big] \\
=& \int \mathrm{p}\Big(\mathbf{x}_T|\mathbf{x}_{T-\mathrm{d}t}\Big) \exp\Big(-\frac{1}{\lambda} q_T \mathrm{d}t\Big)\mathrm{d}\mathbf{x}_T \\
=& \int \mathrm{p}\Big(\mathbf{x}_T|\mathbf{x}_{T-\mathrm{d}t}\Big) \exp\Big(-\frac{\mathrm{d}t}{\lambda}(\mathbf{x}_T - \mathbf{x}_T^{goal})^{\mathrm{T}}\mathbf{Q}(\mathbf{x}_T - \mathbf{x}_T^{goal})\Big)\mathrm{d}\mathbf{x}_T \\
=& \underbrace{\Big|\mathbf{I} + \frac{\mathrm{d}t}{2\lambda}\mathbf{\Sigma}_T\mathbf{Q}\Big|^{-\frac{1}{2}}}_{\mathcal{S}} \exp\Big(-\frac{1}{2}(\boldsymbol{\mu}_T - \mathbf{x}_T^{goal})^{\mathrm{T}} \underbrace{\frac{\mathrm{d}t}{2\lambda}\mathbf{Q}(\mathbf{I} + \frac{\mathrm{d}t}{2\lambda}\lambda\mathbf{\Sigma}_T\mathbf{Q})^{-1}}_{\mathcal{Q}}(\boldsymbol{\mu}_T - \mathbf{x}_T^{goal})\Big) \\
=& \mathcal{S}\exp\Big(-\frac{1}{2}(\boldsymbol{\mu}_T - \mathbf{x}_T^{goal})^{\mathrm{T}}\mathcal{Q}(\boldsymbol{\mu}_T - \mathbf{x}_T^{goal})\Big).
\end{aligned}$$

The above one-step analytic solution is applied to evaluate the desirability Ψ_t recursively (i.e., compute $\Psi_{T-2\mathrm{d}t}, \ldots, \Psi_{t+\mathrm{d}t}, \Psi_t$). The gradient of the desirability with respect to the state can be computed using chain-rule

$$\nabla_{\mathbf{x}_t}\psi_t = \frac{\partial \Psi_{\mathbf{x}_t}}{\partial \mathrm{p}(\mathbf{x}_T)}\frac{\partial \mathrm{p}(\mathbf{x}_T)}{\partial \mathbf{x}_t} = \frac{\partial \Psi_{\mathbf{x}_t}}{\partial \boldsymbol{\mu}_T}\frac{\partial \boldsymbol{\mu}_T}{\partial \mathbf{x}_t} + \frac{\partial \Psi_{\mathbf{x}_t}}{\partial \mathbf{\Sigma}_T}\frac{\partial \mathbf{\Sigma}_T}{\partial \mathbf{x}_t},$$

where

$$\frac{\partial \boldsymbol{\mu}_T}{\partial \mathbf{x}_t} = \Big(\frac{\partial \boldsymbol{\mu}_T}{\partial \boldsymbol{\mu}_{T-\mathrm{d}t}}\frac{\partial \boldsymbol{\mu}_{T-\mathrm{d}t}}{\partial \mathrm{p}(\mathbf{x}_{T-2\mathrm{d}t})} + \frac{\partial \boldsymbol{\mu}_T}{\partial \mathbf{\Sigma}_{T-\mathrm{d}t}}\frac{\partial \mathbf{\Sigma}_{T-\mathrm{d}t}}{\partial \mathrm{p}(\mathbf{x}_{T-2\mathrm{d}t})}\Big) \cdots \frac{\partial \mathrm{p}(\mathbf{x}_{t+\mathrm{d}t})}{\partial \mathbf{x}_t},$$

and $\frac{\partial \mathbf{\Sigma}_T}{\partial \mathbf{x}_t}$ can be computed similarly. We find all partial derivatives analytically, therefore the computational efficiency is significantly improved compared to the model-free PI control framework. Finally, the optimal control is obtained as (8).

4.2 Iterative control improvement scheme

The model-based PI framework introduced in 4.1 relies on samples from the uncontrolled diffusion processes to learn the desirability Ψ_t. However, for control tasks of high-dimensional, complex systems, this sampling strategy is inefficient in practice and degenerates control performances [10]. In

this section we develop an iterative scheme to improve the applicability of the proposed framework. We start our analysis with the stochastic representation of the solution to the backward Chapman Kolmogorov PDE, then apply the Randon Nikodym derivative [14] for Markov diffusion process

$$\Psi_t = \int \exp\Big(-\frac{1}{\lambda}\sum_{j=t}^{T} q_j \mathrm{d}t\Big)\Psi_T \mathrm{dp}(\mathbf{x}_T|\mathbf{x}_t) = \int \exp\Big(-\frac{1}{\lambda}\sum_{j=t}^{T} q_j \mathrm{d}t\Big)\Psi_T \xi \mathrm{dp}(\mathbf{x}_T|\mathbf{x}_t,\mathbf{u}_t), \quad (13)$$

where $\mathrm{dp}(\mathbf{x}_t|\mathbf{x}_t)$ is the path integral representation of the uncontrolled diffusion process $\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t)+\mathbf{B}(\mathbf{x}_t)\mathrm{d}\omega$, while $\mathrm{dp}(\mathbf{x}_t|\mathbf{x}_t,\mathbf{u}_t)$ is the path integral that corresponds to the controlled diffusion process $\mathrm{d}\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t) + \mathbf{G}(\mathbf{x}_t)\mathbf{u}_t^k \mathrm{d}t + \mathbf{B}(\mathbf{x}_t)\mathrm{d}\omega$, where the superscript k is the iteration index. The controlled transition probability $\mathrm{p}(\mathbf{x}_{t+\mathrm{d}t}|\mathbf{x}_t,\mathbf{u}_t)$ is computed similarly as $\mathrm{p}(\mathbf{x}_{t+\mathrm{d}t}|\mathbf{x}_t)$ in section 4.1 (we assume deterministic $\mathbf{u}_t$ in this paper). The ratio of the two probability ξ is the Radon-Nikodym for diffusion processes, which is formulated as

$$\xi = \frac{\mathrm{dp}(\mathbf{x}_T|\mathbf{x}_t)}{\mathrm{dp}(\mathbf{x}_T|\mathbf{x}_t,\mathbf{u}_t)} = \exp\Big(-\frac{1}{2\lambda}\sum_{j=t}^{T}(\mathbf{u}_j^{\mathrm{T}}\mathbf{G}_j^{\mathrm{T}}\mathbf{W}_j^{-1}\mathbf{G}_j\mathbf{u}_j\mathrm{d}t + 2\mathbf{u}_j^{\mathrm{T}}\mathbf{G}_j^{\mathrm{T}}\mathbf{W}_j^{-1}\mathbf{B}_j\mathrm{d}\omega)\Big), \quad (14)$$

where $\mathbf{W}_j = \mathbf{G}_j\mathbf{R}^{-1}\mathbf{G}_j^{\mathrm{T}}$. The desirability will take the form $\Psi_t^k = \mathbb{E}_{\mathrm{p}(\mathbf{x}_T|\mathbf{x}_t,\mathbf{u}_t^k)}\Big[\exp\Big(-\frac{1}{\lambda}\sum_{j=t}^{T}\tilde{q}_j^k\mathrm{d}t\Big)\Psi_T\Big]$, where the path cost $\tilde{q}_j^k = q_j^k + \frac{1}{2}(\mathbf{u}_j^k)^{\mathrm{T}}\mathbf{G}_j^{\mathrm{T}}\mathbf{W}_j^{-1}\mathbf{G}_j\mathbf{u}_j^k + (\mathbf{u}_j^k)^{\mathrm{T}}\mathbf{G}_j^{\mathrm{T}}\mathbf{W}_j^{-1}\mathbf{B}_j\frac{\mathrm{d}\omega}{\mathrm{d}t}$, The gradient of the desirability with respect to the state is evaluated as

$$\nabla_{\mathbf{x}}\Psi_t^k = \nabla_{\mathbf{x}}\mathbb{E}_{\mathrm{p}(\mathbf{x}_T|\mathbf{x}_t,\mathbf{u}_t^k)}\Big[\Big(\exp\Big(-\frac{1}{\lambda}\sum_{j=t}^{T}\tilde{q}_j^k\mathrm{d}t\Big)\Psi_T\Big)\Big] = \frac{1}{\lambda}\Psi_t^k\mathbf{W}_t^{-1}\mathbf{G}_t\mathbf{u}_t^k + \Psi_t^k\frac{\nabla_{\mathbf{x}}\Phi_t^k}{\Phi_t^k},$$

where $\Phi_t^k = \mathbb{E}_{\mathrm{p}(\mathbf{x}_T|\mathbf{x}_t,\mathbf{u}_t^k)}\Big[\exp\Big(-\frac{1}{\lambda}\sum_{j=t}^{T}q_j^k\mathrm{d}t\Big)\Psi_T\Big]$. Finally the optimal control at iteration $k+1$ is obtained as

$$\hat{\mathbf{u}}_t^{k+1} = \lambda\mathbf{R}^{-1}\mathbf{G}_t^{\mathrm{T}}\Big(\frac{\nabla_{\mathbf{x}}\Psi_t^k}{\Psi_t^k}\Big) = \hat{\mathbf{u}}_t^k + \lambda\mathbf{R}^{-1}\mathbf{G}_t^{\mathrm{T}}\Big(\frac{\nabla_{\mathbf{x}}\Phi_t^k}{\Phi_t^k}\Big). \quad (15)$$

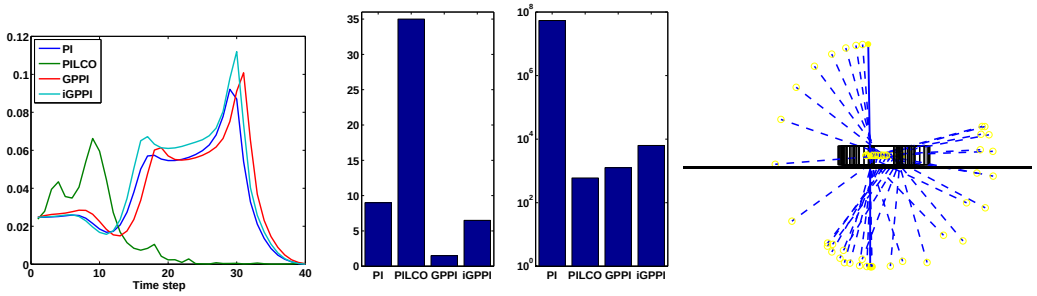
Similar to the case when sampling from the uncontrolled dynamics, $\Phi_t^k, \nabla_{\mathbf{x}}\Phi_t^k$ are obtained by computing integrals recursively and all integrals can be evaluated analytically.

5 Experimental Results

We evaluate the proposed framework in two nontrivial simulated examples: i) cart-pole (CP) swing-up; ii) cart-double pendulum (CDIP) swing-up. We compare the proposed method with the iterative **PI** [10][11] and **PILCO** [6][7], which have demonstrated impressive efficiency and applicability in robotics among model-free and model-based RL/control approaches. We implement our proposed framework in two ways: **GPPI** and **iGPPI** denote the framework based on samples from the uncontrolled dynamics (4.1) and the iterative scheme (4.2), respectively.

Cart-pole swing-up: The CP system is underactuated with 4 state dimensions, 2 degrees of freedom and 1 control input. The target states are inverted position for the pendulum and zero velocity for both cart and pendulum. Fig. 1a and 1b show comparisons of GPPI and iGPPI with PI and PILCO. Both GPPI and iGPPI perform similarly as PI in terms of optimal control, but GPPI and iGPPI require significantly less sampled data (less interactions with the physical system), and less total time to complete the task than PI. PILCO performs very well in terms of data-efficiency, but it is the slowest among all 4 methods. Fig. 1c depicts the postures of CP swing-up using GPPI.

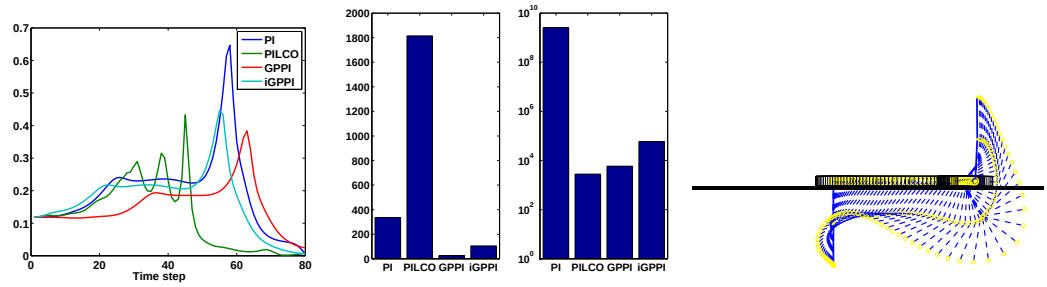
Cart-double inverted pendulum swing-up: The CDIP swing-up is a challenging control task. The system is highly underactuated with 6 state dimensions, 3 degrees of freedom and only 1 control input. The target states are inverted positions for both pendulums and zero velocities for pendulums and the cart. The cost comparison is shown in Fig. 2a. iGPPI outperforms GPPI in terms of terminal cost. GPPI relies on samples from uncontrolled dynamics, while iGPPI updates optimal controls based on samples from controlled dynamics. This iterative strategy shows improved performance for more challenging tasks such as CDIP swing-up. As shown in Fig. 2b, PILCO offers impressive data-efficiency but slow learning speed, while PI costs significantly more sampled data than other approaches. Fig. 2c depicts the postures of CDIP swing-up using iGPPI.



(a) (b) (c)

Figure 1: Cart-pole swing-up task. (a) Cost comparison. (b) Efficiency comparison. The left subfigure shows total computational time required to complete the task (minute). The right subfigure shows the total number of sampled data point required. (c) Postures using GPPI.

Comparative Analysis: Compared to the sampling-based PI, the proposed GPPI/iGPPI are more efficient in terms of data-consumption and learning speed thanks to the analytic representation of path integrals. Compared to PILCO, GPPI/iGPPI learn optimal controls without any policy parameterization and do not rely on any extra optimizer to find the optimal controller, therefore they show significant improvement in terms of learning speed. PILCO shows better performance in terms of total cost reduction over the trajectory. The major reason for this difference is that PI-related approaches are applied in receding horizon modes (e.g., apply current optimal control $\mathbf{u}_t$ then compute $\mathbf{u}_{t+\mathrm{d}t}$) while PILCO optimizes the whole trajectory at every trial. Although GPPI demonstrates higher efficiency for simpler tasks (such as the CP), iGPPI is more applicable to challenging tasks (such as the CDIP) for which sampling form uncontrolled dynamics is insufficient.



(a) (b) (c)

Figure 2: Cart-double inverted pendulum swing-up task. (a) Cost comparison. (b) Efficiency comparison. The left subfigure shows total computational time required to complete the task (minute). The right subfigure shows the total number of sampled data point required. (c) Postures using iGPPI.

6 Conclusions

Motivated by the limitations of sampling-based PI control, we introduced a novel model-based PI control framework. Grounded in the stochastic Hamilton-Jacobi-Bellman equation, the Feynman-Kac formula and Gaussian processes, the proposed approach learns Bayesian nonparametric models and time-varying optimal controls autonomously from sampled data. Thanks to the probabilistic representation of the dynamics model and analytic computations of the optimal controls, the proposed framework showed encouraging learning efficiency compared to the sampling-based PI control and a state-of-the-art GP-based policy search method.

References

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

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

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

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

[5] E. Todorov. Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483, 2009.

[6] M. Deisenroth and C. Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on Machine Learning*, pages 465–472, 2011.

[7] M. Deisenroth, D. Fox, and C. Rasmussen. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transsactions on Pattern Analysis and Machine Intelligence*, 27:75–90, 2014.

[8] H. J. Kappen. Linear theory for control of nonlinear stochastic systems. *Phys Rev Lett*, 95:200–201, 2005.

[9] H. J. Kappen. An introduction to stochastic control theory, path integrals and reinforcement learning. *AIP Conference Proceedings*, 887(1), 2007.

[10] E. Theodorou. *Iterative Path Integral Stochastic Optimal Control: Theory and Applications to Motor Control*. PhD thesis, Los Angeles, CA, USA, 2011.

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

[12] J. Quinonero Candela, A. Girard, J. Larsen, and C. E. Rasmussen. Propagation of uncertainty in bayesian kernel models-application to multiple-step ahead forecasting. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[13] C.K.I Williams and C.E. Rasmussen. *Gaussian processes for machine learning*. MIT Press, 2006.

[14] C. Gardiner. *Stochastic methods*. Springer, 2010.