Continuous time policy evaluation is easier with noisy dynamics

Samuel Robertson SMROBERT@UALBERTA.CA
University of Alberta

Thomas Newton TNEWTON@UWATERLOO.CA
University of Waterloo

Csaba Szepesvári SZEPI@GOOGLE.COM
Google Deepmind and University of Alberta

Editors: Steve Hanneke and Tor Lattimore

Abstract

In this work, we study continuous-time stochastic control problems governed by controlled stochastic differential equations with unknown dynamics. We focus on the discounted infinite-horizon setting and restrict attention to feedback controllers. In general, the continuous time value function is the solution to the nonlinear Hamilton-Jacobi-Bellman (HJB) equation, which typical only admits viscosity solutions with no regularity. Our first contribution is to establish sharp regularity results for value functions using elliptic partial differential equation theory. Under mild growth and regularity assumptions on the controlled dynamics and a uniform ellipticity condition on the diffusion, we show that the value function belongs to a Matérn reproducing kernel Hilbert space (RKHS) that is strictly smoother than the running reward. Building on this analysis, we develop a kernel-based policy evaluation method that estimates value functions directly from online trajectory rollouts of a fixed policy. The resulting algorithm exploits the RKHS structure with a kernel ridge regression technique, reducing the infinite-dimensional learning problem to a finite-dimensional one. Our results establish a direct connection between stochastic control, elliptic regularity theory, and kernel methods, and provide a foundation for online policy evaluation and policy improvement in continuous time.

Keywords: Controlled stochastic differential equations, continuous time reinforcement learning, Hamilton–Jacobi–Bellman equation, kernel ridge regression.

1. Introduction

We study continuous-time stochastic control problems governed by controlled stochastic differential equations (SDEs) of the form

$$dx(t) = \mu(x(t), u(t))\, dt + \sigma(x(t), u(t))\, dW(t), \tag{1a}$$

$$x(0) = x_0, \tag{1b}$$

where $W(t)$ is a d-dimensional Wiener process and the drift μ and diffusion σ are unknown. The objective is to maximize a discounted infinite–horizon reward functional

$$J(x_0, u) := \mathbb{E}\left[\int_0^\infty e^{-\beta t} r(x(t), u(t))\, dt\right], \tag{2}$$

with running reward r and discount rate $\beta > 0$.

Even when the dynamics are known, continuous-time stochastic control problems of this form are notoriously difficult. Establishing the existence of optimal controllers requires delicate technical assumptions, and computing optimal controllers is typically intractable in all but the simplest settings. Two classical theoretical frameworks have been developed to characterize optimality in such systems (see Yong and Zhou (1999) for details). Pontryagin's Maximum Principle provides necessary conditions for optimality in terms of first-order variations of the cost functional, leading to a coupled system of forward-backward stochastic differential equations. While powerful, this approach produces open-loop controllers tailored to a single initial condition and thus is poorly suited to settings where robustness or repeated control over a range of initial states is required. In practice, this lack of closed-loop controllers is particularly problematic, as any numerical solution of optimality conditions necessarily incurs error. Such errors must then be compensated for typically through the introduction of additional trajectory-stabilizing control terms.

An alternative framework is Bellman's Dynamic Programming (DP), which characterizes optimality through the value function

$$V(t_0, x_0) := \sup_u \mathbb{E}_u \left[\int_{t_0}^{\infty} e^{-\beta(t-t_0)} r(x(t), u(t))\, dt \,\middle|\, x(t_0) = x_0 \right]. \tag{3}$$

The DP principle yields a global characterization of optimality and naturally leads to feedback (closed-loop) control laws. However, in continuous time the optimality conditions take the form of the Hamilton-Jacobi-Bellman (HJB) equation,

$$\partial_t V(t,x) = \sup_{u \in U} \left[r(x,u) + \mu(x,u) \cdot \nabla V(t,x) + \frac{1}{2} \operatorname{Tr}\left(\sigma(x,u)\sigma(x,u)^T \nabla^2 V(t,x) \right) \right], \tag{4}$$

which is a fully nonlinear second-order partial differential equation posed over the state space. In general, solutions of the HJB equation lack any regularity and must be interpreted in the viscosity sense, which is technically challenging and problematic for numeric routines. From a computational perspective, solving the HJB equation is effectively intractable in moderate to high dimensions due to the curse of dimensionality, even when the dynamics μ and σ are known.

In this work, we take a different perspective. We focus on establishing sharp regularity conditions under which the value function associated with a given feedback controller is sufficiently smooth to be well approximated using kernel ridge regression. This approach leverages the deep connections between stochastic differential equations, elliptic partial differential equations, and reproducing kernel Hilbert spaces, which is of standalone theoretical interest. We then apply this analysis to develop an online policy evaluation algorithm with performance guarantees, and provide the theoretical foundation for future development of policy improvement algorithms in continuous time.

Contributions

- Under uniform ellipticity and mild growth and regularity conditions on the controlled dynamics, we give a sharp characterization of the regularity gain enjoyed by value functions relative to the running reward.
- We establish that these value functions belong to Matérn reproducing kernel Hilbert spaces whose smoothness is determined by the elliptic operator induced by the closed-loop dynamics.

- We leverage this RKHS characterization to design an online kernel-based policy evaluation algorithm that estimates value functions directly from trajectory rollouts of a policy.

Related literature

Yin et al. (2021) is the original inspiration for our work, and they achieved a similar result in the analogous discrete time (reinforcement learning) setting. Specifically, they studied a discrete time control problem where the transitions had Gaussian noise, and exploited the noise properties for algorithm design, handling policy estimation and improvement. The extension to continuous time is challenging because the random noise is continuously fed into the mean dynamics; even if the noise is a Wiener process, the distribution over a discrete time interval will not be Gaussian.

While most of the RL theory community focuses on discrete time settings (Ayoub et al., 2024; Bakhtiari et al., 2025; Foster and Rakhlin, 2023; Robertson et al., 2025), there have been many works from various communities studying continuous-time stochastic control (Zhu, 2024; Zhu et al., 2025; Jia and Zhou, 2025; Mou, 2025). Like us, Zhu (2024) studied policy evaluation in a controlled diffusion set up, but they assumed access to only discritized observations. They showed that the standard discrete Bellman equation from the reinforcement learning literature gets a suboptimal rate of convergence to the true value function as the discritization size goes to 0, and proposed an improved equation. Without our assumption on the regularity of the noise, their bounds depend on the gradient and Hessian of the mean dynamics. Zhu et al. (2025) extended Zhu (2024) to an algorithm for policy improvement. Jia and Zhou (2025) searched for a continuous-time analog of the Q-function, or action-advantage function, from the RL literature, since the naive extension is not meaningful. The developed algorithms for policy improvement, but ultimately had to discritize time and thus ended up optimizing only over discrete time policies. Mou (2025) studied controlled stochastic differential equations in a finetuning framework, where the goal is to control a diffusion process to minimize its variational distance to another process, with applications to generative models.

2. Preliminaries and problem formulation

We now formalize the problem setting and state the assumptions used throughout the paper. We first lay out quickly some notational conventions that we will use freely below. I_k is the identity matrix of dimension k, $\mathcal{C} \sqsubset \mathcal{X}$ means $\mathcal{C}$ is a sublist of $\mathcal{X}$, $\mathcal{M}_1(\mathcal{X})$ is the probability simplex on $\mathcal{X}$, $\mathcal{B}(\mathcal{X}, B)$ is the ball of radius B on the metric space $\mathcal{X}$, and, given a list $\mathcal{C} \sqsubset \mathcal{X}$ and a kernel k defined on $\mathcal{X}$, $K_{\mathcal{C}} := (k(\mathcal{C}_i, \mathcal{C}_j)_{i,j}$ and $k_{\mathcal{C}}(x) := (k(\mathcal{C}_i, x))_i$.

We consider a d-dimensional controlled stochastic system of the form

$$dx(t) = \mu\big(x(t), u(x(t))\big)\, dt + \sigma\big(x(t), u(x(t))\big)\, dW(t), \tag{5a}$$

$$x(0) = x_0, \tag{5b}$$

where $u : \mathbb{R}^d \to U \subset \mathbb{R}^m$ is a measurable feedback controller. The functions $\mu : \mathbb{R}^d \times U \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \times U \to \mathbb{R}^{d\times d}$ denote the drift and diffusion coefficients, respectively.

To guarantee that the closed-loop dynamics are well posed for every feedback controller, we impose the following standard assumptions.

Assumption 1 (Locally Lipschitz, Linear Growth Dynamics) *For every $\delta > 0$ there exists $L(\delta) > 0$ such that for all $x, y \in \mathbb{R}^d$ with $\|x\|_2, \|y\|_2 \leq \delta$ and all $a \in U$,*

$$\|\mu(x,a) - \mu(y,a)\|_2 + \|\sigma(x,a) - \sigma(y,a)\|_2 \leq L(\delta)\|x - y\|_2. \tag{6}$$

Moreover, there exists $C_D > 0$ such that for all $(x,a) \in \mathbb{R}^d \times U$,

$$\|\mu(x,a)\|_2 + \|\sigma(x,a)\|_2 \leq C_D\big(1 + \|x\|_2\big). \tag{7}$$

Under Assumption 1, the stochastic differential equation (5) admits a strong solution for every measurable feedback controller $u : \mathbb{R}^d \to U$, see for example Theorem 32.3 and Proposition 32.6 of Kallenberg (2021).

Given a running reward $r : \mathbb{R}^d \times U \to \mathbb{R}$, we define the discounted infinite-horizon performance functional

$$J(s,u) := \mathbb{E}\left[\int_0^\infty e^{-\beta t} r(x(t), u(x(t)))\, dt \,\middle|\, x(0) = s\right]. \tag{8}$$

Assumption 2 (Polynomial Growth of Running Reward) *There exist constants $C_r > 0$ and $p \geq 1$ such that*

$$|r(x,a)| \leq C_r\big(1 + \|x\|_2^p\big), \tag{9}$$

for all $(x,a) \in \mathbb{R}^d \times U$.

Under Assumptions 1 and 2, and for a sufficiently large discount rate β, the functional (8) is finite for every feedback controller u. We denote the corresponding value function by

$$V^u(x_0, t_0) := \mathbb{E}_u\left[\int_{t_0}^\infty e^{-\beta(t-t_0)} r(x(t), u(x(t)))\, dt \,\middle|\, x(t_0) = x_0\right]. \tag{10}$$

By an application of the Feynman-Kac formula, the value function V^u satisfies the linear HJB-type equation

$$\partial_t V^u(x,t) = r(x,u(x)) + \mu(x,u(x)) \cdot \nabla V^u(x,t) + \frac{1}{2}\operatorname{Tr}\left(\sigma(x,u(x))\sigma(x,u(x))^T \nabla^2 V^u(x,t)\right). \tag{11}$$

Owing to the infinite–horizon formulation and discounting, the value function is time-independent, and (11) reduces to the stationary elliptic equation

$$\beta V^u(x) = r(x,u(x)) + \mu(x,u(x)) \cdot \nabla V^u(x) + \frac{1}{2}\operatorname{Tr}\left(\sigma(x,u(x))\sigma(x,u(x))^T \nabla^2 V^u(x)\right). \tag{12}$$

This equation falls within the class of elliptic partial differential equations for which strong regularity theory is available under suitable assumptions on the diffusion coefficient. In particular, when the noise is uniformly elliptic, the value function admits classical and Sobolev regularity. In the remainder of the paper, we exploit this regularity together with the relationship between Sobolev spaces and Matérn reproducing kernel Hilbert spaces to derive sharp conditions under which value functions can be efficiently approximated using kernel ridge regression.

Remark 1 *When the HJB equation cannot be written in elliptic form—for instance, in the absence of infinite-horizon, or time-independent dynamics and rewards—many of the regularity results employed here have analogous statements within the framework of parabolic PDE theory, see for example chapter 7.1 of Evans (2010). To keep our exposition simple, we restrict attention to the time-independent elliptic case.*

3. Elliptic PDE Theory

We have identified that the value function satisfies the elliptic partial differential equation given by the stationary HJB Equation (12). To study the regularity properties of its solutions, we briefly recall the necessary elements of classical PDE theory. In particular, we introduce the notions of weak derivatives and Sobolev spaces. Our exposition is intentionally concise and restricted to the concepts required for the subsequent analysis; we refer the interested reader to Evans (2010); Brezis (2011) for a comprehensive treatment.

3.1. Sobolev Spaces

Let $k \in \mathbb{N}^+$ and let $\alpha = (\alpha_1, \dots, \alpha_d)$ be a multi-index satisfying $|\alpha| = \alpha_1 + \cdots + \alpha_d \leq k$. For any function $f \in C^k(\mathbb{R}^d)$, we denote by $D^\alpha f$ the classical partial derivative of order $|\alpha|$, defined as

$$D^\alpha f \;=\; \frac{\partial^{|\alpha|}}{\partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \dots \partial_{x_d}^{\alpha_d}} f. \tag{13}$$

While classical derivatives suffice for smooth solutions, many partial differential equations do not admit solutions possessing all derivatives appearing in the equation. This motivates the use of *weak derivatives*, which extend differentiation to nonsmooth functions via integration against test functions and coincide with classical derivatives whenever the latter exist.

More precisely, let $u, v \in L^1_{loc}(\Omega)$ and let α be a multi-index. We say that v is the α-th weak partial derivative of u, and write $D^\alpha u = v$, if

$$\int_\Omega u \, D^\alpha \phi \, dx \;=\; (-1)^{|\alpha|} \int_\Omega v \, \phi \, dx, \tag{14}$$

for all compactly supported test functions $\phi \in C_c^\infty(\Omega)$. Using weak derivatives, we define the Sobolev space of order k on an open set $\Omega \subseteq \mathbb{R}^d$ by

$$H^k(\Omega) := \left\{ v \in L^2(\Omega) \mid D^\alpha v \in L^2(\Omega) \text{ for all } |\alpha| \leq k \right\}. \tag{15}$$

Equipped with the inner product

$$\langle f, g \rangle_{H^k(\Omega)} := \sum_{|\alpha| \leq k} \langle D^\alpha f, D^\alpha g \rangle_{L^2(\Omega)}, \tag{16}$$

where $D^0 f = f$, the space $H^k(\Omega)$ is a Hilbert space. We further define

$$H_0^k(\Omega) := \left\{ v \in H^k(\Omega) : D^\alpha v|_{\partial\Omega} = 0 \text{ for all } |\alpha| < k \right\}, \tag{17}$$

where boundary values are understood in the sense of the trace operator. On $H_0^1(\Omega)$, several equivalent inner products are commonly used; for instance, one may drop the zeroth-order term and work solely with $\langle \nabla f, \nabla g \rangle_{L^2(\Omega)}$, or equivalently employ a Fourier-based characterization.

We adopt the following Fourier transform conventions. For $f \in L^1(\mathbb{R}^d)$, the Fourier transform and inverse Fourier transform are defined by

$$\mathcal{F}[f](\xi) := \int_{\mathbb{R}^d} f(x) \, e^{-i2\pi x \cdot \xi} \, dx, \qquad \mathcal{F}^{-1}[f](x) := \int_{\mathbb{R}^d} f(\xi) \, e^{i2\pi x \cdot \xi} \, d\xi. \tag{18}$$

These definitions extend to $L^2(\mathbb{R}^d)$ in the usual way with Plancherel's theorem.

On the full space $\mathbb{R}^n$, Sobolev spaces admit an equivalent characterization in Fourier space. For $k \in \mathbb{N}^+$, we have

$$H^k(\mathbb{R}^n) := \left\{ f \in L^2(\mathbb{R}^n) \ : \int_{\mathbb{R}^n} \left(1 + \|\xi\|_2^2\right)^k \left|\mathcal{F}\left[f\right](\xi)\right|^2 d\xi < \infty \right\}, \tag{19}$$

with inner product

$$\langle f, g \rangle_{H^k(\mathbb{R}^n)} := \int_{\mathbb{R}^n} \left(1 + \|\xi\|_2^2\right)^k \mathcal{F}\left[f\right](\xi) \, \overline{\mathcal{F}\left[g\right]}(\xi) \, d\xi. \tag{20}$$

Equivalently, this inner product can be written in physical space as

$$\langle f, g \rangle_{H^k(\mathbb{R}^n)} = \sum_{|\alpha| \leq k} \langle D^\alpha f, D^\alpha g \rangle_{L^2(\mathbb{R}^n)}. \tag{21}$$

Finally, we will also consider functions that are Sobolev-regular only locally. For an open set $\Omega \subseteq \mathbb{R}^n$, we say that $u \in H^k_{loc}(\Omega)$ if $u \in H^k(\omega)$ for every relatively compact subset $\omega \Subset \Omega$. In particular,

$$H^k(\mathbb{R}^d) \subset H^k_{loc}(\mathbb{R}^d). \tag{22}$$

We note that $H^k_{loc}(\Omega)$ is not a Hilbert space.

We will require the boundary of the set Ω to be sufficiently regular, the following notion will be sufficient.

Definition 2 (C^1 Smooth Boundary) *Let $\Omega \subset \mathbb{R}^d$ be open and bounded. We say the boundary $\partial\Omega$ is C^k, and Ω is of class C^k, if for each point $x^0 \in \partial\Omega$ there exists $r > 0$ and $g \in C^k(\mathbb{R}^d)$ such that*

$$\Omega \cap B(x^0, r) = \left\{ x \in B(x^0, r) : x_n > g(x_1, \ldots, x_{d-1}) \right\}. \tag{23}$$

3.2. Reproducing Kernels and Sobolev Spaces

A reproducing kernel Hilbert space is a Hilbert space on which all of the point evaluation functionals are linear bounded operators. These spaces are important in learning theory for their regularity properties, which allow for the derivation of performance guarantees on algorithms, see Sections A and 5 for more details. We now characterize when the Sobolev space $H^k(\mathbb{R}^d)$ admits a reproducing kernel structure.

Proposition 3 *Point evaluation in $H^k(\mathbb{R}^d)$ is continuous if and only if $k > \frac{d}{2}$. In this case, $H^k(\mathbb{R}^d)$ is a reproducing kernel Hilbert space.*

This result follows from the Sobolev embedding theorem and immediately extends to bounded domains via extension operators.

Corollary 4 *Let $\Omega \subset \mathbb{R}^n$ be a nonempty, bounded, and open set with boundary $\partial\Omega \in C^1$. Point evaluation in $H^k(\Omega)$ is continuous if and only if $k > \frac{n}{2}$. In this case, $H^k(\Omega)$ is a reproducing kernel Hilbert space.*

Remark 5 *The conclusion of Corollary 4 extends to smooth, compact manifolds, since analogous Sobolev embedding theorems hold in this setting, see for example Chapter 4.3 of Taylor (2023).*

Having established when $H^k(\mathbb{R}^d)$ is an RKHS, we now identify its reproducing kernel. This can be derived using the Fourier characterization of Sobolev spaces. In particular, one may compute the unique function satisfying the reproducing property by expressing the inner product in the frequency domain. With this characterization, we can identify the kernel of $H^k(\mathbb{R}^d)$ with the Matérn kernel on $\mathbb{R}^d$ with smoothness parameter $\nu = k - d/2$.

Corollary 6 *Let $k > d/2$. The reproducing kernel of $H^k(\mathbb{R}^d)$ is given by*

$$k(x, y) = \mathcal{F}^{-1}\left[\left(1 + \|\xi\|_2^2\right)^{-k}\right](x - y). \tag{24}$$

For $k > d/2$, Proposition 21 implies that for any $\Omega \subset \mathbb{R}^d$, the kernel of $H^k(\mathbb{R}^d)$ restricted to Ω is also a kernel and has the associated RKHS

$$H^k_\Omega := \left\{ f : \Omega \to \mathbb{R} : \exists\, \widehat{f} \in H^k(\mathbb{R}^d) \text{ s.t. } \widehat{f}|_\Omega = f \right\},$$

with norm

$$\|f\|_{H^k_\Omega} := \inf\left\{ \|\widehat{f}\|_{H^k(\mathbb{R}^d)} : \widehat{f} \in H^k(\mathbb{R}^d) \text{ s.t. } \widehat{f}|_\Omega = f \right\}.$$

We note that, although H^k_Ω and $H^k(\Omega)$ are equivalent as sets , they are not always equivalent as Hilbert spaces since their norms may differ. If, however, we assume sufficient regularity on Ω, then the norms on H^k_Ω and $H^k(\Omega)$ are equivalent due to Sobolev Extension Theorem 30 , and so they are isomorphic as Hilbert spaces.

Corollary 7 (Restricted Sobolev RKHS) *Let $k > d/2$ and suppose $\Omega \subset \mathbb{R}^d$ is bounded with C^1 boundary $\partial\Omega$. Then $H^k_\Omega = H^k(\Omega)$ as sets and for all $f \in H^k_\Omega$*

$$c\|f\|_{H^k(\Omega)} \le \|f\|_{H^k_\Omega} \le C\|f\|_{H^k(\Omega)}, \tag{25}$$

where H^k_Ω is the RKHS associated with the restricted kernel $k|_{\Omega\times\Omega}$ for the kernel k of $H^k(\mathbb{R}^d)$.

3.3. Regularity of Elliptic PDEs

We now study conditions under which the value function belongs to a Sobolev space that admits a reproducing kernel representation. To this end, we consider a class of elliptic partial differential equations arising from stationary HJB equations and related control problems.

Let $f \in L^2(\Omega)$ be a given forcing term. We consider the second-order elliptic equation

$$Lv(x) = f(x), \qquad x \in \Omega \subseteq \mathbb{R}^d, \tag{26}$$

where L is a differential operator of the form

$$Lv(x) = \sum_{i,j=1}^{d} a_{ij}(x)\partial_i\partial_j v(x) + \sum_{i=1}^{d} b_i(x)\partial_i v(x) + c(x)v(x), \tag{27}$$

with Lebesgue measurable coefficients $a_{ij}, b_i, c \in L^\infty(\Omega)$.
Throughout, we focus on operators L that are uniformly elliptic, in the sense that there exists $\theta > 0$ such that

$$\theta \|\xi\|_2^2 \leq \sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j, \qquad \text{for all } \xi \in \mathbb{R}^d \text{ and a.e. } x \in \Omega. \tag{28}$$

In this case, we say that Equation (26) is uniformly elliptic. For this class of equations, sharp results are available concerning existence, uniqueness, and regularity of weak solutions. The following classical elliptic regularity theorem quantifies the smoothing effect of uniformly elliptic operators.

Theorem 8 (Elliptic Regularity: Ch 6.3 Theorem 2 Evans (2010)) *Let $k \geq 0$ be an integer and suppose that $a_{ij}, b_j, c \in C^{k+1}(\Omega)$ for $i, j = 1, \ldots, d$ and $f \in H^k(\Omega)$ for an bounded, open set $\Omega \subset \mathbb{R}^d$. Suppose $v \in H^1(\Omega)$ is a weak solution of the elliptic PDE*

$$Lv = f, \qquad x \in \Omega, \tag{29}$$

then $v \in H^{k+2}_{loc}(\Omega)$ and for each compactly contained $\omega \Subset \Omega$ we have

$$\|v\|_{H^{k+2}(\omega)} \leq C\left(\|f\|_{H^k(\Omega)} + \|v\|_{L^2(\Omega)}\right), \tag{30}$$

where C depends only on k, ω, Ω and the coefficients of L.

4. Smoothing Effect of Gaussian Noise

We now combine elliptic regularity theory with Sobolev embedding results to characterize when the value function associated with a fixed policy belongs to a Sobolev space that admits a reproducing kernel representation. This yields sharp sufficient conditions under which continuous-time value functions lie in Sobolev RKHSs.

Assumption 3 (Uniformly elliptic noise) *The diffusion matrix $\sigma(x,u)\,\sigma(x,u)^T$ satisfies the uniform ellipticity condition Equation (28).*

Under these assumptions, the HJB operator associated with a fixed policy is uniformly elliptic and so the value function gains additional regularity over the running reward.

Theorem 9 (Regular policies induce Matérn RKHS value functions) *Let $\omega \Subset \Omega \subset \mathbb{R}^d$ be compactly contained and assume $\partial\omega$ is C^1. Suppose that a stochastic control problem on Ω satisfies Assumption 3, and let u be a policy such that*

$$\mu(\cdot, u(\cdot)), \sigma(\cdot, u(\cdot)) \in C^{k-1}(\Omega), \qquad r(\cdot, u(\cdot)) \in H^{k-2}(\Omega), \tag{31}$$

for some $k > d/2$. Then the corresponding value function $V^{u,\omega}$ solving the HJB equation Equation (12) on Ω satisfies

$$V^{u,\omega} \in H^k(\omega), \tag{32}$$

where $H^k(\omega)$ is an RKHS. Moreover, $H^k(\omega)$ is the RKHS associated with a Matérn kernel of smoothness parameter $\nu = k - d/2$, restricted to $\omega \times \omega$.

Proof Under the stated assumptions, the operator

$$L^u v := \mu(\cdot, u(\cdot)) \cdot \nabla v + \frac{1}{2} \operatorname{Tr}\left(\sigma\left(\cdot, u(\cdot)\right) \sigma\left(\cdot, u(\cdot)\right)^T \nabla^2 v\right) - \beta v,$$

associated with u is uniformly elliptic with coefficients in $C^{k-1}(\Omega)$. Since, we also have $r(\cdot, u(\cdot)) \in H^{k-2}(\Omega)$, Theorem 8 implies

$$V^u \in H^k_{loc}(\Omega). \tag{33}$$

In particular, we have $V^u \in H^k(\omega)$, and by Corollary 7, the space $H^k(\omega)$ identifiable with the desired RKHS. ■

The restriction to bounded domains is primarily technical. If instead $\Omega = \mathbb{R}^d$, it is possible to get $V^u \in H^k(\omega)$ for every compact $\omega \subset \mathbb{R}^d$, that is, $V^u \in H^k_{loc}(\mathbb{R}^d)$. The main obstruction to obtaining $V^u \in H^k(\mathbb{R}^d)$ globally is the lack of control over the $L^2(\mathbb{R}^d)$ component of the Sobolev norm. Indeed, the estimate in Theorem 8 depends explicitly on $\|V^u\|_{L^2(\Omega)}$, which cannot be bounded without additional assumptions. In practice, this issue can be resolved by imposing suitable growth or decay conditions at infinity, or by deriving a priori L^2 estimates from stability or discounting properties of the control problem. In any case, our primary focus is on RKHSs defined over compact domains, where we can apply Mercer's Theorem and the spectral characterizations of the associated kernel integral operators (see below for details).

Remark 10 *The characterization in Theorem 9 is sharp. The regularity guarantees in Theorem 8 are optimal, and relaxing any of the assumptions on ellipticity, coefficient smoothness, or reward regularity allows the construction of examples for which the associated value function fails to possess k weak derivatives. In such cases, the RKHS condition in Proposition 3 is violated, and the value function does not belong to the Matérn RKHS. Consequently, each assumption in Theorem 9 is essential for guaranteeing Sobolev RKHS regularity.*

5. TAR: a policy evaluation algorithm

Given the above result, we can take any value function for an SDE control problem and policy satisfying the above assumptions to be an element of the RKHS $\left(H^k_\Omega, \langle\cdot,\cdot\rangle_{H^k(\mathbb{R}^d)}\right)$ by extending it to an element of $H^k(\mathbb{R}^d)$ then considering its restriction to Ω.

5.1. Continuous-time Markov reward processes

When an controlled SDE is paired with a fixed policy, it induces a Markov reward process (MRP) (Szepesvári, 2010). In short, after a policy π is fixed there is no need of notation for actions, and the system can be viewed as an evolution of state with a reward function. Naturally, controlled SDE induce continuous-time MRPs.[1]

Definition 11 (Markov Reward Process (MRP)) *An MRP consists of a state space $\mathcal{X} = \mathbb{R}^d$, an SDE (μ, σ, ρ) over $\mathcal{X}$, and a measurable, bounded reward function $r : \mathcal{X} \to [-R, R]$.*

1. Not to be confused with "continuous-time Markov chains", which are markov chains with discrete state transitions, but where the transition "times" are separated by an exponential random variable.

The reward function r is a local measure of instantaneous reward, but the real quantity of interest in an MRP is the value function, $v : \mathcal{X} \to \mathbb{R}$, which captures a notion of long term performance:

$$v(s) = \mathbb{E}\left[\int_0^\infty \exp(-\beta t) r(s_t) dt \middle| s_0 = s\right], \tag{34}$$

where $\beta \in \mathbb{R}^{++}$, and $(s_t)_{t\in\mathbb{R}^+}$ is a trajectory of the MRP started at s.

Our aim is to estimate v accurately over the part of the state space that is likely to be reached by a trajectory up until a fixed time $T \in \mathbb{R}^+$, i.e. to produce an estimate $\hat{v}$ such that $|\hat{v}(s_t) - v(s_t)|$ is small for $t \in [0, T]$ with high probability. We assume that μ, σ, ρ, and r are unknown. However, we assume access to samples from the starting state distribution ρ, as well as knowledge of the reward bound R and discount rate β. Moreover, we will be able to access the dynamics of the MRP through a continuous-time simulator:

Assumption 4 (Continuous time local access simulator) *We assume access to a simulator* Sim *that, given a stopping rule $\varphi : \mathbb{R}^+ \times \mathcal{X} \to \{0, 1\}$, produces sampled trajectories $(s_t)_{t\in[0,\tau]}$ of the MRP over the time interval $[0, \tau]$, where $\tau = \min\{t \in \mathbb{R}^+ : \varphi(t, s_t) = 1\}$ is the stopping time associated with φ.*

A call to the simulator will be expressed as $(s_t)_{t\in[0,\tau]} \sim \text{Sim}(\varphi)$.

To make this value estimation problem tractable, we will need to limit our attention to the states with a bounded 2-norm, i.e. to correctly estimating v on the states in $\mathcal{B}(\mathcal{X}, B)$ which are likely to be reached, where $B \in \mathbb{R}^{++}$. Naturally, we will assume that the choice of B is sufficiently large to contain the initial state.

Assumption 5 *There is a known constant $B \in \mathbb{R}^{++}$ such that* $\operatorname{supp} \rho \subset \mathcal{B}(\mathcal{X}, B)$.

We will also make the critical assumption from Section 4, Assumption 3, that the noise is regular in the underlying controlled SDE. By Theorem 9 this implies that $v \in \mathcal{H}$, the reproducing kernel Hilbert space induced by kernel k, where k is the Matérn kernel with parameter $\nu > 0$. We will refer to $k, \mathcal{H}$, and ν throughout this section.

With the assumptions out of the way and the stage set, we will state our algorithm for policy evaluation in MRPs.

5.2. TAR

Algorithm 3 contains our algorithm, Transform And Rollout (TAR). Its execution consists of two phases, which are separated out into subroutines, Algorithms 1 and 2. First, TAR calls Rollout (Algorithm 1) to sample trajectories, or rollouts, from the dynamics while collecting a "core list" of states. Every time a state is discovered that is not in the space spanned by the core list, in an appropriate sense depending on k, that state is added to the core list and the procedure is restarted. In the second phase, TAR calls Estimate (Algorithm 2), which collects Monte Carlo estimates of v at all the states in the core list, and uses these estimates to construct a kernel ridge regression estimate of v (see Section A.1). In this section we give a performance guarantee for TAR, but details are exported to the proofs of propositions in the appendix.

Algorithm 1 Rollout($\mathcal{C}, n, \lambda, T, \epsilon, B$)

inputs: core list $\mathcal{C}$, trial count $n \in \mathbb{N}_+$, regularisation parameter $\lambda \in \mathbb{R}^+$, coverage horizon $T \in \mathbb{R}^{++}$, surprise threshold $\epsilon \in \mathbb{R}^{++}$, state bound $B \in \mathbb{R}^{++}$.
$K_{\mathcal{C}} := [k(s, s')]_{s,s' \in \mathcal{C}}$ // Gram matrix
$k_{\mathcal{C}} := s \mapsto (k(s, s'))_{s' \in \mathcal{C}}$ // Kernel evaluation vector
$\varphi := (t, s) \mapsto \mathbb{I}\left[t \geq T \text{ or } \left(k(s, s) - k_{\mathcal{C}}(s)^\top (K_{\mathcal{C}} + \lambda I_{|\mathcal{C}|})^{-1} k_{\mathcal{C}}(s) \geq \epsilon \text{ and } \|s\|_2 \leq B\right)\right]$ // Stopping rule for rollouts
for $i \in [n]$ **do** // Sample n rollouts to check for surprises
    $(s_t)_{t \in [0,\tau]} \sim \text{Sim}(\varphi)$ // Call the simulator
    if $\tau < T$ **then return** s_τ appended to $\mathcal{C}$ // Found a "surprising" state, core list grows
return $\mathcal{C}$

Algorithm 2 Estimate($\mathcal{C}, \lambda, m, T$)

inputs: core list $\mathcal{C}$, regularisation parameter $\lambda \in \mathbb{R}^+$, rollout count $m \in \mathbb{N}_+$, and horizon $T \in \mathbb{R}^{++}$.
$\mathcal{V} := [\,]$ // Regression target list
$\varphi := (t, s) \mapsto \mathbb{I}[t \geq T]$ // Stopping rule for value estimates
for $s \in \mathcal{C}$ **do** // Estimate the value of states in the core list
    $\bar{v} := 0$ // Monte Carlo estimate of $v(s)$
    for $i \in [m]$ **do**
        $(s_t)_{t \in [0,T]} \sim \text{Sim}(\varphi)$ // Call the simulator
        $\bar{v} := \bar{v} + \frac{1}{m} \int_0^T e^{-\beta t} r(s_t) dt$ // Update estimate
    Append $\bar{v}$ to $\mathcal{V}$ // Store target
$K_{\mathcal{C}} := [k(s, s')]_{s,s' \in \mathcal{C}}$ // Gram matrix
$k_{\mathcal{C}} := s \mapsto (k(s, s'))_{s' \in \mathcal{C}}$ // Kernel evaluation vector
return $s \mapsto k_{\mathcal{C}}(s)(K_{\mathcal{C}} + \lambda I_{|\mathcal{C}|})^{-1} \mathcal{V}$ // Kernel regression solution

5.3. Performance guarantee for TAR

The following is the statement of our main result for the TAR algorithm. It guarantees that, with high probability, TAR outputs an estimate of the value function which is accurate on all the states in $\mathcal{B}(\mathcal{X}, B)$ that are likely to be seen up to time T. This kind of guarantee could serve as the basis for policy improvement algorithms, which we highlight as an interesting area for future work.

Theorem 12 (Performance guarantee for TAR)

If TAR is executed with the following inputs: value estimate rollout count m, discount rate β; regularisation parameter λ; horizon T; surprise threshold ϵ; state bound B; failure probability δ_{fail}; and leaving probability δ_{leave}; then, with probability $1 - \delta_{fail}$, it outputs a value function estimate $\hat{v}$ such that, with probability $1 - \delta_{leave}$, for all $t \in [0, T]$, if $\|s_t\| \leq B$ then

$$|\hat{v}(s_t) - v(s_t)| \leq \sqrt{\epsilon} \left(\sqrt{\lambda} \|v\|_{\mathcal{H}} + \frac{3R}{\beta} \sqrt{\frac{\eta \log(4\eta/\delta_{fail})}{m}} \right), \tag{35}$$

Algorithm 3 Transform And Rollout (TAR)

inputs: value estimate rollout count $m \in \mathbb{N}_+$, discount rate $\beta \in \mathbb{R}^{++}$, regularisation parameter $\lambda \in \mathbb{R}^+$, horizon $T \in \mathbb{R}^{++}$, surprise threshold $\epsilon \in \mathbb{R}^{++}$, state bound $B \in \mathbb{R}^{++}$, failure probability $\delta_{\text{fail}} \in (0,1)$, and leaving probability $\delta_{\text{leave}} \in (0,1)$.
$s \sim \rho$ // Sample a state to initialise the core list
$\mathcal{C} := [s]$ // Initialise the core list
for $i \in \mathbb{N}$ **do** // Sample rollouts to check for surprises
    $\mathcal{C}' :=$ Rollout($\mathcal{C},\ \lceil 2\log(2/\delta_{\text{fail}}\delta_{\text{leave}})/\delta_{\text{fail}} \rceil + i,\ \lambda,\ T,\ \epsilon,\ B$)
    if $\mathcal{C}' = \mathcal{C}$ **then return** Estimate($\mathcal{C},\ \lambda,\ m,\ \log(m)/\beta$) // No surprises during rollouts
    else $\mathcal{C} := \mathcal{C}'$

where $(s_t)_{t\in[0,T]}$ is a trajectory starting at $s_0 \sim \rho$ and

$$\eta = 3 + \frac{C^{2d+d/\nu} B^{2d}}{\log^{(2\nu+d)/\nu}(1+\epsilon/\lambda)} \tag{36}$$

for an absolute constant C.

The following definition is important for the proof of Theorem 12.

Definition 13 (Good set) *The good set $\mathcal{G}_\mathcal{C} \subset \mathcal{X}$ corresponding to core list $\mathcal{C}$ is defined as*

$$\mathcal{G}_\mathcal{C} = \left\{ s \in \mathcal{B}(\mathcal{X}, B) \ :\ k(s,s) - k_\mathcal{C}(s)^\top (K_\mathcal{C} + \lambda I_{|\mathcal{C}|})^{-1} k_\mathcal{C}(s) < \epsilon \right\}. \tag{37}$$

The proof of Theorem 12 can be broken down into the following three propositions. We state these propositions here and then string them together for the proof; each proposition is proved in its own section of the appendix.

Proposition 14 (Runtime bound) *There exists an absolute constant $C \in \mathbb{R}^{++}$ such that, during the execution of TAR, the size of the core list $\mathcal{C}$ is bounded by*

$$|\mathcal{C}| \le 3 + \frac{C^{2d+d/\nu} B^{2d}}{\log^{(2\nu+d)/\nu}(1+\epsilon/\lambda)}\,. \tag{38}$$

Proposition 15 (Bounded error) *Let $\mathcal{C}$ be the core list at the time TAR terminates, and $\hat{v}$ be the value function estimate that TAR returns. Then, with probability $1 - \delta_{fail}/2$, for all $s \in \mathcal{G}_\mathcal{C}$,*

$$\left|\hat{v}(s) - v(s)\right| \le \sqrt{\epsilon}\left(\sqrt{\lambda}\|v\|_\mathcal{H} + \frac{3R}{\beta}\sqrt{\frac{|\mathcal{C}|\log(4|\mathcal{C}|/\delta_{fail})}{m}} \right). \tag{39}$$

Proposition 16 (Coverage) *With probability $1 - \delta_{fail}/2$ over the randomness in the execution of TAR, if $(s_t)_{t\in[0,T]}$ is a trajectory starting at $s_0 \sim \rho$,*

$$P\big(\exists t \in [0,T] \ :\ s_t \in \mathcal{B}(\mathcal{X}, B) \setminus \mathcal{G}_\mathcal{C}\big) \le \delta_{leave}\,, \tag{40}$$

where $\mathcal{G}_\mathcal{C}$ is the good set at the time of termination of TAR.

Be careful when interpreting the above proposition: the $1 - \delta_{\text{fail}}/2$ event is over the randomness in the execution of TAR, and the probability in Equation (40) is over the randomness in $(s_t)_{t\in[0,T]}$, and independently sampled trajectory.

With the lemmata out of the way, we're ready to prove Theorem 12.

Proof [Proof of Theorem 12] Proposition 14 gives us that $|\mathcal{C}| \leq \eta$. Combining this with an application of Proposition 15 produces that, for all $s \in \mathcal{G}_{\mathcal{C}}$,

$$|\hat{v}(s) - v(s)| \leq \epsilon \left(\sqrt{\lambda}\|v\|_{\mathcal{H}} + \frac{3R}{\beta}\sqrt{\frac{\eta \log(4\eta/\delta)}{m}} \right), \tag{41}$$

with probability $1 - \delta_{\text{fail}}/2$. By Proposition 16, with probability $1 - \delta_{\text{fail}}/2$ over the randomness in the execution of TAR, the good set at the end of execution contains all the states in $\mathcal{B}(\mathcal{X}, B)$ that will be visited by a fresh trajectory, with probability δ_{leave} over the randomness in the new trajectory.

Taking a union bound over the two bad events, i.e. the events that the bounds of Propositions 15 and 16 fail to materialize, which each occur with probability at most $\delta_{\text{fail}}/2$, gives the result. ■

6. Conclusions

In this work, we investigated continuous–time stochastic control problems governed by controlled stochastic differential equations in the discounted infinite-horizon setting. Focusing on feedback controllers and the associated policywise Hamilton-Jacobi-Bellman equations, we analysed the regularity properties of value functions arising from fixed policies.

Our main contribution was to show that, under mild growth and regularity assumptions on the controlled dynamics and a uniform ellipticity condition on the diffusion, policywise value functions exhibit a regularity gain relative to the running reward. Using elliptic partial differential equation theory, we established that these value functions belong to Matérn reproducing kernel Hilbert spaces of higher smoothness. This characterization provides a principled functional-analytic foundation for approximating value functions in continuous-time stochastic control.

Leveraging this regularity, we proposed a kernel-based policy evaluation method that estimates value functions directly from online trajectory rollouts of a fixedfeedback controller. By exploiting the RKHS structure induced by the elliptic HJB operator, the resulting algorithm reduces the infinite-dimensional value function estimation problem to a finite-dimensional kernel ridge regression problem, thereby avoiding explicit discretization of the state space and mitigating the curse of dimensionality.

Several directions for future work naturally arise from this framework. One important extension is to integrate the proposed policy evaluation method into full policy improvement and policy iteration schemes, yielding end-to-end continuous-time reinforcement learning algorithms with theoretical guarantees. It would also be of interest to extend the analysis to time-dependent and parabolic HJB equations, as well as to controlled diffusion processes evolving in constrained state spaces. Together, these directions suggest a fruitful interface between stochastic control, partial differential equation theory, and learning-based methods for continuous-time decision making.

References

Robert A. Adams and Joh J. F. Fournier. *Sobolev Spaces*. Academic Press, 2 edition, 2003.

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 *International Conference on Machine Learning (ICML)*, pages 2291–2320. PMLR, 2024. doi: 10.48550/arXiv.2403.05385.

Alireza Bakhtiari, Alex Ayoub, Samuel Robertson, David Janz, and Csaba Szepesvári. Eluder dimension: localise it! In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net.

Dennis S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, second edition, 2011.

Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2011.

Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, 2010.

Dylan J. Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*, 2023. URL https://arxiv.org.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 1963.

Yanwei Jia and Xun Yu Zhou. q-learning in continuous time, 2025. URL https://arxiv.org/abs/2207.00713.

Olav Kallenberg. *Foundations of Modern Probability*. Springer, 3 edition, 2021.

Wenlong Mou. Is RL fine-tuning harder than regression? a PDE learning approach for diffusion models, 2025. URL https://arxiv.org/abs/2509.02528.

Samuel Robertson, Thang Chu, Bo Dai, Dale Schuurmans, Csaba Szepesvári, and Jincheng Mei. Reinforce converges to optimal policies with any learning rate. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://openreview.net.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer, 2008.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Morgan and Claypool Publishers, 2010.

Michael E. Taylor. *Partial Differential Equations I*. Springer, 3 edition, 2023.

Sattar Vakili and Julia Olkhovskaya. Kernelized reinforcement learning with order optimal regret bounds. In *Thirty-sixth Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=pirH9ycaNg.

Dong Yin, Botao Hao, Yasin Abbasi-Yadkori, Nevena Lazic, and Csaba Szepesvári. Efficient local planning with linear function approximation. 2021.

Jiongmin Yong and Xun Yu Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*. Applications of mathematics. Springer Science & Business Media, 1999.

Yuhua Zhu. PhiBE: A PDE-based bellman equation for continuous time policy evaluation, 2024. URL https://arxiv.org/abs/2405.12535.

Yuhua Zhu, Yuming Zhang, and Haoyu Zhang. Optimal-PhiBE: A PDE-based model-free framework for continuous-time reinforcement learning, 2025. URL https://arxiv.org/abs/2506.05208.

Appendix A. Reproducing Kernel Hilbert Spaces

We start with some standard definitions and results about reproducing kernel Hilbert spaces.

Definition 17 (Kernel (Steinwart and Christmann (2008))) *Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel on $\mathcal{X}$ if there exists a real Hilbert space $\mathcal{H}^k$ and a map $\Phi : \mathcal{X} \to \mathcal{H}^k$ such that, for all $x, y \in \mathcal{X}$,*

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}^k} . \tag{42}$$

We refer to Φ as the feature map and $\mathcal{H}^k$ as the feature space of k.

The following standard results will be useful.

Lemma 18 (Restriction of a kernel (Lemma 4.3 of Steinwart and Christmann (2008))) *Let k be a kernel on $\mathcal{X}$, Ω a nonempty set, and $A : \Omega \to \mathcal{X}$. Then the function $\tilde{k} : \Omega \times \Omega \to \mathbb{R}$ defined by $\tilde{k}(x, y) := k(A(x), A(y))$ is a kernel on Ω. In particular, if $\Omega \subset \mathcal{X}$ then $k|_{\Omega \times \Omega}$ is a kernel on Ω.*

Lemma 19 (Theorem 4.16 of Steinwart and Christmann (2008)) *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive semidefinite.*

We are particularly interested in kernels with the *reproducing property*:

Definition 20 (Reproducing Kernel) *Let $\mathcal{X}$ be a nonempty set and $\mathcal{H}$ be a real Hilbert space of functions defined over $\mathcal{X}$. We say that a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel if $k(\,\cdot\,, x) \in \mathcal{H}$ for all $x \in \mathcal{X}$ and the reproducing property,*

$$f(x) = \langle f, k(\,\cdot\,, x) \rangle_{\mathcal{H}} , \tag{43}$$

holds for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$. Conversely, we say that a Hilbert space $\mathcal{H}$ is reproducing kernel Hilbert space (RKHS) over $\mathcal{X}$ if, for all $x \in \mathcal{X}$, the point evaluation functional $E_x \in \mathcal{H}^*$ defined by

$$E_x f := f(x) \tag{44}$$

is bounded and hence continuous.

It is routine to show that every RKHS has a unique reproducing kernel given by the Reisz representation of the point evaluation functionals Steinwart and Christmann (2008). Reproducing kernels are in fact kernels, and conversely, every kernel is the unique reproducing kernel of some RKHS Steinwart and Christmann (2008). Importantly, Lemma 18 implies that the restricted kernel of an RKHS also induces an RKHS, which is related to the original RKHS as follows (Steinwart and Christmann, 2008, pg. 163).

Proposition 21 *Let k be a kernel on $\mathcal{X}$ with RKHS $\mathcal{H}$. For any $\Omega \subset \mathcal{X}$, the RKHS of the restricted kernel $k|_{\Omega\times\Omega}$ is*

$$\mathcal{H}_\Omega := \left\{ f : \Omega \to \mathbb{R} : \exists\, \widehat{f} \in \mathcal{H} \text{ s.t. } \widehat{f}|_\Omega = f \right\},$$

with norm

$$\|f\|_{\mathcal{H}_\Omega} := \inf \left\{ \|\widehat{f}\|_{\mathcal{H}} : \widehat{f} \in \mathcal{H} \text{ s.t. } \widehat{f}|_\Omega = f \right\}, \tag{45}$$

Mercer's theorem is an important result from RKHS theory that gives an eigendecomposition for reproducing kernels supported on a compact metric space.

Theorem 22 (Mercer's Theorem (Theorem 4.49 of Steinwart and Christmann (2008))) *Let $\mathcal{X}$ be a compact metric space and μ be a finite Borel measure with $\operatorname{supp}\mu = \mathcal{X}$. Supposing that $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a continuous kernel, there exists an orthonormal system $(\phi_i)_{i\in\mathbb{N}_+} \sqsubset \mathcal{H}^k$ and a sequence $(\lambda_i)_{i\in\mathbb{N}_+} \sqsubset \mathbb{R}$ such that, for all $x, y \in \mathcal{X}$,*

$$k(x,y) = \sum_{i\in\mathbb{N}_+} \lambda_i \phi_i(x)\phi_i(y)\,, \tag{46}$$

and the above the series converges absolutely and uniformly.

In applications of Mercer's theorem we will refer to $(\phi_i)_{i\in\mathbb{N}_+}$ and $(\lambda_i)_{i\in\mathbb{N}_+}$ as the eigenfeatures and eigenvalues of k, respectively.

A.1. Kernel Ridge Regression

We are primarily interested in reproducing kernels for their utility in the estimation of target functions from noisy data. Suppose we are given a set of data points $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X}\times\mathbb{R}$, and suppose that $\mathcal{H}^k$ is a RKHS over $\mathcal{X}$ with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For some given regularization parameter $\lambda > 0$, our goal is to find a function $f \in \mathcal{H}^k$ that solves the empirical risk minimization

$$\hat{f} := \operatorname*{argmin}_{f\in\mathcal{H}^k} \sum_{i=1}^n \left(f(x_i) - y_i\right)^2 + \lambda \|f\|_{\mathcal{H}^k}. \tag{47}$$

The representer theorem (Theorem 5.5 Steinwart and Christmann (2008)) tells us that the unique solutions to this optimization exists and is of the form

$$\hat{f}(x) = \sum_{i=1}^n \alpha_i k(x, x_i), \tag{48}$$

for scalar coefficients $\alpha_1, \ldots, \alpha_n$. We can find these coefficients by collecting them into the vector $\alpha \in \mathbb{R}^n$, defining the kernel Gram matrix $K_{ij} = k(x_i, x_j)$, and then solving the linear system

$$(K + \lambda I_n)\,\alpha = y, \tag{49}$$

where $y \in \mathbb{R}^n$ is a vector of the collected data points and $I_n \in \mathbb{R}^{n\times n}$ the n-dimensional identity matrix. For any $\lambda > 0$, Lemma 19 tells us that the matrix K is positive semidefinite, so the matrix $K + \lambda I_n$ is positive definite and hence invertible.

In general we assume that the data points are noisy. To model this we assume there is some ground truth function $f_\star \in \mathcal{H}^k$ and then our data points satisfy $y_i = f_\star(x_i) + \epsilon_i$ for some noise term $\epsilon_i \in \mathbb{R}$, which we collect into the error vector $\epsilon \in \mathbb{R}^n$. With this setup, we want to be able to bound the extrapolation error of the kernel ridge regression solution to points within our domain not seen in the data set.

Proposition 23 (Extrapolation Error Bounds) *Let $\mathcal{H}$ be a real RKHS with kernel k over $\mathcal{X}$ and $(x_i, y_i)_{i\in[n]} \sqsubset \mathcal{X} \times \mathbb{R}$ be data points generated by $y_i = f_\star(x_i) + \epsilon_i$ for some $f_\star \in \mathcal{H}$ and $\epsilon \in \mathbb{R}^n$. The empirical risk minimizer $\hat{f} \in \mathcal{H}$ given above satisfies*

$$\left|\hat{f}(x) - f_\star(x)\right| \le \sqrt{k(x,x) - K_X(x)^T(K + \lambda I_n)^{-1}K_X(x)}\left(\sqrt{\lambda}\,\|f_\star\|_{\mathcal{H}} + \|\epsilon\|_2\right), \tag{50}$$

for all $x \in \mathcal{X}$, where $K_X(x) := [k(x_i, x)]_{1\le i\le n}$.

Proof We start by defining the sampling operator $S : \mathcal{H} \to \mathbb{R}^n$ as

$$(Sf)_i := f(x_i) = \langle f, k(x_i, \cdot)\rangle_{\mathcal{H}},$$

with adjoint $S^* : \mathbb{R}^n \to \mathcal{H}$ given by

$$(S^*\alpha)(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot),$$

then define the operator $G := S^*S + \lambda I_{\mathcal{H}}$. Since $\lambda > 0$, G is invertible. Then note $y = Sf_\star + \epsilon$, so we obtain

$$\hat{f} = G^{-1}S^*y = G^{-1}S^*(Sf_\star + \epsilon),$$

and therefore

$$\hat{f} - f_\star = -\lambda G^{-1}f_\star + G^{-1}S^*\epsilon.$$

Fix $x \in \mathcal{X}$ and denote the feature map $\Phi(x) := k(x, \cdot)$. By the reproducing property,

$$\hat{f}(x) - f_\star(x) = \langle \hat{f} - f_\star, \Phi(x)\rangle_{\mathcal{H}}.$$

Applying the triangle inequality yields

$$\left|\hat{f}(x) - f_\star(x)\right| \le \lambda\left|\langle G^{-1}f_\star, \Phi(x)\rangle_{\mathcal{H}}\right| + \left\langle G^{-1}S^*\epsilon, \Phi(x)\right\rangle_{\mathcal{H}}\Big|.$$

For the bias term, applying Cauchy-Schwarz gives

$$\left|\langle G^{-1}f_\star, \Phi(x)\rangle_{\mathcal{H}}\right| = \left|\langle G^{-1/2}f_\star, G^{-1/2}\Phi(x)\rangle_{\mathcal{H}}\right| \le \|f_\star\|_{G^{-1}}\,\|\Phi(x)\|_{G^{-1}},$$

where

$$\|f\|_{G^{-1}} = \langle f, G^{-1} f \rangle_{\mathcal{H}}.$$

Since $G = S^*S + \lambda I \succeq \lambda I_{\mathcal{H}}$, we have $G^{-1} \preceq \lambda^{-1} I_{\mathcal{H}}$ and thus

$$\lambda \left| \langle G^{-1} f_\star, \Phi(x) \rangle_{\mathcal{H}} \right| \leq \sqrt{\lambda} \, \|f_\star\|_{\mathcal{H}} \, \|\Phi(x)\|_{G^{-1}}.$$

For the noise term, using the adjoint and Cauchy-Schwarz in $\mathbb{R}^n$ to get

$$\left| \langle G^{-1} S^* \epsilon, \Phi(x) \rangle_{\mathcal{H}} \right| = \left| \langle \epsilon, S G^{-1} \Phi(x) \rangle_{\mathbb{R}^n} \right| \leq \|\epsilon\|_2 \, \|S G^{-1} \Phi(x)\|_2,$$

then note

$$\|S G^{-1} \Phi(x)\|_2^2 = \langle G^{-1} \Phi(x), S^* S G^{-1} \Phi(x) \rangle_{\mathcal{H}} \leq \langle G^{-1} \Phi(x), G G^{-1} \Phi(x) \rangle_{\mathcal{H}} = \|\Phi(x)\|_{G^{-1}}^2,$$

since $S^*S \preceq G$, and so

$$\left| \langle G^{-1} S^* \epsilon, \Phi(x) \rangle_{\mathcal{H}} \right| \leq \|\epsilon\|_2 \, \|\Phi(x)\|_{G^{-1}}.$$

Combining the above estimates yields

$$\left| \hat{f}(x) - f_\star(x) \right| \leq \|\Phi(x)\|_{G^{-1}} \left(\sqrt{\lambda} \, \|f_\star\|_{\mathcal{H}} + \|\epsilon\|_2 \right).$$

Finally, noting that

$$\|\Phi(x)\|_{G^{-1}}^2 = k(x,x) - K_X(x)^T (K + \lambda I_n)^{-1} K_X(x),$$

to get our bound. ■

A.2. Maximum information gain

In order to control the runtime of our algorithm, we will make use of the concept of maximum information gain from the kernelized bandit literature, which is defined as follows (Srinivas et al., 2009).

Definition 24 (Information gain) *Given a set $\mathcal{X}$, a positive semidefinite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a regularisation parameter $\lambda \in \mathbb{R}^{++}$, and a finite sequence $\mathcal{C} \sqsubset \mathcal{X}$, the information gain of k at $\mathcal{C}$ is defined to be*

$$\gamma_\lambda^k(\mathcal{C}) := \frac{1}{2} \log \det \left(I_{|\mathcal{C}|} + \frac{1}{\lambda} K_{\mathcal{C}} \right). \tag{51}$$

Definition 25 (Maximum information gain) *Let $\mathcal{X}$, k, and λ be as in Definition 24, and $\Omega \subset \mathcal{X}$. The maximum information gain of k over Ω is denoted $\Gamma_\lambda^k(\Omega, \cdot) : \mathbb{N}_+ \to \mathbb{R}$, and defined by*

$$\Gamma_\lambda^k(\Omega, n) := \sup_{\mathcal{C} \sqsubset \Omega \,:\, |\mathcal{C}| = n} \gamma_\lambda^k(\mathcal{C}). \tag{52}$$

We will use a result that bounds the maximum information gain on the d-dimensional hypercube with a given side length. Its statement requires the following definition.

Definition 26 (Polynomial eigendecay (Definition 1 of Vakili and Olkhovskaya (2023))) *Let $\mathcal{Z} \subset \mathbb{R}^d$ be a hypercube with side length $B_{\mathcal{Z}}$, and $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a continuous kernel such that $\sup_{z \in \mathcal{Z}} k(z,z) \leq 1$. Let $(\lambda_i)_{i \in \mathbb{N}_+}$ and $(\phi_i)_{i \in \mathbb{N}_+}$ be respectively the eigenvalues and eigenfeatures of k from Mercer's theorem, sorted in decreasing order of eigenvalues. We say that k has polynomial eigendecay if there exist constants $c, \alpha \in \mathbb{R}^{++}$, $p > 1$, and $\eta \in \mathbb{R}^+$, such that the following hold for all $i \in \mathbb{N}_+$:*

$$\lambda_i \leq c i^{-p} B_{\mathcal{Z}}^{\alpha}\,, \tag{53}$$

$$\sup_{z \in \mathcal{Z}} \phi_i(z) \leq c i^{p\eta}\,. \tag{54}$$

With this definition in place, we are equipped to control the maximum information gain of well-behaved kernels.

Proposition 27 (Bound on maximum information gain (Lemma 2 of Vakili and Olkhovskaya (2023))) *Let k be a kernel with polynomial eigendecay, i.e. satisfying Definition 26. Then there exists an absolute constant $C \in \mathbb{R}$ such that the maximum information gain is bounded by*

$$\Gamma_\lambda^k(\mathcal{Z}, n) \leq C n^{1/p(1-2\eta)} \log^{1-1/p(1-2\eta)}(n) B_{\mathcal{Z}}^{\alpha/p(1-2\eta)}\,. \tag{55}$$

Finally, we specialise the above result to the case where k is a Matérn kernel, swapping the hypercube for a ball.

Proposition 28 (Bound on maximum information gain for Matérn kernels) *Let k be the Matérn kernel on $B(\mathbb{R}^d, B)$ with smoothness parameter ν. There exists an absolute constant C such that the maximum information gain satisfies*

$$\Gamma_\lambda^k(B(\mathbb{R}^d, B), n) \leq C n^{d/(2\nu+d)} \log^{2\nu/(2\nu+d)}(n) (2B)^{2\nu d/(2\nu+d)}\,. \tag{56}$$

Proof First we observe that $B(\mathbb{R}^d, B)$ is contained in a hypercube with side length $2B$. As argued by Vakili and Olkhovskaya (2023) , k has polynomial eigendecay with $\eta = 0, p = (2\nu + d)/d$, and $\alpha = 2\nu$. Applying Proposition 27 gives the result. ■

Appendix B. Sobolev Theory

B.1. General Theory

Theorem 29 (General Sobolev Inequalities: Theorem 6 in section 5.6 of Evans (2010)) *Let $\Omega \subset \mathbb{R}^d$ be a bounded open subset with continuous boundary $\partial\Omega \in C^1$. For any $f \in H^k(\Omega)$ if $k > d/2$ then $f \in C^{k - \lfloor d/2 \rfloor - 1, \gamma}(\Omega)$, where*

$$\gamma = \begin{cases} \lfloor \frac{d}{2} \rfloor + 1 - \frac{d}{2} & \text{if } \frac{d}{2} \notin \mathbb{Z}\,, \\ \text{any element of } (0,1) & \text{if } \frac{d}{2} \in \mathbb{Z}\,. \end{cases} \tag{57}$$

Further, we have the bound

$$\|f\|_{C^{k - \lfloor d/2 \rfloor - 1, \gamma}(\Omega)} \leq C \|f\|_{H^k(\Omega)}\,, \tag{58}$$

where C is a constant that depends only on k, d, γ, and Ω.

Theorem 30 (Stein Extension Theorem (Adams and Fournier, 2003, Thm. 6.5.24)) *Let $\Omega \subset \mathbb{R}^d$ be a bounded set with locally Lipchitz boundary $\partial\Omega$. Then for all $k \in \mathbb{N}$ there exists a bounded linear operator $E : H^k(\Omega) \to H^k(\mathbb{R}^d)$ such that, for all $f \in H^k(\Omega)$:*

i) $Ef = f$ a.e. on Ω,

ii) and there is a $C \in \mathbb{R}^{++}$ depending only on k and Ω such that $\|Ef\|_{H^k(\mathbb{R}^d)} \leq C\|f\|_{H^k(\Omega)}$.

Theorem 31 (Sobolev Embedding: Corollary 9.13 of Brezis (2011)) *Let $k \geq 1$ be an integer such that $k > d/2$. Then for every integer*

$$m < k - \frac{d}{2}, \tag{59}$$

there is a continuous embedding

$$H^k(\mathbb{R}^d) \hookrightarrow C^m(\mathbb{R}^d). \tag{60}$$

That is, for every $f \in H^k(\mathbb{R}^d)$ there exists a representative $\bar{f} \in C^m(\mathbb{R}^d)$ such that $f = \bar{f}$ a.e. and

$$\|\bar{f}\|_{C^m(\mathbb{R}^d)} \leq C\|f\|_{H^k(\mathbb{R}^d)}, \tag{61}$$

where $C > 0$ depends only on k and d.

B.2. RKHS Sobolev Theory

Remark 32 *Although we have only defined the Sobolev spaces corresponding to integer numbers of weak derivatives, it is possible to define the space $H^k(\mathbb{R}^d)$ with non-integer k with the Fourier definition. Doing this does not meaningfully change any of the following calculations, beyond some additional care with the embedding theorems, and so they extend to the non-integer k setting.*

Proof [Proof of Proposition 3] Suppose $k > d/2$, then Theorem 29 implies there exists some $C > 0$ such that

$$\sup_{x \in \mathbb{R}^d} |u(x)| \leq C\|u\|_{H^k(\mathbb{R}^d)}, \tag{62}$$

for all $u \in H^k(\mathbb{R}^d)$. Hence for any $x \in \mathbb{R}^d$, the point evaluation functional

$$E_x : H^k(\mathbb{R}^d) \to \mathbb{R}; u \mapsto u(x), \tag{63}$$

is bounded with norm at most C, and so is continuous.

Let $k \leq d/2$ and suppose for contradiction that point evaluation on $H^k(\mathbb{R}^d)$ was an bounded linear functional, then there must exists some finite $0 < C < \infty$ such that for all $f \in H^k(\mathbb{R}^d)$ we have

$$|f(0)| \leq C\|f\|_{H^k(\mathbb{R}^d)} \implies \frac{|f(0)|}{\|f\|_{H^k(\mathbb{R}^d)}} \leq C\,. \tag{64}$$

Now, for any $0 < R$, let

$$1_{B_R}(x) := \begin{cases} 1, & \text{if } \|x\| \leq R, \\ 0, & \text{else.} \end{cases} \tag{65}$$

be the indicator function of the ball of radius R. Then, we define the function

$$\varphi_R(x) := \mathcal{F}^{-1}\left[\left(1 + \|\xi\|_2^2\right)^{-k} 1_{B_R}(\xi)\right](x). \tag{66}$$

Since

$$\left(1+\|\xi\|^2\right)^{-k} 1_{B_R}(\xi) \in L^1(\mathbb{R}^d)\,, \tag{67}$$

by Plancherel's Theorem, we have $\varphi_R \in L^2(\mathbb{R}^d)$. By definition,

$$I(R)^2 := \|\varphi_R\|^2_{H^k(\mathbb{R}^d)} = \int_{\|\xi\|_2 \le R} \left(1+\|\xi\|_2^2\right)^{-k} d\xi\,, \tag{68}$$

which we can note is the same integral we previously considered. Hence assuming $R > 1$ and changing to polar coordinates, we have

$$I(R) \le \left[\frac{1}{d} + \int_1^R r^{d-1-2k}\, dr\right]^{\frac{1}{2}} < \infty\,, \tag{69}$$

is finite for all finite $1 < R < \infty$, which implies that $\varphi_R \in H^k(\mathbb{R}^d)$. Next, by the Fourier inversion theorem and identification with the continuous representative, we have

$$\varphi_R(0) = \int_{\|\xi\| \le R} \left(1+\|\xi\|_2^2\right)^{-k} d\xi = I(R)^2, \tag{70}$$

which implies

$$\frac{|\varphi_R(0)|}{\|\varphi_R\|_{H^k(\mathbb{R}^d)}} = I(R). \tag{71}$$

Then changing to polar coordinates, we can then calculate

$$\begin{aligned} I(R) &= \left|S^{n-1}\right| \int_0^R \frac{r^{d-1}}{\left(1+r^2\right)^k}\, dr && (72)\\ &\ge \left|S^{n-1}\right| \int_1^R \frac{r^{d-1}}{\left(1+r^2\right)^k}\, dr && (73)\\ &\ge \left|S^{n-1}\right| \int_1^R \frac{r^{d-1}}{kr^{2k}}\, dr && (74)\\ &= \left|S^{n-1}\right| k^{-1} \int_1^R r^{d-1-2k}\, dr\,. && (75) \end{aligned}$$

This final integral diverges if $d-1-2k \ge -1$ which is equivalent to $k \le d/2$, as we assumed. This implies that C cannot be finite and so point evaluation cannot be a bounded linear functional on $H^k(\mathbb{R}^d)$. ■

Proof [Proof for Corollary 4] Because the kernel of an RKHS is unique, we need only show that $k(\cdot,\cdot)$ as defined satisfies the reproducing property. With the given kernel definition, we can apply the basic properties of the Fourier transform to $k_x = k(x,\cdot)$ for

$$\mathcal{F}\left[k_x\right](\xi) = \left(1+\|\xi\|_2^2\right)^{-k} e^{-2\pi i x\cdot\xi}\,. \tag{76}$$

We note that $k_x \in H^k(\mathbb{R}^d)$ since

$$\begin{aligned} \|k_x\|_{H^k(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} \left(1+\|\xi\|_2^2\right)^k \left|\mathcal{F}\left[k_x\right](\xi)\right|^2\, d\xi && (77)\\ &= \int_{\mathbb{R}^d} \left(1+\|\xi\|_2^2\right)^{-k}\, d\xi < \infty\,. && (78) \end{aligned}$$

Then, using the Fourier characterization of the inner product on $H^k(\mathbb{R}^d)$, we have that, for any $f \in H^k(\mathbb{R}^d)$,

$$\langle f, k_x \rangle_{H^k(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \left(1 + \|\xi\|_2^2\right)^k \mathcal{F}[f](\xi) \overline{F(k_x)(\xi)}\, d\xi \tag{79}$$

$$= \int_{\mathbb{R}^d} \left(1 + \|\xi\|_2^2\right)^k \mathcal{F}[f](\xi) \overline{\left(1 + \|\xi\|_2^2\right)^{-k} e^{-2\pi i x \cdot \xi}}\, d\xi \tag{80}$$

$$= \int_{\mathbb{R}^d} F(f)(\xi) e^{2\pi i x \cdot \xi}\, d\xi \tag{81}$$

$$= f(x), \tag{82}$$

where we used the Fourier inversion theorem and identification with the continuous representative, which shows that k_x satisfies the reproducing property. ■

Proof [Proof for Corollary 7] Let $v \in H^k_\Omega$ and define

$$A_v := \left\{ u \in H^k(\mathbb{R}^d) \mid u|_\Omega = v \right\}.$$

It is clear that, for all $u \in H^k(\mathbb{R}^d)$,

$$\|u|_\Omega\|_{H^k(\Omega)} \leq \|u\|_{H^k(\mathbb{R}^d)}.$$

hence

$$\|v\|_{H^k(\Omega)} \leq \inf_{u \in A_v} \|u\|_{H^k(\mathbb{R}^d)} = \|v\|_{H^k_\Omega}.$$

Thus $v \in H^k(\Omega)$. Next, let $v \in H^k(\Omega)$. By Theorem 30, there exists an extension operator $E : H^k(\Omega) \to H^k(\mathbb{R}^d)$ such that $(Ev)|_\Omega = v$ and

$$\|Ev\|_{H^k(\mathbb{R}^d)} \leq C \|v\|_{H^k(\Omega)},$$

for some $C > 0$ depending only on k and Ω. Hence

$$\|v\|_{H^k_\Omega} \leq \|Ev\|_{H^k(\mathbb{R}^d)} \leq C \|v\|_{H^k(\Omega)},$$

and $v \in H^k_\Omega$. Moreover, the norms are equivalent. ■

Appendix C. Proof of Proposition 14, the runtime bound

Proposition 14 (Runtime bound) *There exists an absolute constant $C \in \mathbb{R}^{++}$ such that, during the execution of TAR, the size of the core list $\mathcal{C}$ is bounded by*

$$|\mathcal{C}| \leq 3 + \frac{C^{2d+d/\nu} B^{2d}}{\log^{(2\nu+d)/\nu}(1 + \epsilon/\lambda)}. \tag{38}$$

First we will state a useful lemma from linear algebra.

Lemma 33 (Block matrix determinant lemma (Proposition 2.8.3 of Bernstein (2011))) *For a matrix X with block form*

$$X = \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix}, \tag{83}$$

where A_1 and A_4 are square and A_1 is invertible, $\det(X) = \det(A_1)\det(A_4 - A_3 A_1^{-1} A_2)$.

We are ready to prove Proposition 14.

Proof [Proof of Proposition 14] Throughout the proof, let $K_{\mathcal{C}} = [k(s,s')]_{s,s' \in \mathcal{C}}$ denote the Gram matrix associated with the core list $\mathcal{C}$, and $k_{\mathcal{C}}(s) = (k(s,s'))_{s' \in \mathcal{C}}$ denote the kernel evaluation vector. We will consider the determinant of the matrix $\Sigma_{\mathcal{C}} := K_{\mathcal{C}} + \lambda I_{|\mathcal{C}|}$ as $\mathcal{C}$ evolves during the runtime of TAR. More specifically, we will derive lower and upper bounds on $\det(\Sigma_{\mathcal{C}})$ that can only be jointly satisfied for sufficiently small values of $|\mathcal{C}|$.[2] Starting with the lower bound, we will show that

$$\det(\Sigma_{\mathcal{C}}) \geq \lambda^{|\mathcal{C}|}(1 + \epsilon/\lambda)^{|\mathcal{C}|-1} \tag{84}$$

by induction on $|\mathcal{C}|$. For the base case suppose $|\mathcal{C}| = 1$; then $\mathcal{C} = [s]$ for some state s and $\det(\Sigma_{\mathcal{C}}) = \lambda + k(s,s) \geq \lambda$ as desired. Now suppose that $|\mathcal{C}| > 1$. Examining Algorithm 3, we see that $\mathcal{C}$ has been modified most recently by a call to Rollout, where a state s_τ was appended. Let $\mathcal{C}'$ denote the core list before s_τ was appended, so that $\mathcal{C}_i = \mathcal{C}'_i$ for $0 < i < |\mathcal{C}|$ and $\mathcal{C}_{|\mathcal{C}|} = s_\tau$. We make the inductive hypothesis that Equation (84) holds for $\mathcal{C}'$. We have

$$\begin{aligned} \Sigma_{\mathcal{C}} &= K_{\mathcal{C}} + \lambda I_{|\mathcal{C}|} && (85)\\ &= \begin{bmatrix} K_{\mathcal{C}'} + \lambda I_{|\mathcal{C}'|} & k_{\mathcal{C}'}(s_\tau) \\ k_{\mathcal{C}'}(s_\tau)^\top & k(s_\tau, s_\tau) + \lambda \end{bmatrix} && (86)\\ &= \begin{bmatrix} \Sigma_{\mathcal{C}'} & k_{\mathcal{C}'}(s_\tau) \\ k_{\mathcal{C}'}(s_\tau)^\top & k(s_\tau, s_\tau) + \lambda \end{bmatrix}. && (87) \end{aligned}$$

Since $\det(\Sigma_{\mathcal{C}'}) > 0$, $\Sigma_{\mathcal{C}'}$ is invertible. We can apply Lemma 33 to Equation (87) and conclude that

$$\begin{aligned} \det(\Sigma_{\mathcal{C}}) &= \det(\Sigma_{\mathcal{C}'})\big(k(s_\tau, s_\tau) + \lambda - k_{\mathcal{C}'}(s_\tau)^\top \Sigma_{\mathcal{C}'}^{-1} k_{\mathcal{C}'}(s_\tau)\big) && (88)\\ &\geq \lambda^{|\mathcal{C}|-1}(1+\epsilon/\lambda)^{|\mathcal{C}|-2}\left(\lambda + \left(k(s_\tau, s_\tau) - k_{\mathcal{C}'}(s_\tau)^\top \Sigma_{\mathcal{C}'}^{-1} k_{\mathcal{C}'}(s_\tau)\right)\right), && (89) \end{aligned}$$

where the second inequality follows from the inductive hypothesis and $|\mathcal{C}'| = |\mathcal{C}| - 1$. Since s_τ was added to the core list, the condition of the if statement in Rollout was true, so the stopping rule for the simulator must have been triggered at $\tau < T$, implying $k(s_\tau, s_\tau) - k_{\mathcal{C}'}(s_\tau)^\top \Sigma_{\mathcal{C}'}^{-1} k_{\mathcal{C}'}(s_\tau) \geq \epsilon$. Plugging this in to Equation (89) produces Equation (84), concluding the induction.

We will now upper bound $\det(\Sigma_{\mathcal{C}})$. Since $\operatorname{supp}\rho \subset \mathcal{B}(\mathcal{X}, B)$, and, because the stopping rule in Rollout is only triggered early at states in $\mathcal{B}(\mathcal{X}, B)$, we have that $\mathcal{C} \sqsubset \mathcal{B}(\mathcal{X}, B)$. Thus, by inspection of Definitions 24 and 25, we can upper bound the determinant using the maximum information gain of k restricted to $\mathcal{B}(\mathcal{X}, B)$:

$$\det(\Sigma_{\mathcal{C}}) \leq \lambda^{|\mathcal{C}|} \exp\left(2\Gamma_\lambda^k(\mathcal{B}(\mathcal{X}, B),\, |\mathcal{C}|)\right). \tag{90}$$

2. This technique is inspired by the bound on the eluder dimension of linear function classes from Russo and Van Roy (2013).

Proposition 28 generates an absolute constant C such that

$$\Gamma_\lambda^k(\mathcal{B}(\mathcal{X}, B), |\mathcal{C}|) \leq C|\mathcal{C}|^{d/(2\nu+d)} \log^{2\nu/(2\nu+d)}(|\mathcal{C}|)(2B)^{2\nu d/(2\nu+d)} . \tag{91}$$

Combining Equations (90) and (91) and absorbing 2 into the constant C produces

$$\det(\Sigma_\mathcal{C}) \leq \lambda^{|\mathcal{C}|} \exp\left(C|\mathcal{C}|^{d/(2\nu+d)} \log^{2\nu/(2\nu+d)}(|\mathcal{C}|)(2B)^{2\nu d/(2\nu+d)}\right) . \tag{92}$$

We will now combine the upper and lower bounds to control the maximum possible value attainable by $|\mathcal{C}|$: chaining the inequalities of Equations (84) and (92), dividing both sides by $\lambda^{|\mathcal{C}|}$, then taking logarithms yields

$$(|\mathcal{C}| - 1) \log\left(1 + \frac{\epsilon}{\lambda}\right) \leq C|\mathcal{C}|^{d/(2\nu+d)} \log^{2\nu/(2\nu+d)}(|\mathcal{C}|)(2B)^{2\nu d/(2\nu+d)} . \tag{93}$$

If $|\mathcal{C}| < 3$ then Equation (38) is satisfied, so suppose $|\mathcal{C}| \geq 3$. Then $|\mathcal{C}|/2 \leq |\mathcal{C}| - 1$, so we can clean up Equation (93) (again absorbing a factor of 2 into C) to get

$$|\mathcal{C}| \log\left(1 + \frac{\epsilon}{\lambda}\right) \leq C|\mathcal{C}|^{d/(2\nu+d)} \log^{2\nu/(2\nu+d)}(|\mathcal{C}|)(2B)^{2\nu d/(2\nu+d)} . \tag{94}$$

Dividing the above by $|\mathcal{C}|^{d/(2\nu+d)} \log(1 + \epsilon/\lambda)$ before raising to the power $(2\nu + d)/2\nu$, we have

$$|\mathcal{C}| \leq \frac{C^{(2\nu+d)/2\nu}(2B)^d}{\log^{(2\nu+d)/2\nu}(1 + \epsilon/\lambda)} \log(|\mathcal{C}|) . \tag{95}$$

Temporarily setting $\xi := C^{(2\nu+d)/2\nu}(2B)^d / \log^{(2\nu+d)/2\nu}(1 + \epsilon/\lambda)$ for convenience, Equation (95) becomes

$$|\mathcal{C}| \leq \xi \log(|\mathcal{C}|) . \tag{96}$$

Thus we will upper bound the values of $x > 0$ that satisfy $x \leq \xi \log(x)$. This equation could be solved in closed form using the Lambert W function, but we will not lose much by observing that if $x \geq \xi^2$ then $x > \xi \log(x)$: at $x = \xi^2$ the inequality holds; if $\xi \leq 1$ then $x > \xi \log(x)$ is always true, and otherwise $D_x(\xi \log(x)) = \xi/x$ is less than $D_x(x) = 1$ for all $x \geq \xi^2$. Thus $|\mathcal{C}| \leq \xi^2$, whence expanding ξ and combining constants implies Equation (38). ■

Appendix D. Proof of Proposition 15

Proposition 15 (Bounded error) *Let $\mathcal{C}$ be the core list at the time TAR terminates, and $\hat{v}$ be the value function estimate that TAR returns. Then, with probability $1 - \delta_{fail}/2$, for all $s \in \mathcal{G}_\mathcal{C}$,*

$$\left|\hat{v}(s) - v(s)\right| \leq \sqrt{\epsilon}\left(\sqrt{\lambda}\|v\|_\mathcal{H} + \frac{3R}{\beta}\sqrt{\frac{|\mathcal{C}| \log(4|\mathcal{C}|/\delta_{fail})}{m}}\right) . \tag{39}$$

The proof requires a standard concentration inequality, which was proved by Hoeffding (1963).

Proposition 34 (Hoeffding's inequality) *For $n \in \mathbb{N}_+$ and $b \in \mathbb{R}^+$, let $(X_i)_{i\in[n]}$ be sampled i.i.d. from a distribution supported on $[-b, b]$. Then, for any $c \in \mathbb{R}^+$,*

$$P\left(\left|\sum_{i\in[n]} X_i - \mathbb{E}\left[\sum_{i\in[n]} X_i\right]\right| \geq c\right) \leq 2\exp\left(\frac{-c^2}{2nb^2}\right). \tag{97}$$

The proof of Proposition 15 essentially boils down to showing the following result, which leverages Hoeffding's inequality to guarantee that Estimate (Algorithm 2) probably provides good estimates.

Proposition 35 (Monte Carlo estimates work) *Suppose* Estimate$(\mathcal{C}, \lambda, m, T)$ *is called with input core list $\mathcal{C}$, regularization parameter λ, rollout count m, and horizon T. Let $\mathcal{C}_i$ denote the i^{th} entry of the core list, and $\mathcal{V}_i$ denote the corresponding entry of the regression target list at the end of execution of* Estimate. *For any $\delta \in (0, 1]$, let $\mathcal{E}_\delta$ be the event that, for all $i \in [|\mathcal{C}|]$,*

$$|v(\mathcal{C}_i) - \mathcal{V}_i| \leq \frac{R}{\beta}\sqrt{\frac{4\log(2|\mathcal{C}|/\delta)}{m} + 2\exp(-2\beta T)}\,. \tag{98}$$

Then $P(\mathcal{E}_\delta) \geq 1 - \delta$.

Proof Fix an arbitrary $i \in [|\mathcal{C}|]$, and let $\tilde{v}(\mathcal{C}_i) := \mathbb{E}\left[\int_0^T \exp(-\beta t) r(s_t) dt \middle| s_0 = \mathcal{C}_i\right]$ be the true reward function at state $\mathcal{C}_i$ with the time horizon truncated to T. Then

$$\begin{aligned}
|v(\mathcal{C}_i) - \mathcal{V}_i| &\leq |v(\mathcal{C}_i) - \tilde{v}(\mathcal{C}_i)| + |\tilde{v}(\mathcal{C}_i) - \mathcal{V}_i| && (99)\\
&\leq \int_T^\infty \exp(-\beta t) R dt + |\tilde{v}(\mathcal{C}_i) - \mathcal{V}_i| && (100)\\
&= \frac{\exp(-\beta T) R}{\beta} + |\tilde{v}(\mathcal{C}_i) - \mathcal{V}_i|\,. && (101)
\end{aligned}$$

It remains to bound $|\tilde{v}(\mathcal{C}_i) - \mathcal{V}_i|$. To this end, note that $\mathcal{V}_i$ is of the form $\frac{1}{m}\sum_{j\in[m]} X_j$, where $(X_j)_{j\in[m]}$ are unbiased i.i.d. estimates of $\tilde{v}(\mathcal{C}_i)$. Since these estimates are bounded above, in particular

$$|X_j| \leq \int_0^T \exp(-\beta t) R dt \leq \frac{R}{\beta}\,, \tag{102}$$

we can apply Hoeffding's inequality (Proposition 34) to conclude that, with probability at least $1-\delta$,

$$|\tilde{v}(\mathcal{C}_i) - \mathcal{V}_i| \leq \sqrt{\frac{2R^2\log(2/\delta)}{\beta^2 m}}\,. \tag{103}$$

Taking a union bound over the $|\mathcal{C}|$ elements of $\mathcal{C}$ gives that, with probability $1 - \delta$, for all $i \in [|\mathcal{C}|]$,

$$|\tilde{v}(\mathcal{C}_i) - \mathcal{V}_i| \leq \sqrt{\frac{2R^2\log(2|\mathcal{C}|/\delta)}{\beta^2 m}}\,. \tag{104}$$

Together, Equations (101) and (104) and $\sqrt{x} + \sqrt{y} \leq \sqrt{2x + 2y}$ imply that $P(\mathcal{E}_\delta) \geq 1 - \delta$. ■

Proof [Proof of Proposition 15] Let $\mathcal{C}$ and $\mathcal{V}$ refer respectively to the core list in TAR (Algorithm 3) and the regression target list in Estimate (Algorithm 2) at the end of execution.[3] We will combine the Monte Carlo concentration result above with the principle of kernel ridge regression (see Section A.1) to guarantee that $\hat{v}$ is accurate over the good set.

Applying Proposition 35 with $\delta := \delta_{\text{fail}}/2$ and using that TAR calls Estimate with the horizon $T := \log(m)/\beta \geq \log\left(\frac{m}{2\log(4/\delta_{\text{fail}})}\right)/2\beta$, we have that all $i \in [|\mathcal{C}|]$ simultaneously satisfy

$$|v(\mathcal{C}_i) - \mathcal{V}_i| \leq \frac{3R}{\beta}\sqrt{\frac{\log(4|\mathcal{C}|/\delta_{\text{fail}})}{m}}\,, \tag{105}$$

with probability at least $1 - \delta_{\text{fail}}/2$. On the event that the above display holds, we can apply Proposition 23 with RKHS $\mathcal{H}$, kernel k, and $f^\star = v$, to conclude that, for all $s \in \mathcal{X}$,

$$\begin{aligned}&\left|\hat{v}(s) - v(s)\right| \\ &\quad\leq \left(k(s,s) - k_{\mathcal{C}}(s)^\top (K_{\mathcal{C}} + \lambda I_{|\mathcal{C}|})^{-1} k_{\mathcal{C}}(s)\right)\left(\sqrt{\lambda}\|v\|_{\mathcal{H}} + \frac{3R}{\beta}\sqrt{\frac{|\mathcal{C}|\log(4|\mathcal{C}|/\delta_{\text{fail}})}{m}}\right).\end{aligned} \tag{106}$$

Finally, by the definition of the good set, all $s \in \mathcal{G}_{\mathcal{C}}$ satisfy $k(s,s) - k_{\mathcal{C}}(s)^\top (K_{\mathcal{C}} + \lambda I_{|\mathcal{C}|})^{-1} k_{\mathcal{C}}(s) < \epsilon$. ■

Appendix E. Proof of Proposition 16

Proposition 16 (Coverage) *With probability $1 - \delta_{fail}/2$ over the randomness in the execution of TAR, if $(s_t)_{t\in[0,T]}$ is a trajectory starting at $s_0 \sim \rho$,*

$$P\big(\exists t \in [0,T] \,:\, s_t \in \mathcal{B}(\mathcal{X}, B) \setminus \mathcal{G}_{\mathcal{C}}\big) \leq \delta_{leave}\,, \tag{40}$$

where $\mathcal{G}_{\mathcal{C}}$ is the good set at the time of termination of TAR.

Proof Let p denote the probability of success, i.e. of rolling out and seeing no states in $\mathcal{B}(\mathcal{X}, B) \setminus \mathcal{G}_{\mathcal{C}}$ (at the end of execution of TAR), and let $n := \lceil 2\log(2/\delta_{\text{fail}}\delta_{\text{leave}})/\delta_{\text{fail}}\rceil$.

Note that p is determined by the good set, which is in turn determined by the core list at the time TAR terminates. Since TAR has terminated, a call to Rollout completed without encountering any "surprising" states, i.e. without observing a state in $\mathcal{B}(\mathcal{X}, B) \setminus \mathcal{G}_{\mathcal{C}}$.

By the block matrix inversion lemma, adding a state to $\mathcal{C}$ cannot decrease the size of the good set. Thus over the m calls to Rollout the success probability at each sampled trajectory in each call was no more than p. The probability of the algorithm terminating with $p \leq 1 - \delta_{\text{leave}}$ can thus be upper bounded by the probability that, in an infinite sequence of calls to Rollout, with call $i \in \mathbb{N}$ having a trial count of $n + i$ and a trial success probability of $1 - \delta_{\text{leave}}$, any call to Rollout completes with all trials being successful. Since the probability of m independent trials being successful is $(1 - \delta_{\text{leave}})^m$, a union bound gives us that the probability that **any** of this infinite sequence of calls

3. Note that Estimate is called exactly once during the execution of TAR, and thus the value of $\mathcal{V}$ at the end of execution is well-defined.

to Rollout completes with all trials successful is upper bounded by $\sum_{i\in\mathbb{N}}(1-\delta)^{i+n} = (1-\delta)^n/\delta$. It remains to upper bound this quantity by $\delta_{\text{fail}}/2$. To this end, we have that

$$n \geq \frac{\log(1/(\delta_{\text{fail}}/2)\delta_{\text{leave}})}{\delta_{\text{fail}}/2} \geq \frac{\log(1/(\delta_{\text{fail}}/2)\delta_{\text{leave}})}{\log(1/(1-\delta_{\text{fail}}/2))}\,, \tag{107}$$

where the second inequality comes from $e^x \geq 1 + x$. Rearranging the above display and taking exponents yields $(1-\delta_{\text{fail}}/2)^n \leq (\delta_{\text{fail}}/2)\delta_{\text{leave}}$, and dividing by $\delta_{\text{fail}}/2$ gives the conclusion that $p \geq 1 - \delta_{\text{leave}}$. ■