Journal of Machine Learning Research 23 (2022) 1-46 Submitted 1/22; Revised 8/22; Published 8/22

Nonparametric adaptive control and prediction: theory and randomized algorithms

Nicholas M. Boffi BOFFI@CIMS.NYU.EDU
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012, USA

Stephen Tu STEPHENTU@GOOGLE.COM
Google Brain Robotics
New York, NY 10011, USA

Jean-Jacques E. Slotine JJS@MIT.EDU
Nonlinear Systems Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: George Konidaris

Abstract

A key assumption in the theory of nonlinear adaptive control is that the uncertainty of the system can be expressed in the linear span of a set of known basis functions. While this assumption leads to efficient algorithms, it limits applications to very specific classes of systems. We introduce a novel *nonparametric* adaptive algorithm that estimates an infinite-dimensional density over parameters online to learn an unknown dynamics in a reproducing kernel Hilbert space. Surprisingly, the resulting control input admits an analytical expression that enables its implementation despite its underlying infinite-dimensional structure. While this adaptive input is rich and expressive – subsuming, for example, traditional linear parameterizations – its computational complexity grows linearly with time, making it comparatively more expensive than its parametric counterparts. Leveraging the theory of random Fourier features, we provide an efficient randomized implementation that recovers the complexity of classical parametric methods while provably retaining the expressivity of the nonparametric input. In particular, our explicit bounds only depend *polynomially* on the underlying parameters of the system, allowing our proposed algorithms to efficiently scale to high-dimensional systems. As an illustration of the method, we demonstrate the ability of the randomized approximation algorithm to learn a predictive model of a 60-dimensional system consisting of ten point masses interacting through Newtonian gravitation. By reinterpretation as a gradient flow on a specific loss, we conclude with a natural extension of our kernel-based adaptive algorithms to deep neural networks. We show empirically that the extra expressivity afforded by deep representations can lead to improved performance at the expense of the closed-loop stability that is rigorously guaranteed and consistently observed for kernel machines.

Keywords: Adaptive control, kernel methods, random features, approximation theory.

1. Introduction

One of the fundamental assumptions of nonlinear adaptive systems theory is that the uncertainty of the system can be written as a linear expansion in a set of known basis functions that are nonlinear in the system state. While such linear parameterizations enable the derivation of efficient algorithms with provable guarantees, results outside of this restrictive regime are scarce. Notable examples typically leverage notions of monotonicity (Astolfi and Ortega, 2003; Tyukin et al., 2007)

or convexity (Annaswamy et al., 1998; Fradkov et al., 1999) to make the underlying learning problem tractable.

Here we broaden the applicability of adaptive control by relaxing this classical assumption. In statistical learning, nonlinear function approximation is handled through the use of reproducing kernel Hilbert spaces (RKHSs) (Cucker and Smale, 2002), which are infinite-dimensional function spaces that admit tractable algorithms reminiscent of finite-dimensional linear regression. Inspired by this approach, we develop an adaptive input that learns directly over an RKHS without reference to a finite-dimensional vector of parameters.

One significant drawback of RKHSs is their computational cost. While the representer theorem ensures that estimation in an RKHS can always be cast as a finite expansion over the dataset, the number of parameters grows with its size, which makes learning on large datasets computationally demanding. A key breakthrough in overcoming this difficulty was the theory of random Fourier features, which shows that elements in many RKHSs can be approximated in the linear span of a finite set of *random* basis functions with high probability. Remarkably, the number of random basis functions needed can be shown to scale polynomially (Rahimi and Recht, 2008b) in the function norm and the ambient dimension, which enables efficient computation even in high-dimensional spaces.

In the dynamical systems setting considered in this work, the system trajectory plays the role of the dataset, and the horizon plays the role of its size. Paralleling the statistical learning setting, the complexity of the nonparametric adaptive input that we introduce grows with this horizon. To overcome this complication, we leverage the theory of random features to provide high-probability guarantees on the possibility of uniformly approximating the nonparametric input via a finite-dimensional expansion in random basis functions. Importantly, this approach leads to efficient update laws that match the computational complexity of parametric methods while retaining the expressivity of the RKHS.

We focus on two primary problem settings. The first setting is the classical problem of adaptive control with matched uncertainty, where the uncertainty is assumed to live in the span of the control matrix. Our second application is in adaptive state estimation, where we seek to learn a model of an unknown dynamics governing the evolution of a particular state variable. As a byproduct of our analysis, we exhibit a duality between these two problems reminiscent of the duality between LQR and Kalman filtering in linear control theory. In both settings, we assume that the unmodelled component can be written as the sum of a term that can be linearly parameterized with known physically-motivated basis functions and a term assumed to live in an RKHS. This setup captures the practically relevant setting where a learner can leverage some available physical knowledge of the system but also must perform estimation in a purely unstructured fashion to achieve ideal performance.

The paper is organized as follows. In Section 2, we review related work and summarize our contribution. In Section 3, we formulate the adaptive control and prediction problems. In Section 4, we develop a theory of nonparametric adaptive control, building upon a simple observation reminiscent of the "kernel trick" in machine learning. In Section 5, we review the theory of random Fourier features, which we subsequently apply in Section 6 to design practical adaptive algorithms that asymptotically drive the control or prediction error to a ball around zero. The radius of the ball scales with the approximation error of the random feature expansion, and we give an explicit bound on the number of features needed to ensure that the tracking or prediction error falls below a tolerance threshold ε with high probability. In Section 7, we first study the performance of the nonparametric method in comparison to its randomized approximations on a synthetic adaptive control problem. We subsequently illustrate the effectiveness of its randomized approximations in very high dimension by constructing an adaptive predictor for a 60-dimensional Hamiltonian dynamical system describing the motion of a collection of particles interacting through a $1/r^2$ potential.

2. Related Work and Summary of Contributions

Uniform approximation for adaptive control Most related to the present contribution is a line of work initiated by Kurdila and Lei (2013) and followed by Bobade et al. (2019), who study adaptive control and estimation in RKHSs. In these works, a nonparametric input is treated as an ideal, non-implementable abstraction, and this abstract input is approximated via orthogonal projections or a fixed grid of radial basis functions. Asymptotic convergence results are shown for the approximations, but no finite-sample theory is given, and the grid of centers is chosen in an *ad-hoc* fashion. By gridding the space, these past approaches essentially reduce to a classic line of work by Sanner and Slotine (1992), who approximate an unknown dynamics uniformly with a sum of radial basis functions. These basis functions are spaced on a regular grid, and the grid resolution is chosen based on considerations from sampling theory to ensure a sufficient degree of uniform approximation for the control application. Importantly, while these gridding-based approaches are suitable and highly efficient for low-dimensional systems, they become intractable for higher-dimensional systems. From the perspective of constructing a regular grid, "low-dimensional" is often as restrictive as four-dimensional, which is easily surpassed by modern control applications.

Another closely related line of work is Chowdhary et al. (2012, 2015), who propose to use Gaussian Process (GP) regression for model reference adaptive control. The primary difference with our work is that we derive a control law (cf. Section 4) that operates purely in continuous-time, which obviates the need to take a time derivative of the error signal as supervision. This is important in practice, since it is well-known that computing the time derivative of a signal (e.g. with finite difference approximations) can amplify measurement noise. An additional difference is that our theory quantifies the relation between the number of random features in the function approximation (which governs its quality) and the size of the ball around the desired trajectory to which the system will converge. Finally, our work relies on random feature approximations (Rahimi and Recht, 2007) for tractability, which is simpler to implement in practice than approaches based on sparse GPs.

Randomization and dimensionality-dependence We show that a nonparametric controller can be implemented as the action of a certain kernel integral operator against a known signal over the system trajectory, and we provide an intuitive derivation via the celebrated "kernel trick". This result naturally leads to the randomized approximation methods developed here, which can be seen as a stochastic alternative to a fixed grid of basis functions. The main advantage of randomization is computational: due to concentration of measure, the number of basis functions needed for our construction grows polynomially in the state and input dimension of the underlying control problem. This permits our method to scale to much higher-dimensional systems than prior methods based on gridding, which require a number of basis functions that grows *exponentially* in dimension. Moreover, our work provides a natural path towards developing a theory of adaptive control with more expressive function classes such as single-layer neural networks (Bach, 2017; Bengio et al., 2006), as well as alternative approximation schemes such as the Nyström method (Lu et al., 2016).

Random feature approximations Our randomized algorithm is based on random Fourier features (Rahimi and Recht, 2007, 2008b,a) and their extension to vector-valued functions (Brault et al., 2016; Minh, 2016). We build heavily on the results of Rahimi and Recht (2008b), who prove that the L_∞ approximation error over a compact set for a function f in an RKHS $\mathcal{H}$ decays as $O(1/\sqrt{K})$, where K is the number of features drawn from a particular distribution induced by $\mathcal{H}$. This rate matches that due to Barron (1993) for approximation of functions whose gradients have absolutely integrable Fourier transforms via sums of sigmoidal basis functions.

Control and robotic learning In control and robotics applications, several authors have utilized random features for function approximation in learning stable vector fields (Sindhwani et al., 2018), control contraction metrics (Singh et al., 2020), Lyapunov functions (Boffi et al., 2020), and in velocity gradient-based adaptation (Boffi et al., 2021). However, these works do not analyze the effect of the approximation error introduced by random features on the control performance, nor do

they provide any bounds on the number of random features needed to achieve a specified level of uniform approximation. Adaptive control laws have also been developed for robotic manipulators by exploiting the structure of the governing Euler-Lagrange equations (Slotine and Li, 1987); it is straightforward to extend our results to this setting, or to augment existing robotic adaptive control laws with a nonparametric component to improve robustness to unmodeled disturbances.

Generality of results While the focus of this work is on nonparametric adaptive control and randomized approximation schemes, we have written our results generally to capture a variety of different settings in adaptive control, including Lyapunov-based adaptive control (Krstić et al., 1995), speed/velocity gradient methods (Fradkov et al., 1999; Krstić et al., 1995), mirror descent (Boffi and Slotine, 2021), and contraction metrics (Lopez and Slotine, 2021). We believe that this unification of results represents one of the most general treatments of nonlinear adaptive control available in the literature, and see it to be of independent interest.

3. Problem Formulation

Adaptive control We study nonlinear dynamical systems in *matched uncertainty form*

$$\dot{x} = f(x,t) + g(x,t)\left(u(x,t) - Y(x,t)\alpha_p - h(x)\right), \tag{3.1}$$

where $f : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^n$ is the "nominal dynamics" representing the behavior of the system in the absence of any inputs, $g : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{n\times d}$ is the control matrix describing how an input enters the system, $u : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^d$ is the control input chosen by the learner, $Y : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{d\times p}$ is a matrix of basis functions describing the system's physical structure, $\alpha_p \in \mathbb{R}^p$ is a corresponding vector of physical parameters, and $h \in \mathcal{H}$ is an unknown dynamics in an operator-valued RKHS $\mathcal{H}$ of functions mapping $\mathbb{R}^n \mapsto \mathbb{R}^d$ (Carmeli et al., 2010)[1]. Both h and α_p are unknown, and the goal is to drive $x(t)$ to a bounded desired trajectory $x_d(t)$ by learning a suitable input $u(x,t)$ online. As a supervisory signal, the learner observes an error $e(t) \in \mathbb{R}^s$ at each t with dynamics

$$\dot{e} = f_e(e,t) + g_e(x,t)(u(x,t) - Y(x,t)\alpha_p - h(x)), \tag{3.2}$$

where $f_e : \mathbb{R}^s \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^s$ and $g_e : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{s\times d}$. While the most natural error signal is the trajectory tracking error $e(t) = x(t) - x_d(t)$, we formulate the error signal more abstractly to allow for controllers that only actuate higher order derivatives of the state. This is discussed more in Example 3.6.

Remark 3.1. *Our formulation with h autonomous can be relaxed by considering an RKHS of functions mapping $\mathbb{R}^{n+1} \mapsto \mathbb{R}^d$, i.e., by treating time explicitly as an input variable.*

Adaptive prediction We study nonlinear dynamical systems that can be additively decomposed

$$\dot{x} = f(x,t) = Y(x,t)\alpha_p + h(x),$$

where $f : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^n$ is an unknown dynamics composed of terms that have a similar interpretation to the control setting. The goal is to learn an approximation $\hat{f} : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^n$ of the true dynamics f by designing an estimator

$$\dot{\hat{x}} = \hat{f}(\hat{x},t) + k(\hat{x}, x(t)) \tag{3.3}$$

that will ensure $\hat{x}(t)$ asymptotically approaches $x(t)$. In (3.3), $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is a feedback term satisfying $k(x,x) = 0$ for all x that is used to ensure $\hat{x}(t)$ remains close to $x(t)$ during learning. In this setting, the error signal can be taken as the prediction error $e(t) = \hat{x}(t) - x(t)$. Moreover,

1. A formal definition of an operator-valued RKHS will be provided in Section 4.

the estimator state $\hat{x}(t)$ plays the role of $x(t)$ from the control setting, while $x(t)$ plays the role of the desired trajectory $x_d(t)$. When measurements are no longer available, the open-loop system $\dot{\hat{x}} = \hat{f}(\hat{x}, t)$ may be used to extrapolate the state to make predictions into the future. If measurements are only available at some discrete sampling frequency, the dynamics $\hat{f}$ can be used to interpolate the value of the state between sampling points. This discrete setting is expanded upon in Appendix A.

3.1 Notation

We consider algorithms that update estimates of the physical parameters $\hat{\alpha}_p \in O_p \subseteq \mathbb{R}^p$ and model parameters (when applicable) $\hat{\alpha}_m \in O_m \subseteq \mathbb{R}^m$ online, where O_p and O_m are open convex subsets. We fix twice differentiable mirror maps[2] (potential functions) $\psi_p : O_p \to \mathbb{R}$ (resp. $\psi_m : O_m \to \mathbb{R}$) that are strongly convex with respect to a norm $\|\cdot\|$ on O_p (resp. $\|\cdot\|'$ on O_m) and have locally Lipschitz Hessians. For a potential ψ, we let $d_\psi(\alpha\|\hat{\alpha}) = \psi(\alpha) - \psi(\hat{\alpha}) - \nabla\psi(\hat{\alpha})^\mathsf{T}(\hat{\alpha} - \alpha)$ denote the Bregman divergence associated with ψ. We use $\|\cdot\|_2$ to denote the ℓ_2 norm, $\|\cdot\|_{\mathrm{op}}$ to denote the $\ell_2 \to \ell_2$ operator norm of a matrix, $B_2^n(R)$ to denote the closed ℓ_2 ball of radius R in $\mathbb{R}^n$, $\mathbb{S}^{n-1}$ to denote the unit sphere in $\mathbb{R}^n$, $\mathbb{R}_{\geqslant 0}$ to denote the non-negative reals, and $\mathsf{Sym}_{\geqslant 0}^{n\times n}$ to denote the set of symmetric positive semidefinite $n \times n$ matrices. More generally, for a normed vector space E, $\|\cdot\|_E$ denotes its norm, and $B_E(R)$ denotes a closed ball in E of radius R. For a measure ν, measurable space Θ, and positive integer q, the space $L_2^q(\Theta, \nu)$ denotes the real Hilbert space of square integrable measurable functions $f : \Theta \to \mathbb{R}^q$ with norm $\|f\|_{L_2^q(\Theta,\nu)}^2 = \int_\Theta \|f(\theta)\|_2^2 d\nu(\theta)$. We will often drop the dependence on q when it is clear from the context. Finally, for a positive definite metric $M : \mathbb{R}^n \to \mathsf{Sym}_{\geqslant 0}^{n\times n}$, the Riemannian energy $E_M : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_{\geqslant 0}$ is defined as:

$$E_M(x, y) := \inf_\gamma \int_0^1 \gamma_s(s)^\mathsf{T} M(\gamma(s))\gamma_s(s) ds, \;\; \gamma_s(s) = \frac{d\gamma}{ds}(s),$$

where the infimum ranges over smooth curves γ satisfying $\gamma(0) = x$ and $\gamma(1) = y$.

3.2 Assumptions

To make the above learning problems tractable and to simplify our presentation of results, we require some standard definitions and assumptions. The first requirement is regularity of the nominal dynamics, control matrix, and basis functions.

Definition 3.2. *Let E_1 and E_2 be normed vector spaces. A function $f(x, t)$ mapping $E_1 \times \mathbb{R}_{\geqslant 0} \mapsto E_2$ is said to be locally Lipschitz in x if for every finite $T > 0$ and $R > 0$,*

$$\sup_{t\in[0,T]} \sup_{\substack{\|x\|_{E_1}\leqslant R,\\ \|y\|_{E_1}\leqslant R,\\ x\neq y}} \frac{\|f(x,t) - f(y,t)\|_{E_2}}{\|x - y\|_{E_1}} < \infty.$$

Furthermore, f is said to be locally bounded in x uniformly in t if for every finite $R > 0$,

$$\sup_{t\in\mathbb{R}_{\geqslant 0}} \sup_{\|x\|_{E_1}\leqslant R} \|f(x,t)\|_{E_2} < \infty.$$

Assumption 3.3 (Dynamics regularity)**.** *The functions f, g, and Y are known to the learner. Moreover, f, g, Y, and h are locally Lipschitz in x and locally bounded in x uniformly in t.*

Our second requirement is a set of reasonable conditions on the error to ensure it provides a suitable signal for learning.

2. See e.g. (Bubeck, 2015, Section 4.1) for a definition.

Assumption 3.4 (Error regularity)**.** *f_e and g_e are locally Lipschitz in their first argument and locally bounded in their first argument uniformly in t. Moreover, the following three conditions hold:*

(i) In the absence of the unknown dynamics and any input, zero error is a fixed point,

$$f_e(0,t) = 0 \text{ for all } t \geqslant 0. \tag{3.4}$$

(ii) Bounded error implies a bounded deviation from the desired trajectory,

$$\sup_{t\in[0,T]} \|e(t)\|_2 < \infty \text{ implies } \sup_{t\in[0,T]} \|x(t) - x_d(t)\|_2 < \infty \text{ for all } T > 0. \tag{3.5}$$

(iii) A convergent error signal implies a convergent trajectory

$$\lim_{t\to\infty} \|e(t)\|_2 = 0 \text{ implies } \lim_{t\to\infty} \|x(t) - x_d(t)\|_2 = 0. \tag{3.6}$$

To demonstrate that such error signals can be constructed in practice, we provide a few simple illustrative examples.

Example 3.5 (Systems with regularity)**.** *Consider a system satisfying Assumption 3.3. Then $e(t) = x(t) - x_d(t)$ satisfies the requirements in Assumption 3.4.*

Example 3.6 (Controllable linear time-invariant systems)**.** *Consider the linear time-invariant system $f(x,t) = Ax$ and $g(x,t) = B$ with the pair (A,B) controllable. Let $z(t) \in \mathbb{R}^n$ denote the state of the system expressed in control canonical form, and let $z_d(t) \in \mathbb{R}^n$ denote the corresponding desired trajectory. Define $e(t) = H(s)\left(z_1(t) - z_{d,1}(t)\right)$ where $H(s)$ is a stable transfer function with at most $n-1$ poles and $z_i(t)$ denotes the i^{th} component of z. Then $e(t)$ satisfies the requirements of Assumption 3.4.*

The following stability assumption on the error model is key to our analysis. This assumption is equivalent to requiring that in the absence of the unknown dynamics and adaptive input, the system will nominally tend to the desired trajectory.

Assumption 3.7 (Lyapunov stability of the error)**.** *The error system (3.2) admits a continuously differentiable Lyapunov function $Q : \mathbb{R}^s \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ satisfying for every $e \in \mathbb{R}^s$ and $t \geqslant 0$,*

(i) $\nabla Q(e,t)$ and $\frac{\partial Q}{\partial t}(e,t)$ are locally bounded in e uniformly in t,

(ii) $\nabla Q(e,t)$ is locally Lipschitz in e,

(iii) $\langle \nabla Q(e,t), f_e(e,t)\rangle + \frac{\partial Q}{\partial t}(e,t) \leqslant -\rho(\|e\|_2)$, and

(iv) $\mu_1(\|e\|_2) \leqslant Q(e,t) \leqslant \mu_2(\|e\|_2)$,

where ρ, μ_1, and μ_2 denote class-$\mathcal{K}_\infty$ functions.

While we focus on Lyapunov stability of the error dynamics, our results encompass incremental forms of stability such as contraction (Lohmiller and Slotine, 1998).

Remark 3.8 (Contraction)**.** *We say that the error system is contracting in a metric $M : \mathbb{R}^s \times \mathbb{R}_{\geqslant 0} \to \mathsf{Sym}^{s\times s}_{\geqslant 0}$ if for some $\lambda > 0$,*

$$\frac{\partial f_e}{\partial e}(e,t)^{\mathsf{T}} M(e,t) + M(e,t)\frac{\partial f_e}{\partial e}(e,t) + \dot{M}(e,t) \preccurlyeq -2\lambda M(e,t), \ \ \forall e \in \mathbb{R}^s, t \in \mathbb{R}_{\geqslant 0}. \tag{3.7}$$

Taking the first variation of the Riemannian energy between the error e and the zero trajectory $Q(e,t) = E_{M(\cdot,t)}(e,0)$ shows that $\langle \nabla Q(e,t), f_e(e,t)\rangle + \frac{\partial Q}{\partial t}(e,t) \leqslant -2\lambda Q(e,t)$, so that the energy serves as an exponentially stable Lyapunov function. This correspondence will be used in the prediction setting with $e(t) = \hat{x}(t) - x(t)$.

4. Nonparametric adaptive control and prediction

In this section, we present our primary result in the nonparametric setting. Given a Lyapunov function for the error dynamics as stated in Assumption 3.7, the standard procedure in adaptive nonlinear control is to approximate the unknown dynamics $h(x)$ appearing in (3.1) & (3.2) by an expansion in known basis functions $\Phi : \mathbb{R}^n \to \mathbb{R}^{d\times p}$ (Sanner and Slotine, 1992)

$$\hat{h}(x,t) = \Phi(x)\hat{\alpha}(t), \tag{4.1}$$

and to update the parameter estimates $\hat{\alpha}(t) \in \mathbb{R}^p$ according to a Lyapunov-based update law

$$\dot{\hat{\alpha}}(t) = -\gamma\Phi(x)^\mathsf{T} g_e(x,t)^\mathsf{T}\nabla Q(e,t), \tag{4.2}$$

for $\gamma > 0$ a learning rate.

4.1 Nonparametric form

We start with the following simple observation about the construction in (4.1) & (4.2), which is analogous to the "kernel trick" in machine learning.

Observation 4.1 (Kernel trick)**.** *Assume* $\hat{\alpha}(0) = 0$[3]*. Then the adaptive approximation* (4.1) *with parameters updated according to the algorithm* (4.2) *is equivalent to the nonparametric approximation*

$$\hat{h}(x,t) = \int_0^t \mathsf{K}(x, x(\tau))c(\tau)d\tau, \tag{4.3}$$

where we have defined the kernel function $\mathsf{K} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{d\times d}$ *and coefficients* $c(t) \in \mathbb{R}^d$ *as:*

$$\begin{aligned}\mathsf{K}(x,y) &= \Phi(x)\Phi(y)^\mathsf{T},\\ c(t) &= -\gamma g_e(x(t),t)^\mathsf{T}\nabla Q(e(t),t).\end{aligned}$$

The proof is simple and proceeds by formally writing the solution of (4.2) as an integral over time. Observation 4.1 demonstrates that the function estimates formed by classical adaptive control algorithms only depend on inner products between the basis functions and do not, in principle, require any reference to a vector of parameter estimates. This implies that the basis functions need not be finite-dimensional so long as they admit a computationally inexpensive procedure for computing their inner products, which is precisely the case for an RKHS.

Data-adapted centers Restricting to the case where $\mathsf{K}(\cdot,\cdot)$ is the Gaussian kernel, (4.3) can be seen as leaving a "trail" of Gaussians along the system trajectory $x(\tau)$ for $\tau < t$. In this sense, similar to kernel machines in statistical learning, (4.3) automatically constructs data-adapted centers at which to place spatially-localized basis functions.

Complexity The price paid for the expressivity in the representation (4.3) is that $\hat{h}(x,t)$ now obeys a partial differential equation that must be solved over a horizon of length t at each $x \in \mathbb{R}^n$,

$$\frac{\partial\hat{h}}{\partial t}(x,t) = \mathsf{K}(x,x(t))c(t). \tag{4.4}$$

While (4.4) is decoupled in space so that a global solve is not required, past work from time $\tau < t$ cannot be re-used at time t. Hence, unlike standard parametric methods that incur an $\mathcal{O}(1)$ cost at each timestep, solving (4.4) for the value of $\hat{h}(x,t)$ at a given spatial location x incurs an $\mathcal{O}(t)$ cost at each time t. For most applications, this is prohibitively expensive, and we now turn to efficient approximation schemes that circumvent this difficulty.

3. Note that this is without loss of generality, since any non-zero $\hat{\alpha}(0)$ results in a non-zero $\hat{h}(\cdot,0)$ which can simply be absorbed into h.

4.2 Random feature space

Observation 4.1 motivates us to work with function classes described by kernels. The following definition introduces the notion of an operator-valued kernel.

Definition 4.2 (Operator-valued reproducing kernel, see e.g., Carmeli et al. (2010))**.** *A kernel $\mathsf{K} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{d\times d}$ is said to be an operator-valued reproducing kernel for an RKHS $\mathcal{H}$ if*

(i) For every $\{x_i\}_{i=1}^N \subseteq \mathbb{R}^n$ and $\{w_i\}_{i=1}^N \subseteq \mathbb{R}^d$, it holds that $\sum_{i,j=1}^N \langle w_i, \mathsf{K}(x_i, x_j)w_j\rangle \geqslant 0$.

(ii) $\mathsf{K}(\cdot, x)w \in \mathcal{H}$ for every $x \in \mathbb{R}^n$ and $w \in \mathbb{R}^d$.

(iii) $\mathcal{H}$ can be written $\mathcal{H} = \mathsf{cl}\left\{ f \;\middle|\; \exists\, \{x_i\}_{i=1}^n, \{w_i\}_{i=1}^n \text{ s.t. } f(\cdot) = \sum_{i=1}^n \mathsf{K}(\cdot, x_i)\omega_i \right\}$.

The adaptive algorithms we formulate will be valid for any RKHS $\mathcal{H}$ with a known operator-valued kernel K. However, we focus on RKHSs with specific structure that will enable the design of efficient randomized approximations. These function spaces are described by the following assumption.

Assumption 4.3 (The function class $\mathcal{F}_2$, see e.g., Bach (2017))**.** *The unknown dynamics h lies in an RKHS $\mathcal{H}$ with known operator-valued kernel K. Moreover, K may be written in terms of a feature map $\Phi : \mathbb{R}^n \times \Theta \to \mathbb{R}^{d\times d_1}$ as*

$$\mathsf{K}(x, y) = \int_\Theta \Phi(x, \theta)\Phi(y, \theta)^\mathsf{T} d\nu(\theta), \tag{4.5}$$

with $d_1 \leqslant d$ and where ν is a known probability measure on a measurable space Θ.

In Assumption 4.3, we have overloaded the definition of the feature map Φ as a generalization of the structure of K seen in Observation 4.1. Assumption 4.3 is not very restrictive, as many rich kernels applied in practice – such as the Gaussian and Laplace kernels – can readily written in this form. In particular, the operator-valued generalization of Bochner's theorem (Brault et al., 2016) states that any translation-invariant kernel can be written in the form (4.5) with a feature map

$$\Phi(x, \theta) = B(w)\cos(w^\mathsf{T} x + b), \tag{4.6}$$

where $\Theta \subseteq \mathbb{R}^{n+1}$, $\theta = (w, b)$, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$, and for suitable choices of ν and $B : \mathbb{R}^n \to \mathbb{R}^{d\times d_1}$.

Under Assumption 4.3, it is well-known (c.f. Bach (2017), Appendix A) that $h \in \mathcal{H}$ can be written, for some square-integrable signed density $\alpha : \Theta \to \mathbb{R}^{d_1}$ with respect to the base measure ν, as the integral

$$h(\cdot) = \int_\Theta \Phi(\cdot, \theta)\alpha(\theta) d\nu(\theta), \;\; \|h\|_\mathcal{H}^2 = \|\alpha\|_{L_2(\Theta,\nu)}^2. \tag{4.7}$$

The corresponding Hilbert space is referred to as $\mathcal{F}_2$ (Bach, 2017; Bengio et al., 2006). $\mathcal{F}_2$ is related to the Banach space of single-layer neural networks $\mathcal{F}_1$, which may be obtained by taking the union over all possible base measures for $\mathcal{F}_2$. The space $\mathcal{F}_2$ is convenient for our purposes because it allows us to treat the infinite-dimensional density over parameters α similar to a standard finite-dimensional vector of parameters. To do so, we introduce a second moment regularity condition that will ensure the nonparametric input leads to a stable and convergent feedback system.

Assumption 4.4 (Second moment regularity of Φ)**.** *For every $x \in \mathbb{R}^n$, the second moment of the feature matrix is finite, i.e., $\int_\Theta \|\Phi(x, \theta)\|_{\mathrm{op}}^2 d\nu(\theta) < \infty$. Furthermore, for every $R > 0$,*

$$\sup_{\substack{\|x\|_2 \leqslant R, \|y\|_2 \leqslant R, \\ x \neq y}} \frac{\left(\int_\Theta \|\Phi(x, \theta) - \Phi(y, \theta)\|_{\mathrm{op}}^2 d\nu(\theta)\right)^{1/2}}{\|x - y\|_2} < \infty.$$

To obtain accurate parametric approximations, we may sample points $\theta_i \in \Theta$ i.i.d. from the base measure ν. This has the effect of discretizing the density α into a vector of parameters that can be learned using standard adaptive methods.

4.3 Main results

For simplicity of exposition, we restrict to the case where $\alpha_p = 0$ in (3.1) to focus on convergence of the nonparametric input; the randomized approximations in Section 6 will adapt over both physical and mathematical parameter estimates simultaneously. Moreover, we focus here on the setting of adaptive control. Later, the proof of Theorem 6.7 will highlight a duality between adaptive control and adaptive prediction that immediately implies an analogous result for prediction. The following theorem demonstrates that the nonparametric adaptation algorithm leads to a stable and convergent trajectory.

Theorem 4.5 (Convergence)**.** *Consider system* (3.1) *under Assumptions 3.7, 4.3, and 4.4. Fix* $\alpha_p = 0$ *and let* $\gamma > 0$*. Then the adaptive control input*

$$u(x,t) = -\gamma \int_0^t \mathsf{K}(x, x(\tau)) g_e(x(\tau),\tau)^{\mathsf{T}} \nabla Q(e(\tau),\tau) d\tau$$

ensures that both $x(t)$ *and* $e(t)$ *exist and are uniformly bounded for all* $t \geqslant 0$*. Moreover,* $u(\cdot,t) \in \mathcal{H}$ *for all* $t \geqslant 0$ *and* $\lim_{t\to\infty} \|x(t) - x_d(t)\|_2 = 0$*.*

Next, we study the interpolation properties of the input $u(x,t)$ along the desired trajectory. To this end, we strengthen Definition 3.2 to be uniform in t.

Definition 4.6. *Let* E_1 *and* E_2 *be normed vector spaces. A function* $f(x,t)$ *mapping* $E_1 \times \mathbb{R}_{\geqslant 0} \mapsto E_2$ *is said to be locally Lipschitz in* t *uniformly in* x *if the following two conditions hold for every* $R > 0$*:*

$$\sup_{\|x\|_{E_1} \leqslant R} \sup_{\substack{t_1,t_2 \in \mathbb{R}_{\geqslant 0}, \\ t_1 \neq t_2}} \frac{\|f(x,t_1) - f(x,t_2)\|_{E_2}}{|t_1 - t_2|} < \infty,$$

$$\sup_{t \in \mathbb{R}_{\geqslant 0}} \sup_{\substack{\|x_1\|_{E_1} \leqslant R, \\ \|x_2\|_{E_1} \leqslant R, \\ x_1 \neq x_2}} \frac{\|f(x_1,t) - f(x_2,t)\|_{E_2}}{\|x_1 - x_2\|_{E_1}} < \infty.$$

With Definition 4.6 in hand, we may state the following theorem.

Theorem 4.7 (Interpolation)**.** *Consider the setting of Theorem 4.5. Suppose furthermore that both* $f_e(e,t)$ *and* $g_e(x,t)$ *are locally Lipschitz in their first argument uniformly in* t*. Finally, suppose that for every* $R > 0$*,*

$$\int_\Theta \sup_{\|x\|_2 \leqslant R} \|\Phi(x,\theta)\|_{\mathrm{op}}^2 d\nu(\theta) < \infty.$$

Then the nonparametric input asymptotically interpolates the unknown in the span of the control matrix, $\lim_{t\to\infty} \|g_e(x(t),t)(u(x(t),t) - h(x(t)))\|_2 = 0$*.*

Mirroring the finite-dimensional setting considered by Boffi and Slotine (2021), we now demonstrate that the adaptive input in Theorem 4.5 converges to the minimum RKHS-norm interpolating solution.

Theorem 4.8 (Implicit regularization)**.** *Consider the setting of Theorem 4.5. Define the interpolating set over the trajectory*

$$\mathcal{A} := \left\{ \bar{h} \in \mathcal{H} : \bar{h}(x(t)) = h(x(t)),\ \forall t \geqslant 0 \right\},$$

and assume that $\lim_{t\to\infty} u(\cdot,t) \in \mathcal{A}$*. Then,*

$$\lim_{t\to\infty} u(\cdot,t) \in \operatorname*{argmin}_{\bar{h}\in\mathcal{A}} \|\bar{h}(\cdot)\|_{\mathcal{H}}. \tag{4.8}$$

Given these results for the computationally expensive nonparametric input, we now turn to develop a theory of efficient randomized approximation schemes.

5. Random feature approximation

5.1 Approximation theory

We now demonstrate how the function space $\mathcal{F}_2$ leads to efficient randomized approximation algorithms. These randomized algorithms will enable us to restore the computational advantages of classical finite-dimensional parametric approximations while retaining the expressiveness of the RKHS $\mathcal{F}_2$ with high probability. Roughly speaking, the approach will be to apply the law of large numbers to the expectation (4.7), which leads to a finite-dimensional approximation

$$h(\cdot) \approx \frac{1}{K}\sum_{i=1}^{K} \Phi(\cdot,\theta_i)\alpha_i,$$

where the $\theta_i \sim \nu$ are drawn i.i.d. from the base measure ν and the $\alpha_i = \alpha(\theta_i) \in \mathbb{R}^{d_1}$ are treated as parameters to be learned. K denotes the number of sampling points and will tune the accuracy of the approximation. We provide a bound on the number of random features K needed to ensure that there exists a set of weights $\{\alpha_i\}$ capable of ε-uniformly approximating h on a fixed compact set $X \subset \mathbb{R}^n$. To begin, let $B_\Phi(\delta)$ be any function that satisfies, for any $\delta \in (0,1)$,

$$\mathbb{P}_{\theta\sim\nu}\left(\sup_{x\in X}\|\Phi(x,\theta)\|_{\mathrm{op}} > B_\Phi(\delta)\right) \leqslant \delta.$$

Then, for any $\eta \in (0,1)$, define a truncated version of Φ as

$$\Phi_\eta(x,\theta) := \Phi(x,\theta)\mathbf{1}\left\{\|\Phi(x,\theta)\|_{\mathrm{op}} \leqslant B_\Phi(\eta)\right\}.$$

We will be interested in approximating functions over the subset

$$\mathcal{F}_2(B) = \left\{ f(\cdot) = \int_\Theta \Phi(\cdot,\theta)\alpha(\theta)d\nu(\theta) \;\middle|\; \operatorname*{ess\,sup}_{\theta\in\Theta}\|\alpha(\theta)\|_2 \leqslant B \right\} \subset \mathcal{F}_2,$$

which is dense in $\mathcal{F}_2$ as $B \to \infty$ (Rahimi and Recht, 2008b); this bound on the density $\alpha(\theta)$ is needed to obtain a uniform approximation result. With this notation in hand, we may extend the approximation theory of Rahimi and Recht (2008b) to vector-valued functions.

Proposition 5.1 (Approximation error)**.** *Let $X \subset \mathbb{R}^n$ be compact. Fix $\delta \in (0,1)$, $B_h > 0$, $h \in \mathcal{F}_2(B_h)$, and a positive integer K. Let $\theta_1, ..., \theta_K$ be i.i.d. draws from ν. Put $\eta = \frac{\delta}{2K}$. With probability at least $1-\delta$, there exist weights $\{\alpha_i\}_{i=1}^K \subset \mathbb{R}^{d_1}$ such that $\|\alpha_i\|_2 \leqslant B_h$ for $i = 1, ..., K$, and*

$$\left\|\frac{1}{K}\sum_{i=1}^{K}\Phi(\cdot,\theta_i)\alpha_i - h\right\|_\infty \leqslant \frac{2}{K}\mathbb{E}\left\|\sum_{k=1}^{K}\varepsilon_i\Phi_\eta(\cdot,\theta_i)\alpha(\theta_i)\right\|_\infty + \sqrt{2}B_\Phi(\eta)B_h\sqrt{\frac{\log(2/\delta)}{K}} + B_h\sqrt{\frac{\delta\sup_{x\in X}\mathbb{E}\|\Phi(x,\theta)\|_{\mathrm{op}}^2}{2K}}.$$

Above, each ε_i is an i.i.d. Rademacher random variable[4] *and $\|f\|_\infty := \sup_{x\in X}\|f(x)\|_2$.*

4. That is, $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$.

In order to bound the Rademacher complexity term appearing in Proposition 5.1, we now make a few more assumptions on the structure of $\Phi(x,\theta)$. These assumptions are motivated by the operator-valued Bochner's theorem (Brault et al., 2016).

Assumption 5.2. *The feature space Θ is a subset of $\mathbb{R}^{n+1}$, so that $\theta \in \Theta$ may be written as $\theta = (w,b)$ with $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Moreover, the feature map can be factorized as $\Phi(x,\theta) = \phi(w^\mathsf{T} x + b)M(w)$ for $M : \mathbb{R}^n \to \mathbb{R}^{d\times d_1}$ and a 1-Lipschitz scalar function $\phi : \mathbb{R} \to [-1,1]$.*

Because $|\phi| \leqslant 1$, we may take $B_\Phi(\delta)$ to be any function that satisfies $\mathbb{P}(\|M(w)\|_{\mathrm{op}} > B_\Phi(\delta)) \leqslant \delta$. Accordingly, we have $\Phi_\eta(x,\theta) = M_\eta(w)\phi(w^\mathsf{T}x+b)$ with $M_\eta(w)$ defined as $M_\eta(w) := M(w)\mathbf{1}\{\|M(w)\|_{\mathrm{op}} \leqslant B_\Phi(\eta)\}$. With these extra assumptions in place, we can bound the Rademacher complexity term as follows.

Proposition 5.3 (Rademacher complexity bound)**.** *Let Assumption 5.2 hold, and denote $B_X := \sup_{x\in X}\|x\|_2$. Then for any $\eta \in (0,1)$,*

$$\frac{2}{K}\mathbb{E}\left\|\sum_{i=1}^K \varepsilon_i \Phi_\eta(\cdot;\theta_i)\alpha(\theta_i)\right\|_\infty \leqslant \frac{4B_h B_\Phi(\eta)}{\sqrt{K}}\left[B_X\sqrt{\mathbb{E}\|w_1\|_2^2} + \sqrt{d_1}\right].$$

Combining Proposition 5.1 and Proposition 5.3, we have that with probability $1-\delta$,

$$\begin{aligned}&\inf_{\{\alpha_i\}_{i=1}^K \subseteq \mathbb{R}^{d_1} : \|\alpha_i\|_2 \leqslant B_h} \left\|\frac{1}{K}\sum_{k=1}^K \Phi(\cdot,\theta_i)\alpha_i - h\right\|_\infty \\ &\leqslant \frac{B_h}{\sqrt{K}}\left[2B_\Phi\left(\frac{\delta}{2K}\right)\left(2B_X\sqrt{\mathbb{E}\|w_1\|_2^2} + 2\sqrt{d_1} + \sqrt{\log(2/\delta)}\right) + \sqrt{\frac{\delta}{2}\mathbb{E}\|M(w)\|_{\mathrm{op}}^2}\right]. \qquad (5.1)\end{aligned}$$

To simplify this expression, we now look at some particular choices of kernels.

5.2 Examples of Reproducing Kernels

In what follows, we consider a few examples of vector-valued kernels.

5.2.1 Shift-invariant kernels

First, we consider shift-invariant kernels from Brault et al. (2016) and Minh (2016). Let $\mathsf{k}(x-z)$ be an arbitrary scalar shift-invariant kernel and denote by μ the normalized inverse Fourier transform of $\mathsf{k}(\cdot)$. We will assume generically that $\mathbb{E}_{w\sim\mu}\|w\|_2^2 \asymp n$ where μ denotes the marginal of ν over b.

Decomposable kernels Let $\mathsf{K}(x,z) = A\mathsf{k}(x-z)$ for any positive semidefinite $A = BB^\mathsf{T}$. Then $\Phi(x,\theta) = B\cos(w^\mathsf{T}x + b)$ and $B_\Phi(\delta) = \|B\|_{\mathrm{op}}$. Here, the approximation error bound (5.1) scales as $\frac{B_h\|B\|_{\mathrm{op}}}{\sqrt{K}}\left(B_X\sqrt{n} + \sqrt{d_1}\right)$.

Curl-free kernel Let $n = d$ and set $\mathsf{K}(x,z) = -\nabla^2\mathsf{k}(x-z)$. Then $A(w) = ww^\mathsf{T}$ and $\Phi(x,\theta) = w\cos(w^\mathsf{T}x + b)$. If $\mu \sim N(0,\sigma^2 I)$, then $B_\Phi(\delta) = \sqrt{n} + 2\sigma\sqrt{\log(1/\delta)}$ by standard Gaussian concentration results. The approximation error bound (5.1) then scales as $\frac{B_h(B_x\vee 1)}{\sqrt{K}}(n + \log K)$.

Divergence-free kernel Again let $n = d$. Set $\mathsf{K}(x,z) = (\nabla^2 - I\Delta)\mathsf{k}(x-z)$, where Δ is the Laplacian and I is the identity matrix. Then $A(w) = \|w\|_2^2 P_w^\perp$, where P_M denotes the orthogonal projection onto the range of M and $P_M^\perp = I - P_M$. Hence, $\Phi(x,\theta) = \|w\|_2 P_w^\perp\cos(w^\mathsf{T}x + b)$. If $\nu \sim N(0,\sigma^2 I)$, then $B_\Phi(\delta) = \sqrt{n} + 2\sigma\sqrt{\log(1/\delta)}$. The approximation error bound (4.5) also scales as $\frac{B_h(B_x\vee 1)}{\sqrt{K}}(n + \log K)$.

5.2.2 Other kernels

We now consider some other possible choices of kernels.

Kernels leveraging prior physical information Any known physical structure can easily be combined with reproducing kernels. As a concrete example, suppose the state x decomposes as $x = (x_1, x_2) \in \mathbb{R}^{n_1+n_2}$, and that the unknown dynamics factorizes as $h(x) = h_1(x_1)h_2(x_2)$, where $h_1 : \mathbb{R}^{n_1} \to \mathbb{R}^d$ is a known vector-valued function and $h_2 : \mathbb{R}^{n_2} \to \mathbb{R}$ is an unknown function in an RKHS with scalar kernel k. Then we can set $\mathsf{K}((x_1, x_2), (z_1, z_2)) = h_1(x_1)h_1(z_1)^\mathsf{T}\mathsf{k}(x_2, z_2)$. This type of structural simplification is common in, e.g., robotic applications (Sanner and Slotine, 1995).

The neural tangent kernel The neural tangent kernel (Jacot et al., 2018) was recently developed as an approximation to infinitely wide deep neural networks. Consider a network $h(x, \theta)$, where x denotes the network input and θ denotes the network parameters. The NTK is defined as the following kernel:

$$\mathsf{K}(x, z) = \mathbb{E}_{\theta\sim\mathcal{D}}\left[\frac{\partial h}{\partial\theta}(x, \theta)^\mathsf{T}\frac{\partial h}{\partial\theta}(z, \theta)\right],$$

where $\mathcal{D}$ is the distribution used to initialize the weights of the network. Expressions of the NTK for various common architectures are available in closed form (Arora et al., 2019).

6. Randomized adaptive control and prediction

We now demonstrate how the nonparametric input in Theorem 4.5 can be approximated using the uniform approximation theory of Section 5 to obtain adaptive control and prediction algorithms with high-probability guarantees of convergence. We state completely general results under the assumption that the unknown dynamics $h(\cdot)$ can be uniformly approximated to a desired degree of accuracy, similar to the classical results of Sanner and Slotine (1992) but in a generalized context. Taking $h(\cdot)$ to lie in the function space $\mathcal{F}_2$ and applying the results of Section 5 immediately gives a sufficient bound on the number of random features needed to track the desired trajectory to a given tolerance.

6.1 Deadzones

Before we present our main approximate algorithms, we first introduce the notion of a deadzone. Since any finite-dimensional approximation to $h(\cdot)$ will have some non-zero approximation error, any adaptive algorithm cannot learn below this noise floor; a deadzone allows us to disable adaptation when the only residual error remaining is due to approximation error.

Definition 6.1. *Let $\Delta > 0$. A continuously differentiable function $\sigma_\Delta : \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ is called Δ-admissible deadzone if:*

(i) $0 \leqslant \sigma_\Delta$ *and* $\sigma_\Delta(x) = 0$ *for all* $x \in [0, \Delta]$,

(ii) $0 \leqslant \sigma'_\Delta$ *and* $\sigma'_\Delta(x) = 0$ *for all* $x \in [0, \Delta]$,

(iii) σ'_Δ *is locally Lipschitz.*

The function σ_Δ is called a (Δ, L, B)-admissible deadzone if condition (iii) *is replaced with the condition that σ'_Δ is L-Lipschitz and B-bounded.*

We now give some examples of Δ-admissible deadzones. The first example is a direct extension of the deadzone used in Sanner and Slotine (1992).

Example 6.2. *Fix a scalar $\delta > 0$. Let $s_\delta : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ be defined as $s_\delta(x) := (x-\delta)\mathbf{1}\{x > \delta\}$. For any $\Delta > 0$, the function $x \mapsto s^2_{\sqrt{\Delta}}(\sqrt{x})$ is a $(\Delta, 1/(2\Delta), 1)$-admissible deadzone.*

An issue with a deadzone based on s_δ is that the Lipschitz constant of the derivative diverges with vanishing Δ. This makes it challenging to prove sharp "approximate interpolation" results similar to Theorem 4.7. To remedy this issue, we construct a deadzone where the Lipschitz constant of the derivative is decoupled from Δ. The following construction is directly inspired by smooth approximations to the hinge loss for support vector machines (see e.g. Chapelle (2007)).

Example 6.3. *Fix $\delta > 0$ and $\gamma > 0$. Define $s_{\delta,\gamma}$ as:*

$$s_{\delta,\gamma}(x) := \begin{cases} 0 & \text{if } x \leqslant \delta, \\ \frac{(x-\delta)^2}{4\gamma} & \text{if } x \in (\delta, \delta + 2\gamma), \\ x - (\delta + \gamma) & \text{if } x \geqslant \delta + 2\gamma. \end{cases}$$

For any $\Delta > 0$ and $\gamma > 0$, the function $s_{\Delta,\gamma}$ is a $(\Delta, 1/(2\gamma), 1)$-admissible deadzone.

Worked details of Examples 6.2 and 6.3 may be found in Appendix E. Our results to come will be stated in terms of an arbitrary deadzone according to Definition 6.1, but concrete instantiations can be found via these prescriptions.

6.2 Adaptive control

We are now ready to present our main result in the setting of approximate control. The following is a general result about adaptive control with uniform approximation that can be applied with an arbitrary choice of basis.

Theorem 6.4 (Adaptive control with finite-dimensional approximation)**.** *Suppose that Assumption 3.7 holds. Let $\alpha_{\ell,0} := \arg\min_{\alpha \in O_\ell} \psi_\ell(\alpha)$ for $\ell \in \{p, m\}$. Fix $B_{\alpha_p} > 0$ satisfying $d_{\psi_p}(\alpha_p \| \alpha_{p,0}) \leqslant B_{\alpha_p}$, $B_{\alpha_m} > 0$, and R satisfying*

$$R > \mu_1^{-1}\left(Q(e(0), 0) + B_{\alpha_p} + B_{\alpha_m}\right).$$

Suppose there exists a finite C_e such that for every $T > 0$:

$$\max_{t \in [0,T]} \|e(t)\|_2 \leqslant R \text{ implies } \|x(T) - x_d(T)\|_2 \leqslant C_e R. \tag{6.1}$$

Let $\Psi : \mathbb{R}^n \to \mathbb{R}^{d \times m}$ be a locally Lipschitz feature map. Define the constants

$$\begin{aligned} B_d &:= \sup_{t \geqslant 0} \|x_d(t)\|_2, \\ B_x &:= C_e R + B_d, \\ B_{g_e} &:= \sup_{t \geqslant 0} \sup_{\|x\|_2 \leqslant B_x} \|g_e(x,t)\|_{\mathrm{op}}, \\ B_{\nabla Q} &:= \sup_{t \geqslant 0} \sup_{\|e\|_2 \leqslant R} \|\nabla Q(e,t)\|_2, \\ B_{\mathrm{approx}} &:= \inf_{d_{\psi_m}(\alpha_m \| \alpha_{m,0}) \leqslant B_{\alpha_m}} \sup_{\|x\|_2 \leqslant B_x} \|\Psi(x)\alpha_m - h(x)\|_2. \end{aligned}$$

Let Δ be any positive constant satisfying

$$\Delta \geqslant \mu_2(\rho^{-1}(2B_{g_e} B_{\nabla Q} B_{\mathrm{approx}})),$$

and let σ_Δ be a Δ-admissible deadzone. Then the dynamical system

$$\begin{aligned}
\dot{x} &= f(x,t) + g(x,t)(u(x,t) - Y(x,t)\alpha_p - h(x)),\\
\dot{e} &= f_e(e,t) + g_e(x,t)(u(x,t) - Y(x,t)\alpha_p - h(x)),\\
u(x,t) &= Y(x,t)\hat{\alpha}_p + \Psi(x)\hat{\alpha}_m,\\
\frac{d}{dt}\nabla\psi_p(\hat{\alpha}_p) &= -\sigma_\Delta'(Q(e,t))Y(x,t)^\mathsf{T} g_e(e,t)^\mathsf{T}\nabla Q(e,t),\\
\frac{d}{dt}\nabla\psi_m(\hat{\alpha}_m) &= -\sigma_\Delta'(Q(e,t))\Psi(x)^\mathsf{T} g_e(e,t)^\mathsf{T}\nabla Q(e,t),
\end{aligned}$$

with initial conditions $x(0) = x_0$, $e(0) = m(x_0, 0)$, $\hat{\alpha}_p(0) = \alpha_{p,0}$, and $\hat{\alpha}_m(0) = \alpha_{m,0}$ has a solution $(x(t), e(t), \hat{\alpha}_p(t), \hat{\alpha}_m(t))$ that exists for all $t \geqslant 0$. Furthermore,

$$\limsup_{t\to\infty} \|e(t)\|_2 \leqslant \mu_1^{-1}(\Delta).$$

Theorem 6.4 can be used in conjunction with the results of Section 5 to obtain a high-probability guarantee for control, as illustrated by the following example.

Example 6.5 (Adaptive control with random features)**.** *Suppose for simplicity that $d_{\psi_m}(x\|y) = \frac{1}{2}\|x-y\|_2^2$ is the Euclidean distance. Fix a positive integer K, and let $\delta \in (0,1)$. Assume that $h \in \mathcal{F}_2(B_h)$ under Assumption 5.2, and again for simplicity assume that the kernel is decomposable as in Section 5.2. Set $B_{\alpha_m} = B_h^2/2$. Let $\{\theta_i\}_{i=1}^K$ be i.i.d. draws from ν. Then, by Equation 5.1, with probability at least $1-\delta$ there exists $\alpha_m = (\alpha_{m,1}, ..., \alpha_{m,K}) \in \mathbb{R}^{Kd_1}$ satisfying $\|\alpha_{m,i}\|_2 \leqslant B_h/K$ for $i = 1, ..., K$ and*

$$\sup_{\|x\|_2 \leqslant B_x} \|h(x) - \Psi\left(x; \{\theta_i\}_{i=1}^K\right)\alpha_m\|_2 \leqslant \frac{C(h,\delta)(B_x\sqrt{n} + \sqrt{d_1})}{\sqrt{K}},$$

with $\Psi(x; \{\theta_i\}_{i=1}^K) = \left[\Phi(x,\theta_1), ..., \Phi(x,\theta_K)\right] \in \mathbb{R}^{d\times Kd_1}$. Here, $C(h,\delta) > 0$ is a constant that depends only on h and δ. Note that

$$d_{\psi_m}(\alpha_m\|0) = \frac{1}{2}\|\alpha_m\|_2^2 = \frac{1}{2}\sum_{i=1}^K \|\alpha_{m,i}\|_2^2 \leqslant \sum_{i=1}^K \frac{B_h^2}{2K^2} = \frac{B_h^2}{2K} \leqslant B_{\alpha_m},$$

so that $B_{\text{approx}} \leqslant \frac{C(h,\delta)(B_x\sqrt{n}+\sqrt{d_1})}{\sqrt{K}}$. Hence, to ensure $\limsup_{t\to\infty}\|e(t)\|_2 \leqslant \varepsilon$ for some $\varepsilon > 0$, it suffices to take K satisfying

$$K \geqslant \frac{4B_{g_e}^2 B_{\nabla Q}^2 C(h,\delta)^2 (B_x\sqrt{n} + \sqrt{d_1})^2}{\rho^2(\mu_2^{-1}(\mu_1(\varepsilon)))}.$$

Suppose that $\mu_1(x) = \mu x$, $\mu_2(x) = Lx$, and $\rho(x) = \beta x^5$. Then this bound simplifies to

$$K \geqslant \frac{4}{\beta^2\varepsilon^2}\left(\frac{L}{\mu}\right)^2 B_{g_e}^2 B_{\nabla Q}^2 C(h,\delta)^2 (B_x\sqrt{n} + \sqrt{d_1})^2.$$

Approximation region For simplicity of presentation, we have chosen the approximation region in Theorem 6.4 large enough to cover the variation of the error signal throughout adaptation. Alternatively, the approximation region can be specified *a-priori*, and sliding mode control can be used to force the system to stay inside the approximation region. Such a formulation requires additional technical assumptions on the error dynamics.

5. For $V(t)$ a quadratic Lyapunov function certifying exponential stability, it is a simple calculation to show that one can take $Q(t) = \sqrt{V(t)}$ to obtain such linear functions for μ_1, μ_2 and ρ.

Contraction Assume that the error dynamics is contracting. Then we may take $Q(e,t)$ to be the Riemannian energy as in Remark 3.8 and set $\psi_\ell(\cdot) = \frac{1}{2}\|\cdot\|_2^2$ for $\ell \in \{p, m\}$ to recover the contraction metric-based adaptation law due to Lopez and Slotine (2021)

$$\dot{\hat{\alpha}}_m = -\Psi(x)^\mathsf{T} g_e(x,t)^\mathsf{T} M(e,t)\gamma_s(e,0,t).$$

Here, $\gamma_s(e,0,t)$ denotes the tangent vector to a geodesic in the metric $M(e,t)$ between e and the origin at the endpoint e (a similar metric-based update also holds for $\hat{\alpha}_p$).

Mirror descent By analogy to mirror descent, the choice of potential functions $\psi_p(\cdot)$ and $\psi_m(\cdot)$ can be used to regularize the learned physical and random feature models, or can be used to improve convergence when adapted to the problem geometry (Boffi and Slotine, 2021). The random sinusoidal features considered in Section 5 are uniformly bounded in ℓ_∞ norm independent of the number of parameters. This observation suggests that, for a large number of features, a potential function strongly convex with respect to the ℓ_1 norm such as the hypentropy potential due to Ghai et al. (2020) may lead to improved performance.

Interpolation We conclude our treatment of adaptive control by presenting an approximate version of Theorem 4.7, which demonstrates how the approximation error from finite-dimensional truncation translates into an interpolation error for the learned dynamics approximation. Specifically, if Theorem 6.4 is invoked with a (Δ, L, B)-admissible deadzone, then the following result shows that the interpolation error is bounded by $O\left(\sqrt{\mu_1^{-1}(\Delta)(1+L)}\right)$. This motivates the construction in Example 6.3.

Theorem 6.6 (Approximate interpolation)**.** *Suppose the hypotheses of Theorem 6.4 hold. Let σ_Δ denote a (Δ, L, B)-admissible deadzone, and assume that f_e, g_e, and Y are locally Lipschitz in their first arguments uniformly in t. Then there exist constants $C_1 > 0$ and $C_2 > 0$ not depending on Δ such that*

$$\limsup_{t\to\infty} \|g_e(x(t),t)(u(x(t),t) - Y(x(t),t)\alpha_p - h(x(t)))\|_2 \leqslant C_1\sqrt{\mu_1^{-1}(\Delta)(1+L)} + C_2\mu_1^{-1}(\Delta).$$

6.3 Adaptive prediction

Similar to Theorem 6.4, the following theorem designs a predictor by leveraging the ability to uniformly approximate the unknown dynamics to a suitable degree of accuracy.

Theorem 6.7 (Adaptive prediction with uniform approximation)**.** *Suppose that the trajectory $x(t)$ of the system $\dot{x} = f(x,t)$ is uniformly bounded. Choose a continuous and locally Lipschitz $k(\hat{x}, x)$ such that $f(\hat{x},t) + k(\hat{x}, x(t))$ is contracting in a metric $M : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathsf{Sym}_{\succ 0}^{n\times n}$ with rate $\lambda > 0$, and suppose that the metric M satisfies $\mu I \preccurlyeq M(\hat{x},t) \preccurlyeq LI$ for all $\hat{x}$ and t. Let $\gamma(\cdot;\hat{x},x,t) : [0,1] \to \mathbb{R}^n$ denote a geodesic between $\hat{x}$ and x in the metric $M(\hat{x},t)$, and let $\gamma_s(s;\hat{x},x,t)$ denote the derivative of $s \mapsto \gamma(s;\hat{x},x,t)$. Suppose that the map $(\hat{x},t) \mapsto \|\gamma_s(0;\hat{x},x(t),t)\|_2$ is locally bounded in $\hat{x}$ uniformly in t. Fix any $B_{\alpha_p} > 0$ satisfying $d_{\psi_p}(\alpha_p\|\alpha_{p,0}) \leqslant B_{\alpha_p}$, any $B_{\alpha_m} > 0$, and any R satisfying*

$$R > \sqrt{\frac{Q(\hat{x}(0),0) + B_{\alpha_p} + B_{\alpha_m}}{\mu}}, \;\; Q(\hat{x},t) := E_{M(\cdot,t)}(\hat{x}, x(t)).$$

Let $\Psi : \mathbb{R}^n \to \mathbb{R}^{d\times m}$ be a locally Lipschitz feature map. Define the following constants

$$\begin{aligned} B_x &:= \sup_{t\geqslant 0}\|x(t)\|_2, \\ B_{\hat{x}} &:= R + B_x, \\ B_\gamma &:= \sup_{t\geqslant 0}\sup_{\|\hat{x}\|_2\leqslant B_{\hat{x}}}\|\gamma_s(0;\hat{x},x(t),t)\|_2, \\ B_{\text{approx}} &:= \inf_{d_{\psi_m}(\alpha_m\|\alpha_{m,0})\leqslant B_{\alpha_m}}\sup_{\|\hat{x}\|_2\leqslant B_{\hat{x}}}\|\Psi(\hat{x})\alpha_m - h(\hat{x})\|_2. \end{aligned}$$

Choose any Δ satisfying $\Delta \geqslant \frac{L^2 B_\gamma B_{\text{approx}}}{\lambda\mu}$, and let σ_Δ be a Δ-admissible deadzone. Then the dynamical system

$$\begin{aligned} \dot{\hat{x}} &= \hat{f}(\hat{x},\hat{\alpha}_p,\hat{\alpha}_m,t) + k(\hat{x},x(t)), \\ \hat{f}(\hat{x},\hat{\alpha}_p,\hat{\alpha}_m,t) &= Y(\hat{x},t)\hat{\alpha}_p + \Psi(\hat{x})\hat{\alpha}_m, \\ \frac{d}{dt}\nabla\psi_p(\hat{\alpha}_p) &= -\sigma_\Delta'(Q(\hat{x},t))Y(\hat{x},t)^\mathsf{T}\nabla Q(\hat{x},t), \\ \frac{d}{dt}\nabla\psi_m(\hat{\alpha}_m) &= -\sigma_\Delta'(Q(\hat{x},t))\Psi(\hat{x})^\mathsf{T}\nabla Q(\hat{x},t), \end{aligned}$$

with initial conditions $\hat{x}(0) = \hat{x}_0$, $\hat{\alpha}_p(0) = \alpha_{p,0}$, and $\hat{\alpha}_m(0) = \alpha_{m,0}$ has a solution that exists for all $t \geqslant 0$. Furthermore,

$$\limsup_{t\to\infty}\|\hat{x}(t) - x(t)\|_2 \leqslant \sqrt{\frac{\Delta}{\mu}}.$$

Constructing a metric Theorem 6.7 requires a metric $M(\hat{x},t)$ such that $\bar{f}(\hat{x},t) := f(\hat{x},t) + k(\hat{x},x(t))$ is contracting. One such metric can always be obtained by taking $k(\hat{x},x) = -\zeta(\hat{x}-x)$, in which case $\frac{\partial\bar{f}}{\partial\hat{x}}(\hat{x},t) = \frac{\partial f}{\partial\hat{x}}(\hat{x},t) - \zeta I$. If we further assume that $\frac{\partial f}{\partial\hat{x}}$ is locally bounded in $\hat{x}$ uniformly in t and that $x(t)$ is uniformly bounded, then there exists a finite $\zeta \in (0,\infty)$ such that $\bar{f}$ is contracting in the identity metric $M(\hat{x},t) = I$. In the case where f is known, k can be tailored to the system physics to obtain improved convergence (Chung and Slotine, 2009).

Duality The proof of Theorem 6.7 highlights a duality between the nonlinear adaptive control and nonlinear adaptive prediction problems reminiscent of the duality between LQR and Kalman filtering in linear control theory. Intuitively, any model capable of predicting the time evolution of a system could be used to control the system. Conversely, a model that can be used to control a system could instead be used to predict its evolution.

Interpolation Theorem 6.7 assumes that the true system state $x(t)$ is measured continuously and concludes that the learned prediction $\hat{x}(t)$ will asymptotically become consistent with the observed measurements up to a level specified by the accuracy of the uniform approximation. Applying duality, the interpolation result in Theorem 6.6 shows that the learned model $\hat{f}(\hat{x},\hat{\alpha}_p,\hat{\alpha}_m,t)$ becomes approximately consistent with the true model along the trajectory $x(t)$.

Discrete sampling In practical applications, measurements of the true system state are obtained at discrete instants, and an open-loop predictor with fixed parameters is used to extrapolate beyond them. The parameters are then updated according to a discretized adaptation law when measurements are received. In Appendix A, we demonstrate how the nominal contraction properties required by Theorem 6.7 can be preserved with discrete measurements by taking the feedback term $k(\hat{x},x)$ to have a sufficiently high contraction rate in comparison to the spacing between measurements Δt.

7. Simulations

We now study the empirical performance of the nonparametric method and its randomized approximation. In the control setting we directly compare the kernel and approximate inputs. In prediction we illustrate the ability of the random feature approximation to scale to high-dimensional systems. In addition, we study the convergence of the prediction and interpolation errors as a function of K.

7.1 Adaptive control

Here we consider a synthetic example in adaptive control to compare the nonparametric adaptive input to its randomized approximation.

System dynamics We study the stable linear time-invariant system

$$\dot{x} = A\left(x - \frac{3}{2}\mathbf{1}\right) + u(x,t) - h(x), \;\; x \in \mathbb{R}^5, \;\; h(x) = \sin(x)\mathrm{erf}(x), \tag{7.1}$$

where A is a known matrix with eigenvalues lying entirely in the left half-plane and $\mathbf{1}$ denotes the vector of ones. The operations defining h are applied elementwise to each coordinate. The error signal is set to $e(t) = x(t) - \frac{3}{2}\mathbf{1}$, and the desired trajectory is constant at the nominal equilibrium point $x_d(t) = \frac{3}{2}\mathbf{1}$. This system admits a Lyapunov function $Q(x,t) = \frac{1}{2}(x - x_d(t))^\mathsf{T} P(x - x_d(t))$, where P is the unique positive definite solution to the Lyapunov matrix equation $A^\mathsf{T}P + PA = -I$.

Implementation We apply a nonparametric input generated by the Gaussian kernel

$$\mathsf{K}(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right) I, \;\; \sigma = 0.1.$$

For its randomized approximation, we use the random Fourier features described in Section 5.2. Both the randomized and nonparametric adaptive laws are obtained by forward Euler integration with a fixed timestep $\Delta t = 0.001$. At each time, the kernel input (4.3) is evaluated via a Riemann sum approximation at the same resolution,

$$u(x,t) = \int_0^t K(x, x(\tau))c(\tau)d\tau \approx \sum_{i=0}^{n_t} K(x, x(t_i))c(t_i)\Delta t$$

with $n_t = t/\Delta t$. This corresponds to solving the pointwise-decoupled partial differential equation

$$\frac{\partial u}{\partial t}(x,t) = K(x, x(t))c(t)$$

again via forward Euler integration with a timestep Δt.

Results (Figure 1) Error bars around the random feature curves display the 20% and 80% quantiles, while the solid central curves display the corresponding median. In comparison to each value of K, the kernel input obtains the best tracking performance both transiently and asymptotically by several orders of magnitude. The tracking error at each fixed time decreases monotonically as a function of K (Figure 1A). The overall magnitude of the adaptive control input $\|u(x,t)\|_2$ decreases monotonically as a function of K, and the kernel input consistently applies the lowest magnitude input despite obtaining the best performance (Figure 1B). Similar to the tracking error, the kernel input obtains the best dynamics approximation by several orders of magnitude, and the dynamics interpolation error decreases monotonically as a function of K for each fixed time (Figure 1C).

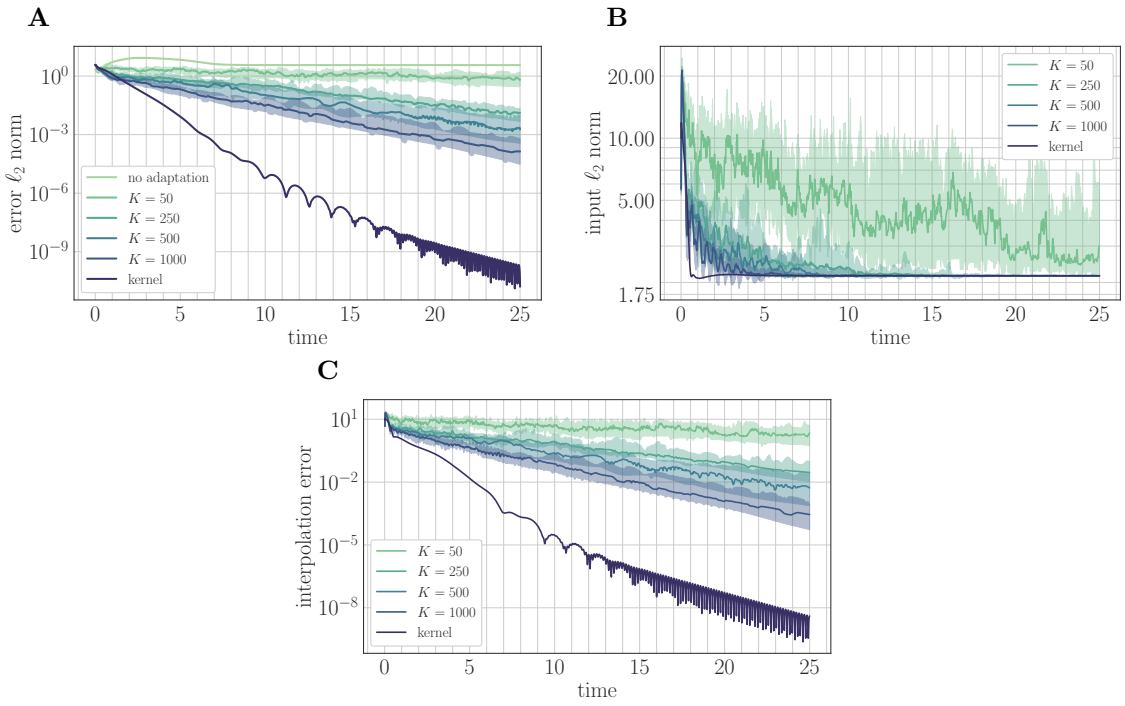

Figure 1: **Adaptive control.** (A) Tracking error as a function of time. Error bars display the 20%/80% quantiles over 20 trials (draws of the θ_i) for each choice of K. Solid lines display the median. The tracking error decreases monotonically with the number of features, and the kernel input obtains the best performance by several orders of magnitude. (B) Magnitude of the adaptive input over time. The kernel input obtains the best performance despite using the lowest input magnitude. (C) Interpolation error as a function of time. Similar to the tracking error, the interpolation error decreases monotonically with increasing K, and the kernel input obtains the best performance by several orders of magnitude.

7.2 Adaptive prediction

The infinite-dimensional input considered in Theorem 4.5 enjoys guarantees that are independent of the system dimension. As shown by (5.1), the accuracy of the random feature approximation only depends polynomially on the system dimension. These observations suggest that the nonparametric input and its randomized approximations should scale well to high-dimensional systems.

Failures of uniform gridding Any gridding-based approach must depend *exponentially* on the system dimension, and as a consequence suffers from the curse of dimensionality. Modern robotic systems, for instance, often have state dimension in the twenties or thirties, which renders such approaches inapplicable for robotic control. For illustration, consider a uniform gridding method as suggested by the calculations in Sanner and Slotine (1992). For a nine-dimensional system, placing only ten basis functions in each direction would require one billion total basis functions, a computationally and statistically intractable number. Here, we study the efficiency of our randomized method in forming a predictive model of a sixty-dimensional system, and find that our randomized approach leads to good accuracy.

m-body system Consider a system of m point masses interacting via Newtonian gravitation in d dimensions, and denote by $q_i \in \mathbb{R}^d$ the position of mass i and $p_i \in \mathbb{R}^d$ the momentum of mass i.

Assuming equal masses, such a system admits a Hamiltonian in non-dimensionalized units

$$H\left(\{p_i\}_{i=1}^m, \{q_i\}_{i=1}^m\right) = \sum_{i=1}^m \frac{\|p_i\|_2^2}{2} - \sum_{i<j}^m \frac{1}{\|q_i - q_j\|_2},$$

and a corresponding symplectic dynamics $\dot{q}_i = \frac{\partial H}{\partial p_i}$, $\dot{p}_i = -\frac{\partial H}{\partial q_i}$. We denote by $x = (q^\mathsf{T}, p^\mathsf{T})^\mathsf{T} \in \mathbb{R}^{2md}$ with $q \in \mathbb{R}^{md}$ and $p \in \mathbb{R}^{md}$ vectors containing the stacked q_i and p_i over i.

Hamiltonian estimation Let $\hat{q}_i \in \mathbb{R}^d$ and $\hat{p}_i \in \mathbb{R}^d$ denote estimates of the coordinates and momenta of the masses. Similar to the original offline method developed in Chen et al. (2019) and the online approach due to Boffi and Slotine (2021), consider learning a model of the Hamiltonian $\widehat{H}\left(\{\hat{p}_i\}_{i=1}^m, \{\hat{q}_i\}_{i=1}^m, t\right)$ by evolving the state estimates according to

$$\dot{\hat{q}}_i = \frac{\partial \widehat{H}}{\partial \hat{p}_i}\left(\{\hat{p}_i\}_{i=1}^m, \{\hat{q}_i\}_{i=1}^m, t\right) + k \cdot \left(q_i(t) - \hat{q}_i\right),$$
$$\dot{\hat{p}}_i = -\frac{\partial \widehat{H}}{\partial \hat{q}_i}\left(\{\hat{p}_i\}_{i=1}^m, \{\hat{q}_i\}_{i=1}^m, t\right) + k \cdot \left(p_i(t) - \hat{p}_i\right),$$

where $k > 0$ denotes a measurement gain and where $q_i(t)$ and $p_i(t)$ denote measurements of the true system state. The error signals $\tilde{q}(t) = \hat{q}_i(t) - q_i(t)$ and $\tilde{p}(t) = \hat{p}_i(t) - p_i(t)$ can be used to update the Hamiltonian estimate $\widehat{H}$ until $\hat{q}_i(t)$ and $\hat{p}_i(t)$ become consistent with $q_i(t)$ and $p_i(t)$.

Symplectic kernel Define the symplectic matrix

$$J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \text{so that} \quad \dot{x} = J\nabla_x H(x)$$

and let $\hat{x} = (\hat{q}^\mathsf{T}, \hat{p}^\mathsf{T})^\mathsf{T}$ with $\hat{q} \in \mathbb{R}^{md}$ and $\hat{p} \in \mathbb{R}^{md}$ the stacked vectors of $\hat{q}_i$ and $\hat{p}_i$ over i. We search for the Hamiltonian estimate $\widehat{H}$ over an RKHS $\mathcal{H}_\mathsf{k}$ corresponding to a scalar-valued translation-invariant kernel $\mathsf{k} : \mathbb{R} \to \mathbb{R}$. Similar to the curl-free kernel seen in Section 5.2, we define the *symplectic kernel*

$$\mathsf{K}(x, y) = -J\nabla^2 \mathsf{k}(x - y)J^\mathsf{T}, \tag{7.2}$$

which describes the RKHS corresponding to the dynamics $J\nabla_{\hat{x}}\widehat{H}(\hat{x})$ for $\widehat{H} \in \mathcal{H}_\mathsf{k}$. Taking $\mathsf{k}(\cdot)$ to be the Gaussian kernel, we may write (7.2) as

$$\mathsf{K}(x, y) = -J\mathbb{E}[ww^\mathsf{T} \cos(w^\mathsf{T}x + b)\cos(w^\mathsf{T}x + b)]J^\mathsf{T}$$

with the expectation taken over $w \sim \mathsf{N}(0, \sigma_w^2 I)$ and $b \sim \mathsf{Unif}(0, 2\pi)$. Let $\Psi : \mathbb{R}^{2md} \to \mathbb{R}^K$ denote a vector of random features. We may take each component $\Psi_i(\hat{x}) = \cos(w_i^\mathsf{T}\hat{x} + b_i)$ with the (w_i, b_i) i.i.d. samples and write, for $\gamma > 0$ a learning rate,

$$\widehat{H}(\hat{x}, t) = \Psi(\hat{x})^\mathsf{T}\hat{\alpha}(t),$$
$$\dot{\hat{\alpha}}(t) = -\gamma\left(\left[\nabla_{\hat{p}}\Psi(\hat{x})\right]^\mathsf{T}\tilde{q}(t) - \left[\nabla_{\hat{q}}\Psi(\hat{x})\right]^\mathsf{T}\tilde{p}(t)\right).$$

Results (Figure 2) We consider the sixty-dimensional ten body problem ($m = 10$) in three dimensions ($d = 3$). With $K = 2500$ features, the prediction and interpolation errors for the positions $\hat{q}(t) - q(t)$, momenta $\hat{p}(t) - p(t)$, and corresponding dynamics $\nabla_{\hat{p}}\widehat{H}$ and $-\nabla_{\hat{q}}\widehat{H}$ are driven to a small ball around zero (Figure 2A/B). In the early stages of learning ($t \lesssim 5$), the trajectory prediction oscillates around the target trajectory. As learning proceeds, the prediction becomes smoother and accurately tracks the true system trajectory (Figure 2C). As the number of random features K increases, the sizes of the asymptotic balls in both the prediction and interpolation errors decrease as a power law in K (Figure 2D).

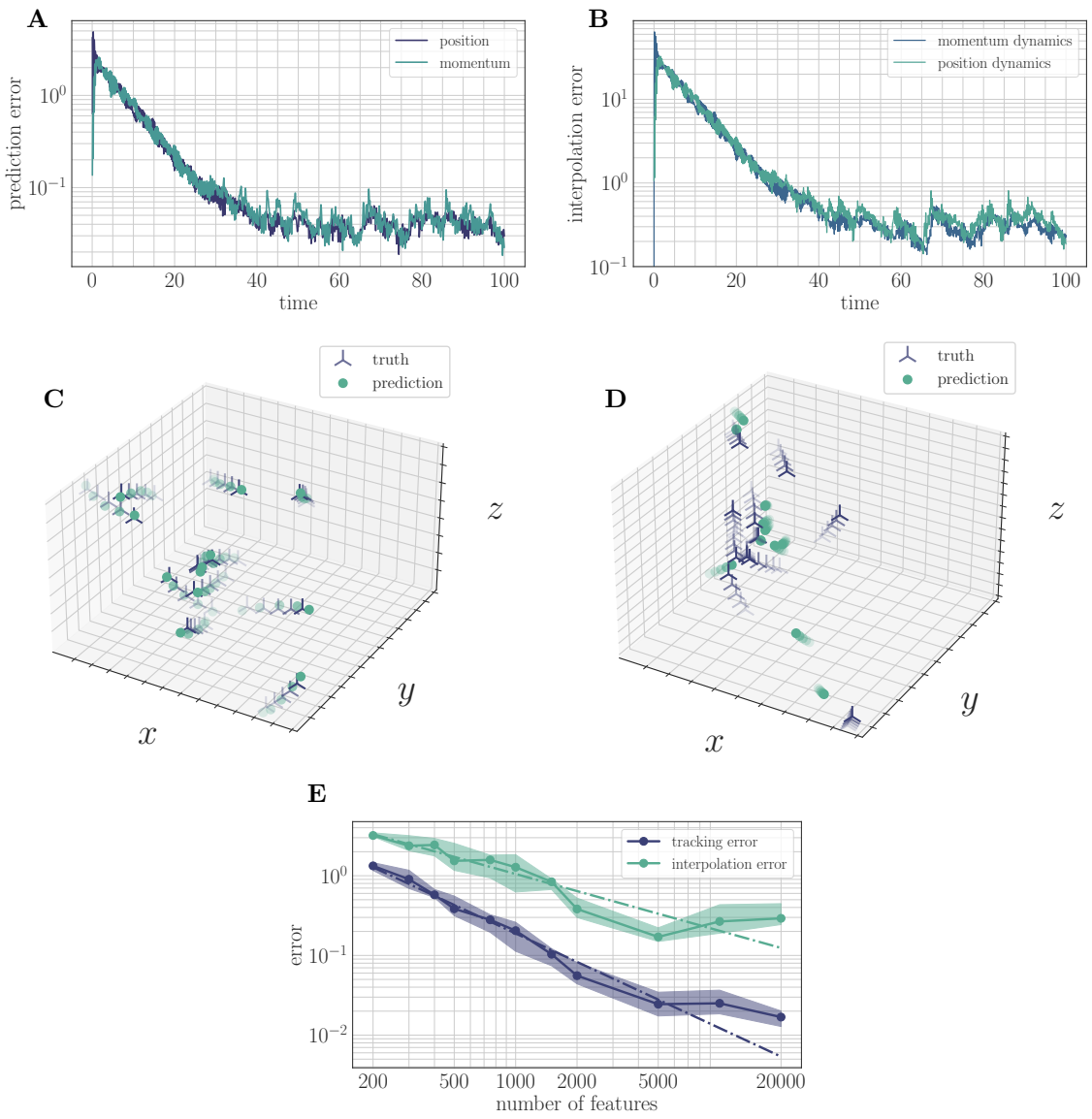

Figure 2: **Adaptive prediction.** All adaptive trajectories use $K = 2500$ unless otherwise stated. (A/B) Prediction error (A) and dynamics interpolation error (B) over time. Both errors smoothly decreases to a ball around zero. (C/D) Example position prediction trajectory for the particles with learning (C) and without learning (D). Low-opacity fade denotes particle trajectories in time. The learned system accurately predicts the ground truth, while the system without learning fails to accurately capture the particle motion. (E) Prediction and interpolation errors at $t = 100$ as a function of the number of features K (solid: median, error bars: 20% / 80% quantiles over 10 trials per K value). As K increases, the asymptotic prediction error decreases as a power law $\sim K^{-\xi}$, and the interpolation error decreases as a distinct power law $\sim K^{-\zeta}$. Best-fit power laws obtained via nonlinear least squares are shown in dashed with $\xi \approx 1.28 \pm 0.03$ and $\zeta \approx 0.77 \pm 0.03$.

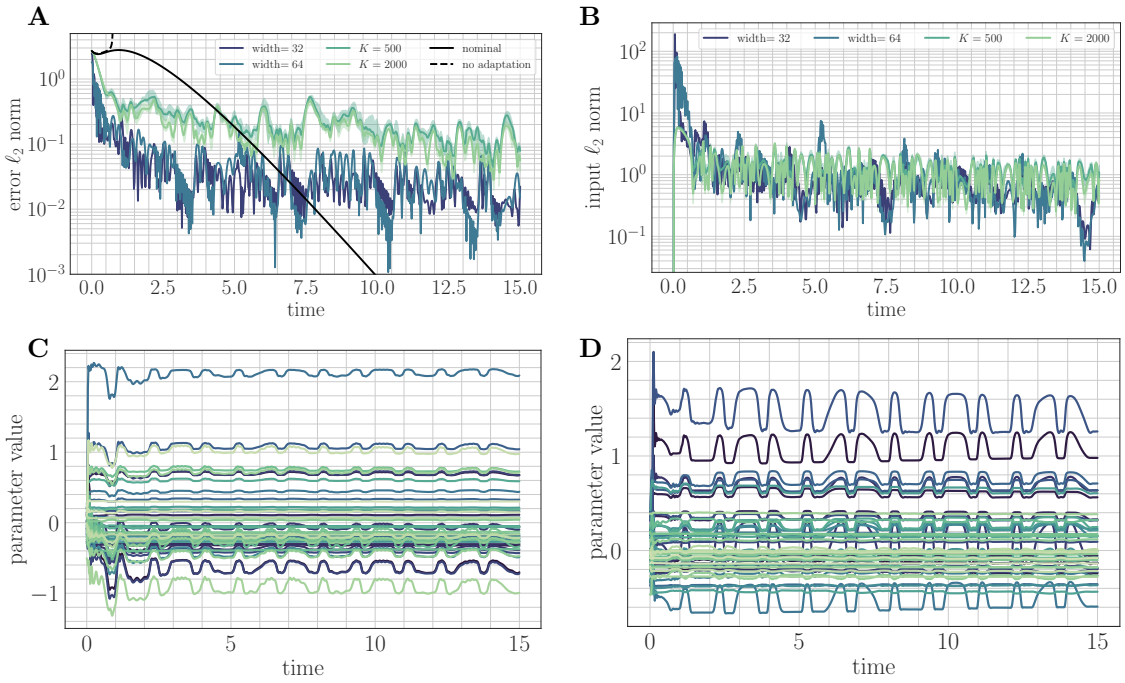

Figure 3: **Adaptive control with multilayer networks.** (A) Tracking error $\|x(t) - x_d(t)\|_2$ for the system without adaptation, the nominal dynamics, an adaptive system with a neural network approximation, and an adaptive system with a random feature approximation. The system without adaptation is driven unstable, while the adaptive systems with neural network and random feature approximations both regulate the actual trajectory to a ball around the desired trajectory. (B) Input norm $\|u(x(t), t)\|_2$. The neural network system undergoes an initial transient with large input. (C/D) Linear (C) and hidden layer (D) weights over time for the multilayer representation with width of 32 neurons. The hidden layer weights change significantly from their initialization, indicating that the network is operating outside of the kernel regime.

Power law exponents Let ξ denote the exponent $\limsup_{t\to\infty} \|\hat{x}(t) - x(t)\|_2 \sim K^{-\xi}$ in the power law for the prediction error, and let ζ denote an analogous quantity for the interpolation error. Nonlinear least-squares fits lead to estimates ($\pm$ denotes 95% confidence intervals) $\xi \approx 1.28 \pm 0.03$ and $\zeta \approx 0.77 \pm 0.03$. For the adaptive predictor considered in this section, a Lyapunov function for the nominal error dynamics is the quadratic $V(t) = \|e(t)\|_2^2$. Moreover, due to the feedback term $k \cdot (x(t) - \hat{x}(t))$, the nominal dynamics is exponentially stable with rate k, and we may take $\rho(\|e\|_2) \propto \|e\|_2^2$. This setting was considered in Example 6.5 and leads to the analytical predictions $\xi = 1/2$ and $\zeta = 1/4$, where $\zeta = \xi/2$ follows after an application of Theorem 6.6. The rates we obtain empirically are faster than the $\mathcal{O}\left(1/\sqrt{K}\right)$ Monte-Carlo rate for random feature approximations predicted by our theory. One plausible explanation for this observation is that more features are required to see the $\mathcal{O}\left(1/\sqrt{K}\right)$ tail behavior, as suggested by the flattening of the curve observed near $K \approx 20,000$.

7.3 Adaptive control with deep neural networks

The preceding sections established the adaptive controllability of high-dimensional systems via randomized approximations of kernel machines. These random feature methods can be viewed as linearizations of neural networks (Ghorbani et al., 2020a; Jacot et al., 2018), and have been shown to suffer from the curse of dimensionality when the target function only depends on a few relevant directions in the input space (Bach, 2017; Ghorbani et al., 2020b). Neural networks with a single hidden layer do not exhibit the same difficulties, which raises the question if it is possible to use deep neural networks for adaptive control.

A gradient flow The Lyapunov-based adaptive law (4.2) that forms the basis for the learning rules in Theorems 6.4 & 6.7 may also be written as the gradient flow (Fradkov et al., 1999)

$$\dot{\hat{\alpha}}(t) = -\gamma \nabla_{\hat{\alpha}} \dot{Q}(e(t), \hat{\alpha}(t), t). \tag{7.3}$$

In (7.3), $\dot{Q}(e(t), \hat{\alpha}, t)$ denotes the time derivative of the Lyapunov function Q for the nominal error dynamics along the actual error trajectory $e(t)$. This formulation of the method shows that the parameters are updated to promote stability by enforcing negativity of $\dot{Q}(e(t), \hat{\alpha}(t), t)$. Importantly, (7.3) leads to a simple algorithm for non-linearly parameterized function approximators such as neural networks. Choosing as adaptive input $u(x,t) = \varphi(x, t, \hat{\alpha}(t))$ a neural network and omitting time arguments for brevity, (7.3) becomes

$$\dot{\hat{\alpha}} = -\gamma \left(\nabla_{\hat{\alpha}} \varphi(x, t, \hat{\alpha})\right)^{\mathsf{T}} g_e(x,t)^{\mathsf{T}} \nabla Q(e,t). \tag{7.4}$$

In (7.4), the linear basis functions $Y(x,t)$ from (4.2) have been replaced by the Jacobian of the neural network evaluated at the current parameter estimates.

In general, it is challenging to obtain practical theoretical guarantees for (7.4) due to the non-convexity, time-dependence, and feedback properties of the resulting online optimization problem. Nevertheless, for expressive classes of functions that empirically exhibit benign optimization landscapes such as neural networks, it is plausible that there exists a target set of parameters α that can render $\dot{Q}(e(t), \alpha, t)$ negative definite and that these parameters can be found via gradient-based optimization.

An unstable system To test the adaptive law (7.4), we study a more difficult variant of (7.1) with a time-varying desired trajectory and an unknown dynamics that renders the system unstable in the absence of adaptation,

$$\begin{aligned} \dot{x} &= A(x - x_d(t)) + \dot{x}_d(t) + u(x,t) - h(x), \quad x \in \mathbb{R}^5, \\ x_d(t) &= \sin\left(2\pi t + \cos\left(\sqrt{2}\pi t\right)\right), \\ h_i(x) &= \frac{1}{4} x_i^4, \end{aligned} \tag{7.5}$$

where A is a known stable matrix and $x_d(t)$ denotes the desired trajectory. We take $Q(x,t) = \frac{1}{2}\left(x - x_d(t)\right)^{\mathsf{T}} P \left(x - x_d(t)\right)$ with $A^{\mathsf{T}}P + PA = -I$ as in Section 7.1. We consider single hidden-layer neural networks with width of 32 or 64 neurons and the `swish` activation function. For comparison, we use the same random Fourier feature approximation of the Gaussian kernel as in Section 7.1. We set $\gamma = 20$ for the random feature adaptation law and $\gamma = 10$ for the neural network.

Results (Figure 3) Both the random feature and neural network representations effectively stabilize the system and regulate the actual trajectory to a ball around the desired trajectory (Figure 3A). The neural network obtains slightly improved performance over the random feature method for both choices of the width despite having similar or fewer parameters. This can be traced to learning of the hidden-layer weights, which indicates that the network is operating outside of the kernel regime (Figure 3C/D).

In exchange for this improved performance, we find that the neural network adaptation law (7.4) is significantly more brittle to choice of hyperparameters: while the random feature approximation is provably stable for any choice of learning rate, the neural network adaptation law empirically renders the system unstable for many choices of learning rate. Moreover, we find that the stable range of learning rates depends on the network architecture. Increasing the network depth without careful tuning of the learning rate often leads to instability. Signatures of this phenomenon can be seen in the input norm $\|u(x,t)\|_2$ even for trajectories that remain stable (Figure 3B), where the magnitude of the neural network input is seen to exceed that of the random feature input by one or two orders of magnitude during an initial transient.

Discussion These observations are consistent with both the theory presented in this work and the approximation properties of neural networks. Neural networks, in principle, can perform better than kernel methods due to greater expressivity. Nevertheless, due to the worst-case difficulty of the corresponding online nonconvex optimization, a stability proof as provided in this work for kernel methods is likely out of reach. For adaptive control systems where stability of the closed-loop dynamics is necessary, these considerations may render kernel methods a more desirable choice than deeper architectures. Nevertheless, understanding if the adaptive law (7.4) can be modified to ensure stability of the closed-loop dynamics – or, to the same end, if a neural network-based adaptive system can be augmented with a kernel-based approach – are interesting directions of future research.

8. Conclusions and future directions

In this work, we introduced a novel nonparametric method for adaptive control and prediction that estimates the unknown dynamics over a reproducing kernel Hilbert space. By restricting to the space $\mathcal{F}_2$, we analyzed efficient finite-dimensional randomized approximations that scale well to high dimension. A promising future direction of work is to study the Banach space $\mathcal{F}_1$ of single-layer neural networks of the form $h(\cdot) = \int_\Theta \Phi(\cdot,\theta)\mu(d\theta)$ for a signed Radon measure μ (Bach, 2017; Bengio et al., 2006). The space $\mathcal{F}_1$ admits convergence analyses for gradient-based optimization via the theory of Wasserstein gradient flows, as well as efficient approximation via particle methods (Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2019). Such approaches could in principle be generalized to the adaptive control setting considered here via the gradient flow algorithm (7.4).

9. Acknowledgments

NMB thanks Eric Vanden-Eijnden and Joan Bruna for many instructive discussions on the function spaces $\mathcal{F}_1$ and $\mathcal{F}_2$. All authors thank Pannag Sanketi and Vikas Sindhwani for helpful feedback.

References

Anuradha M. Annaswamy, Fredrik P. Skantze, and Ai-Poh Loh. Adaptive control of continuous time systems with convex/concave parametrization. *Automatica*, 34(1):33–49, 1998.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Neural Information Processing Systems*, 2019.

Alessandro Astolfi and Romeo Ortega. Immersion and invariance: a new tool for stabilization and adaptive control of nonlinear systems. *IEEE Transactions on Automatic Control*, 48(4):590–606, 2003.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Yoshua Bengio, Nicolas Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In *Neural Information Processing Systems*, 2006.

Parag Bobade, Suprotim Majumdar, Savio Pereira, Andrew J. Kurdila, and John B. Ferris. Adaptive estimation for nonlinear systems using reproducing kernel hilbert spaces. *Advances in Computational Mathematics*, 45:869–896, 2019.

Nicholas M. Boffi and Jean-Jacques E. Slotine. Implicit regularization and momentum algorithms in nonlinearly parameterized adaptive control and prediction. *Neural Computation*, 33(3):590–673, 2021.

Nicholas M. Boffi, Stephen Tu, Nikolai Matni, Jean-Jacques E. Slotine, and Vikas Sindhwani. Learning stability certificates from data. In *Conference on Robot Learning*, 2020.

Nicholas M. Boffi, Stephen Tu, and Jean-Jacques E. Slotine. Regret bounds for adaptive nonlinear control. In *Learning for Dynamics and Control*, 2021.

Romain Brault, Markus Heinonen, and Florence d'Alché-Buc. Random Fourier features for operator-valued kernels. In *Proceedings of the 8th Asian Conference on Machine Learning*, 2016.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(1):19–61, 2010.

Olivier Chapelle. Training a support vector machine in the primal. *Neural Computation*, 19(5): 1155–1178, 2007.

Zhengdao Chen, Jianyu Zhang, Martin Arjovsky, and Léon Bottou. Symplectic recurrent neural networks. *arXiv:1909.13334*, 2019.

Girish Chowdhary, Jonathan How, and Hassan Kingravi. Model reference adaptive control using nonparametric adaptive elements. In *AIAA Guidance, Navigation, and Control Conference*, 2012.

Girish Chowdhary, Hassan A. Kingravi, Jonathan P. How, and Patricio A. Vela. Bayesian nonparametric adaptive control using gaussian processes. *IEEE Transactions on Neural Networks and Learning Systems*, 26(3):537–550, 2015.

Soon-Jo Chung and Jean-Jacques E. Slotine. Cooperative robot control and concurrent synchronization of lagrangian systems. *IEEE Transactions on Robotics*, 25(3):686–700, 2009.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

Bruce K. Driver. *Analysis Tools with Examples*. 2004.

Alexander L. Fradkov, Iliya V. Miroshnik, and Vladimir O. Nikiforov. *Nonlinear and Adaptive Control of Complex Systems*. Springer, 1999.

Udaya Ghai, Elad Hazan, and Yoram Singer. Exponentiated gradient meets gradient descent. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv:1904.12191*, February 2020a.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When Do Neural Networks Outperform Kernel Methods? *arXiv:2006.13409*, June 2020b.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Neural Information Processing Systems*, 2018.

Miroslav Krstić, Ioannis Kanellakopoulos, and Petar Kokotović. *Nonlinear and Adaptive Control Design*. Wiley, 1995.

Andrew Kurdila and Yu Lei. Adaptive control via embedding in reproducing kernel hilbert spaces. In *2013 American Control Conference*, 2013.

Winfried Lohmiller and Jean-Jacques E. Slotine. On contraction analysis for non-linear systems. *Automatica*, 34(6):683–696, 1998.

Winfried Lohmiller and Jean-Jacques E. Slotine. Nonlinear process control using contraction theory. *AIChE Journal*, 46(3):588–596, 2000.

Brett T. Lopez and Jean-Jacques E. Slotine. Adaptive nonlinear control with contraction metrics. *IEEE Control Systems Letters*, 5(1):205–210, 2021.

Jing Lu, Steven C.H. Hoi, Jialei Wang, Peilin Zhao, and Zhi-Yong Liu. Large scale online kernel learning. *Journal of Machine Learning Research*, 17(47):1–43, 2016.

Andreas Maurer. A vector-contraction inequality for Rademacher complexities. *arXiv preprint arXiv:1605.00251*, 2016.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33), 2018.

Há Quang Minh. Operator-valued bochner theorem, fourier feature maps for operator-valued kernels, and vector-valued learning. *arXiv preprint arXiv:1608.05639*, 2016.

Quang-Cuong Pham. Analysis of discrete and hybrid stochastic systems by nonlinear contraction theory. In *2008 10th International Conference on Control, Automation, Robotics and Vision*, 2008.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems*, 2007.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Neural Information Processing Systems*, 2008a.

Ali Rahimi and Benjamin Recht. Uniform approximation of functions with random bases. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, 2008b.

Grant M. Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2019.

Robert M. Sanner and Jean-Jacques E. Slotine. Gaussian networks for direct adaptive control. *IEEE Transactions on Neural Networks*, 3(6):837–863, 1992.

Robert M. Sanner and Jean-Jacques E. Slotine. Stable adaptive control of robot manipulators using "neural" networks. *Neural Computation*, 7:753–790, 1995.

Vikas Sindhwani, Stephen Tu, and Seyed Mohammad Khansari-Zadeh. Learning contracting vector fields for stable imitation learning. *arXiv preprint arXiv:1804.04878*, 2018.

Sumeet Singh, Spencer M. Richards, Vikas Sindhwani, Jean-Jacques E. Slotine, and Marco Pavone. Learning stabilizable nonlinear dynamics with contraction-based regularization. *The International Journal of Robotics Research*, 40(10–11):1123–1150, 2020.

Jean-Jacques E. Slotine and Weiping Li. On the adaptive control of robot manipulators. *The International Journal of Robotics Research*, 6(3), 1987.

Ivan Y. Tyukin, Danil V. Prokhorov, and Cees van Leeuwen. Adaptation and parameter estimation in systems with unstable target dynamics and nonlinear parametrization. *IEEE Transactions on Automatic Control*, 52(9):1543–1559, 2007.

Appendix A. Discrete sampling

Assume that measurements of the true system state $\{x(t_i)\}_{i=0}^{\infty}$ are received at potentially non-uniformly spaced intervals $t_i = t_0 + \sum_{i'=0}^{i-1} \Delta t_{i'}$. Denote $\hat{x}_i = \hat{x}(t_i)$ and let $\phi_{t_i+\Delta t_i}(\hat{x}_i)$ denote the flow from time t_i to time $t_{i+1} = t_i + \Delta t_i$ of the system $\dot{\hat{x}} = f(\hat{x}, t)$ starting at $\hat{x}(t_i) = \hat{x}_i$. We are interested in the contraction properties of the hybrid system

$$\hat{x}_{i+1/2} = \phi_{t_i+\Delta t_i}(\hat{x}_i), \;\; \hat{x}_{i+1} = k_i(\hat{x}_{i+1/2}, x_{i+1}),$$

where x_{i+1} denotes the measurement $x(t_{i+1})$. The following result is similar to Lohmiller and Slotine (2000, Eq. 6).

Proposition A.1. *Suppose that there exists some $\Theta : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{n\times n}$ and $0 < \beta < 1$ such that $y_{i+1} = k_i(y_i, x)$ is contracting as a discrete-time dynamical system with rate β for any x, i.e.,*

$$F_i := \Theta(y_{i+1}, t_{i+1}) \frac{\partial k_i}{\partial y}(y_i, x)\Theta(y_i, t_i)^{-1}, \;\; F_i^{\mathsf{T}} F_i \preccurlyeq \beta I.$$

Assume that $k_i(x, x) = x$ for all $x \in \mathbb{R}^n$, and denote by

$$\bar{\lambda}_i = \sup_{t\in[t_i,t_{i+1}]} \lambda_{\max}\left\{ \mathsf{Sym}\left(\dot{\Theta}(\hat{x}(t), t) + \Theta(\hat{x}(t), t)\frac{\partial \hat{f}}{\partial \hat{x}}(\hat{x}(t), t)\Theta(\hat{x}(t), t)^{-1} \right) \right\}$$

the maximum expansion rate of the open loop dynamics between t_i and t_{i+1} in the metric $M(\hat{x}, t) = \Theta(\hat{x}, t)^{\mathsf{T}}\Theta(\hat{x}, t)$. Then the Riemannian energy in the metric $M(\hat{x}, t)$ obeys

$$E(\hat{x}_{i+1}, x_{i+1}) \leqslant \beta e^{\bar{\lambda}_i \Delta t_i} E(\hat{x}_i, x_i).$$

Proof Let $M_i(\cdot) = M(\cdot, t_i)$. Let $t_{i+1/2} = t_i + \Delta t_i^-$ denote the instant before the measurement. Let $\gamma^{i+1} : [0, 1] \to \mathbb{R}^n$ denote a geodesic in the metric $M_{i+1}(\cdot)$ between $\hat{x}_{i+1}$ and x_{i+1}, and let $\gamma_s^{i+1} = \frac{d}{ds}\gamma^{i+1}$. The Riemannian energy under the metric M_{i+1} is then

$$E(\hat{x}_{i+1}, x_{i+1}) = \int_0^1 \gamma_s^{i+1}(s)^{\mathsf{T}} M_{i+1}\left(\gamma^{i+1}(s)\right)\gamma_s^{i+1}(s) ds.$$

Now let $\gamma^{i+1/2} : [0, 1] \to \mathbb{R}^n$ denote a geodesic in the metric $M_{i+1/2}(\cdot)$ between $\hat{x}_{i+1/2}$ and x_{i+1}. Observe that because $k_i(x, x) = x$ for all x, $\hat{x}_{i+1/2} = x_{i+1}$ is a fixed point. Then, by contraction of $k(\hat{x}, x)$ in $\hat{x}$ with rate $0 < \beta < 1$ (cf. Lohmiller and Slotine (1998), Pham (2008)):

$$\int_0^1 \gamma_s^{i+1}(s)^{\mathsf{T}} M_{i+1}\left(\gamma^{i+1}(s)\right)\gamma_s^{i+1}(s)\, ds \leqslant \beta \int_0^1 \gamma_s^{i+1/2}(s)^{\mathsf{T}} M_{i+1/2}\left(\gamma^{i+1/2}(s)\right)\gamma_s^{i+1/2}(s)\, ds.$$

Let $\gamma^i : [0,1] \to \mathbb{R}^n$ denote a geodesic in the metric $M(\hat{x},t)$ between $\hat{x}_i$ and x_i, let $\psi_t(s) = \Theta\left(\gamma^i(s),t\right)\gamma_s^i(s)$, and define

$$J_t(s) = \dot{\Theta}\left(\gamma^i(s),t\right) + \Theta\left(\gamma^i(s),t\right)\frac{\partial f}{\partial x}\left(\gamma^i(s),t\right)\Theta\left(\gamma^i(s),t\right)^{-1}.$$

Observe that $\frac{d}{dt}\psi_t(s) = J_t(s)\psi_t(s)$. Hence,

$$\begin{aligned}\frac{d}{dt}\left[\gamma_s^i(s)^{\mathsf{T}} M\left(\gamma^i(s),t\right)\gamma_s^i(s)\right] &= 2\psi_t(s)^{\mathsf{T}} J_t(s)\psi_t(s)\\ &\leqslant \bar{\lambda}\psi_t(s)^{\mathsf{T}}\psi_t(s)\\ &= \bar{\lambda}_i\gamma_s^i(s)^{\mathsf{T}} M\left(\gamma^i(s),t\right)\gamma_s^i(s).\end{aligned}$$

Then, by the comparison lemma,

$$\gamma_s^{i+1/2}(s)^{\mathsf{T}} M_{i+1/2}\left(\gamma^{i+1/2}(s)\right)\gamma_s^{i+1/2}(s) \leqslant e^{\bar{\lambda}_i\Delta t_i}\gamma_s^i(s)^{\mathsf{T}} M_i\left(\gamma^i(s)\right)\gamma_s^i(s).$$

Plugging this in to our previous bound,

$$E(\hat{x}_{i+1},x_{i+1}) \leqslant \beta e^{\bar{\lambda}_i\Delta t_i}\int_0^1 \gamma_s^i(s)^{\mathsf{T}} M_i\left(\gamma^i(s)\right)\gamma_s^i(s)\,ds.$$

Observing that $\int_0^1 \gamma_s^i(s)^{\mathsf{T}} M\left(\gamma^i(s),t\right)\gamma_s^i(s)ds = E(\hat{x}_i,x_i)$ completes the proof. ■

Appendix B. Preliminary results

Let E and E' be normed vector spaces. Denote by $\mathcal{L}(E,E')$ the space of linear operators from E to E' equipped with the operator norm. A function $f(x,t)$ mapping $E \times \mathbb{R}_{\geqslant 0} \mapsto F$ with E and F normed vector spaces is said to be locally bounded in x if for every $R > 0$ and $T > 0$,

$$\sup_{t\in[0,T]}\sup_{\|x\|_E\leqslant R}\|f(x,t)\|_F < \infty.$$

Proposition B.1. *Let $\{E_i\}_{i=1}^2$, $\{F_i\}_{i=1}^2$ be normed vector spaces and let $f_i : E_i \times \mathbb{R}_{\geqslant 0} \to F_i$ for $i \in \{1,2\}$ be locally Lipschitz. Then the following hold*

(i) If $E_1 = E_2$ and $F_1 = F_2$, then the sum $(x,t) \mapsto f_1(x,t) + f_2(x,t)$ is locally Lipschitz.

(ii) If $E_1 = E_2$, $F_1 = \mathcal{L}(F_2,F_3)$, and both f_1 and f_2 are locally bounded, then the product $(x,t) \mapsto f_1(x,t)f_2(x,t)$ is locally Lipschitz and locally bounded.

(iii) If $F_1 = E_2$ and both f_1 and f_2 are locally bounded, then the composition $(x,t) \mapsto f_2(f_1(x,t),t)$ is locally Lipschitz and locally bounded.

Proof Let R and T be arbitrary positive constants. Let $C = C(R,T) > 0$ be a finite positive constant such that:

$$\begin{aligned}\sup_{t\in[0,T]}\|f_1(x,t) - f_1(y,t)\|_{F_1} &\leqslant C\|x-y\|_{E_1}\ \ \forall x,y \in B_{E_1}(R),\\ \sup_{t\in[0,T]}\|f_2(x,t) - f_2(y,t)\|_{F_2} &\leqslant C\|x-y\|_{E_2}\ \ \forall x,y \in B_{E_2}(R).\end{aligned}$$

Now, let $x,y \in B_{E_1}(R)$ and $t \in [0,T]$ be arbitrary.

The sum $(x,t) \mapsto f_1(x,t) + f_2(x,t)$**.** Observe that

$$\begin{aligned}\|f_1(x,t) + g_1(x,t) - (f_2(y,t) + g_2(y,t))\|_{F_1} &\leqslant \|f_1(x,t) - f_1(y,t)\|_{F_1} + \|f_2(x,t) - f_2(y,t)\|_{F_1} \\ &\leqslant 2C\|x-y\|_{E_1}.\end{aligned}$$

The product $(x,t) \mapsto f_1(x,t)f_2(x,t)$**.** Let $C' = C'(R,T) > 0$ be a finite positive constant such that:

$$\sup_{t\in[0,T]} \sup_{\|x\|_{E_1}\leqslant R} \max\{\|f_1(x,t)\|_{\mathcal{L}(F_2,F_3)}, \|f_2(x,t)\|_{F_2}\} \leqslant C'.$$

Now observe that

$$\begin{aligned}\|f_1(x,t)f_2(x,t) - f_1(y,t)f_2(y,t)\|_{F_3} &\leqslant \|f_1(x,t) - f_1(y,t)\|_{\mathcal{L}(F_2,F_3)}\|f_2(x,t)\|_{F_2} \\ &\quad + \|f_1(y,t)\|_{\mathcal{L}(F_2,F_3)}\|f_2(x,t) - f_2(y,t)\|_{F_2} \\ &\leqslant 2CC'\|x-y\|_{E_1}.\end{aligned}$$

The fact that the composition is locally bounded is immediate.

The composition $(x,t) \mapsto f_2(f_1(x,t),t)$**.** First, let $C' = C'(R,T)$ be such that:

$$\sup_{t\in[0,T]} \sup_{\|x\|_{E_1}\leqslant R} \|f_1(x,t)\|_{F_1} \leqslant C'.$$

Next, let $C'' = C''(R,T)$ be such that:

$$\sup_{t\in[0,T]} \|f_2(x,t) - f_2(y,t)\|_{F_2} \leqslant C''\|x-y\|_{E_2} \ \forall x,y \in B_{E_2}(C').$$

Then we have:

$$\|f_2(f_1(x,t),t) - f_2(f_1(y,t),t)\|_{F_2} \leqslant C''\|f_1(x,t) - f_1(y,t)\|_{F_1} \leqslant CC''\|x-y\|_{E_1}$$

This shows that the composition is locally Lipschitz. The fact that the composition is locally bounded is immediate.

■

Now, let E and F be normed vector spaces and let $U \subseteq E$. A function $f : U \to F$ is said to be globally Lipschitz (uniformly Lipschitz) if

$$\sup_{x,y\in U, x\neq y} \frac{\|f(x) - f(y)\|_F}{\|x-y\|_E} < \infty.$$

A function $f : U \to F$ is said to be globally bounded (uniformly bounded) if:

$$\sup_{x\in U} \|f(x)\|_F < \infty.$$

Proposition B.2. *Let $\{E_i\}_{i=1}^2$ and $\{F_i\}_{i=1}^2$ be collections of normed vector spaces. Let $f_i : U_i \to F_i$ with $U_i \subseteq E_i$ for $i \in \{1,2\}$ be globally Lipschitz. Then the following hold*

1. *If $E_1 = E_2$ and $F_1 = F_2$, then the sum $x \mapsto f_1(x) + f_2(x)$ is globally Lipschitz.*
2. *If $E_1 = E_2$, $F_1 = \mathcal{L}(F_2, F_3)$, and both f_1 and f_2 are globally bounded, then the product $x \mapsto f_1(x)f_2(x)$ is globally Lipschitz and globally bounded.*

3. *If $F_1 = E_2$ and both f_1 and f_2 are globally bounded, then the composition $(x,t) \mapsto f_2(f_1(x))$ is globally Lipschitz and globally bounded.*

Proof Nearly identical proof as Proposition B.1 ■

The following result generalizes Barbalat's lemma to the case when the limiting value of a function f only converges to a ball. We first state Barbalat's lemma, and then state our generalization.

Proposition B.3 (Barbalat's lemma)**.** *Let $f \in C^1(\mathbb{R}_{\geqslant 0}, \mathbb{R})$ satisfy $\lim_{t\to\infty} f(t) < \infty$. Further assume that f' is uniformly continuous. Then $\lim_{t\to\infty} f'(t) = 0$.*

Proposition B.4 (Generalized Barbalat's lemma)**.** *Let $f \in C^1(\mathbb{R}_{\geqslant 0}, \mathbb{R})$ satisfy $\limsup_{t\to\infty} |f(t) - \alpha| \leqslant \varepsilon$ for some $\alpha \in \mathbb{R}$ and $\varepsilon \geqslant 0$. Further assume that f' is L-Lipschitz. Then,*

$$\limsup_{t\to\infty} |f'(t)| \leqslant 2\sqrt{\varepsilon L}.$$

Proof Suppose for a contradiction that $\limsup_{t\to\infty} |f'(t)| > 2\sqrt{\varepsilon L}$. Then there exists an increasing sequence $\{t_n\}_{n\geqslant 1}$ with $t_n \to \infty$ such that $|f'(t_n)| > 2\sqrt{\varepsilon L}$ for all $n \geqslant 1$. Define $\delta := 2\sqrt{\varepsilon L}/L$. Then for any $n \geqslant 1$, we have

$$\begin{aligned}
\left| \int_{t_n}^{t_n+\delta} f'(t)dt \right| &= \left| \delta f'(t_n) + \int_{t_n}^{t_n+\delta} (f'(t) - f'(t_n))dt \right|, \\
&\geqslant \delta |f'(t_n)| - \int_{t_n}^{t_n+\delta} |f'(t) - f'(t_n)|dt, \\
&> \delta 2\sqrt{\varepsilon L} - L \int_{t_n}^{t_n+\delta} |t_n - t|dt, \\
&= \delta 2\sqrt{\varepsilon L} - \frac{L}{2}\delta^2, \\
&= 2\varepsilon.
\end{aligned}$$

This lower bound implies that for any $n \geqslant 1$

$$|f(t_n + \delta) - f(t_n)| = \left| \int_{t_n}^{t_n+\delta} f'(t)dt \right| > 2\varepsilon.$$

This bound implies

$$\begin{aligned}
2\varepsilon &< \limsup_{n\to\infty} |f(t_n + \delta) - f(t_n)|, \\
&\leqslant \limsup_{t\to\infty} |f(t + \delta) - f(t)|, \\
&\leqslant \limsup_{t\to\infty} |f(t + \delta) - \alpha| + \limsup_{t\to\infty} |f(t) - \alpha|, \\
&\leqslant 2\varepsilon,
\end{aligned}$$

which yields a contradiction. ■

In adaptive control, a typical use of Barbalat's lemma is to conclude (via deadzones) that the error signal tends to a small value. In the sequel, we will use Barbalat's lemma in conjunction with the generalized Barbalat's lemma (Proposition B.4) to argue that both the error signal and the *time derivative* of the error signal are small. The time derivative of the error signal can be written as a nominal term plus the error of the adaptive signal. By controlling this quantity, we will be able to show that the error of the adaptive signal is small as well, allowing us to prove approximate interpolation type results (Theorem 6.6).

Sharpness of the bound Proposition B.4 is sharp in the following sense. Fix any $\varepsilon > 0$ and $\omega \in \mathbb{R}$, and define $f(t) := \varepsilon \sin\left(\sqrt{\frac{\omega}{\varepsilon}} t\right)$. Clearly $\limsup_{t\to\infty} |f(t)| = \varepsilon$, and furthermore

$$f'(t) = \sqrt{\varepsilon\omega} \cos\left(\sqrt{\frac{\omega}{\varepsilon}} t\right), \quad f''(t) = -\omega \sin\left(\sqrt{\frac{\omega}{\varepsilon}} t\right).$$

This shows that the smallest valid global Lipschitz constant for f' is ω. Furthermore,

$$\limsup_{t\to\infty} |f'(t)| = \sqrt{\varepsilon\omega}.$$

Appendix C. Omitted proofs for Section 4

C.1 Proof of Theorem 4.5

We first state the following technical lemma.

Lemma C.1. *Let E denote the Banach space $E := \mathbb{R}^n \times \mathbb{R}^s \times L_2(\Theta, \nu)$ equipped with the norm $\|(x, e, \hat{\alpha})\|_E := \max\{\|x\|_2, \|e\|_2, \|\hat{\alpha}\|_{L_2(\Theta,\nu)}\}$. Write $z = (x, e, \hat{\alpha})$ for $z \in E$ and define the function $F : E \times \mathbb{R}_{\geqslant 0} \to E$ as:*

$$F(z,t) := \begin{bmatrix} f(x,t) + g(x,t)\left(\int_\Theta \Phi(x,\theta)\hat{\alpha}(\theta) d\nu(\theta) - h(x)\right) \\ f_e(e,t) + g_e(x,t)\left(\int_\Theta \Phi(x,\theta)\hat{\alpha}(\theta) d\nu(\theta) - h(x)\right) \\ -\gamma \Phi(x,\cdot)^\mathsf{T} g_e(x,t)^\mathsf{T} \nabla Q(e,t) \end{bmatrix}.$$

Then, under Assumption 4.4, $F(z,t)$ is locally Lipschitz in z with respect to $\|\cdot\|_E$. That is, for each $R > 0$ and $T > 0$, letting $B_E(R) := \{z \in E : \|z\|_E \leqslant R\}$,

$$\sup_{t\in[0,T]} \sup_{z_1, z_2 \in B_E(R)} \frac{\|F(z_1,t) - F(z_2,t)\|_E}{\|z_1 - z_2\|_E} < \infty.$$

Proof By the composition rules for locally Lipschitz functions (cf. Proposition B.1), it suffices to show that the functions $\psi_1 : E \to \mathbb{R}^d$ and $\psi_2 : E \to \mathcal{L}(\mathbb{R}^{d_1}, L_2(\Theta, \nu))$ defined by

$$\begin{aligned} \psi_1((x,e,\hat{\alpha})) &:= \int_\Theta \Phi(x,\theta)\hat{\alpha}(\theta) d\nu(\theta), \\ \psi_2((x,e,\hat{\alpha}))(q) &:= \Phi(x,\cdot)^\mathsf{T} q \ \ \forall q \in \mathbb{R}^{d_1}. \end{aligned}$$

are locally Lipschitz and locally bounded. We view both ψ_1 and ψ_2 as functions defined on E, consistent with their appearance in the definition of $F(z,t)$; however, clearly ψ_1 is independent of e and ψ_2 is independent of both e and $\hat{\alpha}$.

Because ψ_1 and ψ_2 do not depend on time t, locally Lipschitz implies locally bounded. We first show that ψ_1 is locally Lipschitz. Fix an $R > 0$ and let $z_1 = (x_1, e_1, \hat{\alpha}_1)$, $z_2 = (x_2, e_2, \hat{\alpha}_2)$ be contained in $B_E(R)$. By Assumption 4.4, there exists a $C = C(R) > 0$ such that the following conditions hold:

$$\begin{aligned} \sup_{x\in B_2^n(R)} \int_\Theta \|\Phi(x,\theta)\|_{\mathrm{op}}^2 d\nu(\theta) &\leqslant C^2, \\ \int_\Theta \|\Phi(x_1,\theta) - \Phi(x_2,\theta)\|_{\mathrm{op}}^2 d\nu(\theta) &\leqslant C^2 \|x_1 - x_2\|_2^2 \ \ \forall x_1, x_2 \in B_2^n(R). \end{aligned}$$

By the triangle inequality and Cauchy-Schwarz,

$$\begin{aligned}
&\|\psi_1(z_1) - \psi_1(z_2)\|_2 \\
&\leqslant \sqrt{\int_\Theta \|\Phi(x_1,\theta) - \Phi(x_2,\theta)\|_{\mathrm{op}}^2 d\nu(\theta)} \sqrt{\int_\Theta \|\hat{\alpha}_1(\theta)\|_2^2 d\nu(\theta)} \\
&\quad + \sqrt{\int_\Theta \|\Phi(x_1,\theta)\|_{\mathrm{op}}^2 d\nu(\theta)} \sqrt{\int_\Theta \|\hat{\alpha}_1(\theta) - \hat{\alpha}_2(\theta)\|_2^2 d\nu(\theta)} \\
&\leqslant CR\|x_1 - x_2\|_2 + C\|\alpha_1 - \alpha_2\|_{L_2(\Theta,\nu)} \\
&\leqslant C(1+R)\|z_1 - z_2\|_E.
\end{aligned}$$

This shows that ψ_1 is locally Lipschitz. To show that ψ_2 is locally Lipschitz, by Cauchy-Schwarz,

$$\begin{aligned}
\|\psi_2(z_1) - \psi_2(z_2)\|_{\mathcal{L}(\mathbb{R}^{d_1}, L_2(\Theta,\nu))} &= \sup_{\|q\|_2=1} \|(\Phi(x_1,\cdot) - \Phi(x_2,\cdot))^{\mathsf{T}} q\|_{L_2(\Theta,\nu)} \\
&= \sup_{\|q\|_2=1} \left(\int_\Theta \|(\Phi(x_1,\theta) - \Phi(x_2,\theta))^{\mathsf{T}} q\|_2^2 d\nu(\theta) \right)^{1/2} \\
&\leqslant \left(\int_\Theta \|\Phi(x_1,\theta) - \Phi(x_2,\theta)\|_{\mathrm{op}}^2 d\nu(\theta) \right)^{1/2} \\
&\leqslant C\|x_1 - x_2\|_2 \leqslant C\|z_1 - z_2\|_E.
\end{aligned}$$

■

We now require the following result concerned with existence and uniqueness of solutions to ordinary differential equations defined on Banach spaces. This result will be used in conjunction with Lemma C.1 to assert the existence of our nonparametric input over an interval of time. Via a Lyapunov argument, we can then extend the interval to infinity.

Proposition C.2 (Existence of a maximal solution (see e.g., Proposition 11.8 of Driver (2004)))**.** *Let E be a Banach space, U be an open subset of E, $T \subseteq \mathbb{R}$ be an interval of time containing 0, and $F : U \times T \to E$ be a continuous vector field on E. Assume that F is locally Lipschitz in the following sense. For every $x_0 \in U$ and compact $I \subseteq T$, there exists finite positive $L = L(x_0, I)$ and $R = R(x_0, I)$ such that:*

$$\sup_{t\in I} \|f(x,t) - f(y,t)\|_E \leqslant \|x - y\|_E \ \ \forall x, y \in B_E(x_0, R).$$

Then for each $x_0 \in U$, there exists a maximal interval $I(x_0) = (a(x_0), b(x_0)) \subseteq T$ with $a(x_0) \in [-\infty, 0)$ and $b(x_0) \in (0, +\infty]$ such that the ordinary differential equation

$$\dot{x}(t) = F(x(t), t), \quad x(0) = x_0$$

has a unique continuously differentiable solution $x : I(x_0) \to U$.

We may now state our proof of the main nonparametric theorem.

Theorem 4.5 (Convergence)**.** *Consider system* (3.1) *under Assumptions 3.7, 4.3, and 4.4. Fix $\alpha_p = 0$ and let $\gamma > 0$. Then the adaptive control input*

$$u(x,t) = -\gamma \int_0^t \mathsf{K}(x, x(\tau)) g_e(x(\tau),\tau)^{\mathsf{T}} \nabla Q(e(\tau),\tau) d\tau$$

ensures that both $x(t)$ and $e(t)$ exist and are uniformly bounded for all $t \geqslant 0$. Moreover, $u(\cdot, t) \in \mathcal{H}$ for all $t \geqslant 0$ and $\lim_{t\to\infty} \|x(t) - x_d(t)\|_2 = 0$.

Proof By Assumption 4.3, there exists a signed density $\alpha(\theta) \in L_2(\Theta, \nu)$ such that

$$h(\cdot) = \int_\Theta \Phi(\cdot, \theta)\alpha(\theta)d\nu(\theta), \;\; \|h\|_{\mathcal{H}}^2 = \|\alpha\|_{L_2(\Theta,\nu)}^2.$$

Define the signed density $\hat{\alpha} : \Theta \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^{d_1}$ by $\hat{\alpha}(\cdot, 0) = 0$ and the pointwise update for $\theta \in \Theta$,

$$\frac{\partial \hat{\alpha}}{\partial t}(\theta, t) = -\gamma \Phi(x(t), \theta)^\mathsf{T} g_e(x(t), t)^\mathsf{T} \nabla Q(e(t), t).$$

Observe that by Lemma C.1 and Proposition C.2, there exists some maximal $T_{\max} \in (0, \infty]$ such that the curve $t \mapsto (x(t), e(t), \hat{\alpha}(t))$ exists, is unique, and is continuously differentiable. Moreover, we may write the input as

$$u(x, t) = \int_\Theta \Phi(x, \theta)\hat{\alpha}(\theta, t)d\nu(\theta).$$

By means of contradiction, let us suppose that $T_{\max} < \infty$. For $t \in [0, T_{\max})$, define $\tilde{\alpha}(\theta, t) := \hat{\alpha}(\theta, t) - \alpha(\theta)$ so that

$$u(\cdot, t) - h(\cdot) = \int_\Theta \Phi(\cdot, \theta)\tilde{\alpha}(\theta, t)d\nu(\theta).$$

Now consider the Lyapunov-like function $V : [0, T_{\max}) \to \mathbb{R}$,

$$V(t) = Q(e(t), t) + \frac{1}{2\gamma}\|\tilde{\alpha}(\cdot, t)\|_{L_2(\Theta,\nu)}^2.$$

We note that because $L_2(\Theta, \nu)$ is a real Hilbert space, the map $u \mapsto \|u\|_{L_2(\Theta,\nu)}^2$ is (Fréchet) differentiable with derivative $h \mapsto 2\langle u, h\rangle_{L_2(\Theta,\nu)}$ Therefore, by the differentiability of the curve $t \mapsto \tilde{\alpha}(\cdot, t)$ and the chain rule, we have:

$$\frac{d}{dt}\int_\Theta \|\tilde{\alpha}(\theta, t)\|_2^2 d\nu(\theta) = 2\left\langle \tilde{\alpha}(\cdot, t), \frac{\partial \hat{\alpha}}{\partial t}(\cdot, t)\right\rangle_{L_2(\Theta,\nu)}.$$

Computing the time derivative, for any $t \in [0, T_{\max})$,

$$\begin{aligned}
\dot{V}(t) &= \frac{\partial Q}{\partial t}(e(t), t) + \nabla Q(e(t), t)^\mathsf{T}\left(f_e(e(t), t) + g_e(x(t), t)\left(u(x(t), t) - h(x(t))\right)\right) \\
&\quad + \frac{1}{\gamma}\left\langle \tilde{\alpha}(\cdot, t), \frac{\partial \hat{\alpha}}{\partial t}(\cdot, t)\right\rangle_{L_2(\Theta,\nu)}, \\
&\leqslant -\rho\left(\|e(t)\|_2\right) + \nabla Q(e(t), t)^\mathsf{T} g_e(e(t), t)\left(u(x(t), t) - h(x(t))\right) \\
&\quad + \frac{1}{\gamma}\left\langle \tilde{\alpha}(\cdot, t), \frac{\partial \hat{\alpha}}{\partial t}(\cdot, t)\right\rangle_{L_2(\Theta,\nu)},
\end{aligned}$$

where we have applied Assumption 3.7. Now, observe that

$$\left\langle \tilde{\alpha}(\cdot, t), \frac{\partial \hat{\alpha}}{\partial t}(\cdot, t)\right\rangle_{L_2(\Theta,\nu)} = -\gamma\int_\Theta \langle \tilde{\alpha}(\theta, t), \Phi(x(t), \theta)^\mathsf{T} g_e(x(t), t)\nabla Q(e(t), t)\rangle d\nu(\theta)$$

so that the last two terms in $\dot{V}(t)$ cancel, and hence:

$$\dot{V}(t) \leqslant -\rho(\|e(t)\|_2).$$

Now, because $\dot{V}(t) \leqslant 0$ for all $t \in [0, T_{\max})$, $V(t) \leqslant V(0)$. Therefore, since $Q(e(t), t) \geqslant \mu_1(\|e(t)\|_2)$ and μ_1 is a class $\mathcal{K}_\infty$ function,

$$\sup_{t\in[0,T_{\max})} \|e(t)\|_2 < \infty, \quad \sup_{t\in[0,T_{\max})} \|\hat{\alpha}(\cdot,t)\|_{L_2(\Theta,\nu)} < \infty.$$

Furthermore, $\sup_{t\in[0,T_{\max})}\|x(t) - x_d(t)\|_2 < \infty$ by requirement (3.5) on the error signal, and since x_d is uniformly bounded, we also have that $\sup_{t\in[0,T_{\max})}\|x(t)\|_2 < \infty$. This contradicts that $T_{\max}$ is finite, so we conclude that $T_{\max} = \infty$. This implies that

$$\sup_{t\geqslant 0} \max\{\|x(t)\|_2, \|e(t)\|_2, \|\tilde{\alpha}(\cdot,t)\|_{L_2(\Theta,\nu)}\} < \infty,$$

so that $u(\cdot, t) \in \mathcal{H}$ for all $t \geqslant 0$. This proves the first two claims. Now, integrating both sides of $\dot{V}(t)$,

$$\int_0^\infty \rho\left(\|e(t)\|_2\right) dt \leqslant V(0).$$

To complete the proof, we now need to show that $t \mapsto \rho(\|e(t)\|_2)$ is uniformly continuous on $[0, \infty)$ and apply Barbalat's lemma.

We first show that $e(t)$ is uniformly Lipschitz in t. To do so, we bound $\sup_{t\geqslant 0}\|\dot{e}(t)\|_2$ and apply $\|e(t_1) - e(t_2)\|_2 \leqslant \sup_{t\geqslant 0}\|\dot{e}(t)\|_2|t_1 - t_2|$. Let $C_\Phi := \sup_{t\geqslant 0}\left(\int_\Theta \|\Phi(x(t),\theta)\|_2^2 d\nu(\theta)\right)^{1/2}$. Because $x(t)$ is uniformly bounded, C_Φ is finite by Assumption 4.4. Next,

$$\|u(x(t),t) - h(x(t))\|_2 \leqslant \left(\int_\Theta \|\Phi(x(t),\theta)\|_{\text{op}}^2 d\nu(\theta)\right)^{1/2} \|\tilde{\alpha}(\cdot,t)\|_{L_2(\Theta,\nu)} \leqslant C_\Phi\sqrt{2\gamma V(0)}.$$

Now, observe that

$$\begin{aligned}\|\dot{e}(t)\|_2 &\leqslant \|f_e(e(t),t)\|_2 + \|g_e(x(t),t)\|_{\text{op}}\|u(x(t),t) - h(x(t))\|_2 \\ &\leqslant \|f_e(e(t),t)\|_2 + \|g_e(x(t),t)\|_{\text{op}} C_\Phi\sqrt{2\gamma V(0)}.\end{aligned}$$

Because both f_e and g_e are locally bounded in x uniformly in t, $\|\dot{e}(t)\|_2$ is uniformly bounded in t. Therefore, $t \mapsto \|e(t)\|_2$ is uniformly Lipschitz and $t \mapsto \|e(t)\|_2$ is uniformly continuous. Now, because ρ is continuous, it is uniformly continuous on the range of $t \mapsto \|e(t)\|_2$. Since the composition of two uniformly continuous functions remains uniformly continuous, $t \mapsto \rho(\|e(t)\|)$ is uniformly continuous. By Barbalat's lemma, this implies that $\lim_{t\to\infty}\rho(\|e(t)\|_2) = 0$. By continuity of ρ and the fact that $\rho(a) = 0$ if and only if $a = 0$, we conclude that $\lim_{t\to\infty}\|e(t)\|_2 = 0$. From the requirement (3.6) on the error signal, we conclude that $\lim_{t\to\infty}\|x(t) - x_d(t)\|_2$. ■

C.2 Proof of Theorem 4.7

Theorem 4.7 (Interpolation)**.** *Consider the setting of Theorem 4.5. Suppose furthermore that both $f_e(e,t)$ and $g_e(x,t)$ are locally Lipschitz in their first argument uniformly in t. Finally, suppose that for every $R > 0$,*

$$\int_\Theta \sup_{\|x\|_2\leqslant R} \|\Phi(x,\theta)\|_{\text{op}}^2 d\nu(\theta) < \infty.$$

Then the nonparametric input asymptotically interpolates the unknown in the span of the control matrix, $\lim_{t\to\infty}\|g_e(x(t),t)(u(x(t),t) - h(x(t)))\|_2 = 0$.

Proof Recall that the error dynamics satisfy:

$$\dot{e}(t) = f_e(e(t), t) + g_e(x(t), t)(u(x(t), t) - h(x(t))).$$

From the proof of Theorem 4.5, $\lim_{t\to\infty} e(t) = 0$. If we show in addition that $t \mapsto \dot{e}(t)$ is uniformly Lipschitz, then by Barbalat's lemma (applied to each coordinate), $\lim_{t\to\infty} \dot{e}(t) = 0$. Since $f_e(0, t) = 0$ and f_e is locally Lipschitz in e uniformly in t, $\lim_{t\to\infty} \dot{e}(t) = 0$ implies that $\lim_{t\to\infty} \|g_e(x(t), t)(u(x(t), t) - h(x(t)))\|_2 = 0$.

It remains to show that $t \mapsto \dot{e}(t)$ is uniformly Lipschitz. By the composition rule (cf. Proposition B.2), it suffices to show that the functions:

$$t \mapsto f_e(e(t), t), \;\; t \mapsto g_e(x(t), t), \;\; t \mapsto u(x(t), t), \;\; t \mapsto h(x(t)),$$

are all uniformly Lipschitz and bounded. From the proof of Theorem 4.5, both $t \mapsto e(t)$ and $t \mapsto x(t)$ are uniformly bounded, and $t \mapsto e(t)$ is uniformly Lipschitz. A nearly identical argument shows that $t \mapsto x(t)$ is also uniformly Lipschitz. Therefore, since f_e, g_e, and h are all locally Lipschitz and locally bounded uniformly in t, it is clear that $t \mapsto f_e(e(t), t)$, $t \mapsto g_e(x(t), t)$, and $t \mapsto h(x(t))$ are all uniformly Lipschitz and uniformly bounded.

To see that $t \mapsto u(x(t), t)$ is also uniformly Lipschitz, we first choose a finite constant $C > 0$ such that

$$\sup_{t\geqslant 0} \max\{\|x(t)\|_2, \|g_e(x(t), t)\|_{\mathrm{op}}, \|\nabla Q(e(t), t)\|_2\} \leqslant C.$$

Now observe that for every θ and t,

$$\begin{aligned}
\left\| \frac{\partial \hat{\alpha}}{\partial t}(\theta, t) \right\|_2 &= \gamma \left\| \Phi(x(t), \theta)^{\mathsf{T}} g_e(x(t), t)^{\mathsf{T}} \nabla Q(e(t), t) \right\|_2 \\
&\leqslant \gamma \|\Phi(x(t), \theta)\|_{\mathrm{op}} \|g_e(x(t), t)\|_{\mathrm{op}} \|\nabla Q(e(t), t)\|_2 \\
&\leqslant \gamma C^2 \|\Phi(x(t), \theta)\|_{\mathrm{op}}.
\end{aligned}$$

Put $C_\Phi := \left(\int_\Theta \sup_{\|x\|_2 \leqslant C} \|\Phi(x, \theta)\|_{\mathrm{op}}^2 d\nu(\theta) \right)^{1/2}$, which is finite by assumption. Fix t_1, t_2, and for $i \in \{1, 2\}$ define:

$$u_i := u(x(t_i), t_i), \;\; \Phi_i(\cdot) := \Phi(x(t_i), \cdot), \;\; \hat{\alpha}_i(\cdot) := \hat{\alpha}(\cdot, t_i).$$

We have:

$$\begin{aligned}
\|\hat{\alpha}_1 - \hat{\alpha}_2\|_{L_2(\Theta,\nu)} &= \left(\int_\Theta \|\hat{\alpha}(\theta, t_1) - \hat{\alpha}(\theta, t_2)\|_2^2 d\nu(\theta) \right)^{1/2} \\
&\leqslant \left(\int_\Theta \left\| \int_{t_1}^{t_2} \frac{\partial \hat{\alpha}}{\partial t}(\theta, t) dt \right\|_2^2 d\nu(\theta) \right)^{1/2} \\
&\leqslant \left(\int_\Theta \left(\int_{t_1}^{t_2} \left\| \frac{\partial \hat{\alpha}}{\partial t}(\theta, t) \right\|_2 dt \right)^2 d\nu(\theta) \right)^{1/2} \\
&\leqslant \gamma C^2 \left(\int_\Theta \sup_{\|x\|_2 \leqslant C} \|\Phi(x, \theta)\|_{\mathrm{op}}^2 d\nu(\theta) \right)^{1/2} |t_1 - t_2| \\
&\leqslant \gamma C^2 C_\Phi |t_1 - t_2|.
\end{aligned}$$

Next, let C'_Φ be a finite constant such that

$$\left(\int_\Theta \|\Phi(x,\theta)-\Phi(y,\theta)\|_{\mathrm{op}}^2 d\nu(\theta)\right)^{1/2} \leqslant C'_\Phi \|x-y\|_2 \ \ \forall x,y \in B_2^n(C).$$

Then,

$$\begin{aligned}
\|u_1-u_2\|_2 &\leqslant \int_\Theta \|\Phi_1(\theta)\hat{\alpha}_1(\theta)-\Phi_2(\theta)\hat{\alpha}_2(\theta)\|_2 d\nu(\theta) \\
&\leqslant \int_\Theta \|\Phi_1(\theta)-\Phi_2(\theta)\|_{\mathrm{op}}\|\hat{\alpha}_1(\theta)\|_2 d\nu(\theta) + \int_\Theta \|\Phi_2(\theta)\|_{\mathrm{op}}\|\hat{\alpha}_1(\theta)-\hat{\alpha}_2(\theta)\|_2 d\nu(\theta) \\
&\leqslant C\left(\int_\Theta \|\Phi(x(t_1),\theta)-\Phi(x(t_2),\theta)\|_{\mathrm{op}}^2 d\nu(\theta)\right)^{1/2} + C_\Phi \|\hat{\alpha}_1-\hat{\alpha}_2\|_{L_2(\Theta,\nu)} \\
&\leqslant CC'_\Phi \|x(t_1)-x(t_2)\|_2 + \gamma C^2 C_\Phi^2 |t_1-t_2| \\
&\leqslant (C^2 C'_\Phi + \gamma C^2 C_\Phi^2)|t_1-t_2|.
\end{aligned}$$

This shows that $t \mapsto u(x(t),t)$ is uniformly Lipschitz. To conclude, we argue that $t \mapsto u(x(t),t)$ is uniformly bounded:

$$\begin{aligned}
\|u(x(t),t)\|_2 &\leqslant \int_\Theta \|\Phi(x(t),\theta)\|_{\mathrm{op}}\|\hat{\alpha}(t)\|_2 d\nu(\theta) \\
&\leqslant \sqrt{\int_\Theta \|\Phi(x(t),\theta)\|_{\mathrm{op}}^2 d\nu(\theta)}\|\hat{\alpha}(t)\|_2 \\
&\leqslant C_\Phi \|\hat{\alpha}(t)\|_2.
\end{aligned}$$

The proof of Theorem 4.5 shows that $\|\tilde{\alpha}(t)\|_2$ is uniformly bounded, and therefore so is $\|\hat{\alpha}(t)\|_2$ by the triangle inequality. ■

C.3 Proof of Theorem 4.8

Theorem 4.8 (Implicit regularization)**.** *Consider the setting of Theorem 4.5. Define the interpolating set over the trajectory*

$$\mathcal{A} := \left\{\bar{h} \in \mathcal{H} : \bar{h}(x(t)) = h(x(t)),\ \forall t \geqslant 0\right\},$$

and assume that $\lim_{t\to\infty} u(\cdot,t) \in \mathcal{A}$*. Then,*

$$\lim_{t\to\infty} u(\cdot,t) \in \operatorname*{argmin}_{\bar{h}\in\mathcal{A}} \|\bar{h}(\cdot)\|_{\mathcal{H}}. \tag{4.8}$$

Proof From Theorem 4.5, $u(\cdot,t) \in \mathcal{H}$ for all $t \geqslant 0$. Let $\bar{h}(\cdot) \in \mathcal{H}$ be arbitrary. Then by Assumption 4.3 there exists $\bar{\alpha} \in L_2(\Theta,\nu)$ such that

$$\bar{h}(x) = \int_\Theta \Phi(x,\theta)\bar{\alpha}(\theta) d\nu(\theta).$$

Consider the Lyapunov-like function $V : \mathbb{R}_{\geqslant 0} \to \mathbb{R}$,

$$V(t) = \frac{1}{2}\|u(\cdot,t)-\bar{h}\|_{\mathcal{H}}^2 = \frac{1}{2}\|\hat{\alpha}(\cdot,t)-\bar{\alpha}\|_{L_2(\Theta,\nu)}^2,$$

where $\hat{\alpha}(\cdot, t) \in L_2(\Theta, \nu)$ was defined in the proof of Theorem 4.5 by the partial differential equation

$$\frac{\partial \hat{\alpha}}{\partial t}(\theta, t) = -\gamma \Phi(x(t), \theta)^\mathsf{T} g_e(x(t), t)^\mathsf{T} \nabla Q(e(t), t), \;\; \hat{\alpha}(\theta, 0) = 0.$$

Computing the time derivative of V,

$$\dot{V}(t) = \left\langle \hat{\alpha}(\cdot, t) - \bar{\alpha}, \frac{\partial \hat{\alpha}}{\partial t}(\cdot, t) \right\rangle_{L_2(\Theta, \nu)}.$$

Integrating both sides of the above from 0 to t,

$$\frac{1}{2}\|\hat{\alpha}(\cdot, t) - \bar{\alpha}\|^2_{L_2(\Theta,\nu)} = \frac{1}{2}\|\bar{\alpha}\|^2_{L_2(\Theta,\nu)} + \int_0^t \left\langle \hat{\alpha}(\cdot, \tau) - \bar{\alpha}, \frac{\partial \hat{\alpha}}{\partial t}(\cdot, \tau) \right\rangle_{L_2(\Theta,\nu)} d\tau,$$

Define $\hat{\alpha}_\infty(\theta)$ to be the density such that $\lim_{t\to\infty} u(\cdot, t) = \int_\Theta \Phi(\cdot, \theta)\hat{\alpha}_\infty(\theta) d\nu(\theta)$. Taking the limit as $t \to \infty$ of both sides,

$$\begin{aligned}\frac{1}{2}\|\hat{\alpha}_\infty - \bar{\alpha}\|^2_{L_2(\Theta,\nu)} &= \lim_{t\to\infty} \frac{1}{2}\|\hat{\alpha}(\cdot, t) - \bar{\alpha}\|^2_{L_2(\Theta,\nu)} \\ &= \frac{1}{2}\|\bar{\alpha}\|^2_{L_2(\Theta,\nu)} + \int_0^\infty \left\langle \hat{\alpha}(\cdot, \tau) - \bar{\alpha}, \frac{\partial \hat{\alpha}}{\partial t}(\cdot, \tau) \right\rangle_{L_2(\Theta,\nu)} d\tau. \end{aligned} \tag{C.1}$$

Now take $\bar{h}(\cdot) \in \mathcal{A}$. Observe that, by definition of $\mathcal{A}$, for any $\tau \geqslant 0$,

$$\begin{aligned}\left\langle \bar{\alpha}, \frac{\partial \hat{\alpha}}{\partial t}(\cdot, \tau) \right\rangle_{L_2(\Theta,\nu)} &= -\gamma \int_\Theta \bar{\alpha}(\theta)^\mathsf{T} \Phi(x(\tau), \theta)^\mathsf{T} g_e(x(\tau), t)^\mathsf{T} \nabla Q(e(\tau), \tau) \\ &= -\gamma \bar{h}(x(\tau))^\mathsf{T} g_e(x(\tau), \tau)^\mathsf{T} \nabla Q(e(\tau), \tau) \\ &= -\gamma h(x(\tau))^\mathsf{T} g_e(x(\tau), \tau)^\mathsf{T} \nabla Q(e(\tau), \tau).\end{aligned}$$

Hence, (C.1) may be re-written,

$$\begin{aligned}\frac{1}{2}\|\hat{\alpha}_\infty - \bar{\alpha}\|^2_{L_2(\Theta,\nu)} = \frac{1}{2}\|\bar{\alpha}\|^2_{L_2(\Theta,\nu)} &+ \int_0^\infty \left\langle \hat{\alpha}(\cdot, \tau), \frac{\partial \hat{\alpha}}{\partial t}(\cdot, \tau) \right\rangle_{L_2(\Theta,\nu)} d\tau \\ &+ \gamma \int_0^\infty h(x(\tau))^\mathsf{T} g(x(\tau), \tau)^\mathsf{T} \nabla Q(x(\tau), \tau) d\tau,\end{aligned}$$

which has eliminated the dependence of the right-hand side on $\bar{\alpha}$ except for in the first term. Let $\bar{\mathcal{A}} := \{\bar{\alpha} \in L_2(\Theta, \nu) : h(\cdot) = \int_\Theta \Phi(\cdot, \theta)\bar{\alpha}(\theta) d\nu \in \mathcal{A}\}$. Since $\hat{\alpha}_\infty \in \bar{\mathcal{A}}$ by assumption, taking the arg min over both sides of the above equation,

$$\hat{\alpha}_\infty \in \operatorname*{argmin}_{\bar{\alpha} \in \bar{\mathcal{A}}} \|\bar{\alpha}\|_{L_2(\Theta,\nu)}.$$

The claim now follows by the correspondence between $L_2(\Theta, \nu)$ and $\mathcal{H}$. ∎

Appendix D. Omitted proofs for Section 5

D.1 Proof of Proposition 5.1

Proposition 5.1 (Approximation error)**.** *Let $X \subset \mathbb{R}^n$ be compact. Fix $\delta \in (0, 1)$, $B_h > 0$, $h \in \mathcal{F}_2(B_h)$, and a positive integer K. Let $\theta_1, ..., \theta_K$ be i.i.d. draws from ν. Put $\eta = \frac{\delta}{2K}$. With probability*

at least $1-\delta$*, there exist weights* $\{\alpha_i\}_{i=1}^K \subset \mathbb{R}^{d_1}$ *such that* $\|\alpha_i\|_2 \leqslant B_h$ *for* $i = 1, ..., K$*, and*

$$\left\| \frac{1}{K}\sum_{i=1}^K \Phi(\cdot,\theta_i)\alpha_i - h \right\|_\infty \leqslant \frac{2}{K}\mathbb{E}\left\| \sum_{k=1}^K \varepsilon_i \Phi_\eta(\cdot,\theta_i)\alpha(\theta_i) \right\|_\infty + \sqrt{2}B_\Phi(\eta)B_h\sqrt{\frac{\log(2/\delta)}{K}} + B_h\sqrt{\frac{\delta \sup_{x\in X}\mathbb{E}\|\Phi(x,\theta)\|_{\mathrm{op}}^2}{2K}}.$$

Above, each ε_i *is an i.i.d. Rademacher random variable*[6] *and* $\|f\|_\infty := \sup_{x\in X}\|f(x)\|_2$*.*

Proof We first define a truncated target function $h_\eta(x)$ and its truncated approximation $\hat{h}_\eta(x; \{\theta_i\}_{i=1}^K)$

$$h_\eta(x) := \int_\Theta \Phi_\eta(x,\theta)\alpha(\theta)\, d\nu(\theta),$$
$$\hat{h}_\eta(x; \{\theta_i\}_{i=1}^K) := \frac{1}{K}\sum_{i=1}^K \Phi_\eta(x,\theta_i)\alpha(\theta_i).$$

Clearly, for each $x \in \mathbb{R}^n$,

$$\mathbb{E}_{\{\theta_i\}_{i=1}^K} \hat{h}_\eta(x; \{\theta_i\}_{i=1}^K) = h_\eta(x).$$

Now, consider two sets $\{\theta_i\} \subseteq \Theta$ and $\{\tilde{\theta}_i\} \subseteq \Theta$ that differ in only one index i. Observe that

$$\|\hat{h}_\eta(\cdot; \{\theta_i\}_{i=1}^K) - \hat{h}_\eta(\cdot; \{\tilde{\theta}_i\}_{i=1}^K)\|_\infty \leqslant \frac{2B_\Phi(\eta)B_h}{K}.$$

Hence, by McDiarmid's inequality, with probability at least $1-\delta/2$,

$$\begin{aligned}\|\hat{h}_\eta(\cdot; \{\theta_i\}_{i=1}^K) - h_\eta(\cdot)\|_\infty &\leqslant \mathbb{E}\|\hat{h}_\eta(\cdot; \{\theta_i\}_{i=1}^K) - h_\eta(\cdot)\|_\infty + \sqrt{2}B_\Phi(\eta)B_h\sqrt{\frac{\log(2/\delta)}{K}} \\ &\leqslant \frac{2}{K}\mathbb{E}\left\| \sum_{i=1}^K \varepsilon_i \Phi_\eta(\cdot;\theta_i)\alpha_i \right\|_\infty + \sqrt{2}B_\Phi(\eta)B_h\sqrt{\frac{\log(2/\delta)}{K}},\end{aligned}$$

where the last inequality follows by a standard symmetrization argument. Define the event $\mathcal{E}$ as

$$\mathcal{E} := \left\{ \max_{i=1,\ldots,K} \sup_{x\in X} \|\Phi(x,\theta_i)\|_{\mathrm{op}} \leqslant B_\Phi(\eta) \right\}.$$

By our assumption on B_Φ and a union bound, we have that $\mathbb{P}(\mathcal{E}^c) \leqslant \delta/2$. Furthermore, $\Phi(\cdot,\cdot)$ and $\Phi_\eta(\cdot,\cdot)$ agree on $\mathcal{E}$ by definition, so that

$$\begin{aligned}\mathbf{1}\{\mathcal{E}\}\left\| \frac{1}{K}\sum_{i=1}^K \Phi(\cdot,\theta_i)\alpha_i - h(\cdot) \right\|_\infty &= \mathbf{1}\{\mathcal{E}\}\left\| \frac{1}{K}\sum_{i=1}^K \Phi_\eta(\cdot,\theta_i)\alpha_i - h(\cdot) \right\|_\infty \\ &\leqslant \left\| \frac{1}{K}\sum_{i=1}^K \Phi_\eta(\cdot,\theta_i)\alpha_i - h_\eta(\cdot) \right\|_\infty + \|h(\cdot) - h_\eta(\cdot)\|_\infty.\end{aligned}$$

We now focus on bounding the term on the right-hand side. We write

$$h(x) - h_\eta(x) = \int_\Theta \mathbf{1}\{\|\Phi(x,\theta)\|_{\mathrm{op}} > B_\Phi(\eta)\}\Phi(x,\theta)\alpha(\theta)d\nu(\theta).$$

6. That is, $\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = 1/2$.

This implies the estimate

$$\begin{aligned}\|h(\cdot) - h_\eta(\cdot)\|_\infty &\leqslant \sup_{x\in X} B_h \mathbb{E}_{\theta\sim\nu} \mathbf{1}\{\|\Phi(x,\theta)\|_{\mathrm{op}} > B_\Phi(\eta)\}\|\Phi(x,\theta)\|_{\mathrm{op}} \\ &\leqslant B_h\sqrt{\eta}\sqrt{\sup_{x\in X}\mathbb{E}\|\Phi(x,\theta)\|_{\mathrm{op}}^2} \\ &= B_h\sqrt{\frac{\delta \sup_{x\in X}\mathbb{E}\|\Phi(x,\theta)\|_{\mathrm{op}}^2}{2K}}.\end{aligned}$$

The claim now follows by a union bound. ∎

D.2 Proof of Proposition 5.3

To state the proof of the proposition, we will require the following useful result.

Lemma D.1 (Maurer (2016), Corollary 4)**.** *Let $\mathcal{X}$ be any set, let $(x_1,\dots,x_n)\in\mathcal{X}^n$, let $\mathcal{F}$ be a class of functions $f:\mathcal{X}\to\ell_2$, and let $h_i:\ell_2\to\mathbb{R}$ have Lipschitz constant L. Then,*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\sum_i \varepsilon_i h_i(f(x_i)) \leqslant \sqrt{2}L\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i,k}\varepsilon_{i,k}f_k(x_i)$$

where the ε_{ik} are i.i.d. Rademacher random variables and $f_k(x_i)$ is the k^{th} component of $f(x_i)$.

We may now proceed with the proof.

Proposition 5.3 (Rademacher complexity bound)**.** *Let Assumption 5.2 hold, and denote $B_X := \sup_{x\in X}\|x\|_2$. Then for any $\eta\in(0,1)$,*

$$\frac{2}{K}\mathbb{E}\left\|\sum_{i=1}^K \varepsilon_i \Phi_\eta(\cdot;\theta_i)\alpha(\theta_i)\right\|_\infty \leqslant \frac{4B_hB_\Phi(\eta)}{\sqrt{K}}\left[B_X\sqrt{\mathbb{E}\|w_1\|_2^2} + \sqrt{d_1}\right].$$

Proof Put $\alpha_i = \alpha(\theta_i)$ and $M_{\eta,i} := M_\eta(w_i)$. We write, by definition of the $\|\cdot\|_\infty$-norm and duality,

$$\mathbb{E}\left\|\sum_{i=1}^K \varepsilon_i\Phi_\eta(x;\theta_i)\alpha_i\right\|_\infty = \mathbb{E}\sup_{x\in X}\sup_{\psi\in\mathbb{S}^{d_1-1}}\sum_{i=1}^K \varepsilon_i\psi^{\mathsf{T}}M_{\eta,i}\alpha_i\phi(w_i^{\mathsf{T}}x+b_i).$$

Towards applying Lemma D.1, for a tuple $(x,\psi)\in X\times\mathbb{S}^{d_1-1}$, define

$$f_{x,\psi}(w,b) := \begin{pmatrix} w^{\mathsf{T}}x + b \\ \psi\end{pmatrix}.$$

Next, define $h_i:\mathbb{R}\times\mathbb{S}^{d_1-1}\to\mathbb{R}$ as

$$h_i(v_1,v_2) := v_2^{\mathsf{T}}M_{\eta,i}\alpha_i\phi(v_1).$$

We need to show that h_i is Lipschitz continuous. Let $v=(v_1,v_2)$, $w=(w_1,w_2)$, and observe that

$$|h_i(v_1,v_2) - h_i(w_1,w_2)| \leqslant \sqrt{2}B_hB_\Phi(\eta)\|v-w\|_2,$$

where we have applied the triangle inequality. Now, let $\{\xi_i\}_{i=1}^K \subseteq \{\pm 1\}$ and $\{\zeta_i\}_{i=1}^K \subseteq \{\pm 1\}^{d_1}$ be independent random vectors with i.i.d. Rademacher random variables as entries. Then, by Lemma D.1,

$$\begin{aligned}
\mathbb{E}\sup_{x\in X}\sup_{\psi\in\mathbb{S}^{d_1-1}}\sum_{i=1}^K \varepsilon_i \psi^\mathsf{T} M_{\eta,i}\alpha_i\phi(w_i^\mathsf{T}x+b_i) &= \mathbb{E}\sup_{x\in X,\psi\in\mathbb{S}^{d_1-1}}\sum_{i=1}^K \varepsilon_i h_i(F_{x,\psi}(w_i,b_i)),\\
&\leqslant 2B_hB_\Phi(\eta)\mathbb{E}\sup_{x\in X,\psi\in\mathbb{S}^{d_1-1}}\sum_{i=1}^K\left\langle\begin{pmatrix}\xi_i\\ \zeta_i\end{pmatrix},\begin{pmatrix}w_i^\mathsf{T}x+b_i\\ \psi\end{pmatrix}\right\rangle,\\
&= 2B_hB_\Phi(\eta)\left[\mathbb{E}\sup_{x\in X}\sum_{i=1}^K \xi_i w_i^\mathsf{T}x + \mathbb{E}\sup_{\psi\in\mathbb{S}^{q-1}}\sum_{i=1}^K\zeta_i^\mathsf{T}\psi\right],\\
&\leqslant 2B_hB_\Phi(\eta)\left[B_X\mathbb{E}\left\|\sum_{i=1}^K\xi_i w_i\right\|_2 + \mathbb{E}\left\|\sum_{k=1}^K\zeta_i\right\|_2\right],\\
&\leqslant 2\sqrt{K}B_hB_\Phi(\eta)\left[B_X\sqrt{\mathbb{E}\|w_1\|_2^2}+\sqrt{d_1}\right].
\end{aligned}$$

This completes the proof. ∎

Appendix E. Omitted proofs for Section 6

E.1 Details of Example 6.2

First, for any $\delta > 0$, we define the function

$$\bar{s}_\delta(x) := \frac{s_\delta(x)}{x}.$$

Lemma E.1. *For any $\delta > 0$, the function $x \mapsto \bar{s}_\delta(\sqrt{x})$ is $\frac{1}{2\delta^2}$-Lipschitz on $\mathbb{R}_{\geqslant 0}$.*

Proof Fix $x, y \in \mathbb{R}_{\geqslant 0}$. Without loss of generality, suppose that $x \leqslant y$ (otherwise, we may flip the roles of x and y). If $y \leqslant \delta^2$, then $\bar{s}_\delta(\sqrt{x}) = \bar{s}_\delta(\sqrt{y}) = 0$, in which case the claim is trivial.

Now suppose that $x \geqslant \delta^2$. On $[\delta, \infty)$, we have that $\bar{s}_\delta$ coincides with $x \mapsto 1 - \delta/x$, and hence

$$\bar{s}_\delta(\sqrt{x}) - \bar{s}_\delta(\sqrt{y}) = 1 - \frac{\delta}{\sqrt{x}} - \left(1 - \frac{\delta}{\sqrt{y}}\right) = \delta\left(\frac{1}{\sqrt{y}} - \frac{1}{\sqrt{x}}\right).$$

The function $x \mapsto 1/\sqrt{x}$ is $\frac{1}{2\delta^3}$-Lipschitz on $[\delta^2, \infty)$, and hence

$$|\bar{s}_\delta(\sqrt{x}) - \bar{s}_\delta(\sqrt{y})| \leqslant \frac{1}{2\delta^2}|x - y|.$$

Finally, we suppose that $x \leqslant \delta^2 \leqslant y$. By concavity of the square root on $\mathbb{R}_{\geqslant 0}$,

$$\sqrt{y} \leqslant \delta + \frac{1}{2\delta}(y - \delta^2).$$

Therefore,

$$|\bar{s}_\delta(\sqrt{x}) - \bar{s}_\delta(\sqrt{y})| = |\bar{s}_\delta(\sqrt{y})| = 1 - \frac{\delta}{\sqrt{y}} = \frac{\sqrt{y}-\delta}{\sqrt{y}} \leqslant \frac{\sqrt{y}-\delta}{\delta} \leqslant \frac{y-\delta^2}{2\delta^2} \leqslant \frac{y-x}{2\delta^2} = \frac{|y-x|}{2\delta^2}.$$

∎

Example 6.2. *Fix a scalar $\delta > 0$. Let $s_\delta : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ be defined as $s_\delta(x) := (x-\delta)\mathbf{1}\{x > \delta\}$. For any $\Delta > 0$, the function $x \mapsto s^2_{\sqrt{\Delta}}(\sqrt{x})$ is a $(\Delta, 1/(2\Delta), 1)$-admissible deadzone.*

Proof It is straightforward to check that $\frac{d}{dx} s^2_{\sqrt{\Delta}}(\sqrt{x}) = \bar{s}_{\sqrt{\Delta}}(\sqrt{x})$. Conditions (i) and (ii) are immediately satisfied. To check condition (iii), observe that by Lemma E.1, $\bar{s}_{\sqrt{\Delta}}(\sqrt{x})$ is $1/(2\Delta)$-Lipschitz. Finally, $\bar{s}_{\sqrt{\Delta}}(\sqrt{x}) \leqslant 1$ for all $x \geqslant 0$. ■

E.2 Details of Example 6.3

Example 6.3. *Fix $\delta > 0$ and $\gamma > 0$. Define $s_{\delta,\gamma}$ as:*

$$s_{\delta,\gamma}(x) := \begin{cases} 0 & \text{if } x \leqslant \delta, \\ \frac{(x-\delta)^2}{4\gamma} & \text{if } x \in (\delta, \delta + 2\gamma), \\ x - (\delta + \gamma) & \text{if } x \geqslant \delta + 2\gamma. \end{cases}$$

For any $\Delta > 0$ and $\gamma > 0$, the function $s_{\Delta,\gamma}$ is a $(\Delta, 1/(2\gamma), 1)$-admissible deadzone.

Proof It is easy to check that the derivative $s'_{\Delta,\gamma}$ exists, is continuous, and is given by:

$$s'_{\Delta,\gamma}(x) = \begin{cases} 0 & \text{if } x \leqslant \Delta, \\ \frac{x-\Delta}{2\gamma} & \text{if } x \in (\Delta, \Delta + 2\gamma), \\ 1 & \text{if } x \geqslant \Delta + 2\gamma. \end{cases}$$

It is also easy to check that this derivative is $\frac{1}{2\gamma}$-Lipschitz and bounded by 1. ■

E.3 Proof of Theorem 6.4

Proposition E.2. *Fix any $\Delta > 0$. Let σ_Δ be Δ-admissible. Define $F : \mathbb{R}^n \times \mathbb{R}^s \times O_p \times O_m \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}^n \times \mathbb{R}^s \times \mathbb{R}^p \times \mathbb{R}^m$ as:*

$$F(x, e, \hat{\alpha}_p, \hat{\alpha}_m, t) := \begin{bmatrix} f(x,t) + g(x,t)(Y(x,t)\tilde{\alpha}_p + \Psi(x)\hat{\alpha}_m - h(x)), \\ f_e(e,t) + g_e(x,t)(Y(x,t)\tilde{\alpha}_p + \Psi(x)\hat{\alpha}_m - h(x)), \\ -\sigma'_\Delta(Q(e,t))[\nabla^2\psi_p(\hat{\alpha}_p)]^{-1}Y(x,t)^\mathsf{T} g_e(e,t)^\mathsf{T}\nabla Q(e,t), \\ -\sigma'_\Delta(Q(e,t))[\nabla^2\psi_m(\hat{\alpha}_m)]^{-1}\Psi(x)^\mathsf{T} g_e(e,t)^\mathsf{T}\nabla Q(e,t). \end{bmatrix}.$$

The function $F(x, e, \hat{\alpha}_p, \hat{\alpha}_m, t)$ is locally Lipschitz in $(x, e, \hat{\alpha}_p, \hat{\alpha}_m)$.

Proof The functions $f, f_e, g, g_e, h, Y, \Psi, \nabla Q, B_h$, and σ'_Δ are all locally Lipschitz and locally bounded by assumption. As long as we can check that both $\zeta_1(e,t) := \sigma'_\Delta(Q(e,t))$ and $\zeta_{2,\ell}(\hat{\alpha}) := [\nabla^2\psi_\ell(\hat{\alpha})]^{-1}$ for $\ell \in \{p, m\}$ are locally Lipschitz and locally bounded, then the result follows via repeated applications of the sum and product composition rules (Proposition B.1).

We now verify that ζ_1 is locally Lipschitz and locally bounded. Since $\nabla Q(e,t)$ is locally bounded, this means that $Q(e,t)$ is locally Lipschitz. Furthermore, since $0 \leqslant Q(e,t) \leqslant \mu_2(\|e\|_2)$, it is clear that $Q(e,t)$ is locally bounded. Next, σ'_Δ is locally Lipschitz by admissibility. Since σ'_Δ does not depend on time, then it is also locally bounded. This shows that ζ_1 is locally Lipschitz and bounded, since it is the composition of two locally Lipschitz and bounded functions.

For $\zeta_{2,\ell}$, we first observe that, since ψ_ℓ is strongly convex with respect to a norm $\|\cdot\|$ on O_ℓ, there exists a $c > 0$ such that $\nabla^2\psi_\ell(\hat{\alpha}) \succcurlyeq cI$ for all $\hat{\alpha} \in O_\ell$. Next, for any invertible square matrices A, B, we have the algebraic identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$. Therefore, for any two $\hat{\alpha}_1, \hat{\alpha}_2$,

$$\|[\nabla^2\psi_\ell(\hat{\alpha}_1)]^{-1} - [\nabla^2\psi_\ell(\hat{\alpha}_2)]^{-1}\|_{\mathrm{op}} \leqslant c^{-2}\|\nabla^2\psi(\hat{\alpha}_1) - \nabla^2\psi(\hat{\alpha}_2)\|_{\mathrm{op}}.$$

Because the potential ψ_ℓ has locally Lipschitz Hessians, this shows that $\zeta_{2,\ell}$ is locally Lipschitz. Since $\zeta_{2,\ell}$ does not depend on time, it is also locally bounded. ∎

Proposition E.3. *Let $O \subseteq \mathbb{R}^\ell$ be an open convex set, and let $\psi : O \to \mathbb{R}$ be a strongly convex potential with respect to some norm $\|\cdot\|$ on O. Then we have that*

$$\inf_{\substack{\alpha,\hat{\alpha}\in O,\\ \alpha\neq\hat{\alpha}}} \frac{d_\psi\left(\alpha\|\hat{\alpha}\right)}{\|\alpha-\hat{\alpha}\|_2^2} > 0.$$

Proof Because all norms on $\mathbb{R}^\ell$ are equivalent, strong convexity on O says that there exists a $c > 0$ such that $\nabla^2\psi(a) \succcurlyeq cI$ for all $a \in O$. By Taylor's theorem, for any arbitrary $\alpha, \hat{\alpha}$,

$$d_\psi\left(\alpha\|\hat{\alpha}\right) = \frac{1}{2}(\alpha-\hat{\alpha})^\mathsf{T}\left[\int_0^1 \nabla^2\psi((1-t)\hat{\alpha}+\alpha t)dt\right](\alpha-\hat{\alpha}) \geqslant \frac{c}{2}\|\alpha-\hat{\alpha}\|_2^2.$$

The claim now follows. ∎

Theorem 6.4 (Adaptive control with finite-dimensional approximation)**.** *Suppose that Assumption 3.7 holds. Let $\alpha_{\ell,0} := \arg\min_{\alpha\in O_\ell}\psi_\ell(\alpha)$ for $\ell \in \{p, m\}$. Fix $B_{\alpha_p} > 0$ satisfying $d_{\psi_p}\left(\alpha_p\|\alpha_{p,0}\right) \leqslant B_{\alpha_p}$, $B_{\alpha_m} > 0$, and R satisfying*

$$R > \mu_1^{-1}\left(Q(e(0),0) + B_{\alpha_p} + B_{\alpha_m}\right).$$

Suppose there exists a finite C_e such that for every $T > 0$:

$$\max_{t\in[0,T]}\|e(t)\|_2 \leqslant R \text{ implies } \|x(T)-x_d(T)\|_2 \leqslant C_eR. \tag{6.1}$$

Let $\Psi : \mathbb{R}^n \to \mathbb{R}^{d\times m}$ be a locally Lipschitz feature map. Define the constants

$$\begin{aligned}
B_d &:= \sup_{t\geqslant 0}\|x_d(t)\|_2,\\
B_x &:= C_eR + B_d,\\
B_{g_e} &:= \sup_{t\geqslant 0}\sup_{\|x\|_2\leqslant B_x}\|g_e(x,t)\|_{\mathrm{op}},\\
B_{\nabla Q} &:= \sup_{t\geqslant 0}\sup_{\|e\|_2\leqslant R}\|\nabla Q(e,t)\|_2,\\
B_{\mathrm{approx}} &:= \inf_{d_{\psi_m}(\alpha_m\|\alpha_{m,0})\leqslant B_{\alpha_m}}\sup_{\|x\|_2\leqslant B_x}\|\Psi(x)\alpha_m - h(x)\|_2.
\end{aligned}$$

Let Δ be any positive constant satisfying

$$\Delta \geqslant \mu_2(\rho^{-1}(2B_{g_e}B_{\nabla Q}B_{\mathrm{approx}})),$$

and let σ_Δ be a Δ-admissible deadzone. Then the dynamical system

$$\begin{aligned}
\dot{x} &= f(x,t) + g(x,t)(u(x,t) - Y(x,t)\alpha_p - h(x)),\\
\dot{e} &= f_e(e,t) + g_e(x,t)(u(x,t) - Y(x,t)\alpha_p - h(x)),\\
u(x,t) &= Y(x,t)\hat{\alpha}_p + \Psi(x)\hat{\alpha}_m,\\
\frac{d}{dt}\nabla\psi_p(\hat{\alpha}_p) &= -\sigma_\Delta'(Q(e,t))Y(x,t)^\mathsf{T}g_e(e,t)^\mathsf{T}\nabla Q(e,t),\\
\frac{d}{dt}\nabla\psi_m(\hat{\alpha}_m) &= -\sigma_\Delta'(Q(e,t))\Psi(x)^\mathsf{T}g_e(e,t)^\mathsf{T}\nabla Q(e,t),
\end{aligned}$$

with initial conditions $x(0) = x_0$, $e(0) = m(x_0, 0)$, $\hat{\alpha}_p(0) = \alpha_{p,0}$, *and* $\hat{\alpha}_m(0) = \alpha_{m,0}$ *has a solution* $(x(t), e(t), \hat{\alpha}_p(t), \hat{\alpha}_m(t))$ *that exists for all* $t \geqslant 0$. *Furthermore,*

$$\limsup_{t\to\infty} \|e(t)\|_2 \leqslant \mu_1^{-1}(\Delta).$$

Proof By Proposition E.2, the right-hand side of the dynamical system on $(x, e, \hat{\alpha}_p, \hat{\alpha}_m)$ is locally Lipschitz, and therefore there exists a maximal time $T_{\max} > 0$ such that there exists a unique C^1 curve $t \mapsto (x(t), e(t), \hat{\alpha}_p(t), \hat{\alpha}_m(t))$ that satisfies the dynamics on $[0, T_{\max})$.

We now define the candidate Lyapunov function $V : [0, T_{\max}) \to \mathbb{R}_{\geqslant 0}$ as

$$V(t) = \sigma_\Delta(Q(e(t), t)) + d_{\psi_p}\left(\alpha_p \| \hat{\alpha}_p\right) + d_{\psi_m}\left(\alpha_m \| \hat{\alpha}_m\right).$$

where α_m is the minimizing α_m in the definition B_{approx}[7]. Taking the time derivative of V and suppressing dependence on time,

$$\begin{aligned}
\frac{d}{dt}V(t) &= \sigma'_\Delta(Q)\left(\langle \nabla Q, f_e + g_e(Y\tilde{\alpha}_p + \Psi\alpha_m - h)\rangle + \frac{\partial Q}{\partial t}\right) + \left\langle \frac{d}{dt}\nabla\psi_p(\hat{\alpha}_p), \tilde{\alpha}_p \right\rangle + \left\langle \frac{d}{dt}\nabla\psi_m(\hat{\alpha}_m), \tilde{\alpha}_m \right\rangle \\
&\leqslant \sigma'_\Delta(Q)\left(-\rho(\|e\|_2) + \langle \nabla Q, g_e(Y\tilde{\alpha}_p + \Psi\alpha_m - h)\rangle\right) + \left\langle \frac{d}{dt}\nabla\psi_p(\hat{\alpha}_p), \tilde{\alpha}_p \right\rangle + \left\langle \frac{d}{dt}\nabla\psi_m(\hat{\alpha}_m), \tilde{\alpha}_m \right\rangle \\
&= -\sigma'_\Delta(Q)\rho(\|e\|_2) + \sigma'_\Delta(Q)\left\langle g_e^\mathsf{T}\nabla Q, \Psi\hat{\alpha}_m - h\right\rangle - \sigma'_\Delta(Q)\left\langle g_e^\mathsf{T}\nabla Q, \Psi\tilde{\alpha}_m\right\rangle \\
&= -\sigma'_\Delta(Q)\rho(\|e\|_2) + \sigma'_\Delta(Q)\left\langle g_e^\mathsf{T}\nabla Q, \Psi\alpha_m - h\right\rangle \\
&\leqslant -\sigma'_\Delta(Q)\rho(\|e\|_2) + \sigma'_\Delta(Q)\|g_e^\mathsf{T}\nabla Q\|_2\|\Psi\alpha_m - h\|_2.
\end{aligned}$$

Because σ_Δ is a Δ-admissible deadzone, $\sigma'_\Delta > 0$ only when $Q > \Delta$. But since $Q(e, t) \leqslant \mu_2(\|e\|_2)$, we have $\|e\|_2 > \mu_2^{-1}(\Delta)$. Therefore,

$$\frac{d}{dt}V(t) \leqslant -\sigma'_\Delta(Q)\rho(\mu_2^{-1}(\Delta)) + \sigma'_\Delta(Q)\|g_e^\mathsf{T}\nabla Q\|_2\|\Psi\alpha_m - h\|_2. \tag{E.1}$$

Let T_0 be defined as

$$T_0 := \sup\{T \in [0, T_{\max}) \mid \|e(t)\|_2 \leqslant R \ \forall t \in [0, T]\}.$$

Note that since

$$\|e(0)\|_2 \leqslant \mu_1^{-1}(V(0)) \leqslant \mu_1^{-1}\left(Q(e(0), 0) + B_{\alpha_p} + B_{\alpha_m}\right) < R,$$

T_0 is well-defined. Now, by means of contradiction, suppose $T_0 < T_{\max}$. For every $t \in [0, T_0]$, by (6.1), we have that $\|x(t) - x_d(t)\|_2 \leqslant C_e R$, and hence $\|x(t)\|_2 \leqslant C_e R + B_d = B_x$. Hence, by the definition of B_{g_e} and $B_{\nabla Q}$, from (E.1) and the requirement that $\Delta \geqslant \mu_2(\rho^{-1}(2B_{g_e}B_{\nabla Q}B_{\text{approx}}))$, for every $t \in [0, T_0]$,

$$\begin{aligned}
\frac{d}{dt}V(t) &\leqslant -\sigma'_\Delta(Q)\rho(\mu_2^{-1}(\Delta)) + \sigma'_\Delta(Q)B_{g_e}B_{\nabla Q}B_{\text{approx}} \\
&\leqslant -\sigma'_\Delta(Q)\rho(\mu_2^{-1}(\Delta))/2.
\end{aligned}$$

Hence, $V(t) \leqslant V(0)$ for all $t \in [0, T_0]$. On the other hand, since T_0 is maximal, we must have that $\|e(T_0)\|_2 = R$, otherwise, if $\|e(T_0)\|_2 < R$, by continuity of the solution $e(t)$ on $[0, T_{\max})$, there would exist a $\delta > 0$ such that for all $t \in [0, T_0 + \delta]$, we have $\|e(t)\|_2 \leqslant R$. This means then that,

$$V(0) \geqslant V(T_0) \geqslant \mu_1(\|e(T)\|_2) = \mu_1(R) > \mu_1(\mu_1^{-1}(V(0))) = V(0),$$

7. Such a minimizing α_m exists since the function $\alpha_m \mapsto \sup_{\|x\|_2 \leqslant B_x} \|\Psi(x)\alpha_m - h(x)\|_2$ is continuous and the set $\{d_{\psi_m}(\alpha_m \| \alpha_{m,0}) \leqslant B_{\alpha_m}\}$ is closed.

a contradiction. Hence $T_0 = T_{\max}$.

Now we argue that $T_{\max}$ cannot be finite. Suppose towards a contradiction that $T_{\max}$ is finite. We already have $\max_{t\in[0,T_{\max})}\|e(t)\|_2 \leqslant R$. This implies that $\|x(t)\|_2 \leqslant C_e R + B_d = B_x$ for $t \in [0, T_{\max})$. Finally, since $V(t) \leqslant V(0)$ on all $t \in [0, T_{\max})$, this shows that both $\|\hat{\alpha}_p(t)\|_2$ and $\|\hat{\alpha}_m(t)\|_2$ are uniformly bounded for all $t \in [0, T_{\max})$ via Proposition E.3. This contradicts the maximality of $T_{\max}$, showing that $T_{\max} = \infty$.

To continue the proof, we integrate the inequality $\frac{d}{dt}V(t) \leqslant -\sigma'_\Delta(Q)\rho(\mu_2^{-1}(\Delta))/2$ to conclude that

$$\int_0^\infty \sigma'_\Delta(Q(e(t),t))dt \leqslant \frac{2V(0)}{\rho(\mu_2^{-1}(\Delta))}.$$

We now argue that the integrand $t \mapsto \sigma'_\Delta(Q(e(t),t))$ is uniformly continuous.

To do this, we will argue that (a) $t \mapsto Q(e(t),t)$ is uniformly bounded, (b) $t \mapsto e(t)$ is uniformly Lipschitz, and (c) $t \mapsto Q(e(t),t)$ is uniformly Lipschitz. To see (a), we note that $Q(e(t),t) \leqslant V(t) \leqslant V(0)$. To see (b), we note that:

$$\|\dot{e}(t)\|_2 \leqslant \|f_e(e(t),t)\|_2 + \|g_e(x(t),t)\|_{\mathrm{op}}(\|Y(x(t),t)\|_{\mathrm{op}}\|\tilde{\alpha}_p(t)\|_2 + \|\Psi(x(t))\|_{\mathrm{op}}\|\hat{\alpha}_m(t)\|_2 + \|h(x(t))\|_2).$$

Since f_e, g_e, Y, Ψ, and h are locally bounded in the first argument uniformly in t, and since $\|\hat{\alpha}_p(t)\|_2$ and $\|\hat{\alpha}_m(t)\|_2$ are uniformly bounded, this shows that $\|\dot{e}(t)\|_2$ is uniformly bounded, and hence $t \mapsto e(t)$ is uniformly Lipschitz. To see (c), we observe that:

$$\begin{aligned}
&|Q(e(s),s) - Q(e(t),t)| \\
&\qquad \leqslant |Q(e(s),s) - Q(e(s),t)| + |Q(e(s),t) - Q(e(t),t)| \\
&\qquad \leqslant \left[\sup_{t\geqslant 0}\sup_{\|e\|_2\leqslant R}\left|\frac{\partial Q}{\partial t}(e,t)\right|\right]|s-t| + \left[\sup_{t\geqslant 0}\sup_{\|e\|_2\leqslant R}\|\nabla Q(e,t)\|_2\right]\|e(s)-e(t)\|_2.
\end{aligned}$$

Since $\frac{\partial Q}{\partial t}$ and ∇Q are both locally bounded in e uniformly in t, and since $t \mapsto e(t)$ is uniformly Lipschitz, we see that $t \mapsto Q(e(t),t)$ is also uniformly Lipschitz.

We now argue that $t \mapsto \sigma'_\Delta(Q(e(t),t))$ is uniformly continuous. Since σ'_Δ is locally Lipschitz, it is uniformly Lipschitz on $[0, V(0)]$. Therefore, $t \mapsto \sigma'_\Delta(Q(e(t),t))$ is the composition of two Lipschitz functions, and is hence Lipschitz (and therefore uniformly continuous).

From this, we apply Barbalat's lemma to conclude that:

$$\lim_{t\to\infty}\sigma'_\Delta(Q(e(t),t)) = 0.$$

Since σ_Δ is a Δ-admissible deadzone, this implies that

$$\limsup_{t\to\infty}\|e(t)\|_2 \leqslant \mu_1^{-1}(\Delta).$$

■

E.4 Proof of Theorem 6.7

Theorem 6.7 (Adaptive prediction with uniform approximation)**.** *Suppose that the trajectory $x(t)$ of the system $\dot{x} = f(x,t)$ is uniformly bounded. Choose a continuous and locally Lipschitz $k(\hat{x},x)$ such that $f(\hat{x},t)+k(\hat{x},x(t))$ is contracting in a metric $M : \mathbb{R}^n \times \mathbb{R}_{\geqslant 0} \to \mathsf{Sym}^{n\times n}_{\succ 0}$ with rate $\lambda > 0$, and suppose that the metric M satisfies $\mu I \preccurlyeq M(\hat{x},t) \preccurlyeq LI$ for all $\hat{x}$ and t. Let $\gamma(\cdot;\hat{x},x,t) : [0,1] \to \mathbb{R}^n$ denote a geodesic between $\hat{x}$ and x in the metric $M(\hat{x},t)$, and let $\gamma_s(s;\hat{x},x,t)$ denote the derivative of*

$s \mapsto \gamma(s; \hat{x}, x, t)$. Suppose that the map $(\hat{x}, t) \mapsto \|\gamma_s(0; \hat{x}, x(t), t)\|_2$ is locally bounded in $\hat{x}$ uniformly in t. Fix any $B_{\alpha_p} > 0$ satisfying $d_{\psi_p}(\alpha_p \| \alpha_{p,0}) \leqslant B_{\alpha_p}$, any $B_{\alpha_m} > 0$, and any R satisfying

$$R > \sqrt{\frac{Q(\hat{x}(0), 0) + B_{\alpha_p} + B_{\alpha_m}}{\mu}}, \;\; Q(\hat{x}, t) := E_{M(\cdot,t)}(\hat{x}, x(t)).$$

Let $\Psi : \mathbb{R}^n \to \mathbb{R}^{d \times m}$ be a locally Lipschitz feature map. Define the following constants

$$\begin{aligned}
B_x &:= \sup_{t \geqslant 0} \|x(t)\|_2, \\
B_{\hat{x}} &:= R + B_x, \\
B_\gamma &:= \sup_{t \geqslant 0} \sup_{\|\hat{x}\|_2 \leqslant B_{\hat{x}}} \|\gamma_s(0; \hat{x}, x(t), t)\|_2, \\
B_{\text{approx}} &:= \inf_{d_{\psi_m}(\alpha_m \| \alpha_{m,0}) \leqslant B_{\alpha_m}} \sup_{\|\hat{x}\|_2 \leqslant B_{\hat{x}}} \|\Psi(\hat{x})\alpha_m - h(\hat{x})\|_2.
\end{aligned}$$

Choose any Δ satisfying $\Delta \geqslant \frac{L^2 B_\gamma B_{\text{approx}}}{\lambda \mu}$, and let σ_Δ be a Δ-admissible deadzone. Then the dynamical system

$$\begin{aligned}
\dot{\hat{x}} &= \hat{f}(\hat{x}, \hat{\alpha}_p, \hat{\alpha}_m, t) + k(\hat{x}, x(t)), \\
\hat{f}(\hat{x}, \hat{\alpha}_p, \hat{\alpha}_m, t) &= Y(\hat{x}, t)\hat{\alpha}_p + \Psi(\hat{x})\hat{\alpha}_m, \\
\frac{d}{dt}\nabla\psi_p(\hat{\alpha}_p) &= -\sigma'_\Delta(Q(\hat{x}, t)) Y(\hat{x}, t)^\mathsf{T} \nabla Q(\hat{x}, t), \\
\frac{d}{dt}\nabla\psi_m(\hat{\alpha}_m) &= -\sigma'_\Delta(Q(\hat{x}, t)) \Psi(\hat{x})^\mathsf{T} \nabla Q(\hat{x}, t),
\end{aligned}$$

with initial conditions $\hat{x}(0) = \hat{x}_0$, $\hat{\alpha}_p(0) = \alpha_{p,0}$, and $\hat{\alpha}_m(0) = \alpha_{m,0}$ has a solution that exists for all $t \geqslant 0$. Furthermore,

$$\limsup_{t \to \infty} \|\hat{x}(t) - x(t)\|_2 \leqslant \sqrt{\frac{\Delta}{\mu}}.$$

Proof We proceed by reduction to Theorem 6.4. Observe that we may write the predictor (3.3) in the matched uncertainty form (3.1) with $g(\hat{x}, t) = I$

$$\dot{\hat{x}} = f(\hat{x}, t) + k(\hat{x}, x(t)) + \left(\hat{f}(\hat{x}, \hat{\alpha}, t) - f(\hat{x}, t)\right).$$

This is an adaptive control problem with input $\hat{f}(\hat{x}, \hat{\alpha}, t)$ and desired trajectory $x(t)$. The "nominal dynamics" $\bar{f}(\hat{x}, t) := f(\hat{x}, t) + k(\hat{x}, x(t))$ is contracting at rate λ in the metric M by assumption, meaning that

$$\frac{\partial \bar{f}}{\partial \hat{x}}(\hat{x}, t)^\mathsf{T} M(\hat{x}, t) + M(\hat{x}, t)\frac{\partial \bar{f}}{\partial \hat{x}}(\hat{x}, t) + \dot{M}(\hat{x}, t) \preccurlyeq -2\lambda M(\hat{x}, t) \;\; \forall \hat{x} \in \mathbb{R}^n, t \in \mathbb{R}_{\geqslant 0}.$$

Let the error signal $e(t) := \hat{x}(t) - x(t)$. The error dynamics are

$$\begin{aligned}
\dot{e} &= f(\hat{x}, t) - f(x(t), t) + k(\hat{x}, x(t)) + \left(\hat{f}(\hat{x}, \hat{\alpha}, t) - f(\hat{x}, t)\right) \\
&= \bar{f}(\hat{x}, t) - f(x(t), t) + \left(\hat{f}(\hat{x}, \hat{\alpha}, t) - f(\hat{x}, t)\right).
\end{aligned}$$

Hence we can define

$$f_e(e, t) := \bar{f}(e + x(t), t) - f(x(t), t), \;\; g_e(x, t) := I.$$

We first check that $f_e(e,t)$ is locally Lipschitz and locally bounded uniformly in t via Proposition B.1. First we consider $(e,t) \mapsto f(e+x(t),t)$. Write this map as the composition $f(\phi(e,t),t)$ with $\phi(e,t) := e + x(t)$. Since the signal $x(t)$ is uniformly bounded, it is clear that ϕ is both locally Lipschitz and locally bounded uniformly in t. Since the outer function $f(x,t)$ is also locally Lipschitz and locally bounded uniformly in t, the composition remains locally Lipschitz and locally bounded uniformly in t. Next, since $k(\hat{x},x)$ is locally Lipschitz in $\hat{x}$ and continuous, and since $x(t)$ is uniformly bounded, the function $(\hat{x},t) \mapsto k(\hat{x},x(t))$ is locally Lipschitz and locally bounded uniformly in t. By an identical composition argument, so is $(e,t) \mapsto k(e+x(t),x(t))$. Finally, the function $(e,t) \mapsto f(x(t),t)$ is trivially locally Lipschitz and locally bounded uniformly in t since $x(t)$ is bounded. Therefore, f_e is locally Lipschitz and locally bounded uniformly in t.

The Jacobian $\frac{\partial f_e}{\partial e}(e,t) = \frac{\partial \bar{f}}{\partial \hat{x}}(e+x(t),t)$, which shows that $f_e(e,t)$ is contracting at rate λ in the metric $M_e(e,t) := M(e+x(t),t)$. Furthermore, it is easy to check that $e = 0$ is a particular solution to $\dot{e} = f_e(e,t)$, as $k(x,x) = 0$ for all x. Therefore, f_e admits an exponentially stable Lyapunov function $Q(e,t) = E_{M_e(\cdot,t)}(e,0)$ that satisfies:

$$\langle \nabla Q(e,t), f_e(e,t) \rangle + \frac{\partial Q}{\partial t}(e,t) \leqslant -2\lambda Q(e,t) \ \forall e \in \mathbb{R}^n, t \in \mathbb{R}_{\geqslant 0}.$$

Moreover, because $\mu I \preccurlyeq M_e(e,t) \preccurlyeq LI$,

$$\mu \|e\|_2^2 \leqslant Q(e,t) \leqslant L\|e\|_2^2 \ \forall e \in \mathbb{R}^n, t \in \mathbb{R}_{\geqslant 0}.$$

Now, observe that

$$\nabla Q(e,t) = M_e(e,t)\gamma_s(0; e+x(t), x(t), t),$$

so that $B_{\nabla Q} = \sup_{t \geqslant 0} \sup_{\|e\|_2 \leqslant R} \|\nabla Q(e,t)\|_2 \leqslant LB_\gamma$. Furthermore, by the boundedness of M and the assumption that $(\hat{x},t) \mapsto \|\gamma_s(0;\hat{x},x(t),t)\|_2$ is locally bounded in $\hat{x}$ uniformly in t, we have that $\nabla Q(e,t)$ is locally bounded in e uniformly in t. Similarly, since

$$\frac{\partial Q}{\partial t}(e,t) = -\gamma_s(1; e+x(t), x(t), t)^{\mathsf{T}} M_e(e,t) f_e(e,t),$$

by the boundedness of M, the assumption that $(\hat{x},t) \mapsto \|\gamma_s(1;\hat{x},x(t),t)\|_2$ is locally bounded in $\hat{x}$ uniformly in t (since geodesics have constant speed, we have $\|\gamma_s(1;\hat{x},x(t),t)\|_2 = \|\gamma_s(0;\hat{x},x(t),t)\|_2$), we have that $\frac{\partial Q}{\partial t}(e,t)$ is locally bounded in $\hat{x}$ uniformly in t. Hence, we can invoke Theorem 6.4 with

$$C_e = 1, \ B_d = B_x, \ B_x = B_{\hat{x}}, \ B_{g_e} = 1, \ B_{\nabla Q} = LB_\gamma,$$
$$\rho(\|e\|_2) = 2\lambda\mu\|e\|_2^2, \ \mu_1(\|e\|_2) = \mu\|e\|_2^2, \ \mu_2(\|e\|_2) = L\|e\|_2^2.$$

The result now follows. ∎

E.5 Proof of Theorem 6.6

Theorem 6.6 (Approximate interpolation)**.** *Suppose the hypotheses of Theorem 6.4 hold. Let σ_Δ denote a (Δ, L, B)-admissible deadzone, and assume that f_e, g_e, and Y are locally Lipschitz in their first arguments uniformly in t. Then there exist constants $C_1 > 0$ and $C_2 > 0$ not depending on Δ such that*

$$\limsup_{t\to\infty} \|g_e(x(t),t)(u(x(t),t) - Y(x(t),t)\alpha_p - h(x(t)))\|_2 \leqslant C_1\sqrt{\mu_1^{-1}(\Delta)(1+L)} + C_2\mu_1^{-1}(\Delta).$$

Proof From the proof of Theorem 6.4, the solution $t \mapsto (x(t), e(t), \hat{\alpha}_p(t), \hat{\alpha}_m(t))$ exists for $t \geqslant 0$, is unique, and is continuously differentiable. Furthermore, by Proposition E.3 we have the following

uniform estimates

$$\sup_{t\geqslant 0}\|e(t)\|_2 \leqslant R,\ \sup_{t\geqslant 0}\|x(t)\|_2 \leqslant B_x,\ \sup_{t\geqslant 0}\|\hat{\alpha}_\ell(t)\|_2 \leqslant \sqrt{\frac{B_{\alpha_\ell}}{c_\ell}} + \|\alpha_{\ell,0}\|_2 + \sqrt{\frac{V(0)}{c_\ell}},\ \ell \in \{p,m\}.$$

Here, c_p (resp. c_m) is a constant depending only on the ambient dimension p and ψ_p (resp. m and ψ_m).

Now, applying that f, g, f_e, g_e, Y, Ψ, and h are all locally bounded in their first arguments uniformly in t, that $\nabla^2\psi_\ell$ is uniformly bounded from below for $\ell \in \{p,m\}$, and the assumption that σ'_Δ is B-bounded, we conclude that $\dot{x}(t)$, $\dot{e}(t)$, $\dot{\alpha}_p(t)$, and $\dot{\alpha}_m(t)$ are all uniformly bounded. Hence $x(t)$, $e(t)$, $\alpha_p(t)$, and $\alpha_m(t)$ are uniformly Lipschitz with Lipschitz constants that do not depend on Δ and L.

Next, the fact that f_e, g_e, Y, Ψ, and h are locally Lipschitz and locally bounded in their first arguments uniformly in t implies that $\dot{e}(t)$ is uniformly Lipschitz, with a Lipschitz constant that depends affinely on L. Therefore, by Proposition B.4, we have that:

$$\limsup_{t\to\infty}\|\dot{e}(t)\|_2 \leqslant C_1\sqrt{\mu_1^{-1}(\Delta)(1+L)},$$

for a constant C_1 that does not depend on Δ and L. Now for any t,

$$\begin{aligned}\|g_e(x(t),t)(u(x(t),t) - Y(x(t),t)\alpha_p - h(x(t)))\|_2 &\leqslant \|\dot{e}(t)\|_2 + \|f_e(e(t),t)\|_2 \\ &\leqslant \|\dot{e}(t)\|_2 + C_2\|e(t)\|_2,\end{aligned}$$

where C_2 does not depend on Δ and L. Taking the lim sup on both sides yields the claim. ∎