

Optimal Control of Probabilistic Dynamics Models via Mean Hamiltonian Minimization

David Leeftink¹, Çağatay Yıldız², Steffen Ridderbusch³, Max Hinne¹, Marcel van Gerven¹

Abstract—Without exact knowledge of the true system dynamics, optimal control of non-linear continuous-time systems requires careful treatment under epistemic uncertainty. In this work, we translate a probabilistic interpretation of the Pontryagin maximum principle to the challenge of optimal control with learned probabilistic dynamics models. Our framework provides a principled treatment of epistemic uncertainty by minimizing the mean Hamiltonian with respect to a posterior distribution over the system dynamics.

We propose a multiple shooting numerical method that leverages mean Hamiltonian minimization and is scalable to large-scale probabilistic dynamics models, including ensemble neural ordinary differential equations. Comparisons against other baselines in online and offline model-based reinforcement learning tasks show that our probabilistic Hamiltonian approach leads to reduced trial costs in offline settings and achieves competitive performance in online scenarios. By bridging optimal control and reinforcement learning, our approach offers a principled and practical framework for controlling uncertain systems with learned dynamics.

I. INTRODUCTION

Optimal decision-making under uncertainty is a fundamental challenge in learning-based control, especially when the system dynamics are unknown and must be inferred from limited, noisy data. In such settings it is crucial to achieve high data-efficiency, as real-world experiments often come with high costs or risks. To address this, one has to effectively handle epistemic uncertainty, arising from incomplete data, which critically shapes both the robustness of the controller and the balance between exploration and exploitation in learning.

A common approach for decision making under uncertainty involves minimizing the mean cost over the posterior distribution of system dynamics. However, this optimization can become challenging when the posterior distribution is multi-modal or exhibits significant variance. While black-box stochastic optimization methods are often employed to

avoid local minima, they overlook the inherent structure of the optimal control problem [1].

Pontryagin’s maximum principle (PMP) offers a compelling alternative. By defining necessary conditions for optimality, PMP indirectly characterizes optimal control inputs, providing a powerful lens into the structure of optimal solutions [2]. As a cornerstone of optimal control theory, PMP has driven significant advances in deterministic and stochastic control settings [3], [4]. Its ability to leverage the inherent structure of the optimal control problem enables computationally efficient optimization and insights into optimal behavior.

Related work. Optimal control theory provides foundational principles for learning-based control, guiding optimal action selection under uncertainty. Techniques like the iterative Linear Quadratic Regulator (iLQR) [5] and path integral (PI) control [6] have been widely adopted to design optimal policies in uncertain or learning-driven settings. Originally developed for deterministic systems, PMP has also been extended to stochastic [4], robust [7], and, more recently, learning-based control [8], [9]. While these PMP-based methods primarily address deterministic systems or those with aleatoric uncertainty, recent theoretical advances have established optimality conditions for systems with epistemic uncertainty in their dynamics. In [10], conditions are derived for minimizing average terminal costs in general systems with parametric uncertainty, while in [11] similar conditions are developed for control-affine ensemble systems. Our work builds on this foundation by operationalizing the principle of mean Hamiltonian minimization for learned probabilistic dynamics.

Our approach closely aligns with Bayesian control and reinforcement learning (RL) paradigms, where optimal policies explicitly account for uncertainty [12], [13]. The probabilistic inference for learning and control (PILCO) framework [14] exemplifies this by learning a Gaussian process (GP) dynamics model and computing policy gradients via closed-form Gaussian approximations. Similarly, Bayesian extensions to model predictive control (MPC) have shown efficacy by planning under uncertain dynamics, using PMP conditions to propagate Gaussian uncertainty over a moving horizon [15].

In deep model-based reinforcement learning (MBRL), methods like probabilistic ensembles with trajectory sampling (PETS) [16] have advanced high-dimensional task solving through uncertainty-aware planning. Our work bridges these advances with continuous-time RL, where discrete-time measurements inform continuous-time dynamics and control [17], [18]. By integrating PMP with these Bayesian and

*This publication is part of the project ROBUST: Trustworthy AI-based Systems for Sustainable Growth with project number KICH3.LTP.20.006, which is (partly) financed by the Dutch Research Council (NWO), ASMP, and the Dutch Ministry of Economic Affairs and Climate Policy (EZK) under the program LTP KIC 2020-2023. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

¹D. Leeftink, M. Hinne and M. van Gerven are with the Department of Machine Learning and Neural Computing, Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525XZ Nijmegen, the Netherlands. ²Ç. Yıldız is with the Cluster of Excellence Machine Learning, University of Tübingen, 72076 Tübingen, Germany, and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC number 2064/1 ³S. Ridderbusch is with the Department of Engineering, Control Group, University of Oxford, OX1 3PJ Oxford, UK. (Contact: david.leeftink@ru.nl)

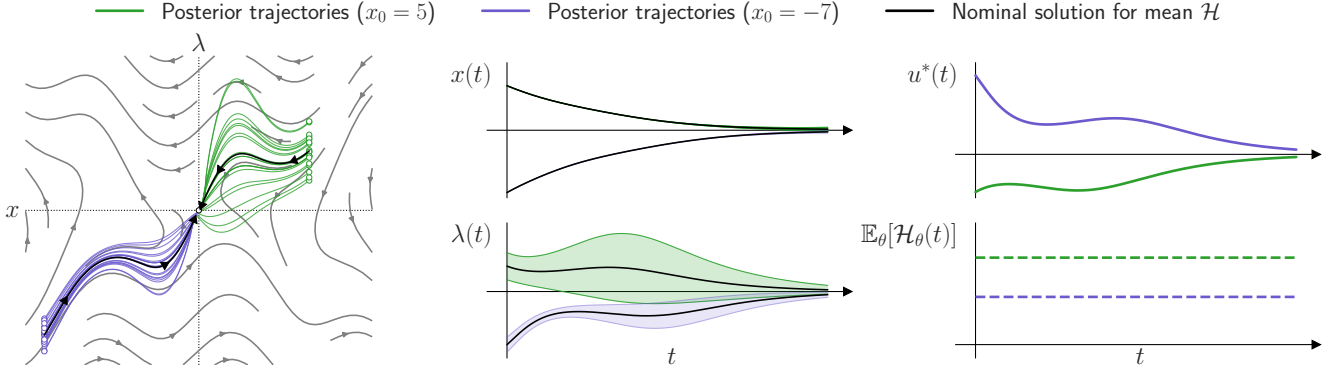


Fig. 1: Optimal control of an uncertain dynamical system using the mean Hamiltonian minimization principle. **(Left)** The Hamiltonian vector field for a system with dynamics depicting state and co-states of the dynamics function $f_\theta(x, u) = \theta \sin(x) + u$. The thin lines show posterior trajectories sampled from the dynamics belief $p(\theta)$, while the bold black line shows the nominal solution of the control function that minimizes the mean Hamiltonian. **(Middle & Right)** The evolution of the state $x(t)$, co-state $\lambda(t)$, optimal control $u^*(t)$, and the mean Hamiltonian $\mathbb{E}_\theta[\mathcal{H}_\theta]$. The optimal control function, derived from minimizing the mean Hamiltonian, successfully steers the system to the origin. As a necessary condition for optimality, the mean Hamiltonian is constant under the optimal control.

continuous-time approaches, we propose a novel framework for optimal planning under epistemic uncertainty, tailored to ensemble-based RL, offering a robust and principled approach to controlling complex, uncertain systems.

Contribution. Our contributions are threefold. First, we formulate the optimal control problem for learned probabilistic dynamics models within the PMP framework, introducing a principled approach to planning based on mean Hamiltonian minimization that explicitly addresses epistemic uncertainty. Second, we develop a practical numerical method combining forward multiple shooting with moving horizon optimization, tailored to probabilistic ordinary differential equation (ODE) models. Third, we validate our approach experimentally, demonstrating improved or competitive performance compared to state-of-the-art methods in both offline and online RL scenarios, highlighting its potential for efficient learning and robust decision-making.

The application of our framework to a representative optimal control problem with uncertain dynamics is illustrated in Figure 1, which visualizes the full probabilistic solution, from the state-costate trajectories to the final optimal control.

II. PROBLEM SETTING

We study a continuous-time deterministic differential equation \mathbf{f} , whose solution is given by:

$$\mathbf{x}(t) = \mathbf{x}_0 + \int_{t_0}^{t_f} \mathbf{f}(\mathbf{x}(\tau), \mathbf{u}(\tau), \tau) d\tau, \quad (\text{II.1})$$

where $\mathbf{x}_0 \in \mathcal{X} \subset \mathbb{R}^{d_x}$ is the initial state, $\mathbf{u}: [t_0, t_f] \rightarrow \mathcal{U} \subset \mathbb{R}^{d_u}$ is the control input, and d_x and d_u denote the dimensions of the state space and the control input. The objective is to find the optimal control input \mathbf{u}^* , that is, the input that minimizes the cost function:

$$C(\mathbf{u}) := \int_{t_0}^{t_f} L(\mathbf{x}(\tau), \mathbf{u}(\tau)) d\tau + \Phi(\mathbf{x}(t_f)), \quad (\text{II.2})$$

where $\Phi(\mathbf{x})$ is the terminal cost associated with ending in state $\mathbf{x}(t_f)$ at time t_f and $L(\mathbf{x}, \mathbf{u})$ is the state cost associated with taking action \mathbf{u} in state \mathbf{x} . Both the terminal cost and state cost are assumed differentiable with respect to their inputs. We are interested in solving the following optimal control problem:

$$\begin{aligned} \mathbf{u}^*(t) &:= \underset{\mathbf{u} \in \mathcal{U}}{\operatorname{argmin}} C(\mathbf{u}) \\ \text{s.t. } \dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t), \quad \text{for } t \in [t_0, t_f] \\ \mathbf{x}(t_0) &= \mathbf{x}_0. \end{aligned} \quad (\text{II.3})$$

Although the true function \mathbf{f} is unknown, it can be evaluated by interacting with the system and collecting data over the course of N trials over the time horizon t_f . Let $\mathcal{D}_n := \{(t_i^n, \mathbf{y}_i^n, \mathbf{u}_i^n)\}$ denote the data collected in trial $n \in [1, N]$. Here, $t_i^n \in [t_0, t_f]$ is the time point corresponding to observation i in trial n , $\mathbf{x}_i^n := \mathbf{x}(t_i^n)$ is the true system state at time t_i^n in trial n , $\mathbf{y}_i^n \sim \mathcal{N}(\mathbf{x}_i^n, \sigma_n^2 \mathbf{I})$ is a noisy measurement of the state with i.i.d. Gaussian noise with variance σ_n^2 , and \mathbf{u}_i^n is the applied control input corresponding to observation i for trial n , assumed to be exactly known. It is assumed that the true control function can be reconstructed for $t \in [t_0, t_f]$ using interpolation techniques.

In MBRL, a dynamics model is learned based on previous interactions with the system. In the online variant, this dataset consists initially of only a single trial generated by a random controller. Then iteratively, a model is trained on the available data, the optimal action is determined using the model, and the newly observed trajectory is added to the data.

A. Probabilistic dynamics models

When learning a dynamics model from limited data, it is crucial to employ efficient exploration strategies that separate epistemic uncertainty, associated with reducible uncertainty

due to lack of knowledge of the true system, from aleatoric uncertainty, associated with irreducible uncertainty due to inherent randomness such as sensor or process noise. A general approach for quantifying uncertainty is to assign a prior distribution over the parameters and a likelihood to the possibly noisy observations, and use Bayes' theorem to obtain the posterior distribution over model parameters:

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}), \quad (\text{II.4})$$

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{i=1}^T \mathcal{N}(\mathbf{y}_i \mid \mathbf{x}_{\boldsymbol{\theta}}(t_i), \sigma^2 \mathbf{I}), \quad (\text{II.5})$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \quad (\text{II.6})$$

Here, T is the number of discrete-time observations throughout $[t_0, t_f]$ and $\boldsymbol{\theta} \in \Theta = \mathbb{R}^p$ are the parameters. For a given realization $\boldsymbol{\theta}$, the following continuous-time model is defined:

$$\dot{\mathbf{x}}_{\boldsymbol{\theta}}(t) = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_{\boldsymbol{\theta}}(t), \mathbf{u}(t), t), \quad (\text{II.7})$$

where $\mathbf{f}_{\boldsymbol{\theta}}$ is differentiable in its first two arguments. The posterior distribution can then be estimated using Bayesian sampling methods such as Hamiltonian Monte Carlo (HMC) [19].

B. Mean cost objective

A common objective in MBRL is the expected cost over the parameter distribution of the dynamics model which reflects the epistemic uncertainty:

$$\bar{C}(\mathbf{u}) := \mathbb{E}_{\boldsymbol{\theta}} [C_{\boldsymbol{\theta}}(\mathbf{u})], \quad (\text{II.8})$$

where

$$C_{\boldsymbol{\theta}}(\mathbf{u}) := \int_0^{t_f} L(\mathbf{x}_{\boldsymbol{\theta}}(\tau), \mathbf{u}(\tau)) d\tau + \Phi(\mathbf{x}_{\boldsymbol{\theta}}(t_f)),$$

and $\mathbb{E}_{\boldsymbol{\theta}}$ is shorthand for $\mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{D})}$. This results in the optimal control problem:

$$\mathbf{u}^*(t) := \underset{\mathbf{u} \in \mathcal{U}}{\operatorname{argmin}} \bar{C}(\mathbf{u}) \quad (\text{II.9})$$

$$\text{s.t. } \dot{\mathbf{x}}_{\boldsymbol{\theta}}(t) = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_{\boldsymbol{\theta}}(t), \mathbf{u}(t), t), \quad \text{for } t \in [t_0, t_f] \\ \mathbf{x}_{\boldsymbol{\theta}}(t_0) = \mathbf{x}_0.$$

III. PROBABILISTIC CONTROL VIA PMP

A. Pontryagin's maximum principle

The mean cost objective $\bar{C}(\mathbf{u})$ can be optimized via the *direct* method or via the *indirect* method. In the former, the control function is discretized and optimized as a real-valued vector, whereas in the latter, first the necessary optimality conditions for the continuous-time optimal control problem are described using PMP and then the control function is discretized. A key function is the Hamiltonian:

$$\mathcal{H}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{u}, t) := L(\mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda}^\top \mathbf{f}(\mathbf{x}, \mathbf{u}, t),$$

where $\mathcal{H}: \mathcal{X} \times \mathbb{R}^{d_x} \times \mathcal{U} \times \mathbb{R} \rightarrow \mathbb{R}$ and $\boldsymbol{\lambda} \in \mathbb{R}^{d_x}$ are the co-states (or adjoints). For a given function $\mathbf{u}(t)$, the state

and costate equations are given as:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t), \quad (\text{III.1})$$

$$\dot{\boldsymbol{\lambda}}(t) = -\nabla_{\mathbf{x}} \mathcal{H}(\mathbf{x}(t), \boldsymbol{\lambda}(t), \mathbf{u}(t), t), \quad (\text{III.2})$$

$$\boldsymbol{\lambda}(t_f) = \nabla \Phi(\mathbf{x}(t_f)), \quad (\text{III.3})$$

for $t \in [t_0, t_f]$, where the notation $(\nabla_y f)_{i,j} := \partial f_j / \partial y_i$ denotes Jacobians, in this case $\nabla_{\mathbf{x}} \mathcal{H} = (\frac{\partial \mathcal{H}}{\partial x_1}, \dots, \frac{\partial \mathcal{H}}{\partial x_{d_x}})^\top$ and $\nabla \Phi = (\frac{\partial \Phi}{\partial x_1}, \dots, \frac{\partial \Phi}{\partial x_{d_x}})^\top$. Equation (III.3) is commonly known as the *transversality condition*, and provides a boundary condition for the co-states. If $\mathbf{u}^*(t)$ is an optimal solution to $C(\mathbf{u})$, the following necessary condition of optimality is derived via PMP:

$$\mathbf{u}^*(t) = \underset{\mathbf{u}}{\operatorname{argmin}} \mathcal{H}(\mathbf{x}^*(t), \boldsymbol{\lambda}^*(t), \mathbf{u}, t), \quad (\text{III.4})$$

where $\mathbf{x}^*(t)$ and $\boldsymbol{\lambda}^*(t)$ satisfy (III.1) – (III.3). We refer to (III.4) as the *minimum principle*. The indirect approach thus provides candidate solutions for the optimal control $\mathbf{u}^*(t)$ that satisfy the necessary optimality conditions for the cost functional. Next, we extend these conditions to the probabilistic setup.

B. Probabilistic Hamiltonian

In the presence of uncertainty, the posterior distribution over system dynamics $p(\boldsymbol{\theta} \mid \mathcal{D})$ in (II.6) implicitly defines a distribution over Hamiltonian functions $p(\mathcal{H} \mid \mathcal{D})$ through its dependence on the dynamics. Sampling a realization $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathcal{D})$ yields a corresponding Hamiltonian:

$$\mathcal{H}_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{u}, t) := L(\mathbf{x}, \mathbf{u}) + \boldsymbol{\lambda}^\top \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{u}, t). \quad (\text{III.5})$$

which extends the classical Hamiltonian by incorporating the parameter-dependent dynamics $\mathbf{f}_{\boldsymbol{\theta}}$. Our goal is to find a control function $\mathbf{u}^*(t)$ that minimizes the mean cost functional $\bar{C}(\mathbf{u})$. For a single realization $\boldsymbol{\theta}$, the state and co-state equations are:

$$\dot{\mathbf{x}}_{\boldsymbol{\theta}}(t) = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}_{\boldsymbol{\theta}}(t), \mathbf{u}(t), t), \quad (\text{III.6})$$

$$\dot{\boldsymbol{\lambda}}_{\boldsymbol{\theta}}(t) = -\nabla_{\mathbf{x}} \mathcal{H}_{\boldsymbol{\theta}}(\mathbf{x}_{\boldsymbol{\theta}}(t), \boldsymbol{\lambda}_{\boldsymbol{\theta}}(t), \mathbf{u}(t), t), \quad (\text{III.7})$$

$$\boldsymbol{\lambda}_{\boldsymbol{\theta}}(t_f) = \nabla \Phi(\mathbf{x}_{\boldsymbol{\theta}}(t_f)). \quad (\text{III.8})$$

Here, $\mathbf{x}_{\boldsymbol{\theta}}(t)$ and $\boldsymbol{\lambda}_{\boldsymbol{\theta}}(t)$ are the state and costate trajectories for a given admissible control $\mathbf{u}(t)$ and realization $\boldsymbol{\theta}$ satisfying (III.6) – (III.8). The necessary optimality condition for a single realization $C_{\boldsymbol{\theta}}(\mathbf{u})$ follows from the minimum principle:

$$\mathbf{u}_{\boldsymbol{\theta}}^*(t) := \underset{\mathbf{u}}{\operatorname{argmin}} \mathcal{H}_{\boldsymbol{\theta}}(\mathbf{x}_{\boldsymbol{\theta}}^*(t), \boldsymbol{\lambda}_{\boldsymbol{\theta}}^*(t), \mathbf{u}, t),$$

for $t \in [t_0, t_f]$, with $\mathbf{x}_{\boldsymbol{\theta}}^*(t)$ and $\boldsymbol{\lambda}_{\boldsymbol{\theta}}^*(t)$ satisfying (III.6) – (III.8). However, optimizing for a single $\boldsymbol{\theta}$ does not address the mean cost $\bar{C}(\mathbf{u})$, which requires a control function that is optimal across the distribution of $\boldsymbol{\theta}$. To this end, we consider the *mean minimum principle*:

$$\mathbf{u}^*(t) = \underset{\mathbf{u}}{\operatorname{argmin}} \mathbb{E}_{\boldsymbol{\theta}} [\mathcal{H}_{\boldsymbol{\theta}}(\mathbf{x}_{\boldsymbol{\theta}}(t), \boldsymbol{\lambda}_{\boldsymbol{\theta}}(t), \mathbf{u}, t)], \quad (\text{III.9})$$

This condition yields a control function that is optimal in expectation, producing a family of state and co-state trajectories $\mathbf{x}_{\boldsymbol{\theta}}(t)$ and $\boldsymbol{\lambda}_{\boldsymbol{\theta}}(t)$ that satisfy (III.6) – (III.8) for each $\boldsymbol{\theta}$.

The mean minimum principle balances performance across all possible dynamics models weighted by their posterior probability, which is illustrated in Figure 2.

C. Necessary optimality condition for the mean cost

We now state the necessary condition for optimality for the mean cost $\bar{C}(\mathbf{u})$. The general principle developed in [10] and [11] requires the optimal control $\mathbf{u}^*(t)$ to minimize the mean Hamiltonian at each point in time over the (possibly bounded) set of admissible controls \mathcal{U} . Here, we restrict ourselves to the case of unconstrained control setting, in which case this simplifies to a first-order condition on the gradient using a variational approach [20].

Assumptions: we assume that $\mathcal{U} = \mathbb{R}^{d_u}$, \mathbf{f}_θ and L are continuously differentiable w.r.t. \mathbf{x} and \mathbf{u} , ensuring \mathcal{H}_θ is differentiable for all $\theta \in \Theta$, and $\nabla_{\mathbf{u}} \mathcal{H}_\theta$ satisfies standard regularity conditions for interchanging \mathbb{E}_θ with differentiation and integration over $t \in [t_0, t_f]$.

Theorem 1 (Necessary Mean Cost Optimality Condition). *Let $\mathbf{u}^*(t)$ be an optimal control that minimizes the mean cost functional $\bar{C}(\mathbf{u})$ over the set of admissible controls \mathcal{U} . Let the mean Hamiltonian be $\bar{\mathcal{H}}^* := \mathbb{E}_\theta[\mathcal{H}_\theta(\mathbf{x}_\theta(t), \boldsymbol{\lambda}_\theta(t), \mathbf{u}^*(t), t)]$, evaluated under $\mathbf{u}^*(t)$. A necessary condition for optimality is that:*

$$\nabla_{\mathbf{u}} \bar{\mathcal{H}}^* = 0 \quad \text{for all } t \in [t_0, t_f]. \quad (\text{III.10})$$

Since $\mathbf{u}^*(t)$ minimizes the mean cost, $\mathbf{x}_\theta(t)$ and $\boldsymbol{\lambda}_\theta(t)$ are not necessarily optimal for each θ .

Proof. We consider the Gateaux derivative of the mean cost:

$$\delta \bar{C}(\mathbf{u}) = \lim_{\epsilon \rightarrow 0} \frac{\bar{C}(\mathbf{u} + \epsilon \delta \mathbf{u}) - \bar{C}(\mathbf{u})}{\epsilon},$$

where $\delta \mathbf{u}(t)$ is an arbitrary variation defined over $[t_0, t_f]$. The first variation can be obtained by isolating the first order terms around the optimal solution $\mathbf{u}^*(t)$:

$$\bar{C}(\mathbf{u}^* + \epsilon \delta \mathbf{u}) - \bar{C}(\mathbf{u}^*) \approx \delta \bar{C}|_{\mathbf{u}^*(t)}(\delta \mathbf{u})\epsilon,$$

where higher-order terms are neglected as $\epsilon \rightarrow 0$. Since by definition $\mathbf{u}^*(t)$ is a minimum for \bar{C} , its first variation must be zero. We first consider the deterministic case for a realization θ , where $\mathbf{u}_\theta^*(t)$ is optimal for $C_\theta(\mathbf{u})$. A well-known result in optimal control theory states that the first variation is given as [3, Eq. (5.1.13)]:

$$\delta C_\theta|_{\mathbf{u}_\theta^*(t)}(\delta \mathbf{u}) = \int_{t_0}^{t_f} [\nabla_{\mathbf{u}} \mathcal{H}_\theta^*] \delta \mathbf{u}(t) dt,$$

where $\mathcal{H}_\theta^* := \mathcal{H}_\theta(\mathbf{x}_\theta^*(t), \boldsymbol{\lambda}_\theta^*(t), \mathbf{u}_\theta^*(t), t)$.

By the fundamental lemma of calculus of variations, it follows that:

$$\nabla_{\mathbf{u}} \mathcal{H}_\theta^* = 0, \quad \text{for } t \in [t_0, t_f].$$

Using this result, we now turn to the mean cost and interchange the derivative and expectation:

$$\delta \bar{C}(\mathbf{u}) = \delta \mathbb{E}_\theta[C_\theta(\mathbf{u})] = \mathbb{E}_\theta[\delta C_\theta(\mathbf{u})].$$

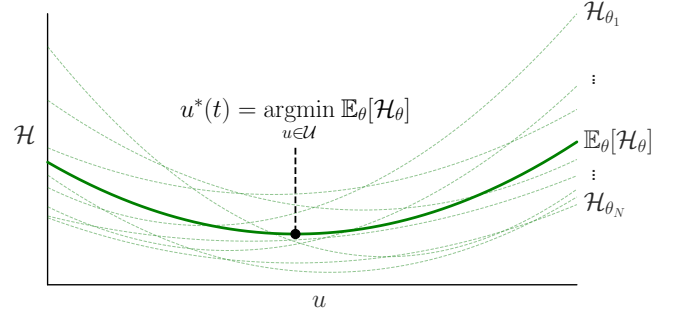


Fig. 2: Illustration of the mean Hamiltonian minimization principle. At a given time t , each of the N samples from a probabilistic dynamics model (indexed by θ) defines a Hamiltonian $\mathcal{H}_\theta := \mathcal{H}_\theta(\mathbf{x}_\theta(t), \boldsymbol{\lambda}_\theta(t), \mathbf{u}, t)$ as a function of the control \mathbf{u} (dashed lines). The optimal control \mathbf{u}^* minimizes the mean Hamiltonian, $\mathbb{E}_\theta[\mathcal{H}_\theta]$ (solid line), providing a control action that is optimal on average with respect to the posterior distribution over system dynamics.

Using the earlier result of the first variation:

$$\begin{aligned} \mathbb{E}_\theta[\delta C_\theta|_{\mathbf{u}^*(t)}(\delta \mathbf{u})] &= \mathbb{E}_\theta \left[\int_{t_0}^{t_f} \nabla_{\mathbf{u}} \mathcal{H}_\theta(\mathbf{x}_\theta(t), \boldsymbol{\lambda}_\theta(t), \mathbf{u}^*(t), t) \delta \mathbf{u}(t) dt \right] \\ &= \int_{t_0}^{t_f} \nabla_{\mathbf{u}} \mathbb{E}_\theta[\mathcal{H}_\theta(\mathbf{x}_\theta(t), \boldsymbol{\lambda}_\theta(t), \mathbf{u}^*(t), t)] \delta \mathbf{u}(t) dt \\ &= \int_{t_0}^{t_f} [\nabla_{\mathbf{u}} \bar{\mathcal{H}}^*] \delta \mathbf{u}(t) dt. \end{aligned}$$

By the fundamental lemma of the calculus of variations, it follows that:

$$\nabla_{\mathbf{u}} \bar{\mathcal{H}}^* = 0 \quad \text{for all } t \in [t_0, t_f]. \quad \blacksquare$$

This shows that the optimal control $\mathbf{u}^*(t)$ is only required to satisfy the optimality conditions in the expectation of $p(\theta | \mathcal{D})$, extending the minimum principle to probabilistic dynamics models that capture epistemic uncertainty.

IV. NUMERICAL METHODS FOR PROBABILISTIC HAMILTONIAN

A. Trajectory sampling

Computing the mean over Hamiltonians $\bar{\mathcal{H}}$ in Eq. (III.9) requires estimating the uncertainty in state and costate trajectories. We adopt the trajectory sampling technique to capture this variability induced by the dynamics posterior [16]. The mean Hamiltonian is approximated by aggregating the sampled Hamiltonian functions:

$$\mathbf{u}^*(t) \approx \argmin_{\mathbf{u}} \frac{1}{M} \sum_{i=1}^M \mathcal{H}_{\theta_i}(\mathbf{x}_{\theta_i}(t), \boldsymbol{\lambda}_{\theta_i}(t), \mathbf{u}, t), \quad (\text{IV.1})$$

where $\theta_i \sim p(\theta | \mathcal{D})$ for $i = 1, \dots, M$, and each conditioned Hamiltonian is computed along the trajectories $\mathbf{x}_{\theta_i}(t)$ and $\boldsymbol{\lambda}_{\theta_i}(t)$, which comply with the PMP conditions.

An alternative formulation is to consider the mean of the optimal controls $\bar{\mathbf{u}}^*(t) = \mathbb{E}_{\theta}[\mathbf{u}_{\theta}^*(t)]$:

$$\bar{\mathbf{u}}^*(t) \approx \frac{1}{M} \sum_{i=1}^M \mathbf{u}_{\theta_i}^*(t). \quad (\text{IV.2})$$

By decoupling the optimization process over different realizations from the dynamics posterior, the mean optimal control approximation is only a valid necessary optimality condition when the Hamiltonian function is convex. However, the solutions can be computed in parallel, allowing for computationally efficient implementation since both approaches leverage Monte Carlo estimates of the state and co-state trajectories to efficiently evaluate the M samples using GPU parallelization.

B. Two-point boundary value problem

For each realization θ_i , the initial state $\mathbf{x}_{\theta_i}(t_0) = \mathbf{x}_0$ and the final co-state $\boldsymbol{\lambda}_{\theta_i}(t_f) = \nabla_{\mathbf{x}}\Phi(\mathbf{x}_{\theta_i}(t_f))$ are shared. This gives rise to a two-point boundary value problem (TPBVP) that satisfies the coupled differential equations $\dot{\mathbf{x}}_{\theta_i}(t) = \mathbf{f}_{\theta_i}(\mathbf{x}_{\theta_i}(t), \boldsymbol{\lambda}_{\theta_i}(t))$ and $\dot{\boldsymbol{\lambda}}_{\theta_i}(t) = -\nabla_{\mathbf{x}}\mathcal{H}_{\theta_i}(\mathbf{x}_{\theta_i}(t), \boldsymbol{\lambda}_{\theta_i}(t), \mathbf{u}(t), t)$, where the initial value of the co-state $\boldsymbol{\lambda}_{\theta_i}(t_0)$ is unknown. Due to the continuous-time formulation, numerical ODE solvers with adaptive step-size solvers can be leveraged to improve stability and control the numerical error.

We solve this using the *forward shooting method* [21], which guesses an initial co-state $\boldsymbol{\lambda}_{\theta_i}(t_0)$, integrates the states and co-states forward, and adjusts the guess based on the terminal errors $\boldsymbol{\lambda}_{\theta_i}(t_f) - \nabla_{\mathbf{x}}\Phi(\mathbf{x}_{\theta_i}(t_f))$. Root-finding methods (e.g., Newton’s method) or least-squares optimization (e.g., Levenberg-Marquardt) refine the guess iteratively. For robustness with non-linear dynamics for long time horizons, we employ the *multiple shooting* formulation [21]. This method breaks down the long time horizon into multiple shorter intervals and enforces continuity conditions between these, to improve convergence and stability.

The mean Hamiltonian approach shares controls across dynamics function realizations, leading to MSd_x decision variables, where S is the number of shooting segments. In contrast, the mean optimal control variant decouples the optimization process of the sample realizations by only averaging *after* the control function is optimized. This results in M independent optimization problems with Sd_x decision variables, which can be solved in parallel.

C. Model predictive control

The TPBVP describes an open-loop trajectory optimization problem. To obtain a closed-loop control, principles from non-linear MPC can be leveraged by solving the trajectory optimization at every time point under a moving horizon of $t \in [t_0, t_0 + H]$ [21]. The optimized control function is then applied until a new system measurement becomes available, and the optimization process is repeated. By using a finite prediction horizon, the control resulting from MPC is not necessarily optimal because it does not account for how future control actions will be re-optimized

based on new measurements. Thus, this approach trades off theoretical optimality for computational tractability and the ability to react to real-time disturbances.

V. EXPERIMENTAL RESULTS

We demonstrate our approach on three classical control problems, with all system parameters, control cost matrices and simulation parameters detailed in the Appendix, Table III. For each task, we adopt a standard quadratic cost functional of the form

$$C(\mathbf{u}) = (\mathbf{x}(t_f) - \mathbf{x}^*)^\top Q_f (\mathbf{x}(t_f) - \mathbf{x}^*) + \int_{t_0}^{t_f} ((\mathbf{x}(t) - \mathbf{x}^*)^\top Q (\mathbf{x}(t) - \mathbf{x}^*) + \mathbf{u}(t)^\top R \mathbf{u}(t)) dt$$

where Q and Q_f are positive semi-definite state-cost matrices, R is a positive definite control-cost matrix, and \mathbf{x}^* is the target state. The code implementing the experiments is available at <https://github.com/DavidLeeftink/probabilistic-pontyagin-control>.

A. Van der Pol stabilization

We first evaluate our method on stabilizing a Van der Pol oscillator, a non-linear dynamical system commonly used in electrical circuit design and computational neuroscience [22]. We consider the following control formulation:

$$\begin{aligned} \dot{x}_1 &= \mu \left(x_2 - \frac{1}{3}x_1^3 - x_1 \right), \\ \dot{x}_2 &= -x_1 + u, \end{aligned}$$

where $\mu = 1.5$. The goal is stabilization of the system at the unstable fixed point at the origin, under a quadratic cost function. This requires precise control actions to prevent repulsion towards the attracting limit cycle.

We formulate an offline RL task where we generate a dataset by simulating the Van der Pol system for $n = 25$ trajectories of length $T = 10$. These simulations are driven by random periodic control inputs (Schroeder sweeps [23]) and initialized from uniformly sampled states. In this offline setting, the system’s underlying dynamics are assumed to be unknown. We then use this data to train a probabilistic dynamics model — an ensemble of neural ODEs — to capture the epistemic uncertainty. Finally, the learned model is used to construct a controller, allowing a direct comparison between the probabilistic Pontryagin approaches and direct optimization approaches of the mean cost.

To explicitly investigate the effect of epistemic uncertainty modeling, we compare our approach using both deterministic and probabilistic dynamics models trained on identical datasets. We benchmark the mean Hamiltonian (PMP-Mean \mathcal{H}) and mean posterior (PMP-Mean u^*) formulations against direct optimization methods: Sequential Quadratic Programming (SQP), Broyden–Fletcher–Goldfarb–Shanno (BFGS) [24], improved Cross-Entropy Method (iCEM) [25], [26], and Adam [27]. We provide further experimental details in the Appendix, Table III.

Table I summarizes the performance in terms of accumulated cost over 15 repetitions. The indirect PMP approach

TABLE I: Comparison of optimization methods for Van der Pol stabilization with true and learned dynamics (mean cost \pm std. dev. over 15 repetitions).

Method	True Model	NODE	Prob. NODE
PMP-Mean \mathcal{H}	9.85	12.71 \pm 2.45	11.57 \pm 1.73
PMP-Mean u^*			10.59 \pm 0.88
iCEM	10.08	13.90 \pm 2.80	14.06 \pm 2.81
SQP	10.04	13.10 \pm 2.52	13.69 \pm 2.14
BFGS	10.03	13.13 \pm 2.59	13.67 \pm 2.75
Adam	10.05	14.03 \pm 2.66	14.27 \pm 3.03

TABLE II: Comparison of optimization methods for the Cart pole control problem with true and learned dynamics (mean cost \pm std. dev. over 15 repetitions).

Method	True Model	NODE	Prob. NODE
PMP-Mean \mathcal{H}	12.57	112.22 \pm 126.99	36.78 \pm 31.45
PMP-Mean u^*			64.44 \pm 77.95
iCEM	7.23	87.55 \pm 109.56	109.51 \pm 183.34
SQP	12.57	111.52 \pm 110.37	77.98 \pm 107.10
BFGS	12.57	147.95 \pm 168.82	81.78 \pm 132.58
Adam	12.89	251.31 \pm 270.43	311.77 \pm 349.75

achieves improved cost performance over its direct counterpart, under both the deterministic and probabilistic dynamics model. The lowest cost is achieved under the probabilistic Pontryagin approaches, where the mean optimal control approximation slightly outperforms the mean Hamiltonian approach, suggesting that explicitly modeling the epistemic uncertainty improves performance.

B. Cart pole swing up

Next, we consider a cart pole swing-up task, where the objective is to balance the pendulum upright from an initial down position under quadratic cost [28]:

$$\ddot{x} = \frac{lm \sin(\theta) \dot{\theta}^2 + u + m g \cos(\theta) \sin(\theta)}{M + m(1 - \cos^2(\theta))},$$

$$\ddot{\theta} = -\frac{lm \cos(\theta) \sin(\theta) \dot{\theta}^2 + u \cos \theta + (M + m) g \sin(\theta)}{lM + lm(1 - \cos^2(\theta))}.$$

Here, x represents the horizontal position of the cart, θ is the angle of the pole (with $\theta = \pi$ corresponding to the upright, balanced position), and the control input u is the horizontal force applied to the cart. The physical parameters are the pole's half-length l , the pole's mass m , the cart's mass M , and the acceleration due to gravity g .

We design an offline RL task in similar fashion to the Van der Pol experiment, where the dynamics of the cart pole are assumed to be unknown and have to be learned from observed trajectories. We generate $n = 50$ trajectories of length $T = 5$ using random controls and uniformly sampled initial conditions, which serve as data. This results in sparse observations near the goal state. We compare the performance of the previously mentioned trajectory optimization methods on deterministic and probabilistic neural ODEs. Further experimental details are described in the Appendix, Table III.

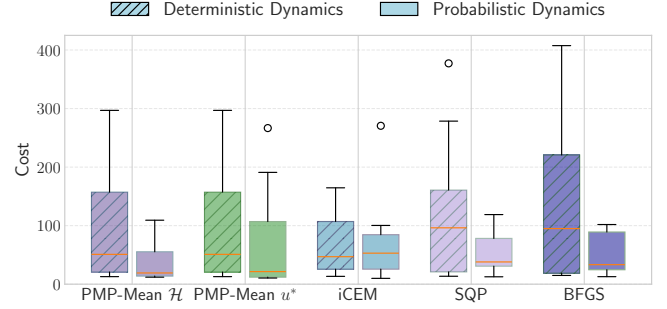


Fig. 3: Comparison of optimization methods for the cart pole control task with deterministic and probabilistic dynamics models. The boxplot shows the distribution of cost over 15 datasets, including the median, interquartile range, and outliers. Due to its high cost average, the Adam optimizer is not visualized.

Table II and Figure 3 present the trial costs averaged over 15 repetitions, comparing deterministic and probabilistic dynamics models across various optimization methods. Under limited available data, adopting the probabilistic dynamics model improves performance for most optimizers, emphasizing the role of capturing epistemic uncertainty in learning-based control tasks. The mean Hamiltonian formulation achieves the lowest mean and median trial costs out of the considered optimization methods, outperforming direct methods. The mean posterior approximation, theoretically suited only for convex Hamiltonians, performs worse than its counterpart with higher variability and average cost.

C. Duffing oscillator stabilization

Lastly, we consider an online RL study involving the Duffing oscillator, a non-linear system modeling a mass in a double well potential, which can be regarded as a model of a steel beam that is deflected towards two magnets [29]:

$$\dot{x}_1 = x_2,$$

$$\dot{x}_2 = -\delta x_2 - \alpha x_1 - \beta x_1^3 + \gamma u.$$

Under periodic forcing, chaotic behavior can be observed. In our configuration, the system has two stable attractors, separated by an unstable fixed point at the origin. The objective is to drive the system to the unstable fixed point.

In the online RL setting, one starts from a single random-policy trial, and iteratively trains the probabilistic dynamics model, updates the controller, and augments the dataset after each trial. We compare the mean Hamiltonian and mean posterior methods to the state-of-the-art PETS algorithm [16], which combines probabilistic dynamics models with MPC and the (i)CEM optimizer. We adapt PETS to the continuous-time setting using probabilistic neural ODEs as dynamics models. Further experimental details are described in the Appendix, Table III.

Figure 4 presents the average trial costs over 10 sequential iterations (averaged over 10 independent repetitions), with the dashed reference line indicating performance under

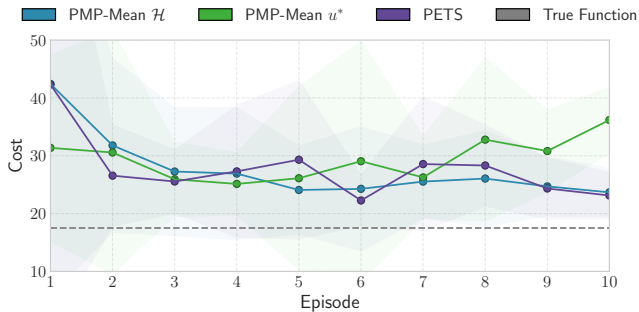


Fig. 4: Comparison of online model-based RL methods for the Duffing oscillator stabilization task (cost \pm std. dev. over 10 independent repetitions).

known dynamics. Both PETS and the mean Hamiltonian approach converge to similar final costs, although PETS shows slight cost oscillations. The mean posterior approximation initially performs better but later becomes unstable, likely due to the inadequate aggregation of independently optimized control functions in the presence of multiple attractors near the goal state.

VI. CONCLUSION AND DISCUSSION

In this work, we have introduced a novel framework for MBRL that addresses the critical challenge of planning under epistemic uncertainty. Our approach translates a probabilistic interpretation of PMP, established in optimal control theory [10], [11], into a practical algorithm for systems with learned dynamics. The core of our method is the principle of minimizing the mean Hamiltonian, which we compute over an ensemble of models representing the posterior distribution over the system dynamics that arises from limited data.

We developed a scalable numerical method, combining forward multiple shooting with moving horizon optimization, to make this principled approach tractable for deep, probabilistic models. A key feature of our framework is that it optimizes trajectories in continuous time—naturally aligning with the physics of many real-world systems—while still only requiring standard discrete-time observations for model learning. While we adopted neural ODE ensembles as a probabilistic dynamics model, the proposed principle generalizes to other probabilistic dynamics models, including Bayesian neural ODEs or systems with known functional forms but uncertain parameters.

Our experiments demonstrate the practical effectiveness of our approach in offline RL tasks under limited data availability, including the Van der Pol oscillator and a cart pole system. Comparisons against direct trajectory optimization methods, such as SQP and (i)CEM, show that our proposed mean Hamiltonian formulation achieves reduced trial costs. Our results further underline the advantage of explicitly modeling the epistemic uncertainty: under identical neural architectures, the probabilistic ensemble model consistently outperforms its deterministic counterpart, indicating that capturing epistemic uncertainty can lead to improved

control performance. Additionally, in online RL settings, such as the stabilization of a Duffing oscillator, the proposed approach achieves competitive trial cost and data efficiency, comparable to the state-of-the-art MBRL method PETS.

Although these results are promising, computational challenges arise in higher-dimensional systems or large ensemble sizes, due to solving the co-state equations via a root-finding task with cubic complexity. However, the sparse Jacobian structure and parallelizable computations of the sampled dynamics provide avenues for future algorithmic improvements or GPU acceleration. Alternatively, the forward-backward iterative approach can be adopted, which avoids the cubic complexity but may require additional iterations until convergence [30].

Future theoretical work could explore extending the optimality conditions to problems with state constraints, which commonly arises in safety critical application. One possible avenue to realize this is by incorporating safety constraints in the Hamiltonian via additional co-state variables. From a practical perspective, our probabilistic PMP approach is especially suited for applications requiring precise control under uncertainty on slower timescales, such as personalized treatment planning in healthcare or process control.

ACKNOWLEDGMENTS

We would like to thank Dr. Yuzhen Qin for his valuable feedback and insightful comments during the review of this manuscript. We are also grateful to the anonymous reviewers for their constructive suggestions, which helped to significantly improve the final version of the paper.

APPENDIX

A. Experimental details

For all experiments, cubic interpolation is used to construct a continuous-time control function from a discretized control vector, and a 5th-order Dormand-Prince solver is used to solve predicted states (and co-states for PMP) [31] for the learned dynamics model. An overview of all the system parameters, control cost matrices and simulation parameters used in the experiments is presented in Table III.

The following settings were used for the respective experiments:

a) Van der Pol: Following the approach described in [16], we consider an ensemble size of $M = 5$ following common practice in model-based RL, with each neural ODE consisting of two hidden layers with 32 units, resulting in parameters $\theta \in \mathbb{R}^{1248}$. Our indirect methods use $S = 4$ shooting segments. All trajectory optimization methods employ moving horizon optimization (horizon $H = 3$) and 15 iterations per timestep.

b) Cart pole: The dynamics model architecture is consistent with the previous Van der Pol setup, using ensemble neural ODEs ($\theta \in \mathbb{R}^{1376}$) with identical control interpolation and ODE solver. Here, we use a shorter MPC horizon of $H = 1$ with 25 iterations per time step, and two shooting segments ($S = 2$) to ensure computational efficiency.

TABLE III: System parameters, control cost matrices and simulation parameters used in the experiments.

System	Parameters	Q	Q_f	R	σ	\mathcal{U}	x_0	x^*	t_0	t_f	Δt (MPC)	H
Van der Pol	$\mu = 1.5$	\mathbb{I}	\mathbb{I}	0.5	0.01	$[-2, 2]$	$(1, 1)^\top$	$(0, 0)^\top$	0	10	0.05	3
Cart pole	$l = m = M = 1$ $g = 9.81$	$\mathbb{I}(1, 1, 0.1, 0.1)^\top$	$\mathbb{I}(1, 5, 1, 1)^\top$	0.05	0	$[-20, 20]$	$(0, 0, 0, 0)^\top$	$(1, \pi, 0, 0)^\top$	0	5	0.02	1
Duffing	$\alpha = -1, \quad \beta = 2$ $\delta = 0.2, \quad \gamma = 1$	$5 \cdot \mathbb{I}$	$5 \cdot \mathbb{I}$	1	0.01	$[-2, 2]$	$(1.5, 1)^\top$	$(0, 0)^\top$	0	5	0.05	2

c) Duffing: We adopt an ensemble of neural ODEs with $M = 5$ as a probabilistic dynamics model, consistent with the previous van der Pol setup ($\theta \in \mathbb{R}^{1248}$) with identical control interpolation and ODE solver. We employ MPC with horizon $H = 2$ and adjusted integration steps suitable for the Duffing oscillator's faster time scale, and use $S = 5$ shooting segments for the indirect methods.

REFERENCES

- [1] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. Wiley, 2011.
- [2] L. S. Pontryagin, *Mathematical Theory of Optimal Processes*, 1st ed. Routledge, 1987.
- [3] D. E. Kirk, *Optimal Control Theory: An Introduction*. Courier Corporation, 2004.
- [4] S. Peng, "A general stochastic maximum principle for optimal control problems," *SIAM Journal on control and optimization*, vol. 28, no. 4, pp. 966–979, 1990.
- [5] W. Li and E. Todorov, "Iterative linear quadratic regulator design for nonlinear biological movement systems," in *First International Conference on Informatics in Control, Automation and Robotics*, vol. 2. SciTePress, 2004, pp. 222–229.
- [6] H. J. Kappen, "Linear theory for control of nonlinear stochastic systems," *Physical Review Letters*, vol. 95, no. 20, p. 200201, 2005.
- [7] T. Başar and P. Bernhard, *H-infinity Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*. Springer Science & Business Media, 2008.
- [8] W. Jin, Z. Wang, Z. Yang, and S. Mou, "Pontryagin differentiable programming: An end-to-end learning and control framework," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7979–7992, 2020.
- [9] O. Eberhard, C. Vernade, and M. Muehlebach, "A Pontryagin perspective on reinforcement learning," in *Proceedings of the Sixth Annual Learning for Dynamics & Control Conference*, 2025.
- [10] P. Bettiol and N. Khalil, "Necessary optimality conditions for average cost minimization problems," pp. 2093–2124, 2019. [Online]. Available: <https://www.aims sciences.org/article/id/1cf0575a-65d3-4533-8dbe-f9d8e5ebe252>
- [11] A. Scagliotti, "Optimal control of ensembles of dynamical systems," *ESAIM: Control, Optimisation and Calculus of Variations*, vol. 29, p. 22, 2023.
- [12] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar, "Bayesian reinforcement learning: A survey," *Foundations and Trends in Machine Learning*, vol. 8, no. 5-6, pp. 359–483, 2015.
- [13] A. Pesare, M. Palladino, and M. Falcone, "Convergence results for an averaged lqr problem with applications to reinforcement learning," *Mathematics of Control, Signals, and Systems*, vol. 33, no. 3, pp. 379–411, 2021.
- [14] M. Deisenroth and C. E. Rasmussen, "Pilco: A model-based and data-efficient approach to policy search," in *Proceedings of the 28th International Conference on machine learning*, 2011, pp. 465–472.
- [15] S. Kamthe and M. Deisenroth, "Data-efficient reinforcement learning with probabilistic model predictive control," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018, pp. 1701–1710.
- [16] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [17] C. Yildiz, M. Heinonen, and H. Lähdesmäki, "Continuous-time model-based reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 009–12 018.
- [18] L. Treven, J. Hübotter, B. Sukhija, F. Dorfler, and A. Krause, "Efficient exploration in continuous-time model-based reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 42 119–42 147, 2023.
- [19] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid monte carlo," *Physics letters B*, vol. 195, no. 2, pp. 216–222, 1987.
- [20] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. Taylor & Francis, 1975.
- [21] S. Gros and M. Diehl, *Numerical Optimal Control (Draft)*, 2022, available online: <https://www.syscop.de/files/2020ss/NOC/book-NOCSE.pdf> (accessed on 28 June 2023).
- [22] B. Van der Pol, "Lxxxviii. on "relaxation-oscillations"," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 978–992, 1926.
- [23] M. Schroeder, "Synthesis of low-peak-factor signals and binary sequences with low autocorrelation (corresp.)," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 85–89, 1970.
- [24] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.
- [25] R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodology and Computing in Applied Probability*, vol. 1, pp. 127–190, 1999. [Online]. Available: <https://doi.org/10.1023/A:1010091220143>
- [26] C. Pinneri, S. Sawant, S. Blaes, J. Achterhold, J. Stueckler, M. Rolinek, and G. Martius, "Sample-efficient cross-entropy method for real-time planning," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 1049–1065.
- [27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.
- [28] M. Kelly, "An introduction to trajectory optimization: How to do your own direct collocation," *SIAM Review*, vol. 59, no. 4, pp. 849–904, 2017.
- [29] F. C. Moon and P. J. Holmes, "A magnetoelastic strange attractor," *Journal of Sound and Vibration*, vol. 65, no. 2, pp. 275–296, 1979.
- [30] S. K. Mitter, "Successive approximation methods for the solution of optimal control problems," *Automatica*, vol. 3, no. 3–4, pp. 135–149, 1966.
- [31] J. R. Dormand and P. J. Prince, "A family of embedded Runge-Kutta formulae," *Journal of Computational and Applied Mathematics*, vol. 6, no. 1, pp. 19–26, 1980.