Lazily Adapted Constant Kinky Inference for Nonparametric Regression and Model-Reference Adaptive Control

Jan-Peter Calliess [a] Stephen J. Roberts [a] Carl Edward Rasmussen [b] Jan Maciejowski [b]

[a] *Dept. of Engineering Science, University of Oxford, UK.*

[b] *Dept. of Engineering, University of Cambridge, UK.*

Abstract

Techniques known as *Nonlinear Set Membership* prediction or *Lipschitz Interpolation* are approaches to supervised machine learning that utilise *presupposed* Lipschitz properties to perform inference over unobserved function values. Provided a bound on the true best Lipschitz constant of the target function is known a priori they offer convergence guarantees as well as bounds around the predictions. Considering a more general setting that builds on Lipschitz continuity, we propose an online method for estimating the Lipschitz constant online from function value observations that are possibly corrupted by bounded noise. Utilising this as a data-dependent hyper-parameter gives rise to a nonparametric machine learning method, for which we establish strong *universal approximation* guarantees. That is, we show that our prediction rule can learn any continuous function on compact support in the limit of increasingly dense data, up to a worst-case error that can be bounded by the level of observational error. We also consider applications of our nonparametric regression method to learning-based control. For a class of discrete-time settings, we establish convergence guarantees on the closed-loop tracking error of our online learning-based controllers. To provide evidence that our method can be beneficial not only in theory but also in practice, we apply it in the context of *nonparametric model-reference adaptive control (MRAC)*. Across a range of simulated aircraft roll-dynamics and performance metrics our approach outperforms recently proposed alternatives that were based on *Gaussian processes* and *RBF-neural networks*.

Key words: Machine Learning; Nonparametric Regression; System Identification; Model-Reference Adaptive Control.

1 Introduction

Among supervised learning methods, nonparametric regression techniques have attracted much attention due to their great flexibility and ability to learn rich function classes. Among many others, popular approaches include kernel methods such as *Gaussian Processes (GPs)* [20], the *NW-estimator* [18,25], local methods such as *LOESS* regression [11] as well as *Lipschitz Interpolation (LI)* [24,26]. In spite a wealth of classic as well as recent work that has shed light on the theoretical and practical properties of these methods, a common limitation remains: typically all results rest on the assumption of the knowledge of a suitable hyper-parameter that encodes a priori knowledge about the underlying learning target. While for some methods, especially for many of the kernel methods with certain choices of kernels, asymptotic consistency guarantees can be given for general classes of target functions, irrespective of the chosen hyper-parameter, in practice, the choice of hyper-parameter markedly impacts the predictive performance of the regression method for finite data sets. In *Lipschitz Interpolation (LI)* or *Nonlinear Set Membership (NSM)* methods [24,15,26], the hyper-parameter is a Lipschitz constant of the predictor. If set too low, the class of learnable target functions is too restrictive. If on the other hand the parameter is set too high, the resulting predictor will tend to overfit to noise in the data and might yield poor generalisation performance. Therefore, a common solution is to resort to hyper-parameter optimisation [20,5]. While often working well in practice, these approaches tend to be too computationally expensive to work with large data and to support online learning and adaptive control. Moreover, to the best of our knowledge, no theoretical insights into the learning-theoretic properties of the inferences with the hyper-parameter optimisers in place exist to date.

For Lipschitz Interpolation (LI), this paper addresses this gap. To this end, we propose a closed-form expres-

Email addresses: jan@robots.ox.ac.uk (Jan-Peter Calliess), sjrob@robots.ac.uk (Stephen J. Roberts), cer54@cam.ac.uk (Carl Edward Rasmussen), jmm@cam.ac.uk (Jan Maciejowski).

sion to estimate the Lipschitz constant from the data that is a modification of Strongin's estimator [23]. It has the benefit ~~to support~~ computationally tractable online updates but also offers robustness to (bounded) observational noise. We then propose to utilise the estimates in the LI rule to make predictions of function values at unobserved inputs. This combination of Lipschitz constant estimator and LI yields a new nonparametric regression method which we refer to as Lazily Adaptive Constant Kinky Inference (LACKI). For our LACKI method, we provide convergence and sample complexity bounds on the worst-case prediction error showing that our method can learn any continuous function both in an online as well as in an offline supervised learning setting. In the second part of the paper, we apply LACKI to learning-based control. We provide theoretical results on the closed-loop dynamics of a plant controlled by a learning-based controller that employs LACKI to learn about uncertain dynamics online. To illustrate some of the benefits and shortcomings of our approach, we compare LACKI with a selection of established regression methods on a model-reference adaptive control task where it outperforms competing approaches across a range of performance metrics and problem setups.

In contrast to most works on learning-based methods, we treat hyper-parameter estimation as part of the nonparametric learning process, both practically and in our theoretical analysis. In the absence of observational errors or if a noise bound is known, our approach is truly hyper-parameter free. If a bound on the noise is unknown, hyperparameter tuning might become necessary. In contrast to most other learning methods, this process merely entails a one-dimensional optimisation problem rather than a multi-dimensional one. Furthermore, our theory still quantifies worst-case error convergence bounds in the presence of a misspecified hyper-parameter, both in batch and online learning settings of continuous functions. Moreover, when our approach is employed in a basic class of online learning-based control settings, we prove convergence bounds of the closed-loop dynamics. We are unaware of existing work with similar theoretical guarantees based on so few a priori assumptions.

2 Lipschitz Interpolation with adapted Lipschitz constant estimates

Setting. Let $\mathcal{X}$ be an input space endowed with (pseudo-) metric $\mathfrak{d} : \mathcal{X}^2 \to \mathbb{R}_{\geq 0}$ and let $\mathcal{Y}$ be an output (vector) space endowed with a translation-invariant pseudo-metric $\mathfrak{d}_{\mathcal{Y}} : \mathcal{Y}^2 \to \mathbb{R}_{\geq 0}$. Let $\mathrm{Lip}(L) = \{\phi : \mathcal{X} \to \mathcal{Y} | \, \mathfrak{d}_{\mathcal{Y}}(\phi(x), \phi(x')) \leq L\, \mathfrak{d}(x, x'), \forall x, x' \in \mathcal{X}\}$ denote the set of Lipschitz continuous functions with Lipschitz constant L. The best Lipschitz constant of a function f is the smallest number L^* such that $f \in \mathrm{Lip}(L^*)$. A function is Lipschitz continuous if it has a finite Lipschitz constant.

Let $f : \mathcal{X} \to \mathcal{Y}$ be a *target* function we desire to learn in a supervised fashion. Often, we consider incremental learning, where over time (indexed by $n \in \mathbb{N}$), an increasing amount of data about f becomes available. To this end, we assume that, at time step n, we have access to a *sample* or *data set* $\mathcal{D}_n := \{(s_i, \tilde{f}_i) \,|\, i = 1, \ldots, N_n\}$ containing $N_n \in \mathbb{N}$ (possibly corrupted) sample values $\tilde{f}_i \in \mathcal{Y}$ of *target function* f at sample input $s_i \in \mathcal{X}$. The sampled function values are allowed to have *observational error* given by an error function $\mathfrak{e} : \mathcal{X} \to \mathcal{Y}$ which may model stochastic noise or systematic error. That is, we assume $\mathfrak{d}_{\mathcal{Y}}(\tilde{f}_i, f(s_i)) \leq \mathfrak{d}_{\mathcal{Y}}(0, \mathfrak{e}(s_i))$. For convenience, we may also write $\mathcal{D}_n = (G_n, \mathcal{Y}_n)$ where $G_n = \{s_i | i = 1, ..., N_n\} \subset \mathcal{X}$ is the collection (or *grid*) of sample inputs and $\mathcal{Y}_n = \{\tilde{f}_i | i = 1, ..., N_n\} \subset \mathcal{Y}$ is the ~~pertaining~~ sequence of observed function values. It is our aim to learn target function f in the sense that we utilise the available data $\mathcal{D}_n$ to infer *predictions* $\hat{\mathfrak{f}}_{\mathfrak{n}}(x)$ of $f(x)$ at unobserved *query inputs* $x \notin G_n$. In our context, the evaluation of $\hat{\mathfrak{f}}_n$ is what we refer to as *(inductive) inference* or *prediction* and $\hat{\mathfrak{f}}_{\mathfrak{n}}$ is referred to as the *predictor*.

Learning rule. We will now consider a simplified version of *Kinky Inference (KI)* [6] – a class of nonparametric learning rules that encompasses a host of other methods such as NSM methods [15] and standard Lipschitz Interpolation [24,3,26]. As a special case, we will then define our proposed method that incorporates an adaptive estimator of the Lipschitz constant of the target.

Definition 1 (Kinky inference (KI) rule) *Given access to a sample set $\mathcal{D}_n$ and an input space pseudo-metric $\tilde{\mathfrak{d}}(\cdot, \cdot; \Xi(n)) : \mathcal{X}^2 \to \mathbb{R}$ parameterised by $\Xi(n)$, we define the KI predictor by $\hat{\mathfrak{f}}_{\mathfrak{n}}(\cdot; \Xi(n), \mathcal{D}_n) : \mathcal{X} \to \mathcal{Y}$ to perform inference over function values as per:*

$$\hat{\mathfrak{f}}_{\mathfrak{n}}(x; \Xi(n), \mathcal{D}_n) := \frac{1}{2}\mathfrak{u}_n(x; \Xi(n)) + \frac{1}{2}\mathfrak{l}_n(x; \Xi(n)). \qquad (1)$$

Here, $\mathfrak{u}_n(\cdot; \Xi(n)), \mathfrak{l}_n(\cdot; \Xi(n)) : \mathcal{X} \to \mathbb{R}^m$ are defined by $\mathfrak{u}_n(x; \Xi(n)) := \min_{i=1,\ldots,N_n} \tilde{f}_i + \tilde{\mathfrak{d}}(x, s_i; \Xi(n))$ and $\mathfrak{l}_n(x; \Xi(n)) := \max_{i=1,\ldots,N_n} \tilde{f}_i - \tilde{\mathfrak{d}}(x, s_i; \Xi(n))$, respectively.

The computational effort for making a prediction is ~~in~~ $\mathcal{O}(N_n\, M)$ where M is the effort for evaluating the pseudo-metric. However, it is possible to apply (generalised) nearest-neighbour techniques to reduce this effort to expected logarithmic growth in the number of sample points [3,6].

A special case arises for the choice of $\tilde{\mathfrak{d}}(x, y; \Xi(n)) = L(n)\, \|x - y\|$ which is referred to as *Lipschitz Interpolation* [3] or as *Nonlinear Set Interpolation* [15]. Here the hyper-parameter $\Xi(n) = L(n)$ is the supposed Lipschitz constant of the target. And, it is easy to show that the predictor $\hat{\mathfrak{f}}_{\mathfrak{n}}(\cdot; L(n), \mathcal{D}_n)$ is Lipschitz continuous with

Lipschitz constant $L(n)$ [6]. Typically, this constant is assumed to be either known a priori or estimated from the data, e.g. [23,15,5]. Unfortunately, little is understood about the effects of the previously proposed parameter estimation techniques on the predictor's performance and about the impact of observational noise.

Similarly to the kernel learning literature, the generality afforded by allowing the specification of pseudo-metrics rather than metrics allows us to support automated relevance determination or taking advantage of peridocity in the data which we can seek to discover in a data-driven fashion by hyperparameter tuning [5]. However, for ease of exposition, we will henceforth make some simplifying assumptions:

Assumption 2 *(1) The output space is an m-dimensional normed space, $\mathcal{Y} \subseteq \mathbb{R}^m$ with $\mathfrak{d}_{\mathcal{Y}}(y, y') = \|y - y'\|_\infty, \forall y, y' \in \mathcal{Y}$.*
(2) The input space is a d-dimensional normed space, $\mathcal{X} \subseteq \mathbb{R}^d$ with $\mathfrak{d}(x, x') = \|x - x'\|_\infty, \forall x, x' \in \mathcal{X}$.
(3) Furthermore, observational errors are bounded by some $\bar{\mathfrak{e}} := \sup_{x \in \mathcal{X}} \|\mathfrak{e}(x)\|_\infty$.

Under these simplifying assumptions our *Lazily Adapted Kinky Inference (LACKI)* learning rule can be defined as follows:

For notational convenience, for two sets $S, S' \subset \mathcal{X}$ of inputs we define $U(S, S') := \{(s, s') \in S \times S' | \, \|s - s'\|_\infty > 0\}$ and let $U_n := U(G_n, G_n)$ be the set of all pairs of distinct sample inputs.

Definition 3 (LACKI rule) *The Lazily Adapted Lipschitz Constant Kinky Inference (LACKI) rule computes a KI predictor $\hat{\mathfrak{f}}_{\mathfrak{n}}$ as per ~~Defn.~~ 1, but where $\tilde{\mathfrak{d}}(x, x'; L(n)) = L(n)\, \mathfrak{d}(x, x')$ and where we set*

$$L(n) := \max\Big\{0, \max_{(s,s') \in U_n} \frac{\left\|\tilde{f}(s) - \tilde{f}(s')\right\|_\infty - \lambda}{\|s - s'\|_\infty}\Big\}. \quad (2)$$

where ordinarily we set $\lambda := 2\bar{\mathfrak{e}}$.

Remark 4 *Note, $L(n)$ is a modified Strongin estimate [23] of a Lipschitz constant. Here λ compensates for the influence of observational noise and thereby, is guaranteed to prevent $L(n)$ from diverging as n grows, if $\lambda \geq 2\bar{\mathfrak{e}}$ [4]. If set freely, λ is a design parameter acting as a (hyper-) hyper-parameter of the nonparametric LACKI prediction rule. Being a Lipschitz constant of the predictor, boundedness of $L(n)$ can cause the predictor to smooth out observational noise. On the other hand, if we erroneously set λ set to a value below $2\bar{\mathfrak{e}}$ (e.g. because we underestimate the observational error level) $L(n)$ might become infinite in the limit of dense data (effectively learning the noise gradient). In this case, the input distance terms $L(n)\, \mathfrak{d}(x, s_i)$ become dominant in the LACKI prediction rule (1) and LACKI effectively starts predicting like 1-nearest neighbour regression (inheriting its convergence properties). For an illustration, cf. Fig. 1 (lacki vs lacki-nonoise). Similar to the case of Gaussian processes and other nonparametric machine learning approaches, in the absence of a good guess of $\bar{\mathfrak{e}}$, we can attempt to tune hyper-parameter λ by cross-validation, minimising empirical risk (in lieu ~~to~~ the approach considered in [5]). Note, since we make no distributional assumptions about the observational noise (in particular it could be systematic error), our convergence guarantees we will derive below will generally have to depend on it and, our worst-case prediction error bounds we will derive below could become arbitrarily poor in the case of unbounded observational errors. This is unavoidable in worst-case analysis without restrictive assumptions on the noise. In the case of i.i.d. stochastic noise that is unbounded, we recommend the reader to consider employing the POKI-LC estimator of the Lipschitz hyperparameters introduced in [5].*

Next, consider an online learning situation where the available data grows incrementally such that $G_{n+1} = G_n \cup \{s_{n+1}\}, \forall n$. We can define an incremental update rule inductively as follows: $L(n+1) :=$

$$\max\Big\{L(n), \max_{(s,s') \in U(G_n, \{s_{n+1}\})} \frac{\left\|\tilde{f}(s) - \tilde{f}(s')\right\|_\infty - \lambda}{\|s - s'\|_\infty}\Big\} \quad (3)$$

for $n \in \mathbb{N}$ and where $L(0) := 0$.

The effort for computing $L(n+1)$ in time step $n+1$ based on the newly arrived sample point and the previous Lipschitz constant estimate $L(n)$ is in $\mathcal{O}(M\, N_n)$. It is easy to see that the incremental update rule yields estimates consistent with the batch estimate defined in Eq. 2.

3 Learning Theoretical Analysis

3.1 Properties

We will now establish several properties of the LACKI rules including boundedness of the predictors, sample-consistency and Lipschitz continuity. Most importantly however, we will show that the LACKI is a universal approximator, in the sense that it can be be set to learn any continuous function with arbitrarily low worst-case error up to a bound that depends on the observational errors in the data. First, we establish Lipschitz continuity and sample-consistency. This allows us to prove that LACKI can learn any Lipschitz function. Note, some universal approximators, such as *radial basis function networks (RBFNs)* with Gaussian kernels, are provably Lipschitz. Therefore, learning any continuous function can be interpreted as learning some Gaussian RBFN with an observational error level that absorbs the discrepancy between the RBFN and the ground truth. Since a finite RBFN with smooth, bounded-derivative kernel is provably Lipschitz and since we can learn any Lipschitz function with LACKI up to the level of observational error,

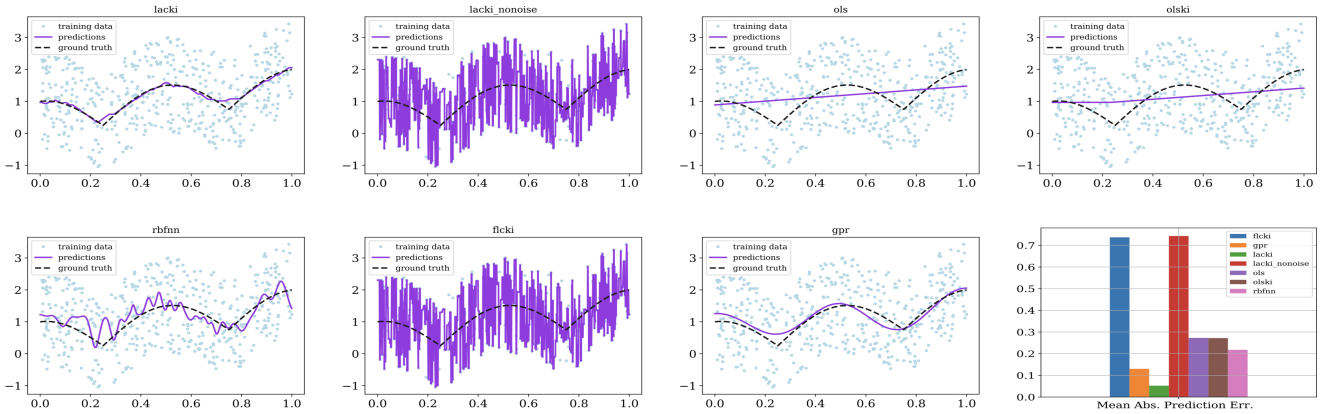

Fig. 1. The predictors of several regression methods for a target function $f : x \mapsto |\cos(2\pi x)| + x$ (dashed line). The $N_n = 500$ observations (light blue dots) in the sample were perturbed by uniform noise drawn i.i.d. from the interval $[-1.5, 1.5]$. The predictions of the trained models are plotted in purple. From top-left to bottom-right: LACKI: Our LACKI method with correctly set noise parameter, i.e. $\lambda = 1.5$. LACKI-nonoise: LACKI with falsely set noise parameter $\lambda = 0$. OLS: Ordinary least squares regression. OLSKI: Kinky inference with $L(n)$ inferred as gradient norm of fitted weights of OLS (following [15]). The inability of OLS to model the high-variation nonlinearity extends to OLSKI via a Lipschitz constant estimate that is too low. RBFNN: Radial-basis function neural network fitted with 20 neurons. FLCKI: Kinky inference with $L(n)$ set to the fixed Lipschitz constant of the fitted RBFNN (following [15]). Note, how the approach exacerbates the fitting issues of the RBFNN due to the capacity increase. GPR: posterior mean of a Gaussian process regressor with tuned hyper-parameters using an RBF kernel prior. The plot shows the best result out of 100 restarts of the tuning procedure. Bar plots: Absolute mean predictions error on test set containing 2000 independently drawn samples.

we can learn the continuous ground-truth up to the approximation error of the RBFN.

Following this outline, we will now proceed to establish the desired properties formally.

Lemma 5 (Boundedness) *Irrespective of the boundedness of input space $\mathcal{X}$ and assuming finite sample size $N_n = |\mathcal{D}_n| < \infty$, the predictor $\hat{\mathfrak{f}}_\mathfrak{n} : \mathcal{X} \to \mathcal{Y}$ is bounded. In particular, we have for all $x \in \mathcal{X}$:*

$$\left\|\hat{\mathfrak{f}}_\mathfrak{n}(x)\right\|_\infty \leq \max_i \left\|\tilde{f}_i\right\|_\infty + \frac{L(n)}{2} \max_{i,j} \|s_i - s_j\|_\infty < \infty.$$

Proof: *Let $D = \max_{i,j=1,\dots,N_n} \|s_i - s_j\|$ and for the kth output dimension let $F_k = \max_{i=1,\dots,N_n} \left|\tilde{f}_{i,k}\right|$. Utilising the definition of the predictor and the triangle inequality we see that, for any $x \in \mathcal{X}$ and any output dimension k, there are some $i, j \in \{1, ..., N_n\}$ such that we have:* $\hat{\mathfrak{f}}_{n,k}(x) = \frac{\tilde{f}_{j,k}+\tilde{f}_{i,k}}{2} + \frac{L(n)}{2}\left(\|x - s_i\|_\infty - \|x - s_j\|_\infty\right) \leq \frac{\tilde{f}_{j,k}+\tilde{f}_{i,k}}{2} + \frac{L(n)}{2}\|s_j - s_i\|_\infty \leq F_k + \frac{L(n)}{2} D < \infty.$ □

As promised, we establish that the predictors of the LACKI inference rule are Lipschitz continuous:

Lemma 6 (Lipschitz continuity) *The prediction functions $\hat{\mathfrak{f}}_\mathfrak{n}$ $(n \in \mathbb{N})$ are Lipschitz continuous with Lipschitz constant $L(n)$. That is, $\forall n \in \mathbb{N} : \hat{\mathfrak{f}}_\mathfrak{n} \in Lip(L(n))$.*

Proof: *It is easy to show that the one-dimensional mappings of the form $x \mapsto \ell \|x - x'\|$ are $\ell-$ Lipschitz continuous for any choices of ℓ and inputs x'. Furthermore, taking point-wise* max, min *as well as averages of Lipschitz continuous functions is known to not change their Lipschitz properties (e.g. cf. [6]). Therefore, the output-component predictors $\hat{\mathfrak{f}}_{n,j}$ $(j = 1, ..., m)$ are $L(n)$- Lipschitz.* □

We now establish how well our LACKI rule can interpolate the training data as function of the noise bound and regularisation parameter λ:

Lemma 7 (Sample-consistency) *The LACKI rule is sample-consistent (up to $\frac{\lambda}{2}$). That is,*

$$\forall q \in \{1, \dots, N_n\} : \hat{\mathfrak{f}}_\mathfrak{n}(s_q) \in \mathfrak{B}_{\frac{\lambda}{2}}\left(\tilde{f}_q\right)$$

where $\mathfrak{B}_{\frac{\lambda}{2}}\left(\tilde{f}_q\right) = \{x \in \mathcal{Y} | \left\|x - \tilde{f}_q\right\|_\infty \leq \frac{\lambda}{2}\}$ denotes the $\frac{\lambda}{2}$-ball around the observation $\tilde{f}_q$;
and, $\left\|f(s_q) - \hat{\mathfrak{f}}_\mathfrak{n}(s_q)\right\|_\infty \leq \frac{\lambda}{2} + \|\mathfrak{e}(s_q)\|_\infty \leq \frac{\lambda}{2} + \bar{\mathfrak{e}}$.

Proof: *For ease of notation, we will confine our proof to the case of one-dimensional outputs ($d = 1$). The multi-dimensional case follows trivially from the one-dimensional result by applying it to each output component function. Let $n \in \mathbb{N}$ be fixed and, for ease of notation, write $L := L(n)$. Let $j, k \in \{1, \dots, N_n\}$ with $j \in argmin_i \tilde{f}_i + L\|s_i - s_q\|_\infty$, $k \in argmax_i \tilde{f}_i - L\|s_i - s_q\|_\infty$. By definition of $\hat{\mathfrak{f}}_\mathfrak{n}$ we*

have:

$$\hat{\mathfrak{f}}_{\mathfrak{n}}(s_q) = \frac{1}{2}\underbrace{\left(\tilde{f}_j + L\left\|s_j - s_q\right\|_\infty\right)}_{:=B} + \frac{1}{2}\underbrace{\left(\tilde{f}_k - L\left\|s_k - s_q\right\|_\infty\right)}_{=:A}.$$

(i) Firstly, we show $A \in [\tilde{f}_q, \tilde{f}_q + \lambda]$: If $k = q$, this holds trivially true since then $A = \tilde{f}_q$. So, assume $k \neq q$. We have $\tilde{f}_k \geq \tilde{f}_k - L\left\|s_k - s_q\right\|_\infty \geq \tilde{f}_q - L\left\|s_q - s_q\right\|_\infty = \tilde{f}_q$ where the second inequality holds due to $k \in argmax_i \tilde{f}_i - L\left\|s_i - s_q\right\|_\infty$. That is,

$$A = \tilde{f}_k - L\left\|s_k - s_q\right\|_\infty \geq \tilde{f}_q. \tag{4}$$

On the other hand, since $L \geq \max_{(s,s')\in U_n} \frac{\left|\tilde{f}(s) - \tilde{f}(s')\right| - \lambda}{\|s - s'\|_\infty}$ we have in particular: $L \geq \frac{\left|\tilde{f}_k - \tilde{f}_q\right| - \lambda}{\|s_k - s_q\|_\infty}$. Thus, $L\left\|s_k - s_q\right\|_\infty + \lambda \geq \left|\tilde{f}_k - \tilde{f}_q\right| = \tilde{f}_k - \tilde{f}_q$. Hence, $\tilde{f}_q + \lambda \geq \tilde{f}_k - L\left\|s_k - s_q\right\|_\infty = A$. In conjunction with (4), we have shown: $A \in [\tilde{f}_q, \tilde{f}_q + \lambda]$.

(ii) The proof of $B \in [\tilde{f}_q - \lambda, \tilde{f}_q]$ is completely analogous to that of (i) and hence, is omitted.

(iii) Together, the statements in (i) and (ii) entail $\hat{\mathfrak{f}}_{\mathfrak{n}}(s_q) = \frac{1}{2}A + \frac{1}{2}B \in [\tilde{f}_q - \frac{\lambda}{2}, \tilde{f}_q + \frac{\lambda}{2}]$. Hence, $\left\|\hat{\mathfrak{f}}_{\mathfrak{n}}(s_q) - \tilde{f}(s_q)\right\|_\infty \leq \frac{\lambda}{2}$. Moreover, for any sample input s_q, we have: $\hat{\mathfrak{f}}_{\mathfrak{n}}(s_q) = f(s_q) + \phi_q + \psi_q$ with $\|\psi_q\|_\infty \leq \frac{\lambda}{2}, \|\phi_q\|_\infty \leq \|\mathfrak{e}(s_q)\|_\infty \leq \bar{\mathfrak{e}}$. Hence, $\left\|\hat{\mathfrak{f}}_{\mathfrak{n}}(s_q) - f(s_q)\right\|_\infty = \|\phi_q + \psi_q\|_\infty \leq \frac{\lambda}{2} + \|\mathfrak{e}(s_q)\|_\infty \leq \frac{\lambda}{2} + \bar{\mathfrak{e}}$. □

3.1.1 Prediction error bounds and consistency

To asses our learning rule, we might be interested the discrepancy $\mathfrak{d}_{\mathcal{F}}(\hat{\mathfrak{f}}_{\mathfrak{n}}, f)$ between the predictor $\hat{\mathfrak{f}}_{\mathfrak{n}}$ and the target function f relative to some metric $\mathfrak{d}_{\mathcal{F}}$ between functions in the space $\mathcal{F}$ of continuous functions. In statistics, a typical choice is the mean-square error metric assessed with respect to some distribution over inputs, the function space and the noise. However, in many safety-critical applications, often arising in control, worst-case error considerations are of greater value, leading to a worst-case metric $\mathfrak{d}_{\mathcal{F}}(f, g) = \sup_{x\in I}\left\|f(x) - g(x)\right\|_\infty$ for some subset $I \subseteq \mathcal{X}$ of queries one finds interesting to take into consideration.

Therefore, we will now establish worst-case consistency guarantees of our LACKI inference rules. That is, we shall study the worst-case error sequence $\mathcal{E}^\infty := \left(\mathcal{E}_n^\infty\right)_{n\in\mathbb{N}}$, $\mathcal{E}_n^\infty := \sup_{x\in I}\left\|\hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x)\right\|_\infty$ for data $\mathcal{D}_n$ that becomes increasingly dense over time relative to a set of query inputs $I \subseteq \mathcal{X}$. To clarify the latter concept, consider the sequence of grids $\left(G_n\right)_{n\in\mathbb{N}}$. We say this sequence converges to a set that *becomes dense relative to a set I in the limit of large n* if we can use points in the sequence to approximate any points in I with increasing accuracy. That is, if $\forall\epsilon > 0, x \in I \exists n_0 \forall n \geq n_0 \exists g \in G_n : \left\|x - g\right\|_\infty < \epsilon$. If the rate at which this happens is independent of x then we say that the grid sequence becomes dense *uniformly*. This is the case iff $\forall\epsilon > 0 \exists n_0 \forall n \geq n_0, x \in I \exists g \in G_n : \left\|x - g\right\|_\infty < \epsilon$. To make the rates explicit in our notation, we list the following general definitions:

Definition 8 (Becoming dense, rates, $\overset{r}{\longrightarrow}, \overset{r}{\rightsquigarrow}, \overset{r}{\twoheadrightarrow}$) *Let $\mathcal{X}$ be a space endowed with a norm $\|\cdot\|$. Let $r : \mathbb{N} \to \mathbb{R}$ be a "rate" function. that vanishes, that is, with $\lim_{n\to\infty} r(n) = 0$ (i.e. $r \in o(1)$).*

- *The sequence $s = \left(s_n\right)_{n\in\mathbb{N}}$ of points in $\mathcal{X}$ is said to converge to a point $x \in \mathcal{X}$* with rate *r (denoted by $s \overset{r}{\longrightarrow} x$) iff $\forall n \in \mathbb{N} : \|x - s_n\| \leq r(n)$ and $r(n) \overset{n\to\infty}{\longrightarrow} 0$.*
  - *The sequence s is said to converge to a set $\mathbb{S} \subset \mathcal{X}$ with rate $r : \mathbb{N} \to \mathbb{R}$ (denoted by $s \overset{r}{\longrightarrow} \mathbb{S}$) iff $\forall n \in \mathbb{N} : \inf_{x\in\mathbb{S}}\left\|x - s_n\right\|_\infty \leq r(n)$ and $r(n) \overset{n\to\infty}{\longrightarrow} 0$.*
- *A sequence of sets $S = \left(S_n\right)_{n\in\mathbb{N}}$ is said to* become dense *relative to $x \in \mathcal{X}$* with rate *r (denoted by $S \overset{r}{\rightsquigarrow} x$) iff S contains a point sequence that converges to x with that rate. That is, iff $\exists s = \left(s_n\right)_{n\in\mathbb{N}} : s \overset{r}{\longrightarrow} x \wedge \forall n : s_n \in S_n$.*
  - *Similarly, the sequence of sets S is said to become dense relative to a* set *of points $\mathbb{S} \subset \mathcal{X}$ (denoted by $S \rightsquigarrow \mathbb{S}$) iff it becomes dense relative to all points of $\mathbb{S}$, i.e. iff $\forall x \in \mathbb{S} : S \overset{r_x}{\rightsquigarrow} x$ for some vanishing rate $r_x : \mathbb{N} \to \mathbb{R}$.*
  - *The sequence is becoming dense relative to $\mathbb{S}$* uniformly *(denoted by $S \twoheadrightarrow \mathbb{S}$) iff there is a single vanishing rate for all $x \in \mathbb{S}$. That is, if $\exists r : \mathbb{N} \to \mathbb{R} : \lim_{n\to\infty} r(n) = 0 \wedge \sup_{x\in\mathbb{S}} \inf s_n \in S_n \left\|s_n - x\right\| \leq r(n), \forall n$. Function r is referred to as the convergence rate and we write $S \overset{r}{\twoheadrightarrow} \mathbb{S}$ to denote that S becomes dense relative to $\mathbb{S}$ with uniform rate r.*

Theorem 9 (Lipschitz learnability) *Assume observational errors are bounded by $\bar{\mathfrak{e}} < \infty$ and that the target $f : \mathcal{X} \to \mathcal{Y}$ is Lipschitz continuous, that is $\exists L^* \in \mathbb{R} : f \in Lip(L^*)$. Then we have:*

***(A)** If the grid becomes dense (pointwise), the point-wise worst-case error vanishes up to $\frac{\lambda}{2} + \bar{\mathfrak{e}}$: If $\forall x \in I \subset \mathcal{X} \exists r_x \in o(1) : L(\cdot)r_x(\cdot) \in o(1) \wedge \left(G_n\right)_{n\in\mathbb{N}} \overset{r_x}{\rightsquigarrow} x$ then we have:*

$$\forall x \in I : \left(\left\|\hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x)\right\|_\infty\right)_{n\in\mathbb{N}} \xrightarrow{\varrho_x} [0, \frac{\lambda}{2} + \bar{\mathfrak{e}}]$$

where for the error convergence rate ϱ_x we have

$\varrho_x(n) \leq (L(n) + L^*)r_x(n), \forall n \in \mathbb{N}$.

(B) *If the grid becomes dense in $I \subset \mathcal{X}$ uniformly, then the worst-case prediction error vanishes uniformly (up to $\frac{\lambda}{2} + \bar{\mathfrak{e}}$). That is,*

if $\exists r \in o(1) : L(\cdot)r_x(\cdot) \in o(1) \wedge (G_n) \overset{r}{\twoheadrightarrow} I$ then we have:

$$\mathcal{E}^\infty \xrightarrow{\varrho} [0, \frac{\lambda}{2} + \bar{\mathfrak{e}}]$$

where for the uniform error convergence rate ϱ we have $\varrho(n) \leq (L(n) + L^*)r(n), \forall n \in \mathbb{N}$.

Proof: *We have established that the predictors $\hat{\mathfrak{f}}_{\mathfrak{n}}(\cdot)$ of the LACKI rule are $L(n)$- Lipschitz (Lemma 6) and sample-consistent up to level $\frac{\lambda}{2}$ (Lemma 7). For any input $x \in \mathcal{X}$ let ξ_n^x denote a nearest neighbour of x in grid G_n. That is, $\xi_n^x \in \arg\inf_{s \in G_n} \|x - s\|_\infty$. Since G_n is assumed to become dense in the input domain $\mathcal{X}$, for any input x there is a rate function $r_x : \mathbb{N} \to \mathbb{R}_{\geq 0}$ such that $r_x(n) \overset{n\to\infty}{\longrightarrow} 0$ and $\|x - \xi_n^x\|_\infty \leq r_x(n), \forall n \in \mathbb{N}$. In the case of uniform convergence a rate function can be chosen independently of x and will be denoted by r rather than r_x.*

(A) *For all $n \in \mathbb{N}$ and $x \in \mathcal{X}$ we have:*

$$\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(\xi_n^x) \right\|_\infty \overset{(i)}{\leq} \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x), \hat{\mathfrak{f}}_{\mathfrak{n}}(\xi_n^x) \right\|_\infty + \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(\xi_n^x), f(\xi_n^x) \right\|_\infty \tag{5}$$

$$\overset{(ii)}{\leq} \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - \hat{\mathfrak{f}}_{\mathfrak{n}}(\xi_n^x) \right\|_\infty + \frac{\lambda}{2} + \|\mathfrak{e}(\xi_n^x)\|_\infty \tag{6}$$

$$= \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - \hat{\mathfrak{f}}_{\mathfrak{n}}(\xi_n^x) \right\|_\infty + \frac{\lambda}{2} + \bar{\mathfrak{e}} \tag{7}$$

$$\overset{(iii)}{\leq} L(n) \|x - \xi_n^x\|_\infty + \frac{\lambda}{2} + \bar{\mathfrak{e}} \tag{8}$$

Here, (i) follows from the triangle inequality, (ii) leverages Lemma 7 and (iii) follows by Lipschitz continuity of the predictors (Lemma 6). Thus, for $x \in \mathcal{X}, n \in \mathbb{N}$:

$$\begin{aligned} 0 &\leq \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty \\ &\leq \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(\xi_n^x) \right\|_\infty + \|f(\xi_n^x) - f(x)\|_\infty \\ &\overset{(\dagger)}{\leq} (L(n) + L^*) \|x - \xi_n^x\|_\infty + \frac{\lambda}{2} + \bar{\mathfrak{e}} \end{aligned}$$

where (†) follows from (8) and the presupposed Lipschitz continuity of f.
Since by assumption, $\|x - \xi_n^x\|_\infty \leq r_x(n), \forall n$, this implies:

$$\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty \in \left[0, (L(n) + L^*)r_x(n) + \frac{\lambda}{2} + \bar{\mathfrak{e}}\right], \forall n.$$

By assumption, $r_x(n), L(n)r_x(n) \overset{n\to\infty}{\to} 0, \forall x$ and hence, $\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty$ converges to $[0, \frac{\lambda}{2} + \bar{\mathfrak{e}}], \forall x$ with rate $\varrho_x \leq (L(n) + L^*)r_x(n)$.

(B) *Proceeding analogously as before, but utilising uniform convergence with rate r, we obtain:*

$$\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty \in \left[0, (L(n) + L^*)r(n) + \frac{\lambda}{2} + \bar{\mathfrak{e}}\right], \forall x \forall n.$$

By assumption, $L(n)r(n) \in o(1)$ and thus, $\lim_{n\to\infty} L(n)r(n) = 0$. Hence,

$$\mathcal{E}^\infty = \left(\sup_{x \in I} \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty \right)_{n \in \mathbb{N}} \xrightarrow{\varrho} [0, \frac{\lambda}{2} + \bar{\mathfrak{e}}]$$

with rate ϱ such that $\varrho(n) \leq (L(n) + L^*)r(n), \forall n$. □

Note a necessary condition was that the product of $L(n)$ and the rate was in $o(1)$, that is, vanishing in the limit of $n \to \infty$. A sufficient condition for this to hold is if $L(n)$ is guaranteed to be bounded (assuming the rate is vanishing). It is easy to show that $\exists \bar{L} < \infty \forall n \in \mathbb{N} : L(n) \leq \bar{L}$, as long as parameter $\lambda \geq 2\bar{\mathfrak{e}} + q$ for any $q \geq 0$ [4]. This yields the following result:

Corollary 10 *With definitions and assumption as before, if the parameter λ is chosen to be $2\bar{\mathfrak{e}} + q$ for some arbitrary $q \geq 0$ then convergence to the ground truth is guaranteed (up to an twice the observational error and a term dependent on q). In particular, if the data becomes dense uniformly in $I \subseteq \mathcal{X}$ with a rate of $r(n)$ then, for some $\bar{L} \in [0, L^*]$ and any $n \in \mathbb{N}$, we have*

$$\sup_{x \in I} \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\| \leq (\bar{L} + L^*)r(n) + \frac{q}{2} + 2\bar{\mathfrak{e}} \overset{n\to\infty}{\longrightarrow} \frac{q}{2} + 2\bar{\mathfrak{e}}. \tag{9}$$

Of course in the absence of observational errors, one can choose $\lambda = 0$. In this case, the corollary implies that LACKI will learn the ground-truth arbitrarily well in the limit of infinitely dense data.

Having established that our LACKI rule can learn any Lipschitz function with any Lipschitz constant, we will now attend to extend the results to non-Lipschitz functions. In preparation of the necessary derivations we will first rehearse universality and Lipschitz properties of radial basis function networks. Park and Sandberg derived universal approximation guarantees for radial-basis function networks [19]. In particular, on page 252 in their article the authors make an assertion that translates to our notation as follows:

Lemma 11 (Expressiveness of RBFNs) *Assume $\mathcal{X} \subseteq \mathbb{R}^d$ is compact. Given parameter vector $\theta :=$*

$(w_1, \ldots, w_m, \sigma_1, ..., \sigma_m, c_1, \ldots, c_m)$ and kernel function $K : \mathcal{X} \to \mathcal{Y}$ let $\beta(\cdot;\theta) = \sum_{i=1}^{m} w_i K(\frac{\cdot - c_i}{\sigma_i})$ denote a radial basis function network (RBFN). Assume $K : \mathbb{R}^d \to \mathbb{R}$ is continuous and has non-vanishing integral, i.e. $\int_{\mathbb{R}^d} K(x)\, dx \neq 0$. Then, the set $S_K := \{\beta(\cdot;\theta) | m \in \mathbb{N}, \theta \in \mathbb{R}^{3m}\}$ of all RBFNs is uniformly dense in the set $C(\mathcal{X})$ of continuous functions on compact domain $\mathcal{X}$. That is, $\forall f \in C(\mathcal{X}) \forall \epsilon > 0 \exists m, \theta \in \mathbb{R}^{3m} : \sup_{x \in \mathcal{X}} |f(\cdot) - \beta(\cdot;\theta)| < \epsilon$.

Remark 12 *We note that, for any finite-dimensional parameter θ, any RBFN $\beta(\cdot;\theta)$ is Lipschitz continuous as long as the kernel K is. This can be seen by applying Lipschitz arithmetic (see appendix of [4]) which allows us to conclude that the Lipschitz constant of RBFN $\beta(\cdot;\theta) = \sum_{i=1}^{m} w_i K(\frac{\cdot - c_i}{\sigma_i})$ is given by $L_\beta = \sum_{i=1}^{m} \left|\frac{w_i}{\sigma_i}\right| L_K$ where $L_K \in \mathbb{R}_{\geq 0}$ denotes a Lipschitz constant of K. By the same Lemma it is easy to see that choosing the Gaussian kernel for K satisfies both the Lipschitz requirement as well as the integrability requirements of Lemma 11. As a by-product this means that on a compact support, any continuous function can be approximated by some Lipschitz function with arbitrarily small, positive worst-case error $\epsilon > 0$. Note, it may well be the case that the Lipschitz constant of the approximator grows with decreasing approximation error bound ϵ. We consider this to be inevitable when the approximated function is not Lipschitz.*

Harnessed with these preparatory statements we can move on to show that the LACKI rule can be set up to learn any continuous function up to arbitrary low error.

Theorem 13 (Universality and consistency) *Assume we are given a sequence $\left(\mathcal{D}_n\right)_{n \in \mathbb{N}}$ of samples with observational errors bounded by $\bar{\mathfrak{e}} \in \mathbb{R}_{\geq 0}$. We set the parameters of the LACKI rule to and $\lambda := 2\bar{e} + q$ for some arbitrary $q > 0$. In this theorem, we assume that the set of interest $I \subseteq \mathcal{X}$ is compact.* ***Then, we have:***

The LACKI rule as per Def. 3 is a universal approximator in the following sense: If the sequence of input grids $\left(G_n\right)_{n \in \mathbb{N}}$ converges to I (uniformly) then the sequence of predictors $\left(\hat{\mathfrak{f}}_{\mathfrak{n}}\right)_{n \in \mathbb{N}}$ computed by the LACKI rule (uniformly) converges to any continuous target $f : \mathcal{X} \to \mathbb{R}$ up to error $2\bar{\mathfrak{e}} + \frac{3q}{2}$. That is, the following holds true:

- *(I) Let $x \in I$. If $\exists r_x \in o(1) : (G_n) \overset{r_x}{\rightsquigarrow} x$ then*

$$\exists C \in \mathbb{R} : \left(\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty\right) \overset{Cr_x}{\longrightarrow} [2\bar{\mathfrak{e}} + \frac{3q}{2}].$$

- *(II) If $\exists r \in o(1) : (G_n) \overset{r}{\twoheadrightarrow} I$ then:*

$$\exists C \in \mathbb{R} : \mathcal{E}^\infty \overset{Cr}{\longrightarrow} [2\bar{\mathfrak{e}} + \frac{3q}{2}].$$

Proof: We choose any parameter $\lambda = 2\bar{\mathfrak{e}} + q$ with $q > 0$. As observed in Rem. 12, Lemma 11 allows us to infer that there exists a Lipschitz function h that approximates the target with worst-case error of at most $\frac{q}{2}$. That is, $\sup_{x \in \mathcal{X}} \|h(x) - f(x)\|_\infty \leq \frac{q}{2}$.

Consequently, there exists a function $\phi' : \mathcal{X} \to \mathcal{Y}$ with $\sup_x \|\phi'(x)\|_\infty \leq \frac{q}{2}$ accounting for the discrepancy between the Lipschitz function h and the target f:

$$f = h + \phi'.$$

Furthermore, we define ϕ to be the bounded observational noise. Hence, we have $\tilde{f} = f + \phi$ and $\sup_x \|\phi(x)\|_\infty \leq \bar{\mathfrak{e}}$. Combining both functions into $\psi := \phi + \phi'$, we can write $\tilde{f} = h + \psi$ with $\sup_x \|\psi(x)\|_\infty \leq \frac{q}{2} + \bar{\mathfrak{e}} =: \bar{\nu}$.

This can be interpreted as follows: Instead of viewing the given sample as being generated by target f (with some observational error ϕ) we can view the sample as being generated by the Lipschitz function h corrupted by the extended "observational noise" ψ accounting for both the original observational error and the discrepancy between the target and Lipschitz function h. This gives us a reduction to the case of learning Lipschitz functions with observational error bounded by $\bar{\nu}$. Firstly, we note that $\lambda = 2\bar{\mathfrak{e}} + q = 2\bar{\nu}$ (which entails that the sequence $\left(L(n)\right)_{n \in \mathbb{N}}$ is bounded by some constant $\bar{L} = \sup_{x \neq x'} \frac{\|h(x) - h(x')\|_\infty - q}{\|x - x'\|_\infty} \leq L_h$). Linking this with Theorem 9, we obtain all the desired statements with regard to learning h. These can easily be converted into statements about learning f by adding the worst-case difference $\frac{q}{2}$ between f and h to all error bounds. For example, leveraging $\sup_x \|\phi'(x)\|_\infty \leq \frac{q}{2}$ and $\lambda = 2\bar{\mathfrak{e}} + q$ and going through analogous steps as in the previous theorem we obtain:

$$\begin{aligned}
\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty &= \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - h + \phi'(x) \right\|_\infty \\
&\leq \left\| (\hat{\mathfrak{f}}_{\mathfrak{n}}(x) - h(x) \right\|_\infty + \|\phi'(x)\|_\infty \\
&\leq (\bar{L} + L_h) \|x - \xi_n^x\|_\infty + \frac{\lambda}{2} + \bar{\nu} + \frac{q}{2} \\
&\overset{(*)}{\leq} (\bar{L} + L_h) \|x - \xi_n^x\|_\infty + 2\bar{\mathfrak{e}} + \frac{3q}{2}
\end{aligned}$$

where $\xi_n^x := \arg\inf_{s \in G_n} \|x - s\|_\infty$ denotes a nearest neighbour of x in the input sample G_n. So, convergence (pointwise or uniform) of the grid to the input space with a rate of at most $r(n)$ implies that the right-hand side of Ineq. (*) and hence, the prediction error, converges (pointwise or uniformly) to the interval $[0, 2\bar{\mathfrak{e}} + \frac{3q}{2}]$ with a rate of at most $(\bar{L} + L_h) r(n)$ as $n \to \infty$.

□

3.1.2 Sample complexity bounds and worst-case consistency for uniformly distributed inputs

Above we have given guarantees relative to the deterministic convergence rates of the input sample to the domain. In this subsection, we shall study probabilistic convergence rates as a function of the sample size in situations where the sample is obtained by drawing inputs independently from a uniform probability distribution on $I = \mathcal{X} := [0,1]^d$.

We can show that the worst-case prediction error vanishes (up to the usual worst-case bounds in the presence of observational errors) in probability:

Theorem 14 *Let $\mathcal{X} = [0,1]^d$ be the domain of target function $f \in Lip(L^*)$. Assume the input data $G_n = \{s_1, \ldots, s_n\}$ contains n data sample inputs which are drawn independently at random from a uniform distribution over $\mathcal{X}$. Furthermore, assume there are no observational errors, i.e. $\bar{\mathfrak{e}} = 0$. The worst-case error of our LACKI predictor vanishes in probability: $\forall \epsilon > 0 \forall \delta \in (0,1) \exists N \in \mathbb{N} \forall n \geq N$:*

$$\Pr[\sup_{x \in \mathcal{X}} \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty > \epsilon] \leq \delta. \tag{10}$$

In particular, for all $\delta \in (0,1)$, (10) holds for :

(1) any $\epsilon \geq 2L^*$, provided that $n \geq 1$;
(2) any $\epsilon < 2L^*$, provided that
$n \geq N := \left\lceil \frac{\log(\delta\, 2^{-kd})}{\log(1-2^{-kd})} \right\rceil$ *with* $k = \left\lceil \frac{\log(\epsilon^{-1} 2L^*)}{\log 2} \right\rceil$.

Proof: *Let $r_n := \sup_{x \in \mathcal{X}} \min_{s \in G_n} \|x - s\|_\infty \leq 1$ and let $P_n^\epsilon := \Pr[\sup_{x \in \mathcal{X}} \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty > \epsilon]$ which we intend to bound from above. From Cor. 10, remember that $\sup_x \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty \leq 2L^* r_n$. Hence, for $\epsilon \geq 2L^*$, $P_n^\epsilon \leq 0, \forall n \in \mathbb{N}$.*

So, it suffices to focus on the case where $\epsilon < 2L^*$. Now, $\sup_x \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty \leq \epsilon$ is implied by $\sup_x \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty \leq 2L^* r_n$, provided that $r_n \leq \frac{\epsilon}{2L^*}$. So, we define an event E_n that ensures r_n satisfies the latter inequality with a probability that grows as n increases. To this end, we introduce a partition of the domain into m hyper-rectangles $H_1, ..., H_m$ of equal size, each having edge length $l_k = \frac{1}{2^k}$ where k is a natural number such that $l_k \leq \frac{\epsilon}{2L^*}$. As a valid choice, we set $k := \left\lceil \frac{\log(\epsilon^{-1} 2L^*)}{\log 2} \right\rceil$. Note, $\Pr[s_i \in H_j] = l_k^d = \frac{1}{2^{dk}}$. By construction, in the event that each hyper-rectangle ends up containing at least one sample input of G_n, we have $r_n \leq \frac{\epsilon}{2L^*}$. We define the complement of this event as $\bar{E}_n := \{(s_1, ..., s_n) \in \mathcal{X}^n | \exists j \in \{1, ..., m\} \forall i \in \{1, ..., n\} : s_i \notin H_j\}$.

Let $W := \{s = (s_1, ..., s_n) | \sup_x \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty > \epsilon\}$ be the event that the sample inputs are located in such a fashion that they give rise to a worst-case error larger than ϵ. We have: $s \notin \bar{E}$ implies that $r(n) \leq \frac{\epsilon}{2L^*}$ which in turn implies $\sup_x \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x) - f(x) \right\|_\infty \leq \epsilon$, i.e. that $s \notin W$. Hence, $W \subseteq \bar{E}_n$ and thus, $P_n^\epsilon = \Pr[W] \leq \Pr[\bar{E}_n]$. So, to bound P_n^ϵ from above it suffices to bound $\Pr[\bar{E}_n]$ from above which we will do next: We can employ the union bound, utilise that $m = 2^{kd}$ and the fact that the s_i are drawn i.i.d. from a uniform to see that $\Pr[\bar{E}_n] \leq \sum_{j=1}^m \prod_{i=1}^n \Pr[s_i \notin H_j] = 2^{kd}(1 - \frac{1}{2^{dk}})^n \overset{n \to \infty}{\longrightarrow} 0$ which shows the main statement of the theorem. To find an n sufficently large to ensure $\Pr[W] \leq \delta$ we consider the inequality $2^{kd}(1 - \frac{1}{2^{dk}})^n \leq \delta$. Taking the log *on both sides and rearranging yields the sufficient condition: $n \geq \frac{\log(\delta\, 2^{-kd})}{\log(1-2^{-kd})}$.* □

3.2 Online learning guarantees

Above, we considered the worst-case asymptotics for the case where the data becomes dense in the domain with high probability. Here the error was evaluated on the entire input domain under an i.i.d. uniform input distribution. In online learning and control however, imposing such distributional assumptions is typically unrealistic. Therefore, we will now consider an online learning setting where we incrementally get to observe samples along the trajectory of inputs $\left(x_n\right)_{n \in \mathbb{N}}$ and are interested in the long-term one-step-lookahead prediction errors on this trajectory irrespective of distributional assumptions. That is, we are interested in the evolution of worst-case prediction errors, where the predictor $\hat{\mathfrak{f}}_n(\cdot)$ is based on $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{\left(x_{n-1}, \tilde{f}(x_{n-1})\right)\}, \forall n > 1$.

We will show that this error trajectory vanishes (up to observational errors), provided that the input sequence $\left(x_n\right)_{n \in \mathbb{N}}$ is bounded. In preparation of these considerations, we will establish the following facts:

Lemma 15 *Assume we are given a trajectory $\left(x_n\right)_{n \in \mathbb{N}}$ of inputs with $x_n \in \mathcal{X}$ where input space $\mathcal{X}$ can be endowed with a shift-invariant measure. Furthermore, assume the sequence is bounded, i.e. $\|x_n\|_\infty \leq \beta$ for some $\beta \in \mathbb{R}_+$ and all $n \in \mathbb{N}$. Finally assume the inputs of the available data coincide with the complete history of past inputs, i.e. $G_n = \{x_i | i \in \mathbb{N}, i < n\}$. Then we have:*

$$dist(G_n, x_n) = \min\{\|g - x_n\|_\infty \,|\, g \in G_n\} \overset{n \to \infty}{\longrightarrow} 0.$$

Proof: *The intuition behind the following proof is that if the distances were not to converge, there was an infinite number of disjoint balls around the input points that summed up to infinite volume. This however, would be a contradiction to the presupposed boundedness of the*

sequence. We formalise this intuition as follows: We can rephrase the desired convergence statement as $\forall \epsilon > 0 \exists n \in \mathbb{N} \forall m \geq n : dist(x_m, G_m) \leq \epsilon$.

For contradiction, assume that $\exists \epsilon > 0 \forall n \in \mathbb{N} \exists m(n) \geq n : dist(x_{m(n)}, G_{m(n)}) > \epsilon$. *Hold such an* $\epsilon > 0$ *fixed and choose any* $n \in \mathbb{N}$. *By definition of* $G_{m(n)} = \{x_i | i < m(n)\}$ *we have:*

$$\forall i < m(n) : \left\| x_{m(n)} - x_i \right\|_\infty > \epsilon. \tag{11}$$

Let $C_n := \bigcup_{i<n} \mathfrak{B}_{\frac{\epsilon}{2}}(x_i)$ *be the union of all* $\frac{\epsilon}{2}$*-balls around each point in* G_n *and define* $\bar{I} = \bigcup_{n \in \mathbb{N}} C_n$. *By definition, each* x_n *is contained in* $\bar{I}$. *Since sequence* $(x_n)_{n \in \mathbb{N}}$ *is bounded,* $\bar{I}$ *has a finite volume relative to some positive, shift-invariant measure* μ. *I.e.* $\mu(\bar{I}) < \infty$ *(e.g. choose the Lebesgue measure for* μ*). Furthermore,* $\mu(C_n) \leq \sum_{i<n} \mu(B_i) \leq \mu(\bar{I}) < \infty$ *where* $B_i := \mathfrak{B}_{\frac{\epsilon}{2}}(x_i)$. *Owing to the assumed shift-invariance, we can assign the same measure* M *each ball, i.e.* $M := \mu(B_1) = \mu(B_n) \forall n \in \mathbb{N}$. *Thus,* $\mu(C_n) \leq nM$. *Define* $q := \left\lceil \frac{\mu(\bar{I})}{M} \right\rceil \in \mathbb{N}$. *This is an upper bound on the number of disjoint balls of measure* M *that can be contained in* $\bar{I}$. *Intuitively, since this number is finite, there cannot be an infinite number of non-intersecting balls around the elements of the sequence* $(x_n)_{n \in \mathbb{N}}$. *More formally our argument proceeds as follows: Choose* $n > q + 1$. *Statement (11) yields:*

$$\forall i \in \{1, ..., n\} \exists p(i) \geq i \forall j \leq p(i) : \left\| x_{p(i)} - x_j \right\|_\infty > \epsilon. \tag{12}$$

Define a permutation π *such that* $\pi(p(1)) \leq \ldots \leq \pi(p(n))$. *With Statement (12) it follows that* $\mathfrak{d}_{\mathcal{X}}(x_{\pi(p(i))}, x_{\pi(p(j))}) > \epsilon$, $\forall i, j = 1, ..., n, i < j$. *Thus, we conclude the disjointness conditions* $B_{\pi(p(i))} \cap B_{\pi(p(j))} = \emptyset, \forall i, j = 1, ..., n, i \neq j$. *Hence,* $\mu(\bar{I}) \geq \mu(C_{\pi(p(n))}) \geq \mu(C_{\pi(p(1))}) + \sum_{i=1}^{n} \mu(B_{\pi(p(i))})$ $= \mu(C_{\pi(p(1))}) + nM > \mu(C_{\pi(p(1))}) + (q+1)M \geq \mu(C_{\pi(p(1))}) + \mu(\bar{I})$, *where the last inequality follows from the fact that* $Mq = M \left\lceil \frac{\mu(\bar{I})}{M} \right\rceil \geq \mu(\bar{I})$. *Since* $\mu(C_{\pi(p(1))}) \geq 0$, *we have concluded the false statement* $\mu(\bar{I}) > \mu(\bar{I})$. □

Theorem 16 *Assume that, for some* $q \geq 0$, *we chose* $\lambda = 2\bar{\mathfrak{e}} + q$ *in our LACKI prediction rule. And, let* f *be Lipschitz continuous up to some error level* $\bar{E}_h$. *That is,* $f = \phi + \psi$ *with* $\phi \in Lip(L^*)$ *and a function* ψ *such that* $\sup_x \|\psi(x)\|_\infty \leq \bar{E}_h \in \mathbb{R}$.

Assume we are given a trajectory $\left(x_n\right)_{n \in \mathbb{N}}$ *of inputs that is bounded, i.e. where* $\|x_n\|_\infty \leq \beta$ *for some* $\beta \in \mathbb{R}_+$ *and all* $n \in \mathbb{N}$. *Furthermore, assume* $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{(x_n, \tilde{f}(x_n))\}$ *and thus,* $G_n = \{x_i | i \in \mathbb{N}, i < n\}$. *Then the prediction error on the sequence vanishes up to the level of sample-consistency and Lipschitz continuity in the following sense:*

$$\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - f(x_n) \right\|_\infty \overset{n \to \infty}{\longrightarrow} [0, \frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h].$$

In particular, in case the observations are error-free ($\tilde{f} = f$*) and assuming the target is Lipschitz continuous then, when choosing* $\lambda = 0$, *the prediction error is guaranteed to vanish. That is,*

$$\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - f(x_n) \right\|_\infty \overset{n \to \infty}{\longrightarrow} 0.$$

Proof: *Let* $\xi_n \in argmin_{g \in G_n} \|x_n - g\|_\infty$ *denote the nearest neighbour of* x_n *in* $G_n = \{x_1, ..., x_{n-1}\}$. *Since sequence* (x_n) *is bounded, Lemma 15 is applicable and hence: (i)* $\lim_{n \to \infty} \|x_n - \xi_n\|_\infty = 0$. *In [4], Lemma 2.7, it was shown that* $\left\| f(s_q) - \hat{\mathfrak{f}}_{\mathfrak{n}}(s_q) \right\|_\infty \leq \frac{\lambda}{2} + \|\mathfrak{e}(s_q)\|_\infty \leq \frac{\lambda}{2} + \bar{\mathfrak{e}}$. *Therefore, if we set* $\lambda = 2\bar{\mathfrak{e}} + q$ *then* $\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(\xi_n) - f(\xi_n) \right\|_\infty \leq 2\bar{\mathfrak{e}} + \frac{q}{2}$. *Hence, appealing to the triangle inequality, we see that*

$$(ii) \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - f(\xi_n) \right\|_\infty \leq \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - \hat{\mathfrak{f}}_{\mathfrak{n}}(\xi_n) \right\|_\infty + 2\bar{\mathfrak{e}} + \tfrac{q}{2}.$$

Moreover we note that the predictors $\hat{\mathfrak{f}}_{\mathfrak{n}}$ *have Lipschitz constants* $L(n)$ *and that the* $L(n)$ *are bounded from above by some* $\bar{L} \in \mathbb{R}$. *Thus, (iii)* $\exists \bar{L} \in \mathbb{R} \forall n \in \mathbb{N} : \hat{\mathfrak{f}}_{\mathfrak{n}} \in Lip(\bar{L})$.

In conclusion,

$$\begin{aligned}
\left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - f(x_n) \right\|_\infty &\leq \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - f(\xi_n) \right\|_\infty + \|f(\xi_n) - f(x_n)\|_\infty \\
&\overset{(ii)}{\leq} \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - \hat{\mathfrak{f}}_{\mathfrak{n}}(\xi_n) \right\|_\infty + 2\bar{\mathfrak{e}} + \tfrac{q}{2} + \|f(\xi_n) - f(x_n)\|_\infty \\
&\leq \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - \hat{\mathfrak{f}}_{\mathfrak{n}}(\xi_n) \right\|_\infty + 2\bar{\mathfrak{e}} + \tfrac{q}{2} + \|\phi(\xi_n) - \phi(x_n)\|_\infty + 2\bar{E}_h \\
&\overset{(iii)}{\leq} (\bar{L} + L^*) \|x_n - \xi_n\|_\infty + 2\bar{\mathfrak{e}} + \tfrac{q}{2} + 2\bar{E}_h \overset{n \to \infty}{\longrightarrow} 2\bar{\mathfrak{e}} + \tfrac{q}{2} + 2\bar{E}_h.
\end{aligned}$$

□

4 Online Learning-Based Control

In (discrete-time) control, many classical control tasks desire to turn certain aspects of the dynamics of a plant into a contraction. For example, consider tracking control where one wishes the state $x_n \in \mathcal{X}$ at time step $n \in \mathbb{N}$ to follow a reference trajectory ξ_n. Defining the error as $e_n = \xi_n - x_n$ one might wish to define a control law that ideally would cause the error to satisfy $e_{n+1} = \phi(e_n)$ $(n \in \mathbb{N})$ for some contraction ϕ with a desirable fixed point e_* (normally $e_* = 0$), ensuring exponentially fast convergence of the error to (approximately) e_*. Unfortunately, most of these control designs, such as linearising controllers or model-predictive controllers, require an accurate model of the original dynamics. In absence of good model knowledge, we might

wish learn the dynamics model online with a learning approach. With the control inputs based on predictions of the learned model, the prediction errors enter the closed-loop dynamics as a sequence of disturbances $(d_n)_{n\in\mathbb{N}}$. Fortunately, we can translate our convergence results on the prediction errors derived above to guarantees on the closed-loop error dynamics resulting from employing the learning-based controller.

Theorem 17 *With notation as before, let $r = \frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h$ and $\|\cdot\|$ be some norm on $\mathcal{X}$. Assume the reference ξ_n is bounded and that a plant's state x_n at time $n \in \mathbb{N}$ satisfies the recurrence relation $x_{n+1} = f(x_n, u_n)$; here u_n denotes the control input applied at time n and f is an a priori uncertain, continuous function we desire to learn online utilising LACKI. We assume online learning is performed and that after each time step n, LACKI has access to (possibly erroneous with error up to $\bar{\mathfrak{e}}$) samples of the function values $f(x_i, u_i)$ for all past time steps $i < n$. That is we can compute a LACKI predictor $\hat{\mathfrak{f}}_{\mathfrak{n}}(x_n, u_n; \mathcal{D}_n)$ on the basis of a data set $\mathcal{D}_n = \{\big((x_i, u_i), \tilde{f}_i\big) | i < n\}$. Assume that a learning-based control law $u_n = u\big(x_n; \hat{\mathfrak{f}}_{\mathfrak{n}}(\cdot; \mathcal{D}_n)\big)$ is defined such that, utilising $\hat{\mathfrak{f}}_{\mathfrak{n}}$, we obtain the closed-loop error dynamics*

$$e_n = \phi(e_n) + d_n$$

where $d_n = f(x_n, u_n) - \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n, u_n)$ is the one-step prediction error and ϕ is a contraction with fixed-point $e_* \in \mathcal{X}$ and Lipschitz constant $\lambda_\phi \in [0, 1)$. Then we have:

$$\|e_n - e_*\| \overset{n\to\infty}{\to} \Big[0, \frac{r}{1-\lambda_\phi}\Big].$$

Proof: *Define the nominal reference error $\bar{e}_n$ inductively by $\bar{e}_0 = e_0, \bar{e}_{n+1} = \phi(\bar{e}_n)$, $(n \geq 0)$. Note, we can define*

$$\sigma := \sum_{i=0}^{\infty} \lambda_\phi^i = \lim_{n\to\infty} \sum_{i=0}^{n-1} \lambda_\phi^{n-1-i} = \frac{1}{1-\lambda_\phi}.$$

where we have applied a change of variables and the geometric series formula in the last step.

The assumptions of our theorem assure that the assumptions of Theorem 16 are met. Hence, $\|d_n\| \overset{n\to\infty}{\to} [0, r]$. Let $\epsilon > 0$. We desire to show:

$$\exists M \forall m \geq M : \|e_m - e_*\| \leq \epsilon + \sigma r.$$

To this end, firstly, we note that due to convergence of x_n to the fixed point x_*, we can find n_0 such that

$$\forall n \geq n_0 : \|\bar{e}_n - e_*\| < \frac{\epsilon}{3}. \tag{13}$$

Secondly, we note that, by induction, it is easy to show that for all $k, n \in \mathbb{N}$, we have

$$\|e_{k+n} - \bar{e}_{k+n}\| \leq \lambda_\phi^n \|\bar{e}_k - y_k\| + \sum_{i=0}^{n-1} \lambda_\phi^{n-1-i} \|d_{k+i}\| \tag{14}$$

$$\leq \lambda_\phi^n \|\bar{e}_k - e_k\| + \bar{\mathfrak{N}}_{k,n}\sigma \tag{15}$$

where $\bar{\mathfrak{N}}_{k,n} := \max\{\|d_k\|, \ldots, \|d_{k+n-1}\|\}$. Due to the convergence property of the disturbances, we know that there exists k_0 such that

$$\forall k \geq k_0, n \in \mathbb{N} : \bar{\mathfrak{N}}_{k,n} \leq \frac{\epsilon}{3\sigma} + r. \tag{16}$$

Now choose any $m_0 := \max\{k_0, n_0\}$. We notice that there exists $q_0 \in \mathbb{N}$ such that

$$\lambda_\phi^n \|\bar{e}_{m_0} - e_{m_0}\| < \frac{\epsilon}{3}, \forall n \geq q_0. \tag{17}$$

Now, let $m > M := m_0 + q_0$. Then we can find $n \geq q_0$ such that $m = m_0 + n$. And, for any $m = m_0 + n$, $n \geq q_0$, we have :

$$\begin{aligned}
\|e_m - e_*\| &\leq \|e_* - \bar{e}_m\| + \|e_m - \bar{e}_m\| \\
&\overset{(13)}{\leq} \frac{\epsilon}{3} + \|e_m - \bar{e}_m\| \\
&\overset{(15)}{\leq} \frac{\epsilon}{3} + \lambda_\phi^n \|\bar{e}_{m_0} - e_{m_0}\| + \bar{\mathfrak{N}}_{m_0,n}\sigma \\
&\overset{(16)}{\leq} \frac{\epsilon}{3} + \lambda_\phi^n \|\bar{e}_{m_0} - e_{m_0}\| + (\frac{\epsilon}{3\sigma} + r)\sigma \\
&\overset{(17)}{\leq} \frac{\epsilon}{3} + \frac{\epsilon}{3} + (\frac{\epsilon}{3\sigma} + r)\sigma = \epsilon + \sigma r.
\end{aligned}$$

□

Note, λ_ϕ will have to be quantified on a case-by case basis. Below, we will do so, considering the special case of tracking in model-reference adaptive control where feedback-linearisation will be employed to yield closed-loop error dynamics $e_{n+1} = \phi(e_n) + d_n$ with $\phi(e) = Me$ for some Schur matrix M. To see this is a special case, note ϕ is an eventually contracting map with fixed point 0 (cf. [12], Corollary 3.3.5) and hence, a contraction relative to some metric $\tilde{d}$ uniformly equivalent to the metric $\mathfrak{d} : (x, x') \mapsto \|x - x'\|$ [12]. Finally, we note that convergence to an interval relative to a metric entails convergence relative to all equivalent metrics (including the canonical metric derived from the maximums norm) and hence, Theorem 17 is applicable to give a bound on the Euclidean norm of the long-term tracking error. Next, we will consider this special case in greater detail.

4.1 Application to Online Learning-Based Model Reference Adaptive Control

While there are many different controllers where LACKI might be applicable, we consider *model-reference adaptive control (MRAC)* [1] of a feedback-linearisable control-affine system with unknown drift vector field as considered in [9]. This setting is a special case of the general online learning-based control framework considered above.

We begin this section by rehearsing the problem setting and assumptions made by [9], before proving that LACKI applied to MRAC in discrete-time versions of this setting can control a plant successfully to a given reference trajectory. We will conclude the section by extending experiments by other authors, who have applied machine learning methods to MRAC to a discrete-time approximation of fighter jet roll dynamics under wing rock. Our experiments demonstrate the advantages of utilising LACKI over previously proposed solutions for this benchmark scenario.

4.2 MRAC– definitions and assumptions

Assume $m \in \mathbb{N}$ to be the dimensionality of a configuration of the system in question and define $d = 2m$ to be the dimensionality of the pertaining state space $\mathcal{X}$.

Let $x = [x_1; x_2] \in \mathcal{X}$ denote the state of the plant to be controlled. Given the feedback linearisable control-affine system

$$\dot{x}_1 = x_2,\ \dot{x}_2 = a(x) + b(x)\, u(x) \tag{18}$$

it is desired to find a control law $u(x)$ such that the closed-loop dynamics exhibit a desired reference behaviour: $\dot{\xi}_1 = \xi_2, \dot{\xi}_2 = f_r(\xi, r)$ where r is a reference command, f_r some desired response and $t \mapsto \xi(t)$ is the reference trajectory.

If a priori a and b are believed to coincide with $\hat{a}_0, \hat{b}_0$ respectively, the inversion control $u = \hat{b}_0^{-1}(-\hat{a}_0 + u')$ is applied. This reduces the closed-loop dynamics to $\dot{x}_1 = x_2, \dot{x}_2 = u' + \tilde{a}(x, u)$ where $\tilde{a}(x, u)$ captures the modelling error of the dynamics:

$$\tilde{a}(x, u) = a(x) - \hat{a}_0(x) + \big(b(x) - \hat{b}_0(x)\big)u. \tag{19}$$

Let $I_d \in \mathbb{R}^{d\times d}$ denote the identity matrix. If b is perfectly known, then $b - \hat{b}_0^{-1} = 0$ and the model error can be written as $\tilde{a}(x) = a(x) - \hat{a}_0(x)$. In particular, $\tilde{a}$ has lost its dependence on the control input.

In this situation [9,8] propose to set the pseudo control as follows: $u'(x) := \nu_r + \nu_{pd} - \nu_{ad}$ where $\nu_r = f_r(\xi, r)$ is a feed-forward reference term, ν_{ad} is a yet to be defined output of a learning module *adaptive element* and $\nu_{pd} = [K_1 K_2]e$ is a feedback error term designed to decrease the *tracking error* $e(t) = \xi(t) - x(t)$ by defining $K_1, K_2 \in \mathbb{R}^{m\times m}$ as described in what is to follow.

Inserting these components, we see that the resulting *error dynamics* are:

$$\dot{e} = \dot{\xi} - [x_2; \nu_r + \nu_{pd} + \tilde{a}(x)] = Me + B\big(\nu_{ad}(x) - \tilde{a}(x)\big) \tag{20}$$

where $M = \begin{pmatrix} O_m & I_m \\ -K_1 & -K_2 \end{pmatrix}$ and $B = \begin{pmatrix} O_m \\ I_m \end{pmatrix}$. If the feedback gain matrices K_1, K_2 parametrising ν_{pd} are chosen such that M is stable then the error dynamics converge to zero as desired, provided the learning error E_λ vanishes: $E_\lambda(x(t)) = \|\nu_{ad}(x(t)) - a(x(t))\| \overset{t\to\infty}{\longrightarrow} 0$.

It is assumed that the adaptive element is the output of a learning algorithm that is tasked to learn $\tilde{a}$ online. This is done by continuously feeding it training examples of the form $\big(x(t_i), \tilde{a}(x(t_i)) + \varepsilon_i\big)$ where ε_i is observational noise.

Intuitively, assuming the learning algorithm is suitable to learn target $\tilde{a}$ (i.e. $\tilde{a}$ is close to some element in the hypothesis space [16] of the learner) and that the controller manages to keep the visited state space bounded, the learning error (as a function of time t) should vanish.

Substituting different learning algorithms yields different adaptive controllers. *RBFN-MRAC* [13] utilises radial basis function neural networks for this purpose whereas *GP-MRAC* employs Gaussian process learning [20] to learn $\tilde{a}$ [9,8].

In what is to follow, we utilise our LACKI method as the adaptive element. Following the nomenclature of the previous methods we name the resulting adaptive controller *LACKI-MRAC*.

4.2.1 Convergence Guarantees

We now provide guarantees for LACKI-MRAC controller in the discrete-time setting where LACKI is allowed to perform online learning. That is, we assume that at time step $n + 1$, the controller gets to see an additional sample of the uncertain drift at the state visited in the previous time step n. That is, the predictor $\hat{\mathfrak{f}}_{n+1}(\cdot)$ is based on $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{\big(x_n, \tilde{f}(x_n)\big)\}, \forall n$. Let $\mathcal{X}$ denote state space endowed with a norm $\|\cdot\|$. We consider a first-order Euler time-discretised version of the dynamics described in the previous subsection.

Here the error dynamics become :

$$e_{n+1} = Me_n + \Delta F(x_n) \tag{21}$$

where $\Delta \in \mathbb{R}_+$ is a positive time increment, $F_n := F(x_n) = E_\lambda(x_n) = f(x_n) - \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n)$ 1-step look-ahead prediction error of the LACKI model utilised to compute the feedback-linearising control law at time step $n \in \mathbb{N}_0$. Remember, we have shown that the prediction error vanishes up to a level that depends on the observational and representational error levels. Furthermore,

$$F_n := F(x_n) = f(x_n) - \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) = B\big(\nu_{ad}(x_n) - \tilde{a}(x_n)\big)$$

is a disturbance due to the model error of the learner, $B = \begin{pmatrix} O_m \\ \Delta I_m \end{pmatrix}$ and $M = \begin{pmatrix} I_m & \Delta I_m \\ -\Delta K_1 & I_m - \Delta K_2 \end{pmatrix}$ is the (error state) transition matrix. Here, $m = \frac{d}{2}$ is half the dimensionality of the state space, I_m denotes the $m \times m$ identity matrix and K_1, K_2 are gain matrices that can be freely chosen by the designer of the linear pseudo controller. In particular, they can be set to ensure that M is a stable matrix with spectral radius strictly less than 1, i.e $\rho(M) < 1$ and $\forall n : \|F_n\| < \bar{\mathfrak{N}}$ for some upper bound $\bar{\mathfrak{N}}$ on the disturbance. By induction, it is easy to show that for all $k \in \mathbb{N}_0, n \in \mathbb{N}$ we have: $e_{k+n} = M^n e_k + \sum_{i=0}^{n-1} M^{n-1-i} F_{i+k}$. Thus,

$$\|e_{k+n}\| \leq |||M^n||| \, \|e_k\| + \sum_{i=0}^{n-1} \big|||M^{n-1-i}\big||| \, \|F_{i+k}\| \quad (22)$$

$$\leq |||M^n||| \, \|e_{k+n}\| + \bar{\mathfrak{N}}_{k,n} \sum_{i=0}^{n-1} \big|||M^{n-1-i}\big||| \quad (23)$$

where $|||\cdot|||$ denotes the spectral norm and

$$\bar{\mathfrak{N}}_{k,n} := \max_{i=0,\ldots,n-1} \|F_{i+k}\| \leq \bar{\mathfrak{N}}.$$

Since M is stable, the terms in (23) are bounded and convergent as $n \to \infty$ (see e.g. [6]). In particular, with Gelfand's formula and the standard root test for series it is easy to establish convergence of the series: That is, there exists $\sigma \in \mathbb{R}$ with $\lim_{k\to\infty} \sum_{i=0}^{k-1} \big|||M^{k-1-i}\big||| = \sum_{i=0}^{\infty} \big|||M^i\big||| =: \sigma$.[1] And, $\sum_{i=0}^{n-1} \big|||M^{n-1-i}\big||| \leq \sigma, \forall n$. Hence,

$$\|e_{k+n}\| \leq |||M^n||| \, \|e_k\| + \sigma \, \bar{\mathfrak{N}}_{k,n}, \forall n \in \mathbb{N}, k \in \mathbb{N}_0. \quad (24)$$

Above, we have seen that any continuous function can be approximated by some Lipschitz continuous LACKI predictor up to an arbitrarily small error. For convenience, we will establish the following definition:

Definition 18 *We say that a continuous function f is L^*-Lipschitz* up to *error $\bar{E}_h \in \mathbb{R}$ on domain $\mathcal{X}$ iff there is an L^*-Lipschitz function $\phi \in Lip(L^*)$ and a function ψ such that:*
$\forall x : f(x) = \phi(x) + \psi(x)$, $\sup_{x\in\mathcal{X}} \|\psi(x)\|_\infty \leq \bar{E}_h$.

Theorem 19 (Tracking error convergence) *Assume that, for some $q \geq 0$, we choose $\lambda = 2\bar{\mathfrak{e}} + q$ in our LACKI prediction rule and that the sequence of prediction errors $\Big(F_n(x_n)\Big)_{n\in\mathbb{N}}$ as well as the reference $\Big(\xi_n\Big)_{n\in\mathbb{N}}$ are bounded. If the initial error innovation function is bounded, i.e. if $\exists b \in \mathbb{R} \forall x : \|F_0(x)\|_\infty \leq b$, and, if M is a stable matrix, i.e. if $\rho(M) < 1$, then the tracking error converges to the interval $\sigma[0, \frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h]$. That is,*

$$\|e_n\|_\infty \overset{n\to\infty}{\longrightarrow} [0, \sigma(\frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h)]$$

where $\sigma := \Delta \sum_{i=0}^{\infty} \big|||M^i\big||| < \infty$.

Proof: Let $\|\cdot\| := \|\cdot\|_\infty$ with accociated matrix norm $|||\cdot||| := \sqrt{d}\|\cdot\|_2$. Let $\epsilon > 0$. We desire to show:

$$\exists N \in \mathbb{N} \forall n \geq N : \|e_n\| \leq \epsilon + \frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h. \quad (25)$$

If sequence $\Big(F_n(x_n)\Big)_{n\in\mathbb{N}}$ is bounded then, owing to M being stable, $\Big(e_n\Big)_{n\in\mathbb{N}}$ is bounded. That is, $\exists b \in \mathbb{R} \forall n : \|e_n\| \leq \beta$. Knowing that the error dynamics are bounded by some $\beta \geq 0$ we see that $\big|||M^k\big||| \, \|e_n\| \leq \big|||M^k\big|||\beta \overset{k\to\infty}{\longrightarrow} 0$. Here, the convergence to zero follows from the assumption that M is a stable matrix. Hence, we have:

$$(I) \; \forall n \exists k_0(n) \in \mathbb{N} \forall k \geq k_0(n) : \big|||M^k\big||| \, \|e_n\| \leq \frac{\epsilon}{2}.$$

If in addition, the reference is bounded this implies that the sequence (x_n) is bounded, too. Theorem 16 implies convergence of the innovations and hence, assuming $\mathfrak{d}_{\mathcal{Y}}(f, f') = \|f - f'\|$, we have:

$$\forall \varepsilon > 0 \exists n_0 \forall n \geq n_0 : \|F_n(x_n)\| \leq \varepsilon + \frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h. \quad (26)$$

We can convert Ineq. (23) to state that for all $k \in \mathbb{N}, n \in \mathbb{N}_0$ we have:

$$\|e_{n+k}\| \leq \big|||M^k\big||| \, \|e_n\| + \Delta Q_{n:n+k} \sum_{i=0}^{k-1} \big|||M^{k-1-i}\big||| \quad (27)$$

$Q_{n:n+k} := \max\{\|F_n(x_n)\|, \ldots, \|F_{k+n-1}(x_{k+n-1})\|\}$. With Gelfand's formula and the standard root test for series it is easy to establish convergence of the series: That is, $\sigma = \lim_{k\to\infty} \Delta \sum_{i=0}^{k-1} \big|||M^{k-1-i}\big||| < \infty$. And, we have $\Delta \sum_{i=0}^{k-1} \big|||M^{k-1-i}\big||| \leq \sigma, \forall k$. Hence,

$$\|e_{n+k}\| \leq \big|||M^k\big||| \, \|e_n\| + \sigma \, Q_{k:n+k}, \forall n \in \mathbb{N}_0, k \in \mathbb{N}. \quad (28)$$

[1] In [6], a practically computable upper bound on σ can be found.

With (26) follows that there exists $n_0 \in \mathbb{N}_0$ such that we have:

$$(II)\ \forall k \in \mathbb{N} : Q_{n_0:n_0+k} \leq \frac{\epsilon}{2\sigma} + \frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h.$$

Combining (I) and (II) with Eq. 28 allows us to conclude that for any $n \geq N := n_0 + k_0(n_0)$ we have

$$\|e_n\| \leq \frac{\epsilon}{2} + \sigma\Big(\frac{\epsilon}{2\sigma} + \frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h\Big) = \epsilon + \sigma(\frac{q}{2} + 2\bar{\mathfrak{e}} + 2\bar{E}_h\Big).$$

□

Note, since the error converges to a bounded set the state will converge to the target trajectory. So, if the target trajectory is bounded, the continuity of the control law (as a function of state) implies that the control is bounded as well.

Corollary 20 *In the special case of error-free observations of a Lipschitz continuous target function, choosing a parameter $\lambda = 0$ implies that the tracking error vanishes, i.e. :*

$$\|e_n\|_\infty \overset{n\to\infty}{\longrightarrow} 0.$$

The control action sequence $\Big(u(x_n)\Big)_{n\in\mathbb{N}}$ converges, provided the reference trajectory $\Big(\xi_n\Big)_{n\in\mathbb{N}}$ converges.

Proof: The convergence statement is an immediate consequence of the preceding theorem. Remember from Sec. 4.1 that the control action at time n is of the form $u_n := u(x_n) = -\hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - Ke_n + c$ for some constant c. We show that (u_n) is a Cauchy sequence, provided that the reference sequence ξ_n is. Since $\mathcal{X}$ is a Hilbert space, the desired convergence result follows.

So, let $\epsilon > 0$. Since $(e_n), (\xi_n)$ converge, also the state sequence (x_n) converges. Hence, all three are convergent Cauchy sequences. In particular, there is an N such that for all $n, m > N$: $\|e_n - e_m\| < \frac{\epsilon}{2|||K|||}$ and $\|x_n - x_m\| < \frac{\epsilon}{2\bar{L}}$. Hence, utilising the definition of the control law and the fact that all predictors are Lipschitz continuous with Lipschitz constant $\bar{L}$, for all $m, n > N$: $\|u_n - u_m\| \leq |||K||| \, \|e_n - e_m\| + \Big\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x_n) - \hat{\mathfrak{f}}_m(x_m) \Big\| \leq \frac{\epsilon}{2} + \bar{L} \, \|x_n - x_m\| \leq \epsilon$. Therefore, (u_n) is a Cauchy sequence and hence, convergent. □

Next, we will illustrate the performance of LACKI-MRAC in a simulated application scenario that fits the theory developed up to this point.

4.2.2 Learning-based tracking control of an F-4 fighter jet under wing rock

As pointed out in [10], modern fighter aircraft designs are susceptible to lightly damped oscillations in roll known as "wing rock". Commonly occurring during landing [21], removing wing rock from the dynamics is crucial for precision control of such aircraft. Precision tracking control in the presence of wing rock is a nonlinear problem of practical importance and has served as a test bed for a number nonlinear adaptive control methods [9,17,10].

For comparison, we replicated the experiments of Chowdhary et. al. [9,8]. [2] Using a realistic model of the roll dynamics of an F-4 fighter jet, the authors examined the task of using a model-reference adaptive controller (MRAC) to perform a roll manoeuvre under uncertain wing rock. Within a time span between t_0 and t_f, the task was to control the aircraft's ailerons in order to cause the aircraft's state trajectory $x : [t_0, t_f] \to \mathbb{R}^2$ to closely follow a roll manoeuvre prescribed by the reference trajectory $\xi(\cdot)$. Here the first component of the state and reference was the roll angle and the second was the angular velocity.

Since wing rock can destabilise the dynamics, the authors proposed to utilise a (budgeted) Gaussian process approach to learn a model of the wing rock dynamics online and demonstrated this could significantly improve tracking performance over competing methods. They compared their Gaussian process based approach, called *GP-MRAC*, to the more established adaptive model-reference control approach based on RBF neural networks [22,13], referred to as *RBFN-MRAC*. As the controller was meant to adapt to the uncertain wing rock dynamics online during runtime, computational real time constraints necessitated to fix the kernel hyperparameters of the GP. Furthermore, they also proposed to limit the GP to a fixed budget of training examples which would be incrementally updated online.

Replacing the GP by our LACKI learner, we readily obtain an analogous learning-based controller which we call *LACKI-MRAC*. For baseline comparison, we also examined the performance of a simple PD-controller.

We created 700 randomised test runs of the wing rock tracking problems and tested each control algorithm on each one of them. The initial state $x(t_0)$ was drawn uniformly at random from $[0, 7] \times [0, 7]$, the initial kernel length scales were drawn uniformly at random from $[0.05, 2]$, and used both for RBF-MRAC and GP-MRAC. For LACKI, we chose $\lambda = 0$ and $L(0) = 0$. The parameter weights W of the system dynamics (cf. [9]) were multiplied by a constant drawn uniformly at random from the interval $[0, 2]$. To allow for better predictive performance of GP-MRAC, we set the maximal budget to 100 training examples (as in the experiments of [9]) as well as GP2-MRAC using a budgeted GP with up to 1000 training examples. The feedback gains of the linear pseudo

[2] We are grateful to the authors for kindly providing their code.

controller were chosen to be $K_1 = K_2 = 1$ (see [9] for more explanations). As a baseline comparison, we also tested the performance of a simple $PD-$ controller with just these feedback gains.

The performance of all controllers across these randomised trials is depicted in Fig. 2. Each data point of each boxplot represent a performance measurement for one particular trial.

For each method, the figures show the boxplots of the following recorded quantities:

- *log-XERR*: cummulative angular position error (log-deg), i.e. $\log(\int_{t_0}^{t_f} \|\xi_1(t) - x_1(t)\| \, dt)$.
- *log-XDOTERR*: cummulative roll rate error (log-deg/sec.), i.e. $\log(\int_{t_0}^{t_f} \|\xi_2(t) - x_2(t)\| \, dt)$.
- *log-PREDERR*: log-prediction error, i.e. $\log(\int_{t_0}^{t_f} \left\| \hat{\mathfrak{f}}_{\mathfrak{n}}(x(t)) - f(x(t)) \right\| \, dt)$ where f is a vector field affected by the wing rock.
- *log-CMD*: cumulative control magnitude (log-scale), i.e. $\log(\int_{t_0}^{t_f} \|u(t)\| \, dt)$.
- *log-max. RT (predictions)*: the log of the maximal run time (within time span $[t_0, t_f]$) each method took to generate a prediction ν_{ad} within the time span.
- *log-max. RT (learning)*: the log of the maximal run time (within time span $[t_0, t_f]$) it took each method to incorporate a new training example of the drift $\tilde{a}$.

Discussion: All adaptive methods outperformed the simple $PD-$ controller in terms of tracking error. With regard to prediction run time, RBFN-MRAC outperformed all nonparametric learning-based controllers GP-MRAC, GP2-MRAC and LACKI-MRAC. This is hardly surprising. After all, RBFN-MRAC is a parametric method with constant prediction time. By contrast, non-parametric methods have prediction times growing with the number of training examples. That is, it would be the case if GP-MRAC were given an infinite training size budget. Indeed one might argue whether GP-MRAC, if operated with a finite budget, actually is a parametric approximation where the parameter consists of the hyper-parameters along with the fixed-size training data matrix. When comparing the (maximum) prediction and learning run times one should also bear in mind that GP-MRAC predicted with up to 100 and GP2-MRAC with 1000 examples in the training data set. By contrast, fast enough to process large online data, LACKI-MRAC undiscerningly had incorporated all 10001 training points by the end of each trial. In spite of having significantly more training data incorporated, LACKI's prediction times were competitive with the budgeted Gaussian processes. Across the remaining metrics, LACKI-MRAC markedly outperformed all other methods. Note, we also attempted comparisons to a non-budgeted Gaussian process learning-based controller, as well as to one utilising GPs with hyperparameter optimisation. However, the resulting learning-based approach ran into conditioning problems (without extensive manual tweaking of initial conditions for each problem instance) and performed poorly both in terms of runtime and predictive performance. In fact, we would argue that one of the advantages of LACKI is its numerical simplicity and independence from a priori hyperparameter choices. The simplicity facilitates potential embeddability where the controller needs to run on RISC micro-controllers. Furthermore, LACKI robustness and performance without any manual fine tuning seems to afford it with greater black-box learning capabilities even under computational real-time constraints.

5 Conclusions

We have introduced *Lazily Adapted Constant Kinky Inference (LACKI)* as an approach to nonparametric machine learning. Our method was built on the framework of *Kinky Inference* which is a generalisation of well-known approaches such as LI and NSM methods that have become popular in numerical mathematics and learning-based control. Our approach inherits the numerical simplicity of these methods but does not require a priori knowledge of a Lipschitz constant of the underlying target function. Of course, this is of great practical interest since it endows LACKI with substantially improved black-box learning capabilities. In contrast to competing approaches based on Lipschitz constant estimation [15,5,20], LACKI is fast enough to support online learning and we can still give theoretical guarantees on the learning performance showing that LACKI can learn any continuous function. Being a nonparametric regression method that is simple but can learn rich function classes, LACKI hits a sweet spot between efficiency on the one hand and high learning capacity on the other. Furthermore, it is fast enough to be utilised in an online learning setting. This is in contrast to other methods, for instance in Gaussian process regression, that rely on hyper-parameter optimisation but which involve large computational overhead to work well. In turn, this allows LACKI to be utilised in model-reference adaptive control where we can convert our learning guarantees into guarantees on tracking success.

Our theoretical guarantees assume the observational errors to be bounded. Knowledge of such a bound is a common assumption in learning-based control [7,2], albeit not always a realistic one in practice. And, other common assumptions, such as white-noise disturbances, are physically unrealistic. Nonetheless, ongoing work investigates probabilistic consistency proofs in the presence of stochastic, potentially unbounded noise. For first results where the Lipschitz hyperparameter is directly tuned via empirical risk minimisation see [5] .

The illustrations of our control applications have focussed on model reference adaptive control. In recent

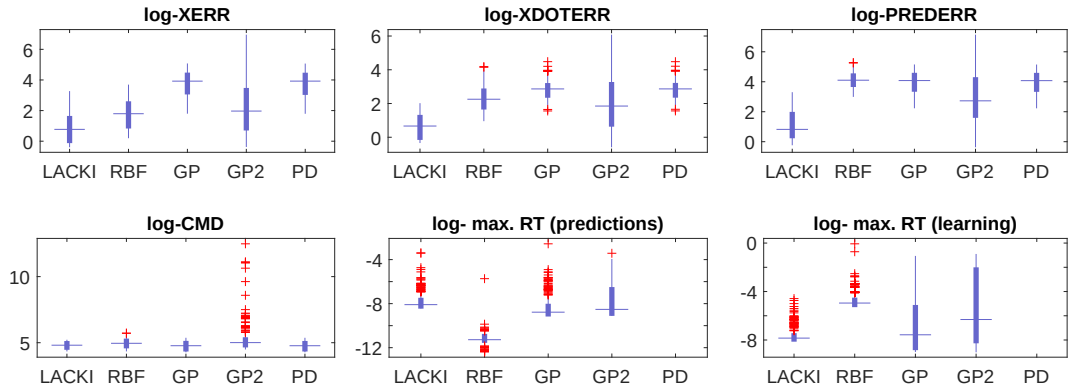

Fig. 2. Performance of the different online controllers over a range of 700 trials with randomised parameter settings and initial conditions. LACKI-MRAC outperformed all other methods with respect to all performance measures, except for prediction run time (where the parametric learner RBFN-MRAC performed best).

work, has considered applications to data-based model-predictive control [14] restricted to an offline learning setting. However, combining the stability results provided therein with our online guarantees derived in this paper could provide novel online learning-based MPC guarantees in the increasing data limit.

References

[1] K. J. Aström and B. Wittenmark. *Adaptive Control*. Addison-Wesley, 2nd edition, 2013.

[2] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin. Provably safe and robust learning-based model predictive control. *Automatica*, 2013.

[3] G. Beliakov. Interpolation of Lipschitz functions. *Journal of Computational and Applied Mathematics*, 2006.

[4] J. Calliess. Lazily Adapted Constant Kinky Inference for Nonparametric Regression and Model-Reference Adaptive Control. *Arxiv preprint arXiv:1701.00178*, 2016.

[5] J. Calliess. Lipschitz Optimisation for Lipschitz Interpolation. In *American Control Conference*, 2017.

[6] Jan-Peter Calliess. *Conservative decision-making and inference in uncertain dynamical systems*. PhD thesis, University of Oxford, 2014.

[7] M. Canale, L. Fagiano, and M. C. Signorile. Nonlinear model predictive control from data: a set membership approach. *Int. J. Robust Nonlinear Control*, 2014.

[8] G. Cho, G. Chowdhary, A. Kingravi, J. P. . How, and A. Vela. A Bayesian nonparametric approach to adaptive control using Gaussian processes. In *CDC*, 2013.

[9] Girish Chowdhary, H.A. Kingravi, J.P. How, and P.A. Vela. Bayesian nonparametric adaptive control using Gaussian processes. Technical report, MIT, 2013.

[10] Girish Chowdhary, Hassan A. Kingravi, Jonathan How, and Patricio A. Vela. Nonparametric adaptive control of time-varying systems using Gaussian processes. In *American Control Conference (ACC)*, 2013.

[11] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 1979.

[12] B. Hasselblatt and A. Katok. *A First Course in Dynamics with a Panorama of Recent Developments*. Cambridge University Press, 2003.

[13] Y. H. Kim and F. Lewis. High-level feedback control with neural networks. *Robotics and Intelligent Systems*, 1998.

[14] J.M. Manzano, D. Limon, D. Munoz de la Pena, and J. Calliess. Output Feedback MPC based on Smoothed Projected Kinky Inference. *Control Theory and Applications*, To appear in 2019.

[15] M. Milanese and C. Novara. Set membership identification of nonlinear systems. *Automatica*, 2004.

[16] T. Mitchell. *Machine Learning*. Mc Graw Hill, 1997.

[17] M.M. Monahemi and M. Krstic. Control of wingrock motion using adaptive feedback linearization. *J. of. Guidance Control and Dynamics.*, 1996.

[18] E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications.*, 1964.

[19] J. Park and J. W. Sandberg. Universal approximation using radial-basis function networks. *Neural Computation*, 1991.

[20] C.E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[21] A. A. Saad. *Simulation and Analysis of wing rock physics for a generic fighter model with three degrees of freedom*. PhD thesis, Air Force Institute of Technology, Air University, 2000.

[22] R. Sanner and J.-J. Slotine. Gaussian networks for direct adaptive control. *Trans. on Neural Networks*, 1992.

[23] R. G. Strongin. On the convergence of an algorithm for finding a global extremum. *Engineering in Cybernetics*, 1973.

[24] A.G. Sukharev. Optimal method of constructing best uniform approximation for functions of a certain class. *Comput. Math. and Math. Phys.*, 1978.

[25] G. S. Watson. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics*, 1964.

[26] Z. B. Zabinsky, R. L. Smith, and B. P. Kristinsdottir. Optimal estimation of univariate black-box Lipschitz functions with upper and lower bounds. *Computers and Operations Research*, 2003.