Operator Models for Continuous-Time Offline Reinforcement Learning

Nicolas Hoischen
TU Munich

Petar Bevanda
TU Munich

Max Beier
TU Munich

Stefan Sosnowski
TU Munich

Boris Houska
ShanghaiTech University

Sandra Hirche
TU Munich

Abstract

Continuous-time stochastic processes underlie many natural and engineered systems. In healthcare, autonomous driving, and industrial control, direct interaction with the environment is often unsafe or impractical, motivating offline reinforcement learning from historical data. However, there is limited statistical understanding of the approximation errors inherent in learning policies from offline datasets. We address this by linking reinforcement learning to the Hamilton–Jacobi–Bellman equation and proposing an operator-theoretic algorithm based on a simple dynamic programming recursion. Specifically, we represent our world model in terms of the infinitesimal generator of controlled diffusion processes learned in a reproducing kernel Hilbert space. By integrating statistical learning methods and operator theory, we establish global convergence of the value function and derive finite-sample guarantees with bounds tied to system properties such as smoothness and stability. Our theoretical and numerical results indicate that operator-based approaches may hold promise in solving offline reinforcement learning using continuous-time optimal control.

Figure 1: Overview of the O-CTRL algorithm: a generator world model based on an RKHS representation of state–action data enables dynamic programming for optimal value functions, illustrated on the swing-up pendulum task from Gymnasium (Towers et al., 2024).

1 INTRODUCTION

A wide variety of phenomena, from the motion of molecules, the nerve activation in our brains, the value of a stock in the financial market, to the dynamics of robotic systems, can be modeled as random processes governed by a stochastic differential equation (Särkkä and Solin, 2019). Consequently, making decisions in these p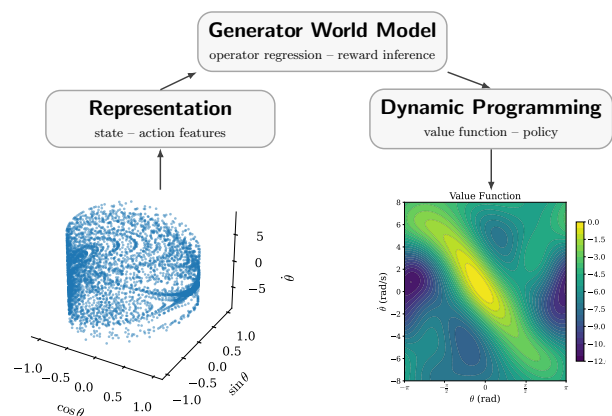rocesses that optimize a quantity of interest is a common task across various disciplines. Examples include the expected return-on-investment in finance, the distance to a goal in robotics, or the yield in chemical engineering. Tools that have been immensely successful are optimal control when accurate reduced models are available or reinforcement learning, when interaction with the environment (or a simulator) is possible. However, some problems do not allow for interaction and have no accurate simulators or surrogate models. Offline reinforcement learning (RL) attempts to build a decision-making policy given a reward signal from historical observation data. This enables policy learning without requiring online environment interaction. The paradigm seeks to transfer the data-first approach behind breakthroughs in computer vision and natural language to decision-making contexts in which online interaction is costly, time-consuming, or infeasible.

Offline RL   Despite substantial empirical progress in offline reinforcement learning (Levine et al., 2020; Prudencio et al., 2023), its theory is well-developed only for discrete-time Markov Decision Processes (MDPs) (Williams et al., 2017; Yu et al., 2020, 2021). In terms of the natural SDE description, we lack an end-to-end learning-theoretic understanding of offline RL for these systems:

How much data is necessary? When is reliable offline RL in principle possible?

In the discrete-time setting, various offline RL methods exist both for the model-based (Yu et al., 2020; Kidambi et al., 2020; Yu et al., 2021; Williams et al., 2017) and model-free case (Fujimoto et al., 2019; Kostrikov et al., 2021, 2022; Burns et al., 2023). The available methods for continuous-time are mostly model-free (Jia and Zhou, 2023; Wiltzer et al., 2024) or do not provide a learning theory (Holt et al., 2023). For a comprehensive review of offline RL, we refer to (Prudencio et al., 2023).

Optimal Control   Under certain technical assumptions, performing RL is equivalent to solving an optimal control problem. It is, therefore, natural to turn to optimal control, as it is a well-studied field. The connection between RL and the numerical analysis of the optimal control problem has been recognized and investigated in Munos (2000); Doya (2000). However, the communities have diverged significantly, with recent works trying to reconnect them still focusing on the LQR case (Wang et al., 2020) or requiring interaction with the environment (Jia and Zhou, 2022a). Meanwhile, numerical schemes constructed to solve optimal control problems lack a statistical learning theory explaining their dependence on the problem and sample complexity (Blessing et al., 2025; Lutter et al., 2020, 2023; Shilova et al., 2024; Halperin, 2024; Meng et al., 2024).

Operator Learning   Operator models provide an approach to learning relationships in feature space (Grünewälder et al., 2012; Li et al., 2022). They enable a fundamentally different perspective on reinforcement learning and optimal control: rather than modeling the system dynamics through stochastic differential equations or transition kernels, they model the evolution of probability densities via Markov semigroups (Engel et al., 2000; Korda and Mezić, 2018; Brunton et al., 2022; Kostic et al., 2022, 2024a, 2025). This shift is attractive because it directly captures the distributional evolution of states, which is central in policy evaluation and optimization/planning, and it allows one to exploit tools from functional analysis and operator theory. Recent operator-learning approaches

highlight the promise of modeling conditional expectations/distributions for control and reinforcement learning (Wenliang et al., 2024). In discrete time, this has led to policy mirror descent (Novelli et al., 2024) and LQ control extensions (Caldarelli et al., 2025), However, the above discrete-time methods do not extend naturally to infinite action spaces or beyond quadratic rewards and linear models. In continuous time, (Bevanda et al., 2025b) proposed a kernel-based formulation that yields promising results but puts unrealistic assumptions on data collection and lacks a systematic error analysis. Our work overcomes these limitations while maintaining the strengths of operator modeling.

Contribution   In this paper, we adopt an operator-based perspective to decompose the offline continuous-time RL problem, including world model operator learning and policy optimization. This separation enables us to construct an Operator-based Continuous-Time offline Reinforcement Learning (O-CTRL) algorithm (Figure 1) that solves the RL problem for the optimal value function. Utilizing statistical learning theory for linear operator learning and convex optimization, we establish global convergence of the value function and derive finite-sample guarantees with bounds tied to system properties such as smoothness and stability.

Organization of the Paper   Section 2 reviews the necessary background and formalizes the continuous offline RL setting. In Section 3, we present the key steps in deriving O-CTRL and give a theoretical analysis in section 4. Section 5 illustrates the theory with numerical examples. The appendix contains additional related work, proofs, and more details on experimental results.

2   PRELIMINARIES

Notation   We write $[n] := \{1, \ldots, n\}$ for integers, $\odot$ for the Hadamard product. Let the state space be $\mathbb{S} \subseteq \mathbb{R}^{n_s}$, the action space be $\mathbb{A} \subseteq \mathbb{R}^{n_a}$ and $\mathbb{Z} := \mathbb{S} \times \mathbb{A}$. The set of k-times continuously differentiable functions is denoted by $C^k(\mathbb{S})$. For a measure μ on $\mathbb{S}$, we denote by $L^2(\mathbb{S})$ the space of square-integrable functions and by $H^k(\mathbb{S})$ the Sobolev space of k-times weakly differentiable functions with square-integrable derivatives. $\mathrm{HS}(\mathcal{F}, \mathcal{Y})$ denotes the space of Hilbert–Schmidt operators with norm $\|A\|^2_{\mathrm{HS}(\mathcal{F}, \mathcal{Y})} := \sum_i \|Ae_i\|^2_{\mathcal{Y}}$ for any orthonormal basis $\{e_i\}$ of $\mathcal{F}$. We write $\langle \cdot, \cdot \rangle$ for the inner product. Finally, the symbols $\nabla_{\boldsymbol{s}}$ and $\Delta_{\boldsymbol{s}}$ denote the gradient and Laplacian with respect to the vector-valued variable $\boldsymbol{s}$.

Continuous-Time Markov Decision Processes
A continuous-time Markov Decision Process is de-

scribed by an SDE influenced by actions $\boldsymbol{a}_t$, typically on an unbounded domain $\mathbb{S} = \mathbb{R}^{n_s}$,

$$d\boldsymbol{S}_t = \left(\boldsymbol{f}(\boldsymbol{S}_t) + \boldsymbol{G}(\boldsymbol{S}_t)\,\boldsymbol{a}_t\right) dt + \sqrt{2\epsilon}\,dW_t, \quad (1)$$

for $t \geq 0$, where W_t is a standard Wiener process and ϵ the diffusion parameter. Here, we focus on action-affine systems with state-independent process noise (1), because they are often sufficient to capture the behavior of many continuous-time cyber-physical systems (Nijmeijer and van der Schaft, 1996). Additionally, we introduce a regularity assumption:

Assumption 1 (Dynamics). *We assume that* $\boldsymbol{f} \in C^1(\mathbb{S})^{n_s}$, $\boldsymbol{G} \in C^1(\mathbb{S})^{n_s \times n_a}$ *and* $\epsilon > 0$.

Operator Models Conditional expectations of observables $\phi \in L^2(\mathbb{S})$ are described by conditional expectation operators (Kallenberg, 1997). The conditional expectation operators with respect to $\boldsymbol{S}_t$ form an evolution family $\Gamma_{u,t}$ where $u \geq t \geq 0$.

$$[\Gamma_{u,t}\phi](\boldsymbol{s}, \boldsymbol{a}(\tau)) = \mathbb{E}\left[\phi(\boldsymbol{S}_u) \mid \boldsymbol{S}_t{=}\boldsymbol{s}, \boldsymbol{a}(\tau)\right], \tau{\in}[u,t] \quad (2)$$

Its infinitesimal generator $\mathcal{G}_t$ at time t is defined on a domain with sufficient regularity $D(\mathcal{G})$ (Engel et al., 2000). In particular, the generator associated with (1) is

$$[\mathcal{G}_t\phi](\boldsymbol{s}, \boldsymbol{a}) = [\mathcal{A}\phi](\boldsymbol{s}) + [\mathcal{B}_t\phi](\boldsymbol{s}, \boldsymbol{a}), \quad (3)$$

which splits into the autonomous and action-dependent dynamics, respectively:

$$[\mathcal{A}\phi](\boldsymbol{s}) = \nabla_{\boldsymbol{s}}\phi(\boldsymbol{s})^\top \boldsymbol{f}(\boldsymbol{s}) + \epsilon\,\Delta_{\boldsymbol{s}}\phi(\boldsymbol{s}), \quad (4a)$$

$$[\mathcal{B}_t\phi](\boldsymbol{s}, \boldsymbol{a}) = \nabla_{\boldsymbol{s}}\phi(\boldsymbol{s})^\top \boldsymbol{G}(\boldsymbol{s})\,\boldsymbol{a}_t. \quad (4b)$$

As the time dependence in $\mathcal{G}_t$ is not from the transition dynamics but from actions $\boldsymbol{a}_t$, we drop their subscripts t when in the infinitesimal setting.

Reinforcement Learning For any reward $r(\boldsymbol{s}, \boldsymbol{a}) : \mathbb{R}^{n_s} \times \mathbb{R}^{n_a} \to \mathbb{R}$ over state-action pairs, the goal of reinforcement learning is to find the stationary policy $\boldsymbol{\pi}$ maximizing the expected discounted return – the value function,

$$V(\boldsymbol{s}) \coloneqq \mathbb{E}\left[\int_0^\infty e^{-\rho t} r\left(\boldsymbol{S}_t, \boldsymbol{\pi}(\boldsymbol{S}_t)\right) dt \middle| \boldsymbol{S}_0 = \boldsymbol{s}\right], \quad (5)$$

with a discount exponent $\rho > 0$. To connect (5) with the operator model in (3) we introduce the substitution operator $P^{\boldsymbol{\pi}} : L^2(\mathbb{S} \times \mathbb{A}) \to L^2(\mathbb{S})$. It replaces open-loop actions with the output of a policy $\boldsymbol{a} = \boldsymbol{\pi}(\boldsymbol{s})$

$$[P^{\boldsymbol{\pi}}r](\boldsymbol{s}) = r(\boldsymbol{s}, \boldsymbol{\pi}(\boldsymbol{s})). \quad (6)$$

Hence, the generator of (1) under a policy reads

$$P^{\boldsymbol{\pi}}\mathcal{G} = \mathcal{A} + P^{\boldsymbol{\pi}}\mathcal{B}. \quad (7)$$

Now, instead of sampling the SDE (1) and approximating the expected return (5), we can use Fubini's theorem to model the conditional expectation directly.

$$V(\boldsymbol{s}) = \int_0^\infty e^{-\rho t}\left[P^{\boldsymbol{\pi}}\Gamma_{t,0}(P^{\boldsymbol{\pi}}r)\right](\boldsymbol{s})dt \quad (8)$$

$$= \left[\left(\rho I - P^{\boldsymbol{\pi}}\mathcal{G}\right)^{-1}(P^{\boldsymbol{\pi}}r)\right](\boldsymbol{s}) \quad (9)$$

The inverse in (9) is exactly the definition of the *resolvent operator* (Engel et al., 2000, 1.10)

$$R_\rho^{\boldsymbol{\pi}}(\mathcal{G}) \coloneqq \left(\rho I - P^{\boldsymbol{\pi}}\mathcal{G}\right)^{-1}$$

applied to the reward function under any policy $\boldsymbol{\pi}$ for which the SDE (1) admits a well-defined solution. As the resolvent *linearly maps* the reward to the value function, this establishes the infinitesimal generator as a fundamental object in continuous-time Markov decision processes.

Hamilton-Jacobi-Bellman Equation Before connecting to optimal control theory, we recognize that the identity in (9) characterizes V for any *fixed* $\boldsymbol{\pi}$. To obtain the *optimal* value function, we optimize over policies. The resulting optimal value function $V^\star$ must satisfy the stationary discounted HJB

$$\max_{\|\boldsymbol{\pi}\|_{L^\infty} < \infty} \left\{[P^{\boldsymbol{\pi}}\mathcal{G}V^\star](\boldsymbol{s}) + r(\boldsymbol{s}, \boldsymbol{\pi}(\boldsymbol{s}))\right\} = \rho V^\star(\boldsymbol{s}), \quad (10)$$

expressed in terms of the generator and the substitution operator. Precise conditions under which the HJB (10) has a well-defined solution $V^\star$ can be found in Houska (2025); Fleming and Soner (2006).

2.1 Problem Statement

Given the recorded data of a continuous-time system in infinitesimal form

$$\mathbb{D}_N = \left\{\dot{\boldsymbol{s}}^{(i)}, \left(\boldsymbol{s}^{(i)}, \boldsymbol{a}^{(i)}\right)\right\}_{i \in [N]}, \quad (11)$$

our objective is to find the optimal value function

$$V^\star \coloneqq \max_{\|\boldsymbol{\pi}\|_{L^\infty} < \infty} \left(\rho I - P^{\boldsymbol{\pi}}\mathcal{G}\right)^{-1} P^{\boldsymbol{\pi}}r, \quad (12)$$

which maximizes the expected cumulative discounted rewards in (5). We aim to derive an algorithm O-CTRL based on the following properties:

(**P1**) **World Model:** By learning an approximation of the infinitesimal generator $\mathcal{G}$, the knowledge of the *reward is required only at inference.*

(**P2**) **Formal Guarantees:** Provide explicit, *interpretable error bounds* on the algorithm's output.

(**P3**) **Simple Algorithm:** Optimization is carried out through recursive dynamic programming updates, *implemented by a single for-loop* (scan).

In a nutshell, our goal is an end-to-end pipeline: we start from dynamical system data to construct a value-function approximation with guarantees (**P2**). These are obtained by relating operator world model (**P1**) learning errors to value-function errors. Building on RKHS-based operator models, we formulate a dynamic-programming recursion that jointly updates value and policy and consists only of a single for-loop scan (**P3**). The introduced building blocks are key in making our approach modular: reward shaping or task changes become plug-and-play without retraining the world model.

3 OFFLINE RL MEETS OPTIMAL CONTROL

After defining the policy evaluation as a linear operator (9), we aim to optimize the policy, thereby maximizing the value function. However, this is challenging for several reasons.

Policy Optimization is Nonlinear Unlike the reward-to-value relationship, the mapping of policies to value functions $\boldsymbol{\pi} \to V$ is nonlinear, making it challenging to optimize in general. Moreover, for our continuous-time setting, the Q-function is ill-defined as the discretization vanishes (Tallec et al., 2019; Kim et al., 2021). Using operator theory and optimizing in Hilbert spaces, we will eliminate any dependence on an arbitrarily chosen time increment or surrogate Q-function (Doya, 2000; Tallec et al., 2019; Jia and Zhou, 2023). The aforementioned is due to an inherent connection to Hamilton-Jacobi-Bellman (HJB) equations that are the continuous-time analogues of the Bellman equation.

3.1 The Optimal Control Perspective

Under the conditions on Assumption 1 and a positive discount $\rho > 0$, (10) has a unique viscosity solution that coincides with the optimal value function (Fleming and Soner, 2006).

Optimal Control for Policy Optimization To compute the steady state $V^\star$ solving (10), we evolve the value-iteration flow for $t \geq 0$ on $\mathbb{S} = \mathbb{R}^{n_s}$,

$$\dot{V}(t, \boldsymbol{s}) = \mathcal{T}\big(V(t, \cdot)\big)(\boldsymbol{s}), V(0, \boldsymbol{s}) = V_0(\boldsymbol{s}) \in H^1, \quad (13)$$

under standard regularity assumptions. To pave the way for a simple DP recursion and to enable a joint value–policy update (**P3**), we require a closed-form expression for the operator $\mathcal{T} : H^1(\mathbb{S}) \to L^2(\mathbb{S})$. For this purpose, and to guarantee that $\mathcal{T}$ is well-defined, we impose the reward structure stated in the following Assumption. This choice ensures that the maximization

in (10) can be expressed as a Fenchel-conjugate term (Houska, 2025).

Assumption 2 (Reward). *We model the reward as a function $r(\boldsymbol{s}, \boldsymbol{a}) = r_{\boldsymbol{s}}(\boldsymbol{s}) - c_{\boldsymbol{a}}(\boldsymbol{s}, \boldsymbol{a})$, which is continuously differentiable in both $\boldsymbol{s}$ and $\boldsymbol{a}$. We require strong-convexity of the action penalty $c_{\boldsymbol{a}}(\boldsymbol{s}, \boldsymbol{a})$ w.r.t. actions. The state reward $r_{\boldsymbol{s}}(\boldsymbol{s})$ is either known/defined a priori or unknown but provided in the dataset* (11).

This reward structure is ubiquitous, with quadratic or smoothed p-norm penalties on control effort (Anderson and Moore, 2007; Tassa et al., 2014) and separable state terms (Doya, 2000; Lillicrap et al., 2015). It is widely used in physics-based benchmarks (Todorov et al., 2012; Towers et al., 2024) and modern RL tasks such as locomotion and racing (Hwangbo et al., 2019; Kaufmann et al., 2023).

Using Assumption 2, we get an explicit *policy update rule* from the Fenchel conjugate of the action penalty[1]

$$\mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda}) \coloneqq \max_{\boldsymbol{a}}\{\langle \boldsymbol{a}, \boldsymbol{\lambda} \rangle - c_{\boldsymbol{a}}(\boldsymbol{s}, \boldsymbol{a})\}, \quad (14)$$

which is well-defined and admits an unique maximizer $\boldsymbol{a}^\star(\boldsymbol{\lambda}) = \nabla \mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda})$ (Boyd, 2004).

After isolating the state reward from the action maximization and substituting the Fenchel conjugate $\mathcal{D}_{\boldsymbol{a}}$ in (10), we get an *infinitesimal HJB formulation*

$$\boxed{\mathcal{T}(V) = -\big(\rho I - \mathcal{A}\big)V + r_{\boldsymbol{s}} + \mathcal{D}_{\boldsymbol{a}}\big(\mathcal{B}V\big)} \quad (15)$$

for $V \in H^1(\mathbb{S})$, where we also leveraged the action-affixnity of the dynamics (1). As the maximum in (14) is attained at $\boldsymbol{a}^\star(\boldsymbol{\lambda})$, we find a, perhaps unsurprising, structure

$$\mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda}) = \underbrace{\langle \boldsymbol{a}^\star(\boldsymbol{\lambda}), \boldsymbol{\lambda} \rangle}_{\substack{\text{improvement w/} \\ \text{optimal policy}}} - \underbrace{c_{\boldsymbol{a}}(\boldsymbol{s}, \boldsymbol{a}^\star(\boldsymbol{\lambda}))}_{\text{regularizer/penalty}}, \quad (16)$$

where the *costate* $\boldsymbol{\lambda} \coloneqq \mathcal{B}V$ is the continuous-time analogue of the advantage signal in discrete-time RL. The first term in (16) describes the improvement under the optimal policy, a *deterministic, continuous-time, analogue* to expected advantage. The second term, on the other hand, serves as the regularization and constraint-enforcing term for the actions. A structural analogue to (16) is found in many policy optimization schemes, such as policy mirror descent (PMD) (Tomar et al., 2022), proximal policy optimization (PPO) (Schulman et al., 2017), trust region policy optimization (TRPO) (Schulman et al., 2015).

3.2 Infinitesimal World Models

To build our world model and solve the HJB via a simple dynamic programming recursion (**P3**), we seek

---

[1]Also commonly written as $c_{\boldsymbol{a}}^*(\boldsymbol{\lambda})$ in the literature.

a data-driven approximation of the infinitesimal generator $\mathcal{G} : D(\mathcal{G}) \to L^2$, with $D(\mathcal{G})$ a space with sufficient regularity, such as H^1. This, in turn, allows for obtaining data-based surrogates of the operators $\mathcal{A}$ and $\mathcal{B}$ (4a)-(4b). To ensure computational tractability, we restrict both the value function and policy to reproducing kernel Hilbert spaces, enabling tractable computation within a rich (infinite-dimensional) parameterization.

Reproducing kernel Hilbert spaces We consider RKHSs $\mathcal{H}_\mathbb{S}/\mathcal{H}_\mathbb{Z}$ that are a subset of H^1/L^2-integrable functions (Steinwart and Christmann, 2008, Chapter 4.3) with associated canonical feature maps $\phi_\mathsf{S} : \mathbb{S} \to \mathcal{H}_\mathbb{S}$ and $\phi_\mathsf{Z} : \mathbb{Z} \to \mathcal{H}_\mathbb{Z}$. To perform the differential calculus required for infinitesimal generators, we consider $k : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ to be a symmetric and positive definite kernel function such that $k \in C^{2,2}(\mathbb{S} \times \mathbb{S})$ and $\mathcal{H}_\mathbb{S}$ the corresponding RKHS (Steinwart and Christmann, 2008), with norm denoted as $\| \cdot \|_{\mathcal{H}_\mathbb{S}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{H}_\mathbb{S}}}$. Moreover, $\forall s, s' \in \mathbb{S}$, we have that $k(s, s') = \langle \phi_\mathsf{S}(s), \phi_\mathsf{S}(s') \rangle_{\mathcal{H}_\mathbb{S}} = \langle k(\cdot, s), k(\cdot, s') \rangle_{\mathcal{H}_\mathbb{S}}$ and the reproducing property $h(s) = \langle h, k(\cdot, s) \rangle_{\mathcal{H}_\mathbb{S}}$ holds for all $s \in \mathbb{S}$ and all observables $h \in \mathcal{H}_\mathbb{S}$. We further assume that we are working with universal kernels and that $k(s, s') < \infty$.

Specifically, we look for an RKHS approximation $G : \mathcal{H}_\mathbb{S} \to \mathcal{H}_\mathbb{Z}$. Yet, population-level quantities in such problems are typically unavailable; thus, we approximate them using historical data samples $\mathbb{D}_N$ defined in (11).

Empirical Risk Minimization A standard approach (Kostic et al., 2022, 2024a; Novelli et al., 2024; Bevanda et al., 2025a) to obtain an estimator G is to construct an empirical risk formulation. We first define the action of the infinitesimal generator $\mathcal{G}$ on the canonical feature map. The RKHS-valued generator representer evaluated at the samples $(s^{(i)}, \dot{s}^{(i)})$ is

$$\mathrm{d}\phi_\mathsf{S}(s^{(i)}; \dot{s}^{(i)}) := (\mathcal{G}\phi_\mathsf{S})(s^{(i)})$$
$$= \sum_{k=1}^{n_s} \dot{s}_k^{(i)} \partial_{s_k} \phi_\mathsf{S}(s^{(i)}) + \epsilon \partial_{s_k s_k}^2 \phi_\mathsf{S}(s^{(i)}). \quad (17)$$

We first introduce sampling operators

$$\widehat{S}_\mathsf{Z} : \mathcal{H}_\mathbb{Z} \to \mathbb{R}^N, \qquad (\widehat{S}_\mathsf{Z}\phi_\mathsf{Z})_i := \phi_\mathsf{Z}((s^{(i)}, a^{(i)})),$$
$$\widehat{S}_\mathsf{S} : \mathcal{H}_\mathbb{S} \to \mathbb{R}^N, \qquad (\widehat{S}_\mathsf{S}\phi_\mathsf{S})_i := \phi_\mathsf{S}(s^{(i)}),$$
$$\widehat{S}_\mathsf{d} : \mathcal{H}_\mathbb{S} \to \mathbb{R}^N, \qquad (\widehat{S}_\mathsf{d}\phi_\mathsf{S})_i := \mathrm{d}\phi_\mathsf{S}(s^{(i)}; \dot{s}^{(i)}),$$
$$\widehat{U} : \mathbb{R}^{n_a} \to \mathbb{R}^N, \qquad (\widehat{U}a)_i := \langle a^{(i)}, a \rangle.$$

The estimator can be obtained by minimizing the squared loss, known as the empirical risk. It is given

by

$$\widehat{\mathcal{R}}(G) := \frac{1}{N} \sum_{i=1}^N \|\mathrm{d}\phi_\mathsf{S}(s^{(i)}; \dot{s}^{(i)}) - G^* \phi_\mathsf{Z}((s^{(i)}, a^{(i)}))\|_{\mathcal{H}_\mathbb{S}}^2$$
$$= \|\widehat{S}_\mathsf{d} - \widehat{S}_\mathsf{Z}G\|_{\mathrm{HS}}^2, \quad \text{for} \quad G \in \mathrm{HS}(\mathcal{H}_\mathbb{S}, \mathcal{H}_\mathbb{Z}). \quad (18)$$

To ensure stability and prevent overfitting in this typically ill-posed estimation problem, is to add a Tikhonov regularization term to (18)

$$\widehat{G}_\gamma := \arg\min_{G \in \mathrm{HS}} \widehat{\mathcal{R}}(G) + \gamma \|G\|^2 = \widehat{S}_\mathsf{Z}^* \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathsf{d} = \begin{bmatrix} \widehat{A}_\gamma \\ \widehat{B}_\gamma \end{bmatrix}, \quad (19)$$

where $\boldsymbol{K}_\gamma = \widehat{S}_\mathsf{Z}\widehat{S}_\mathsf{Z}^* + N\gamma\boldsymbol{I}$ denotes the regularized gram matrix. Equation (19) is the *Kernel Ridge Regression* (KRR) approximation of $\mathcal{G}$ over $\mathcal{H}_\mathbb{S} \to \mathcal{H}_\mathbb{Z}$.

Ultimately, we construct a tractable approximation to (13), using the world model (**P1**), derived from (19), to obtain a value function estimate $\widehat{V} \in \mathcal{H}_\mathbb{S}$ via

$$\widehat{\mathcal{T}}(\widehat{V}) = -(\rho I - \widehat{A}_\gamma)\widehat{V} + \widehat{r}_s + \widehat{\mathcal{D}}_a(\widehat{B}_\gamma\widehat{V}), \quad (20)$$

where $\widehat{\mathcal{T}}$ can be interpreted as an approximation of the HJB operator $\mathcal{T}$ in (15), with the data-based approximations $\widehat{r}_s$ and $\widehat{\mathcal{D}}_a$ defined in the proposition below.

Proposition 1. *Let $\widehat{r}_s = \widehat{S}_\mathsf{S}^* r, \widehat{\mathcal{D}}_a = \widehat{S}_\mathsf{S}^* D_a(\cdot)$, where $\widehat{S}_\mathsf{S}^* : \mathbb{R}^N \to \mathcal{H}_\mathbb{S}$ is the adjoint of the sampling operator $\widehat{S}_\mathsf{S}$. Let the transition dynamics be described by (19). Then the flow of (20) resides in a finite-dimensional subspace*

$$\langle \dot{v}(t, \cdot), k_\mathsf{S}(s) \rangle = \langle \boldsymbol{T}(v(t, \cdot)), k_\mathsf{S}(s) \rangle, \quad (21)$$

with

$$\boldsymbol{T}(v) := -(\rho\boldsymbol{I} - \boldsymbol{A})v + (r + D_a(\boldsymbol{B}v)), \quad (22)$$

where $\widehat{S}_\mathsf{S}\phi_\mathsf{S}(s) = k_\mathsf{S}(s)$ is the sampled canonical map, $\boldsymbol{A} := \boldsymbol{K}_\gamma^{-1}\widehat{S}_\mathsf{d}\widehat{S}_\mathsf{S}^*$, $\boldsymbol{B} := \left[\mathrm{diag}(\widehat{U}e_k)\boldsymbol{A}\right]_{k \in [n_a]}$. The reward/Fenchel-dual terms in the RKHS are $r = \boldsymbol{K}_\gamma^{-1}[r_s(s^{(1)}), \ldots, r_s(s^{(N)})]^\top$ and $D_a(\boldsymbol{\lambda}) = \boldsymbol{K}_\gamma^{-1}[\mathcal{D}_a(\boldsymbol{\lambda}(s^{(i)}))]_{i \in [N]}$, respectively.

This proposition follows by substituting the estimators (19) and the representation $\widehat{V}(t, \cdot) = \widehat{S}_\mathsf{S}^* v(t, \cdot)$ into (20) and test against $\phi_\mathsf{S}(s)$ to obtain (21).

The last step toward obtaining a tractable dynamic-programming recursion (**P3**) is to discretize (21) in time. To this end, we treat the stiff linear part $(\rho\boldsymbol{I} - \boldsymbol{A})$ *implicitly* to ensure unconditional numerical stability of the linear part (He, 2013), while the nonlinear term $D_a(\boldsymbol{B}v)$ is worked out *explicitly* to avoid costly nonlinear solves.

Algorithm 1 OPERATOR-BASED CONTINUOUS-TIME OFFLINE REINFORCEMENT LEARNING (O-CTRL)

Require: $\{\dot{s}^{(i)}, (s^{(i)}, a^{(i)}), r_s(s^{(i)})\}_{i=1}^N$; state kernel $k_S(s, s') = \langle \phi_S(s), \phi_S(s') \rangle$ with $\phi_S : \mathbb{S} \to \mathcal{H}_S$; regularization $\gamma > 0$; diffusion $\epsilon > 0$; discount $\rho > 0$; timestep $\Delta t > 0$; TOL > 0; $k_{\max} \in \mathbb{N}$; dual $a^\star(\cdot)$.

REPRESENTATION

set: $(k_S(s))_i := k_S(s, s^{(i)})$, $(K_S)_{ij} := k_S(s^{(i)}, s^{(j)})$, $U := [a^{(1)}, \ldots, a^{(N)}]^\top$

WORLD MODEL

compute: $K_\gamma := K_S + K_S \odot UU^\top + N\gamma I$ ▷ Gram
compute: $(K_d)_{ij} := \langle \dot{s}^{(i)}, \nabla_{s^{(i)}} k(s^{(i)}, s^{(j)}) \rangle + \epsilon \operatorname{Tr} \nabla^2_{s^{(i)}} (k(s^{(i)}, s^{(j)}))$ ▷ target
compute: $A := K_\gamma^{-1} K_d$, $B := [\operatorname{diag}(Ue_i)A]_{i \in [n_a]}$ ▷ dynamics
compute: $r := K_\gamma^{-1} y_r$, where $y_r := [r_s(s^{(1)}), \ldots, r_s(s^{(N)})]^\top$ ▷ reward

DYNAMIC PROGRAMMING

set: $\lambda(s) := \langle I_{n_a} \otimes k_S(s), \lambda \rangle$, $M := (I + \Delta t(\rho I - A))^{-1}$ ▷ costate & propagator
set: $D_a(\lambda) := K_\gamma^{-1}[\langle a^\star(\lambda(s^{(i)})), \lambda(s^{(i)}) \rangle - c_a(s^{(i)}, a^\star(\lambda(s^{(i)})))]_{i=1}^N$ ▷ Fenchel dual
$v^{(0)} \leftarrow 0$, $k \leftarrow 0$
repeat
$\quad v^{(k+1)} \leftarrow M[v^{(k)} + \Delta t(r + D_a(Bv^{(k)}))]$ ▷ IMEX
$\quad k \leftarrow k + 1$
until $\|v^{(k)} - v^{(k-1)}\| \le \text{TOL}$ or $k = k_{\max}$

Ensure: $\widehat{V}_k(s) = \langle v^{(k)}, k_S(s) \rangle$ and $\widehat{\pi}_k(s) := a^\star(\langle I_{n_a} \otimes k_S(s), B\,v^{(k)} \rangle)$ ▷ value & policy

---

Corollary 1. *Let the step-size $\Delta t > 0$. Then the implicit-explicit IMEX flow (He, 2013; Koto, 2008; Sebastiano, 2023) update reads*

$$v^{(k+1)} = M\left[v^{(k)} + \Delta t(r + D_a(Bv^{(k)}))\right], \quad (23)$$

with $M = (I + \Delta t(\rho I - A))^{-1}$ naturally following from the implicit discretization.

This yields the operator-theoretic recursion used in Algorithm 1; each step requires one linear solve with $(I + \Delta t(\rho I - A))$ and one evaluation of $D_a(Bv^{(k)})$.

4 END-TO-END LEARNING RATES

In this section, we aim to bound $\|V^\star - \widehat{V}_k\|_{L^2}$, where $V^\star$ is the optimal value function (10) and $\widehat{V}_k$, the output of Algorithm 1, is a numerical approximation (e.g., IMEX, Corollary 1) of $\widehat{V}_T$. The latter is obtained by evolving the approximated HJB flow using $\widehat{\mathcal{T}}$ (20) for a large time $T > 0$. To capture the *end-to-end* pipeline from data to value function approximation, we also define $\widehat{V}^* = \lim_{T \to \infty} \widehat{V}_T$, the steady state of $\widehat{\mathcal{T}}$ satisfying $\widehat{\mathcal{T}}(\widehat{V}^\star) = 0$. Despite our nonlinear HJB setting, convergence of the full HJB can still be analyzed using linear operator convergence analysis (Li et al., 2022; Talwai et al., 2022; Kostic et al., 2023, 2024c). We use operator-norm error analysis (Kostic et al., 2023), which captures worst-case behavior essential for reliable policies, unlike the average-case nature of Hilbert-Schmidt bounds. In particular, the operator norm error can be written as $\mathcal{E}(\widehat{G}_\gamma) := \|\mathcal{G} - \widehat{G}_\gamma\|_{\mathcal{H}_S \to L^2}$ and for every $\delta > 0$ there exists a *finite–rank* $G_\gamma \in \operatorname{HS}(\mathcal{H}_S, \mathcal{H}_\mathbb{Z})$ with

$$\mathcal{E}(\widehat{G}_\gamma) < B(\mathcal{H}_\mathbb{Z}) + \delta.$$

If $\mathcal{H}_S$ is chosen as C_0-universal RKHS, we can find arbitrarily good finite-rank approximations of the infinitesimal generator, and thus the representation bias vanishes $B(\mathcal{H}_\mathbb{Z}) = 0$ (Mollenhauer and Koltai, 2020; Kostic et al., 2023; Bevanda et al., 2025a), which we therefore assume below.

Assumption 3. *We assume that $\mathcal{H}_S$ is a C_0-universal RKHS, $\mathcal{D}_a$ is twice continuously differentiable and globally Lipschitz on $\mathcal{H}_S$, and the state reward satisfies $r_s \in \mathcal{H}_S$. Moreover, we assume that the exact discounted HJB (10), with $\rho > 0$, has a well-defined solution $V^\star \in H^1(\mathbb{S})$ that coincides with the optimal expected discounted return, as discussed in Section 2.*

Naturally, under Assumption 3, it follows that the operators $\widehat{A}_\gamma, \widehat{B}_\gamma$ satisfy $\max\{\mathcal{E}(\widehat{A}_\gamma), \mathcal{E}(\widehat{B}_\gamma)\} \le \mathcal{E}(\widehat{G}_\gamma) = \delta$. Furthermore, the requirements on $\mathcal{D}_a$ and r_s ensure that the iterates of Algorithm 1 satisfy $\widehat{V}_k \in \mathcal{H}_S$. This, in turn, allows us to bound the L^2-norm of the approximation error $\|\widehat{V}_k - V^\star\|_{L^2}$.

Theorem 1. *Suppose that Assumptions 1 to 3 and the conditions of Proposition 1 and Corollary 1 hold. Then*

under the zero-initial condition $\widehat{V}_0 = 0$, and provided that $\delta = \mathcal{E}(\widehat{G}_\gamma)$ is sufficiently small,

$$\|V^\star - \widehat{V}_k\|_{L^2} \le \underbrace{(\widehat{\lambda}_{\text{gap}}+\rho)^{-1}\delta}_{\text{learning}} + \underbrace{\mathcal{O}((\Delta t)^p)}_{\text{discretization}} \quad (24a)$$

$$+ \underbrace{\kappa\, e^{-(\widehat{\lambda}_{\text{gap}}+\rho)k\Delta t}\|\widehat{V}^\star\|_{L^2}}_{\text{convergence}}, \quad (24b)$$

where p is the discretization order, and $\kappa > 0$ a constant. Here, $\widehat{\lambda}_{\text{gap}}$ denotes the spectral gap of the estimated closed-loop generator $P^{\widehat{\pi}^\star}\widehat{G}_\gamma$ in (3) under the stationary policy $\widehat{\pi}^\star$.

Proof sketch. By the triangle inequality, $\|V^\star - \widehat{V}_k\|_{L^2} \le \|V^\star - \widehat{V}^\star\|_{L^2} + \|\widehat{V}^\star - \widehat{V}_T\|_{L^2} + \|\widehat{V}_T - \widehat{V}_k\|_{L^2}$, for the learning, convergence and discretization error respectively. The discretization error, for a method of order p, is $\mathcal{O}((\Delta t)^p)$. For the learning and convergence error terms, the exact HJB operator $\mathcal{T}$ in (13) and its approximation $\widehat{\mathcal{T}}$ (20) are Lipschitz in their norms (L^2 and $\mathcal{H}_{\mathbb{S}}$ respectively), and the exact HJB flow is globally exponentially convergent (Appendix). Using local Lipschitz continuity of the spectral gap under small perturbations (Kloeckner, 2018; Kato, 2013), the approximate HJB remains locally exponentially convergent for small $(\mathcal{E}(\widehat{A}_\gamma), \mathcal{E}(\widehat{B}_\gamma))$, which yields the result. □

While a universal RKHS guarantees there are arbitrarily accurate value function approximations, the convergence to the true value function can be arbitrarily slow without additional assumptions. For that, we will use classical source conditions on the regularity of the inverse problems (Engl and Ramlau, 2015) to provide a convergence result.

Corollary 2. *Let $\alpha \in (1, 2]$ denote the regularity of $\mathcal{G}$ and $\beta \in (0, 1]$ the spectral decay rate of $\mathcal{H}_{\mathbb{Z}}$, and choose $\gamma \asymp N^{-\frac{1}{\alpha+\beta}}$. Then, for any $\xi \in (0, 1)$ there exists $c > 0$ such that, with probability at least $1 - \xi$ over an i.i.d. sample $\mathbb{D}_N$, the learning error in (24a) satisfies a finite-sample bound.*

$$\|V^\star - \widehat{V}^\star\|_{L^2} \le c\,(\widehat{\lambda}_{\text{gap}}+\rho)^{-1} N^{-\frac{\alpha}{2(\alpha+\beta)}} \ln \xi^{-1}. \quad (25)$$

This follows directly from the non-asymptotic error bound of (Kostic et al., 2023) for the KRR estimator $\widehat{G}_\gamma$, which yields a *finite-sample rate of convergence* to the optimal value function in Theorem 1. Together with Corollary 2, this result highlights interpretable quantities that govern the difficulty of continuous-time offline RL, such as horizon length, discretization accuracy, and the amount of data. It also links environment properties and world-model complexity (**P1**) to errors in attaining global optimality. Informally, a larger α

indicates better RKHS alignment with the true operator image, while a smaller β reflects faster spectral decay. Both effects improve sample complexity, and in the favorable limit, the error reaches the $N^{-1/2}$ rate.

5 NUMERICAL EXAMPLES

Implementation We evaluate our learning error rates on linear and nonlinear process dynamics (Figure 2). While these are often studied using different policy classes, linear and nonlinear (Tu and Recht, 2019), our convergence analysis covers both on equal footing – without any parametric assumptions (Recht, 2019). We run our proposed algorithm (1) over 8 seeds and gather i.i.d. data to form the quantiles and means in Figure 2. In all cases, we use a squared exponential (SE) kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\|\boldsymbol{x}-\boldsymbol{y}\|^2/2\sigma^2\right)$ while the data samples are drawn randomly from the set $[3, 3] \times [3.5, 3.5]$ for the two first examples.

Linear SDE with Additive Action Linear-quadratic (LQ) control problems play an important role in the control literature. They provide explicit solutions and, often, nonlinear ones can be approximated by LQ ones (Zhao et al., 2023). To validate our findings, we study the value function convergence for an Ornstein-Uhlenbeck process $dS_t = (-S_t + a_t)dt + \sqrt{2\epsilon}\, dW_t$ (Houska, 2025) where the optimal value function is $V_\infty(s) = s^2$ when setting $\rho = 0$, for any ϵ, we set $\epsilon = 0.01$, and a reward function $r(s, a) = -3s^2 - a^2$. The hyperparameter for the used SE kernel is $\sigma = 10$.

Nonlinear SDE with Affine Action We transfer to a nonlinear setting, using an action-affine benchmark system $dS_t = (f(S_t) + g(S_t)a_t)dt + \sqrt{2\epsilon}\, dW_t$, where $g(s) = \frac{1}{2} + \sin(s)$ and $f(s) = -\frac{1}{2}(1 - g(s)^2)$ (Doyle et al., 1996). Here, the reward is $r(s, a) = -s^2 - a^2$ and, practically, the optimal value function approaches $V_\infty = s^2$ as ϵ tends to zero. The hyperparameter for the used SE kernel is $\sigma = 1$ and we set $\epsilon = 0.01$.

Pendulum-Gym We evaluate Algorithm 1 on Gymnasium `Pendulum-v1` (Towers et al., 2024). The action is a torque $a \in [-2, 2]$, and the observation is $\boldsymbol{s} = (\cos\theta, \sin\theta, \dot{\theta})$ with $\cos\theta, \sin\theta \in [-1, 1]$ and $\dot{\theta} \in [-8, 8]$. Episodes truncate at 200 steps. Following the official reward, we split the state term and action cost as $r_{\boldsymbol{s}}(\theta, \dot{\theta}) = -(\theta^2 + 0.1\,\dot{\theta}^2)$ and $c_{\boldsymbol{a}}(a) = 0.001\,a^2$, so the maximum achievable reward is 0 (upright, zero velocity, zero torque). We run Algorithm 1 with $\rho = 0.1$, $\sigma = 3$, $\gamma = 10^{-7}$, and $k_{\max} = 1000$. For training, we use two offline `d3rlpy` datasets ("Random", "Replay"), subsample 8000 (state, action, next state) tuples, and obtain infinitesimal samples via finite differences. The resulting value function is shown in Figure 1. Table 1
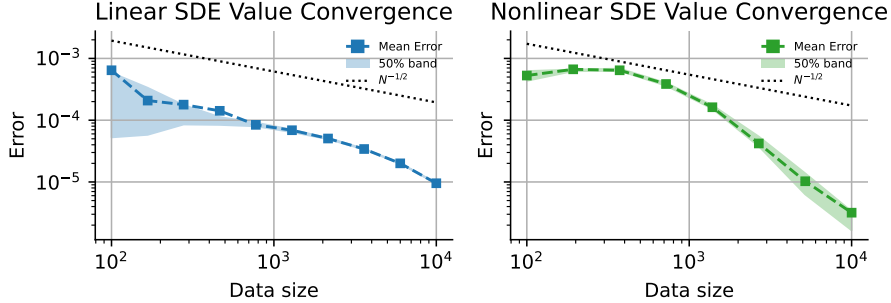
Figure 2: Value function learning. *Left:* A linear SDE with quadratic costs. *Right:* Convergence for a nonlinear SDE with quadratic costs. The error convergence confirms our worst-case analysis.

reports mean ± standard deviation of episode return over 50 i.i.d. rollouts of length 200. Baselines were used largely with default settings and limited tuning; our method was likewise not heavily tuned and trained on the same 8k samples per dataset.

Discussion Offline RL faces distribution shift and extrapolation error, where policies trained on fixed data may query actions outside the data support, causing overestimation and instability. *CQL* (Kumar et al., 2020) uses a pessimistic objective suited to suboptimal data and, in our results, does better on Random, where the experience is collected with a random policy, than on Replay (Table 1). *TD3+BC* (Fujimoto and Gu, 2021) anchors the policy to the behavior data and is sensitive to dataset quality, resulting in overall underperformance, as neither dataset is expert-level, although it improves on Replay. *IQL* (Kostrikov et al., 2022) favors actions that look better than average within the dataset, consistent with stronger results on Replay. These methods provide a practical baseline suite for evaluating O-CTRL under identical offline data and evaluation (Table 1). As additional context, we also report a Soft Actor Critic expert trained online and evaluated with the same protocol. O-CTRL is competitive on Random and surpasses the expert on Replay. Note that, unlike value-based offline RL methods tied to the behavior distribution, O-CTRL first learns a reward-free world model from fixed data and then optimizes the task objective in that model. This decoupling makes performance less sensitive to whether the dataset is expert, medium, or random, and allows reusing the same learned model for different rewards without additional data collection. These properties make O-CTRL a promising direction for offline RL.

6 CONCLUSION

We studied offline reinforcement learning with continuous-time dynamics and continuous state-action

---

	Mean Episode Reward (± std)	
Algorithm	Random	Replay
Ours	$\mathbf{-141.71 \pm 84.23}$	$\mathbf{-104.36 \pm 83.00}$
CQL	-293.09 ± 206.68	-451.00 ± 530.91
IQL	-313.00 ± 222.91	-284.89 ± 196.59
TD3+BC	-1181.08 ± 347.40	-748.11 ± 420.18
Expert	-130.98 ± 79.40	

Table 1: Pendulum results on `d3rlpy` offline datasets (Random, Replay). Higher (less negative) is better. The expert policy is a Stable-Baselines3 SAC agent (Pendulum-v1)[3] trained online (dataset N/A).

spaces. We introduced a model-based, reward-free algorithm that splits value function learning into two steps. The first step is learning the generator of the controlled diffusion process from data through operator regression in RKHS. The second step is to solve the Hamilton–Jacobi–Bellman PDE in the RKHS. We analyze the method by propagating operator learning errors through the empirical HJB and prove that the resulting time-stepping scheme is consistent and convergent. Operator-norm error analysis yields explicit learning rates that relate data size and kernel smoothness to value function suboptimality. The bounds reveal how the generator's spectral gap, discount factor, and regularity influence the offline RL problem. Examples confirm the predicted rates. Our work presents the first end-to-end learning theory for continuous-time offline reinforcement learning, highlighting the power of operator methods as an analytical tool.

Limitations and Open Questions Although we obtain novel insights into continuous-time offline RL, some key questions remain open.

Data The proposed generator learning scheme requires independent data, which, depending on the application, might be costly to obtain. Extending the analysis to dependent (trajectory) data as recently proposed in Mirzaei et al. (2025) would increase its applicability.

---

[3] `https://huggingface.co/sb3/sac-Pendulum-v1`

The datasets we use include state measurements. A natural question is whether similar results still hold for partially observed data.

Analysis In general, we provide upper bounds on the errors, thus providing a pessimistic analysis of the error. Lower bounds could help uncover different effects and lead to novel algorithms. In particular, data-dependent bounds could shed light on the maximum suboptimality of concrete actions and help collect new data points in episodic learning.

Representations In our analysis, we assume a bound on the RKHS norm of the exact value function and resort to a universal RKHS. While this approach allows for elegant analysis, learning finite-dimensional features for a specific problem and dealing with misspecification is often practical.

Algorithm Our upper bound requires $\rho + \lambda_{\mathrm{gap}} > L_D \mathcal{E}(\widehat{G})$ – sufficient stability, and discount to overcome learning errors for consistency within the infinite horizon problem. Ensuring this condition algorithmically remains an open question as we don't know the value of λ_{gap} a priori.

Acknowledgements

Supplementary Materials

The supplementary material is organized as follows:

- Section A summarizes important notations used throughout the paper and the appendix, in the form of a reference table in Table 2.

- Section B provides a comparison of this work with the studies we consider most relevant to the analysis of offline RL.

- Section C provides a detailed review of key concepts from stochastic optimal control that are used throughout the paper and are essential for proving Theorem 1.

- Section D offers background information on operator regression and kernel-based learning, leading to the derivations of our world model. Additionally, it reviews recent results on operator-norm error analysis, which can be incorporated into Theorem 2 to derive an explicit and interpretable finite-sample error bound.

- Section E presents the proofs for Proposition 1 and Corollary 1, which are necessary for our O-CTRL algorithm outlined in Algorithm 1. In the second part, we prove Theorem 1 and Corollary 2.

A Extended Notation

Table 2 summarizes the notation used throughout the paper.

Table 2: Summary of relevant notation

Notation	Meaning
$[n]$	Interval Set $\{1, \dots, n\}$ for an integer n
$C^k(\mathbb{S})$	k-times continuously differentiable functions on $\mathbb{S}$
$L^k_\mu(\mathbb{S})$	L^k-integrable functions w.r.t. a measure μ
$H^k_\mu(\mathbb{S})$	k-times weakly differentiable functions with $L^2_\mu(\mathbb{S})$-integrable derivatives
$M_+(\mathbb{S})$	Set of Borel probability measures on $\mathbb{S}$
$\odot$	Hadamard (element-wise) product
$\circledcirc$	Khatri-Rao product
∇, ∇^2, Δ	Gradient operator, Hessian Matrix and Laplacian respectively
Tr	Trace operator
ρ	Discount
$r, r_s, c_{\boldsymbol{a}}$	Reward, state reward and action penalty
$\mathcal{D}_{\boldsymbol{a}}(\cdot)$	Fenchel conjugate of the action penalty
$\mathcal{G}$	Infinitesimal generator of the SDE
$\mathcal{A}$	Autonomous part of $\mathcal{G}$
$\mathcal{B}$	Controlled part of $\mathcal{G}$
$P^{\boldsymbol{\pi}}$	Substitution operator for a policy $\boldsymbol{\pi}$
$\mathcal{T}$	Infinitesimal Operator for the HJB flow
$\mathfrak{T}(t, \cdot)$	Nonlinear semigroup associated with the HJB such that $\mathfrak{T}(t, V_0) = V(t, \cdot)$
$V^\star$	Optimal value function (steady state satisfying $\mathcal{T}(V^\star) = 0$)
$\widehat{\mathcal{T}}$	Empirical equivalent of $\mathcal{T}$
$\widehat{V}_T$	Approximated value function obtained by evolving the HJB flow with $\widehat{\mathcal{T}}$ for a large time T
$\widehat{V}_k$	Numerical approximation of $\widehat{V}_T$, output of the O-CTRL Algorithm at time step k
$\widehat{V}^\star$	Approximated optimal value function (steady state satisfying $\widehat{\mathcal{T}}(\widehat{V}^\star) = 0$)
$\mathrm{HS}(\mathcal{F}, \mathcal{Y})$	Hilbert–Schmidt operators $A : \mathcal{F} \to \mathcal{Y}$
$\mathcal{H}_\mathbb{S}, \mathcal{H}_\mathbb{Z}$	Input/output RKHSs on $\mathbb{S}$ and $\mathbb{Z}$
k_S	Symmetric, bounded, positive-definite kernel $\mathbb{S} \times \mathbb{S} \to \mathbb{R}$ associated with $\mathcal{H}_\mathbb{S}$
$\phi_\mathsf{S}(\boldsymbol{s})$	Canonical feature map associated to $\boldsymbol{s} \in \mathbb{S}$ (analogously $k_\mathsf{S}(\cdot, \boldsymbol{s}) \in \mathcal{H}_\mathbb{S}$)
$\phi_\mathsf{Z}(\boldsymbol{z})$	Canonical feature map associated to $\boldsymbol{z} \in \mathbb{Z}$ (analogously $k_\mathsf{Z}(\cdot, \boldsymbol{z}) \in \mathcal{H}_\mathbb{Z}$)
S_μ	Canonical inclusion $\mathcal{H}_\mathbb{S} \hookrightarrow H^1_\mu(\mathbb{S})$ of the input RKHS into $H^1_\mu(\mathbb{S})$
S_η	Canonical inclusion $\mathcal{H}_\mathbb{Z} \hookrightarrow L^k_\eta(\mathbb{Z})$ of the output RKHS into $L^k_\eta(\mathbb{Z})$
$C_{\mathsf{ZZ}}, C_{\mathsf{Zd}}$	Covariance and Cross-covariance operators
$G_\gamma, A_\gamma, B_\gamma$	Population KRR estimators
$\widehat{S}_\mathsf{Z}$	State-action sampling operator $\mathcal{H}_\mathbb{Z} \to \mathbb{R}^N$
$\widehat{S}_\mathsf{S}$	State sampling operator $\mathcal{H}_\mathbb{S} \to \mathbb{R}^N$
$\widehat{S}_\mathsf{d}$	Target sampling operator $\mathcal{H}_\mathbb{S} \to \mathbb{R}^N$
$\widehat{G}_\gamma, \widehat{A}_\gamma, \widehat{B}_\gamma$	Empirical KRR estimators
$\boldsymbol{K}_\gamma$	Regularized Gram matrix $\in \mathbb{R}^{N \times N}$
$\boldsymbol{K}_\mathsf{d}$	Target kernel matrix $\in \mathbb{R}^{N \times N}$
$\mathcal{E}(\cdot)$	Operator norm error
δ	Upper bound on the Operator norm $\mathcal{E}(\widehat{G}_\gamma)$
$\widehat{\lambda}_{\mathrm{gap}}$	Spectral gap of the estimated closed loop generator under the approximated optimal policy

B   Extended Related Work

In this section, we focus on situating our end-to-end learning theory for continuous-time offline RL within the landscape of related theoretical results in offline RL. Rather than providing a complete survey of offline RL, we focus on works with formal guarantees and direct readers to Levine et al. (2020); Prudencio et al. (2023) for broader overviews. A particularity of our approach is that, unlike standard offline RL, it separates the learning of system dynamics from the downstream task defined by the reward, enabling modular and reward-free task optimization. This separation enables the subsequent optimization of arbitrary task objectives without requiring additional environment interaction, highlighting a *reward-free, modular learning* paradigm within the offline RL setting. At present, to the best of our knowledge, no related work offers a non-asymptotic end-to-end analysis for continuous-time, offline, reward-free RL. By end-to-end, we mean: from dynamical system data to value function approximation that holds for any continuously differentiable state reward in the RKHS hypothesis. In other words, our end-to-end guarantees aim to quantify the gap between an approximated optimal value function produced by a reinforcement learning algorithm and the optimal value function. Below and in Table 3 is a comparison of theoretical RL results in continuous and discrete time.

Jia and Zhou (2022a,b, 2023) analyze the convergence, but provide no insight into the convergence rate of RL with respect to data and system properties. For example, while the works of Jia and Zhou (2022a,b, 2023) explicitly handle time discretization and prove convergence to an underlying solution as $\Delta t \to 0$, they do not analyze the errors made by choosing a parametric function class for approximation and having limited data. Further, their statements are asymptotic and do not include quantitative rates, including constants related to the general offline RL problem. In contrast, our proposed operator learning framework enables the joint analysis of data, environment, and task, allowing for the derivation of non-asymptotic rates. Nonparametric analysis, such as that for Neural Tangent Kernels (Jacot et al., 2018), has been proven useful for understanding overparameterized networks. Extending such analysis to general infinite-dimensional settings is more subtle, as it involves two distinct spaces: L^2 (the regression space) and $\mathcal{H}_{\mathbb{S}}$ (encoding the nonlinear representation) (Meunier et al., 2025). A key advantage of nonparametric approaches is that, by using suitable infinite-dimensional RKHS (Steinwart and Christmann, 2008), one can approximate arbitrary operators to arbitrary accuracy, providing flexibility beyond fixed parametric models. Building on this perspective, in the discrete-time setting, Novelli et al. (2024) provides convergence rates with data, but cannot handle continuous action spaces and requires stronger assumptions.

Paper	Time Model	$\mathbb{S}/\mathbb{A}$	Data setting	Learning Method	Guarantee	Assumptions
Ours	**(C)**	**(C)**	**Offline**	HJB solution with learned operator model	Sub-optimality dependent on (data, environment, task, and discretization) jointly	Continuous dynamics, data density on $\mathbb{S} \times \mathbb{A}$, reward in (RKHS) $\mathcal{H}_{\mathbb{S}}$
Jia and Zhou (2022a,b, 2023)	**(C)**	**(C)**	Online / Episodic	TD policy evaluation, Actor-critic	Convergence with time-discretization, entropy reg.	Continuous dynamics, Markov Diffusion Process
Novelli et al. (2024)	(D)	**(C)**/Finite	**Offline** / Episodic	Policy mirror-descent with learned operator model	Convergence of Primal-dual gap given CME error rate with data	Linear MDP, data density on $\mathbb{S} \times \mathbb{A}$, reward in (RKHS) $\mathcal{H}_{\mathbb{S}}$

Table 3: Comparison of theoretical results for related RL frameworks. Time model: (C) = Continuous, (D) = Discrete.

Other offline RL methods providing guarantees or rates are in discrete-time and continuous state/action spaces with excess risk bounds guarantees (Farahmand, 2011), sup-norm error (Antos et al., 2007), or are restricted to finite state/action spaces (Kidambi et al., 2020; Chen and Jiang, 2019; Ayoub et al., 2024). Below is a table summarizing the settings and learning methods of those works.

Paper	Time Model	$\mathbb{S}/\mathbb{A}$	Data setting	Learning Method
Farahmand (2011) 2011	(D)	**(C)**	Online & **Offline**	FQI/LSTD with regularization
Antos et al. (2007)	(D)	**(C)**	**Offline**	Continuous-action fitted-Q
Chen and Jiang (2019)	(D)	Finite	**Offline**	Distribution alignment w/ FQI
Ayoub et al. (2024)	(D)	Finite	**Offline**	FQI via log-loss
Kidambi et al. (2020)	(D)	Finite	**Offline**	Pessimistic model-based policy

Table 4: Comparison of related RL frameworks, (C) = Continuous, (D) = Discrete. FQI = Fitted Q-Iteration, LSTD = Least Squares Temporal Difference Learning.

C   CONTINUOUS-TIME OPTIMAL CONTROL

We begin by recalling fundamental results from continuous-time optimal control, with emphasis on the Hamilton–Jacobi–Bellman (HJB) equation. We derive the spectral gap of the infinitesimal generator of the optimally controlled Markov process and establish exponential convergence of the HJB flow under our assumptions, a key step in proving Theorem 1 in Section E.

C.1   Stationary HJB

Recall that our objective is to find the optimal value function maximizing the expected cumulative discounted rewards in (5), such that

$$V^\star := \max_{\|\boldsymbol{\pi}\|_{L^\infty} < \infty} \left(\rho I - P^{\boldsymbol{\pi}}\mathcal{G}\right)^{-1} P^{\boldsymbol{\pi}} r.$$

We also defined the operator $P^{\boldsymbol{\pi}} : L^2(\mathbb{S} \times \mathbb{A}) \to L^2(\mathbb{S})$. substituting open-loop actions with the output of a policy $\boldsymbol{a} = \boldsymbol{\pi}(\boldsymbol{s})$, e.g. $[P^{\boldsymbol{\pi}} r](\boldsymbol{s}) = r(\boldsymbol{s}, \boldsymbol{\pi}(\boldsymbol{s}))$, for a measurable $\boldsymbol{\pi} \in L^\infty(\mathbb{S})$. This allows us to rewrite the generator associated to (1) under a policy as

$$[P^{\boldsymbol{\pi}}\mathcal{G}\phi](\boldsymbol{s}) = [(\mathcal{A} + P^{\boldsymbol{\pi}}\mathcal{B})\phi](\boldsymbol{s}) = [\mathcal{A}\phi](\boldsymbol{s}) + [P^{\boldsymbol{\pi}}\mathcal{B}\phi](\boldsymbol{s}),$$

for $\phi \in D(\mathcal{G}) \subseteq H^1(\mathbb{S})$, a space with sufficient regularity and the shorthand definitions

$$[\mathcal{A}\phi](\boldsymbol{s}) = \nabla_{\boldsymbol{s}}\phi(\boldsymbol{s})^\top \boldsymbol{f}(\boldsymbol{s}) + \epsilon \Delta_{\boldsymbol{s}}\phi(\boldsymbol{s}), \qquad \text{and} \qquad [P^{\boldsymbol{\pi}}\mathcal{B}\phi](\boldsymbol{s}) = \nabla_{\boldsymbol{s}}\phi(\boldsymbol{s})^\top \boldsymbol{G}(\boldsymbol{s})\,\boldsymbol{\pi}(\boldsymbol{s}).$$

Now, we will consider action-affine dynamics (cf. (1)), covering a broad range of cyber-physical systems (Nijmeijer and van der Schaft, 1996) and most tasks relevant to robotic Foundation Models (Tölle et al., 2025).

By the principle of optimality, under Assumption 1, the optimal value function $V^\star \in H^1(\mathbb{S})$—if it exists—satisfies the discounted HJB equation on $\mathbb{S} = \mathbb{R}^{n_s}$ (Fleming and Soner, 2006; Doya, 2000; Lutter et al., 2020),

$$\max_{\|\boldsymbol{\pi}\|_{L^\infty} < \infty} \left\{ [P^{\boldsymbol{\pi}}\mathcal{G}V^\star](\boldsymbol{s}) + r(\boldsymbol{s}, \boldsymbol{\pi}(s)) \right\} = \rho V^\star(\boldsymbol{s}), \tag{26}$$

Recall that the system (1) is affine in the actions, and suppose that the reward structure in Assumption 2 holds.

We define the $\mathbb{A}$-valued operator $\mathsf{B} : D(\mathcal{G}) \to L^2(\mathbb{S}; \mathbb{A})$ such that $[\mathsf{B}\phi](\boldsymbol{s}) := \nabla_{\boldsymbol{s}}\phi(\boldsymbol{s})^\top \boldsymbol{G}(\boldsymbol{s})$, where $L^2(\mathbb{S}; \mathbb{A})$ is the Bochner L^2 space of $\mathbb{A}$-valued, square-integrable function on $\mathbb{S}$. In particular, with $\mathbb{A} = \mathbb{R}^{n_a}$, we have the natural isomorphism $L^2(\mathbb{S}; \mathbb{R}^{n_a}) \cong (L^2(\mathbb{S}))^{n_a} \cong \mathbb{R}^{n_a} \otimes L^2(\mathbb{S})$. Using the inner product in $\mathbb{R}^{n_a}$, we "curry" B to obtain a scalar map on $\mathbb{S} \times \mathbb{R}^{n_a}$:

$$[\mathcal{B}\phi](\boldsymbol{s}, \boldsymbol{a}) = \langle [\mathsf{B}\phi](\boldsymbol{s}), \boldsymbol{a} \rangle_{\mathbb{R}^{n_a}}$$

Setting $\boldsymbol{a} = \boldsymbol{\pi}(\boldsymbol{s})$ yields

$$[P^{\boldsymbol{\pi}}\mathcal{B}\phi](\boldsymbol{s}) = \langle [\mathsf{B}\phi](\boldsymbol{s}), \boldsymbol{\pi}(\boldsymbol{s})\rangle_{\mathbb{R}^{n_a}} = \nabla_{\boldsymbol{s}}\phi(\boldsymbol{s})^{\top}\boldsymbol{G}(\boldsymbol{s})\,\boldsymbol{\pi}(\boldsymbol{s}), \quad \text{a.e.}$$

Note that if $D(\mathcal{G}) \subset H^y(\mathbb{S})$ with $y > \frac{n_s}{2} + 1$, then by Sobolev Embedding $H^y(\mathbb{S}) \hookrightarrow C^1(\mathbb{S})$, $\mathsf{B}\phi$ has a continuous representative and the identities above hold pointwise. Finally, (26) can be rewritten as

$$[\mathcal{A}V^{\star}](\boldsymbol{s}) + r_{\boldsymbol{s}}(\boldsymbol{s}) + \max_{\|\boldsymbol{\pi}\|_{L^{\infty}} < \infty} \{\langle \boldsymbol{\pi}(\boldsymbol{s}), [\mathsf{B}V^{\star}](\boldsymbol{s})\rangle - c_{\boldsymbol{\pi}}(\boldsymbol{s}, \boldsymbol{\pi}(\boldsymbol{s}))\} = \rho V^{\star}(\boldsymbol{s}) \tag{27}$$

To facilitate readability, we will henceforth write $\mathcal{B}$ instead of B, implicitly using the isomorphism to switch between the $\mathbb{R}^{n_a}$-valued and the scalarized view, when no confusion can arise. Next, we identify the maximization term with the Fenchel conjugate of the action penalty, which under Assumption 2 is well defined,

$$\mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda}) \coloneqq \max_{\boldsymbol{a}}\{\langle \boldsymbol{a}, \boldsymbol{\lambda}\rangle - c_{\boldsymbol{a}}(\boldsymbol{s}, \boldsymbol{a})\}, \tag{28}$$

and admits an unique maximizer $\boldsymbol{a}^{\star}(\boldsymbol{\lambda}) = \nabla \mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda})$ (Boyd, 2004). Note that this also follows directly from the Fenchel-Young inequality (Boyd, 2004), which states:

$$c_{\boldsymbol{a}}(\boldsymbol{s}, \boldsymbol{a}) + \mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda}) \geq \langle \boldsymbol{a}, \boldsymbol{\lambda}\rangle \tag{29}$$

Equality holds if and only if $\boldsymbol{a} = \nabla \mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda})$. Furthermore, on a compact convex set $\mathbb{A}$, $\mathcal{D}_{\boldsymbol{a} \in \mathbb{A}}$ satisfies the following property.

Proposition 2 (Fenchel Lipschitzness). *$\mathcal{D}_{\boldsymbol{a} \in \mathbb{A}}$ is globally Lipschitz continuous, with $L_{\mathcal{D}} \leq \sup_{\boldsymbol{a} \in \mathbb{A}}\|\boldsymbol{a}\|$.*

Proof. Given action constraints defined by the compact convex set $\mathbb{A}$, the gradient of the Fenchel conjugate satisfies $\nabla \mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda}) \in \mathbb{A}$ and is bounded. Thus, $\|\nabla \mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda})\|_2 \leq \sup_{\boldsymbol{a} \in \mathbb{A}}\|\boldsymbol{a}\|$, and the Lipschitzness of $\mathcal{D}_{\boldsymbol{a}}$ follows. $\square$

Remark 1 (Smooth Approximations). *In cases where $c_{\boldsymbol{a}}$ is not sufficiently smooth, or box constraints are imposed via an indicator function, smooth approximations of $\mathcal{D}_{\boldsymbol{a} \in \mathbb{A}}$ can be constructed using regularization techniques such as the Moreau envelope or Nesterov smoothing (Moreau, 1965; Nesterov, 2005) enabling perturbation analysis of the HJB.*

Remark 2 (Explicit Fenchel conjugate). *Note that if $c_{\boldsymbol{a}}$ is a Linear Quadratic Regulator (LQR)-style tracking objective, with quadratic state and action penalty terms, we can explicitly work out the Fenchel conjugate. Another explicit example is obtained for quadratic action penalties with a diagonal weight matrix and simple action bounds, yielding simple saturation functions for the components of $\boldsymbol{a}^{\star}$.*

C.2 THE HAMILTON-JACOBI-BELLMAN PDE

For each $t \geq 0$, we define the nonlinear semigroup $\mathfrak{T}(t, \cdot) : L^2 \to L^2$ by $\mathfrak{T}(t, V_0) = V(t, \cdot)$, where V_0 is the initial cost function and $V(t, \cdot) \in L^2$ is the unique solution of

$$\dot{V}(t, \boldsymbol{s}) = -(\rho I - \mathcal{A})V(t, \boldsymbol{s}) + r_{\boldsymbol{s}} + D_{\boldsymbol{a}}(\mathcal{B}V(t, \boldsymbol{s})), \quad V(0, \boldsymbol{s}) = V_0(\boldsymbol{s}) \tag{30}$$

on $\mathbb{S} = \mathbb{R}^{n_s}$ and understood in the Bochner space $W(0, T; L^2(\mathbb{S}), H^1(\mathbb{S}))$ (Hinze et al., 2009, Chapter 1.3). We introduce the shorthand $\mathcal{T} : H^1(\mathbb{S}) \to L^2(\mathbb{S})$ to denote the infinitesimal generator of $\mathfrak{T}$ – the right-hand side operator in the first equation in (30) and this notation makes the sign explicit to distinguish from $\mathcal{T}$.

We are now ready to characterize the spectral gap λ_{gap}, which is essential to prove the exponential convergence of the HJB discussed in the proof sketch of Theorem 1. The optimal policy $\boldsymbol{\pi}^{\star}$ can be recovered as $\boldsymbol{\pi}^{\star}(t, \boldsymbol{s}) = \boldsymbol{a}^{\star}(\mathcal{B}V(t, \boldsymbol{s}))$. Hence, along optimal trajectories,

$$\mathcal{T}(V) = (\mathcal{A} - \rho I)V + r_{\boldsymbol{s}} + \mathcal{D}_{\boldsymbol{a}}(\mathcal{B}V) = (\mathcal{A}^* + \mathcal{B}^*\boldsymbol{\pi}^{\star})^*V - \rho V + r_{\boldsymbol{s}} - c_{\boldsymbol{a}}(\cdot, \boldsymbol{\pi}^{\star}). \tag{31}$$

Here, the operator $\mathcal{A}^* + \mathcal{B}^*\boldsymbol{\pi}^{\star}$ can be interpreted as the Fokker–Planck–Kolmogorov (FPK) operator of the optimally controlled SDE; see Houska (2025). Moreover, if Assumptions 2 and 1 hold, we obtain (since $\boldsymbol{\pi}^{\star}$ is optimal and thus stationary) that the Fréchet derivative of $\mathcal{T}$ at $V^{\star}$ satisfies

$$\frac{\partial \mathcal{T}(V^{\star})}{\partial V} = (\mathcal{A}^* + \mathcal{B}^*\boldsymbol{\pi}^{\star})^* - \rho I, \tag{32}$$

where $\bar{\lambda}$ denotes the complex conjugate of λ, since (32) uses the adjoint of $(\mathcal{A}^* + \mathcal{B}^* \boldsymbol{\pi}^\star)$. Consequently, with ,

$$\sigma\big(\tfrac{\partial \mathcal{T}(V^\star)}{\partial V}\big) = \big\{ -\rho + \overline{\lambda} \; : \; \lambda \in \sigma(\mathcal{A}^* + \mathcal{B}^* \boldsymbol{\pi}^\star) \big\}. \tag{33}$$

where $\sigma(\cdot)$ denotes the spectrum of a linear operator or matrix. As seen in (33), the positive discount $\rho > 0$ shifts the spectrum, thereby acting as a stabilizing term. This also explains why we run the HJB equation using the descent dynamics $\dot{V} = \mathcal{T}(V)$, which exhibits stable eigenvalues (33). Under conditions ensuring the HJB equation is well-posed (existence and uniqueness of viscosity solutions) (see Houska (2025); Fleming and Soner (2006) for details), the operator $\mathcal{A}^* + \mathcal{B}^* \boldsymbol{\pi}^\star$ is the infinitesimal generator of the optimally controlled Markov process and therefore has non-positive eigenvalues, i.e., $\mathrm{Re}(\lambda) \leq 0$.

Remark 3. *The HJB is generally written as a final–value problem in the physical time τ,*

$$-\dot{V}(\tau, \boldsymbol{s}) = \mathcal{T}\big(V(\tau, \cdot)\big)(\boldsymbol{s}), \qquad V(T, \boldsymbol{s}) = V_T(\boldsymbol{s}),$$

*and is solved from T down to 0. Equivalently, reparametrizing with the time–to–go $t = T - \tau$ and (reusing V for the reparametrized function) gives the initial-value problem in (30) with $V(0, \boldsymbol{s}) = V_T(\boldsymbol{s})$. This "descent" evolution shares the same stationary solution (satisfying $\mathcal{T}(V^\star) = 0$) and preserves the correct parabolic sign, enabling **stable** forward integration in t (c.f. (33)).*

We denote the spectral gap of the operator $\mathcal{A}^* + \mathcal{B}^* \boldsymbol{\pi}^\star$ by

$$\lambda_{\mathrm{gap}} := -\mathrm{Re}\big(\lambda_2(\mathcal{A}^* + \mathcal{B}^* \boldsymbol{\pi}^\star)\big) \geq 0, \tag{34}$$

the negative real part of the second largest eigenvalue of the FPK operator $\mathcal{A}^* + \mathcal{B}^* \boldsymbol{\pi}^\star$.

Remark 4. *Note that our definition of the spectral gap is always based on the second eigenvalue of an operator, as the first eigenvalue of $\mathcal{A}^* + \mathcal{B}^* \boldsymbol{\pi}^\star$ is equal to 0.*

As discussed in (Houska, 2025, Theorem 2), and under Assumptions 1 to 3 , λ_{gap} coincides with the global exponential convergence rate of the undiscounted HJB. Therefore, under strictly positive discount $\rho > 0$, we get $\lambda_{\mathrm{gap}} + \rho > 0$ and the following result.

Lemma 1 (Exponential convergence of the HJB)**.** *Suppose the discounted HJB equation $\mathcal{T}(V^\star) = 0$ admits a solution $V^\star \in H^1(\mathbb{S})$ such that*

$$\lim_{t \to \infty} \mathfrak{T}(t, V) = V^\star \quad \text{in } H^1(\mathbb{S}).$$

Then there exist constants $C < \infty$, $\lambda_{\mathrm{gap}} \geq 0$, and $\rho > 0$ such that

$$\big\| \mathfrak{T}(t, V_0) - V^\star \big\|_{L^2} \; \leq \; C \, e^{-(\lambda_{\mathrm{gap}} + \rho)\, t} \, \big\| V_0 - V^\star \big\|_{L^2}, \tag{35}$$

with $\lambda_{\mathrm{gap}} + \rho > 0$.

Remark 5. *The statement of Lemma 1 can be further strengthened if additional regularity assumptions are satisfied. For instance, on open bounded convex domains $\mathbb{S}$, with Neumann boundary conditions for V, $\nabla V(\cdot, t)\eta \mid_{\partial \mathbb{S}} = 0$, for an outer normal η of $\partial \mathbb{S}$, the forward Fokker–Planck–Kolmogorov operator $\mathcal{G}^*$ (the adjoint of (3)) under the stationary policy $\boldsymbol{\pi}^\star$ is exponentially ergodic with a positive spectral gap $\lambda_{\mathrm{gap}} > 0$; see Pinsky (2005) and (Houska, 2025, Lemma 3). Moreover, on unbounded domains $\mathbb{S}$, we can sometimes still establish the existence of a strictly positive spectral gap. For example, if a Bakry–Emery-type curvature condition hold and if a suitable Wonham–Hasminskii–Lyapunov function exists, we also have $\lambda_{\mathrm{gap}} > 0$ on potentially unbounded (but convex) domains, see (Wonham, 1966; Has'minskii, 1960; Houska, 2025).*

D OPERATOR MODELS

D.1 Operator Regression and Kernel-Based Learning

We now define the state measure $\mu \in M_+(\mathbb{S})$ and the joint state-action measure $\eta \in M_+(\mathbb{S} \times \mathbb{A})$, where $M_+(\mathbb{S})$ denotes the space of Borel probability measures on $\mathbb{S}$. We assume that both measures have full support on their respective domains, i.e., $\mathrm{supp}(\mu) = \mathbb{S}$ and $\mathrm{supp}(\eta) = \mathbb{S} \times \mathbb{A}$.

To simplify the notation, we also introduce $\boldsymbol{z} = [\boldsymbol{s}^\top, \boldsymbol{a}^\top]^\top$. To approximate the operator $\mathcal{G} : D(\mathcal{G}) \to L^2_\eta$, with $D(\mathcal{G})$ a space with sufficient regularity such as H^1_μ, we look for an RKHS approximation $G : \mathcal{H}_{\mathbb{S}} \to \mathcal{H}_{\mathbb{Z}}$ based on

its RKHS restriction $\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} \colon \mathcal{H}_{\mathbb{S}} \to L_\eta^2$. Even though $\mathcal{H}_{\mathbb{S}} \subset H_\mu^1, \mathcal{H}_{\mathbb{Z}} \subset L_\eta^2$, they have a different norm, which we account for via the *inclusions*

$$S_\eta : \mathcal{H}_{\mathbb{Z}} \hookrightarrow L_\eta^2 \qquad \text{and} \qquad S_\mu : \mathcal{H}_{\mathbb{S}} \hookrightarrow H_\mu^1.$$

Let us recall the definition of the state symmetric positive definite kernel function $k_{\mathsf{S}} \in C^{2,2}(\mathbb{S} \times \mathbb{S})$, satisfying $k_{\mathsf{S}}(s, s') = \langle \phi_{\mathsf{S}}(s), \phi_{\mathsf{S}}(s') \rangle_{\mathcal{H}_{\mathbb{S}}} = \langle k_{\mathsf{S}}(\cdot, s), k(\cdot, s') \rangle_{\mathcal{H}_{\mathbb{S}}}$ for all $s, s' \in \mathbb{S}$. Then we can define the action-affine kernel $k_{\mathsf{Z}} : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$.

Proposition 3. *[Action-affine kernel, (cf. Bevanda et al., 2025a)] Let $k_{\mathsf{Z}} : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ be a continuous, bounded kernel with RKHS $\mathcal{H}_{\mathbb{Z}}$. Then the tensor product space $\mathcal{H}_{\mathbb{Z}} := \mathbb{V} \otimes \mathcal{H}_{\mathbb{S}}$ with $\boldsymbol{v} = (1, \boldsymbol{a}) \in \mathbb{V}$ has reproducing kernel $k(\boldsymbol{z}, \boldsymbol{z}') = k(\boldsymbol{s}', \boldsymbol{s})(1 + \langle \boldsymbol{a}, \boldsymbol{a}' \rangle)$.*

We impose the following requirements on the previously defined RKHSs and kernels:

1. $\mathcal{H}_{\mathbb{Z}}, \mathcal{H}_{\mathbb{S}}$ are separable: this is satisfied if $\mathbb{S}$ and $\mathbb{Z}$ are Polish spaces and the kernels defining $\mathcal{H}_{\mathbb{Z}}, \mathcal{H}_{\mathbb{S}}$ are continuous;

2. ϕ_{S} and ϕ_{Z} are measurable for μ'-almost all $\boldsymbol{s} \in \mathbb{S}$ and η-almost all $\boldsymbol{z} \in \mathbb{Z}$;

3. $k_{\mathsf{S}}(\boldsymbol{s}, \boldsymbol{s})$ and $k_{\mathsf{Z}}(\boldsymbol{z}, \boldsymbol{z})$ are bounded for μ'-almost all $\boldsymbol{s} \in \mathbb{S}$ and η-almost all $\boldsymbol{z} \in \mathbb{Z}$, respectively.

D.2 Operator Model Risk Objective

We define the risk for an approximation $G \in HS(\mathcal{H}_{\mathbb{S}}, \mathcal{H}_{\mathbb{Z}})$ as:

$$\mathcal{R}(G) = \|\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} - S_\eta G\|_{\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, L_\eta^2)}^2, \tag{37}$$

where $\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} \colon \mathcal{H}_{\mathbb{S}} \to L_\eta^2$ is the target operator, i.e., the restriction of the infinitesimal generator to $\mathcal{H}_{\mathbb{S}}$. For $\mathcal{R}(G)$ to be well-defined, $\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}}$ must be a Hilbert-Schmidt operator in $\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, L_\eta^2)$. Due to the submultiplicativity of the operator norm, we have that $\|\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}}\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2} \leq \|\mathcal{G}\|_{H_\mu^1 \to L_\eta^2} \|S_\mu\|_{\mathcal{H}_{\mathbb{S}} \to H_\mu^1}$. Since $\mathcal{G}$ is bounded on (H_μ^1, L_η^2), we require both S_η and S_μ to be Hilbert–Schmidt. While S_η is known to be Hilbert–Schmidt (Steinwart and Christmann, 2008; Kostic et al., 2022), ensuring that S_μ is Hilbert–Schmidt motivates the following compatibility assumption. [4]

Assumption 4. *The RKHS $\mathcal{H}_{\mathbb{S}}$ is norm-equivalent to $H^y(\mathbb{S})$ with $y > \frac{n_s}{2} + 1$.*

Regularized Risk Minimization By the Pythagorean theorem, (37) decomposes into two parts

$$\underbrace{\|[I - P_{\mathcal{H}_{\mathbb{Z}}}]\mathcal{G}\|_{\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, L_\eta^2)}^2}_{\text{representation risk}} + \underbrace{\|P_{\mathcal{H}_{\mathbb{Z}}}\mathcal{G} - G\|_{\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, L_\eta^2)}^2}_{\text{projected risk}}, \tag{38}$$

where $P_{\mathcal{H}_{\mathbb{Z}}}$ is the orthogonal projector in L_η^2 onto $\mathcal{H}_{\mathbb{Z}}$. Using suitably defined infinite-dimensional RKHSs (Steinwart and Christmann, 2008), the *representation risk* can vanish, leaving the *projected risk* depending on the learned $G \in \mathrm{HS}(\mathcal{H}_{\mathbb{S}}, \mathcal{H}_{\mathbb{Z}})$. By identifying the canonical orthogonal projection with $S_\eta(S_\eta^* S_\eta)^\dagger S_\eta^*$, we see the projected risk is equivalent to

$$\|P_{\mathcal{H}_{\mathbb{Z}}}\mathcal{G} - G\|_{\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, L_\nu^2)}^2 = \|S_\eta(S_\eta^* S_\eta)^\dagger S_\eta^* \mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} - S_\eta G\|_{\mathrm{HS}}^2$$

whose unique minimizer $G^\dagger = (S_\eta^* S_\eta)^\dagger S_\eta^* \mathcal{G}|_{\mathcal{H}_{\mathbb{S}}}$ can be understood as a Galerkin projection onto $\mathcal{H}_{\mathbb{Z}}$. Yet, to ensure stability and prevent overfitting in this typically ill-posed problem, a natural approach is to add a Tikhonov regularization term to (37), so that

$$G_\gamma := \underset{G \in \mathrm{HS}(\mathcal{H}_{\mathbb{S}}, \mathcal{H}_{\mathbb{Z}})}{\arg\min} \mathcal{R}(G) + \gamma\|G\|_{\mathrm{HS}}^2 = (C_{\mathsf{ZZ}} + \gamma \operatorname{Id})^{-1} C_{\mathsf{Zd}}, \quad \gamma > 0 \tag{39}$$

---

[4] By Maurin's theorem (Adams and Fournier, 2003), $H^y \hookrightarrow H^1$ is Hilbert–Schmidt if $y > n_s/2 + 1$, a mild assumption met by common kernels (Gaussian, Matérn).

which corresponds to the *Kernel Ridge Regression* (KRR) approximation of $\mathcal{G}$ over $\mathcal{H}_\mathbb{S} \to \mathcal{H}_\mathbb{Z}$. The covariance and cross-covariance operators are defined as $C_{\mathsf{ZZ}} := S_\eta^* S_\eta : \mathcal{H}_\mathbb{Z} \to \mathcal{H}_\mathbb{Z}$ and $C_{\mathsf{Zd}} := S_\eta^* \mathcal{G}|_{\mathcal{H}_\mathbb{S}} : \mathcal{H}_\mathbb{S} \to \mathcal{H}_\mathbb{Z}$ respectively. Formally,

$$C_{\mathsf{ZZ}} := S_\eta^* S_\eta = \int_\mathbb{Z} \phi_\mathsf{z} \otimes \phi_\mathsf{z} \eta(d\boldsymbol{z}) : \mathcal{H}_\mathbb{Z} \to \mathcal{H}_\mathbb{Z}, \quad \text{and} \quad C_{\mathsf{Zd}} := S_\eta^* \mathcal{G}|_{\mathcal{H}_\mathbb{S}} = \int_\mathbb{Z} \phi_\mathsf{z} \otimes d\phi_\mathsf{s} \eta(d\boldsymbol{z}) : \mathcal{H}_\mathbb{S} \to \mathcal{H}_\mathbb{Z}.$$

where $d\phi_\mathsf{s} : \mathbb{S} \to \mathcal{H}_\mathbb{S}$ is the embedding of the generator in the RKHS $\mathcal{H}_\mathbb{S}$, satisfying the reproducing property $\langle d\phi_\mathsf{s}, h \rangle_{\mathcal{H}_\mathbb{S}} = [\mathcal{G}|_{\mathcal{H}_\mathbb{S}} h](\boldsymbol{s})$ for observables $h \in \mathcal{H}_\mathbb{S}$.

Empirical Risk Minimization   Population-level quantities in (39) are typically unavailable; thus, we approximate them using data samples $\mathbb{D}_N$ defined in (11). To construct the regularized empirical risk, we introduce the sampling operators

$$\widehat{S}_\mathsf{Z} : \mathcal{H}_\mathbb{Z} \to \mathbb{R}^N, \quad (\widehat{S}_\mathsf{Z}\phi_\mathsf{z})_i := \phi_\mathsf{z}((\boldsymbol{s}^{(i)}, \boldsymbol{a}^{(i)})), \quad \text{and} \quad \widehat{S}_\mathsf{S} : \mathcal{H}_\mathbb{S} \to \mathbb{R}^N, \quad (\widehat{S}_\mathsf{S}\phi_\mathsf{s})_i := \phi_\mathsf{s}(\boldsymbol{s}^{(i)}),$$
$$\widehat{S}_\mathrm{d} : \mathcal{H}_\mathbb{S} \to \mathbb{R}^N, \quad (\widehat{S}_\mathrm{d}\phi_\mathsf{s})_i := d\phi_\mathsf{s}(\boldsymbol{s}^{(i)}; \dot{\boldsymbol{s}}^{(i)}), \quad \text{and} \quad \widehat{U} : \mathbb{R}^{n_a} \to \mathbb{R}^N, \quad (\widehat{U}\boldsymbol{a})_i := \langle \boldsymbol{a}^{(i)}, \boldsymbol{a} \rangle.$$

with the adjoints $\widehat{S}_\mathsf{S}^* : \mathbb{R}^N \to \mathcal{H}_\mathbb{S}$, $\widehat{S}_\mathsf{Z}^* : \mathbb{R}^N \to \mathcal{H}_\mathbb{Z}$, $\widehat{S}_\mathrm{d} : \mathbb{R}^N \to \mathcal{H}_\mathbb{S}$ and $\widehat{S}_\mathrm{d} : \mathbb{R}^N \to \mathbb{R}^{n_a}$, called the sampled embedding operators (Smale and Zhou, 2007).

Thus, the empirical risk approximation of (39) reads

$$\widehat{G}_\gamma := \underset{G \in \mathrm{HS}(\mathcal{H}_\mathbb{S}, \mathcal{H}_\mathbb{Z})}{\arg\min} \widehat{\mathcal{R}}(G) + \gamma \|G\|_{\mathrm{HS}}^2 = \widehat{C}_{\mathsf{ZZ},\gamma}^{-1} \widehat{C}_{\mathsf{Zd}} \tag{41}$$

with $\widehat{C}_{\mathsf{ZZ},\gamma} = \widehat{C}_{\mathsf{ZZ}} + \gamma \,\mathrm{Id}$. Although infinite-dimensional RKHSs make direct computations infeasible, finite data enable finite-rank approximations of (19) and make the application of $\widehat{G}_\gamma$ to an observable computational. Hence, we introduce the kernel (Gram) matrices $\boldsymbol{K}_\mathsf{A} = \widehat{U}\widehat{U}^*$, $\boldsymbol{K}_\mathsf{S} := \widehat{S}_\mathsf{S}\widehat{S}_\mathsf{S}^* = [k_\mathsf{S}(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)})]_{i,j\in[N]}$, $\boldsymbol{K}_\mathsf{Z} := \widehat{S}_\mathsf{Z}\widehat{S}_\mathsf{Z}^* = \boldsymbol{K}_\mathsf{S} + \boldsymbol{K}_\mathsf{A}\odot\boldsymbol{K}_\mathsf{S}$ and $\boldsymbol{K}_\gamma = \boldsymbol{K}_\mathsf{Z} + N\gamma\boldsymbol{I}$. Moreover, we introduce the Khatri-Rao Product $\widehat{U}^*\odot\widehat{S}_\mathsf{S}^*$ defined by $(\widehat{U}^* \odot \widehat{S}_\mathsf{S}^*)e_i = (\widehat{U}^* e_i) \otimes (\widehat{S}_\mathsf{S}^* e_i)$ for an orthonormal basis $\{e_i\}$, where $\otimes$ denotes the elementary tensor product defined on the associated Hilbert spaces.

We now present the closed-form solution to the regularized empirical risk minimization problem.

Proposition 4 (Minimizer of (41)). *The minimizer of* (41), *denoted by* $\widehat{G}_\gamma \in \mathrm{HS}(\mathcal{H}_\mathbb{S}, \mathcal{H}_\mathbb{Z})$, *is given by*

$$\widehat{G}_\gamma = \widehat{S}_\mathsf{Z}^* \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d} = \begin{bmatrix} \widehat{S}_\mathsf{S}^* \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d} \\ (\widehat{U}^* \odot \widehat{S}_\mathsf{S}^*) \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d} \end{bmatrix} = \begin{bmatrix} \widehat{A}_\gamma \\ \widehat{B}_\gamma \end{bmatrix} \tag{42}$$

with $\widehat{A}_\gamma \in \mathrm{HS}(\mathcal{H}_\mathbb{S}, \mathcal{H}_\mathbb{S})$, $\widehat{B}_\gamma \in \mathrm{HS}(\mathcal{H}_\mathbb{S}, \mathbb{R}^{n_a} \otimes \mathcal{H}_\mathbb{S})$ *and* $\boldsymbol{K}_\gamma^{-1} := (\boldsymbol{K}_Z + N\gamma\boldsymbol{I})^{-1} \in \mathbb{R}^{N\times N}$.

Proof. Recall the expression in (39) and substitute the empirical covariances expressions for $\widehat{C}_{\mathsf{ZZ},\gamma}$ and $\widehat{C}_{\mathsf{Zd}}$, then it follows that

$$\widehat{G}_\gamma := \left(\tfrac{1}{N} \widehat{S}_\mathsf{Z}^* \widehat{S}_\mathsf{Z} + \gamma \,\mathrm{Id} \right)^{-1} \left(\tfrac{1}{N} \widehat{S}_\mathsf{Z}^* \widehat{S}_\mathrm{d} \right) = \widehat{S}_\mathsf{Z}^* \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d}$$

by using the push-through identity (derived via the Woodbury formula). Applying $\widehat{G}_\gamma$ to an observable $y \in \mathcal{H}_\mathbb{S}$ and evaluating it at $\boldsymbol{z}$ via RKHS inner product yields

$$\dot{\hat{y}}(\boldsymbol{z}) := [\widehat{G}_\gamma y](\boldsymbol{z}) = \langle \widehat{G}_\gamma y, \phi_\mathsf{z}(\boldsymbol{z}) \rangle_{\mathcal{H}_\mathbb{Z}} = \langle \widehat{S}_\mathsf{Z}^* \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d} y, \phi_\mathsf{z}(\boldsymbol{z}) \rangle_{\mathcal{H}_\mathbb{Z}}$$
$$= \langle \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d} y, \widehat{S}_\mathsf{Z}([\begin{smallmatrix}1\\\boldsymbol{a}\end{smallmatrix}] \otimes \phi_\mathsf{s}(\boldsymbol{s})) \rangle_{\mathcal{H}_\mathbb{Z}} = \langle \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d} y, \widehat{S}_\mathsf{S}\phi_\mathsf{s}(\boldsymbol{s}) + \widehat{U}\boldsymbol{a} \odot \widehat{S}_\mathsf{S}\phi_\mathsf{s}(\boldsymbol{s}) \rangle_{\mathcal{H}_\mathbb{S} \oplus (\mathbb{R}^{n_a} \otimes \mathcal{H}_\mathbb{S})} \tag{43}$$
$$= \langle \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d} y, \widehat{S}_\mathsf{S}\phi_\mathsf{s}(\boldsymbol{s}) \rangle_{\mathcal{H}_\mathbb{S}} + \langle \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d} y, (\widehat{U}^* \odot \widehat{S}_\mathsf{S}^*)^* (\boldsymbol{a} \otimes \phi_\mathsf{s}(\boldsymbol{s})) \rangle_{\mathbb{R}^{n_a} \otimes \mathcal{H}_\mathbb{S}} \tag{44}$$
$$= \langle \underbrace{\widehat{S}_\mathsf{S}^* \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d}}_{\widehat{A}_\gamma} y, \phi_\mathsf{s}(\boldsymbol{s}) \rangle_{\mathcal{H}_\mathbb{S}} + \langle \underbrace{(\widehat{U}^* \odot \widehat{S}_\mathsf{S}^*) \boldsymbol{K}_\gamma^{-1} \widehat{S}_\mathrm{d}}_{\widehat{B}_\gamma} y, (\boldsymbol{a} \otimes \phi_\mathsf{s}(\boldsymbol{s})) \rangle_{\mathbb{R}^{n_a} \otimes \mathcal{H}_\mathbb{S}}$$

which follows from the structure of $\mathcal{H}_{\mathbb{Z}}$ and recent results in Bevanda et al. (2025a). In particular, from (43) to (44), we used the property of the Khatri-Rao product, namely that $(\widehat{U}^* \odot \widehat{S}_{\mathsf{S}}^*)^*(v \otimes w) = (\widehat{U}v) \odot (\widehat{S}_{\mathsf{S}}w)$. This yields the infinite-dimensional estimators

$$\widehat{A}_\gamma = \widehat{S}_{\mathsf{S}}^* \boldsymbol{K}_\gamma^{-1} \widehat{S}_{\mathrm{d}} \in \mathrm{HS}\left(\mathcal{H}_{\mathbb{S}}, \mathcal{H}_{\mathbb{S}}\right),$$

$$\widehat{B}_\gamma = (\widehat{U}^* \odot \widehat{S}_{\mathsf{S}}^*) \boldsymbol{K}_\gamma^{-1} \widehat{S}_{\mathrm{d}} \in \mathrm{HS}\left(\mathcal{H}_{\mathbb{S}}, \mathbb{R}^{n_a} \otimes \mathcal{H}_{\mathbb{S}}\right),$$

$\square$

Remark 6. *We can also obtain the Principal Component Regression (PCR) estimator by projecting the input data onto the r-dimensional principal subspace of the covariance matrix $\widehat{C}_{\mathsf{ZZ}}$ (Kostic et al., 2022, 2023). This yields the estimator $\widehat{G}_\gamma^r = [\![\widehat{C}_{\mathsf{ZZ},\gamma}]\!]_r^{-1} \widehat{T} = \widehat{S}_{\mathsf{Z}}^* U_r V_r^{\mathsf{T}} \widehat{S}_{\mathrm{d}}$ where $[\![\boldsymbol{K}_{\mathsf{Z}}]\!]_r = V_r \boldsymbol{\Sigma}_r V_r^*$ is the r-truncated SVD of $\boldsymbol{K}_{\mathsf{Z}}$, and $U_r = V_r \boldsymbol{\Sigma}_r^\dagger$ (Bevanda et al., 2025a).*

D.3 Operator Regression - Learning Error Bounds

We are motivated by using sharper error bounds under the operator norm (Talwai et al., 2022; Kostic et al., 2023, 2024c), which not only provide stronger theoretical guarantees but are also essential in practice, such as applications in safety-critical systems or for robust RL, as the operator norm characterizes worst-case performance.

Remark 7 (Well-posedness of the risk and arbitrary accuracy)**.** *This remark follows the line of (Mollenhauer and Koltai, 2020; Mollenhauer et al., 2022; Kostic et al., 2022, 2023; Bevanda et al., 2025a). Recall the bias-variance decomposition of the risk in (38) with $P_{\mathcal{H}_{\mathbb{Z}}}$ being the orthogonal projector onto the closure of $\mathrm{Im}(S_\eta) \subseteq L_\eta^2(\mathbb{Z})$.*

1. *Representation bias: The representation bias $\|[I - P_{\mathcal{H}_{\mathbb{Z}}}]\mathcal{G}\|_{\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, L_\eta^2)}^2$ quantifies how well the target operator $\mathcal{G}\mid_{\mathcal{H}_{\mathbb{S}}}$ can be approximated by any operator within the model class structure and vanishes to zero when choosing a C_0-universal RKHS $\mathcal{H}_{\mathbb{S}}$ inducing $\mathcal{H}_{\mathbb{Z}}$, i.e. $\mathrm{Im}(\mathcal{G}\mid_{\mathcal{H}_{\mathbb{S}}}) \subset cl(\mathrm{Im}(S_\eta))$ (Steinwart and Christmann, 2008, Chapter 4).*

2. *Arbitrary Accuracy: The estimation error satisfies $\|P_{\mathcal{H}_{\mathbb{Z}}}\mathcal{G} - G\|_{\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, L_\eta^2)}^2 < \epsilon$. This follows because $\mathcal{G}\mid_{\mathcal{H}_{\mathbb{S}}}$ can be approximated arbitrarily well in the HS norm by elements of the form $S_\eta G$, and such elements can, in turn, be approximated arbitrarily well, since finite-rank operators from $\mathcal{H}_{\mathbb{S}} \to \mathcal{H}_{\mathbb{Z}}$ are dense in $\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, \mathcal{H}_{\mathbb{Z}})$.*

3. *Risk well-posedness: The squared operator norm $\|\mathcal{G} - G\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2}^2$, is bounded by the Hilbert-Schmidt risk i.e. $\|\mathcal{G} - G\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2}^2 \leq \mathcal{R}(G) = \|\mathcal{G} - G\|_{\mathrm{HS}(\mathcal{H}_{\mathbb{S}}, L_\eta^2)}^2$. Thus, minimizing the HS risk $\mathcal{R}(G)$ also drives down an upper bound on the operator norm error.*

Bounding the operator norm error for the full estimator The operator norm error can be written as $\mathcal{E}(\widehat{G}_\gamma) := \|\mathcal{G}\mid_{\mathcal{H}_{\mathbb{S}}} - S_\eta \widehat{G}_\gamma\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2}$ and decomposed into

$$\mathcal{E}(\widehat{G}_\gamma) \leq \underbrace{\|\mathcal{G}\mid_{\mathcal{H}_{\mathbb{S}}} - S_\eta G_\gamma\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2}}_{\text{bias due to regularization}} + \underbrace{\|S_\eta(G_\gamma - G_\gamma^r)\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2}}_{\text{rank reduction bias}} + \underbrace{\|S_\eta(G_\gamma - \widehat{G}_\gamma)\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2}}_{\text{variance of the estimator}} \tag{45}$$

where G_γ is the minimizer of the Tikhonov regularized risk and the population version of the empirical estimator $\widehat{G}_\gamma$ and G_γ^r the reduced rank population version of the empirical estimator $\widehat{G}_\gamma^r$ obtained via *Reduced Rank Regression* (RRR) or *Principal Components Regression* (PCR).

We first recall key results from Kostic et al. (2023) on the operator norm error of $\widehat{G}_\gamma$, adopting the same assumptions and notation for consistency.

(RC) *Regularity of $\mathcal{G}$*: for some $\alpha \in (1, 2]$ there exists $a > 0$ such that $C_{\mathsf{Zd}} C_{\mathsf{Zd}}^* \preceq a^2 C_{\mathsf{ZZ}}^{1+\alpha}$;

(BK) *Boundedness.* There exists $c_{\mathcal{H}_{\mathbb{S}}} > 0$ such that $\underset{\boldsymbol{s} \sim \mu}{\mathrm{ess\,sup}} \|\phi(\boldsymbol{s})\|^2 \leq c_{\mathcal{H}_{\mathbb{S}}}$, i.e. $\phi \in L_\mu^\infty(\mathbb{S}, \mathcal{H}_{\mathbb{S}})$

and $c_{\mathcal{H}_{\mathbb{Z}}} > 0$ such that $\underset{\boldsymbol{z} \sim \eta}{\mathrm{ess\,sup}} \|\psi(\boldsymbol{z})\|^2 \leq c_{\mathcal{H}_{\mathbb{Z}}}$, i.e. $\psi \in L_\eta^\infty(\mathbb{Z}, \mathcal{H}_{\mathbb{Z}})$

(SD) *Spectral Decay.* There exists $\beta \in (0, 1]$ and $b > 0$ such that $\lambda_j(C_{\mathsf{ZZ}}) \leq b\, j^{-1/\beta}$, for all $j \in J$.

We define $J := 1, 2, \ldots \subseteq \mathbb{N}$. Informally, **(RC)** on $\mathcal{G}$, adapted from Kostic et al. (2023), ensures that $\mathcal{G}$ is well-aligned with the RKHS structure and quantifies the relationship between bounded operators in our RKHS hypothesis class and the target operator. Assumption **(BK)** ensures that functions under the kernel embedding have bounded norms, controlling the complexity and stability of the estimator. Assumption **(SD)** controls the eigenvalue decay of C_{ZZ}, where faster decay (smaller β) favors better estimation rates.

Theorem 2. *[Kostic et al. (2023)] Let ((SD)) and ((RC)) hold for some $\beta \in (0, 1]$ and $\alpha \in (1, 2]$, respectively, and let $\mathrm{cl}(\mathrm{Im}(S_\eta)) = L_\eta^2(\mathbb{Z})$ and (BK) be satisfied. Let the regularization parameter be chosen as $\gamma \asymp N^{-\frac{1}{\alpha+\beta}}$. Then, for any $\xi \in (0, 1)$, there exists a constant $c > 0$, depending only on $\mathcal{H}_{\mathbb{Z}}$, such that with probability at least $1 - \xi$ over an i.i.d. sample $\mathbb{D}_N$ the following holds for the KRR estimator $\widehat{G}_\gamma$:*

$$\mathcal{E}(\widehat{G}_\gamma) \le c \, N^{-\frac{\alpha}{2(\alpha+\beta)}} \, \ln \xi^{-1}. \tag{46}$$

Proof Sketch. This result follows directly from Kostic et al. (2023), by bounding each term of the error decomposition in 45. The regularization bias consists of a term depending on the choice of γ and a term reflecting the alignment between $\mathcal{H}_{\mathbb{Z}}$ and $\mathrm{Im}(\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}})$. The latter can be set to zero by choosing a universal kernel for which $\mathrm{Im}(\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}}) \subseteq \mathrm{cl}(\mathrm{Im}(S_\eta))$ (Kostic et al., 2022; Li et al., 2022). The bias due to rank reduction is 0 for the KRR estimator, while for the PCR estimator, it can be derived by using an orthogonal projector onto the subspace of the leading r eigenfunctions of C_{ZZ}, which yields the upper bound $\sigma_{r+1}(S_\eta)$. Finally, bounding the variance of the estimator follows (Kostic et al., 2023, Appendix D.3). Combining the bias due to regularization and variance terms, for both estimators, we obtain the optimal regularization parameter γ and the rates. $\qquad\square$

Remark 8. *When clear from the context, for conciseness, we drop the explicit inclusions, e.g. for $\mathcal{E}(\widehat{G}_\gamma) = \|\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} - S_\eta \widehat{G}_\gamma\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2}$, we write $\mathcal{E}(\widehat{G}_\gamma) = \|\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} - \widehat{G}_\gamma\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2}$ as the mapping is implied from the norm definition.*

We now turn to bounding $\mathcal{E}(\widehat{A}_\gamma, \widehat{B}_\gamma)$. Recall that $\widehat{A}_\gamma \in \mathrm{HS}(\mathcal{H}_{\mathbb{S}}, \mathcal{H}_{\mathbb{S}})$, $\widehat{B}_\gamma \in \mathrm{HS}(\mathcal{H}_{\mathbb{S}}, \mathbb{R}^{n_a} \otimes \mathcal{H}_{\mathbb{S}})$ and that $\widehat{G}_\gamma \in \mathrm{HS}(\mathcal{H}_{\mathbb{S}}, \mathcal{H}_{\mathbb{Z}})$. To compare errors in a common output space, we introduce the bounded inclusions $S_A : \mathcal{H}_{\mathbb{S}} \to L_\mu^2$ and $S_B : \mathbb{R}^{n_a} \otimes \mathcal{H}_{\mathbb{S}} \to \mathbb{R}^{n_a} \otimes L_\mu^2$. We begin from

$$\mathcal{E}(\widehat{G}_\gamma) := \|\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} - S_\eta \widehat{G}_\gamma\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2} = \|\begin{bmatrix} \mathcal{A}|_{\mathcal{H}_{\mathbb{S}}} \\ \mathcal{B}|_{\mathcal{H}_{\mathbb{S}}} \end{bmatrix} - S_\eta \begin{bmatrix} \widehat{A}_\gamma \\ \widehat{B}_\gamma \end{bmatrix}\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2} = \|\begin{bmatrix} \mathcal{A}|_{\mathcal{H}_{\mathbb{S}}} - S_A \widehat{A}_\gamma \\ \mathcal{B}|_{\mathcal{H}_{\mathbb{S}}} - S_B \widehat{B}_\gamma \end{bmatrix}\|_{\mathcal{H}_{\mathbb{S}} \to L_\eta^2} \tag{47}$$

Let $\phi_{\mathsf{S}} \in \mathcal{H}_{\mathbb{S}}$, $\phi_{\mathsf{Z}} \in \mathcal{H}_{\mathbb{Z}}$ and $\boldsymbol{z} = [\boldsymbol{s}, \boldsymbol{a}]^{\mathsf{T}}$, then

$$\|(\mathcal{A}|_{\mathcal{H}_{\mathbb{S}}} - S_A \widehat{A}_\gamma)\phi_{\mathsf{S}}(\boldsymbol{s})\|_{L_\mu^2}^2 + \|(\mathcal{B}|_{\mathcal{H}_{\mathbb{S}}} - S_B \widehat{B}_\gamma)(\boldsymbol{a} \otimes \phi_{\mathsf{S}}(\boldsymbol{s}))\|_{\mathbb{R}^{n_a} \otimes L_\mu^2}^2 = \|(\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} - S_\eta \widehat{G}_\gamma)\phi_{\mathsf{Z}}(\boldsymbol{z})\|_{L_\eta^2}^2 \tag{48}$$

$$\implies \|(\mathcal{A}|_{\mathcal{H}_{\mathbb{S}}} - S_A \widehat{A}_\gamma)\phi_{\mathsf{S}}(\boldsymbol{s})\|_{L_\mu^2}^2 \le \|(\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} - S_\eta \widehat{G}_\gamma)\phi_{\mathsf{Z}}(\boldsymbol{z})\|_{L_\eta^2}^2 \tag{49}$$

$$\text{and} \quad \|(\mathcal{B}|_{\mathcal{H}_{\mathbb{S}}} - S_B \widehat{B}_\gamma)\phi_{\mathsf{S}}(\boldsymbol{s})\|_{\mathbb{R}^{n_a} \otimes L_\mu^2}^2 \le \|(\mathcal{G}|_{\mathcal{H}_{\mathbb{S}}} - S_\eta \widehat{G}_\gamma)\phi_{\mathsf{Z}}(\boldsymbol{z})\|_{L_\eta^2}^2 \tag{50}$$

From the operator-norm definition, it follows that,

$$\mathcal{E}(\widehat{A}_\gamma) := \|\mathcal{A}|_{\mathcal{H}_{\mathbb{S}}} - \widehat{A}_\gamma\|_{\mathcal{H}_{\mathbb{S}} \to L_\mu^2}^2 \le \mathcal{E}(\widehat{G}_\gamma)$$

$$\mathcal{E}(\widehat{B}_\gamma) := \|\mathcal{B}|_{\mathcal{H}_{\mathbb{S}}} - S_B \widehat{B}_\gamma\|_{\mathcal{H}_{\mathbb{S}} \to \mathbb{R}^{n_a} \otimes L_\mu^2}^2 \le \mathcal{E}(\widehat{G}_\gamma) \tag{51}$$

This leads to the result $\max\{\mathcal{E}(\widehat{A}_\gamma), \mathcal{E}(\widehat{B}_\gamma)\} \le \mathcal{E}(\widehat{G}_\gamma)$.

E   PROOFS

In the first part of this section, we present the derivation and construction of O-CTRL (Algorithm 1), detailing how to obtain a finite-dimensional representation of the value function (Proposition 1) and a tractable dynamic programming recursion (Corollary 1) that solves the HJB for the estimated operator world model. In the second part, we provide the proofs of Theorem 1 and Corollary 2, highlighting how smoothness, the spectral gap, and time discretization shape the difficulty of offline RL in continuous time.

E.1 Derivations of O-CTRL Algorithm

Proposition 1. *Let $\widehat{r}_{\boldsymbol{s}} = \widehat{S}_{\mathsf{S}}^{*}\,\boldsymbol{r}, \widehat{\mathcal{D}}_{\boldsymbol{a}} = \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{D}_{\boldsymbol{a}}(\cdot)$, where $\widehat{S}_{\mathsf{S}}^{*} : \mathbb{R}^{N} \to \mathcal{H}_{\mathsf{S}}$ is the adjoint of the sampling operator $\widehat{S}_{\mathsf{S}}$. Let the transition dynamics be described by* (19). *Then the flow of* (20) *resides in a finite-dimensional subspace*

$$\langle \dot{\boldsymbol{v}}(t,\cdot),\, \boldsymbol{k}_{\mathsf{S}}(\boldsymbol{s}) \rangle = \langle \boldsymbol{T}\big(\boldsymbol{v}(t,\cdot)\big),\, \boldsymbol{k}_{\mathsf{S}}(\boldsymbol{s}) \rangle, \tag{21}$$

with

$$\boldsymbol{T}(\boldsymbol{v}) := -(\rho \boldsymbol{I} - \boldsymbol{A})\,\boldsymbol{v} + \big(\boldsymbol{r} + \boldsymbol{D}_{\boldsymbol{a}}(\boldsymbol{B}\,\boldsymbol{v})\big), \tag{22}$$

where $\widehat{S}_{\mathsf{S}}\phi_{\mathsf{S}}(\boldsymbol{s}) = \boldsymbol{k}_{\mathsf{S}}(\boldsymbol{s})$ is the sampled canonical map, $\boldsymbol{A} := \boldsymbol{K}_{\gamma}^{-1} \widehat{S}_{\mathsf{d}} \widehat{S}_{\mathsf{S}}^{*}$, $\boldsymbol{B} := \big[\operatorname{diag}(\widehat{U}\,\boldsymbol{e}_{k})\,\boldsymbol{A}\big]_{k \in [n_{a}]}$. The reward/Fenchel-dual terms in the RKHS are $\boldsymbol{r} = \boldsymbol{K}_{\gamma}^{-1}\big[r_{\boldsymbol{s}}(\boldsymbol{s}^{(1)}), \ldots, r_{\boldsymbol{s}}(\boldsymbol{s}^{(N)})\big]^{\top}$ and $\boldsymbol{D}_{\boldsymbol{a}}(\boldsymbol{\lambda}) = \boldsymbol{K}_{\gamma}^{-1}\big[\mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda}(\boldsymbol{s}^{(i)}))\big]_{i \in [N]}$, respectively.

Proof. We start by defining $\boldsymbol{\lambda}(\boldsymbol{s}) := \langle \boldsymbol{\lambda}, \boldsymbol{I}_{n_{a}} \otimes \boldsymbol{k}_{\mathsf{S}}(\boldsymbol{s}) \rangle$. Then we use the definitions for $\widehat{r}_{\boldsymbol{s}} = \widehat{S}_{\mathsf{S}}^{*}\,\boldsymbol{r}, \widehat{\mathcal{D}}_{\boldsymbol{a}}(\boldsymbol{\lambda}) = \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{K}_{\gamma}^{-1}\big[\mathcal{D}_{\boldsymbol{a}}(\boldsymbol{\lambda}(\boldsymbol{s}^{(i)}))\big]_{i \in [N]}$ and the approximated observable $\widehat{V} = \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{v}$ and substitute them into (20), leading to

$$
\begin{aligned}
\widehat{\mathcal{T}}(\widehat{V}) &= -(\rho I - \widehat{A}_{\gamma})\widehat{S}_{\mathsf{S}}^{*}\boldsymbol{v} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{r} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{K}_{\gamma}^{-1}\Big[\mathcal{D}_{\boldsymbol{a}}\Big([\widehat{B}_{\gamma}\widehat{S}_{\mathsf{S}}^{*}\boldsymbol{v}](\boldsymbol{s}^{(i)})\Big)\Big]_{i \in [N]} \\
&= -(\rho I - \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{K}_{\gamma}^{-1}\widehat{S}_{\mathsf{d}})\widehat{S}_{\mathsf{S}}^{*}\boldsymbol{v} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{r} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{K}_{\gamma}^{-1}\Big[\mathcal{D}_{\boldsymbol{a}}\Big([(\widehat{U}^{*}\!\circledcirc\widehat{S}_{\mathsf{S}}^{*})\boldsymbol{K}_{\gamma}^{-1}\widehat{S}_{\mathsf{d}}\widehat{S}_{\mathsf{S}}^{*}\boldsymbol{v}](\boldsymbol{s}^{(i)})\Big)\Big]_{i \in [N]} \\
&= -\widehat{S}_{\mathsf{S}}^{*}(\rho\boldsymbol{I} - \boldsymbol{K}_{\gamma}^{-1}\widehat{S}_{\mathsf{d}}\widehat{S}_{\mathsf{S}}^{*})\boldsymbol{v} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{r} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{K}_{\gamma}^{-1}\Big[\mathcal{D}_{\boldsymbol{a}}\Big([(\widehat{U}^{*}\!\circledcirc\widehat{S}_{\mathsf{S}}^{*})\boldsymbol{K}_{\gamma}^{-1}\widehat{S}_{\mathsf{d}}\widehat{S}_{\mathsf{S}}^{*}\boldsymbol{v}](\boldsymbol{s}^{(i)})\Big)\Big]_{i \in [N]} \\
&= -\widehat{S}_{\mathsf{S}}^{*}(\rho\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{r} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{K}_{\gamma}^{-1}\Big[\mathcal{D}_{\boldsymbol{a}}\Big(\langle[\operatorname{diag}(\widehat{U}\boldsymbol{e}_{k})\boldsymbol{A}]_{k \in n_{a}}\boldsymbol{v},\, \boldsymbol{I}_{n_{a}} \otimes \boldsymbol{k}_{\mathsf{S}}(\boldsymbol{s}^{(i)})\rangle\Big)\Big]_{i \in [N]} \\
&= -\widehat{S}_{\mathsf{S}}^{*}(\rho\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{r} + \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{D}_{\boldsymbol{a}}(\boldsymbol{B}\boldsymbol{v})
\end{aligned}
$$

where we substituted the expression for the KRR estimators $(\widehat{A}_{\gamma}, \widehat{B}_{\gamma})$ of (42), and used the Khatri-Rao product, as well as the structure of $\mathcal{H}_{\mathbb{Z}}$ (cf. Proposition 4). Finally,

$$
\begin{aligned}
\langle \widehat{\mathcal{T}}(\widehat{V}), \phi_{\mathsf{S}}(\boldsymbol{s}) \rangle &= \langle \widehat{S}_{\mathsf{S}}^{*}\left(-(\rho\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v} + \boldsymbol{r} + \boldsymbol{D}_{\boldsymbol{a}}(\boldsymbol{B}\boldsymbol{v})\right), \phi_{\mathsf{S}}(\boldsymbol{s}) \rangle \\
&= \langle \underbrace{-(\rho\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v} + \boldsymbol{r} + \boldsymbol{D}_{\boldsymbol{a}}(\boldsymbol{B}\boldsymbol{v})}_{\boldsymbol{T}(\boldsymbol{v})}, \boldsymbol{k}_{\mathsf{S}}(\boldsymbol{s}) \rangle
\end{aligned}
$$

Note that, while the approximate observable $\widehat{V} = \widehat{S}_{\mathsf{S}}^{*}\boldsymbol{v} \in \mathcal{H}_{\mathsf{S}}$ is in the full RKHS and $\widehat{G}_{\gamma}$ only a finite rank operator in $\operatorname{HS}(\mathcal{H}_{\mathsf{S}}, \mathcal{H}_{\mathbb{Z}})$, the following results show that we only require finite-dimensional computations through the use of the reproducing property ("kernel-trick").

The expression for $\boldsymbol{A}$ and $\boldsymbol{B}$ read

$$\boldsymbol{A} = \boldsymbol{K}_{\gamma}^{-1}\widehat{S}_{\mathsf{d}}\widehat{S}_{\mathsf{S}}^{*} = \boldsymbol{K}_{\gamma}^{-1}\boldsymbol{K}_{\mathsf{d}}, \quad \text{and} \quad \boldsymbol{B} = [\operatorname{diag}(\boldsymbol{U}\boldsymbol{e}_{i})\boldsymbol{A}]_{i \in [n_{a}]}.$$

The target kernel matrix $\boldsymbol{K}_{\mathsf{d}} = \widehat{S}_{\mathsf{d}}\widehat{S}_{\mathsf{S}}^{*}$ is computed via the Itō formula and the derivative reproducing property (Arnold, 1974; Kostic et al., 2024a; Zhang et al., 2023)

$$(\boldsymbol{K}_{\mathsf{d}})_{ij} := \langle \dot{\boldsymbol{s}}^{(i)}, \nabla_{\boldsymbol{s}^{(i)}} k(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)}) \rangle + \epsilon \operatorname{Tr} \nabla_{\boldsymbol{s}^{(i)}}^{2}\Big(k(\boldsymbol{s}^{(i)}, \boldsymbol{s}^{(j)})\Big).$$

$\square$

Corollary 1. *Let the step-size $\Delta t > 0$. Then the implicit-explicit IMEX flow (He, 2013; Koto, 2008; Sebastiano, 2023) update reads*

$$\boldsymbol{v}^{(k+1)} = \boldsymbol{M}\left[\boldsymbol{v}^{(k)} + \Delta t\big(\boldsymbol{r} + \boldsymbol{D}_{\boldsymbol{a}}(\boldsymbol{B}\boldsymbol{v}^{(k)})\big)\right], \tag{23}$$

with $\boldsymbol{M} = (\boldsymbol{I} + \Delta t(\rho\boldsymbol{I} - \boldsymbol{A}))^{-1}$ naturally following from the implicit discretization.

Proof. The proof starts with the direcretization of the descent flow $\dot{\widehat{V}} = \widehat{\mathcal{T}}(\widehat{V})$, whose linearization has stable eigenvalues (c.f. Section C.2), namely

$$\dot{\boldsymbol{v}}(t) = \underbrace{-(\rho\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v}(t)}_{\text{implicit}} + \underbrace{\boldsymbol{r} + \boldsymbol{D_a}(\boldsymbol{B}\boldsymbol{v}(t))}_{\text{explicit}}.$$

Then,

$$\frac{\boldsymbol{v}^{(k+1)} - \boldsymbol{v}^{(k)}}{\Delta t} = -(\rho\boldsymbol{I} - \boldsymbol{A})\boldsymbol{v}^{(k+1)} + \boldsymbol{r} + \boldsymbol{D_a}(\boldsymbol{B}\boldsymbol{v}^{(k)}),$$

and it follows that,

$$(\boldsymbol{I} + \Delta t(\rho\boldsymbol{I} - \boldsymbol{A}))\boldsymbol{v}^{(k+1)} = \Delta t \left(\boldsymbol{r} + \boldsymbol{D_a}(\boldsymbol{B}\boldsymbol{v}^{(k)}) \right). \tag{52}$$

□

E.2   DERIVATIONS OF THE END-TO-END LEARNING RATES

In this section, we prove Theorem 1 and Corollary 2.

Theorem 1. *Suppose that Assumptions 1 to 3 and the conditions of Proposition 1 and Corollary 1 hold. Then under the zero-initial condition $\widehat{V}_0 = 0$, and provided that $\delta = \mathcal{E}(\widehat{G}_\gamma)$ is sufficiently small,*

$$\|V^\star - \widehat{V}_k\|_{L_\mu^2} \leq \underbrace{(\widehat{\lambda}_{\text{gap}} + \rho)^{-1}\delta}_{\text{learning}} + \underbrace{\mathcal{O}((\Delta t)^p)}_{\text{discretization}} \tag{24a}$$

$$+ \underbrace{\kappa\, e^{-(\widehat{\lambda}_{\text{gap}} + \rho)k\Delta t}\|\widehat{V}^\star\|_{L_\mu^2}}_{\text{convergence}}, \tag{24b}$$

where p is the discretization order, and $\kappa > 0$ a constant. Here, $\widehat{\lambda}_{\text{gap}}$ denotes the spectral gap of the estimated closed-loop generator $P^{\widehat{\boldsymbol{\pi}}^\star}\widehat{G}_\gamma$ in (3) under the stationary policy $\widehat{\boldsymbol{\pi}}^\star$.

Remark 9. *Regarding Assumption 3, note that if $r_{\boldsymbol{s}} \notin \mathcal{H}_{\mathbb{S}}$, we can augment the RKHS with the kernel $k^r(\boldsymbol{s}, \boldsymbol{s}') \coloneqq k(\boldsymbol{s}, \boldsymbol{s}') + \langle r_{\boldsymbol{s}}, r_{\boldsymbol{s}'} \rangle$, ensuring $r_{\boldsymbol{s}}$ lies in the resulting space (still denoted $\mathcal{H}_{\mathbb{S}}$).*

Proof. To arrive at the final result, we will apply the tools from Section C.2. Recall that the total error can be bounded by using the triangle inequality, which leads to three terms

$$\|V^\star - \widehat{V}_k\|_{L_\mu^2} \leq \underbrace{\|V^\star - \widehat{V}^\star\|_{L_\mu^2}}_{\text{learning}} + \underbrace{\|\widehat{V}^\star - \widehat{V}_T\|_{L_\mu^2}}_{\text{convergence}} + \underbrace{\|\widehat{V}_T - \widehat{V}_k\|_{L_\mu^2}}_{\text{discretization}}. \tag{53}$$

This notation is introduced in Section 4 and summarized again in Table 2. To help visualize the different components of our end-to-end error analysis, Figure 3 organizes the results and relates them to our upper bound on the value function error, as discussed in Theorem 1.
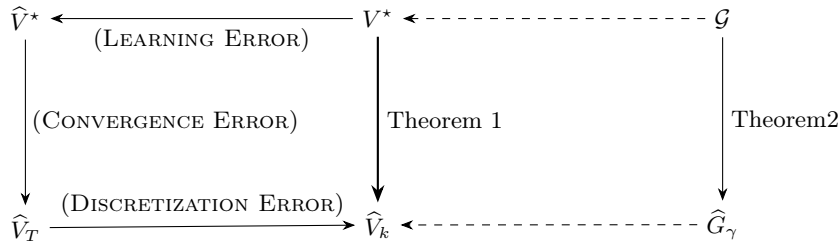

Figure 3: Error decomposition diagram for bounding the term $\|V^\star - \widehat{V}_k\|_{L_\mu^2}$

We first start to analyze the learning error, $\|V^\star - \widehat{V}^\star\|_{L_\mu^2}$.

Infinite Horizon Learning Error   In addition to the conditions stated in Theorem 1, suppose that Lemma 1 also holds. Let $L_{\mathcal{D}}$ denote the Lipschitz constant of $\mathcal{D}_a$ (see Proposition 2 and Assumption 3).

We begin our analysis with an essential property of the exact and approximate HJBs. The exact HJB satisfies $\|\mathcal{T}(V_1) - \mathcal{T}(V_2)\| \le \|\mathcal{A} - \rho I\|_{H^1_\mu \to L^2_\mu} \|V_1 - V_2\|_{L^2_\mu} + \|\mathcal{D}_a(\mathcal{B}V_1) - \mathcal{D}_a(\mathcal{B}V_2)\|_{L^2_\mu}$. Thus, using the Lipschitz continuity of $\mathcal{D}_a(\cdot)$, the Lipschitz constant $L_{\mathcal{T}}$ of the exact HJB and the Lipschitz constant $L_{\widehat{\mathcal{T}}}$ of the approximated HJB read

$$L_{\mathcal{T}} = \|\mathcal{A} - \rho I\|_{H^1_\mu \to L^2_\mu} + L_{\mathcal{D}} \|\mathcal{B}\|_{H^1_\mu \to \mathbb{R}^{n_a} \otimes L^2_\mu}$$
$$L_{\widehat{\mathcal{T}}} = \|\widehat{A}_\gamma - \rho I\|_{\mathcal{H}_{\mathbb{S}} \to \mathcal{H}_{\mathbb{S}}} + L_{\mathcal{D}} \|\widehat{B}_\gamma\|_{\mathcal{H}_{\mathbb{S}} \to \mathbb{R}^{n_a} \otimes \mathcal{H}_{\mathbb{S}}}.$$

But then, since the exact HJB operator $\mathcal{T}$ and its approximation $\widehat{\mathcal{T}}$ are Lipschitz continuous in their respective norms, L^2_μ and $\mathcal{H}_{\mathbb{S}}$, their Fréchet derivatives, denoted as

$$D_F^\star(V^\star) := \frac{\partial}{\partial V} \mathcal{T}(V^\star) \qquad \text{and} \qquad \widehat{D}_F(\widehat{V}) := \frac{\partial}{\partial \widehat{V}} \widehat{\mathcal{T}}(\widehat{V})$$

are bounded linear operators. Moreover, since we assume that the Fenchel conjugates of the action penalties are twice differentiable, the above Frechet derivatives are themselves Lipschitz continuous, satisfying the bound:

$$\|D_F^\star(V^\star) - \widehat{D}_F(\widehat{V})\|_{L^2_\mu} \le c_1(\mathcal{E}(\widehat{A}_\gamma) + L_{\mathcal{D}}\mathcal{E}(\widehat{B}_\gamma)) + c_2\|\widehat{V} - V^\star\|_{L^2_\mu}.$$

This can be seen using the triangle inequality, which yields

$$\begin{aligned}
\|D_F^\star(V^\star) - \widehat{D}_F(\widehat{V})\|_{L^2_\mu} &\le \|D_F^\star(V^\star) - \widehat{D}^F(V^\star)\|_{L^2_\mu} + \|\widehat{D}^F(V^\star) - \widehat{D}_F(\widehat{V})\|_{L^2_\mu} \\
&\le c_1(\mathcal{E}(\widehat{A}_\gamma) + L_{\mathcal{D}}\mathcal{E}(\widehat{B}_\gamma)) + c_2\|\widehat{V} - V^\star\|_{L^2_\mu}.
\end{aligned}$$

where for the first term of the second inequality, we have used that

$$\begin{aligned}
\|\mathcal{T}(V^\star) - \widehat{\mathcal{T}}(V^\star)\|_{L^2_\mu} &\le \|\mathcal{A} - \widehat{A}_\gamma\|\|V^\star\| + \|\mathcal{D}_a(\mathcal{B}V^\star) - \mathcal{D}_a(\widehat{B}_\gamma V^\star)\|_{L^2_\mu} \\
&\le \left(\mathcal{E}(\widehat{A}_\gamma) + L_D\mathcal{E}(\widehat{B}_\gamma)\right)\|V^\star\|_{L^2_\mu}
\end{aligned}$$

Here, $c_1 < \infty$ and $c_2 < \infty$ are the Lipschitz constants of $\widehat{D}_F$ with respect to the operators and the arguments, respectively. Moreover, since the exact HJB is exponentially stable, we know that $D_F^\star$ satisfies $\mathrm{SpectralGap}(D_F^\star) = \lambda_{\mathrm{gap}} + \rho$. Because the spectral gap of a linear operator is locally Lipschitz continuous with respect to small perturbations (Kloeckner, 2018; Kato, 2013), we further find that

$$\mathrm{SpectralGap}(\widehat{D}_F(\widehat{V})) = \lambda_{\mathrm{gap}} + \rho - C_2\left((\mathcal{E}(\widehat{A}_\gamma) + L_{\mathcal{D}}\mathcal{E}(\widehat{B}_\gamma)) + \|\widehat{V} - V^\star\|_{L^2_\mu}\right)$$

for some constant $C_2 < \infty$ under the additional assumption that the approximation errors $\mathcal{E}(\widehat{A}_\gamma)$ and $\mathcal{E}(\widehat{B}_\gamma)$ are sufficiently small. Consequently, Banach's fixed point theorem (or Schauder's extension for bounded operators) (Dontchev and Rockafellar, 2009) implies that there exists a solution of $\widehat{V}^\star \in \mathcal{H}_{\mathbb{S}}$ of the equilibrium equation $\widehat{\mathcal{T}}(\widehat{V}^\star) = 0$. It is further known that $\widehat{V}^\star$ is a locally exponentially stable equilibrium of $\widehat{\mathcal{T}}$, which must be in a local neighborhood of $V^\star$. Finally, Banach's fixed-point theorem implies that we have

$$\|\widehat{V}^\star - V^\star\|_{L^2_\mu} \le \frac{C_1(\mathcal{E}(\widehat{A}_\gamma) + L_D\mathcal{E}(\widehat{B}_\gamma))}{\lambda_{\mathrm{gap}} + \rho - C_2\left((\mathcal{E}(\widehat{A}_\gamma) + L_D\mathcal{E}(\widehat{B}_\gamma))\right)} \tag{54}$$

for some constant or some constants $C_1, C_2 \in (0, \infty)$, recalling that $\mathcal{E}(\widehat{A}_\gamma)$, $\mathcal{E}(\widehat{B}_\gamma)$ need to be sufficiently small in order to ensure that the denominator in (54) is strictly positive. Recall that we showed that $\max\{\mathcal{E}(\widehat{A}_\gamma), \mathcal{E}(\widehat{B}_\gamma)\} \le \mathcal{E}(\widehat{G}_\gamma) \le \delta$ in Section (D.3). We define

$$\widehat{\lambda}_{\mathrm{gap}} = \lambda_{\mathrm{gap}} - C_2\left((\mathcal{E}(\widehat{A}_\gamma) + L_D\mathcal{E}(\widehat{B}_\gamma))\right)$$

for $C_2 > 0$. Thus, (54) further simplifies to

$$\|\widehat{V}^\star - V^\star\|_{L^2_\mu} \le C(\widehat{\lambda}_{\mathrm{gap}} + \rho)^{-1}\delta. \qquad C > 0, \qquad\qquad (\textsc{Learning Error})$$

where δ is the upper bound on $\mathcal{E}(\widehat{G}_\gamma)$, and recalling that $\max\{\mathcal{E}(\widehat{A}_\gamma), \mathcal{E}(\widehat{B}_\gamma)\} \le \mathcal{E}(\widehat{G}_\gamma)$ (Section D.3).

Exponential Convergence of the Approximate Value Function A bound on the term $\|\widehat{V}^\star - \widehat{V}_T\|_{L^2_\mu}$ can be found by using the above considerations, where we already established local exponential convergence of the approximate HJB. Since the exact HJB is globally exponentially stable (see Lemma 1), the statement of this result follows directly using a variant of Gronwall's lemma for operators (Kostic et al., 2024b) and the result from the previous theorem. Then, given $\widehat{V}_0 = \widehat{V}(0)$ we have

$$\|\widehat{V}^\star - \widehat{V}_T\|_{L^2_\mu} \le M e^{-(\widehat{\lambda}_{\mathrm{gap}} + \rho)\,T} \|\widehat{V}^\star - \widehat{V}_0\|_{L^2_\mu}, \qquad \text{(CONVERGENCE ERROR)}$$

for some constant $M < \infty$ and $\widehat{\lambda}_{\mathrm{gap}} = \lambda_{\mathrm{gap}} - C_2\left(\mathcal{E}(\widehat{A}_\gamma) + L_{\mathcal{D}}\,\mathcal{E}(\widehat{B}_\gamma)\right)$. In detail, since both the exact HJB as well as the approximate HJB are globally Lipschitz continuous, Gronwall's lemma implies that there exists for every given $T' < \infty$ a constant $C_3 < \infty$ such that

$$\|V_{T'} - \widehat{V}_{T'}\| \;\le\; C_3\left((\mathcal{E}(\widehat{A}_\gamma) + L_D\mathcal{E}(\widehat{B}_\gamma))\right),$$

where $V_{T'}$ is in the local neighborhood of $V^\star$ in which we have exponential convergence. Here, we recall our assumption that the error on the right-hand side of this inequality is sufficiently small, which then also implies that (CONVERGENCE ERROR) holds.

Discretization Error Constructing $\widehat{V}_T$ involves infinite-dimensional but finite-rank operators, so that the discretization error satisfies

$$\|\widehat{V}_T - \widehat{V}_k\|_{L^2_\mu} \le C\|\boldsymbol{v}_T - \boldsymbol{v}_k\|_2, \qquad \boldsymbol{v}_T, \boldsymbol{v}_k \in \mathbb{R}^N,$$

for some $C > 0$, where we use the bounded operator $\widehat{S}^*_S : \mathbb{R}^N \to \mathcal{H}_\mathbb{S}$ to define $\widehat{V}_T := \widehat{S}^*_S \boldsymbol{v}_T$. Following Frontin and Darmofal (2022); Wanner and Hairer (1996); Viswanath (2001), we establish the global discretization error at final time T, namely

$$\|\boldsymbol{v}_T - \boldsymbol{v}_k\|_2 \le \mathcal{E}(T, p)\,(\Delta t)^p,$$

where p is the order of accuracy of the discretization method and $\mathcal{E}(T, p)$ remains bounded uniformly in T for all finite horizons. Then, under the assumptions derived in Theorem 1, the global discretization error satisfies

$$\|\widehat{V}_T - \widehat{V}_k\|_{L^2_\mu} \le C\,\mathcal{E}(T, p)\,(\Delta t)^p \propto \mathcal{O}((\Delta t)^p). \qquad \text{(DISCRETIZATION ERROR)}$$

Finally summing (LEARNING ERROR), (CONVERGENCE ERROR) and (DISCRETIZATION ERROR) yields the result of Theorem 1. $\qquad\square$

Corollary 2. *Let $\alpha \in (1, 2]$ denote the regularity of $\mathcal{G}$ and $\beta \in (0, 1]$ the spectral decay rate of $\mathcal{H}_\mathbb{Z}$, and choose $\gamma \asymp N^{-\frac{1}{\alpha+\beta}}$. Then, for any $\xi \in (0, 1)$ there exists $c > 0$ such that, with probability at least $1 - \xi$ over an i.i.d. sample $\mathbb{D}_N$, the learning error in (24a) satisfies a finite-sample bound.*

$$\|V^\star - \widehat{V}^\star\|_{L^2_\mu} \le c\,(\widehat{\lambda}_{\mathrm{gap}} + \rho)^{-1} N^{-\frac{\alpha}{2(\alpha+\beta)}}\,\ln \xi^{-1}. \tag{25}$$

Proof. The result directly follows from substituting the rates derived for $\mathcal{E}(\widehat{A}_\gamma)$ and $\mathcal{E}(\widehat{B}_\gamma)$ from Theorem 2 into (LEARNING ERROR) for δ. Looking at (54), we could even extract the factor $(L_D + 1)$ from the constant c, where L_D is the Lipschitz constant of the Fenchel conjugate, to obtain a refined bound

$$\|V^\star - \widehat{V}^\star\|_{L^2_\mu} \le c'\,(L_D + 1)(\widehat{\lambda}_{\mathrm{gap}} + \rho)^{-1} N^{-\frac{\alpha}{2(\alpha+\beta)}}\,\ln \xi^{-1}, \qquad c' > 0.$$

$\qquad\square$

F   ADDITIONAL EXPERIMENTS

Complexity of Algorithm 1 The computational cost of the World Model Construction is dominated by the $\mathcal{O}(N^3)$ term arising from the factorization of K_γ and the subsequent solves required to obtain $\boldsymbol{A}$. Building the state kernel matrix K_S requires $\mathcal{O}(N^2 n_s)$ flops for typical kernels that depend on pairwise Euclidean distances or dot products in $\mathbb{R}^{n_s}$ (e.g., squared-exponential or Matérn kernels). Forming K_γ through the interaction term

$UU^\top$ adds another $\mathcal{O}(N^2(n_s + n_a))$ operations. Similarly, evaluating the target matrix K_d scales as $\mathcal{O}(N^2 n_s)$, provided closed-form expressions for the kernel derivatives are available, which is the case for the aforementioned popular kernels.

The Dynamic Programming Recursion is likewise dominated by an $\mathcal{O}(N^3)$ factorization of $\boldsymbol{M}$. Per iteration, evaluating the product $\boldsymbol{Bv}$ costs $\mathcal{O}(N^2 n_a)$, while computing the Fenchel operator $D_{\boldsymbol{a}}$ pointwise over all N samples scales as $\mathcal{O}(N\,C(n_a))$, where $C(n_a)$ denotes the cost of evaluating one Fenchel conjugate $D_{\boldsymbol{a}}(\boldsymbol{\lambda}_i)$. Applying the pre-factorized matrix $\boldsymbol{M}$ then adds $\mathcal{O}(N^2)$ operations per iteration. The resulting overall complexities are summarized in Table 5. These results correspond to dense matrix operations. Lower complexity can be achieved when exploiting sparsity. Importantly, the computational cost scales linearly with the state dimension (and with the action dimension in the case of separable action penalties) and can be further reduced using sketching techniques such as Nyström approximations (Rudi et al., 2017).

Task	Complexity
Operator World Model Construction	$\mathcal{O}(N^3 + N^2(n_s + n_a))$
Dynamic Programming	$\mathcal{O}(N^3 + k_{\max}(N^2 n_a + N\,C(n_a)))$

Table 5: Computational complexity of Algorithm 1. $C(n_a)$ denotes the complexity of evaluating the Fenchel conjugate $\boldsymbol{D_a}(\boldsymbol{\lambda})$ on one element of $\boldsymbol{\lambda}$, e.g. $C(n_a) = n_a$ for separable action penalties or n_a^2 for quadratic (coupled) action penalties

F.1 Proof-of-Concept Examples

We evaluate our learning error rates on linear and nonlinear process dynamics (Figure 2). While these are often studied using different policy classes, linear and nonlinear (Tu and Recht, 2019), our convergence analysis covers both on equal footing – without any parametric assumptions (Recht, 2019). We run our proposed Algorithm 1 over different seeds and gather i.i.d data to form the quantiles and means in Figure 2. In both cases, we use a squared exponential (SE) kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{y}\|^2 / 2\sigma^2\right)$ while the data samples are drawn randomly from a uniform distribution. The test dataset for the value function was from a uniform grid of 1000 points on the interval $[-3, 3]$. For all our runs, Algorithm 1 discretization parameters were set to $k_{max} = 1000$ and $\Delta t = 0.01s$.

Linear SDE with Additive Action Linear-quadratic (LQ) control problems play an important role in the control literature. They provide explicit solutions and, often, nonlinear ones can be approximated by LQ ones (Zhao et al., 2023). To validate our findings, we study the value function convergence for an Ornstein-Uhlenbeck process $dS_t = (-S_t + a_t)dt + \sqrt{2\epsilon}\,dW_t$ (Houska, 2025) where the optimal value function is $V^\star(s) = s^2$ when setting $\rho = 0$, for any ϵ. We set $\epsilon = 0.01$, and a reward function $r(x, a) = -3s^2 - a^2$. The hyperparameter for the used SE kernel is $\sigma = 10$ with regularizer $\gamma = 10^{-10}$. The statistics in Figure 2 (left) are obtained from 10 iid runs. $[-1, 1] \times [-3.5, 3.5]$. Figure 4 shows how well our learned policy and value functions (using 200 training data) compare to the ground truth.

Nonlinear SDE with Affine Action We transfer to a nonlinear setting, using an action-affine benchmark system $dS_t = (f(S_t) + g(S_t)a_t)dt + \sqrt{2\epsilon}\,dW_t$, where $g(s) = \frac{1}{2} + \sin(s)$ and $f(s) = -\frac{1}{2}(1 - g(s)^2)$ (Doyle et al., 1996). Here, the reward is $r(s, a) = -s^2 - a^2$ and, practically, the optimal value function approaches $V_\infty = s^2$ as ϵ tends to zero. The hyperparameter for the used SE kernel is $\sigma = 1$ and we set $\epsilon = 0.01$ with regularizer $\gamma = 10^{-8}$. The statistics in Figure 2 (right) are obtained from 8 iid runs. The data is drawn from the state-action set $[-3, 3] \times [-3.5, 3.5]$. Figure 5 shows how well our learned policy and value functions (using 200 training data) compare to the reference ground truth.

Pendulum-Gym We evaluate Algorithm 1 on Gymnasium `Pendulum-v1` (Towers et al., 2024). The action is a torque $a \in [-2, 2]$, and the observation is $\boldsymbol{s} = (\cos\theta, \sin\theta, \dot{\theta})$ with $\cos\theta, \sin\theta \in [-1, 1]$ and $\dot{\theta} \in [-8, 8]$. Episodes truncate at 200 steps. Following the official reward, we split the state term and action cost as $r_{\boldsymbol{s}}(\theta, \dot{\theta}) = -(\theta^2 + 0.1\,\dot{\theta}^2)$ and $c_{\boldsymbol{a}}(a) = 0.001\,a^2$, so the maximum achievable reward is 0 (upright, zero velocity, zero torque). We run Algorithm 1 with $\rho = 0.1$, $\sigma = 3$, $\gamma = 10^{-7}$, and $k_{\max} = 1000$. The resulting value function and policy are shown in Figure 6 on the offline `d3rlpy` (Seno and Imai, 2022) dataset ("Replay"), with 8000 subsamples of (state, action, next state) tuples, for which we obtain infinitesimal samples via finite differences.
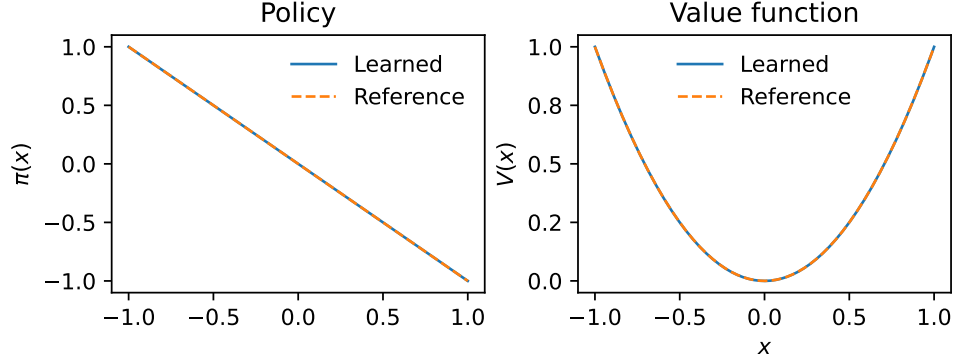
Figure 4: Comparison of learned (200 datapoints) and reference (ground truth) value function and policy linear SDE with additive action, demonstrating that we can effectively learn unknown value functions and policies without parametric assumptions.
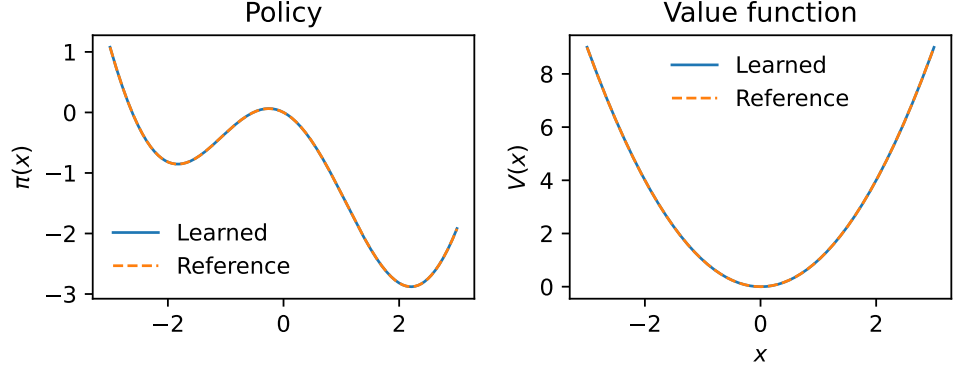

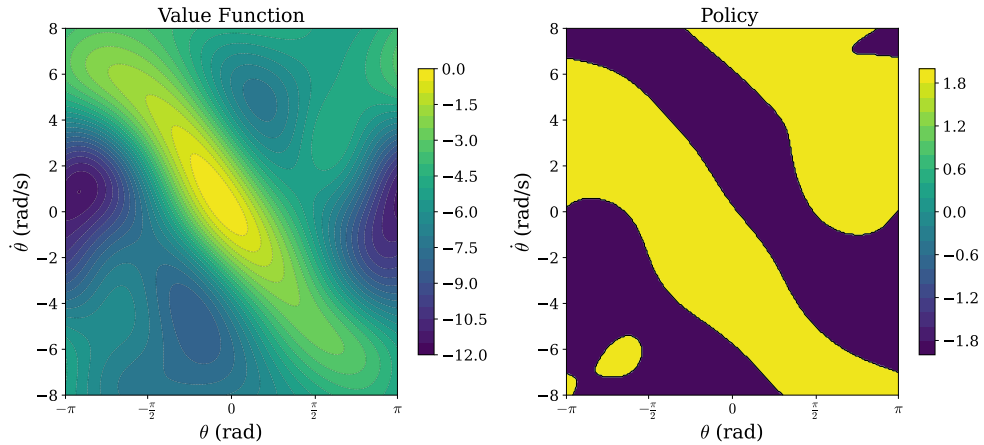Figure 5: Comparison of learned (200 datapoints) and reference (ground truth as ϵ goes to zero) value function for the nonlinear SDE with affine action, demonstrating that we can effectively learn unknown value functions and policies without parametric assumptions.

Figure 6: Value function (left) and policy (right) output from O-CTRL, trained on 8000 datapoints from the offline `d3rlpy` dataset ("Replay").

References

Robert A Adams and John JF Fournier. *Sobolev spaces*, volume 140. Elsevier, 2003.

Brian DO Anderson and John B Moore. *Optimal control: linear quadratic methods*. Courier Corporation, 2007.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in neural information processing systems*, 20, 2007.

Ludwig Arnold. *Stochastic differential equations: theory and applications*, volume 2. John Wiley & Sons, 1974.

Alex Ayoub, Kaiwen Wang, Vincent Liu, Samuel Robertson, James McInerney, Dawen Liang, Nathan Kallus, and Csaba Szepesvári. Switching the loss reduces the cost in batch reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 2135–2158, 2024.

Petar Bevanda, Bas Driessen, Lucian Cristian Iacob, Stefan Sosnowski, Roland Tóth, and Sandra Hirche. Nonparametric control Koopman operators. *arXiv preprint arXiv:2405.07312*, 2025a.

Petar Bevanda, Nicolas Hoischen, Tobias Wittmann, Jan Brudigam, Sandra Hirche, and Boris Houska. Kernel-based optimal control: An infinitesimal generator approach. In *Proceedings of the 7th Annual Learning for Dynamics and Control Conference*, volume 283, pages 1038–1052. PMLR, 04–06 Jun 2025b.

Jonas Blessing, Lianzi Jiang, Michael Kupper, and Gechun Liang. Convergence rates for chernoff-type approximations of convex monotone semigroups. *Stochastic Processes and their Applications*, page 104700, 2025.

Stephen Boyd. Convex optimization. *Cambridge UP*, 2004.

Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern Koopman theory for dynamical systems. *SIAM Review*, 64(2):229–340, 2022.

Kaylee Burns, Tianhe Yu, Chelsea Finn, and Karol Hausman. Offline reinforcement learning at multiple frequencies. In *Conference on robot learning*, pages 2041–2051. PMLR, 2023.

Edoardo Caldarelli, Antoine Chatalic, Adrià Colomé, Cesare Molinari, Carlos Ocampo-Martinez, Carme Torras, and Lorenzo Rosasco. Linear quadratic control of nonlinear systems with Koopman operator learning and the nyström method. *Automatica*, 177:112302, 2025.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pages 1042–1051. PMLR, 2019.

Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.

Kenji Doya. Reinforcement learning in continuous time and space. *Neural computation*, 12(1):219–245, 2000.

John Doyle, James A Primbs, Benjamin Shapiro, and Vesna Nevistic. Nonlinear games: examples and counterexamples. In *35th Conference on Decision and Control*, volume 4, pages 3915–3920. IEEE, 1996.

Klaus-Jochen Engel, Rainer Nagel, and Simon Brendle. *One-parameter semigroups for linear evolution equations*, volume 194. Springer, 2000.

Heinz W Engl and Ronny Ramlau. Regularization of inverse problems. In *Encyclopedia of applied and computational mathematics*, pages 1233–1241. Springer, 2015.

Amir-massoud Farahmand. *Regularization in Reinforcement Learning*. Phd thesis, University of Alberta, 2011.

Wendell H. Fleming and H. Mete Soner. *Controlled Markov Processes and Viscosity Solutions*. Springer, New York, 2006.

Cory V Frontin and David L Darmofal. Output error behavior for discretizations of ergodic, chaotic systems of ordinary differential equations. *Physics of Fluids*, 34(10), 2022.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Silvana Patterson, Arthur Gretton, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *International Conference on Machine Learning*, pages 1823–1830. Omnipress, 2012.

Igor Halperin. Distributional offline continuous-time reinforcement learning with neural physics-informed pdes (sciphy rl for doctr-l). *Neural Computing and Applications*, 36(9):4643–4659, 2024.

R. Z. Has'minskii. Ergodic properties of recurrent diffusion processes and stabilization of the solution of the Cauchy problem for parabolic equations. *Theory of Probability and its Applications*, 5:179–196, 1960.

Yinnian He. Euler implicit/explicit iterative scheme for the stationary navier–stokes equations. *Numerische Mathematik*, 123(1):67–96, 2013.

M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE Constraints*. Springer, 2009.

Samuel Holt, Alihan Hüyük, Zhaozhi Qian, Hao Sun, and Mihaela van der Schaar. Neural laplace control for continuous-time delayed systems. In *International Conference on Artificial Intelligence and Statistics*, pages 1747–1778. PMLR, 2023.

Boris Houska. Convex operator-theoretic methods in stochastic control. *Automatica*, 177:112274, 2025. ISSN 0005-1098.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26):eaau5872, 2019.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Yanwei Jia and Xun Yu Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55, 2022a.

Yanwei Jia and Xun Yu Zhou. Policy gradient and actor-critic learning in continuous time and space: Theory and algorithms. *Journal of Machine Learning Research*, 23(275):1–50, 2022b.

Yanwei Jia and Xun Yu Zhou. q-learning in continuous time. *Journal of Machine Learning Research*, 24(161): 1–61, 2023.

Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987, 2023.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. Morel: Model-based offline reinforcement learning. *Advances in neural information processing systems*, 33:21810–21823, 2020.

Jeongho Kim, Jaeuk Shin, and Insoon Yang. Hamilton-jacobi deep q-learning for deterministic continuous-time systems with lipschitz continuous controls. *Journal of Machine Learning Research*, 22(206):1–34, 2021.

Benoit R. Kloeckner. Effective perturbation theory for simple isolated eigenvalues of linear operators. *Journal of Operator Theory*, 81(1):175–194, 2018.

Milan Korda and Igor Mezić. On convergence of extended dynamic mode decomposition to the koopman operator. *Journal of Nonlinear Science*, 28(2):687–710, 2018.

Vladimir Kostic, Pietro Novelli, Andreas Maurer, Carlo Ciliberto, Lorenzo Rosasco, and Massimiliano Pontil. Learning dynamical systems via Koopman operator regression in Reproducing Kernel Hilbert Spaces. *Advances in neural information processing systems*, 35:4017–4031, 2022.

Vladimir Kostic, Karim Lounici, Pietro Novelli, and Massimiliano Pontil. Sharp spectral rates for Koopman operator learning. *Advances in neural information processing systems*, 36:32328–32339, 2023.

Vladimir Kostic, Hélène Halconruy, Timothée Devergne, Karim Lounici, and Massimiliano Pontil. Learning the infinitesimal generator of stochastic diffusion processes. *Advances in Neural Information Processing Systems*, 37:137806–137846, 2024a.

Vladimir Kostic, Pietro Novelli, Riccardo Grazzi, Karim Lounici, and Massimiliano Pontil. Learning invariant representations of time-homogeneous stochastic dynamical systems. In *International Conference on Learning Representations*, 2024b.

Vladimir R. Kostic, Karim Lounici, Prune Inzerilli, Pietro Novelli, and Massimiliano Pontil. Consistent long-term forecasting of ergodic dynamical systems. In *International Conference on Machine Learning*. PMLR, 2024c.

Vladimir R. Kostic, Karim Lounici, Hélène Halconruy, Timothée Devergne, Pietro Novelli, and Massimiliano Pontil. Laplace transform based low-complexity learning of continuous Markov semigroups. In *International Conference on Machine Learning*, pages 31560–31589. PMLR, 2025.

Ilya Kostrikov, Rob Fergus, Jonathan Tompson, and Ofir Nachum. Offline reinforcement learning with fisher divergence critic regularization. In *International Conference on Machine Learning*, pages 5774–5783. PMLR, 2021.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022.

Toshiyuki Koto. Imex runge–kutta schemes for reaction–diffusion equations. *Journal of Computational and Applied Mathematics*, 215(1):182–195, 2008.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Zhu Li, Dimitri Meunier, Mattes Mollenhauer, and Arthur Gretton. Optimal rates for regularized conditional mean embedding learning. *Advances in Neural Information Processing Systems*, 35:4433–4445, 2022.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Michael Lutter, Boris Belousov, Kim Listmann, Debora Clever, and Jan Peters. HJB optimal feedback control with deep differential value functions and action constraints. In *Conference on Robot Learning*, pages 640–650. PMLR, 2020.

Michael Lutter, Boris Belousov, Shie Mannor, Dieter Fox, Animesh Garg, and Jan Peters. Continuous-time fitted value iteration for robust policies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5): 5534–5548, 2023.

Yiming Meng, Ruikun Zhou, Amartya Mukherjee, Maxwell Fitzsimmons, Christopher Song, and Jun Liu. Physics-informed neural network policy iteration: algorithms, convergence, and verification. In *International Conference on Machine Learning*, pages 35378–35403. PMLR, 2024.

Dimitri Meunier, Zhu Li, Arthur Gretton, and Samory Kpotufe. Nonlinear meta-learning can guarantee faster rates. *SIAM Journal on Mathematics of Data Science*, 7(4):1594–1615, 2025.

Erfan Mirzaei, Andreas Maurer, Vladimir R. Kostic, and Massimiliano Pontil. An empirical bernstein inequality for dependent data in hilbert spaces and applications. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2025.

Mattes Mollenhauer and Péter Koltai. Nonparametric approximation of conditional expectation operators. *arXiv preprint arXiv:2012.12917*, 2020.

Mattes Mollenhauer, Nicole Mücke, and T. J. Sullivan. Learning linear operators: infinite-dimensional regression as a well-behaved non-compact inverse problem. *arXiv preprint arXiv:2211.08875*, 2022.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

Rémi Munos. A study of reinforcement learning in the continuous case by the means of viscosity solutions. *Machine Learning*, 40:265–299, 2000.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103:127–152, 2005.

H. Nijmeijer and A. J. van der Schaft. *Nonlinear Dynamic Control Systems*. Springer, 1996.

Pietro Novelli, Marco Pratticò, Massimiliano Pontil, and Carlo Ciliberto. Operator world models for reinforcement learning. *Advances in Neural Information Processing Systems*, 37:111432–111463, 2024.

Ross G Pinsky. Spectral gap and rate of convergence to equilibrium for a class of conditioned brownian motions. *Stochastic processes and their applications*, 115(6):875–889, 2005.

Rafael Figueiredo Prudencio, Marcos ROA Maximo, and Esther Luna Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):253–279, 2019.

Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. Falkon: An optimal large scale kernel method. *Advances in neural information processing systems*, 30, 2017.

Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. In *International Conference on Machine Learning*, pages 4651–4660. PMLR, 2017.

Boscarino Sebastiano. High-order semi-implicit schemes for evolutionary partial differential equations with higher order derivatives. *Journal of Scientific Computing*, 96(1):11, 2023.

Takuma Seno and Michita Imai. d3rlpy: an offline deep reinforcement learning library. *Journal of Machine Learning Research*, 23(315):1–20, 2022.

Alena Shilova, Thomas Delliaux, Philippe Preux, and Bruno Raffin. *Learning HJB viscosity solutions with PINNs for continuous-time reinforcement learning*. PhD thesis, Inria Lille–Nord Europe, CRIStAL–Centre de Recherche en Informatique, Signal et Automatique de Lille, 2024.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, NY, first edition, 2008.

Corentin Tallec, Léonard Blier, and Yann Ollivier. Making deep q-learning methods robust to time discretization. In *International Conference on Machine Learning*, pages 6096–6104. PMLR, 2019.

Prem Talwai, Ali Shameli, and David Simchi-Levi. Sobolev norm learning rates for conditional mean embeddings. In *International conference on artificial intelligence and statistics*, pages 10422–10447. PMLR, 2022.

Yuval Tassa, Nicolas Mansard, and Emo Todorov. Control-limited differential dynamic programming. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1168–1175. IEEE, 2014.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: a physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

Maximilian Tölle, Theo Gruner, Daniel Palenicek, Tim Schneider, Jonas Günster, Joe Watson, Davide Tateo, Puze Liu, and Jan Peters. Towards safe robot foundation models using inductive biases. *arXiv preprint arXiv:2505.10219*, 2025.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

Stephen Tu and Benjamin Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on learning theory*, pages 3036–3083. PMLR, 2019.

Divakar Viswanath. Global errors of numerical ode solvers and lyapunov's theory of stability. *IMA journal of numerical analysis*, 21(1):387–406, 2001.

Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous time and space: A stochastic control approach. *Journal of Machine Learning Research*, 21(198):1–34, 2020.

Gerhard Wanner and Ernst Hairer. *Solving ordinary differential equations II*, volume 375. Springer Berlin Heidelberg New York, 1996.

Li Kevin Wenliang, Gregoire Deletang, Matthew Aitchison, Marcus Hutter, Anian Ruoss, Arthur Gretton, and Mark Rowland. Distributional bellman operators over mean embeddings. In *International Conference on Machine Learning*, pages 52839–52868. PMLR, 2024.

Grady Williams, Nolan Wagener, Brian Goldfain, Paul Drews, James M Rehg, Byron Boots, and Evangelos A Theodorou. Information theoretic mpc for model-based reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 1714–1721. IEEE, 2017.

Harley Wiltzer, Marc Bellemare, David Meger, Patrick Shafto, and Yash Jhaveri. Action gaps and advantages in continuous-time distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 47815–47848, 2024.

W. M. Wonham. Liapunov criteria for weak stochastic stability. *Journal of Differential Equations*, 2:195–207, 1966.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33: 14129–14142, 2020.

Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems*, 34: 28954–28967, 2021.

Zhiyue Zhang, Hongyuan Mei, and Yanxun Xu. Continuous-time decision transformer for healthcare applications. In *International Conference on Artificial Intelligence and Statistics*, pages 6245–6262. PMLR, 2023.

Hanyang Zhao, Wenpin Tang, and David Yao. Policy optimization for continuous reinforcement learning. *Advances in Neural Information Processing Systems*, 36:13637–13663, 2023.