On the Theory of Lipschitz Continuous Machine Learning

Julien Walden Huang

St John's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

This thesis is dedicated to my family:
my parents, Daphne and Mathieu.

Acknowledgements

Firstly, I would like to express my gratitude to my first supervisor, Professor Jan-Peter Calliess, with whom I worked closely throughout my DPhil journey. Our meetings and brain-storming sessions were always both incredibly intellectually intense and fun. I enjoyed them tremendously. Your mentorship was instrumental in shaping my research and growth over the past four years.
Secondly, I would like to thank my second supervisor, Professor Stephen Roberts, for your crucial contributions in providing a broader perspective to my work and guiding my research to its best possible outcome. Your guidance and support throughout the DPhil were invaluable.

I also extend my gratitude to my viva committee members, Professor Kostas Margellos and Professor Daniel Limon for examining my thesis. Your constructive feedback and insightful comments have significantly enhanced the overall quality of this thesis.

I am incredibly grateful to the Oxford-Man Institute of Quantitative Finance for their funding and support during the DPhil.

I was fortunate to have been able to work alongside amazing friends at the OMI. In particular, I would like to thank my cubicle neighbour, Daniel, for his friendship as we travelled our DPhil journeys together. Outside of the OMI, many thanks to my friends in Oxford, especially the members of the Ping Pong lunch group and Romain, for all the laughs and great moments.

To Isabelle, thank you for your unwavering love, support and kindness during this DPhil. Your presence and belief in me made every step of this academic adventure more meaningful.

Last but certainly not least, I want to thank my family. To my sister Daphne and brother Mathieu, for always being present and for the fun times we shared during the COVID-19 pandemic. To my Grandfather, for your love and for encouraging me to start the DPhil. And finally, to my parents for their unwavering support and love throughout my life, which have been the foundation of my success. Dad, you were the main source of inspiration for this whole academic journey and Mom, your constant encouragement and belief in me made this adventure possible. Thank you from the bottom of my heart for everything.

Abstract

The field of machine learning theory plays an essential role in establishing the mathematical foundations and performance boundaries of data-driven modelling techniques. By providing a rigorous analysis of the underlying properties of an algorithm, theoretical machine learning guides the development of reliable methods that can be utilised in real-world applications. In this context, Lipschitz regularity has been a particularly useful tool, aiding in establishing robustness, worst-case error bounds, and generalisation capabilities for a wide range of machine learning frameworks. Building on this foundation, this thesis explores the theoretical properties of the general class of Lipschitz continuous machine learning frameworks with a specific focus on dynamical system identification.

The first part of this thesis investigates a fundamental problem of this class of machine learning frameworks which is the estimation of the Lipschitz constant of the target function from data. We derive optimal sample complexity rates for this problem in both the noiseless and the noisy settings under minimal parametric assumptions on the target function. A novel Lipschitz constant estimation technique shown to be computationally efficient and sample optimal is also proposed.

The second part of the thesis focuses on a popular non-parametric system identification method utilised in control: Lipschitz interpolation. It derives a series of theoretical results on the asymptotic properties of the framework under a bounded noise assumption. More specifically, general asymptotic consistency and precise upper bounds on the uniform non-parametric convergence rates are obtained. These established bounds can serve as theoretical tools for comparing Lipschitz interpolation against alternative non-parametric regression methods. Various extensions of these results in the context of online learning, online leaning-based control, and a fully data-driven extension of the classical Lipschitz interpolation framework proposed by Calliess et al. [2020] are also obtained.

The final part of the thesis considers the use of Lipschitz regularity properties in conjunction with neural network-based identification methods in the field of time series analysis. We utilise relaxed Lipschitz-type regularity assumptions on the dynamics of a general class of non-linear autoregressive processes to obtain a characterisation of mean reversion through theoretical results on geometric ergodicity and tight upper bounds on the first hitting times of these processes as they revert back to mean. The utility of these results is demonstrated in a financial application on improving trading decision rules where the theoretical results are harnessed to develop learning-based pairs trading strategies with probabilistic guarantees on their profitability.

Statement of Originality

I hereby declare that, except where made explicitly clear, the contents of this dissertation are my own original work, and to the best of my knowledge do not contain materials submitted in whole, or in part, for consideration for any other degree or qualification at the University of Oxford or any other academic or professional institution.

Signature *Date*

Table of Contents

List of Figures

List of Tables

1 | Introduction

Contents

1.1 Overview

Over the past two decades, the increasing availability of vast amounts of data and rapid advancements in computational capacity have led to the proliferation of machine learning frameworks in real-world applications. In many industries such as finance (Dixon et al. [2020]), robotics (Brunke et al. [2022]), or the automotive industry (Shukla et al. [2019]), this has meant the popularisation of data-driven approaches to identifying and controlling dynamical systems based on input-output observations.

While system identification methods are not novel, historical efforts in this field have focused on relatively simplistic parametric models (Ljung [2010]) due to computational limitations. The increase in computational power has therefore allowed for the use of more powerful non-linear parametric and non-parametric learning-based frameworks which relax the strict structural assumptions inherent in classical approaches. Popular examples include neural networks (Goodfellow et al. [2016]),

kernel methods (Thomas Hofmann [2008]) and Gaussian Processes (Williams and Rasmussen [2006]).

A desideratum of these modern system identification methods in industrial applications is that they provide mechanisms for uncertainty quantification or worst-case guarantees. This ensures that a principled approach to validating safety constraints and certifying satisfactory performance is possible. In the context of learning-based control, two prevailing paradigms have typically been associated with these methods in order to achieve this goal, as outlined in recent surveys on safe learning in predictive control by Hewing et al. [2020], Mesbah et al. [2022]. The first is probabilistic in nature and leverages a Bayesian methodology that generally relies on Gaussian process regression, see for instance Berkenkamp and Schoellig [2015], Hewing et al. [2019]. The second adopts a deterministic approach based on worst-case guarantees that typically relies on the Lipschitz regularity of estimated systems. This can be observed in works such as Milanese and Novara [2004], Calliess et al. [2020] which use Lipschitz interpolation methods or Knuth et al. [2021], Zhou et al. [2022] which develop neural network-based approaches. Machine learning frameworks that are consistent with either of these paradigms are therefore of particular interest for real-world applications of system identification.

In general, the empirical performance of the machine learning-based methods discussed above has been extensively studied. In contrast, due to the absence of underlying structural assumptions, a general theoretical characterisation of these methods has been more challenging. Given the growing use of these data-driven frameworks in practice, obtaining a profound understanding of these methods is becoming increasingly relevant and a growing number of streams of research have started to focus on this issue. These efforts include deriving asymptotic minimax rates of convergence for deep neural networks with both Sigmoid and ReLU activation functions (Bauer and Kohler [2019], Schmidt-Hieber [2020]), non-asymptotic high probability bounds for neural networks (Farrell et al. [2021]) or establishing information and convergence rates for Gaussian process methods (van der Vaart and van Zanten [2008], Van Der Vaart and Van Zanten [2011]) and various extensions (Burt et al.

[2019], Wynne et al. [2021]) to mention but a few. While significant progress has been made, closing the gap between the theoretical and empirical understanding of modern machine learning frameworks remains one of the main current challenges within the field.

This thesis seeks to contribute to the effort to narrow this gap. As the scope of the problem is extensive, our focus centres on the class of machine learning models that offer Lipschitz regularity properties with the aim of deriving meaningful theoretical results in the context of system identification. More specifically, the two following over-arching objectives are considered:

- **A comprehensive theoretical investigation of the non-parametric regression framework known as Lipschitz interpolation.**
- **The development of practical theoretical tools that leverage Lipschitz regularity properties for modern system identification.**

1.2 Research Objectives and Contributions

The first part of the thesis will focus on a relatively simple but fundamental question, which is one of the main practical drawbacks of Lipschitz interpolation frameworks and Lipschitz constant-based computational methods in general. That is the dependence on prior knowledge of the Lipschitz constant or, in the case that this assumption is not made, the necessity of learning a precise Lipschitz constant estimate. To solve this issue, a number of Lipschitz constant estimation methods (also called Lipschitz learning algorithms) have been constructed (see in particular; Strongin [1973], Beliakov [2005] and Calliess et al. [2020]) and applied in practice. Unfortunately, few of these methods provide guarantees of convergence and there is little theoretical insight into the problem itself. As the precise estimation of Lipschitz constants is crucial for the applications of any Lipschitz constant-based computational method, two reasonable questions that can be asked are:

- *"How difficult is the estimation of a Lipschitz constant?".*

and in light of the answer to this first question;

- *"Do existing estimation methods reasonably solve this problem?"*.

The first chapter in this part of the thesis will answer the first question by providing lower bounds on the sample complexity and convergence rate bounds on the problem and will answer the second question by constructing a novel theoretically-backed algorithm that addresses the shortcomings of existing Lipschitz constant estimation methods. While the initial motivation for this research was an application to the context of system identification and more specifically Lipschitz interpolation frameworks which are popular in predictive control, the research done in this part holds in a more general context. In particular, computational applications in adaptive global optimisation (Malherbe and Vayatis [2017]) or reinforcement learning (Chakrabarty et al. [2019]) have explicitly made use of Lipschitz constant estimation methods and will therefore also be interested in the findings of this part.

The second part of the thesis will continue the proposed theoretical investigation into Lipschitz interpolation frameworks by focusing on a minimax convergence rate analysis of a general version of the method. We aim to provide a satisfying answer to the following question:

- *"Can a precise theoretical characterisation of the convergence obtained by Lipschitz interpolation frameworks be established?"*.

As noted above, this type of analysis already exists for the other popular non-parametric regression methods, see in particular Steinwart et al. [2009] for convergence rates for kernel methods, Van Der Vaart and Van Zanten [2011] for GPs or Schmidt-Hieber [2020] for deep neural networks, and we take special care in formulating our results such that they can serve as a theoretical tool for comparing Lipschitz interpolation to these alternative methods. Specifically, our proposed solution provides an explicit condition on the observed tail behaviour of the noise, indicating when Lipschitz interpolation should be expected to outperform or underperform other non-parametric system identification approaches asymptotically.

Finally, the last part of the thesis will focus on the separate topic of leveraging non-

linear machine learning and Lipschitz continuity to characterise mean reversion[1] which is a fundamental topic in econometrics and finance. Most existing approaches that study mean reversion utilise classical econometric models that rely on simple parametric forms such as AR(p) or STAR(p) to model the system dynamics (see Taylor et al. [2001] and discussion therein) and do not consider the recent proliferation of non-linear machine learning-based time series models. As these methods have been shown to perform better than classical models in various forecasting tasks (Christensen et al. [2022], Hsu et al. [2016]), one can wonder whether they could not also be leveraged in the context of mean reversion. The main research question of this part can therefore be stated as:

- *"How can machine learning-based system identification methods and Lipschitz-type regularity conditions be utilised to provide a more precise estimation of mean reversion in non-linear systems?".*

This part of the thesis will provide an answer to this question by developing theoretical results on geometric ergodicity, stationarity and first hitting times based on Lipschitz-type conditions on the dynamics of the time series and by explaining how these results can be applied in practice to neural time series models. It will conclude by providing a case study of the proposed approach in the context of finance where optimal trading rules for statistical arbitrage strategies will be derived. As a side note, we remark that the Lipschitz constant estimator developed in the first part of the thesis could be applied in order to verify the Lipschitz-type conditions used to establish mean reversion in this part.

1.3 Outline

The thesis will be structured as follows:

Chapter 2 [Background] will define the general theoretical setting considered in this chapter and introduce the relevant machine learning frameworks.

[1] We note that the key assumption utilised in this chapter is based on Lipschitz continuity.

Chapter 3 [Lipschitz Constant Estimation] will study the problem of learning a Lipschitz constant of a target function from data. Theoretical lower and upper bounds are derived for the sample complexity of the problem and a novel algorithm based on local linear regression is proposed.

This chapter is based on the following paper:

- J. Huang, S. Roberts, J. Calliess, On the Sample Complexity of Lipschitz Learning, Accepted in *Transactions of Machine Learning Research*[2], 2023

Chapter 4 [Lipschitz Interpolation: Asymptotics] will provide an asymptotic analysis of Lipschitz interpolation focused on deriving consistency and asymptotic convergence rate results. Various theoretical extensions of these results to online learning, online learning-based control and a fully data-driven Lipschitz interpolation extension proposed by Calliess et al. [2020] will also be derived.

This chapter is based on the following paper:

- J. Huang, S. Roberts, J. Calliess, Asymptotic Analysis of Lipschitz Interpolation under Bounded Stochastic Noise, *(Under review)*, 2023.

Chapter 5 [Non-linear Mean Reversion] will provide a non-linear characterisation of mean reversion utilising modern machine learning methods and Lipschitz continuity properties and will explore a financial application on the development of trading decision rules in statistical arbitrage strategies.

This chapter is based on the following papers:

- J. Huang, J. Calliess, First Hitting Time Guarantees for Non-linear Time Series Models, *Time Series Workshop, International Conference on Machine Learning (ICML)*, 2021.
- J. Huang, S. Roberts, J. Calliess, First Hitting Time Guarantees for Contractive Non-linear Systems, *2023 IEEE 62nd Conference on Decision and Control (CDC)*, 2023.

[2]Featured Certification was awarded for very high quality paper (∼2.5% of papers).

- J. Huang, S. Roberts, J. Calliess, Non-linear Mean Reversion: A Machine Learning Perspective, (*submitted*), 2023.

Chapter 6 [Conclusion] will provide the concluding remarks of this thesis. A summary of the contributions made and a discussion on future work will be given.

2 | Background

Contents

This introductory background chapter provides an overview of the theoretical and practical concepts central to this thesis. We begin by describing the general regression setting, typically considered in non-parametric regression estimation, which will be utilised to derive the main theoretical results of this thesis. We discuss various aspects and assumptions of relevance to this setting, how they tie into the existing literature, and how they will be considered in subsequent chapters. We conclude the first part of Chapter 2 by formally defining the space of Lipschitz continuous functions with respect to metrics on $\mathbb{R}^d$, highlighting why the Lipschitz continuity property has often been considered in the design of computational methods.

The second part of this chapter will then introduce the class of machine learning frameworks that offer Lipschitz regularity properties considered in this thesis, that is

Lipschitz interpolation and neural networks. For both methods, we provide context for their relation to Lipschitz regularity-based safe learning and discuss several recent developments. For Lipschitz interpolation, we also examine some important existing theoretical properties that have motivated its use.

2.1 Theoretical setting:

2.1.1 General Regression Setting & Non-parametric Estimation

Non-parametric regression estimation techniques[1] constitute an essential subset of machine learning algorithms, allowing flexible and versatile modelling capable of capturing complex dynamics between variables. Theoretically, this flexibility implies that the non-parametric regression problem can consider a general target function space subject to regularity conditions, avoiding the more restrictive functional form-type assumptions[2] made in parametric regression. Note that certain parametric regression models can also be considered in this theoretical setting; e.g. neural networks can be viewed as non-parametric estimators if the number of layers and/or neurons is not fixed (Bauer and Kohler [2019], Schmidt-Hieber [2020]).

More formally, we consider the following general regression setting.

Let $\mathcal{X} \subset \mathbb{R}^d$ for some $d \in \mathbb{N}$ and $\mathcal{Y} \subset \mathbb{R}$. The goal of non-parametric regression is to learn a target function f that is assumed to belong to a regularity class of functions $\mathcal{C}$. In order to do so, a set of (possibly noisy) observations $D := (G^{\mathcal{X}}, G^{\mathcal{Y}})$ where $G^{\mathcal{X}} := \{x_i\}_{i=1,\dots,N} \subset \mathcal{X}$ represents a set of sample inputs that can be either deterministically or randomly queried and $G^{\mathcal{Y}} := \{\tilde{y}_i\}_{i=1,\dots,N} \subset \mathcal{Y}$ denotes a set of (noise-corrupted) values of the target function f associated with the inputs in $G^{\mathcal{X}}$.

[1] Classical examples of such frameworks include Gaussian processes (Williams and Rasmussen [2006]), various Kernel methods (Thomas Hofmann [2008]), certain classes of Neural networks (Goodfellow et al. [2016]) and Lipschitz interpolation frameworks (Milanese and Novara [2004], Beliakov [2006]).

[2] More, precisely, parametric regression will restrict the target function to a pre-defined paramerized family of functions.

In this thesis, we will generally assume that elements of $G^{\mathcal{Y}}$ are of the form

$$\tilde{y}_k = f(x_k) + e_k \tag{2.1}$$

where $\{e_i\}_{i=1,\dots,N}$ is a collection of random variables denoting the additive observational noise. As noted above, this type of setting is standard in the field of non-parametric estimation (Györfi et al. [2002], Tsybakov [2004]), and additional assumptions on $\{e_i\}_{i=1,\dots,N}$ are generally needed to ensure that the non-parametric estimation error converges to 0 as the number of observations increases. In order to explicitly state the dependence on the data, we denote $(D_n)_{n\in\mathbb{N}}$ as a stream of observations such that $D_n \subset D_{n+1}$ and consider the sequence of predictors $(\hat{f}_n)_{n\in\mathbb{N}}$ generated by a non-parametric estimation framework with $(D_n)_{n\in\mathbb{N}}$.

This thesis can essentially be summarised as a theoretical analysis of various problems related to the non-parametric estimation of the dynamics given in equation (2.1) for different choices of settings and assumptions. While certain assumptions have become standard and are commonly used, these are by no means fixed, and there is a large amount of flexibility in setting them. In particular, assumptions on the behaviour of the noise, the sampling of the input space, the regularity class $\mathcal{C}$ of the target function, and the characterisation of the properties of the chosen non-parametric estimation framework will need to be specified in order to derive the desired theoretical results. As some of these assumptions will vary throughout the chapters of this thesis and can be relatively technical, we provide a brief high-level discussion that aims to give an intuition on their selection.

Distributional Assumptions on the Noise

Assumptions on the distributional properties of the noise variables $\{e_i\}_{i=1,\dots,N}$ will vary significantly across this thesis depending on the aims and technical requirements of each part. This is in line with existing literature and stems from the fact that small changes in these noise assumptions can be shown to have a substantial influence on the resulting theoretical findings, e.g. see recent work by Han and Wellner [2019] on the impact of heavy-tailed distributions on non-parametric conver-

gence rates. In general, our approach will be to establish an initial set of theoretical results under general noise conditions before incorporating additional distributional assumptions in order to refine these outcomes. In certain chapters (see Chapter 3 and Chapter 5), the choice of assumptions will be straightforward and will depend on either standard moment-based assumptions on $\{e_i\}_{i=1,\dots,N}$ or on more restrictive assumptions on the probability function, i.e. sub-Gaussianity in Chapter 3 and approximate knowledge of the cumulative density function in Chapter 5. These types of assumptions are well-established and can be found in various research articles (e.g. Stone [1982]) and academic textbooks (e.g. Györfi et al. [2002], Ibragimov and Has' Minskii [2013]). In contrast, other chapters (Chapter 4) will necessitate the use of more technical and less well-known assumptions on the noise. Specifically, precise distributional assumptions on the bounded tails of the noise variables $\{e_i\}_{i=1,\dots,N}$ near the endpoints[3] of their support are made in order to quantify the effect of the behaviour of $\{e_i\}_{i=1,\dots,N}$ on the asymptotic performance of Lipschitz interpolation frameworks. While these assumptions are non-standard, we note they have become popular in sub-fields such as non-parametric boundary regression (Hall and Van Keilegom [2009] and ensuing articles) and that, under certain conditions[4], non-parametric estimation convergence rates based on these assumptions (Müller and Wefelmeyer [2010], Meister and Reiß [2013]) can be shown to improve on the classical optimal convergence rates derived under Gaussian-type noise assumptions by (Stone [1982]).

Sampling Assumptions on the Data

This assumption pertains to the description of the sets of sample inputs: $G^{\mathcal{X}}$. Classical research on the convergence rate and general theoretical properties of non-parametric estimation methods has generally considered input samples that have been "well sampled" on $\mathcal{X}$ either through an independently and identically distributed sampling distribution or through a deterministic and regular sampling procedure (Stone [1982], Tsybakov [2004]). However, in many applications, particularly

[3] The noise variables will be assumed to be bounded in this chapter.

[4] We will show in Chapter 4 that these improved convergence results hold for Lipschitz interpolation methods.

in control or in finance where the target function f models the dynamics of a stochastic system, these assumptions no longer hold. We will therefore aim, when possible, to utilise the assumption that f models the dynamics of a semi-autoregressive stochastic system;

$$\tilde{y}_n = f(x_n) + e_n \tag{2.2}$$

where $x_n = (\tilde{y}_{n-d_y}, ..., \tilde{y}_{n-1}, u_{n-d_u}, ..., u_n)$ with $\tilde{y}_i \in \mathcal{Y} \subset \mathbb{R}^l$ denoting the past autoregressive inputs and $u_i \in \mathcal{U} \subset \mathbb{R}^s$ denoting a vector of past or current control inputs for $d_y, d_u, s, l \in \mathbb{N}$. The more practical system assumption provided by equation (2.2) will allow us to both extend certain theoretical convergence results obtained for Lipschitz interpolation in the standard sampling setting (Chapter 4) and to derive theoretical properties for non-linear systems under a more realistic sampling setting (Chapter 5). Finally, we note that considering dependency assumptions on the regressors is not novel and that a significant amount of research has been done in this type of setting, see for example the work done to derive uniform convergence rates for various non-parametric estimation frameworks which Chapter 4 will extend to Lipschitz interpolation: (Hansen [2008]): kernel methods, (Chen and Christensen [2015]): splines and wavelets series regression, (Masry [1996]): local polynomial regression. As noted above, Chapters 4 and 5 will consider settings of this type in order to derive their respective results.

Regularity Assumptions on the Target Function

In the majority of existing research on non-parametric regression, the characterisation of the target function's regularity class $\mathcal{C}$ has predominantly been considered to be either a Hölder space (Stone [1982]) or a Sobolev space (Nemirovskij et al. [1985]) of functions defined over $\mathcal{X}$ (see Tsybakov [2004] for an overview). This choice stems from the fact that these function classes provide a general representation of the target function applicable to most real-world settings while offering an elegant means of measuring the smoothness through a concise set of fixed constants. Understanding the impact of these "regularity constants" on the theoretical properties of non-parametric frameworks holds significant interest, as it allows for a more precise application and analysis of these methods. Moreover, it is important

to highlight that in certain research areas of non-parametric estimation, such as convergence rate analysis, these types of assumptions on the regularity class $\mathcal{C}$ are crucial. Without them, obtaining meaningful theoretical results becomes unfeasible, as shown in Theorem 3.1 of Györfi et al. [2002]. In this thesis, we will primarily consider the various Hölder or generalised Lipschitz continuity classes of functions for $\mathcal{C}$. The implications of this choice on the convergence of the Lipschitz interpolation framework are discussed in more detail in Section 2.1.2, see subsection on Hölder Continuous Functions.

Characteristics of the Non-parametric Regression Methods

An apriori understanding of the characteristics of the predictors of the non-parametric estimation framework is crucial in order to derive theoretical properties in the context of system identification. This is non-trivial, as in contrast to parametric model estimators which can utilise the properties of the associated pre-defined parameterised functional form, it is generally significantly more difficult to determine any function properties of the non-parametric predictors $(\hat{f}_n)_{n\in\mathbb{N}}$. Instead, the theoretical characterisation of these models relies on the regularity properties of the framework or on properties of underlying hyperparameter choices. These characteristics can be utilised implicitly to derive general theoretical results, e.g. upper bounds on convergence rates through regularity properties of the method (see Tsybakov [2004], Wynne et al. [2021]), or explicitly to derive theoretical tools that can be used in practice in conjunction with the non-parametric estimation, e.g. utilising Gaussian process kernel properties in order to obtain safety guarantees for non-linear system identification (see Berkenkamp et al. [2016] or Berkenkamp et al. [2017]). Chapter 4 and Chapter will consider the former and will utilise various characteristics of the Lipschitz interpolation framework in order to derive theoretical results pertaining to general consistency, convergence rates, and local guarantees. Chapter 5 focuses on the latter in the context of time series analysis by studying non-parametric frameworks that provide Lipschitz continuous-type properties with the goal of deriving theoretical properties related to stationarity, ergodicity and first hitting times. We refer to the beginning of Chapter 5 for a more detailed discussion on this topic.

2.1.2 The Space of Lipschitz Continuous Functions.

Lipschitz continuity is a fundamental notion in mathematical analysis that characterizes the smoothness of a function by measuring the extent to which it varies in its domain. Understanding Lipschitz continuity is crucial in modelling real-world phenomena, where smoothness assumptions can often be observed to hold for underlying data. Formally, consider the input space $\mathcal{X}$ defined in the previous section and let $\mathfrak{d} : \mathcal{X}^2 \to \mathbb{R}_{\geq 0}$ denote a metric on $\mathbb{R}^d$. Similarly, consider $\mathcal{Y}$ defined in the previous section and let $\mathfrak{d}_{\mathcal{Y}} : \mathcal{Y}^2 \to \mathbb{R}_{\geq 0}$ denote a metric on $\mathbb{R}$. The class of L-Lipschitz (continuous) functions with respect to $\mathfrak{d}_{\mathcal{X}}$, $\mathfrak{d}_{\mathcal{Y}}$ and some $L \in \mathbb{R}_+$ is then defined as follows:

$$Lip(L, \mathfrak{d}) := \{h : \mathcal{X} \to \mathcal{Y} | \, \mathfrak{d}_{\mathcal{Y}}(h(x), h(x')) \leq L \, \mathfrak{d}(x, x'), \forall x, x' \in \mathcal{X}\}.$$

We call the smallest non-negative number L^* for which f is L^*-Lipschitz the *best* Lipschitz constant of f, i.e. $L^* = \min\{L \in \mathbb{R}_{\geq 0} | f \in Lip(L, \mathfrak{d})\}$. The Lipschitz constant L serves as an upper bound on the rate of change of the function and can be interpreted as a measure of the function's smoothness or stability. Typically, smaller values of L imply a smoother function, while larger values indicate a more rapidly changing function. This can be formally observed by noting the following elementary lemma that states that the Lipschitz constant is either exactly equal or upper bounds the maximum gradient of f:

Lemma 2.1.1 *Let $\mathcal{X}$ be a compact and convex*[5]*. If $f \in C^1(\mathcal{X})$, then f is L_p^*-Lipschitz*[6] *with respect to $\|.\|_p$. Furthermore, if f can be extended on an open set $\bar{\mathcal{X}}$ such that $\mathcal{X} \subset \bar{\mathcal{X}}$, then $L_p^* = \max_{x \in \mathcal{X}}\{\|\nabla f(x)\|_q\}$ for $p = 1, 2$ and $L_p^* \leq \max_{x \in \mathcal{X}}\{\|\nabla f(x)\|_q\}$ for $p > 2$ where q is the Hölder conjugate of p.*

Proof (For the sake of completeness, we include the proof of the result).

The fact that f is L_p^*-Lipschitz follows directly from the fact that $\mathcal{X}$ is compact and

[5]Note: the assumption of convexity on $\mathcal{X}$ is slightly unusual but is necessary as we consider $f \in C^1(\mathcal{X})$ and not $f \in C^1(\mathbb{R}^d)$.

[6]Here, the p index is used to explicitly denote the dependence on the $\|.\|_p$ norm.

$f \in C^1(\mathcal{X})$.

$\forall p \in \mathbb{N}$, $L_p^* \leq \max_{x\in\mathcal{X}}\{\|\nabla f(x)\|_q\}$ follows from the multidimensional mean-value theorem, an application of the Hölder inequality and the fact that $\mathcal{X}$ is convex.

For $p = 1, 2$, we show $L_p^* \geq \max_{x\in\mathcal{X}}\{\|\nabla f(x)\|_q\}$. Consider : $\forall x \in \mathcal{X}$ consider the Frechet derivative of f at x; $\lim_{t\to 0} \frac{|f(x+th)-f(x)-\nabla f(x)^\top(th)|}{t} = 0 \ \forall h \in \mathbb{R}^d$. Then, choosing $h = \nabla f(x)$ for $p = 2$ and $h = e_{i^*}$ such that $|\nabla f(x)^\top e_{i^*}| = \|\nabla f(x)\|_\infty$ for $p = 1$ gives $L_p^* \geq \max_{x\in\mathcal{X}}\{\|\nabla f(x)\|_q\}$. Note that this reasoning is well defined because $f \in C^1(\tilde{\mathcal{X}})$ and $\tilde{\mathcal{X}}$ is an open set that contains $\mathcal{X}$ which implies that $f(x+th)$ is well-defined for any $h \in \mathbb{R}^d$ and small enough t.

■

As noted in the introduction, Lipschitz continuity has been utilised explicitly in a variety of different fields ranging from non-parametric estimation to global optimisation and adversarial robustness. The basis of these methods is generally the same and utilises the following consequence of Lipschitz continuity. Assume[7] that the value of the target function f is known at a sample input $x \in \mathcal{X}$ and an estimate of its value at an input $x^* \in \mathcal{X}$ is desired. Then, the Lipschitz continuity of f directly implies:

$$f(x^*) \in [f(x) - L\,\mathfrak{d}(x, x^*), f(x) + L\,\mathfrak{d}(x, x^*)] \tag{2.3}$$

providing bounds on potential values of f at x^*. This information can be utilised in practice for various applications such as the elimination of certain exploration regions when searching for a global maximum (Shubert [1972]) or to design decision-making frameworks that are robust to worst-case errors (Manzano et al. [2020]). It is relatively intuitive that the tightness of the Lipschitz constant to the best Lipschitz constant L^* of f has a significant impact on the tightness of the bounds given in (2.3) and by consequence on the performance of these computational applications as can be observed in the convergence rates of the Lipschitz optimisation methods (Bachoc et al. [2021]) and the bounds on the propagation of the estimation error (Manzano et al. [2020]) for the two applications given above. This connection has also been

[7] Here, $\mathfrak{d}_\mathcal{Y}$ is the metric on $\mathbb{R}$ induced by $|.|$.

extensively studied in the context of neural networks where the size of the Lipschitz constant of a network has been shown to be directly linked to its generalisation (e.g. Bartlett et al. [2017], Negrini et al. [2021]) and robustness capabilities (e.g. Szegedy et al. [2013], Pauli et al. [2021]). For the Lipschitz interpolation frameworks introduced in the following section and studied in Chapter 4, the relation given in (2.3) is fundamental for both the design of the method and the derivation of its theoretical properties.

The Space of Hölder Continuous Functions.

The definition given above on the space of Lipschitz continuous functions is relatively non-standard as the usual definition is only defined with respect to norms on $\mathbb{R}^d$. This difference implies that the general Lipschitz continuity classes defined above could be considered with respect to a metric of the form $\|.\|^\alpha$ where $\|.\|$ is a norm on $\mathbb{R}^d$ and $\alpha \in (0, 1)$ is an exponent which is more commonly known as the class of Hölder continuous functions with parameters (L, α) with respect to $\|.\|$:

$$Hol(L, \alpha, \|.\|) := \{h : \mathcal{X} \to \mathcal{Y} || h(x) - h(x')| \leq L\|x - x'\|^\alpha, \forall x, x' \in \mathcal{X}\}.$$

As noted in the previous section, considering Hölder-type regularity is fairly standard practice in the literature on the convergence of non-parametric frameworks as it allows for an accurate characterisation of the asymptotic convergence rates, see Stone [1982], Van Der Vaart and Van Zanten [2011], Schmidt-Hieber [2020] for a selection of relevant research articles. More precisely, the α exponent allows for an intuitive understanding of fractional-power changes in smoothness of the target function without having to rely on more technical theoretical machinery and can be used to measure the effect of smoothness on the convergence speed of non-parametric estimation in a quantifiable manner; as done in the convergence rates proposed by the papers cited above which depend explicitly on α. Chapter 4 will consider this class of functions in order to derive similar asymptotic convergence rate results that depend on α for Lipschitz interpolation.

2.1.3 Reminder: Useful Theoretical Notions

In this brief subsection, we recall various mathematical definitions that will be used repeatedly throughout the thesis. This chapter is non-essential and can be skipped if the reader is already familiar with the presented concepts.

In various chapters, we will consider the analysis of algorithmic convergence rates and sample complexities. As our investigation primarily revolves around asymptotic considerations and the limiting behavior of the algorithms of interest, we will characterize our results utilising the Bachmann-Landau notations in the form of O, Ω, and Θ. We define these notations formally as follows.

In this brief subsection, we recall a few mathematical definitions that will be used repeatedly throughout the thesis. This chapter is non-essential and can be skipped if the reader is already familiar with the presented concepts.

In various chapters, we will consider the analysis of algorithmic convergence rates and sample complexities. As our investigation primarily revolves around asymptotic considerations and the limiting behaviour of the algorithms of interest, we will characterise our results utilising the Bachmann-Landau notations in the form of O, Ω, and Θ. We define these notations formally as follows.

Definition 2.1.2 *Let $U \subset \mathbb{R}_+$ be unbounded and consider $f_1,\ f_2 : U \to \mathbb{R}_+$, two functions defined on U. Then, we define:*

- $f_1(x) \in O(f_2(x))$ *as* $x \to \infty$ *(with* $x \in U$*) if* $\limsup_{x\to\infty} \frac{f_1(x)}{f_2(x)} < \infty$.
- $f_1(x) \in \Omega(f_2(x))$ *as* $x \to \infty$ *(with* $x \in U$*) if* $\liminf_{x\to\infty} \frac{f_1(x)}{f_2(x)} > 0$.
- $f_1(x) \in \Theta(f_2(x))$ *as* $x \to \infty$ *(with* $x \in U$*) if* $f_1(x) \in O(f_2(x))$ *and* $f_1(x) \in \Theta(f_2(x))$.

We also recall the notion of ϵ-cover that will be used in multiple proofs of this thesis. In particular, see the notion of (ϵ, η)-covered used in Definition 3.3.6 of Chapter 3 and various proofs of Chapter 4.

Definition 2.1.3 *(ϵ-Cover) Let $d \in \mathbb{N}$, $\epsilon > 0$ and consider a set $\mathcal{X} \subset \mathbb{R}^d$ and a metric $\mathfrak{d}$ on $\mathbb{R}^d$. Denoting $B_\epsilon(x)$ the ball of radius ϵ centred in $x \in \mathcal{X}$ with respect to $\mathfrak{d}$, we define an ϵ-cover of $\mathcal{X}$ as a discrete subset $Cov(\epsilon) \subset \mathbb{R}^d$ such that $\mathcal{X} \subset \bigcup_{x \in Cov(\epsilon)} B_\epsilon(x)$ and the associated set of balls as $\mathcal{B} := \{B_\epsilon(x) | x \in Cov(\epsilon)\}$. We say furthermore that $Cov(\epsilon)$ is a ϵ-minimal cover of $\mathcal{X}$ if $|Cov(\epsilon)| = \min\{n : \exists \epsilon\text{-cover over } \mathcal{X} \text{ of size } n\}$.*

We note that an ϵ-cover is not necessarily unique nor finite and that under certain assumptions, the ϵ-minimal cover can always be obtained.

2.2 Lipschitz Continuous Machine Learning

So far, this chapter has revolved around a general discussion on non-parametric estimation with the goal of providing a conceptual intuition of the theoretical setting and assumptions that will be used later on in this thesis. As stated in the previous chapter, a majority of this thesis will be dedicated to deriving the properties of a specific class of non-parametric estimation methods known as Lipschitz Interpolation. In the following section, we define the framework and revisit certain of its key established theoretical properties.

2.2.1 Lipschitz Interpolation

Consider the setting described in the previous section and assume that the target function f belongs to the class of Lipschitz continuous functions $Lip(L^*, \mathfrak{d})$. Furthermore, assume the following basic noise assumption which will be expanded on in future chapters.

Assumption 1 *(Simple Noise Assumption.) Let $(e_k)_{k \in \mathbb{N}}$ denote the collection of noise variables defined in Section 2.1.1. Then, we assume that there exists noise bounds $\bar{\mathfrak{e}} > 0$ such that for all $k \in \mathbb{N}$, $e_k \in [-\bar{\mathfrak{e}}, \bar{\mathfrak{e}}]$.*

Utilising these two assumptions, we can formally define the general Lipschitz interpolation framework that is the basis of the majority of Lipschitz constant-based non-parametric estimators considered in the literature; see Milanese and Novara [2004], Beliakov [2006], Calliess et al. [2020], Manzano et al. [2021] for a selection.

Definition 2.2.1 *(Lipschitz interpolation) Using the setting described in Section 2.1.1, we define the sequence of predictors* $(\hat{f}_n)_{n\in\mathbb{N}}$, $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$ *associated with* $(D_n)_{n\in\mathbb{N}}$, *as*

$$\hat{\mathfrak{f}}_{\mathfrak{n}}(x) := \frac{1}{2}\mathfrak{u}_n(x) + \frac{1}{2}\mathfrak{l}_n(x),$$

where $\mathfrak{u}_n, \mathfrak{l}_n : \mathcal{X} \to \mathcal{Y}$ *are defined as*

$$\mathfrak{u}_n(x) = \min_{i=1,\dots,N_n} \tilde{f}_i + L\,\mathfrak{d}(x, s_i)$$
$$\mathfrak{l}_n(x) = \max_{i=1,\dots,N_n} \tilde{f}_i - L\,\mathfrak{d}(x, s_i)$$

and $L \in \mathbb{R}_{\geq 0}$ *is a selected hyper-parameter.*

Ideally, the hyper-parameter $L \in \mathbb{R}_{\geq 0}$ can be set to be larger than the best Lipschitz constant L^* of the unknown target function. This can generally be achieved by assuming some prior knowledge of the properties of the system dynamics or by utilising a principled Lipschitz constant estimation approach; see Chapter 3 for an in-depth discussion on the subject. In this case, a series of finite sample and worst-case guarantees that depend on the density of the grid $G^{\mathcal{X}}$ on $\mathcal{X}$ can be derived (Calliess et al. [2020]). These are stated in the theorem given below.

Theorem 2.2.2 *(Calliess et al. [2020]) Assume that the target function* $f \in Lip(L^*, \mathfrak{d})$ *for* $L^* > 0$ *and that* $(D_n)_{n\in\mathbb{N}}$ *becomes uniformly dense in* $\mathcal{X}$ *with at a rate* $(r(n))_{n\in\mathbb{N}} \in o(1)$ *(see Definition 4.3.1 for a precise definition). Then, denoting* $(\hat{f}_n)_{n\in\mathbb{N}}$ *the predictors generated by the Lipschitz interpolation framework with a hyper-parameter* $L > L^*$, *we have*

- *(Finite Sample)* $\forall n \in \mathbb{N}\ \forall x \in \mathcal{X},\ |\hat{f}_n(x) - f(x)| \leq (L^* + L)r(n) + 2\bar{\mathfrak{e}}$
- *(Asymptotic)* $\lim_{n\to\infty} \|\hat{f}_n - f\|_\infty \in [0, 2\bar{\mathfrak{e}}]$

- *(Lipschitz Continuity) The predictors* $(\hat{f}_n)_{n\in\mathbb{N}}$ *are Lipschitz continuous with Lipschitz constant* L.

Proof These results follow from Lemma 6, Theorem 9 and Corollary 10 of Calliess et al. [2020] which provides proofs in a more general context[8].

■

The properties stated in Theorem 2.2.2 and their more general formulation given in Calliess et al. [2020] have been particularly useful in the context of control where applications of a fully data-driven[9] extension of Lipschitz interpolation have been considered in model reference adaptive control (Calliess et al. [2020]) and model predictive control (Manzano et al. [2020]).

While this first set of theoretical results showcases the usefulness of the Lipschitz interpolation framework, the presence of noise in the observation implies that the estimation bounds discussed in Section 2.1.2 which follow from the Lipschitz continuity of the target function do not hold. As these deterministic error bounds can be useful in practice, one would want to extend the Lipschitz interpolation framework in order for them to hold. In settings where bounds on the support of the noise are known, an alternative framework Lipschitz interpolation framework that proposes worst-case error bounds can be considered.

Remark 2.2.3 *(Alternative Formulation) In some works (see in particular Milanese and Novara [2004]) an alternative formulation is given for the Lispchitz interpolation predictors. In this case, an upper bound* $\bar{\mathfrak{e}}'$ *on* $\bar{\mathfrak{e}}$ *is assumed known and is explicitly used in the formulation of* $\mathfrak{u}^2, \mathfrak{l}_n^2 : \mathcal{X} \to \mathcal{Y}$*:*

$$\mathfrak{u}_n^2(x) = \min_{i=1,\dots,N_n} \tilde{f}_i + L\,\mathfrak{d}(x, s_i) + \bar{\mathfrak{e}}'$$
$$\mathfrak{l}_n^2(x) = \max_{i=1,\dots,N_n} \tilde{f}_i - L\,\mathfrak{d}(x, s_i) - \bar{\mathfrak{e}}'.$$

[8] (Calliess et al. [2020]) provides general worst-case bounds for a fully data-driven approach which hold even when the Lipschitz constant estimate utilised in the Lipschitz interpolation framework is smaller than L^*.

[9] By "fully data-driven", we mean that the Lipschitz interpolation extension includes an in-built Lipschitz constant estimator.

This formulation is useful for computing tight worst-case upper and lower bound guarantees in practice and extended by considering asymmetric error bounds, i.e. $e \in [\bar{\mathfrak{e}}_1, \bar{\mathfrak{e}}_2]$ *with probability* 1 *where* $\bar{\mathfrak{e}}_1 < 0 < \bar{\mathfrak{e}}_2 \in \mathbb{R}$.

In the context of this thesis, these two frameworks can be treated equivalently as all the theoretical properties derived for the classical Lipschitz interpolation framework hold for the alternative framework. In fact, obtaining theoretical results for the alternative formulation is slightly easier if the noise bounds ($\bar{\mathfrak{e}}$) of Assumption 1 are tight with respect to the support of the noise distribution of $(e_k)_{k\in\mathbb{N}}$ and $\bar{\mathfrak{e}}' = \bar{\mathfrak{e}}$. This follows from the fact that under these additional assumptions, one can study the noise present in each of the $\mathfrak{u}_n^2$, $\mathfrak{l}_n^2$ estimators individually by leveraging the provided noise bound. By contrast, for the $\mathfrak{u}_n$, $\mathfrak{l}_n$ estimators, the theoretical examination must ensure that the noise present in each estimator counterbalances in the mean predictor $\hat{\mathfrak{f}}_n$.

While the alternative formulation of the Lipschitz interpolation method necessitates the additional assumption that one has prior knowledge of the bounds on the noise, in exchange it offers optimality guarantees and the capability to establish bounds on the set of all dynamics that are consistent with observed data set and provided Lipschitz constant and noise bounds (Milanese and Novara [2004]). This is described formally in the following theorem.

Theorem 2.2.4 *(Milanese and Novara [2004]) Let* $f \in Lip(L^*, \|.\|)$ *for* L^* *and denote* $(\hat{f}_n^2)_{n\in\mathbb{N}}$ *the predictors generated by the alternative Lipschitz interpolation framework with a hyper-parameter* $L > L^*$ $\bar{\mathfrak{e}}' > \bar{\mathfrak{e}}$. *Finally, define the feasible systems set as follows:*

$$FSS_n := \{g \in Lip(L^*, \|.\|) \big| \forall (x, \tilde{y}) \in D_n,\ |g(x) - \tilde{y}| \leq \bar{\mathfrak{e}}'\}$$

with worst-case bounds; $\bar{f}_n : \bar{f}_n(w) := \sup_{g\in FSS_n} g(w)$ *and* $\underline{f}_n : \underline{f}_n(w) := \inf_{g\in FnSS_n} g(w)$. *Then,* $(\hat{f}_n^2)_{n\in\mathbb{N}}$ *has the following finite sample properties:*

- $\forall n \in \mathbb{N}$, *the worst case bounds correspond to* $\mathfrak{u}_n^2$, $\mathfrak{l}_n^2$*:* $\bar{f}_n \equiv \mathfrak{u}_n^2$, $\underline{f}_n \equiv \mathfrak{l}_n^2$.

- $\forall n \in \mathbb{N},\ \hat{f}_n^2 \in \operatorname{arginf}_{\hat{f}} \sup_{g \in FSS} \|\hat{f} - g\|$.

Proof This proof of this result follows from Theorem 2 and Theorem 7 of Milanese and Novara [2004] who considers a more general setting[10].

■

In essence, Theorem 2.2.4 states that Lipschitz interpolation provides the best worst-case bounds that are consistent with the known information: $(L, \bar{\mathfrak{e}}_2, D_n)$. It also asserts that the predictors $(\hat{f}_n^2)_{n\in\mathbb{N}}$ generated by the Lipschitz interpolation framework are the optimal solutions that minimise the worst-case loss. The optimality of both the worst-case bounds and the predictors suggest that Lipschitz interpolation is a particularly useful method within the class of Lipschitz regularity-based safe-learning approaches described in Chapter 1. Various applications utilising the alternative Lispchitz interpolation framework and the results stated in Theorem 2.2.4 have been pursued in predictive control (see Canale et al. [2009], Canale et al. [2014]).

Remark 2.2.5 *It is important to note that in the setting considered so far where the noise bounds $\bar{\mathfrak{e}}$ and upper bound $\bar{\mathfrak{e}}'$ are considered constant, the non-parametric predictors $(\hat{f}_n)_{n\in\mathbb{N}}$ and $(\hat{f}_n^2)_{n\in\mathbb{N}}$ are equivalent. This implies that the predictors $(\hat{f}_n)_{n\in\mathbb{N}}$ defined without knowledge of $\bar{\mathfrak{e}}$ or $\bar{\mathfrak{e}}'$ also coincide with the optimal solutions that minimise the worst-case loss (in this case $\bar{\mathfrak{e}}$ can be utilised instead of $\bar{\mathfrak{e}}'$ in the definition of FSS_n).*

In addition to the alternative Lipschitz interpolation framework, several extensions of the classical Lipschitz interpolation approach have been proposed. Calliess et al. [2020] relax the assumption of prior knowledge of the Lipschitz constant in favour of a fully data-driven approach, Maddalena and Jones [2020a] propose an equivalent smooth formulation which is more suited for controllers that rely on gradient computations, (Blaas et al. [2019] extend the framework by incorporating localised Lipschitz constants, and Manzano et al. [2022] propose a computationally more ef-

[10] In particular, in Milanese and Novara [2004], the noise bounds $\bar{\mathfrak{e}}$ are assumed to be input-dependent.

ficient approach that retains key properties of the original Lipschitz interpolation framework that are used in the context of model predictive control.

2.2.2 Lipschitz Constant Estimation

The main drawback of the Lipschitz interpolation described in the previous section is the necessity of prior knowledge of the Lipschitz constant in order to set the L hyper-parameter of the framework. Unfortunately, the problem of estimating the Lipschitz constant of an unknown target function is far from trivial and has been treated in the context of Lipschitz interpolation in numerous research articles. For example, Milanese and Novara [2004] utilise a shallow neural network-based approach to obtain an upper bound estimate, Beliakov [2006], Calliess [2017] consider an optimisation-based framework or more recently Calliess et al. [2020] define a "LACKI rule" estimator that can deal with online data in a computationally efficient way to obtain real-time Lipschitz constant estimates. More generally, the problem has attracted multidisciplinary research[11] interest in fields such as global optimisation with the development of Lipschitz constant estimators in adaptive Lipschitz optimisation algorithms (e.g. Malherbe and Vayatis [2017]) or in the context of designing safe initial policies for reinforcement learning via approximate dynamic programming techniques and kernelized Lipschitz constant estimators (Chakrabarty et al. [2020]). An in-depth literature review of existing Lipschitz constant estimation methods can be found in Chapter 3.

It is important to emphasize that although obtaining a loose[12] Lipschitz constant estimate can be fairly straightforward in some cases, this imprecision often leads to sub-optimal performance of the Lipschitz constant dependent application; this follows from the discussion given in Section 2.1.2 on the practical usefulness of the Lipschitz condition given in (2.3). As a consequence, it is necessary for Lipschitz constant estimation methods to guarantee a certain degree of accuracy in order to facilitate the effective use of these applications. As a last remark, we note that

[11] A more extensive list of applications is provided in Chapter 3.
[12] We utilise "loose" to signify Lipschitz constants L such that $L >> L^*$.

the theoretical aspects of this problem such as deriving convergence rates or sample complexity bounds, have been largely ignored as existing research focused on developing practical algorithms. Chapter 3 will focus on addressing this gap.

2.2.3 Connections to Artificial Neural Networks

The last chapter of the thesis considers a general theoretical framework that can be applied to any Lipschitz continuous machine learning method. While this implies that the Lipschitz interpolation framework described above could be used, an alternative approach based on the application of feed-forward neural networks and its Lipschitz continuity properties is proposed. For completeness, we briefly discuss this class of parametric estimators and its connection to Lipschitz regularity. A more extensive overview of neural networks can be found in Goodfellow et al. [2016].

A feed-forward neural network can be modelled as a recursive application of an activation function and a series of matrix-vector products. More formally, let $L \in \mathbb{N}$ denote the number of layers, $\{d_i\}_{i=0,...,L}$ $d_i \in \mathbb{N}$ for all $i \in \{0, ..., L\}$ denote the number neurons on each layer, $\{\sigma_i\}_{i=1,...,L}$ $\sigma_i : \mathbb{R} \to \mathbb{R}$ for all $i \in \{1, ..., L\}$ denote the activation functions and $\{W_i\}_{i=1,...,L}$, $\{b_i\}_{i=1,...,L}$ with $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ and $b_i \in \mathbb{R}^{d_i}$ for all $i \in \{1, ..., L\}$ denote the weights of the neural network. The following representation ($\hat{f}$) of a fully connected feed-forward neural network can then be given:

$$\hat{f}(x) = \sigma_L(W_L\sigma_{L-1}(W_{L-1}\sigma_{L-2}(\ ...\ \sigma_1(W_1x + b_1)\ ...\ ) + b_{L-1}) + b_L) \tag{2.4}$$

where σ is applied component-wise at each iteration. The initial and final (L-th) layers can be referred to as the input and output layers, while all other layers are termed hidden layers. Note that, in the context of the estimation problem described in (2.1), $d_0 = d$, $d_L = 1$, and the rest of the layer dimensions can be set arbitrarily. The number of layers (L), the dimensions of the weights ($\{d_i\}_{i=0,...,L}$), and the activation functions ($\{\sigma_i\}_{i=1,...,L}$) together fully characterise the neural network.

The estimation properties of these types of networks are well-known, in particular,

we highlight Hornik et al. [1989] for foundational work on the universal approximation properties of single-layered neural networks as the number of neurons goes to infinity and Bauer and Kohler [2019], Schmidt-Hieber [2020] for recent research which utilises a non-parametric estimation setting to derive asymptotic convergence rates for feed-forward neural networks as the number of layers increases with the size of the data set.

Lipschitz Continuity of Neural Networks

As noted in Section 2.1.2, establishing the Lipschitz continuity of a neural network and knowing its Lipschitz constant can yield several advantageous properties pertaining to adversarial robustness (e.g. Pauli et al. [2021] and references cited therein) and generalisation (e.g. Bartlett et al. [2017] and ensuing works). Furthermore, in the context of system identification and control, several recent results have utilised Lipschitz regularity properties in conjunction with neural networks. Some examples include Zhou et al. [2022] who utilise the Lipschitz continuity (and bounds on the Lipschitz constant) of both the learned networks and the unknown dynamics in order to derive closed-loop stability for a dual neural network approach that learns the system dynamics, a valid Lyapunov function and a stabilising controller or Knuth et al. [2021] who bound the difference between the learned and target system dynamics by utilising a Lipschitz constant estimate of this difference and obtains conditions ensuring the existence of a one-step feedback law, therefore, enabling the design of a planner that provides safety and performance guarantees.

The Lipschitz regularity of the neural network described in (2.4) can be trivially derived if the activation functions are assumed to be Lipschitz continuous[13] and a loose Lipschitz constant estimate can obtained from upper bounds on the weights $\{W_i\}_{i=1,\dots,L}$, $\{b_i\}_{i=1,\dots,L}$ and the Lipschitz constant of the activation functions $\{\sigma_i\}_{i=1,\dots,L}$. Estimating a tighter[14] bound on the Lipschitz constant of the neural network is however far less straightforward in general and significant research has been carried out in this direction. These approaches are generally computationally intensive and ne-

[13] Most popular choices of activation function are Lipschitz continuous.

[14] Estimating a tight bound is in fact NP-hard even for simple neural networks Virmaux and Scaman [2018].

cessitate the computation of optimisation problems: notably, Fazlyab et al. [2019] solve a convex optimisation problem to obtain an accurate Lipschitz constant estimate of the neural network while Jordan and Dimakis [2020] utilise a Mixed-Integer Problem (MIP) based approach. An alternative computationally quick but significantly more approximative estimate of the Lipschitz constant can be obtained for the $\|.\|_1$ and $\|.\|_2$ norms if the dimension of the input space is small. This approach was described by Scaman and Virmaux [2018] and consists of directly computing the input-output gradients through back-propagation and utilising a simple grid search in order to find the maximum of the norm[15] of the gradient values; note that this strategy follows naturally from Lemma 2.1.1 which implies that this maximum corresponds to the best Lipschitz constant of the target function. While this approach is less principled than the other aforementioned methods, it can be useful for practical reasons depending on the standing assumptions. In Chapter 5, we will base our empirical analysis on this procedure in order to apply our theoretical results in the context of neural network-based auto-regressive time series models for financial time series analysis.

[15] For a properly selected norm.

3 | On the Sample Complexity of Lipschitz Constant Estimation

Contents

3.1 Introduction

This chapter will study the problem of estimating a Lipschitz constant of a target function from data under minimal parametric assumptions. As discussed in Chapter 2, this is primarily motivated by Lipschitz interpolation frameworks, whose main practical drawback is the dependency on the assumption of prior knowledge of the Lipschitz constant or, in the case that this assumption is not made, the necessity of having to learn a precise Lipschitz constant estimate. More generally, this issue is shared with a number of computational applications which are based on the Lipschitz continuity of an objective or target function and depend explicitly on the value of a Lipschitz constant. Examples include: robustness analysis in control settings which rely on Lipschitz interpolation and therefore utilise Lipschitz constants to characterise worst-case behaviour (Limon et al. [2017], Canale et al. [2014]), global optimization algorithms which rely on precise Lipschitz constant estimates to ensure speedy convergence (Jones et al. [1993], González et al. [2016], Malherbe and Vayatis [2017]), multi-armed bandit problems which utilise the Lipschitz constants to obtain asymptotic lower bounds and design algorithms (Magureanu et al. [2014]) or reinforcement learning which utilises the Lipschitz constant to construct safe initial policies (Chakrabarty et al. [2020]). For these applications, it is critical that the Lipschitz constant estimate used is sufficiently precise in order to ensure satisfactory performance in their specified goal.

Consequently, a number of Lipschitz constant estimation methods (also called Lipschitz learning algorithms) have been developed. For target functions belonging to families of parametric models, Lipschitz learning approaches generally utilise the structure of the model class to obtain precise estimates, see for example the discussion given in Section 2.2.3 on Lipschitz constant estimation and neural networks. For frameworks that don't consider a particular parametric family, a majority of the existing approaches are black-box methods that utilise and extend Strongin's classical

estimator (Strongin [1973]): $\hat{L} := r \max_{i \neq j} \frac{|f_i - f_j|}{\|x_i - x_j\|}$ where $r \in \mathbb{R}$ is a multiplicative factor and (x_i, f_i) is a data sample with $f_i = f(x_i)$. In particular, we highlight: Wood and Zhang [1996] builds on Strongin's estimator by fitting an approximate reverse Weibull distribution to the Lipschitz estimate in the one-dimensional case and using the location parameter as a Lipschitz estimate, Sergeyev [1995] utilises Strongin's approach to compute local Lipschitz constant estimates and extends the approach to the multidimensional case by using space-filling curves in order to solve a global optimization problem and a more recent approach Strongin et al. [2019] proposes dual Strongin Lipschitz estimates: with two differing "local" and "global" multiplicative factors. We remark that the class of Lipschitz learning algorithms described so far does not consider the possibility of observational noise and can explode in value if it exists. In this case, we can consider the Lispchitz constant estimator proposed by both Novara et al. [2013] and Calliess et al. [2020] which specifically extends Strongin's estimate to deal with bounded observational noise.

Alternative Lipschitz learning approaches that do not directly utilise Strongin's estimate have also been developed. These generally can consider the case of observational noise and include: Beliakov [2005] which utilises a short optimisation problem and cross validation/sample splitting to obtain Lipschitz constant estimates, Bubeck et al. [2011] which employs a similar idea to Strongin's estimate in order to propose a Lipschitz constant estimator in the context of the Lipschitz multi-armed bandit problem, González et al. [2016] which generates Lipschitz constant estimates using the mean function of the gradient function estimate of a fitted Gaussian Process (GP) and is directly computable using the GP-associated covariance function, and Calliess [2017] which obtains Lipschitz constant estimates by optimising a Lipschitz interpolation method. Unfortunately, while this class of Lipschitz learning algorithms tends to work well in practice, they generally do not guarantee asymptotic convergence and are of limited theoretical interest.

Despite the wide range of proposed Lipschitz learning algorithms, there has been little theoretical investigation into the Lipschitz constant estimation problem other than various consistency proofs of Strongin-based estimators. It is generally under-

stood that this learning problem, without making further restrictive assumptions on the underlying space of target functions, is inevitably subject to the curse of dimensionality. However, to the best of our knowledge, this intuition has not been explored formally. A first goal of this chapter is therefore to provide a theoretical investigation into the Lipschitz learning problem by deriving lower bounds on the sample complexity in the case of both noiseless and noisy sampling settings. We confirm the general intuition on the difficulty of the Lipschitz learning problem by demonstrating that the problem has a sample complexity lower bound that scales at least exponentially on the function input dimension in both the noiseless case and the noisy sampling case when the noise is assumed to be Gaussian.

Remark 3.1.1 *The closest work proposing related theoretical learning results can be found in the literature on global optimisation, e.g. see Bull [2011] and Wang et al. [2018], as Lipschitz learning can be conceptualised as a type of global optimisation problem defined on the gradient of a target function. However, we note that the convergence and sample complexity rates derived in these papers cannot be directly extended to Lipschitz constant estimation, necessitating additional analysis. A second relevant sub-field connected to the theoretical properties of Lipschitz constant estimation is the study of non-parametric convergence rates such as those derived by Stone [1982]. This connection is discussed more precisely in Remark 3.2.8.*

Our theoretical results imply that a precise estimation of the Lipschitz constant requires a significant number of samples. This is computationally problematic for classical Strongin-based Lipschitz learning algorithms due to the fact that the computational complexity of these methods can be shown to be quadratic in the number of samples: $O(n_{samples}^2)$. While the non-Strongin based estimators discussed above could potentially be less computationally expensive, they only provide heuristic or experimental convergence guarantees and are generally difficult to study from a theoretical perspective. Therefore, in light of our lower bounds on sample complexity, we propose a novel algorithm for Lipschitz learning called LCLS (short for *Lipschitz Constant estimation by Least Squares regression*) that has a linear computational complexity in the number of sample points and for which we can derive theoretical

guarantees on asymptotic convergence and finite sample behaviour in a general noisy sampling setting. The optimality of the lower bounds on the sample complexity of the Lipschitz learning problem in both the noiseless sampling setting and in the noisy sampling setting under Gaussian noise assumptions derived in the first part of the chapter follow from these theoretical results.

In practice, the proposed LCLS algorithm provides a theoretically motivated and computationally quick way of estimating the Lipschitz constant. With minimal fine-tuning, LCLS can be plugged into any computational method that utilises a Lipschitz constant estimate – see above discussion – under any sampling noise assumptions. We provide an example of such a procedure in the context of non-parametric regression for system identification in control by combining the LCLS algorithm with a classical nonlinear set-membership/Lipschitz interpolation framework and illustrating the empirical performance of the combined regression method through a series of short experiments.

A more comprehensive list of the contributions of this chapter is given below.

3.1.1 Contributions & Outline of Chapter

In this chapter, we provide a rigorous treatment of the Lipschitz constant estimation problem discussed above. In particular, we make the following contributions:

1. In Section 3.2, we provide novel theoretical lower bounds on the sample complexity of the general Lipschitz learning problem (for $p \in \{1, 2\}$) in the noiseless sampling setting (see Theorem 3.2.3) and of a slightly modified version of the problem[1] in the noisy sampling setting (see Theorem 3.2.6) when the target function $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ satisfies a regularity condition $C^2(\mathcal{X}, K)$ defined in Assumption 3. We note that these bounds can be equivalently stated as lower bounds on the convergence rate of the general Lipschitz learning problem (see Corollaries 3.2.4 and 3.2.7). As far as the authors are aware, the sample complexity and convergence bounds derived in this chapter are the first theoretical

[1] Which can be shown to be equivalent for the majority of existing Lipschitz learning algorithms.

results pertaining to the convergence of the general Lipschitz learning problem. We show in Section 3.3 that our proposed lower bound on the sample complexity rate is optimal in both the noiseless sampling setting and the noisy sampling setting under a Gaussian assumption.

2. In Section 3.3.1, we propose a least squares-based Lipschitz learning (LCLS) approach that utilises a partition of the input space $\mathcal{X}$ and local least squares estimates in order to generate a Lipschitz constant estimate. As discussed in the introduction, our motivation for developing the LCLS algorithm rests on the following two points:

    - both theoretically and computationally tractable.
    - directly applicable across all noise settings considered in the literature and implementable without any prior knowledge of target function properties or of the noise structure.

3. In Sections 3.3.2 and 3.3.3, we investigate the theoretical properties of the proposed algorithm:

    - Asymptotic convergence for general partition choice in noiseless and general noisy sampling set-ups (Section 3.3.2, see Theorem 3.3.7).
    - Finite sample guarantees in the noiseless and general noisy sampling set-ups when the partition is constructed using regular hypercubes (Section 3.3.3, see Theorem 3.3.10, Corollary 3.3.16). These guarantees can be used to provide an upper bound on the sample complexity of the Lipschitz learning problem and show that the complexity rates derived in the first part of the chapter match in both the noiseless and noisy settings under a Gaussian assumption. (Section 3.3.3, see Remark 3.3.11, 3.3.12, Theorem 3.3.15).

4. In Section 3.3.4, we illustrate and compare the empirical performance of the LCLS algorithm against Strongin-based Lipschitz learning algorithms on a set of test functions. We consider both the noiseless and noisy sampling settings. We find that while the benchmark Strongin-based algorithms converge slightly

faster in terms of number of samples, our proposed algorithm converges faster in terms of computation time for all functions in the test set (see Figure 3.4). This is despite the fact that we consider noise settings for which the benchmark algorithms are specifically designed in our experiments.

5. In Section 3.4, we explore the application of the various theoretical results and the LCLS algorithm derived in this chapter to the fields of *Global Optimisation* and *Non-parametric Regression for System Identification*. More specifically, we propose a lower bound on the sample complexity of adaptive Lipschitz optimisation algorithms that follows from one of the theoretical results stated in Section 3.2 and a new non-parametric regression method constructed by combining the LCLS algorithm of Section 3.3 with a classical nonlinear set membership framework (see Milanese and Novara [2004]).

3.2 Assumptions & Sample Complexity Lower Bound

In this section, we provide the standing assumptions of the chapter and state the main results pertaining to theoretical lower bounds on the sample complexity of Lipschitz learning algorithms.

3.2.1 Basic Assumptions

Let $p \in \mathbb{N}$, $d \in \mathbb{N}$. In this chapter, we will consider Lipschitz continuous functions with respect to $\|.\|_p$ norms on $\mathbb{R}^d$. In order to alleviate notation and improve clarity, we will make the following additions to the notation given in Chapter 2 on Lipschitz continuity.

Notation 3.2.1 *Let $p \in \mathbb{N}$, $d \in \mathbb{N}$. We utilise the following notation:*

1. *For $f \in Lip(L, \|.\|_p)$ with best Lipschitz constant L^*, we make explicit the dependence of L^* on f and p, i.e. $L_p^*(f) := L^*$.*

2. *We classify the set of Lipschitz continuous functions according to the best Lipschitz constants:*

$$\mathcal{F}_p(L) := \{h : \mathcal{X} \to \mathbb{R} | h \text{ is Lipschitz } \wedge L_p^*(h) = L\}.$$

The Lipschitz learning problem considers the estimation of the best Lipschitz constant $L_p^*(f)$ where f is an unknown target function. As described in the introduction of this chapter, we consider a general version of this problem where f is considered black-box and can only be accessed through queries to a, possibly noisy, oracle. As f is not assumed to belong to any parametric family, other assumptions are needed[2] in order to derive theoretical bounds on the sample complexity. For our results, we make the following two assumptions on the input space $\mathcal{X}$ and the regularity of f.

Assumption 2 *(Domain) The domain $\mathcal{X}$ of the target function f is a convex and compact sub-set of $\mathbb{R}^d$.*

Assumption 3 *(Functional) The target function $f \in C^2(\mathcal{X})$ and there exists an upper bound $K \in \mathbb{R}_+$ on the second-order partial derivatives of f, i.e. $|\frac{\partial f}{\partial x_i \partial x_j}| \leq K$ for all $x \in \mathcal{X}$ and $i, j \in \{1, ..., d\}$. Furthermore, f can be extended on an open set $\bar{\mathcal{X}}$ such that $\mathcal{X} \subset \bar{\mathcal{X}}$.*

For a given $K \in \mathbb{R}_+$, we denote by $C^2(\mathcal{X}, K)$ the class of functions that satisfies Assumption 3 with an upper bound K on the second degree partial derivatives. It is important to point out that this bound does need to be tight and that if Assumption 2 holds then any $f \in C^2(\mathcal{X})$ automatically belongs to $C^2(\mathcal{X}, \bar{K})$ for some $\bar{K} \in \mathbb{R}_+$. Finally, we assume that we have access to the target function f through an oracle $\Omega : \mathcal{X} \to \mathbb{R}$ – defined formally below for each sampling setting – which can be queried in order to generate observations of f. In particular, this oracle can be freely used by any Lipschitz learning algorithm as described in the following definition.

Definition 3.2.2 *(Lipschitz Learning Algorithms) We define $\mathcal{L}_{n,p}(\mathcal{X})$ as the set of all $\|.\|_p$-Lipschitz learning algorithms that utilise at most $n \in \mathbb{N}$ queries to the Oracle*

[2]Otherwise, a theoretical characterisation of the Lipschitz learning problem is not feasible.

Ω with inputs in $\mathcal{X}$. The sampling procedure is considered to be a part of the Lipschitz learning algorithm and $\forall \hat{L} \in \mathcal{L}_{n,p}(\mathcal{X})$ we denote the set of generated samples by $D_{\hat{L}} = \{(x_i^{\hat{L}}, \Omega(x_i^{\hat{L}}))_{i=1,\dots,n}\}$.

We note that Definition 3.2.2 defines a general class of Lipschitz learning algorithms without any structural specifications and that the inclusion of the sampling procedure in the algorithm is common for applications in both control and global optimisation.

3.2.2 Noiseless Sampling Setting

Assumptions (2)-(3) are sufficient to formulate a lower bound on the sample complexity rate of the Lipschitz learning problem in the case where one has access to an oracle[3] Ω that can be queried to obtain noiseless observations of the underlying target function. Formally, the noiseless Oracle is described by

$$\begin{aligned} \Omega: \ & \mathcal{X} \to \mathbb{R} \\ & x \overset{\Omega}{\mapsto} f(x). \end{aligned}$$

The lower bound on the sample complexity of any Lipschitz learning algorithm is given in the following theorem.

Theorem 3.2.3 *(Sample Complexity Bound – Noiseless) Let $M \in \mathbb{R}_+$, $d \in \mathbb{N}$, $p \in \{1, 2\}$ and suppose $\mathcal{X} := [0, M]^d$. Assume that Assumption (3) holds and that a noiseless Oracle Ω, (described above) is available. $\forall L^* \geq 0, \forall \epsilon \in (0, MK)$, if*

$$\inf_{\hat{L} \in \mathcal{L}_{n,p}(\mathcal{X})} \sup_{f \in C^2(\mathcal{X},K) \cap \mathcal{F}_p(L^*)} |\hat{L}(f) - L^*| < \epsilon$$

then

$$n \geq \left(C(d,p) \frac{MK}{\epsilon} \right)^d$$

In this chapter, we find $C(d,p) = \frac{1}{20 d^{\frac{1}{p} - \frac{1}{2}}}$, however this value has not been optimized.

[3]Note: in the noiseless case, the oracle and the target function are equivalent.

Theorem 3.2.3 provides a lower bound on the minimum number of oracle queries that are needed in order for a Lipschitz learning algorithm to ensure a precise estimate of the Lipschitz constant for all underlying target functions in $C^2(\mathcal{X}, K)$. As speculated in the introduction, it shows that the Lipschitz Learning problem is a computationally expensive one that depends heavily on the input dimension. The lower bounding expression is given as a function of the size of the input space (M), the assumed bound on the second order partial derivatives (K) and the precision parameter (ϵ) but is independent of the true Lipschitz constant (L^*) of the target function. The product MK can be understood as a bound on the maximum change in the gradient values of functions in $C^2(\mathcal{X}, K)$. In Section 3.3, the proposed LCLS algorithm will be shown capable of estimating the Lipschitz constant of all functions f in $C^2(\mathcal{X}, K)$ using $O((\frac{MK}{\epsilon})^d)$ queries to the noiseless sampling oracle Ω implying that the lower bound on the sample complexity rate stated in Theorem 3.2.3 is optimal.

An equivalent reformulation of Theorem 3.2.3 in the form of a lower bound on the convergence rate of Lipschitz learning algorithms is provided in the following corollary.

Corollary 3.2.4 *(Convergence Rate Bound – Noiseless) Assume the same setting as Theorem 3.2.3. Then,* $\forall L^* \geq 0$,

$$\inf_{\hat{L} \in \mathcal{L}_{n,p}(\mathcal{X})} \sup_{f \in C^2(\mathcal{X},K) \cap \mathcal{F}_p(L^*)} |\hat{L}(f) - L^*| \geq C(d,p) \frac{MK}{\sqrt[d]{n}}$$

where $C(d,p)$ *is defined in Theorem 3.2.3.*

Corollary 3.2.4 is generally more practical to use then Theorem 3.2.3 when considering convergence properties of applications of Lipschitz constant estimators. In Section 3.4, we show how Corollary 3.2.4 can be applied in conjunction with recent theoretical results (Bachoc et al. [2021]) to derive lower bounds on the sample complexity of adaptive Lipschitz optimisation algorithms.

3.2.3 Noisy Setting

In many instances, such as the general regression setting described in (2.1), the sampling oracle cannot be assumed reliable and only approximate observations of the target function are obtainable. In this case, we model $\Omega : \mathcal{X} \to \mathbb{R}$ as being corrupted by additive noise and a new lower bound on the sample complexity of Lipschitz learning algorithms can be derived. In order to do so, additional assumptions must be made on the additive observational noise process.

Assumption 4 *(Noisy Oracle – Gaussian Noise) Let $\sigma^2 > 0$. We define a noisy sampling oracle as*

$$\begin{aligned} \tilde{\Omega} : \ & \mathcal{X} \to \mathbb{R} \\ & x \overset{\tilde{\Omega}}{\mapsto} \tilde{f}_x := f(x) + \gamma_x \end{aligned}$$

where γ_x are independent Gaussian random variables ($x \in \mathcal{X}$) with mean 0 and variance σ^2. Note: γ_x is an abuse of notation as the noise is not dependent on the input x. In other words: if $x \in \mathcal{X}$ is sampled twice, then $\gamma_x^1 \neq \gamma_x^2$.

As the class of Lipschitz learning has been loosely defined so far, with no parametric or functional assumptions, a slight reformulation of the Lipschitz learning problem is needed in order to derive lower bounds on the sample complexity. Consider the function class $C^2(\mathcal{X}, K)$ as defined above and $p \in \mathbb{N}$. It is known (e.g. see Lemma 2.1.1) that for all $f \in C^2(\mathcal{X}, K)$, $L_p^*(f) = \max_{\mathcal{X}}\{\|\nabla f(x)\|_q\}$ if $p = 1, 2$ and $L_p^*(f) \leq \max_{\mathcal{X}}\{\|\nabla f(x)\|_q\}$ otherwise, where q is the Hölder conjugate of p. Then, instead of directly considering the estimation error $|\hat{L}(f) - L_p^*|$ with a Lipschitz learning algorithm $\hat{L}(f) \in \mathcal{L}_{n,p}(\mathcal{X})$ as done in Theorem 3.2.3, one can consider the problem of obtaining $x^{\hat{L}(f)} \in \mathcal{X}$ such that $|\|\nabla f(x^{\hat{L}(f)})\|_q - L_p^*|$ is minimised. In this case, we say that the algorithm $\hat{L}(f)$ belongs to the class $\overline{\mathcal{L}}_{n,p}(\mathcal{X})$ of *Lipschitz Learning search algorithms* which is defined formally as follows.

Definition 3.2.5 *(Lipschitz Learning Search Algorithms) We define $\overline{\mathcal{L}}_{n,p}(\mathcal{X})$ as the set of all $\|.\|_p$-Lipschitz learning search algorithms that utilise at most $n \in \mathbb{N}$ queries*

to the Oracle $\tilde{\Omega}$ *with inputs in* $\mathcal{X}$ *in order to produce an estimate* $\hat{x}$ *that aims to minimise:* $Loss(\hat{x}, f) := |\|\nabla f(\hat{x})\|_q - L_p^*|$.

This type of paradigm is similar to the one considered in the literature on minimax global optimisation where one generally aims to obtain the minimising argument $\hat{x} \in \mathcal{X}$ of a target function rather than directly estimating the minimum (e.g. Bull [2011]). We stress that if a good estimate of $x^{\hat{L}(f)}$ can be obtained, then it is relatively straightforward to obtain an accurate Lipschitz constant estimate by computing a local gradient or slope estimate of the target function near $x^{\hat{L}(f)}$. In fact, the majority of Lipschitz algorithms either directly or implicitly operate by maximising gradient or slope estimates (e.g. Strongin [1973], Calliess et al. [2020], the LCLS algorithm proposed in this chapter) and could therefore be trivially modified to generate a search estimate: $x^{\hat{L}(f)}$. In particular, see Theorem 3.3.15).

Using Assumption 4, the following lower bound on the sample complexity rate can be derived in the noisy sampling setting.

Theorem 3.2.6 *(Sample Complexity Bound – Noisy) Let* $M \in \mathbb{R}_+$, $d \in \mathbb{N}$, $p \in \{1, 2\}$ *and suppose* $\mathcal{X} := [0, M]^d$. *Assume that Assumptions (3)-(4) hold, that one has access to a noisy oracle* $\tilde{\Omega} : \mathcal{X} \to \mathbb{R}$ *as specified in Assumption (4) and that the sample inputs are uniformly and independently sampled on* $\mathcal{X}$. *Define* $\overline{\mathcal{L}}_{n,p}(\mathcal{X})$ *as in Definition 3.2.5. If there exists* $\delta \in (0, 1)$ *such that*

$$\lim_{\epsilon \to 0^+} \inf_{\hat{L} \in \overline{\mathcal{L}}_{n,p}(\mathcal{X})} \sup_{f \in C^2(\mathcal{X}, K)} \mathbb{P}(Loss(x^{\hat{L}(f)}, f) > \epsilon) \leq \delta$$

then,

$$n \in \Omega\left(\frac{\sigma^2 M^d K^{d+2} \log(\frac{MK}{\epsilon})}{\epsilon^{d+4}}\right).$$

In contrast to the sample complexity bounds obtained in the noiseless sampling setting, the bounds proposed in Theorem 3.2.6 only hold asymptotically and can therefore not be utilised in order to obtain finite sample guarantees. Furthermore, the Lipschitz learning (search) algorithms considered in Theorem 3.2.6 are assumed

to be passive[4] (as the sampling inputs are sampled randomly) as opposed to the active algorithms considered in Theorem 3.2.3. Nonetheless, the obtained bounds provide insight into the necessary sampling requirements needed to ensure convergence for the Lipschitz learning (search) problem in the noisy sampling setting. In Section 3.3.3, we will show that the LCLS algorithm matches the convergence rate stated in Theorem 3.2.6 under the same assumptions implying that the rate $\Theta\left(\sigma^2 \frac{M^d K^{d+2} \log(\frac{MK}{\epsilon})}{\epsilon^{d+4}}\right)$ is optimal.

Finally, as in the noiseless sampling setting, an equivalent reformulation of Theorem 3.2.6 is provided in the form of a probabilistic lower bound on the convergence rate of Lipschitz learning (search) algorithms.

Corollary 3.2.7 *(Sample Complexity Bound – Noisy) Assume the same setting as Theorem 3.2.6. Then, for all $\delta \in (0,1)$, there exists $C > 0$ such that*

$$\lim_{n\to\infty} \inf_{\hat{L}\in\overline{\mathcal{L}}_{n,p}(\mathcal{X})} \sup_{f\in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}(f)}, f) > CMK\left(\frac{\log(MKn)}{nM^4K^2\sigma^2}\right)^{\frac{1}{d+4}}) > \delta.$$

Theorem 3.2.6 and Corollary 3.2.7 are particularly interesting in the context of system identification for control applications (e.g. Milanese and Novara [2004], Calliess et al. [2020]) where robustness properties depend explicitly on estimating a feasible Lipschitz constant from noisy data. These frameworks often ignore the modelling error arising from the Lipschitz constant estimation which is problematic when the goal is to provide worst-case guarantees. One source of usefulness for the two bounds stated in this subsection is therefore to provide a theoretical intuition of the worst-case estimation of Lipschitz constants in this context and therefore to make possible a more realistic robustness analysis of Lipschitz constant-based system identification methods in practice. A short illustrative comparison of the convergence rate given in Corollary 3.2.7 with the convergence of existing Lipschitz learning algorithms used for system identification purposes is given in Figure 3.3.

[4]Note: we are not aware of any existing active Lipschitz learning algorithms as Lipschitz constant estimation is usually computed as a secondary task to a main objective (e.g. optimisation, non-parametric regression).

Remark 3.2.8 *(Comparison to Non-parametric Estimation) The optimal convergence rates of non-parametric estimation in the noisy sampling setting with Gaussian noise are well-known (Tsybakov [2004]). In particular, the uniform*[5] *convergence rate of the non-parametric estimation of first-order partial derivatives on* $C^2(\mathcal{X}, K)$ *for some* $K > 0$ *is given by* $\Theta\left((\frac{log(n)}{n})^{\frac{1}{d+4}}\right)_{n\in\mathbb{N}}$ *which corresponds exactly to the lower bounds derived in Corollary 3.2.7. This implies that although Lipschitz learning seems more straightforward than partial derivative estimation, asymptotically their sample complexities are equivalent. Note that this observation is somewhat unsurprising as similar results hold in the context of global optimisation (Bull [2011], Wang et al. [2018]).*

3.3 Lipschitz Constant estimation by Least Squares regression (LCLS)

The theoretical results of Section 3.2 imply that a significant number of samples must be used in order to obtain a precise estimation of the best Lipschitz constant. As noted in the introduction, this is problematic computationally for classical Strongin-based Lipschitz learning algorithms due to the fact that the computational complexity of these methods can be shown to be quadratic in the number of samples. Using existing non-Strongin based methods could resolve this computational problem, however obtaining convergence guarantees would then be difficult as this class of methods is generally complicated to study from a theoretical perspective. Therefore, in the goal of obtaining a Lipschitz learning approach that can provide asymptotically consistency guarantees, has low computational complexity and for which we can derive upper bounds on the sample complexity (in the goal of comparing with the sample complexity lower bounds derived in Section 3.2), we define a new estimator: *Lipschitz Constant estimation by Least Squares regression* (LCLS).

[5] With respect to $\|.\|_\infty$.

3.3.1 Overview

The general intuition behind the Lipschitz learning algorithm proposed in this chapter follows from the simple observation that the coefficients from a least squares regression can be interpreted as a local approximation of the gradient and that the maximum q-norm of the gradient of f on $\mathcal{X}$ coincides[6] (for certain values of $p \in \mathbb{N}$) with the best Lipschitz constant associated to the p-norm, where q is the Hölder-conjugate of p, i.e. $\frac{1}{p} + \frac{1}{q} = 1$. Therefore, by using a partition $\mathcal{H}$ of the input space $\mathcal{X}$ that is sufficiently refined to properly capture the gradient variation of f and computing the maximum q-norm of the least squares coefficients associated to each subset of $\mathcal{H}$, a precise estimate of the Lipschitz constant is obtainable. Practically, in order to ensure that the refinement of the partition suffices[7], the proposed estimation framework is designed as an iterative method that utilises a sequence of increasingly fine convex partitions $(\mathcal{H}_I)_{I\in\mathbb{N}}$ that are given as input. A brief technical description of an iteration of the algorithm can be described as follows: *For a given iteration, indexed by $I \in \mathbb{N}$, a set of observations $D_I^H := \{(x_{H_i}, \tilde{f}_{H_i})\}_{i\in\{1,\dots,N_I^H\}}$ is generated by an oracle $\Omega : \mathcal{X} \to \mathbb{R}$ (defined in Section 3.2) for each subset H of the partition $\mathcal{H}_I$ and used individually to compute the coefficients $\{\beta_I^H\}_{H\in\mathcal{H}}$ of an ordinary least squares regression for each subset $H \in \mathcal{H}_I$. The Lipschitz constant estimate can then be directly computed: $\hat{L}_I := \max_{H\in\mathcal{H}_I}\{\|\beta_I^H\|_q\}$ where q is the Hölder-conjugate of p.*

Definition 3.3.1 *(Notation Overview) For a partition $\mathcal{H}_I$ of $\mathcal{X}$ and a set of samples $D_I := \{(x_i, \tilde{f}_i)\}_{i\in\{1,\dots,N_I\}}$ as described above, we utilise the following notation.*

1. *The subset of samples that belongs to $H \in \mathcal{H}_I$ is denoted $D_I^H := \{(x_{H_i}, \tilde{f}_{H_i})\}_{i\in\{1,\dots,N_I^H\}}$. Note: samples can only belong to one subset H. If a sample point is on the border between two sets, then it can be included in either design matrix.*
2. *We denote the design matrices of the least squares regressions $X_I^H \in \mathbb{R}^{N_I^H\times(d+1)}$*

[6] See Lemma 2.1.1 in Appendix for a formal statement.

[7] In the case where the upper bound K given in Assumption 3 is known beforehand it is possible to directly partition at the required refinement level (See Theorem 3.3.10 for example).

[7] The method described in this algorithm corresponds to the specific case where the (K, σ^2) variables are known.

Algorithm 1 General LCLS

Input: $\tilde{\Omega}$ (Oracle), $(\mathcal{H}_I)_{I\in\mathbb{N}}$ (Partition Sequence)
Output: $\{\hat{L}_I\}$ (Lipschitz Estimates)
procedure: LCLS($\tilde{\Omega}$, $(\mathcal{H}_I)_{I\in\mathbb{N}}$)
 initialise: $I \leftarrow 1$
 repeat
  $\hat{L}_I \leftarrow 0$
  for $H \in \mathcal{H}_\mathcal{I}$ **do**
   $(X_I^H, \tilde{f}_I^H) \leftarrow D_I^H$ generated by $\tilde{\Omega}$
   $\hat{\beta}_I^H \leftarrow (X_I^{H\top} X_I^H)^{-1} X_I^{H\top} \tilde{f}_I^H$
   $\hat{L}_I \leftarrow \max(\|\hat{\beta}_I^H\|_q, \hat{L}_I)$
  end
  $I \leftarrow I + 1$
return $\{\hat{L}_I\}_{I\in\mathbb{N}}$

Algorithm 2 Hypercube LCLS[8] on $[0, M]^d$

Input: $\tilde{\Omega}$ (Oracle), K (Bound from (3)), η (covering constant), σ^2 (noise variance), (ϵ, δ) (precision)
Output: $\hat{L}$ (Lipschitz Constant Estimation)
procedure: LCLS($\tilde{\Omega}$, K, η, σ^2, (ϵ, δ))
 initialise: $\hat{L} \leftarrow 0$ $I \leftarrow (C_1(d)\frac{MK}{\sqrt{\eta}\epsilon})$,
 $N_I \leftarrow (C_2(d, q)\frac{\sigma^2}{\eta\delta}\frac{I^{d+2}}{M^2\epsilon^2})$
 $\mathcal{H} \leftarrow$ hypercube partition of $[0, M]^d$ with side-length $\frac{M}{I}$
 for $H \in \mathcal{H}$ **do**
  $(X^H, \tilde{f}^H) \leftarrow D^H$ generated by $\tilde{\Omega}$
  $\hat{\beta}^H \leftarrow (X^{H\top} X^H)^{-1} X^{H\top} \tilde{f}^H$
  $\hat{L} \leftarrow \max(\|\hat{\beta}^H\|_q, \hat{L})$
 end
 return $\hat{L}$

Figure 3.1: *Algorithm 1 details the implementation of the LCLS algorithm for a general input space and partition choice. Algorithm 2 details the theoretical implementation given in Theorem 3.3.10 of the LCLS algorithm when the input space is a hypercube $[0, M]^d$ and the partitions are regular. In practice, $I \in \mathbb{N}$ and $N_I \in \mathbb{R}_+$ can be set heuristically in order to improve convergence. We note that the generated data points X_I^H used by the two algorithms are selected arbitrarily in each $H \in \mathcal{H}_I$. In order to ensure convergence of the LCLS algorithm, X_I^H will need to verify Assumption 5 for all $I \in \mathbb{N}$ and $H \in \mathcal{H}_I$.*

and the observation vectors $\tilde{f}_I^H \in \mathbb{R}^{N_I^H}$;

$$X_I^H = \begin{bmatrix} 1 & x_{H_1}^\top \\ 1 & x_{H_2}^\top \\ \vdots & \vdots \\ 1 & x_{H_{N_I^H}}^\top \end{bmatrix}, \quad \tilde{f}_I^H = \begin{bmatrix} \tilde{f}_{H_1} \\ \tilde{f}_{H_2} \\ \vdots \\ \tilde{f}_{H_{N_I^H}} \end{bmatrix} = \underbrace{\begin{bmatrix} f_{H_1} \\ f_{H_2} \\ \vdots \\ f_{H_{N_I^H}} \end{bmatrix}}_{f_I^H :=} + \underbrace{\begin{bmatrix} \gamma_{H_1} \\ \gamma_{H_2} \\ \vdots \\ \gamma_{H_{N_I^H}} \end{bmatrix}}_{\gamma_I^H :=}, \quad \text{where } \forall k \in$$

$\{1, ..., N_I^H\}$ and where $(x_{H_k}, \tilde{f}_{H_k})$ is a sample point contained in D_I^H with $\tilde{f}_{H_k} := \tilde{\Omega}(x_{H_k})$ and by abuse of notation; $\gamma_{H_k} = \gamma_{x_{H_k}}$.

3. *We denote by $[\hat{b}_I^H, \hat{\beta}_I^H] \in \mathbb{R}^{d+1}$ (where $\hat{b}_I^H \in \mathbb{R}$ is the intercept) the least squares coefficients associated to $H \in \mathcal{H}_I$ and computed using X_I^H and $\tilde{f}_I^H$.*

The LCLS algorithm is described here in its most general form in order to allow flexibility in the choice of the input space partitions and sampling scheme. Algorithm 1 provides an algorithmic description of this approach. A more specific implementa-

tion of the LCLS algorithm which utilises a regular hypercube partition of the input space is given in Algorithm 2 and discussed later on in this section in Theorem 3.3.10 and ensuing discussions. As one might expect, the structure of $(\mathcal{H}_I)_{I\in\mathbb{N}}$ is a key part of the LCLS estimator. In practice, these partitions can be defined using domain or functional knowledge in order to better estimate the gradient variation and therefore speed up the convergence of the algorithm. The distribution of the sample points given by $(N_I^H)_{n\in\mathbb{N}}$ should also be considered carefully and can be selected in a partition dependent way to take advantage of any prior knowledge of f or of the underlying noise distribution. We note that the relation between the structure of $(\mathcal{H}_I)_{I\in\mathbb{N}}$ and $(N_I^H)_{n\in\mathbb{N}}$ is essential in the proofs of Theorem 3.3.7 and Theorem 3.3.10.

The following variables are used to formally describe a partition belonging to $(\mathcal{H}_I)_{I\in\mathbb{N}}$.

Notation 3.3.2 *Let $\delta(A) = \sup_{x,y\in A} \|x-y\|_2$ denote the diameter function and $B_r(x)$ the d-dimensional ball centered in $x \in \mathcal{X}$ and with radius r with respect to $\|.\|_2$.*

Definition 3.3.3 *(Partition Variables) Let $\mathcal{H}_J \in (\mathcal{H}_I)_{I\in\mathbb{N}}$. We define the following two $\mathcal{H}_J$ related quantities: the maximum diameters of $\mathcal{H}_J$: $\{\Delta_J^H\}_{H\in H_J}$, $\Delta_J^H := \delta(H)$ and the minimum diameters of the biggest subset-inscribed balls of $\mathcal{H}_J$: $\{\delta_J^H\}_{H\in H_J}$, $\delta_J^H := 2\max\{r \in \mathbb{R}_+ | \exists x \in H \text{ such that } B_r(x) \subset H\}$.*

The quantities $\{\Delta_J^H\}_{H\in H_J}$ and $\{\delta_J^H\}_{H\in H_J}$ are used in Definition 3.3.5 and in Theorem 3.3.7 to define sufficient conditions on the structure of the $(\mathcal{H}_I)_{I\in\mathbb{N}}$ partitions in order for the general version of the LCLS algorithm to converge.

We conclude this subsection by giving a result on the computational complexity of the proposed algorithm.

Proposition 3.3.4 *(Computational Complexity of LCLS) The computational complexity of the Lipschitz Constant Least Squares Estimator is $O(d^2 n_{samples})$ where $n_{samples}$ denotes the number of observations sampled by the algorithm and the $d \in \mathbb{N}$ is the input dimension of the target function.*

The computational complexity derived in Proposition 3.3.4 is significantly smaller than the complexity of Strongin-based approaches which is $O(dn^2_{samples})$. The difference in computation speed is illustrated empirically on a set of test functions in Section 3.3.4.

3.3.2 General Theoretical Analysis

An investigation of the theoretical behaviour and performance of the proposed LCLS algorithm is carried out in this section. This analysis provides an understanding of the design constraints required for the construction of the input space partitions and for the choice of sampling schemes in order to ensure satisfactory performance – see Remark 3.3.8. We begin by stating an asymptotic convergence result for the general form of the algorithm in the noiseless and general noisy sampling settings before stating and discussing finite sample results for a more concrete application of LCLS when the partition of the input space is constructed to be a set of regular hypercubes.

The following definition defines two quantities $(a_I)_{I\in\mathbb{N}}$, $(b_I)_{I\in\mathbb{N}}$ as a function of $\{\Delta_J^H\}_{H\in H_J}$, $\{\delta_J^H\}_{H\in H_J}$, $(\{N_I^H\}_{H\in\mathcal{H}_\mathcal{I}})_{I\in\mathbb{N}}$ and $(|\mathcal{H}_I|)_{I\in\mathbb{N}}$ in order to alleviate notation[9]. They will be used to describe the conditions on the structure of the input partition sequence needed in order to ensure asymptotic consistency.

Definition 3.3.5 *For any sequence of convex and compact partitions, $(\mathcal{H}_I)_{I\in\mathbb{N}}$, we construct the following sequences:*

- $(a_I)_{I\in\mathbb{N}}$, $a_I = \max_{H\in\mathcal{H}_I}\left(\frac{(\Delta_I^H)^2}{\delta_I^H}\right)$
- $(b_I)_{I\in\mathbb{N}}$, $b_I = \max_{H\in\mathcal{H}_I}\left(\frac{|\mathcal{H}_I|}{N_I^H(\delta_I^H)^2}\right)$.

Before stating the first main result of this section, a condition on the sampling procedure used by the LCLS algorithm and a generalisation of the Oracle noise assumption of Section 3.2 are given.

[9]Here. $|.|$ denotes the cardinality operator.

Definition 3.3.6 *Let $H \subseteq \mathcal{X}$ be compact and convex and denote by $D_I^H := \{(x_{H_i}, \hat{f}_{H_i})\}_{i\in\{1,\dots,N_I^H\}}$ the subset of generated or archived data samples in H. We say that H is (ϵ, η)-covered for $\epsilon > 0, \eta \in]0, 1]$ if there exists an ϵ-cover of H (with respect to $\|.\|_2$) with at least ηN_I^H samples of D_I^H in each of the balls associated to an element in the ϵ-cover.*

Assumption 5 *(Sampling) For a given $\eta \in (0, 1]$, the sampling scheme selected for LCLS is such that $\forall I \in \mathbb{N}$ and $\forall H \in \mathcal{H}_I$, H is $(\frac{\delta_I^H}{8}, \eta)$-covered.*

The sampling condition stated in Assumption 5 is used in order to ensure stability of the least squares coefficient as the sequence of partitions becomes increasingly refined and can be satisfied by using quasi-Monte Carlo schemes in practice. In essence, it ensures that the samples are sufficiently well distributed in each subset H of the partition $\mathcal{H}_I$, avoiding extreme cases such as when a single sample input gets repeatedly sampled. A possible (conservative) value for η in the regular hypercube implementation of the LCLS algorithm of Section 3.3.3 is given by $\eta = \frac{vol(B_1(0))}{2^{3d+2}}$ with $vol(B_1(0))$ denoting the volume of the d-dimensional unit ball. In the proof of Theorem 3.3.15, under a uniform and independent sampling (on $\mathcal{X}$) assumption, Assumption 5 is shown to hold for $\eta = \frac{vol(B_1(0))}{2^{3d+1}3^d}$ with high probability when the number of observations is sufficiently large.

The second assumption of this section generalises the Gaussian noise assumption made in Assumption 4 to include all distributions with zero mean and finite variance and weaken the assumption of independence.

Assumption 6 *(Noisy Oracle – General) Let $\sigma^2 > 0$ and denote by $\mathcal{D}(0, \sigma^2)$ the set of all probability distributions on $\mathbb{R}$ with zero mean and finite variance $\sigma^2 > 0$. We define a general noisy sampling oracle as*

$$\begin{aligned} \tilde{\Omega} : \ & \mathcal{X} \to \mathbb{R} \\ & x \overset{\tilde{\Omega}}{\mapsto} \tilde{f}_x := f(x) + \gamma_x \end{aligned}$$

where $\gamma_x \sim D_{\gamma_x} \in \mathcal{D}(0, \sigma^2)$ are uncorrelated random variables ($x \in \mathcal{X}$).

The general nature of Assumption 6 ensures that the convergence results obtained for the proposed LCLS algorithm hold for a wide range of noise distributions and therefore removes the necessity of having to verify a sub-Gaussian noise assumption when applying LCLS in practice.

Theorem 3.3.7 formalizes the consistency of the proposed Lipschitz learning framework for the general noisy sampling setting.

Theorem 3.3.7 *(General Convergence Rate) If Assumptions (2),(3),(5),(6) (for a given $\eta \in (0, 1]$) hold and the following conditions are verified:*

1. $\forall I \in \mathbb{N}$, $\mathcal{H}_I$ *is a convex partition of* $\mathcal{X}$,
2. $\lim_{I\to\infty} a_I = 0$, $\lim_{I\to\infty} b_I = 0$, $\lim_{I\to\infty} \max_{H\in\mathcal{H}_I}(\Delta_I^H) = 0$,

then $\forall\ D_\gamma^n \in \mathcal{D}(0, \sigma^2)$, $f \in C^2(\mathcal{X}, K)$,

$$\hat{L}_I(f) \xrightarrow[I\to\infty]{\mathbb{P}} L_p(f)$$

where $L_p(f) = L_p^*(f)$ for $p = 1, 2$, $L_p \geq L_p^*(f)$ for $p > 2$. $\mathbb{P}$ denotes convergence in probability and $(\hat{L}_I(f))_{I\in\mathbb{N}}$ is the sequence of Lipschitz constant estimates generated by the LCLS estimator.

Remark 3.3.8 *(Design Constraints) Condition 2 of Theorem 3.3.7 specifies the design constraints needed in the construction of the partition sequence $(\mathcal{H}_I)_{I\in\mathbb{N}}$ and the number of sample points $(\{N_I^H\}_{H\in\mathcal{H}_\mathcal{I}})_{I\in\mathbb{N}}$ required per hypercube in order to ensure convergence. In particular:*

1. $\lim_{I\to\infty} a_I = 0$ *provides the limitations on the shape of the sets in each partition $\mathcal{H}_I$ as I goes to infinity. In particular, as $I \to \infty$, $(\Delta_I^H)^2 << \delta_I^H < \Delta_I^H$.*
2. *In the noisy sampling setting, $\lim_{I\to\infty} b_I = 0$ specifies a condition on the number of samples needed per hypercube. As $I \to \infty$, $N_I^H >> \frac{|\mathcal{H}_I|}{(\delta_I^H)^2}$. This is made precise in the next section in Remark 3.3.11.*
3. $\lim_{I\to\infty} \max_{H\in\mathcal{H}_I}(\Delta_I^H) = 0$ *ensures that the partitions are increasingly refined.*

In practice, applying the theoretical conditions used in Theorem 3.3.7 produces an overly conservative estimator in terms of required number of queries made to the

oracle – see Section 3.3.4 for an illustration of the empirical convergence of the LCLS estimator. This is due to the fact that the LCLS estimator makes minimal functional assumptions and therefore has to explore all of $\mathcal{X}$ to generate a precise Lipschitz estimate. In order to avoid this issue, the number of samples per hypercube as measured by $(b_I)_{I\in\mathbb{N}}$ can be set heuristically in order to improve the empirical performance.

In the noiseless sampling setting, the stopping and sampling rules given in Theorem 3.3.7 and Remark 3.3.8 can be modified in order to obtain a quicker convergence. This is detailed in the following corollary.

Corollary 3.3.9 *(Noiseless Sampling) If Assumptions (2),(3),(5) (for a given $\eta \in (0,1]$) hold, a noiseless oracle $\Omega : \mathcal{X} \to \mathbb{R}$ is available and the following conditions are verified:*

1. $\forall I \in \mathbb{N}$, $\mathcal{H}_I$ *is a convex partition of* $\mathcal{X}$,
2. $\lim_{I\to\infty} a_I = 0$, $\lim_{I\to\infty} \max_{H\in\mathcal{H}_I}(\Delta_I^H) = 0$,
3. $\forall I \in \mathbb{N}, H \in \mathcal{H}_I$, $N_I^H \geq d+1$,

then $f \in C^2(\mathcal{X}, K)$,

$$\hat{L}_I(f) \xrightarrow[I\to\infty]{} L_p(f)$$

where $L_p(f) = L_p^*(f)$ for $p = 1, 2$, $L_p(f) \geq L_p^*(f)$ for $p > 2$ and the right arrow denotes deterministic convergence.

While, the conditions on the design constraints of the partition sequence needed to ensure asymptotic convergence of the LCLS algorithm remain the same as in Theorem 3.3.7, the sampling conditions specified in Corollary 3.3.9 imply that a much smaller number of samples are required per hypercube. More precisely, the only sampling condition stated in Corollary 3.3.9 is related to the minimum number of samples needed to ensure that the local linear regressions computed by the LCLS algorithm are well-defined.

Using the general results developed in this section, we now explore a more specific application to the $[0, M]^d$ input space. Theorem 3.3.10 provides finite-sample sam-

ple complexity bounds for the LCLS in the general noise sampling setting that can be utilised when limited information on the noise distribution is available. As a consequence of Theorem 3.3.10, sample complexity rates in the noiseless and Gaussian noise setting can also be derived and compared to the lower bounds proposed in Theorem 3.2.3 and Theorem 3.2.6. This is discussed in Remark 3.3.11, Remark 3.3.12 and Theorem 3.3.15.

3.3.3 LCLS with Regular Partitions & Sample Complexity Upper Bound

In the previous section, we considered a general form of the LCLS algorithm and stated the conditions on the design constraints of the input partition sequence and the sampling scheme required to ensure convergence. Here, we assume that the input space is the d-dimensional hypercube $[0, M]^d$ and consider the case where every input partition $\mathcal{H}_I$ is a regular hypercube partition of side-length $\frac{M}{I}$. The associated sampling scheme is then defined based on the sampling condition given in Assumption 5 and the desired precision of the Lipschitz constant estimate.

Under these additional constraints, the following finite sample guarantee can be obtained for the LCLS algorithm.

Theorem 3.3.10 *(Finite Sample Guarantee) Let $\mathcal{X} := [0, M]^d$ and $(\mathcal{H}_I)_{I\in\mathbb{N}_{>1}}$ denote the regular partition of sub-hypercubes of $\mathcal{X}$ with side-length $\frac{M}{I}$. If Assumptions (3)-(5) (for a given $\eta \in (0, 1]$) hold and if $\forall\epsilon > 0$, $\delta \in (0, \frac{1}{2}]$, the LCLS algorithm is set with a hypercube partition indexed by $I \geq \left(C_1(d)\frac{MK}{\sqrt{\eta}\epsilon}\right)$ and with $\forall H \in \mathcal{H}_I$ $N_I^H \geq \left(C_2(d, q)\frac{\sigma^2}{\eta\delta}\frac{I^{d+2}}{M^2\epsilon^2}\right)$ for $C_1(d), C_2(d, q) \in \mathbb{R}_+$, then $\forall\ D_\gamma^n \in \mathcal{D}(0, \sigma^2)$:*

$$\inf_{f\in C^2(\mathcal{X},K)} \mathbb{P}(|L_p(f) - \hat{L}_I(f)| \leq \epsilon) \geq 1 - \delta. \tag{3.1}$$

where $L_p(f) = L_p^*(f)$ for $p = 1, 2$ and $L_p \geq L_p^*$ for $p > 2$. Here $C_1(d) = 8d^2\sqrt{d}d^{\max\{\frac{1}{q}-\frac{1}{2},0\}}$ and $C_2(d, q) = 2^5 d^{\max\{\frac{2}{q},1\}}$ however these constants have not been optimized.

The theoretical guarantees of Theorem 3.3.10 can be extended to include any $\mathcal{X} \subset \mathbb{R}^d$ that satisfies Assumption 2. Indeed, trivially there exists a hypercube $[a,b]^d \subset \mathbb{R}^d$ with $a,b \in \mathbb{R}$ such that $\mathcal{X} \subset [a,b]^d$ which can be partitioned according to the iterative regular hypercube partitioning approach. The partition sequence inputted into the LCLS algorithm then consists of the regular hypercube subsets partitions of $[a,b]^d$ that intersect with $\mathcal{X}$. In this case, under Assumptions (2)-(5), a modified version of Theorem 3.3.10 holds: the condition on I remains the same, but the lower bound condition on N_I^H can be weakened to become $N_I^H \geq \left(C_2(d,q)\frac{\sigma^2}{\eta\delta}\frac{I^{d+2}-\Gamma}{(b-a)^2\epsilon^2}\right)$, $\forall H \in \mathcal{H}_I, I \in \mathbb{N}$, where $\Gamma = |\{H \in \mathcal{H}_I \big| H \cap \mathcal{X} = \emptyset\}|$.

Since Theorem 3.3.10 holds under Assumption 6, i.e. for any $D_\gamma^n \in \mathcal{D}(0,\sigma^2)$, $n \in \mathbb{N}$ and any $f \in C^2([0,M]^d, K)$ it also holds for $\sup_{D_\gamma^n \in \mathcal{D}(0,\sigma^2)} \sup_{f \in C^2([0,M]^d,K)}$. Therefore, using Theorem 3.3.7, we can obtain the following general sample complexity rate for the LCLS algorithm.

Remark 3.3.11 *(Sample Complexity – Noisy) For $p = 1, 2$, assuming that the lower bounds: $I \geq \left(C_1(d)\frac{MK}{\sqrt{\eta}\epsilon}\right)$ and $\forall H \in \mathcal{H}_I$ $N_I^H \geq \left(C_2(d,q)\frac{\sigma^2}{\eta\delta}\frac{I^{d+2}}{M^2\epsilon^2}\right)$ are satisfied with an equality, the total number n_1 of points required to ensure $\mathbb{P}(|L_p - \hat{L}_I| \leq \epsilon) \geq 1-\delta$ is given by*

$$n_1 = |\mathcal{H}_I| N_I^H = C_2(d,q)\frac{\sigma^2}{\eta\delta}\left(\frac{C_1(d)}{\sqrt{\eta}}\frac{MK}{\epsilon}\right)^{2d+2}\frac{1}{M^2\epsilon^2} = O\left(\left(\frac{MK}{\epsilon}\right)^{2d+2}\frac{1}{M^2\epsilon^2}\right).$$

This sample complexity differs significantly from the lower bound on the sample complexity derived in Theorem 3.2.6. This is expected given the more general noise assumptions made in Theorem 3.3.10.

By slightly modifying the necessary conditions used in Theorem 3.3.10, we can also compare the sample complexity of the LCLS algorithm implied by Theorem 3.3.10 in the noiseless sampling setting to the lower bound on the sample complexity of the noiseless Lipschitz learning problem stated in Theorem 3.2.3. In order to do so, we define

$$N(I) := \max_{H \in \mathcal{H}_\mathcal{I}} \min\{|D_I^H| : D_I^H \text{ contains a disjointed } \delta_I^H\text{-cover of } H\}$$

which is constant $\forall I \in \mathbb{N}$ when $(\mathcal{H}_I)_{I\in\mathbb{N}}$ is defined as a sequence of regular hypercube partitions on $[0, M]^d$. In this case, we remove the dependence on I and write $N := N(I)$. We note that the following two inequalities hold: (1) $\eta \leq \frac{1}{N}$ (tight) and (2) $N < \sqrt{d}^d$ (loose).

Remark 3.3.12 *(Sample Complexity – Noiseless) In the case of noiseless sampling, the lower bound on N_I^H stated in Theorem 3.3.10 can be replaced by condition 3. of Corollary 3.3.9 and the definition of N given above, i.e. $\forall I \in \mathbb{N}, H \in \mathcal{H}_I$, $N_I^H = \max(d+1, N)$. Proceeding as in Remark 3.3.11, we have in this case:*

$$n_2 = |\mathcal{H}_I| N_I^H = \max(d+1, N)\big(C_1(d)\frac{MK}{\sqrt{\eta}\epsilon}\big)^d = O\left(\big(\frac{MK}{\epsilon}\big)^d\right).$$

This convergence rate corresponds exactly to the lower bound on the noiseless sample complexity rate stated in the Theorem 3.2.3 and therefore implies that the sample complexity rate $\big(\frac{MK}{\epsilon}\big)^d$ is optimal (up to constant factors dependent on d and p) in the sense that it characterises the minimum number of samples that are needed to obtain an ϵ-precise Lipschitz constant estimate for any $f \in C^2(\mathcal{X}, K)$.

As in Section 3.2, we can reformulate the sample complexity rates of the LCLS algorithm given in Remarks 3.3.11 and 3.3.12 as convergence rates and therefore as upper bounds on the convergence rate of the general Lipschitz learning problem. This is done in the following corollary.

Corollary 3.3.13 *(Convergence Rate Comparison)*

1. *(Noiseless) Assume the same setting as Remark 3.3.12. Then,*

$$\inf_{\hat{L}\in\mathcal{L}_{n,p}(\mathcal{X})} \sup_{f\in C^2(\mathcal{X},K)} |\hat{L}(f) - L_p^*(f)| \leq C(d,\eta)\frac{MK}{\sqrt[d]{n}}$$

   where $C(d,\eta) := \max(d+1, N)\big(\frac{C_1(d)}{\sqrt{\eta}}\big)^d$ and $C_1(d)$ is given in Theorem 3.3.10.

2. *(Noisy) Assume the same setting as Remark 3.3.11. Then, $\forall$ distribution*

$D_\gamma^n \in \mathcal{D}(0, \sigma^2)$:

$$\sup_{\hat{L} \in \mathcal{L}_{n,p}(\mathcal{X})} \inf_{f \in C^2(\mathcal{X}, K)} \mathbb{P}_{D_\gamma^n}(|\hat{L}(f) - L_p^*(f)| < C(\sigma^2, \delta, d, \eta) \frac{M^{\frac{d}{d+2}} K^{\frac{d+1}{d+2}}}{\sqrt[2d+4]{n}}) \geq 1 - \delta$$

where $C(\sigma^2, \delta, d, \eta) = C_2(d, q) \frac{\sigma^2}{\eta\delta} \left(\frac{C_1(d)}{\sqrt{\eta}} \right)^{2d+2}$ *and* $C_1(d)$, $C_2(d, q)$ *are given in Theorem 3.3.10.*

An interesting consequence of Corollary 3.3.13 is that it provides a way of generating a sequence of feasible[10] Lipschitz constant estimates that converge to the best Lipschitz constant if a potentially loose upper bound on the second degree partial derivatives is known. More precisely, one can consider the Lipschitz constant estimates:

- $\hat{L}_{up}(f) := \hat{L}(f) + C(d, p) \frac{MK}{\sqrt[d]{n}}$ in the noiseless sampling setting
- $\hat{L}_{up}(f) := \hat{L}(f) + C(\sigma^2, \delta, d) \frac{M^{\frac{d}{d+2}} K^{\frac{d+1}{d+2}}}{\sqrt[2d+4]{n}}$ in the general noisy sampling setting

where $\hat{L}(f)$ denotes the Lipschitz constant estimate generated by the LCLS algorithm. Such an approach is useful in practice as Lipschitz constant-based computational frameworks often rely on the assumption that the estimated Lipschitz constant used is feasible. This is briefly discussed further in Section 3.4 where a direct application of the LCLS algorithm in the context of non-parametric regression for system identification is developed.

Remark 3.3.14 *(Knowledge of K and Assumption 3) The theoretical results of the LCLS algorithm of this section have been stated under Assumption 3 and the knowledge of a tight upper bound on the second-order partial derivatives: K. This tightness is in fact not necessary and all result pertaining to LCLS hold for any upper bound $K' \geq K$. In this case, the LCLS algorithm simply ensures convergence for a larger class of functions, $C^2(\mathcal{X}, K) \subset C^2(\mathcal{X}, K')$, then required at a slightly slower rate of convergence. Furthermore, while knowing an upper bound on K is necessary in order for the theoretical properties of the LCLS algorithm to hold, the*

[10] I.e. which upper bound the best Lipschitz constant and satisfy the Lipschitz continuity condition.

algorithm can still be implemented heuristically in practice without it.

We conclude this section by stating the asymptotic sample complexity rates of the LCLS algorithm under Gaussian noise assumptions and providing a finite sample guarantee.

Theorem 3.3.15 *(Asymptotic Sample Complexity – Gaussian Noise) Let $M \in \mathbb{R}_+$, $d \in \mathbb{N}$, $p \in \{1, 2\}$ and $(\mathcal{H}_I)_{I \in \mathbb{N}}$ denote the regular partition of sub-hypercubes of $\mathcal{X}$ with side-length $\frac{M}{I}$. Assume that Assumption (3) holds, that one has access to a noisy oracle $\tilde{\Omega} : \mathcal{X} \to \mathbb{R}$ as specified in Assumption (4) and that the sample inputs are uniformly and independently sampled on $\mathcal{X}$. Setting the LCLS algorithm with a hypercube partition indexed by $I = \lceil C_1(d)\frac{MK}{\epsilon} \rceil$ for $\epsilon > 0$ (see below for definition of $C_1(d)$), there exists $C > 0$ such that if*

$$n \geq C \frac{\sigma^2 M^d K^{d+2} \log(\frac{MK}{\epsilon})}{\epsilon^{d+4}}$$

Then,

$$\lim_{\epsilon \to 0^+} \sup_{f \in C^2(\mathcal{X}, K)} \mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon) = 0,$$

where $x^{\hat{L}_I(f)}$ denotes the the center of the hypercube associated to $argmax_{H \in \mathcal{H}_I} \|\hat{\beta}^H\|_q$ computed in Algorithm 2. Here, $C_1(d) = \frac{16 d^2 \sqrt{d} d^{\max\{\frac{1}{q} - \frac{1}{2}, 0\}}}{\sqrt{\eta}}$ with $\eta = \frac{vol(B_1(0))}{2^{3d+1} 3^d}$[11].

The asymptotic sample complexity rates derived in Theorem 3.3.15 match exactly the rates derived in Theorem 3.2.6. This implies that $\Theta\left(\frac{\sigma^2 M^d K^{d+2} \log(\frac{MK}{\epsilon})}{\epsilon^{d+4}}\right)$ is the optimal asymptotic sample complexity rate of the Lipschitz learning (search) problem and that the LCLS algorithm is sample optimal in the noisy setting when the noise is assumed to follow a Gaussian distribution. As done in Corollary 3.2.7, we can modify Theorem 3.3.15 in order to show that the optimal asymptotic convergence rate is $\Theta(MK\left(\frac{\log(MKn)}{nM^4K^2\sigma^2}\right)^{\frac{1}{d+4}})$. We note that Theorem 3.3.15 holds more generally for any sub-Gaussian noise assumption on the sampling noise. In particular, the same convergence rate holds in the settings where this noise is assumed to be bounded which are often considered in Lipschitz-constant based applications (Canale et al.

[11] $vol(B_1(0))$ denotes the d-dimensional unit ball.

[2014], Sergeyev et al. [2020]).

Replacing the random uniform sampling assumption of Theorem 3.3.15 with Assumption 5 as done in Theorem 3.3.10, a small modification of the proof of Theorem 3.3.15 yields the following result on the finite-sample guarantees of the LCLS algorithm in the noisy sampling setting with sub-Gaussian noise.

Corollary 3.3.16 *(Finite Sample Guarantee – Gaussian Noise) Consider the setting of Theorem 3.3.15. Assume that Assumptions (3), (5) (for a given $\eta \in (0,1]$) hold, that one has access to a noisy oracle $\tilde{\Omega} : \mathcal{X} \to \mathbb{R}$ as specified in Assumption (4). $\forall \epsilon \in (0, \frac{C_1(d)MK}{3})$, $\delta \in (0, \frac{1}{2})$, setting $I = \lceil C_1(d)\frac{MK}{\epsilon} \rceil$ and $\forall H \in \mathcal{H}_I$: $N_I^H \geq C^*(\eta, d)\frac{\sigma^2 K^2}{\epsilon^4} \log(\frac{2^{\frac{2}{d}} I}{\log(\frac{1}{1-\delta})^{\frac{1}{d}}})$ implies*

$$\sup_{f \in C^2(\mathcal{X}, K)} \mathbb{P}(|\hat{L}_I(f) - L_p^*(f)| > \epsilon) \leq \delta.$$

Here, C_1 is defined as in Theorem 3.3.15 and $\tilde{C}^*(\eta, d) := \frac{2^{10} n_q^2 C_1(d)^2 d^2}{\eta}$ however these constants have not been optimized.

As done for the general noise setting in Corollary 3.3.13, convergence rates for the LCLS in the Gaussian noise setting can be obtained by reformulating the finite sample guarantees stated in Corollary 3.3.16. Then, following the approach described above, a sequence of feasible Lipschitz constant estimates converging to the best Lipschitz constant can be constructed:

- $\hat{L}_{up}^{Gauss} := \hat{L}(f) + C\frac{M^{\frac{d}{d+4}} K^{\frac{d+2}{d+4}}}{\sqrt[d+4]{n\sigma^2 \log(MKn)^{-1}}}$ in the Gaussian noise setting

where $\hat{L}(f)$ denotes the LCLS algorithm with the hyperparameters set in Corollary 3.3.16 and $C \in \mathbb{R}_+$ is a constant that can be computed from $C_1(d), \tilde{C}^*(d, \eta)$. We observe that this sequence of feasible Lipschitz constant estimates converges significantly faster than the one constructed above for the general noisy setting.

3.3.4 Empirical Performance

The focus so far in this section has been on developing the theoretical properties of the LCLS algorithm. While that discussion is useful in itself as it provides performance guarantees for LCLS as well as upper bounds on the sample complexity of the general Lipschitz learning problem, we are also interested in how the proposed algorithm performs empirically. In particular, we would like to compare the convergence speed of the LCLS algorithm to other theoretically well-behaved methods and to verify whether the theoretical computational advantage of LCLS (see Proposition 3.3.4) is observed in practice. In this subsection, we investigate these questions by illustrating the convergence rate and computation time of the proposed Lipschitz constant estimation method and comparing it against existing Strongin-based algorithms on a set of test functions with interesting properties in noiseless, bounded noise and unbounded noise sampling settings.

3.3.4.1 Experimental Setup

Table 3.1 provides an overview of the four test functions that are used in the experiments discussed in this section. The choice of these functions represents different testing points that are of interest: Function (a) reaches the maximum of the normed gradient in a single unique point of the input space, Function (b) is a classical optimisation testing function which we have also defined to have large second degree partial derivatives, Function (c) is a trigonometric function which provides an illustration of convergence for simple target functions and finally, Function (d) is a higher dimensional version of Function (a) with 3 dimensional inputs. We do not explore higher dimensional versions (>3) of Function (a) as the convergence speed with respect to computation time of the Strongin-based benchmark algorithms is already very slow for Function (d) - see Figure 3.4 and ensuing discussion.

As benchmarks we utilise the classical Strongin Lipschitz learning algorithm (Strongin [1973]) in the noiseless setting and the popular modified Strongin-based Lipschitz constant estimator in the bounded noise setting (see in particular Novara et al.

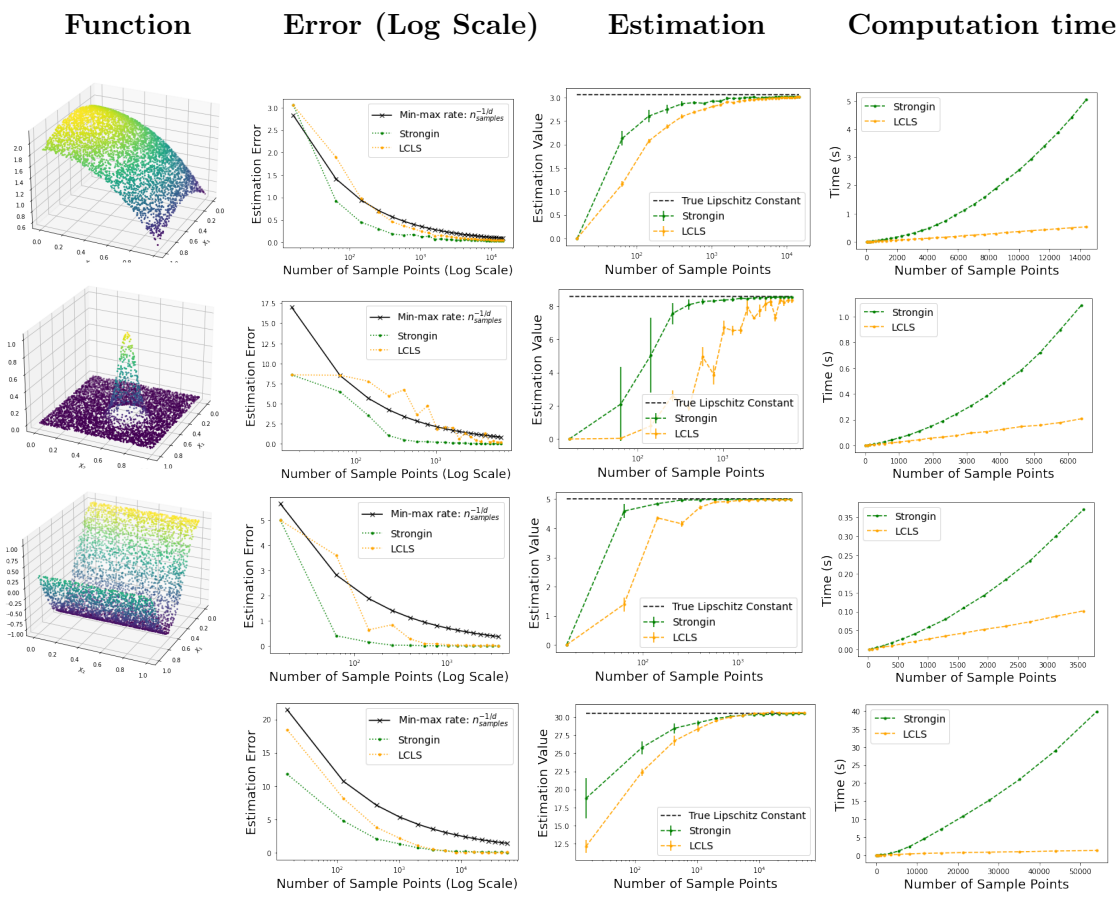


Figure 3.2: Comparison between the performance of the LCLS algorithm (in orange) and the classical Strongin algorithm (in green) in the noiseless setting. Each row corresponds to a different test function ((a) - (d)) and each column represents a different point of comparison between the two algorithms. From left to right: Column 1: Illustration of the target function where applicable. Column 2: Error of Lipschitz constant estimate - the bound on the sample complexity rate derived in Corollary 3.2.4 is plotted (in black). Column 3: Behaviour of the sequence of Lipschitz constant estimates. Column 4: Computation time required for each algorithm.

[2013], Calliess et al. [2020] and Khajenejad et al. [2021] for applications in control problems). We note that this modified Strongin estimator is strongly dependent on a precise estimate of the smallest upper bound of the noise $\bar{b} > 0$ in order to properly specify $\bar{e} \in \mathbb{R}_+$ hyper-parameter. Indeed, if $\bar{e}$ is smaller than $\bar{b}$, then the Lipschitz constant estimates generated by the modified Strongin estimator converge to $+\infty$ as the number of observations increases. In contrast, if $\bar{e}$ is bigger then $\bar{b}$ then the generated Lipschitz constant estimates will converge to an underestimate of $L_p^*(f)$ and never be feasible.

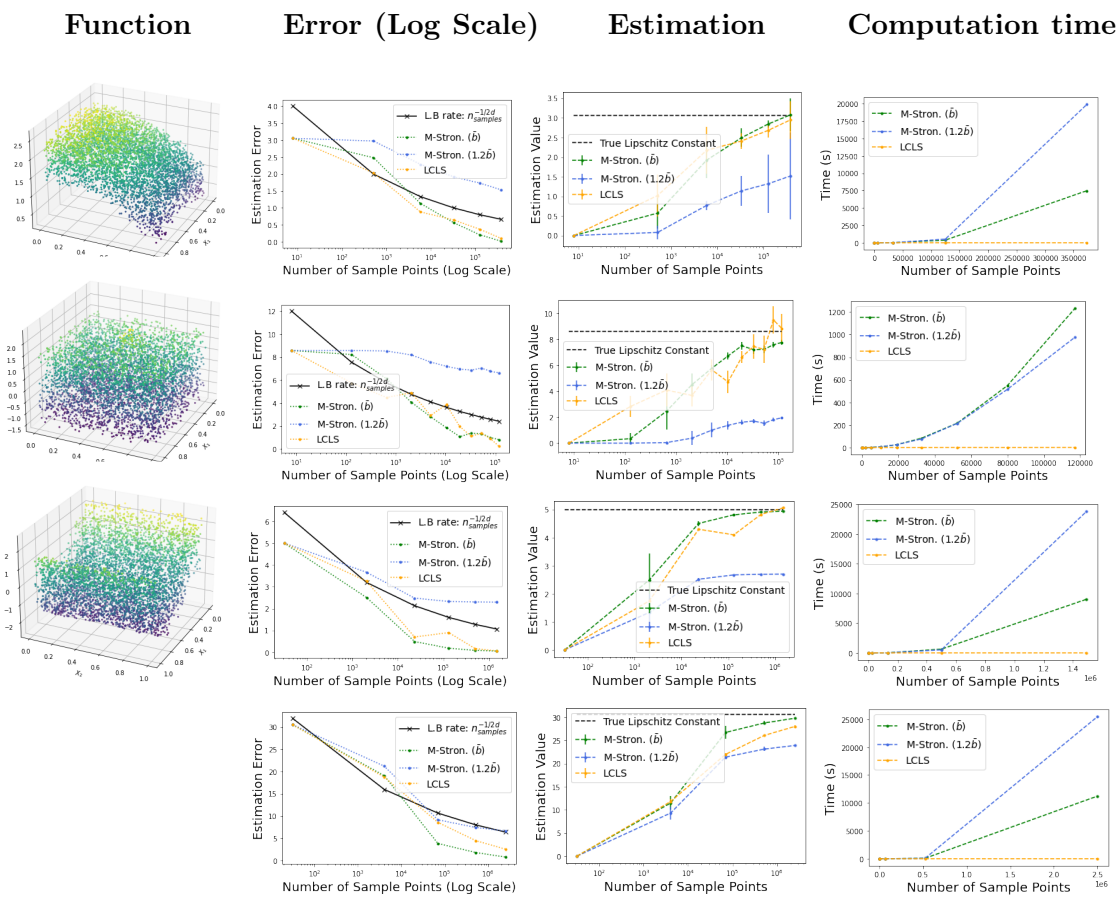


Figure 3.3: Comparison between the performance of the LCLS algorithm (in orange) and the modified-Strongin algorithm with a correctly (in green) and incorrectly (in blue) specified hyper-parameter in the noisy setting. Each row corresponds to a different test function ((a) - (d)) and each column represents a different point of comparison between the two algorithms. From left to right: Column 1: Illustration of the target function where applicable. Column 2: Error of Lipschitz constant estimate - the bound on the sample complexity rate derived in Corollary 3.2.4 is plotted (in black). Column 3: Behaviour of the sequence of Lipschitz constant estimates. Column 4: Computation time required for each algorithm.

Benchmarking algorithms:

- *(Noiseless Setting) Strongin Estimator:*

$$\hat{L} := \max_{i \neq j} \frac{|\tilde{f}_i - \tilde{f}_j|}{\|x_i - x_j\|}.$$

- *(Noisy Setting) Modified Strongin Estimator:*

$$\hat{L} := \max_{i \neq j} \frac{|\tilde{f}_i - \tilde{f}_j| - 2\bar{e}}{\|x_i - x_j\|}$$

where $\bar{e}$ is a hyper-parameter that estimates the tightest upper bound $\bar{b}$ on the noise. We consider modified Strongin Lipschitz estimators with a correctly specified hyper-parameter ($\bar{e} = \bar{b}$) and a hyper-parameter that is slightly larger than the true upper bound ($\bar{e} = 1.2\bar{b}$) as benchmarks.

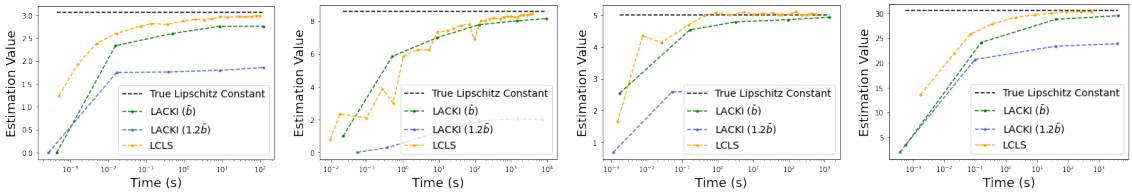
3.3.4.2 Discussion

In Figure 3.2, we illustrate the performance of the LCLS algorithm against the classical Strongin algorithm on the proposed set of test functions. We plot the theoretical lower bounds on the sample complexity rate found in Section 3.2 in order to provide an intuition for the theoretical bounds. As one would expect, due to the fact that the Strongin algorithm was specifically designed for the noiseless setting, our proposed approach converges more slowly in terms of number of samples on all four test functions. However the difference in convergence speed is not significant and is mitigated by the substantial divergence in computation time. We also note that the plotted sample complexity rate implied by the lower bound of Section 3.2 does not appear to be tight which is unsurprising as it represents a min-max type bound.

Remark 3.3.17 *(Link between the proof of Theorem 3.3.7 and convergence of LCLS) From the proof of Theorem 3.3.7, we have that the convergence of the LCLS algorithm depends on two factors:*

1. *($I \in \mathbb{N}$) the diameter of the subsets of the regular partition (upper bounded theoretically using a Taylor expansion).*
2. *(N_I, $I \in \mathbb{N}$) the number of samples in each subset (upper bounded theoretically using a multivariate Chebyshev inequality).*

The relation between these two factors is essential for ensuring quick convergence of the LCLS algorithm. In particular, for cases where the second derivatives of the target function f are large, N_I can be decreased and I increased so that the LCLS algrithm considers a finer partition of $\mathcal{X}$ (without having to increase the number of sample points). This type of modification improves the linear approximation of the



(a) **(b)** **(c)** **(d)**

Figure 3.4: Illustration of convergence speed relative to computation time in the bounded noise setting using the set of test functions given in Table 3.1. We compare the LCLS algorithm (in orange) and the modified-Strongin algorithm with a correctly (in green) and incorrectly (in blue) specified hyper-parameter. We observe that the LCLS algorithm performs better on all test functions.

gradient of f at the cost of increasing the noise in the local least squares estimates (see Function (b) - Easom function in Figure 3.3).

In Figure 3.3, we observe the performance of the LCLS algorithm in the bounded noise setting. Here, the convergence speed relative to sample size of the LCLS method differs more significantly from the convergence speed of the correctly specified modified Strongin benchmark algorithm. This is again unsurprising as the correctly specified modified Strongin algorithm makes use of additional information on the noise distribution and the choice of a uniform noise distribution in the experiment is beneficial towards its convergence speed[12]. We note that the modified Strongin algorithm with a slightly incorrectly specified tightest upper bound fails to show any sign of convergence and that the difference in computation time is more significant than in the noiseless setting. The relation between computational complexity and convergence rate of the LCLS and modified Strongin Lipschitz constant estimators is illustrated more precisely in Figure 3.4 by plotting the convergence rate relative to computation time. We observe that the LCLS estimator performs better on all functions in the test set despite the fact that the modified Strongin estimator utilises additional information on the noise distribution. In particular, for Function (d) which takes inputs in $\mathbb{R}^3$, the LCLS algorithm needs 8.5 seconds to generate estimates with an estimation error < 0.5, while the Strongin approach

[12]If a truncated Gaussian distribution had been used instead, the convergence speed of the modified Strongin estimator could have been arbitrarily slowed by decreasing the variance of the distribution.

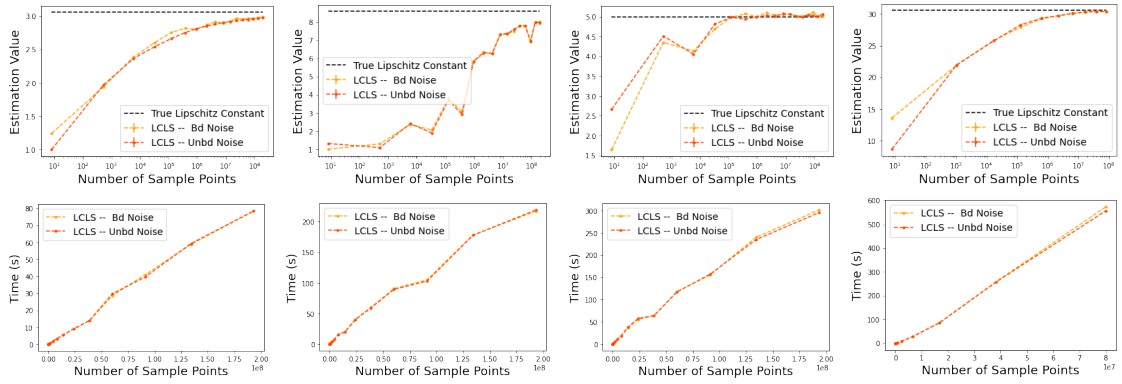


Figure 3.5: Illustration of the LCLS algorithm in bounded (in light orange) and unbounded noise settings (in dark orange) using the set of test functions given in Table 3.1. We observe no significant impact of the unboundedness of the noise distribution on the Lipschitz constant estimates produced by LCLS.

needs approximately 4000 seconds. This suggests that for application settings with high sampling capacity and time constraints, the LCLS method should be used even when the modified Strongin algorithm can be properly specified.

In Figures 3.2, 3.3 and 3.4, the performance of the LCLS method seems to be more dependent on the value of the maximum of the second degree partial derivatives than the Strongin-based methods. This can be observed by noting the difference in convergence performance for Function (b) relative to the other test functions[13] and is due to the fact that the LCLS algorithm depends on the maximum to define a sufficiently refined partition of the input space in order to "localise" the computations and generate local Lipschitz constant estimates, see Remark 3.3.8 and Theorem 3.3.10 for a precise characterisation of the relationship. In some sense, the stronger dependency on the maximum of the second degree partial derivatives of the target function can be interpreted as a trade-off for the improvement in computation time obtained by the LCLS algorithm.

The last illustration, provided in Figure 3.5, shows the convergence and computation time of the LCLS algorithm in the unbounded noise setting. We do not provide a benchmark as no alternative theoretically backed approaches exist in this setting: the approaches of Beliakov [2005] and Calliess [2017] could be used but do not have

[13] See also Figure 3.7 and ensuing discussion.

any asymptotic convergence guarantees. Instead, we compare the convergence rate to the one obtained by LCLS in the bounded noise setting and observe the fact that no significant performance loss has occurred when the noise is unbounded on any of the test functions.

We conclude this section by remarking that throughout the experiments, our proposed method has been relatively unaffected by the changes in sample setting assumptions and can be used with minimal fine-tuning. Indeed, only the relation between the number of samples in each subset and the diameter of each of these subsets needs to be modified (see Remark 3.3.17). This relation can be set in a theoretically principled manner by considering the results given in Remarks 3.3.11, 3.3.12 and Corollary 3.3.16 or treated as a hyper-parameter and set more heuristically. The flexibility of the LCLS algorithm is in contrast to existing asymptotically consistent Lipschitz learning algorithms such as the benchmark approaches used in this section which either only consider noiseless sampling settings or require prior knowledge of the noise distribution in order to be applied.

3.4 Connections to Machine Learning & Related Fields

The theoretical results derived in Section 3.2 are fundamental in nature. They can be used as a benchmark when developing novel Lipschitz constant estimators or more generally to provide a better theoretical understanding of algorithms that depend explicitly on Lipschitz constant estimates of an underlying target function. Utilising Corollaries 3.2.4 and 3.2.7 the worst-case estimation errors of Lipschitz constant estimation can be better understood and their negative impact on overall performance mitigated. This is particularly important as Lipschitz constant dependent algorithms often rely on heuristic or experimental arguments which might not always hold in practice to justify the Lipschitz constant estimation step.

In some settings, the LCLS estimator developed in Section 3.3 can be directly ap-

plied to improve existing computational frameworks in which case the finite sample guarantees derived in Theorem 3.3.10 and Corollary 3.3.13 can be used. In particular, when a (loose) bound on the second order partial derivatives of the target function is known, a sequence of feasible Lipschitz constants converging to the best Lispchitz constant at a known convergence speed is obtainable. Unfortunately, while this approach is possible in all the sampling set-ups considered in this chapter, the convergence rates obtained for the noisy sampling set-up (see Corollaries 3.3.13 and 3.3.16) can be too slow to be useful in some practical applications. In these cases, the LCLS estimator can be applied directly to estimate the Lipschitz constant without feasibility guarantees.

In the section below, we briefly discuss how the results and algorithms derived in this chapter can be used in the fields of system identification and global optimisation.

3.4.1 Global Optimisation

A major subfield of global optimisation research focuses on sequential search methods that explicitly utilise the Lipschitz constant of the target function to remove large sets in the search space and enhance the efficiency of exploration (Shubert [1972], Mladineo [1986]). As a good estimate of the Lipschitz constant is not always available in practice, work arounds must be be found (Jones et al. [1993]). In particular, a number of these optimisation frameworks make use of a Lipschitz constant estimator (Kvasov and Sergeyev [2012] and references therein, D'Agostino [2022]). The computation of these estimates is generally done heuristically without convergence analysis or error-certificate of the Lipschitz constant estimates. Therefore, the minimax bounds derived in Theorem 3.2.3 of Section 2 provide a context for the expected performance of these methods. More precisely, given recent work by Malherbe and Vayatis [2017] and Bachoc et al. [2021] which derives optimal sample complexity rates for Lipschitz Optimisation when a Lipschitz constant is known, it becomes possible to derive a lower bound on the sample complexity of adaptive Lipschitz Optimisation algorithms that separate the optimisation procedure and the Lipschitz constant estimation. We derive such a lower bound below as an example

of how this can be done.

Following the set-up of certified online learning algorithms described in Bachoc et al. [2021], we assume that we have access to a black-box target function f that can be queried to obtain noiseless observations. The goal of certified global optimisation is to design an algorithm that systematically queries f in order to generate an output sequence $((x_n, f(x_n^*), \zeta_n))_{n\in\mathbb{N}}$ where x_n is the n-th query point, $f(x_n^*)$ is the generated estimate of $\max_{x\in\mathcal{X}} f(x)$ after n queries and $\zeta_n \geq 0$ is an error certificate that guarantees: $\max_{x\in\mathcal{X}} f(x) - f(x_n^*) \leq \zeta_n$.

Given an accuracy $\epsilon \in \mathbb{R}_+$, we can then define the sample complexity [14] $N(A, f, \epsilon)$ of a certified global optimisation algorithm A as the smallest number of queries needed in order to obtain an error certificate smaller than ϵ for all f belonging to a function class $\mathcal{C}$, or in other words:

$$N(A, f, \epsilon) := \min\{n \in \mathbb{N} \cup \{+\infty\} | \zeta_n < \epsilon\}.$$

Utilising this theoretical set-up, we can then combine the theoretical results of Bachoc et al. [2021] with Corollary 3.2.4 in order to obtain the following statement on the worst case lower sample complexity bound of the adaptive Lipschitz optimisation problem.

Proposition 3.4.1 *(Sample Complexity - Adaptive Lipschitz Optimisation) Assume that $\mathcal{X}$ is the hypercube and consider the set $\mathcal{A}$ of adaptive Lipschitz optimisation algorithms which combine classical Lipschitz optimisation methods with a separable[15] feasible Lipschitz constant estimator. There exists constants $C_1, C_2 > 0$ such that $\forall L^* \geq 0$, $A \in \mathcal{A}$ and $\epsilon \in (0, \epsilon_0)$ where $\epsilon_0 \in (0, 2^{d-1}ML^*)$:*

$$\sup_{f\in C^2(\mathcal{X},K)\cap\mathcal{F}_p(L^*)} N(A, f, \epsilon)$$
$$\geq C_1\alpha_d(M, L^*, K)((1 + C_2 \max(\min(\frac{3}{C_2}, \frac{1}{\lceil(1 + \log_2(\frac{\epsilon_0}{\epsilon})\rceil)^{\frac{1}{d}} + \beta(L^*, K, \epsilon)}),$$

[14] Note: this differs slightly from the definition used in Bachoc et al. [2021].

[15] In other words, only knowledge of the Lipschitz constant estimate is used in the optimisation part of the algorithm.

$$\frac{\gamma_d(M, L^*, \epsilon)}{\beta(L^*, K, \epsilon)}) \ )^{\frac{1}{2}} - 1)^d \tag{3.2}$$

where $m := \max_{y \in \mathcal{X}} f(y)$, $V_{\mathcal{X}} = M^d$, K *is as defined in Assumption 3 and*

- $\alpha_d(M, L^*, K) := (\frac{ML^*}{K})^d$ *represents the problem's general dependency on the input space size, true best Lipschitz constant of the target function and desired precision of the optimisation algorithm and second degree partial derivatives.*
- $\beta(L^*, K, \epsilon) := (1 + \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil)^{1/d} \frac{K\epsilon}{L^{*2}}$ *represents the dependency on the true best Lipschitz constant of the target function, the second degree partial derivatives and and desired precision of the optimisation algorithm.*
- $\gamma_d(M, L^*, \epsilon) := \sqrt[d]{\frac{L^*\epsilon(d-1)}{M}}$ *represents the dependency on the input space size, true best Lipschitz constant of the target function and desired precision of the optimisation algorithm and second degree partial derivatives.*

To our knowledge, (3.2) is the first lower bound on the sample complexity of adaptive Lipschitz optimisation frameworks (see Malherbe and Vayatis [2017] for a possible sample complexity upper bound provided by the adaLIPO algorithm). It depends on the input space, desired precision and upper bounds on the first two orders of differentiation of f. The structure of the proof of Proposition 3.4.1 as well as the two terms contained in the max expression of the lower bound can be interpreted as a comparison between the sample complexity arising from the optimisation procedure and the one arising from the Lipschitz constant estimation. In particular, $\frac{\gamma_d(L^*, M, \epsilon)}{\beta(L^*, K, \epsilon)}$ is computed by considering the subset of linear functions of $C^2(\mathcal{X}, K) \cap \mathcal{F}_p(L^*)$ which is trivial to optimise in the case where the Lipschitz constant is known but becomes complicated to certify if the Lipschitz constant estimation is difficult.

Unfortunately, the proposed bound is loose as the sampling scheme for the Lipschitz optimisation algorithm can differ significantly from the sampling scheme of Lipschitz constant estimator and is of moderate interest as it only considers a subset of adaptive Lipschitz optimisation algorithms. It does however provide an example of how the lower bounds derived in Section 3.2 of this chapter can be utilised to theoretically analyse existing computational frameworks that rely on Lipschitz learning and future work will consider refining the lower bound given in (3.2).

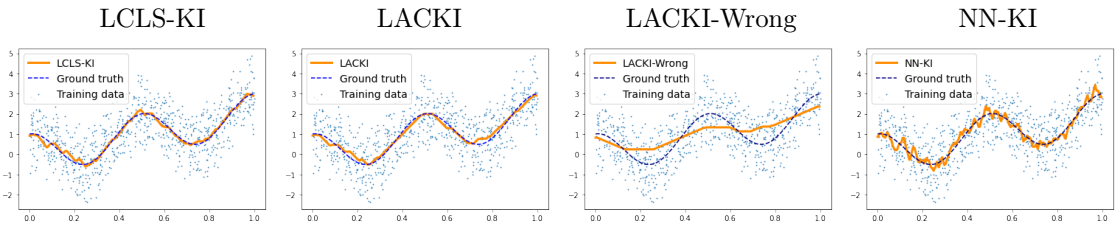


Figure 3.6: Illustration of several non-parametric methods applied to noisy data. The target function is $f : x \mapsto cos(4\pi x) + 2x$ and the noise is distributed according to a truncated Gaussian distribution (std: 1, upper/lower bound: $-2/2$). The predictions of the trained methods are plotted in orange and the training data in light blue (800 observations). From left to write: LCLS-KI: Kinky inference using the Lipschitz constant estimate generated by the LCLS algorithm. LACKI: Adaptive Kinky Inference proposed by Calliess et al. [2020] with correctly set error bounds. LACKI-wrong: LACKI method with error bounds set at the wrong observational error bound (i.e. at 1.3× the true error bound). NN-KI: Kinky inference method using the Lipschitz constant of a fitted Neural Network model with sigmoid activation as proposed by Milanese and Novara [2004].

Finally, we note that the lower bounds derived in Section 3.2 can also be considered in the application of recently proposed batch Bayesian optimisation frameworks (González et al. [2016], Alvi et al. [2019]). Indeed, while these methods provide interesting experimental results, the convergence bound stated in Corollary 3.2.4 shows that in the worst case the Lipschitz constant estimate generated from the fitted Gaussian Process can differ significantly from the true Lipschitz constant - severely impacting the performance of the algorithm in high dimensional settings. At best, the Lipschitz constant estimate used in these chapters: $\max_{x\in\mathcal{X}} \|\mu_\nabla(x)\|$ must be replaced by $\max_{x\in\mathcal{X}} \|\mu_\nabla(x)\| + C(d,p)\frac{MK}{\sqrt[d]{n}}$ in order to ensure that the estimated value is a feasible Lipschitz constant. Here μ_∇ denotes the mean function of the gradient function estimate associated to the fitted GP which can be computed efficiently using the covariance function of the GP.

3.4.2 Non-parametric Regression for System Identification

As discussed in Chapter 2, Lipschitz interpolation frameworks (Milanese and Novara [2004], Beliakov [2006], Calliess et al. [2020]) explicitly utilise the Lipschitz constant of an underlying Lipschitz continuous target function to define the smallest set of all possible systems that is consistent with the observed data and to provide optimal[16]

[16] See Theorem 2.2.4 and (Milanese and Novara [2004]).

point estimates. In the relevant literature, a number of approaches have been used to estimate the Lipschitz constant however these either rely on heuristic estimation (Milanese and Novara [2004], Calliess [2017]) or on knowledge of often unavailable hyper-parameters such as tight bounds on the noise (Novara et al. [2013],Calliess et al. [2020]) which underestimate the true Lipschitz constant. Utilising the LCLS algorithm developed in Section 3.3 would therefore be an interesting alternative approach to constructing an adaptive Nonlinear Set Membership framework. As noted at the beginning of the section, we directly utilise the Lipschitz estimate produced by the LCLS estimator as the worst-case error guarantees stated in Corollaries 3.3.13 and 3.3.16 are too conservative to be useful in the considered use case.

In Figure 3.6, we illustrate the performance of a hybrid LCLS - Kinky Inference method in comparison to other non-parametric methods that depend explicitly on the Lipschitz constant of the target function. The variation of the plotted non-parametric predictors is a direct function of the Lipschitz constant estimated from the data – when the Lipschitz constant estimate underestimates the true Lipschitz constant flatter prediction curves that do not fully capture the nonlinearity of the target function are produced while Lipschitz constant estimates that overestimate the true Lipschitz constant produce overly input sensitive predictions. In fact, the kinky inference framework converges to a nearest neighbour estimator as the Lipschitz constant goes to infinity (Maddalena and Jones [2020b]).

Remark 3.4.2 *In the hybrid regression method considered in this section, we apply the LCLS and the Kinky inference algorithms sequentially, effectively separating the computation process. It is worth noting that the same methodology also allows for the creation of more sophisticated regression techniques by integrating the LCLS algorithm with either the projected or local extensions of the Kinky inference algorithm.*

In Figure 3.7, we observe that under the truncated Gaussian noise assumptions, the proposed LCLS-KI approach seems to perform best in comparison to the other non-parametric methods as long as the bound on the second derivative K (see Assumption 3) is not too large relative to the number of observations in the training

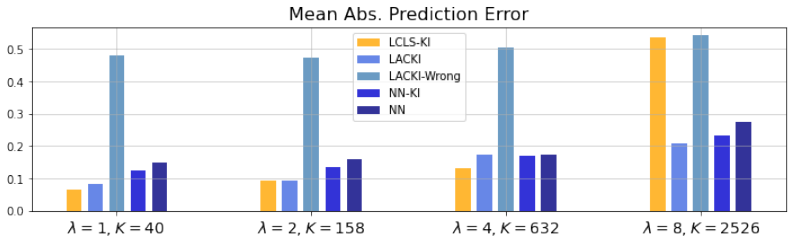


Figure 3.7: Mean absolute error of the non-parametric methods discussed in Figure 3.6 and the neural network utilised to estimate the Lipschitz constant in the NN-KI method. The target functions are given by $f_\lambda : x \mapsto cos(2\lambda\pi x)+\lambda x$, for $\lambda = 1, 2, 4, 10$ and associated maximum second derivative: $K := \max_{x\in\mathcal{X}} f''_\lambda(x) = 40, 158, 632, 2526$. The values shown in the plot are computed on a test set containing 500 independently sampled observations.

data. As noted in Section 3.3.4, this is due to the fact that the LCLS algorithm is more dependent on K than other classes of Lipschitz learning algorithms and can significantly underestimate the true Lipschitz constant when K is too large. Therefore, when the second order derivatives are moderate and upper bounds on the noise on the noise are not precisely known, the LCLS-KI algorithm provides an interesting alternative to existing nonlinear set membership/Lipschitz interpolation methods. Applications of LCLS-KI to the common-use case of such methods, e.g. in learning-based model predictive control (Canale et al. [2014], Limon et al. [2017]), could be pursued in future work.

3.5 Conclusions

In this chapter, we have established precise lower and upper bounds on the sample complexity of the estimation of Lipschitz constants under minimal parametric constraints on the target function. Instead, our bounds rely on the assumption of C^2 regularity of the target function which, given a compact input space, implies the existence of an upper bound on the second degree partial derivatives; this type of assumption is unavoidable as if the second degree partial derivatives are not assumed bounded, then the sample complexity can not be guaranteed to be finite and any theoretical characterisation of the general Lipschitz learning problem is trivial. The obtained bounds on the sample complexity are shown to be optimal in the

noiseless sampling setting and in the noisy sampling setting for a slightly modified but generally equivalent version of the problem under a Gaussian noise assumption. These results can be used to provide a theoretical baseline for the Lipschitz learning problem and to help drive the design of future black-box Lipschitz constant estimators.

In order to derive the upper bound on the sample complexity, we have proposed a new algorithm for Lipschitz learning based on local least squares regression that is sample-optimal in the noiseless setting and in the noisy setting with Gaussian noise. We have thoroughly investigated the theoretical properties of this algorithm showing asymptotic consistency, guarantees on finite sample behaviour and computational complexity in both noiseless and general noisy sampling settings.

A series of brief empirical experiments illustrate how these theoretical results translate into practice and how the LCLS algorithm compares to existing classical Lipschitz constant estimators. The proposed method provides a suitable solution for Lipschitz constant estimation when a theoretically principled and computationally flexible approach is needed.

3.6 Overview of Empirical Test Functions

Function	Expression	Lipschitz Const.	Key Property
(a)	See Lemma 3.A.1	3.054	Lipschitz constant reached in a unique point.
(b)	$e^{-(x_1^2+x_2^2)}cos(x_1)cos(x_2)$	8.5776	Large second degree partial derivatives (K).
(c)	$cos(5x_1)$	5	Simple test function.
(d)	See Lemma 3.A.1	30.5399	Higher dimensional input ($\mathbb{R}^3$).

Table 3.1: Test Functions. Note: Functions (a), (d) are based on the function set utilised in proofs of the sample complexity lower bounds of Section 3.2.

Appendices

Contents

Appendix 3.A Proofs: Lower bounds on Sample Complexity

Lemma 3.A.1 *(Properties of $\mathcal{F}$) For $C_1, C_2 \in \mathbb{R}$, define the function $g_0 : \mathbb{R}^d \to \mathbb{R}$,*

$$g_0(x) = \begin{cases} C_1 e^{-\frac{1}{1-C_2 \sum_{j=1}^d x_j^2}} & \text{if } C_2 \sum_{j=1}^d x_j^2 < 1 \\ 0 \text{ otherwise.} \end{cases}.$$

The following properties of g_0 can be shown;

1. $\max_{x \in \mathbb{R}^d} \|\nabla g_0(x)\|_2 \approx 0.8 C_1 \sqrt{C_2}$
2. $\max_{x \in \mathbb{R}^d, i,j \in \{1,\dots,d\}} \left|\frac{\partial^2 g_0}{\partial x_i x_j}(x)\right| \approx 7.75 C_1 C_2$.

Proof Let g_0 be as described above. It follows from construction that g_0 is a radial function and that there exists $u : [0, +\infty) \to \mathbb{R}$, such that $\forall x \in \mathbb{R}^d$, $u(\sum_{i=1}^d x_i^2) = g_0(x)$ (and in other terms, $\forall r \in [0, +\infty)$, $u(r) := g_0(\sqrt{r}, 0, ..., 0)$). We can compute the maximum magnitude (in $\|.\|_2$) of the gradient of g_0 as follows:

$$\max_{x\in\mathbb{R}^d} \|\nabla g_0(x)\|_2 = \max_{x\in\mathbb{R}^d} \| \left(2x_1 u' \left(\sum_{i=1}^d x_i^2 \right), 2x_2 u' \left(\sum_{i=1}^d x_i^2 \right), ... \right) \|_2$$

$$= \max_{x\in\mathbb{R}^d} \{ \left| 2u' \left(\sum_{i=1}^d x_i^2 \right) \right| \|x\|_2 \} = \max_{r\in\mathbb{R}_+} |2u'(r^2)r| = \max_{0\le r\le \frac{1}{\sqrt{C_2}}} \left\{ 2C_1C_2 r \frac{e^{-\frac{1}{1-C_2r^2}}}{(1-C_2r^2)^2} \right\}$$

$$= C_1\sqrt{C_2} \max_{0\le r\le 1} \left\{ 2r \frac{e^{-\frac{1}{1-r^2}}}{(1-r^2)^2} \right\}.$$

Computing $\max_{0\le r\le 1}\{2r\frac{e^{-\frac{1}{1-r^2}}}{(1-r^2)^2}\}$ gives $\max_{0\le r\le 1}\{2r\frac{e^{-\frac{1}{1-r^2}}}{(1-r^2)^2}\} = \frac{6\sqrt[4]{3^3}e^{-\frac{1}{1-\frac{1}{\sqrt{3}}}}}{(\sqrt{3}-3)^2}$. Since g_0 is continuously differentiable and the support of ∇g_0 is compact, we have that there exists $x^* \in \mathbb{R}^d$ such that $\|\nabla g_0(x^*)\|_2 = C_1\sqrt{C_2}\frac{6\sqrt[4]{3^3}}{(\sqrt{3}-3)^2}e^{-\frac{\sqrt{3}}{\sqrt{3}-1}} \approx 0.8C_1\sqrt{C_2}$.

Similarly, we have that for $i \in \{1, ..., d\}$, $x \in \mathbb{R}^d$;

$$\frac{\partial^2 g_0}{\partial x_i^2}(x) = 2u' \left(\sum_{i=1}^d x_i^2 \right) + 4x_i^2 u'' \left(\sum_{i=1}^d x_i^2 \right).$$

Here it is clear that for $x^* \in \text{argmax}_{x\in\mathbb{R}^d} |\frac{\partial^2 g_0}{\partial x_i^2}(x)|$ either (1) $x_i^* = 0$ or (2) $x_j^* = 0$, $\forall i \neq j$. In the first case; we can compute $\max_{r\in\mathbb{R}_+} |2u'(r)| = \frac{8C_1C_2}{e^2} \approx 1.08C_1C_2$. In the second case, setting $x = re_i$, we consider the computation of $\max_{r\in\mathbb{R}_+} |2u'(r^2) + 4r^2u''(r^2)|$. We have

$$2u'(r^2) + 4r^2u''(r^2) = 2C_1C_2e^{-\frac{1}{1-r^2}} \frac{3C_2^2r^4 - 1}{(1-C_2r^2)^4}$$

and can compute

$$\max_{r\in\mathbb{R}_+} |2u'(r^2) + 4r^2u''(r^2)| = C_1C_2 \max_{r\in\mathbb{R}_+} |2e^{-\frac{1}{1-r^2}} \frac{3C_2^2r^4 - 1}{(1-C_2r^2)^4}| \approx 7.75C_1C_2.$$

Therefore, we have $\max_{x\in\mathbb{R}^d}|\frac{\partial^2 g_0}{\partial x_i^2}(x)| \approx 7.75C_1C_2$. Finally, we check $\forall i \neq j \in \{1,...,d\}$, $\max_{x\in\mathbb{R}^d}|\frac{\partial^2 g_0}{\partial x_i x_j}(x)| = \max_{x\in\mathbb{R}^d}|4x_ix_ju''(\sum_{i=1}^d x_i^2)|$. Clearly, we can set $x = re_i + se_j$ for $r, s \in \mathbb{R}_+$. Computing this quantity gives:

$$\max_{x\in\mathbb{R}^d}\left|4x_ix_ju''(\sum_{i=1}^d x_i^2)\right| = C_1C_2^2\max_{(r,s)\in\mathbb{R}_+\times\mathbb{R}_+}\left|\frac{4rse^{-\frac{1}{1-C_2(r^2+s^2)}}}{(1-C_2(r^2+s^2))^3}\right| = C_1C_2\frac{8\sqrt{2}e^{-2-\sqrt{2}}}{(\sqrt{2}-2)^3}.$$

We obtain $\forall i \neq j \in \{1,...,d\}$, $\max_{x\in\mathbb{R}^d}|\frac{\partial^2 g_0}{\partial x_i x_j}(x)| \approx 1.85C_1C_2 \leq \max_{x\in\mathbb{R}^d}|\frac{\partial^2 g_0}{\partial x_i^2}(x)|$.

■

Proof of Theorem 3.2.3 (Sample Complexity Bound – Noiseless).
Let $p = 2$. If we can show that there exists a set $\mathcal{F} \subset C^2(\mathcal{X}, K) \cap \mathcal{F}_p(L^*)$ of functions that can be constructed such that

$$\forall \hat{L} \in \mathcal{L}_{n,p}(\mathcal{X}), \sup_{f\in\mathcal{F}}|\hat{L}(f) - L_p^*(f)| > \epsilon$$

when $n < \big(C(d,p)\frac{MK}{\epsilon}\big)^d$, then Theorem 3.2.3 follows directly. This expression can be simplified to the equivalent statement:

$$\forall \hat{L} \in \mathcal{L}_{n,p}(\mathcal{X}), \exists f \in \mathcal{F} \text{ such that } |\hat{L}(f) - L_p^*(f)| > \epsilon.$$

Consider the following functional family. For $C_1 \in \mathbb{R}, C_2 \in \mathbb{R}_+$,

$$\mathcal{F}_0(C_1, C_2)$$
$$:= \Big\{g_z : \mathbb{R}^d \to \mathbb{R} | z \in \mathcal{X}, g_z(x) = \begin{cases} C_1 e^{-\frac{1}{1-C_2\sum_{j=1}^d (x_j-z_j)^2}} \text{ if } C_2\sum_{j=1}^d (x_j-z_j)^2 < 1 \\ 0 \text{ otherwise.} \end{cases}\Big\}.$$

For any L^*, we can consider the family $\mathcal{F}_{L^*}(C_1, C_2)$ by adding a linear component, e.g. L^*x_1 to $g_z \in \mathcal{F}_0(C_1, C_2)$. In this case, we have by the construction of $\mathcal{F}_0(C_1, C_2)$ that for all $g_z^0 \in \mathcal{F}_0(C_1, C_2)$ with support in $\mathcal{X}$ and $g_z^{L^*} \in \mathcal{F}_{L^*}(C_1, C_2)$, $z \in \mathcal{X}$, $\max_{x\in\mathbb{R}^d}\|\nabla g_z^{L^*}(x)\|_2 = \max_{x\in\mathbb{R}^d}\|\nabla g_z^0(x)\|_2 + L^*$ and $\max_{x\in\mathbb{R}^d, i,j\in\{1,...,d\}}\big|\frac{\partial^2 g_z^{L^*}(x)}{\partial x_i x_j}\big| = \max_{x\in\mathbb{R}^d, i,j\in\{1,...,d\}}\big|\frac{\partial^2 g_z^0(x)}{\partial x_i x_j}\big|$. The second relation is obvious while the first follows from

the fact that all $g_z^0 \in \mathcal{F}(C_1, C_2)$ are radial functions which implies that all gradients of g_z^0 are either pointing towards or away from z with equal magnitude along any hypersphere of fixed radius. Therefore, by the properties g_z^0 shown in Lemma 3.A.1, for any choice of linear component $l : \mathbb{R}^d \to \mathbb{R}$, $l(x) = ax + b$ for $a \in \mathbb{R}^d, b \in \mathbb{R}$ such that $\|a\|_2 = L^*$, there exists $x^* \in \mathcal{X}$ such that $\max_{x\in\mathbb{R}^d} \|\nabla g_z^0(x)\|_2 = \|\nabla g_z^0(x^*)\|_2$ and $\nabla g_z^0(x^*)$ and a have the same direction (if they have opposite direction it suffices to take x^* that is diametrically opposed on the same hypersphere). With this construction,

$$\max_{x\in\mathbb{R}^d} \|\nabla g_z^{L^*}(x)\|_2 = \|\nabla g_z^0(x^*) + a\|_2 = \|\nabla g_z^0(x)\| + \|a\|_2 = \max_{x\in\mathbb{R}^d} \|\nabla g_z^{L^*}(x)\| + L^*.$$

We can therefore restrict our proof to considering the case where $L^* = 0$.

In the first part of the proof, we will show that for carefully selected values $C_1^*, C_2^* \in \mathbb{R}$, $\mathcal{X}$ contains $\sim (\frac{MK}{\epsilon})^d$ disjointed $\|.\|_2$-hyperspheres $\mathcal{B} := \{B_i\}_{i\in\{1,\dots,(\frac{MK}{\epsilon})^d\}}$ of radius $\frac{1}{\sqrt{C_2^*}}$ such that; $\forall B_i, B_j \in \mathcal{B}$, $B_i \subset \mathcal{X}$, $B_i \cap B_j = \emptyset$ if $i \neq j$ and a set $\mathcal{F} \subset \mathcal{F}_0$ of associated functions with the following properties; $\forall g_{\bar{z}_i} \in \mathcal{F}$ associated to $B_i \in \mathcal{B}$,

1. $supp(g_{\bar{z}_i}) = B_i$,
2. $\max_{x\in\mathcal{X}} \|\nabla g_{\bar{z}_i}(x)\|_2 \geq 2\epsilon \quad (+L^* \text{ if } L^* \neq 0)$,
3. $\forall k, j \in \{1, ..., d\}$, $\max_{x\in\mathcal{X}} |\frac{\partial^2 g_{\bar{z}_i}}{\partial x_k x_j}(x)| \leq K$.

To do so we consider the gradient and second order partial derivatives of the functions in $\mathcal{F}$. Let $g \in \mathcal{F}$, applying Lemma 3.A.1, we have :

1. $\max_{x\in\mathbb{R}^d} \|\nabla g(x)\|_2 \approx 0.8 C_1\sqrt{C_2} \quad (+L^* \text{ if } L^* \neq 0)$
2. $\max_{x\in\mathbb{R}^d, i,j\in\{1,\dots,d\}} \left|\frac{\partial^2 g(x)}{\partial x_i x_j}\right| \approx 7.75 C_1 C_2$.

Using these values, we can define the values of C_1^* and C_2^* discussed earlier in the proof. Firstly, in order to have $g \in C^2(\mathbb{R}^d, K)$, we need $\max_{x\in\mathbb{R}^d, i,j\in\{1,\dots,d\}} |\frac{\partial^2 g}{\partial x_i x_j}(x)| \leq K$. This implies the relation $C_1 = \frac{K}{7.75 C_2}$. Secondly, we set C_2 such that $\max_{x\in\mathcal{X}}$

$\|\nabla g(x)\|_2 = 0.8C_1\sqrt{C_2} = 2\epsilon$. Plugging in the relation for C_1 given above;

$$\frac{0.1K}{\sqrt{C_2}} = 2\epsilon \Leftrightarrow (\frac{K}{20\epsilon})^2 = C_2^* \text{ and } C_1^* = \frac{51\epsilon^2}{K}.$$

Setting $l = \frac{1}{\sqrt{C_2^*}} = \frac{20\epsilon}{K}$ we have

$$supp(g)^{17} := \left\{ x \in \mathbb{R}^d | C_2^* \sum_{i=1}^{d} x_i^2 < 1 \right\} = B_l(c)$$

where $B_l(z)$ denotes the d-dimensional ball of radius l defined with respect to $\|.\|_2$ and centred in $c \in \mathcal{X}$. The last step before defining $\mathcal{F}$ is to count how many balls of radius l can fit[18] in $\mathcal{X} = [0, M]^d$. Here, we use a lower bound that is obtained by considering the regular hypercube partition of $\mathcal{X}$ of side-length $\tilde{l}$, defined by; $N := \lfloor \frac{M}{l} \rfloor$ and $\tilde{l} = \frac{M}{N}$. Let $\mathcal{B}$ denote the set of balls of radius l that can be inscribed in a subset belonging to the hypercube partition of $\mathcal{X}$. Then, for all $B_i, B_j \in \mathcal{B}$, $B_i \subset \mathcal{X}$ and $B_i \cap B_j = \emptyset$. Furthermore, we have $|\mathcal{B}| = (\frac{M}{\tilde{l}})^d \approx \left(\frac{MK}{20\epsilon}\right)^d$ (+ constant).

The associated set $\mathcal{F}$ of functions can be constructed by utilising the set $\mathcal{Z}$ of ball centers z_i for $B_i \in \mathcal{B}$ and the values C_1^*, C_2^* computed above to define

$$\mathcal{F} := \{ g_z \in \mathcal{F}_0(C_1^*, C_2^*) | z \in \mathcal{Z} \} \cup \{ f^0 \}$$

where $f^0 \equiv 0$. Suppose that $n < \frac{1}{20^d}\left(\frac{MK}{\epsilon}\right)^d$ and consider an arbitrary $\hat{L} \in \mathcal{L}_{n,p}(\mathcal{X})$. By construction, there exists a ball B in the set $\mathcal{B}$ with associated ball center $z \in \mathcal{Z}$ (as defined above) such that no observations are sampled in B. Therefore, if the unknown target function $f \in \{g_z^0, f^0\}$ then $\forall (x, \Omega(x)) \in D_{\hat{L}}$, $\Omega(x) = 0$. This implies that we can freely set the target function to either g_z^0 or f^0 with no change to the Lipschitz constant estimate generated by $\hat{L}$. It then suffices to select g_z^0 if $\hat{L}$ generates a prediction that is smaller than ϵ and f^0 otherwise. As the choice of $\hat{L}$

[17] Which we define as the subset of $\mathcal{X}$ where g is non-zero.

[18] This is often referred to as a "packing" of $\mathcal{X}$ by balls of radius l and the maximum cardinality of such a set is called the "packing number" denoted $N(\mathcal{X}, l)$.

was arbitrary, this implies that:

$$\forall \hat{L} \in \mathcal{L}_{n,p}(\mathcal{X}), \exists f \in \mathcal{F} \text{ such that } |\hat{L}(f) - L_p^*(f)| > \epsilon$$

and therefore:

$$\inf_{\hat{L} \in \mathcal{L}_{n,p}(\mathcal{X})} \sup_{f \in C^2(\mathcal{X},K) \cap \mathcal{F}_p(L^*)} |\hat{L}(f) - L_p^*(f)| > \epsilon.$$

Utilising norm equivalences and Lemma 2.1.1, we can apply similar arguments to the ones given above to obtain that in the case: $p = 1$, the sample complexity of the Lipschitz learning problem can be lower bounded by $\left(C(d)\frac{MK}{\epsilon}\right)^d$, where $C(d) = \frac{1}{20d^{\frac{1}{2}}}$.

■

Proof of Theorem 3.2.6 (Sample Complexity Bound – Noisy).
Let $\epsilon > 0$ be sufficiently small such that $\frac{40\epsilon}{K} < M$ (which implies that the packing number $N(\mathcal{X}, \frac{20\epsilon}{K}) > 0$). Consider the maximal packing $\mathcal{B}_\epsilon$ of $\mathcal{X}$ of radius $\frac{20\epsilon}{K}$ with respect to $\|.\|_2$ and the associated class of functions $\mathcal{F}_0$ defined in Lemma 3.A.1 which we denote $\mathcal{F}_\epsilon$ in this proof in order to explicitly mark the dependence on ϵ (we only consider $L^* = 0$). We recall that for all $B \in \mathcal{B}_\epsilon$ and associated $f_B \in \mathcal{F}_\epsilon$; $\max_{x \in B} \|\nabla f_B(x)\|_q = 2\epsilon$ and $\max_{x \in \mathcal{X} \setminus B} \|\nabla f_B(x)\|_q = 0$. Therefore, by construction of $\mathcal{F}_\epsilon$, we have for any distinct pair of functions $f_1, f_2 \in \mathcal{F}_\epsilon$ and $\forall x \in \mathcal{X}$

$$\max\{Loss(x, f_1), Loss(x, f_2)\} = \max\{|\|\nabla f_1(x)\|_q - L^*|, |\|\nabla f_2(x)\|_q - L^*|\} > \epsilon.$$

with $L^* := 2\epsilon$. This implies that

$$\inf_{\hat{L} \in \mathcal{L}_{n,p}(\mathcal{X})} \sup_{f \in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}(f)}, f) > \epsilon) \geq \inf_{\hat{A} \in \mathcal{H}} \sup_{f \in \mathcal{F}_\epsilon} \mathbb{P}(f^{\hat{A}} \neq f).$$

where $\mathcal{H}$ denotes the class of algorithms that utilise the data samples $D_{\hat{L}}$ in order to select the correct f in $\mathcal{F}_\epsilon$. In order to lower bound the right hand side of the above equation, Fano's Lemma can be applied. To do so, we first estimate $\log(|\mathcal{F}_\epsilon|)$ and $\sup_{f_1, f_2 \in \mathcal{F}_\epsilon} \mathrm{KL}\left(p_{f_1} || p_{f_2}\right)$ where p_f denotes the density defined on $(\mathcal{X}, \mathcal{Y})$ of a noisy sample $(x, \tilde{f}_x)$ associated to $f \in \mathcal{F}_\epsilon$ (defined more precisely below). The first term:

$\log(|\mathcal{F}_\epsilon|)$ follows directly from the proof of Theorem 3.2.3 where we obtained that $|\mathcal{F}_\epsilon| = |\mathcal{B}_\epsilon| \geq \left(\frac{MK}{20\epsilon}\right)^d$ which implies

$$\log(|\mathcal{F}_\epsilon|) \geq d \log(\frac{MK}{20\epsilon}).$$

Let $f \in \mathcal{F}_\epsilon$. Denoting the density of the uniform measure on $\mathcal{X}$ as λ_u and the density of the Gaussian measure on $\mathbb{R}$ with mean $\mu \in \mathbb{R}$ and variance σ^2 as ν_{μ,σ^2}, we have $\forall x \in \mathcal{X},\ y \in \mathbb{R},\ p_f(x,y) = \lambda_u(x)\nu_{f(x),\sigma^2}(y)$. Then, the second term can be upper bounded as follows:

$$\sup_{f_1,f_2 \in \mathcal{F}_\epsilon} \mathrm{KL}\left(p_{f_1}||p_{f_2}\right) = \int_{\mathcal{X}\times\mathbb{R}} p_{f_1}(x,y) \log\left(\frac{p_{f_1}(x,y)}{p_{f_2}(x,y)}\right) d(x,y)$$

$$= \int_{\mathcal{X}} \int_{\mathbb{R}} \nu_{f_1(x),\sigma^2}(y) \log\left(\frac{\nu_{f_1(x),\sigma^2}(y)}{\nu_{f_2(x),\sigma^2}(y)}\right) dy \lambda_u(x) dx \overset{\textbf{(i)}}{=} \frac{1}{2\sigma^2} \int_{\mathcal{X}} |f_1(x) - f_2(x)|^2 \lambda_u(x) dx$$

$$\overset{\textbf{(ii)}}{\leq} \frac{1}{\sigma^2} \|f_1 - f_2\|_\infty^2 \int_{vol(B_\epsilon)} \lambda_u(x) dx$$

where B_ϵ denotes an arbitrary ball in the packing defined by $\mathcal{B}_\epsilon$, **(i)** follows from the well-known KL-divergence of univariate Gaussians and **(ii)** follows by construction of $\mathcal{F}_\epsilon$. By the proof of Theorem 3.3.7, we have for all $B \in \mathcal{B}_\epsilon$ and associated $f_B \in \mathcal{F}_\epsilon$, $\sup_{x\in B} |f_B(x)| = C_1^* = \frac{51\epsilon^2}{K}$. Furthermore, $\int_{vol(B_\epsilon)} \lambda_u(x) dx = \tilde{c}\frac{vol(B_\epsilon)}{M^d} = \tilde{c}' \left(\frac{20\epsilon}{MK}\right)^d$ for some constants $\tilde{c}, \tilde{c}' := c(d) > 0$. Therefore, there exists a constant $c > 0$ such that

$$\sup_{f_1,f_2 \in \mathcal{F}_\epsilon} \mathrm{KL}\left(p_{f_1}||p_{f_2}\right) \leq \frac{c}{\sigma^2} \frac{\epsilon^{d+4}}{M^d K^{d+2}}.$$

Applying Fano's Lemma, we obtain for an arbitrary ordering of $\mathcal{F}_\epsilon$:

$$\inf_{\hat{L} \in \mathcal{L}_{n,p}(\mathcal{X})} \sup_{f \in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}(f)}, f) > \epsilon)$$

$$\geq 1 - \frac{\log(2) + n \sup_{f_1,f_2 \in \mathcal{F}_\epsilon} \mathrm{KL}\left(p_{f_1}||p_{f_2}\right)}{\log(|\mathcal{F}_\epsilon|)} \geq 1 - \frac{\log(2) + n\frac{c}{\sigma^2}\frac{\epsilon^{d+4}}{M^d K^{d+2}}}{d\log(\frac{MK}{20\epsilon})}.$$

Therefore, for ϵ sufficiently small and any arbitrary $\delta \in (0,1)$,

$$\inf_{\hat{L}\in\mathcal{L}_{n,p}(\mathcal{X})} \sup_{f\in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}(f)}, f) > \epsilon) \leq \delta \implies 1 - \frac{\log(2) + n\frac{c}{\sigma^2}\frac{\epsilon^{d+4}}{M^d K^{d+2}}}{d\log(\frac{MK}{20\epsilon})} \leq \delta.$$

Taking the limit as ϵ goes to 0, we have that if $n \notin \Omega\left(\frac{M^d K^{d+2}\log(\frac{MK}{\epsilon})}{\epsilon^{d+4}}\right)$, then $\lim_{\epsilon\to 0^+} 1 - \frac{\log(2)+n\frac{c}{2\sigma^2}\frac{\epsilon^{d+4}}{M^d K^{d+2}}}{d\log(\frac{MK}{20\epsilon})} = 1 > \delta$. This implies that

$$n \in \Omega\left(\frac{\sigma^2 M^d K^{d+2}\log(\frac{MK}{\epsilon})}{\epsilon^{d+4}}\right)$$

must necessarily hold in order for $\inf_{\hat{L}\in\mathcal{L}_{n,p}(\mathcal{X})} \sup_{f\in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}(f)}, f) > \epsilon) \leq \delta$ to hold and concludes the proof.

■

Appendix 3.B Proofs: Theoretical Properties of LCLS

Proof of Proposition 3.3.4 (Computational Complexity of LCLS).
Follows directly from the computational complexity of the linear least squares regression algorithm which is $O(n_{samples})$.

■

3.B.1 Technical Lemmas

The proof of Theorem 3.3.7 relies on the following technical lemmas.

Lemma 3.B.1 *(Fundamental logarithm inequalities) For all $x > 0$,*

$$1 - \frac{1}{x} \leq \log(x) \leq x - 1.$$

Lemma 3.B.2 *Let $\delta \in (0, 1)$, then $\forall x \geq 2\log(\frac{1}{1-\frac{\delta}{2}})$:*

$$\frac{1}{1 - \sqrt[x]{1 - \frac{\delta}{2}}} \leq \frac{2x}{\log(\frac{1}{1-\frac{\delta}{2}})}.$$

Proof Let $x > 2\log(\frac{1}{1-\frac{\delta}{2}})$ be arbitrary, we have

$$\frac{1}{1 - \sqrt[x]{1 - \frac{\delta}{2}}} \leq \frac{2x}{\log(\frac{1}{1-\frac{\delta}{2}})} \iff 1 \leq \frac{2x}{\log(\frac{1}{1-\frac{\delta}{2}})}(1 - \sqrt[x]{1 - \frac{\delta}{2}}).$$

Then,

$$\frac{2x}{\log(\frac{1}{1-\frac{\delta}{2}})}(1 - \sqrt[x]{1 - \frac{\delta}{2}}) = \frac{2x}{\log(\frac{1}{1-\frac{\delta}{2}})}(1 - e^{-\frac{1}{x}\log(\frac{1}{1-\frac{\delta}{2}})}).$$

Utilising the fact that $e^y \leq 1 + y + y^2$ for all $y < 1$, we have that the above expression is greater or equal to:

$$\frac{2x}{\log(\frac{1}{1-\frac{\delta}{2}})}\left(\frac{1}{x}\log(\frac{1}{1-\frac{\delta}{2}}) - \frac{1}{x^2}\log(\frac{1}{1-\frac{\delta}{2}})^2\right) = 2\left(1 - \frac{1}{x}\log(\frac{1}{1-\frac{\delta}{2}})\right) \geq 1$$

where last inequality follows from the fact that $x > 2\log(\frac{1}{1-\frac{\delta}{2}})$.

■

Lemma 3.B.3 *Consider a sequence of partitions $(\mathcal{H}_I)_{I\in\mathbb{N}}$ used by LCLS and assume that for a given $\eta \in (0,1]$ the sampling distribution satisfies Assumption 5. Then,*

$$\forall I \in \mathbb{N},\ \forall H \in \mathcal{H}_I,\ \|\left({X_I^H}^\top X_I^H\right)^{-1}\|_2 \leq \frac{16}{\eta {\delta_I^H}^2 N_I^H}.$$

Proof Let $\lambda_{max}(M)$ denote the maximum eigenvalue of a matrix M if it exists. $\forall I \in \mathbb{N}, \forall H \in \mathcal{H}_I$, we have $\|\left({X_I^H}^\top X_I^H\right)^{-1}\|_2 = \frac{1}{\sigma_{min}({X_I^H}^\top X_I^H)}$ where $\sigma_{min}({X_I^H}^\top X_I^H)$ denotes the smallest singular value of ${X_I^H}^\top X_I^H$. Therefore, we can focus on showing the following relation that implies the Lemma statement:

$$\sigma_{min}({X_I^H}^\top X_I^H) \geq \frac{\eta N_I^H}{16}{\delta_I^H}^2.$$

Let $\bar{X}_I^H$ be the design matrix without the first column of ones, ie. $\bar{X}_I^H = \begin{bmatrix} x_{H_1}^\top \\ x_{H_2}^\top \\ \vdots \\ x_{H_{N_I^H}}^\top \end{bmatrix}$,

we have

$$\sigma_{min}({X_I^H}^\top X_I^H) = \sigma_{min}({X_I^H}^\top)^2 = \min_{\|u\|_2=1} \|{X_I^H}^\top u\|_2^2 = \min_{\|u\|_2=1} \| \begin{bmatrix} \mathbf{1}_{N_I^H}^\top \\ \bar{X_I^H}^\top \end{bmatrix} u\|_2^2$$

$$\geq \min_{\|u\|_2=1} \| \begin{bmatrix} \mathbf{0}_{N_I^H}^\top \\ \bar{X_I^H}^\top \end{bmatrix} u\|_2^2 = \min_{\|u\|_2=1} \|\bar{X_I^H}^\top u\|_2^2 = \sigma_{min}(\bar{X_I^H}^\top \bar{X_I^H}).$$

Therefore we can consider the smallest singular value of $\bar{X_I^H}$ instead of X_I^H which allows for a direct use of Assumption 5. We have

$$\sigma_{min}(\bar{X_I^H}^\top \bar{X_I^H}) = \lambda_{min}(\bar{X_I^H}^\top \bar{X_I^H}) = \min_{u\in\mathbb{R}^d, \|u\|_2=1} u^\top \bar{X_I^H}^\top \bar{X_I^H} u = \min_{u\in\mathbb{R}^d, \|u\|_2=1} \sum_{i=1}^{N_I^H} \langle x_{H_i}, u\rangle^2.$$

Let $u_{min}^* \in \mathbb{R}^d$ denote the eigenvector associated to $\lambda_{min}(\bar{X_I^H}^\top \bar{X_I^H})$ with $\|u_{min}^*\|_2 = 1$.

Then, u^*_{min} satisfies $\sum_{i=1}^{N_I^H} \langle x_{H_i}, u^*_{min} \rangle^2 = \min_{\|u\|_2=1} \sum_{i=1}^{N_I^H} \langle x_{H_i}, u \rangle^2$.

Using u^*_{min}, we can construct an orthonormal basis of $\{u^*_{min}, u_2, ..., u_d\}$ of $\mathbb{R}^d$ (such a basis can be constructed using the Gramm-Schmitt algorithm). Then, since D_I^H contains ηN_I^H points in each ball associated to an element of an $\frac{\delta_I^H}{8}$-cover of H, we have that there exists at least ηN_I^H pairs of datapoints $(x_{H_i}, \tilde{f}_{H_j})$, $(x_{H_j}, \tilde{f}_{H_j}) \in D_I^H$ such that $\exists \{\alpha_i\}_{i \in \{1,...,d\}}, \alpha_i \in \mathbb{R}$ with $|\alpha_1| > \frac{\delta_I^H}{2}$ and $(x_{H_i} - x_{H_j}) = \alpha_1 u^*_{min} + \sum_{k=2}^{d} \alpha_k u_k$. This implies that $\max(|\langle x_{H_i}, u^*_{min} \rangle|, |\langle x_{H_j}, u^*_{min} \rangle|) \geq \frac{\delta_I^H}{4}$. Indeed, if $|\langle x_{H_i}, u^*_{min} \rangle| < \frac{\delta_I^H}{4}$, then

$$|\langle x_{H_j}, u^*_{min} \rangle| = |\langle x_{H_j} - x_{H_i} + x_{H_i}, u^*_{min} \rangle| = |\langle x_{H_j} - x_{H_i}, u^*_{min} \rangle + \langle x_{H_i}, u^*_{min} \rangle|$$

$$\geq \frac{\delta_I^H}{2} - \frac{\delta_I^H}{4} = \frac{\delta_I^H}{4}.$$

Using this inequality ηN_I^H times we can conclude, $\sigma_{min}(\bar{X_I^H}^\top \bar{X_I^H}) = \sum_{i=1}^{N_I^H} \langle x_{H_i}, u^*_{min} \rangle^2 \geq \eta N_I^H (\frac{\delta_I^H}{4})^2$.

■

Lemma 3.B.4 *Consider the constructions of Definition 3.3.3. The following relationship holds for all $I \in \mathbb{N}$,*

$$\frac{V(\mathcal{X})\Gamma(\frac{d}{2}+1)2^d}{\pi^{\frac{d}{2}} \max_{H \in \mathcal{H}_I} (\Delta_I^H)^d} \leq |\mathcal{H}_I| \leq \frac{V(\mathcal{X})\Gamma(\frac{d}{2}+1)2^d}{\pi^{\frac{d}{2}} \min_{H \in \mathcal{H}_I} (\delta_n)^d}$$

where $V(\mathcal{X})$ denotes the volume of $\mathcal{X}$.

Proof Follows directly from the definition of $\{\delta_I^H\}_{H \in \mathcal{H}_I}$, $\{\Delta_I^H\}_{H \in \mathcal{H}_I}$ and volume formula for the d-dimensional ball.

■

Lemma 3.B.5 *Let the notation and assumptions be as described in Theorem 3.3.7*

and define $x^* \in \mathcal{X}$ *as* $x^* \in argmax_{x\in\mathcal{X}} \|\nabla f(x)\|_q$. *Then,* $\forall I \in \mathbb{N}$,

$$\left| \|\nabla f(x^*)\|_q - \max_{H\in\mathcal{H}_I} \{\|\mathbb{E}\left[\hat{\beta}_I^H\right]\|_q\} \right| \leq \frac{4\sqrt{d}dn_qK}{\sqrt{\eta}} a_I.$$

where $n_q = d^{\max\{\frac{1}{q}-\frac{1}{2},0\}}$.

Proof Note: such an x^* exists by compactness of $\mathcal{X}$ and the fact that $f \in C^2(\mathcal{X})$.

By definition, $[\hat{b}_I^H, \hat{\beta}_I^H]^\top = (X_I^{H^\top} X_I^H)^{-1} X_I^{H^\top} \tilde{f}_I^H$. Computing the expectation of this expression yields

$$\mathbb{E}\Big[[\hat{b}_I^H, \hat{\beta}_I^H]^\top\Big] = \mathbb{E}\Big[(X_I^{H^\top} X_I^H)^{-1} X_I^{H^\top} \tilde{f}_I^H\Big]$$

$$= \mathbb{E}\Big[(X_I^{H^\top} X_I^H)^{-1} X_I^{H^\top} f_I^H\Big] + \mathbb{E}\Big[(X_I^{H^\top} X_I^H)^{-1} X_I^{H^\top} \gamma_I^H\Big] = (X_I^{H^\top} X_I^H)^{-1} X_I^{H^\top} f_I^H.$$

$\forall H \in \mathcal{H}_I$, let $c_H \in \overline{H}$ (closure of H) be such that $\|\nabla f(c_H)\|_q = \max_{x\in\bar{H}}\{\|\nabla f(x)\|_q\}$ which exists by compactness of $\bar{H}$ and the fact that $f \in C^2(\mathcal{X})$. Then, using the second order Taylor expansion of f around c_h, every coordinate f_{H_k} of f_I^H can be re-expressed as

$$f_{H_k} = f(c_H) + (x_{H_k} - c_H)^\top \nabla f(c_H) + (x_{H_k} - c_H)^\top Hess(c_H + r_{H_k}(x_{H_k} - c_H))(x_{H_k} - c_H), \tag{3.3}$$

where $r_{H_k} \in [0,1]$ and $Hess$ denotes the Hessian matrix of f. To alleviate notation, let $\|.\|_{\tilde{q}}$ denote a pseudo-norm on $\mathbb{R}^{d+1}$ defined by; $x \in \mathbb{R}^{d+1}$, $\|x\|_{\tilde{q}} := \sqrt[q]{\sum_{i=2}^{d+1} x_i^q}$ if $q < \infty$ and $\|x\|_{\tilde{\infty}} := \max_{i\in\{2,\dots,d+1\}} |x_i|$ otherwise. Then, using the definition of X_I^H,

$$\|\mathbb{E}[\hat{\beta}_I^H]\|_q = \| \begin{bmatrix} \mathbb{E}[\hat{b}_I^H] \\ \mathbb{E}[\hat{\beta}_I^H] \end{bmatrix} \|_{\tilde{q}} = \|(X_I^{H^\top} X_I^H)^{-1} X_I^{H^\top} f_I^H\|_{\tilde{q}}$$

$$= \|(X_I^{H^\top} X_I^H)^{-1} X_I^{H^\top} (X_I^H \begin{bmatrix} f(c_H) - c_H^\top \nabla f(c_H) \\ 0 \\ \vdots \\ 0 \end{bmatrix} + X_I^H \begin{bmatrix} 0 \\ \nabla f(c_H) \end{bmatrix} + \begin{bmatrix} (x_{H_1} - c_H)^\top Hess(r_{H_1})(x_{H_1} - c_H) \\ (x_{H_2} - c_H)^\top Hess(r_{H_2})(x_{H_2} - c_H) \\ \vdots \end{bmatrix}) \|_{\tilde{q}}$$

$$= \| \begin{bmatrix} f(c_H) - c_H^\top \nabla f(c_H) \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \nabla f(c_H) \end{bmatrix} + \underbrace{(X_I^{H\top} X_I^H)^{-1} X_I^{H\top} (\begin{bmatrix} (x_{H_1} - c_H)^\top Hess(r_{H_1})(x_{H_1} - c_H) \\ (x_{H_2} - c_H)^\top Hess(r_{H_2})(x_{H_2} - c_H) \\ \vdots \end{bmatrix})}_{=: J(H)} \|_{\tilde{q}}.$$

Plugging this expression into the theorem statement yields:

$$\left| \|\nabla f(x^*)\|_q - \max_{H \in \mathcal{H}_I} \{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\} \right| = \|\nabla f(x^*)\|_q - \max_{H \in \mathcal{H}_I} \{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\}$$

$$\leq \|\nabla f(x^*)\|_q - (\max_{H \in \mathcal{H}_I} \{\| \begin{bmatrix} 0 \\ \nabla f(c_H) \end{bmatrix} \|_{\tilde{q}}\} - \max_{H \in \mathcal{H}_I} \{\|J(H)\|_{\tilde{q}}\})$$

$$\leq \|\nabla f(x^*)\|_q - \max_{H \in \mathcal{H}_I} \{\|\nabla f(c_H)\|_q\} + \max_{H \in \mathcal{H}_I} \{\|J(H)\|_q\} = \max_{H \in \mathcal{H}_I} \{\|J(H)\|_q\}$$

where the last equality follows from the fact that there exists $H \in \mathcal{H}_I$ such that $x^* \in H$. As $f \in C^2(\mathcal{X}, K)$, we have $\forall i, j \in \{1, ..., d\}$, $\forall x \in \mathcal{X}$ that $|\frac{\partial^2 f}{\partial x_i \partial x_j}(x)| < K$. This implies that $\|Hess(x)\|_1 \leq dK \ \forall x \in \mathcal{X}$ and by matrix norm equivalence; $\|Hess(x)\|_2 \leq \sqrt{d}\|Hess(x)\|_1 \leq d\sqrt{d}K$, $\forall x \in \mathbb{R}^d$. Therefore, since matrix p-norms are sub-multiplicative;

$$\|J(H)\|_q \leq n_q \|J(H)\|_2 \leq n_q \|(X_I^{H\top} X_I^H)^{-1} X_I^{H\top}\|_2 \| \begin{bmatrix} (x_{H_1} - c_1)^\top Hess(r_{H_1})(x_{H_1} - c_H) \\ (x_{H_2} - c_2)^\top Hess(r_{H_2})(x_{H_2} - c_H) \\ \vdots \end{bmatrix} \|_2$$

where $n_q = d^{\max\{\frac{1}{q} - \frac{1}{2}, 0\}}$. Using Lemma 3.B.3 we have

$$\|(X_I^{H\top} X_I^H)^{-1} X_I^{H\top}\|_2 = \|X_I^H ((X_I^{H\top} X_I^H)^\top)^{-1}\|_2 = \sqrt{\lambda_{max}((X_I^{H\top} X_I^H)^{-1})}$$

$$= \sqrt{\|(X_I^{H\top} X_I^H)^{-1}\|_2} \leq \frac{4}{\delta_I^H \sqrt{\eta N_I^H}}$$

where λ_{max} denotes the maximum eigenvalue of $(X_I^{H\top} X_I^H)^{-1}$. Furthermore, the components of the vector on the right-hand side can be upper bounded by;

$$|(x_{H_1} - c_1)^\top Hess(r_{H_1})(x_{H_1} - c_H)| \leq \|(x_{H_1} - c_1)^\top\|_2 \|Hess(r_{H_1})(x_{H_1} - c_H)\|_2$$

$$\leq \|(x_{H_1} - c_1)^\top\|_2 \|Hess(r_{H_1})\|_2 \|(x_{H_1} - c_H)\|_2 \leq \Delta_I^{H^2} \sqrt{d} d K.$$

Combining these two upper bounds we can conclude:

$$\left| \|\nabla f(x^*)\|_q - \max_{H \in \mathcal{H}_I} \{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\} \right| \leq \max_{H \in \mathcal{H}_I} \frac{4\sqrt{d} d K n_q}{\sqrt{\eta}} \frac{\Delta_I^{H^2}}{\delta_I^H} = \frac{4\sqrt{d} d n_q K}{\sqrt{\eta}} a_I.$$

■

Lemma 3.B.6 *If the Assumptions of Theorem 3.3.7 hold, then $\forall I \in \mathbb{N}$, the difference between the Lipschitz estimate generated by the LCLS method with noisy sampling $\hat{L}_I$ and the Lipschitz estimate generated by the LCLS method with noiseless sampling $\bar{L}_I$ can be upper bounded by;*

$$\mathbb{P}(|\bar{L}_I - \hat{L}_I| > \frac{\epsilon}{2}) \leq 1 - \prod_{H \in \mathcal{H}_I} (1 - \frac{2^6 \sigma^2 d^{\max\{\frac{2}{q}, 1\}}}{\eta \epsilon^2} \frac{1}{N_I^H \delta_I^{H^2}}). \tag{3.4}$$

Proof Let $I \in \mathbb{N}$. $\forall H \in \mathcal{H}_I$ denote by $[b_I^H, \beta_I^H]$ the least squares coefficients computed using (X_I^H, f_I^H) (instead of $(X_I^H, \tilde{f}_I^H)$), i.e. the noiseless least squares coefficients. Then,

$$\mathbb{E}\left[[\hat{b}_I^H, \hat{\beta}_I^H]^\top\right] = \mathbb{E}\left[(X_I^{H^\top} X_I^H)^{-1} X_I^{H^\top} f_I^H\right] = [b_I^H, \beta_I^H]^\top.$$

Therefore, we can write (with $n_q = d^{\max\{\frac{1}{q} - \frac{1}{2}, 0\}}$)

$$\mathbb{P}\left(\left|\bar{L}_I - \hat{L}_I\right| > \frac{\epsilon}{2}\right) = \mathbb{P}\left(|\max_{H \in \mathcal{H}_I} \{\|[\beta_I^H]\|_q\} - \max_{H \in \mathcal{H}_I} \{\|\hat{\beta}_I^H\|_q\}| > \frac{\epsilon}{2}\right)$$

$$= \mathbb{P}\left(\left|\max_{H \in \mathcal{H}_I} \{\|\mathbb{E}\left[\hat{\beta}_I^H\right]\|_q\} - \max_{H \in \mathcal{H}_I} \{\|\hat{\beta}_I^H\|_q\}\right| > \frac{\epsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\left|\max_{H \in \mathcal{H}_I} \{\|\mathbb{E}\left[\hat{\beta}_I^H\right]\|_q - \|\hat{\beta}_I^H\|_q\}\right| > \frac{\epsilon}{2}\right) \leq \mathbb{P}\left(\max_{H \in \mathcal{H}_I} \{\|\mathbb{E}\left[\hat{\beta}_I^H\right] - \hat{\beta}_I^H\|_q\} > \frac{\epsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\max_{H \in \mathcal{H}_I} \{\|\mathbb{E}\left[\hat{\beta}_I^H\right] - \hat{\beta}_I^H\|_2\} > \frac{\epsilon}{2n_q}\right) = 1 - \mathbb{P}\left(\max_{H \in \mathcal{H}_I} \{\|\mathbb{E}\left[\hat{\beta}_I^H\right] - \hat{\beta}_I^H\|_2\} < \frac{\epsilon}{2n_q}\right)$$

$$\leq 1 - \prod_{H \in \mathcal{H}_I} \mathbb{P}\left(\|\mathbb{E}\left[\hat{\beta}_I^H\right] - \hat{\beta}_I^H\|_2 < \frac{\epsilon}{2n_q}\right)$$

$$= 1 - \prod_{H \in \mathcal{H}_I} \left(1 - \mathbb{P}\left(\|\mathbb{E}\left[\hat{\beta}_I^H\right] - \hat{\beta}_I^H\|_2 \geq \frac{\epsilon}{2n_q}\right)\right).$$

In order to upper bound the term given in product: $\mathbb{P}(\|\mathbb{E}[\hat{\beta}_I^H] - \hat{\beta}_I^H\|_2 \geq \frac{\epsilon}{2n_q})$, we use the covariance matrix: $var([\hat{b}_I^H, \hat{\beta}_I^H]) = \sigma^2 (X_I^H{}^\top X_I^H)^{-1}$ which follows from the fact that the components of γ_I^H are assumed to be uncorrelated with mean 0 and variance σ^2. We also denote by $Tr(M)$ the trace of a matrix $M \in \mathbb{R}^{d \times d}$. Then, by applying an extension of Chebyshev's inequality to finite dimensional vectors (Ferentios [1982]) and Lemma 3.B.3, we have

$$\mathbb{P}\left(\|\mathbb{E}\left[\hat{\beta}_I^H\right] - \hat{\beta}_I^H\|_2 \geq \frac{\epsilon}{2n_q}\right) \overset{\substack{Chebychev's \\ Inequality}}{\leq} \frac{4n_q^2 \sigma^2 Tr((X_I^H{}^\top X_I^H)^{-1})}{\epsilon^2}$$

$$\leq \frac{4n_q^2 \sigma^2 d \|\left(X_I^H{}^\top X_I^H\right)^{-1}\|_2}{\epsilon^2} \overset{Lemma\ 3.B.3}{\leq} \frac{4n_q^2 \sigma^2 d}{\epsilon^2} \frac{16}{\eta \delta_I^H{}^2 N_I^H} = \frac{2^6 n_q^2 \sigma^2 d}{\eta \epsilon^2} \frac{1}{N_I^H \delta_I^H{}^2}.$$

Plugging this expression into the product given above concludes the proof.

$$\mathbb{P}(|\bar{L}_I - \hat{L}_I| > \frac{\epsilon}{2}) \leq 1 - \prod_{H \in \mathcal{H}_I} \left(1 - \frac{2^6 \sigma^2 d^{\max\{\frac{2}{q}-1, 0\}} d}{\eta \epsilon^2} \frac{1}{N_I^H \delta_I^H{}^2}\right).$$

■

3.B.2 Proof of Main Theoretical Properties of LCLS

Proof of Theorem 3.3.7 (General Convergence Rate).

We recall that $\forall I \in \mathbb{N}$, the Lipschitz estimate $\hat{L}_I$ is obtained by considering the partition $\mathcal{H}_I$ and computing $\max_{H \in \mathcal{H}_I}\{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\}$. Let $\epsilon > 0$ be arbitrary. We need to show for $p = 1, 2$:

$$\lim_{I \to \infty} \mathbb{P}(|L_p^* - \hat{L}_I| > \epsilon) = 0$$

and for $p > 2$

$$\lim_{I \to \infty} \mathbb{P}(|L_p - \hat{L}_I| > \epsilon) = 0 \text{ with } L_p \in \mathbb{R}_{\geq L_p^*}.$$

Since f verifies Assumption 3 and $\mathcal{X}$ is convex and compact, Lemma 2.1.1 guarantees the existence of $x^* \in \mathcal{X}$ such that $\|\nabla f(x^*)\|_q = L_p^*$ for $p = 1, 2$ and $L_p := \|\nabla f(x^*)\|_q = \max_{x\in\mathcal{X}} \|\nabla f(x)\|_p \geq L_p^*$ for $p > 2$.

Therefore, for all $p \geq 1$, we can consider the statement;

$$\lim_{I\to\infty} \mathbb{P}(|\|\nabla f(x^*)\|_q - \hat{L}_I| > \epsilon) = 0.$$

Let $I \in \mathbb{N}$ and consider $\mathbb{P}(|\|\nabla f(x^*)\|_q - \hat{L}_I| > \epsilon)$. This expression can be split into two terms:

$$\begin{aligned}&\mathbb{P}(|\|\nabla f(x^*)\|_q - \hat{L}_I| > \epsilon)\\ &\leq \underbrace{\mathbb{P}\left(|\|\nabla f(x^*)\|_q - \max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\right]\|_q\}| > \frac{\epsilon}{2}\right)}_{(I)} + \underbrace{\mathbb{P}\left(|\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\right]\|_q\} - \hat{L}_I| > \frac{\epsilon}{2}\right)}_{(II)}.\end{aligned}$$

In the following, we show that both (I) and (II) converge to 0 when I goes to infinity.

(I): From Lemma 3.B.5, $\left|\|\nabla f(x^*)\|_q - \max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\right]\|_q\}\right| \leq \frac{4\sqrt{d}dn_qK}{\sqrt{\eta}} a_I$. Plugging this upper bound into the above expression, we have

$$\mathbb{P}\left(|\|\nabla f(x^*)\|_q - \max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\right]\|_q\}| > \frac{\epsilon}{2}\right) \leq \mathbb{P}(\frac{4\sqrt{d}dn_qK}{\sqrt{\eta}} a_I > \frac{\epsilon}{2}).$$

By hypothesis 2. $\lim_{I\to\infty} a_I = 0$ and therefore there exists $I_1 \in \mathbb{N}$ sufficiently large such that $\frac{4\sqrt{d}dn_qK}{\sqrt{\eta}} a_{I_1} \leq \frac{\epsilon}{2}$ and therefore $\mathbb{P}(\frac{4\sqrt{d}dn_qK}{\sqrt{\eta}} a_{I_1} > \frac{\epsilon}{2}) = 0$.

(II): We show that $\mathbb{P}(|\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\} - \hat{L}_I| > \frac{\epsilon}{2})$ converges to 0 as I goes to infinity. Let $\bar{L}$ denote the Lipschitz constant estimate generated by LCLS with noiseless samples. Then, applying Lemma 3.B.6, we have the following upper bound on $\mathbb{P}(|\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\} - \hat{L}_I| > \frac{\epsilon}{2})$:

$$\mathbb{P}(|\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\right]\|_q\} - \hat{L}_I| > \frac{\epsilon}{2}) = \mathbb{P}(|\bar{L}_I - \hat{L}_I| > \frac{\epsilon}{2})$$

$$\leq 1 - \prod_{H \in \mathcal{H}_I} (1 - \frac{16\sigma^2 d^{\max\{\frac{2}{q},1\}}}{\eta\epsilon^2} \frac{1}{N_I^H \delta_I^{H^2}}) \leq 1 - (1 - \frac{2^6\sigma^2 d^{\max\{\frac{2}{q},1\}}}{\eta\epsilon^2 \min_{H \in \mathcal{H}_I}(N_I^H \delta_I^{H^2})})^{|\mathcal{H}_I|}.$$

As by Theorem hypothesis 2 $\lim_{I\to\infty} \max_{H\in\mathcal{H}_I}(\Delta_I^H) = 0$, applying Lemma 3.B.4 implies that $\lim_{I\to\infty} |\mathcal{H}_I| = \infty$. Therefore using the fact that $\lim_{I\to\infty} b_I = 0$, we have $\lim_{I\to\infty} \max_{H\in\mathcal{H}_I} \left(\frac{1}{N_I^H \delta_I^{H^2}}\right) = \lim_{I\to\infty} \frac{1}{\min_{H\in\mathcal{H}_I} N_I^H \delta_I^{H^2}} = 0$.

To alleviate notation, let $(\alpha_I)_{I\in\mathbb{N}}$ be the sequence defined by $\alpha_I := \frac{2^6\sigma^2 d^{\max\{\frac{2}{q},1\}}}{\eta\epsilon^2 \min_{H\in\mathcal{H}_I} N_I^H \delta_I^{H^2}}$, then $\lim_{I\to\infty} \frac{1}{\min_{H\in\mathcal{H}_I} N_I^H \delta_I^{H^2}} = 0$ implies that $\exists \bar{I} \in \mathbb{N}$ such that $\forall I \geq \bar{I}$, $\alpha_I < 0.5$. Utilising fundamental logarithm inequalities, we obtain:

$$1 - (1 - \alpha_I)^{|\mathcal{H}_I|} \leq |\mathcal{H}_I| \log(\frac{1}{1-\alpha_I}) \leq |\mathcal{H}_I| \frac{\alpha_I}{1-\alpha_I} \leq |\mathcal{H}_I| \frac{\alpha_I}{2}$$

$$= \frac{2^5\sigma^2 d^{\max\{\frac{2}{q},1\}}}{\eta\epsilon^2} \frac{|\mathcal{H}_I|}{\min_{H\in\mathcal{H}_I} N_I^H \delta_I^{H^2}} = (\frac{2^5\sigma^2 d^{\max\{\frac{2}{q},1\}}}{\eta\epsilon^2}) b_I \overset{I\to\infty}{\longrightarrow} 0.$$

■

Proof of Corollary 3.3.9 (Noiseless Oracle).

As in the proof of Theorem 3.3.7, we can consider the statement; $p \in \mathbb{N}$,

$$\lim_{I\to\infty} \mathbb{P}(|\|\nabla f(x^*)\|_q - \hat{L}_I| > \epsilon) = 0$$

where $x^* \in \text{argmax}_{x\in\mathcal{X}} \|\nabla f(x)\|_p$. Since the data samples contain no noise, $\hat{L}_I = \max_{H\in\mathcal{H}_I} \{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\}$ and

$$\mathbb{P}(|\|\nabla f(x^*)\|_q - \hat{L}_I| > \epsilon) = \mathbb{P}(|\|\nabla f(x^*)\|_q - \max_{H\in\mathcal{H}_I} \{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\}| > \epsilon).$$

Then, applying Lemma 3.B.5 and using $\lim_{I\to\infty} a_I = 0$ as in the proof of Theorem 3.3.7 gives the desired convergence result.

(Note: that the least squares estimation is well defined as $N_I^H \geq d+1$ and Assumption 5 holds.)

■

Proof of Theorem 3.3.10 (Finite Sample Guarantee).

We show the equivalent statement; $\mathbb{P}(|L_p - \hat{L}_I| > \epsilon) \leq \delta$. where as in proof of Theorem 3.3.7, $L_p := \|\nabla f(x^*)\|_q$ with $x^* := \text{argmax}_{x\in\mathcal{X}}\|\nabla f(x)\|_p$ and $L_p = L_p^*$ for $p = 1, 2$, $L_p \geq L_p^*$ for $p > 2$.

In the hypercube set-up, we have $\forall I \in \mathbb{N}_{>1}$, $\forall H \in \mathcal{H}_I$, $\Delta_I^H = \frac{\sqrt{d}M}{I}$, $\delta_I^H = \frac{M}{I}$ and $|\mathcal{H}_I| = I^d$. Let $\epsilon > 0$, $\delta \in (0, \frac{1}{2}]$: From the proof of Theorem 3.3.7 we have that three following inequalities need to be satisfied in order for (3.1) to hold. From **(I)** we need $\frac{4\sqrt{d}dn_qK}{\sqrt{\eta}}\frac{{\Delta_I^H}^2}{\delta_I^H} \leq \frac{\epsilon}{2}$ in order for $\mathbb{P}(|\|\nabla f(x^*)\|_q - \max_{H\in\mathcal{H}_I}\{\|\mathbb{E}[\hat{\beta}_I^H]\|_q\}| > \frac{\epsilon}{2}) = 0$. This implies that;

$$I \geq \frac{8d^2\sqrt{d}n_q}{\sqrt{\eta}}\frac{MK}{\epsilon}.$$

From **(II)**, we have the following two inequalities that need to be satisfied;

$$(1) \qquad \alpha_I = \frac{2^6\sigma^2 d^{\max\{\frac{2}{q},1\}}}{\eta\epsilon^2 \min_{H\in\mathcal{H}_I} N_I^H \delta_I^{H2}} < 0.5$$

$$(2) \qquad \frac{2^5\sigma^2 d^{\max\{\frac{2}{q},1\}}}{\eta\epsilon^2}\frac{|\mathcal{H}_I|}{\min_{H\in\mathcal{H}_I} N_I^H \delta_I^{H2}} < \delta.$$

The first implies that

$$\frac{2^7 d^{\max\{\frac{2}{q},1\}}\sigma^2}{\eta}\frac{I^2}{M^2\epsilon^2} < \min_{H\in\mathcal{H}_I} N_I^H$$

and the second expression gives

$$\frac{2^5 d^{\max\{\frac{2}{q},1\}}\sigma^2}{\eta}\frac{I^2}{M^2\epsilon^2}\frac{|\mathcal{H}_I|}{\delta} < \min_{H\in\mathcal{H}_I} N_I^H.$$

Since $|\mathcal{H}_I| = I^d$, $I \in \mathbb{N}_{>1}$ and $\delta \in (0, \frac{1}{2}]$, we have that if the $\min_{H\in\mathcal{H}_I} N_I^H$ satisfies (2) then (1) is true as well. Therefore, we have $\forall H \in \mathcal{H}_I$;

$$\frac{2^5 d^{\max\{\frac{2}{q},1\}}\sigma^2}{\eta}\frac{I^{d+2}}{\delta M^2\epsilon^2} < \min_{H\in\mathcal{H}_I} N_I^H.$$

Setting $C_1(d) = 8d^2\sqrt{d}d^{\max\{\frac{1}{q}-\frac{1}{2},0\}}$ and $C_2(d,q) = 2^5 d^{\max\{\frac{2}{q},1\}}d$ concludes the proof.

■

Proof of Theorem 3.3.15 (Asymptotic Sample Complexity – Gaussian Noise).

Consider the setting described by Theorem 3.3.15 with $I = \lceil C_1(d)\frac{MK}{\epsilon} \rceil$ when $\epsilon > 0$. As described in the proof of Theorem 3.3.10: in the the hypercube set-up we have $\forall H \in \mathcal{H}_I$, $\Delta_I^H = \frac{\sqrt{d}M}{I}$, $\delta_I^H = \frac{M}{I}$ and $|\mathcal{H}_I| = I^d$.

For all $\epsilon > 0$, let $\mathcal{A}_\epsilon$ denote the event that every $H \in \mathcal{H}_I$ contains a number of samples N_I^H equal or greater than $C(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4}$ for a constant $C(d) > 0$ that depends on d (see $(\star)$ for the explicit definition of $C(d)$) and is $(\frac{\delta_I^H}{8}, \eta)$-covered where $\eta = \frac{vol(B_1(0))}{2^{3d+1}3^d}$ where $vol(B_1(0))$ denotes the volume of the d-dimensional unit ball. More precisely,

$$\mathcal{A}_\epsilon := \{\forall H \in \mathcal{H}_I : N_I^H \geq C(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4} \wedge H \text{ is } (\frac{\delta_I^H}{8}, \eta)\text{-covered}\}.$$

Let us assume that there exists $\bar{\epsilon} > 0$ such that $\forall \epsilon \in (0, \bar{\epsilon})$, $\mathbb{P}(\mathcal{A}_\epsilon) > 0$ (this will follow from $(\star\star)$ given at the end of the proof). Then,

$$\sup_{f \in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon) \leq \sup_{f \in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon | \mathcal{A}_\epsilon) + \mathbb{P}(\mathcal{A}_\epsilon^c).$$

Therefore, in order to show Theorem 3.3.15, it suffices to show that both terms of the right-hand expression given above converge to 0 as ϵ goes to 0. The first part of the proof considers $\sup_{f \in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon | \mathcal{A}_\epsilon)$. We will show that for all $\delta > 0$, there exists $\bar{\epsilon}^*$ such that $\forall \epsilon \in (0, \bar{\epsilon}^*)$,

$$\sup_{f \in C^2(\mathcal{X},K)} \mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon | \mathcal{A}_\epsilon) < \delta.$$

Notation 3.B.7 *To alleviate notation, we omit the conditional dependence on $\mathcal{A}_\epsilon$ in the following computations.*

Fix an arbitrary $\delta > 0$ and define $\forall H \in \mathcal{H}_I$, $\bar{\beta}_I^H := [b_I^H, \beta_I^H]^\top = ({X_I^H}^\top X_I^H)^{-1} {X_I^H}^\top f_I^H$. As the noise and the sampling distribution are independent and every input sample is selected independently, we have

$$\bar{\beta}_I^H = \mathbb{E}\left[[\hat{b}_I^H, \hat{\beta}_I^H]^\top \middle| X_I^H\right] = \mathbb{E}\left[[\hat{b}_I^H, \hat{\beta}_I^H]^\top \middle| G_I^{\mathcal{X}}\right]$$

where we recall (from Chapter 2) that $G_I^{\mathcal{X}}$ denotes the set of all sample inputs. (Note that $\bar{\beta}_I^H$ is a random variable as the sample inputs are randomly sampled). We have

$$\mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon) = \mathbb{P}\left(\left|L_p^* - \|\nabla f(x^{\hat{L}_I(f)})\|_q\right| > \epsilon\right)$$

$$\leq \mathbb{P}\left(\left|L_p^* - \hat{L}_I(f)\right| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(\left|\hat{L}_I(f) - \|\nabla f(x^{\hat{L}_I(f)})\|_q\right| \geq \frac{\epsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\left|\|\nabla f(x^*)\|_q - \max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H \middle| G_I^{\mathcal{X}}\right]\|_q\}\right| > \frac{\epsilon}{4}\right)$$

$$+\mathbb{P}\left(\left|\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H \middle| G_I^{\mathcal{X}}\right]\|_q\} - \hat{L}_I(f)\right| > \frac{\epsilon}{4}\right)$$

$$+\mathbb{P}\left(\left|\|\nabla f(x^{\hat{L}_I(f)})\|_q - \|\mathbb{E}\left[\hat{\beta}_I^{H^{\hat{L}_I(f)}} \middle| G_I^{\mathcal{X}}\right]\|_q\}\right| > \frac{\epsilon}{4}\right)$$

$$+\mathbb{P}\left(\left|\|\mathbb{E}\left[\hat{\beta}_I^{H^{\hat{L}_I(f)}} \middle| G_I^{\mathcal{X}}\right]\|_q - \hat{L}_I(f)\right| > \frac{\epsilon}{4}\right).$$

where $x^* := \text{argmax}_{x\in\mathcal{X}}|\|\nabla f(x)\|_q = L_p^*$ by Lemma 2.1.1 (for $p = 1, 2$) and $\hat{\beta}_I^{H^{\hat{L}_I(f)}}$ denotes parameters of the linear regression associated to the hypercube $\text{argmax}_{H\in\mathcal{H}_I}\|\hat{\beta}^H\|_q$. By the arguments given at the beginning of Lemma 3.B.6, we have

$$\left|\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H \middle| G_I^{\mathcal{X}}\right]\|_q\} - \hat{L}_I(f)\right| \leq n_q \max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H \middle| G_I^{\mathcal{X}}\right] - \hat{\beta}_I^H\|_2\}$$

where we recall $n_q = d^{\max\{\frac{1}{q}-\frac{1}{2},0\}}$. Similarly, by construction of LCLS, the reverse triangle inequality and norm equivalence,

$$\left|\|\mathbb{E}\left[\hat{\beta}_I^{H^{\hat{L}_I(f)}} \middle| G_I^{\mathcal{X}}\right]\|_q - \hat{L}_I\right| = \left|\|\mathbb{E}\left[\hat{\beta}_I^{H^{\hat{L}_I(f)}} \middle| G_I^{\mathcal{X}}\right]\|_q - \|\hat{\beta}_I^{H^{\hat{L}_I(f)}}\|_q\right|$$

$$\leq n_q\|\mathbb{E}\left[\hat{\beta}_I^{H^{\hat{L}_I(f)}} \middle| G_I^{\mathcal{X}}\right] - \hat{\beta}_I^{H^{\hat{L}_I(f)}}\|_2 \leq n_q \max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H \middle| G_I^{\mathcal{X}}\right] - \hat{\beta}_I^H\|_2\}.$$

Therefore,

$$\mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon) \leq \underbrace{\mathbb{P}\left(\left|\|\nabla f(x^*)\|_q - \max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H \middle| G_I^{\mathcal{X}}\right]\|_q\}\right| > \frac{\epsilon}{4}\right)}_{\textbf{(i)}}$$

$$+\underbrace{\mathbb{P}\left(\left|\|\nabla f(x^{\hat{L}_I(f)})\|_q - \|\mathbb{E}\left[\hat{\beta}_I^{H^{\hat{L}_I(f)}} \middle| G_I^{\mathcal{X}}\right]\|_q\right| > \frac{\epsilon}{4}\right)}_{\textbf{(ii)}}$$

$$+2\,\underbrace{\mathbb{P}\left(\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\middle|G_I^{\mathcal{X}}\right]-\hat{\beta}_I^H\|_2\}>\frac{\epsilon}{4n_q}\right)}_{\textbf{(iii)}}.$$

The terms **(i)**, **(ii)** in the above expression can be shown to be equal to 0 with similar arguments. For **(i)**, Lemma 3.B.5 can be utilised (as $\mathcal{A}_\epsilon$ is assumed to hold) to obtain:

$$\left|\|\nabla f(x^*)\|_q-\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\middle|G_I^{\mathcal{X}}\right]\|_q\}\right|\leq\max_{H\in\mathcal{H}_I}\frac{4\sqrt{d}dKn_q}{\sqrt{\eta}}\frac{{\Delta_I^H}^2}{\delta_I^H}\leq\frac{4\sqrt{d}dn_qK}{\sqrt{\eta}}a_I$$

and it follows from applying the same approach as the one used in the proof of Lemma 3.B.5 (as $\mathcal{A}_\epsilon$ is assumed to hold), that

$$\left|\|\nabla f(x^{\hat{L}_I(f)})\|_q-\|\mathbb{E}\left[\hat{\beta}_I^{H^{\hat{L}_I(f)}}\middle|G_I^{\mathcal{X}}\right]\|_q\right|\leq\frac{4\sqrt{d}dKn_q}{\sqrt{\eta}}\frac{{\Delta_I^{H^{\hat{L}_I(f)}}}^2}{\delta_I^{H^{\hat{L}_I(f)}}}\leq\frac{4\sqrt{d}dn_qK}{\sqrt{\eta}}a_I.$$

Note that η is defined in the omitted conditioning on A_ϵ. Then, we have by definition of $I=\lceil C_1(d)\frac{MK}{\epsilon}\rceil$,

$$\frac{4\sqrt{d}dn_qK}{\sqrt{\eta}}a_I=\frac{4\sqrt{d}d^2n_qK}{\sqrt{\eta}}\frac{M}{I}\leq\frac{4\sqrt{d}d^2n_q}{\sqrt{\eta}}\frac{\epsilon}{C_1(d)}=\frac{\epsilon}{4}.$$

where the last line follows from the fact $C_1(d)=\frac{16d^2\sqrt{d}n_q}{\sqrt{\eta}}$. Therefore, conditional on $\mathcal{A}_\epsilon$, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\|\nabla f(x^*)\|_q-\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\middle|G_I^{\mathcal{X}}\right]\|_q\}\right|>\frac{\epsilon}{4}\right)\\&=\mathbb{P}\left(\left|\|\nabla f(x^{\hat{L}_I(f)})\|_q-\|\mathbb{E}\left[\hat{\beta}_I^{H^{\hat{L}_I(f)}}\middle|G_I^{\mathcal{X}}\right]\|_q\right|>\frac{\epsilon}{4}\right)=0.\end{aligned}$$

In order to show that **(iii)** converges to 0 as ϵ goes to 0, we define for all $H\in\mathcal{H}_I$: $E^H:=\{\|\mathbb{E}[\hat{\beta}_I^H\middle|G_I^{\mathcal{X}}]-\hat{\beta}_I^H\|_2\leq\frac{\epsilon}{4n_q}\}$, consider an arbitrary ordering of $\mathcal{H}_I:=\{H_1,...,H_{|\mathcal{H}_I|}\}$ and apply similar arguments as the ones utilised in the proof of Lemma 3.B.6 to obtain

$$\mathbb{P}\left(\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\middle|G_I^{\mathcal{X}}\right]-\hat{\beta}_I^H\|_2\}>\frac{\epsilon}{4n_q}\middle|\mathcal{A}_\epsilon\right)=1-\mathbb{P}(\forall H\in\mathcal{H}_I,\ E^H\middle|\mathcal{A}_\epsilon)$$

$$= 1 - \mathbb{P}\left(E^{H_1}\middle|\mathcal{A}_\epsilon\right)\prod_{i=2}^{|\mathcal{H}_I|}\mathbb{P}\left(E^{H_i}\middle|\mathcal{A}_\epsilon, E^{H_1}, ..., E^{H_{i-1}}\right).$$

The computation of the local linear regressions parameters is done independently with no data overlap implying that the conditioning expression: $\{E^{H_1}, ..., E^{H_{i-1}}\}$ (for $i = 1, ..., |\mathcal{H}_I|$) can only impact the probability by affecting the number of samples contained in H_i: $N_I^{H_i}$ which are utilised in the local linear regression. As the probabilities are also each conditioned on $\mathcal{A}_\epsilon$ which provides a fixed lower bound on N_I^H for all $H \in \mathcal{H}_I$ and the remaining arguments for this part of the proof will only utilise this fact, we use a slight abuse of notation in order to alleviate notation and omit the dependencies on $\{E^{H_1}, ..., E^{H_{i-1}}\}$ (for $i = 1, ..., |\mathcal{H}_I|$) in the remainder of this part of the proof. Therefore, we can consider

$$1 - \prod_{H\in\mathcal{H}_I}\left(1 - \mathbb{P}\left(\|\bar{\beta}_I^H - [\hat{b}_I^H, \hat{\beta}_I^H]^\top\|_2 \geq \frac{\epsilon}{4n_q}\middle|\mathcal{A}_\epsilon\right)\right).$$

In order to upper bound $\mathbb{P}(\|\bar{\beta}_I^H - [\hat{b}_I^H, \hat{\beta}_I^H]^\top\|_2 \geq \frac{\epsilon}{4n_q}|\mathcal{A}_\epsilon)$ a more refined bound than the general Chebyshev inequality used in the proof of Lemma 3.B.6 is utilised. Instead, we apply Corollary 3 of Pinelis and Sakhanenko [1986] to obtain an alternative bound (see **(iv)** below).

Remarking that $0 < \epsilon < \frac{C_1(d)MK}{3}$ implies $I = \left\lceil C_1(d)\frac{MK}{\epsilon}\right\rceil \leq \sqrt{2}C_1(d)\frac{MK}{\epsilon}$, we set $\bar{\epsilon}_1 := \frac{C_1(d)MK}{3}$. Then, for all $\epsilon \in (0, \bar{\epsilon}_1)$,

$$\mathbb{P}\left(\|\bar{\beta}_I^H - [\hat{b}_I^H, \hat{\beta}_I^H]^\top\|_2 \geq \frac{\epsilon}{4n_q}\middle|\mathcal{A}_\epsilon\right) = \mathbb{P}\left(\|(X_I^{H\top}X_I^H)^{-1}X_I^{H\top}\gamma_L^H\|_2 \geq \frac{\epsilon}{4n_q}\middle|\mathcal{A}_\epsilon\right)$$

$$\overset{\textbf{(iv)}}{\leq} 2e^{-\frac{(\frac{\epsilon}{4n_q})^2}{2\mathbb{E}[\|(X_I^{H\top}X_I^H)^{-1}X_I^{H\top}\gamma_L^H\|_2^2|\mathcal{A}_\epsilon]}}.$$

As the Gaussian vector $(X_I^{H\top}X_I^H)^{-1}X_I^{H\top}\gamma_L^H$ has covariance matrix $\sigma^2(X_I^{H\top}X_I^H)^{-1}$, we can utilise the tower rule to observe that

$$\mathbb{E}\left[\|(X_I^{H\top}X_I^H)^{-1}X_I^{H\top}\gamma_L^H\|_2^2\middle|\mathcal{A}_\epsilon\right] = \mathbb{E}\left[\mathbb{E}[\|(X_I^{H\top}X_I^H)^{-1}X_I^{H\top}\gamma_L^H\|_2^2|G_I^{\mathcal{X}}]\middle|\mathcal{A}_\epsilon\right]$$

$$= \mathbb{E}\left[\sigma^2 Tr((X_I^{H\top} X_I^H)^{-1})\Big|\mathcal{A}_\epsilon\right] \leq \mathbb{E}\left[d\sigma^2 \|(X_I^{H\top} X_I^H)^{-1}\|_2\Big|\mathcal{A}_\epsilon\right] \leq \frac{16d\sigma^2}{\eta\delta_I^{H^2}\bar{N}_I}$$

where $\bar{N}_I := C(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4}$ is given by $\mathcal{A}_\epsilon$ (with $C(d)$ explicitly determined below). The first inequality follows from the fact that the trace of a matrix is equal to the sum of its eigenvalues and the second inequality follows from Lemma 3.B.3 which can be applied by definition of $\mathcal{A}_\epsilon$. This implies:

$$\mathbb{P}\left(\|\bar{\beta}_I^H - [\hat{b}_I^H, \hat{\beta}_I^H]^\top\|_2 \geq \frac{\epsilon}{4n_q}\Big|\mathcal{A}_\epsilon\right) \leq 2e^{-\frac{\epsilon^2\eta\delta_I^{H^2}\bar{N}_I}{2^9 n_q^2 d\sigma^2}} = 2e^{-\frac{\epsilon^2\eta M^2\bar{N}_I}{2^9 n_q^2 I^2 d\sigma^2}} \leq 2e^{-\frac{\epsilon^4\eta\bar{N}_I}{2^{10}n_q^2K^2C_1(d)^2d\sigma^2}}.$$

Therefore, denoting $C_2(d) := \frac{\eta}{2^{10}n_q^2C_1(d)^2d}$ and substituting the above expression into the initial upper bound, we obtain

$$\mathbb{P}\left(Loss(x^{\hat{L}_I(f)}, f) > \epsilon\Big|\mathcal{A}_\epsilon\right) \leq 2\mathbb{P}\left(\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\Big|G_I^{\mathcal{X}}\right] - \hat{\beta}_I^H\|_2 > \frac{\epsilon}{4n_q}\Big|\mathcal{A}_\epsilon\right)$$

$$\leq 2 - 2\prod_{H\in\mathcal{H}_I}(1 - 2e^{-C_2(d)\frac{\epsilon^4\bar{N}_I}{\sigma^2K^2}}) = 2 - 2(1 - 2e^{-C_2(d)\frac{\epsilon^4\bar{N}_I}{\sigma^2K^2}})^{|\mathcal{H}_I|}.$$

Then, setting $2 - 2(1 - 2e^{-C_2(d)\frac{\epsilon^4\bar{N}_I}{\sigma^2K^2}})^{|\mathcal{H}_I|} \leq \delta$, we obtain that if

$$\bar{N}_I \geq \frac{\sigma^2K^2}{C_2(d)\epsilon^4}\log(\frac{2}{1 - \sqrt[|\mathcal{H}_I|]{1-\frac{\delta}{2}}})$$

then $\mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon\big|\mathcal{A}_\epsilon) \leq \delta$. As $|\mathcal{H}_I|$ is monotonically increasing and converges to infinity as ϵ goes to 0, there exists $\bar{\epsilon}_2 > 0$ such that $\forall\epsilon \in (0, \bar{\epsilon}_2)$, $|\mathcal{H}_I| \geq 2\log(\frac{1}{1-\frac{\delta}{2}})$. This implies that we can apply Lemma 3.B.2 to obtain:

$$\frac{2}{1 - \sqrt[|\mathcal{H}_I|]{1-\frac{\delta}{2}}} \leq \frac{4|\mathcal{H}_I|}{\log(\frac{1}{1-\frac{\delta}{2}})}.$$

Therefore, we have that the following stronger condition on $\bar{N}_I$ implies $\mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon) \leq \delta$:

$$\bar{N}_I \geq \frac{\sigma^2K^2}{C_2(d)\epsilon^4}\log(\frac{2}{1 - \sqrt[|\mathcal{H}_I|]{1-\frac{\delta}{2}}}) \impliedby \bar{N}_I \geq \frac{\sigma^2K^2}{C_2(d)\epsilon^4}\log(\frac{4|\mathcal{H}_I|}{\log(\frac{1}{1-\frac{\delta}{2}})}).$$

We can rewrite this lower bound with $|\mathcal{H}_I|$ expressed in terms of ϵ:

$$\frac{\sigma^2 K^2}{C_2(d)\epsilon^4}\log(\frac{4|\mathcal{H}_I|}{\log(\frac{1}{1-\frac{\delta}{2}})}) = \frac{\sigma^2 K^2}{C_2(d)\epsilon^4}\log(2^{d+2}C_1(d)^d(\frac{MK}{\epsilon})^d\log(\frac{1}{1-\frac{\delta}{2}})^{-1})$$

$$= \frac{\sigma^2 K^2}{C_2(d)\epsilon^4} d\log(C_3(d)\log(\frac{1}{1-\frac{\delta}{2}})^{-\frac{1}{d}}\frac{MK}{\epsilon})$$

where $C_3(d) := 2^{1+\frac{2}{d}}C_1(d)$. Finally, there exists $\bar{\epsilon}_3 > 0$ such that $\forall \epsilon \in (0, \bar{\epsilon}_3)$,

$$\frac{C_3(d)}{\log(\frac{1}{1-\frac{\delta}{2}})^{\frac{1}{d}}} \leq \frac{MK}{\epsilon}$$

which implies

$$\bar{N}_I \geq C^*(d)\frac{\sigma^2 K^2}{\epsilon^4}\log(\frac{MK}{\epsilon}) \implies \mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon) \leq \delta$$

where $C^*(d) := \frac{2d}{C_2(d)}$. Therefore, as $C^*(d)$ only depends on d (note that η depends only on d) we can set $C(d) = C^*(d)$ $(\star)$.

Selecting $\bar{\epsilon}^* := \min(\epsilon_1, \epsilon_2, \epsilon_3)$, we have $\forall \epsilon \in (0, \bar{\epsilon}^*)$

$$\sup_{f \in C^2(\mathcal{X}, K)} \mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon | \mathcal{A}_\epsilon) < \delta.$$

As the choice of $\delta > 0$ was arbitrary, this concludes the first part of the proof.

$(\star\star)$ We now show $\lim_{\epsilon \to 0^+} \mathbb{P}(\mathcal{A}_\epsilon^c) = 0$ with $C^*(d)$ as defined above in $(\star)$. Let $\epsilon \in (0, C_1(d)MK)$ be arbitrary and define the following events:

$$\mathcal{A}_\epsilon^1 := \left\{ \forall H \in \mathcal{H}_I : N_I^H \geq C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2 K^2}{\epsilon^4} \right\}$$

$$\mathcal{A}_\epsilon^2 := \left\{ \forall H \in \mathcal{H}_I : H \text{ is } (\frac{\delta_I^H}{8}, \eta)\text{-covered} \right\}.$$

We recall:

$$\mathcal{A}_\epsilon := \left\{ \forall H \in \mathcal{H}_I : N_I^H \geq C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2 K^2}{\epsilon^4} \wedge H \text{ is } (\frac{\delta_I^H}{8}, \eta)\text{-covered} \right\}$$

$$= \left\{ \mathcal{A}_\epsilon^1 \wedge \mathcal{A}_\epsilon^2 \right\}.$$

We can write:

$$\mathbb{P}(\mathcal{A}_\epsilon^c) = 1 - \mathbb{P}(\mathcal{A}_\epsilon) = 1 - \mathbb{P}(\mathcal{A}_\epsilon^2 | \mathcal{A}_\epsilon^1)\mathbb{P}(\mathcal{A}_\epsilon^1)$$

which is well defined as we will show that $\mathbb{P}(\mathcal{A}_\epsilon^1) > 0$ for sufficiently small ϵ. Therefore, if we can show that

$$\lim_{\epsilon \to 0^+} \mathbb{P}(\mathcal{A}_\epsilon^2 | \mathcal{A}_\epsilon^1) = \lim_{\epsilon \to 0^+} \mathbb{P}(\mathcal{A}_\epsilon^1) = 1$$

then the proof of $(\star\star)$ is concluded.

(I) We begin by showing $\lim_{\epsilon \to 0^+} \mathbb{P}(\mathcal{A}_\epsilon^1) = 1$. For all $H \in \mathcal{H}_I$, we define

$$\mathcal{E}_\epsilon^H(n) := \left\{ H \text{ contains } \geq C^*(d) \frac{\log(\frac{MK}{\epsilon})\sigma^2 K^2}{\epsilon^4} \text{ for total sample points equal to } n \right\}$$

where n denotes the total number of samples which was assumed to satisfy: $n \geq C \frac{\sigma^2 M^d K^{d+2} \log(\frac{MK}{\epsilon})}{\epsilon^{d+4}}$ for a fixed constant $C > 0$ (defined explicitly below). Then, considering an arbitrary ordering of $\mathcal{H}_I := \{H_1, ..., H_{|\mathcal{H}_I|}\}$, we have

$$\mathbb{P}(\mathcal{A}_\epsilon^1) = \mathbb{P}\left(\forall H \in \mathcal{H}_I, \mathcal{E}_\epsilon^H(n)\right) = \left(\mathcal{E}_\epsilon^{H_1}(n)\right) \prod_{i=2}^{|\mathcal{H}_I|} \mathbb{P}\left(\mathcal{E}_\epsilon^{H_i}(n) \middle| \mathcal{E}_\epsilon^{H_1}(n), ..., \mathcal{E}_\epsilon^{H_{i-1}}(n)\right).$$

As the inputs are sampled independently and the elements of $\mathcal{H}_I$ are disjointed by construction, we have $\forall i \in \{2, ..., |\mathcal{H}_I|\}$

$$\mathbb{P}\left(\mathcal{E}_\epsilon^{H_i}(n) \middle| \mathcal{E}_\epsilon^{H_1}(n), ..., \mathcal{E}_\epsilon^{H_{i-1}}(n)\right) = \mathbb{P}\left(\mathcal{E}_\epsilon^{H_i}(n - (i-1)C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2 K^2}{\epsilon^4})\right).$$

It trivial to see that $\forall i \in \{1, ..., |H_i|\}$, $\mathbb{P}(\mathcal{E}_\epsilon^{H_i}(n))$ is increasing in n. Thus, we have

$$\mathbb{P}(\mathcal{A}_\epsilon^1) \geq \prod_{i=1}^{|\mathcal{H}_I|} \mathbb{P}\left(\mathcal{E}_\epsilon^{H_i}(n - |\mathcal{H}_I|C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2 K^2}{\epsilon^4})\right).$$

For $\epsilon > 0$ satisfying $\epsilon < C_1(d)MK$, we have $I \leq 2C_1(d)\frac{MK}{\epsilon}$. Therefore, using

$|\mathcal{H}_I| = I^d \leq \frac{(2C_1(d)MK)^d}{\epsilon^d}$, we have

$$|\mathcal{H}_I|\frac{C^*(d)\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4} \leq \frac{(2C_1(d)MK)^d}{\epsilon^d}\frac{C^*(d)\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4}$$

$$= (2C_1(d))^dC^*(d)\frac{M^dK^{d+2}\log(\frac{MK}{\epsilon})}{\epsilon^{d+4}} \leq \frac{(2C_1(d))^dC^*(d)}{C}n$$

where the last inequality follows from the theorem assumption: $n \geq C\frac{\sigma^2M^dK^{d+2}\log(\frac{MK}{\epsilon})}{\epsilon^{d+4}}$. Therefore, defining $\bar{C}_1 := 2(2C_1(d))^dC^*(d)$, setting $C \geq \bar{C}_1$, and utilising the upper bound derived above, we obtain

$$\prod_{i=1}^{|\mathcal{H}_I|}\mathbb{P}\left(\mathcal{E}_\epsilon^{H_i}(n - |\mathcal{H}_I|C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4})\right) \geq \prod_{i=1}^{|\mathcal{H}_I|}\mathbb{P}\left(\mathcal{E}_\epsilon^{H_i}(n - \frac{n}{2})\right) = \prod_{i=1}^{|\mathcal{H}_I|}\mathbb{P}\left(\mathcal{E}_\epsilon^{H_i}(\frac{n}{2})\right).$$

We now consider for all $H \in \mathcal{H}_I$ the computation of $\mathbb{P}(\mathcal{E}_\epsilon^H(\frac{n}{2}))$ for which we will derive a lower bound.

For all $H \in \mathcal{H}_I$, denote $M_I^H(n) := |\{i \in \{0, ..., n\} : x_i \in H\}|$ the random variable[19] that counts the number of sample inputs in H. As every sample input is sampled uniformly on $\mathcal{X}$ and for all $H \in \mathcal{H}_I$ $vol(H) = (\frac{M}{I})^d \geq \frac{\epsilon^d}{(2C_1(d)K)^d}$, we have that the probability of a sample input being in $H \in \mathcal{H}_I$ can be modelled using a Bernouilli random variable with success probability $p = \frac{vol(H)}{vol(\mathcal{X})} \geq \frac{\epsilon^d}{(2C_1(d)MK)^d}$. Therefore, $M_I^H(n)$ can be modelled as a sum of independent Bernoulli variables with success probability p. From Lemma 1 of Stone [1982], we have

$$\mathbb{P}(M_I^H(n) \leq \frac{\mathbb{E}[M_I^H(n)]}{2}) \leq (\frac{2}{e})^{\frac{\mathbb{E}[M_I^H(n)]}{2}}.$$

In order to apply this result, we observe that as $C \geq \bar{C}_1 = 2(2C_1(d))^dC^*(d)$ (by construction), the following relations hold:

$$\mathbb{E}[M_I^H(\frac{n}{2})] = \frac{n}{2}p \geq \frac{\sigma^2K^2\log(\frac{MK}{\epsilon})}{\epsilon^4}\frac{C}{C_1(d)^d2^{d+1}} \geq C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4}$$

where the rightmost term corresponds to the bound stated in the definition of the

[19] In essence, $M_I^H(n) = N_I^H$ but makes explicit the dependency on n.

$\mathcal{E}_\epsilon^H(\frac{n}{2})$ events. This implies that $\mathbb{P}(\mathcal{E}_\epsilon^H(\frac{n}{2}))$ can be lower bounded as follows:

$$\mathbb{P}\left(\mathcal{E}_\epsilon^H(\frac{n}{2})\right) = \mathbb{P}\left(M_I^H(\frac{n}{2}) \geq C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4}\right)$$

$$= 1 - \mathbb{P}\left(M_I^H(\frac{n}{2}) \leq C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4}\right) \geq 1 - \mathbb{P}\left(M_I^H(\frac{n}{2}) \leq \frac{\mathbb{E}[M_I^H(\frac{n}{2})]}{2}\right)$$

$$\geq 1 - (\frac{2}{e})^{\frac{\mathbb{E}[M_I^H(\frac{n}{2})]}{2}}.$$

Plugging this expression into the initial bound, we obtain:

$$\prod_{i=1}^{|\mathcal{H}_I|} \mathbb{P}\left(\mathcal{E}_\epsilon^{H_i}(\frac{n}{2})\right) \geq \prod_{i=1}^{|\mathcal{H}_I|}\left(1 - (\frac{2}{e})^{\frac{\mathbb{E}[M_I^H(\frac{n}{2})]}{2}}\right).$$

As $\mathbb{E}[M^H(\frac{n}{2})] \geq \frac{\sigma^2K^2\log(\frac{MK}{\epsilon})}{\epsilon^4}\frac{C}{C_1(d)^d 2^{d+1}}$ and $|\mathcal{H}_I| = I^d$ (i.e. both terms increase polynomially with respect to $\frac{1}{\epsilon}$), the above expression can be shown to go to 1 as ϵ goes to 0. This implies that if $C \geq \bar{C}_1$, then

$$\lim_{\epsilon \to 0^+} \mathbb{P}(\mathcal{A}_\epsilon^1) = 1.$$

(II) We now show that $\lim_{\epsilon\to 0^+} \mathbb{P}(\mathcal{A}_\epsilon^2|\mathcal{A}_\epsilon^1) = 1$.

By the law of total probability, we can derive:

$$\mathbb{P}(\mathcal{A}_\epsilon^2|\mathcal{A}_\epsilon^1) = \sum_{\{\bar{N}_I^H\}_{H\in\mathcal{H}_I} \in V_I(n)} \mathbb{P}\left(\mathcal{A}_\epsilon^2|\{\bar{N}_I^H\}_{H\in\mathcal{H}_I}\right)\mathbb{P}\left(\{N_I^H\}_{H\in\mathcal{H}_I} = \{\bar{N}_I^H\}_{H\in\mathcal{H}_I}\middle|\mathcal{A}_\epsilon^1\right)$$

$$= \sum_{\{\bar{N}_I^H\}_{H\in\mathcal{H}_I} \in V_I(n)} \left(\prod_{H\in\mathcal{H}_I} \mathbb{P}(H \text{ is } (\frac{\delta_I^H}{8}, \eta)\text{-covered}|\bar{N}_I^H)\right)\mathbb{P}\left(\{N_I^H\}_{H\in\mathcal{H}_I} = \{\bar{N}_I^H\}_{H\in\mathcal{H}_I}\middle|\mathcal{A}_\epsilon^1\right)$$

where $V_I(n) := \{\{\bar{N}_I^H\}_{H\in\mathcal{H}_I} \in \mathbb{N}_{\geq c(\epsilon)}{}^{|\mathcal{H}_I|} : \sum_{H\in|\mathcal{H}_I|} N_I^H = n\}$ with $c(\epsilon) := C^*(d)\frac{\log(\frac{MK}{\epsilon})\sigma^2K^2}{\epsilon^4}$ defined as the bound stated in $\mathcal{A}_\epsilon^1$. The second equality follows from the definition of $(\frac{\delta_I^H}{8}, \eta)$-covered and the disjointness of the partition. Note that $V_I(n)$ is non-empty, i.e. $n \geq c(\epsilon)|\mathcal{H}_I|$, due to the inequality: $C \geq \bar{C}_1$ set in **(I)**.

For an arbitrary $\{\bar{N}_I^H\}_{H\in\mathcal{H}_I} \in V_I(n)$, we now focus on lower bounding $\prod_{H\in\mathcal{H}_I}\mathbb{P}(H$ is $(\frac{\delta_I^H}{8},\eta)$-covered $|\bar{N}_I^H)$.

In order to do so, for each $H \in \mathcal{H}_I$ we define $\mathcal{C}_I^H$: the minimal cover of H with balls of radius $\frac{\delta_I^H}{8}$ with respect to $\|.\|_2$ and the associated set of hyperballs: $\mathcal{B}_I^H$.

Let $I \in \mathbb{N}$, $H \in \mathcal{H}_I$ be arbitrary, without loss of generality, we can assume that $\mathcal{C}_I^H \subset H$ as H is a hypercube. This implies that for all $B \in \mathcal{B}_I^H$, $vol(B \cap H) \geq 2^{-d}vol(B) = 2^{-d}vol(B_1(0))(\frac{\delta_I^H}{8})^d$ where $vol(B_1(0))$ corresponds to the volume of the unit ball and is a constant[20] that depends on d. Utilising $\{\mathcal{B}_I^H\}_{H\in\mathcal{H}_I}$, we construct the set $\tilde{\mathcal{B}}_I^H$ as follows

$$\tilde{\mathcal{B}}_I^H := \left\{\tilde{B}_I^H \subset H : \exists B_I^H \in \mathcal{B}_I^H \text{ such that } \tilde{B}_I^H = H \cap B_I^H\right\}.$$

We have $\bigcup_{B\in\tilde{\mathcal{B}}_I^H} B = H$, $|\tilde{\mathcal{B}}_I^H| = |\mathcal{C}_I^H|$ and for all $B \in \tilde{\mathcal{B}}_I^H$, $vol(B) \geq vol(B_1(0))(\frac{\delta_I^H}{16})^d$. Furthermore, by Theorem 14.2 of Wu [2017], we have

$$|\tilde{\mathcal{B}}^H| = |\mathcal{C}_I^H| \leq (\frac{2^3 3}{\delta_I^H})^d \frac{{\delta_I^H}^d}{vol(B_1(0))} = \frac{2^{3d}3^d}{vol(B_1(0))}.$$

Clearly, if each set in $\tilde{\mathcal{B}}_I^H$ contains ηN_I^H sample inputs, then H is $(\frac{\delta_I^H}{8},\eta)$-covered.

For all $B^H \in \tilde{\mathcal{B}}^H$, we define the event:

$$\mathcal{E}^{B^H}(N) := \{B^H \text{ contains } \geq \eta N \text{ inputs if } N \text{ samples are in } H.\}.$$

Then, we can apply the same arguments as the ones given in (**I**) to obtain:

$$\mathbb{P}\left(H \text{ is } (\frac{\delta_I^H}{8},\eta)\text{-covered}|\bar{N}_I^H\right) \geq \mathbb{P}\left(\forall B^H \in \tilde{\mathcal{B}}^H : \mathcal{E}^{B^H}(\bar{N}_I^H)\right)$$

$$\geq \prod_{B^H\in\tilde{\mathcal{B}}^H} \mathbb{P}\left(\mathcal{E}^{B^H}(\bar{N}_I^H - \eta|\tilde{\mathcal{B}}^H|\bar{N}_I^H)\right) \geq \prod_{B^H\in\tilde{\mathcal{B}}^H} \mathbb{P}\left(\mathcal{E}^{B^H}\left(\bar{N}_I^H(1-\eta\frac{2^{3d}3^d}{vol(B_1(0))})\right)\right)$$

$$= \prod_{B^H\in\tilde{\mathcal{B}}^H} \mathbb{P}\left(\mathcal{E}^{B^H}(\frac{1}{2}\bar{N}_I^H)\right)$$

[20] In fact, a closed form for $vol(B_1(0))$ is known, $vol(B_1(0)) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$.

where the last equality follows from the fact that by assumption, $\eta = \frac{vol(B_1(0))}{2^{3d+1}3^d}$. Following the same approach as in (**I**): we consider the random variables $M^{B^H}(N) := |\{i \in \{1, ..., N\} : x_i \in B^H\}|$ where x_i are samples that are selected uniformly on H. For all $B^H \in \tilde{\mathcal{B}}^H$, $M^{B^H}(N)$ can be modelled as the sum of independent Bernouilli variables with success probability: $p = \frac{vol(B^H)}{vol(H)} \geq \frac{vol(B_1(0))(\frac{\delta_I^H}{16})^d}{{\delta_I^H}^d} = \frac{vol(B_1(0))}{16^d}$ and satisfying

$$\mathbb{E}\left[M^{B^H}(\frac{\bar{N}_I^H}{2})\right] = \frac{\bar{N}_I^H}{2}p \geq \bar{N}_I^H \frac{vol(B_1(0))}{2^{4d+1}}.$$

Using this inequality, we observe:

$$\eta\bar{N}_I^H = \frac{vol(B_1(0))}{2^{3d+1}3^d}\bar{N}_I^H \leq \bar{N}_I^H \frac{vol(B_1(0))}{2^{4d+1}} \leq \mathbb{E}[M^{B^H}(\frac{\bar{N}_I^H}{2})].$$

Therefore, leveraging the same arguments as the ones utilised in (**I**), we can apply Lemma 1 of Stone [1982] to obtain:

$$\mathbb{P}\left(\mathcal{E}^{B^H}(\frac{1}{2}\bar{N}_I^H)\right) \geq 1 - (\frac{2}{e})^{\bar{N}_I^H \frac{vol(B_1(0))}{2^{4d+2}}}.$$

By construction, we have that for all $\epsilon > 0$ and $I = I(\epsilon) \in \mathbb{N}$, $\{\bar{N}_I^H\}_{H\in\mathcal{H}_I} \in V_I(n)$ which implies that for all $H \in \mathcal{H}_I$, $\bar{N}_I^H \geq c(\epsilon)$. Combining this bound with the lower bound derived above, we obtain for all $H \in \mathcal{H}_I$:

$$\mathbb{P}(H \text{ is } \left(\frac{\delta_I^H}{8}, \eta\right)\text{-covered}|\bar{N}_I^H\Big) \geq \prod_{B^H\in\tilde{\mathcal{B}}^H} \mathbb{P}\left(\mathcal{E}^{B^H}(\frac{1}{2}\bar{N}_I^H)\right) \geq \prod_{B^H\in\tilde{\mathcal{B}}^H} 1-(\frac{2}{e})^{\bar{N}_I^H \frac{vol(B_1(0))}{2^{4d+2}}}$$

$$\geq \left(1 - (\frac{2}{e})^{\bar{c}(\epsilon)\frac{vol(B_1(0))}{2^{4d+2}}}\right)^{|\tilde{\mathcal{B}}^H|} \geq \left(1 - (\frac{2}{e})^{\bar{c}(\epsilon)\frac{vol(B_1(0))}{2^{4d+2}}}\right)^{\frac{2^{3d}3^d}{vol(B_1(0))}}.$$

As the lower bound derived above does not depend on $\{\bar{N}_I^H\}_{H\in\mathcal{H}_I} \in V_I(n)$, we plug it into the initial expression to obtain

$$\sum_{\{\bar{N}_I^H\}_{H\in\mathcal{H}_I}\in V_I(n)} \left(\prod_{H\in\mathcal{H}_I} \mathbb{P}(H \text{ is } (\frac{\delta_I^H}{8}, \eta)\text{-covered}|\bar{N}_I^H)\right) \mathbb{P}(\{N_I^H\}_{H\in\mathcal{H}_I} = \{\bar{N}_I^H\}_{H\in\mathcal{H}_I}|\mathcal{A}_\epsilon^1)$$

$$\geq \prod_{H\in\mathcal{H}_I}\left(1-(\frac{2}{e})^{\bar{c}(\epsilon)\frac{vol(B_1(0))}{2^{4d+2}}}\right)^{\frac{2^{3d}3^d}{vol(B_1(0))}} \sum_{\{\bar{N}_I^H\}_{H\in\mathcal{H}_I}\in V_I(n)} \mathbb{P}(\{N_I^H\}_{H\in\mathcal{H}_I}=\{\bar{N}_I^H\}_{H\in\mathcal{H}_I}\big|\mathcal{A}_\epsilon^1)$$

$$= \prod_{H\in\mathcal{H}_I}\left(1-(\frac{2}{e})^{\bar{c}(\epsilon)\frac{vol(B_1(0))}{2^{4d+2}}}\right)^{\frac{2^{3d}3^d}{vol(B_1(0))}} \geq \left(1-(\frac{2}{e})^{\bar{c}(\epsilon)\frac{vol(B_1(0))}{2^{4d+2}}}\right)^{\frac{2^{4d}3^d(C_1(d)MK)^d}{vol(B_1(0))\epsilon^d}}$$

where the last inequality follows from: for all $\epsilon > 0$ satisfying $\epsilon < C_1(d)MK$, we have $I \leq 2C_1(d)\frac{MK}{\epsilon}$ implying $|\mathcal{H}_I| = I^d \leq \frac{(2C_1(d)MK)^d}{\epsilon^d}$. It is relatively straightforward to see that the lower bound derived above converges to 1 as ϵ goes to 0. Therefore:

$$1 \geq \lim_{\epsilon\to 0^+}\mathbb{P}(\mathcal{A}_\epsilon^2|\mathcal{A}_\epsilon^1) \geq \lim_{\epsilon\to 0^+}\left(1-(\frac{2}{e})^{\bar{c}(\epsilon)\frac{vol(B_1(0))}{2^{4d+2}}}\right)^{\frac{2^{4d}3^d(C_1(d)MK)^d}{vol(B_1(0))\epsilon^d}} = 1.$$

This shows:

$$\lim_{\epsilon\to 0^+}\mathbb{P}(\mathcal{A}_\epsilon^c) = 1 - \lim_{\epsilon\to 0^+}\mathbb{P}(\mathcal{A}_\epsilon^2|\mathcal{A}_\epsilon^1)\lim_{\epsilon\to 0^+}\mathbb{P}(\mathcal{A}_\epsilon^1) = 1 - 1\cdot 1 = 0$$

and concludes the proof of Theorem 3.3.15.

■

Proof of Corollary 3.3.16 (Finite Sample Guarantee – Gaussian Noise).

The assumptions of Corollary 3.3.16 imply that the event $\mathcal{A}_\epsilon$ defined in the proof of Theorem 3.3.15 holds with constants specified in the statement of the corollary: $\eta \in (0,1)$ and $C_1(d) = \frac{16d^2\sqrt{d}d^{\max\{\frac{1}{q}-\frac{1}{2},0\}}}{\sqrt{\eta}}$.

Therefore, the same arguments as the ones used in the first part of the proof of Theorem 3.3.15 can be applied in order to obtain $\forall\epsilon \in (0, \frac{C_1(d)MK}{3})$:

$$\sup_{f\in C^2(\mathcal{X},K)}\mathbb{P}(|\hat{L}_I(f) - L_p^*(f)| > \epsilon) \leq \mathbb{P}\left(\max_{H\in\mathcal{H}_I}\{\|\mathbb{E}\left[\hat{\beta}_I^H\middle|G_I^{\mathcal{X}}\right] - \hat{\beta}_I^H\|_2\} > \frac{\epsilon}{4n_q}\right)$$

where we note that as we consider $\mathbb{P}(|\hat{L}_I(f) - L_p^*(f)| > \epsilon)$ instead of $\mathbb{P}(Loss(x^{\hat{L}_I(f)}, f) > \epsilon)$, a factor 2 disappears.

This implies that we can consider the statement $\forall \delta \in (0, \frac{1}{2})$:

$$\mathbb{P}\left(\max_{H \in \mathcal{H}_I} \{\|\mathbb{E}\left[\hat{\beta}_I^H \middle| G_I^{\mathcal{X}}\right] - \hat{\beta}_I^H\|_2\} > \frac{\epsilon}{4n_q}\right) \leq \delta$$

in order to show Corollary 3.3.16.

Let $\epsilon \in (0, \frac{C_1(d)MK}{3})$ and $\delta \in (0, \frac{1}{2})$ be arbitrary. Again applying the arguments of Theorem 3.3.15, with δ, we obtain that the above expression holds if $\epsilon \leq \bar{\epsilon}^* := \min(\epsilon_1, \epsilon_2, \epsilon_3)$ where $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ depend on δ and ϵ. We consider each of the epsilon separately:

$\underline{(\epsilon_1)}$. By construction, $\epsilon_1 = \frac{C_1(d)MK}{3}$ and by assumption: $\epsilon \in (0, \frac{C_1(d)MK}{3})$. Therefore, $\epsilon < \epsilon_1$ holds.

$\underline{(\epsilon_2)}$. ϵ_2 is set such that the relation: $|\mathcal{H}_I| \geq 2\log(\frac{1}{1-\delta})$ holds (in order to apply Lemma 3.B.2). Substituting $|\mathcal{H}_I| = I^d$ and $\delta \leq \frac{1}{2}$ into the expression yields:

$$|\mathcal{H}_I| \geq 2\log(\frac{1}{1-\delta}) \impliedby I^d \geq 2\log(2).$$

As $\epsilon \in (0, \frac{C_1(d)MK}{3})$ and I is defined to be $I = \lceil C_1(d)\frac{MK}{\epsilon} \rceil$, $I \geq 3$ and the above holds. Therefore, $\epsilon < \epsilon_2$ holds.

$\underline{(\epsilon_3)}$. ϵ_3 is defined such that the following relation holds:

$$\forall H \in \mathcal{H}_I,\ \bar{N}_I^H \geq \frac{\sigma^2 K^2}{C_2(d)\epsilon^4} \log(\frac{4|\mathcal{H}_I|}{\log(\frac{1}{1-\delta})})$$

where $\bar{N}_I^H$ is guaranteed number of samples in $H \in \mathcal{H}_I$. By the assumptions of Corollary 3.3.16, we have $\bar{N}_I^H = \tilde{C}^*(\eta, d)\frac{\sigma^2 K^2}{\epsilon^4} \log(\frac{4^{\frac{1}{d}} I}{\log(\frac{1}{1-\delta})^{\frac{1}{d}}})$ for all $H \in \mathcal{H}_I$. By construction, $\tilde{C}^*(d) = \frac{2^{10} n_q^2 C_1(d)^2 d^2}{\eta} = \frac{d}{C_2(d)}$. Therefore:

$$\bar{N}_I^H \geq \frac{\sigma^2 K^2}{C_2(d)\epsilon^4} \log(\frac{4|\mathcal{H}_I|}{\log(\frac{1}{1-\delta})}) \iff \bar{N}_I^H \geq \frac{d\sigma^2 K^2}{C_2(d)\epsilon^4} \log(\frac{4^{\frac{1}{d}} I}{\log(\frac{1}{1-\delta})^{\frac{1}{d}}})$$

holds by design for all $\epsilon \in (0, \frac{C_1(d)MK}{3})$. Therefore, $\epsilon < \epsilon_3$ holds.

Thus, we have shown: $\forall \epsilon \in (0, \frac{C_1(d)MK}{3})$, $\delta \in (0, \frac{1}{2})$

$$\sup_{f \in C^2(\mathcal{X}, K)} \mathbb{P}(|\hat{L}_I(f) - L_p^*(f)| > \epsilon) \leq \delta.$$

■

Appendix 3.C Proofs: Sample Complexity of Adaptive Lipschitz Optimisation

In this section we prove the lower bound on the sample complexity of certified adaptive Lipschitz optimisation algorithms given in Section 3.4.

Proof of Proposition 3.4.1 (Sample Complexity of Adaptive Lipschitz Optimisation).

Fix $\epsilon > 0$, $L^* \geq 0$ and let A be a non-adaptive certified optimisation algorithm which takes a given Lipschitz constant $\bar{L} > L^*$ as a hyperparameter. Using the notation given in Section 3.4: with n-queries to the oracle, A outputs a triplet $((x_n, f(x_n^*), \zeta_n))_{n \in \mathbb{N}}$ where x_n is the n-th query point, $f(x_n^*)$ is the generated estimate of $\max_{x \in \mathcal{X}} f(x)$ after n queries and $\zeta_n \geq 0$ is an error certificate that guarantees: $\max_{x \in \mathcal{X}} f(x) - f(x_n^*) \leq \zeta_n$. From Theorem 3 of Bachoc et al. [2021] with $\epsilon_0 < 2^{d-1} M L^*$ (this follows from the fact that $\mathcal{X}$ is a hypercube), we have that for all $f \in \{h : \mathcal{X} \to | h \text{ is Lipschitz cont. and } L_p^*(h) < \bar{L}\}$:

$$N(A, f, \epsilon) \geq \frac{c_d L^{*d} (1 - \frac{L^*}{\bar{L}})^d}{1 + \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil} \int_{\mathcal{X}} \frac{dx}{(f(x^*) - f(x) + \epsilon)^d}. \tag{3.5}$$

where $c_d > 0$ (It is important to note that the term $c_d L^{*d}$ is not optimised in Bachoc et al. [2021] and could be improved in future work). Now, consider an adaptive Lipschitz optimisation algorithm $\tilde{A}$ with a separable Lipschitz constant estimator $\tilde{L}_{\tilde{A}}(f)$. If $\tilde{L}_{\tilde{A}}(f)$ can be guaranteed to be feasible (e.g. see discussion after Corollary 3.3.13) then equation (3.5) holds for $\tilde{A}$ and $\forall f \in C^2(\mathcal{X}, K) \cap \mathcal{F}_p(L^*)$ with

$\bar{L}$ replaced by $\tilde{L}_{\tilde{A}}(f)$ [21]. The precision at which $\tilde{L}_{\tilde{A}}(f)$ estimates $L^*(f)$ therefore directly impacts the lower bound on $N(\tilde{A}, f, \epsilon)$. From the Corollary 3.2.4 given in Section 3.2, we have that $\forall n \in \mathbb{N}$, any Lipschitz learning algorithm $\tilde{L} \in \mathcal{L}_{n,p}$ that guarantees feasible Lipschitz constants must satisfy

$$\sup_{f \in C^2(\mathcal{X},K) \cap \mathcal{F}_p(L^*)} \tilde{L}(f) - L^* \geq C \frac{MK}{\sqrt[d]{n}}.$$

for some $C > 0$. This implies that for all $A \in \mathcal{A}$, there exists a non-empty set $\mathcal{G}_A \subset C^2(\mathcal{X}, K) \cap \mathcal{F}_p(L^*)$ such that $\forall f^* \in \mathcal{G}_{\mathcal{A}}$, $\tilde{L}_A(f^*) - L^* \geq \frac{C}{2} \frac{MK}{\sqrt[d]{n}}$. Then, denoting $I(f) := \frac{c_d L^{*d}}{1+\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil} \int_{\mathcal{X}} \frac{dx}{(f(x^*)-f(x)+\epsilon)^d}$ in order to alleviate notation, we have $\forall A \in \mathcal{A}$,

$$N(A,\epsilon) := \sup_{f \in C^2(\mathcal{X},K) \cap \mathcal{F}_p(L^*)} N(A, f, \epsilon) \geq \sup_{f \in C^2(\mathcal{X},K) \cap \mathcal{F}_p(L^*)} \left\{ (1 - \frac{L^*}{\tilde{L}_A(f)})^d I(f) \right\}$$

$$\geq \left(1 - \frac{L^*}{L^* + \frac{C}{2} \frac{MK}{\sqrt[d]{N(A,\epsilon)}}} \right)^d \sup_{f \in \mathcal{G}_A} \left\{ I(f) \right\}.$$

Re-arranging the terms in the above expression, we can obtain:

$$\frac{C}{2} MK \sup_{f \in \mathcal{G}_A} \left\{ \sqrt[d]{I(f)} \right\} \leq L^* (\sqrt[d]{N(A,\epsilon)})^2 + \frac{C}{2} MK \sqrt[d]{N(A,\epsilon)}$$

which can be solved to give the lower bound

$$\sqrt[d]{N(A,\epsilon)} \geq C_1 \frac{MK}{L^*} \left(\sqrt{1 + C_1 \frac{L^* \sup_{f \in \mathcal{G}_A} \left\{ \sqrt[d]{I(f)} \right\}}{MK}} - 1 \right)$$

where $C_1 > 0$ is a constant. In order to conclude the proof, a lower bound on $\sup_{f \in \mathcal{G}_A} \left\{ \sqrt[d]{I(f)} \right\}$ is needed. To do so, we note that $I(f)$ is minimised when f is constant. We therefore consider the set of functions $\mathcal{F}_0$ defined in the proof of Theorem 3.2.3. From the proof of Theorem 3.2.3, we have that if $N(A,\epsilon) \leq (\frac{MK}{L^*})^d (\frac{C_2}{2})^d$, then $L^* \leq \frac{C_2}{2} \frac{MK}{\sqrt[d]{N(A,\epsilon)}}$ which implies $\mathcal{F}_0(\frac{(L^*)^2}{0.8K}, \frac{0.8}{7.75}(\frac{K}{L^*})^2) \subset \mathcal{G}_A$. Using

[21] Note: this is only possible as we are considering adaptive Lipschitz optimization algorithms which are separable.

$f(x^*) = \frac{(L^*)^2}{0.8K}$, $\forall f \in \mathcal{F}_0(\frac{(L^*)^2}{0.8K}, \frac{0.8}{7.75}(\frac{K}{L^*})^2)$, we obtain the lower bound:

$$\sup_{f \in \mathcal{G}_A} \{I(f)\} \geq \frac{c_d L^{*d}}{1 + \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil} \frac{\mathcal{V}_{\mathcal{X}}}{(\epsilon + \frac{(L^*)^2}{0.8K})^d}.$$

Therefore, if $N(A, \epsilon) \leq (\frac{MK}{L^*})^d (\frac{C_2}{2})^d$, the above expression can be plugged into the lower bound on $\sqrt[d]{N(A, \epsilon)}$. We obtain

$$\sqrt[d]{N(A, \epsilon)} \geq C_1 \frac{MK}{L^*} \left(\sqrt{1 + C_3 \frac{1}{\sqrt[d]{(1 + \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil)(\frac{\epsilon K}{L^{*2}} + 1)}}} - 1 \right)$$

(for some constant $C_3 > 0$) which corresponds to the first half of the lower bound stated in the Proposition 3.4.1. In order to derive the second part of the expression, we consider the case where $N(A, \epsilon) > (\frac{MK}{L^*})^d (\frac{C_2}{2})^d$. In this case, an alternative lower bound on $\sup_{f \in \mathcal{G}_A} \{I(f)\}$ needs to be derived. In order to do so, we consider the following class of functions,

$$\{g : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R} | \forall x \in \mathcal{X}, g(x) = f(x) + (L^* - \frac{C_2}{2} \frac{MK}{\sqrt[d]{N(A, \epsilon)}}) x_1$$

$$\text{where } f \in \mathcal{F}_0 \left(\frac{(L^*)^2}{0.8K}, \frac{0.8}{7.75} (\frac{K}{L^*})^2 \right) \}$$

which belongs to $\mathcal{G}_A$ by construction. However, as obtaining a tight lower bound on $\sup_{f \in \mathcal{G}_A} \{\sqrt[d]{I(f)}\}$ is technically infeasible for this class, we simplify the problem by removing the functional input from $\mathcal{F}_0(\frac{(L^*)^2}{0.8K}, \frac{0.8}{7.75}(\frac{K}{L^*})^2)$ and considering the simple linear function $f^* : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$, $f^*(x) = L^* x_1$ which belongs trivially to $\mathcal{G}_A$. In this case, we can compute the lower bound

$$\sup_{f \in \mathcal{G}_A} \{I(f)\} \geq c_d \frac{L^{*d+1} M^{d-1}}{(1 + \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil) \epsilon^{d-1}} (d-1) \frac{(\frac{LM}{\epsilon} + 1)^{d-1} - 1}{(\frac{LM}{\epsilon} + 1)^{d-1}}$$

$$\geq \frac{c_d (d-1)}{2} \frac{L^{*d+1} M^{d-1}}{(1 + \lceil log_2(\frac{\epsilon_0}{\epsilon}) \rceil) \epsilon^{d-1}}$$

where the last inequality follows from the fact that $LM \geq \epsilon$ since $\epsilon \in (0, \epsilon_0)$.

Plugging this expression into the lower bound on $\sqrt[d]{N(A,\epsilon)}$, we obtain

$$\sqrt[d]{N(A,\epsilon)} \geq C_1 \frac{MK}{L^*}\left(\sqrt{1 + C_4 \frac{L^{*2}}{\epsilon K}\sqrt[d]{\frac{L^*(d-1)\epsilon}{M(1+\lceil log_2(\frac{\epsilon_0}{\epsilon})\rceil)}}} - 1\right)$$

(for some constant $C_4 > 0$) which corresponds to the second half of the lower bounding expression. Note: in the statement of the proposition we simply set $C_2 = \min(C_3, C_4)$ as the used constant.

■

4 | Lipschitz Interpolation: Asymptotic Analysis

Contents

4.1 Introduction

Given the growing use of Lipschitz interpolation frameworks in control (Mesbah et al. [2022]), obtaining a strong theoretical understanding of this method is essential. While, as discussed in Chapter 2, several finite sample guarantees and worst-case error bounds already exist (see in particular Milanese and Novara [2004] and Calliess et al. [2020]), few asymptotic results have been derived and, to the best of our knowledge, almost none under stochastic noise. By contrast, numerous

asymptotic guarantees and convergence rates have been obtained for other popular non-parametric methods. In particular, for alternative safe-learning frameworks based on Gaussian processes, both the pointwise convergence of the posterior mean function (Seeger et al. [2008], Yang et al. [2017]) and the contraction rate of the posterior distribution, which provide a measure of uncertainty quantification (van der Vaart and van Zanten [2008], Van Der Vaart and Van Zanten [2011]), of the fitted Gaussian processes have been derived.

These types of asymptotic properties are crucial for adaptive control applications as they guarantee that the learned dynamics and error bounds accurately converge to the true underlying system dynamics while also providing a characterisation of the long-run performance of the regression method. This in turn ensures that the controllers built on these data-driven frameworks become increasingly more successful the longer the interaction with the underlying plant progresses. Considering the computational advantages of Lipschitz interpolation over Gaussian process regression (Calliess et al. [2020]), deriving analogous asymptotic guarantees for Lipschitz interpolation is therefore strongly desirable and constitutes the main motivation of this chapter. Specifically, the following contributions to the literature are made:

- In the case of independent input sampling, general consistency and upper bounds on the asymptotic convergence rates are obtained for both the prediction function (Theorem 4.3.5) and the worst-case error bounds (Corollary 4.3.6) of the general Lipschitz constant interpolation framework. While convergence lower bounds do not exist for the exact setting considered in this chapter signifying that the optimality of our bounds is not (yet) established, the obtained rates are consistent with the optimal convergence rates for non-parametric regression in related settings; e.g. with the classical convergence rate results derived by (Stone [1982]).

- In the case of discrete-time non-linear and noisy dynamical systems, we show that the Lipschitz interpolation framework and worst-case bounds converge point-wise in moments (Corollary 4.4.1 and ensuing discussion) and that, under an additional sampling assumption, the convergence rates match the ones

derived in the first part of the chapter. The first result can be directly applied in the context of the existing non-linear controllers discussed above (e.g. Canale et al. [2014], Manzano et al. [2020]) and we provide a theoretical illustration in the context of online learning-based trajectory tracking control (see Section 4.6).

- In a general sampling setting, probabilistic consistency is shown (Theorem 4.5.3) for the fully data-driven LACKI (*Lazily Adapted Constant Kinky Inference*) estimator (Calliess et al. [2020]) that extends the general Lipschitz interpolation framework by removing the key assumption of prior knowledge of the Lipschitz constant. This result improves on Theorem 16 of (Calliess et al. [2020]) which derives the consistency of the LACKI estimator in the noise-free setting.

We note that in the goal of obtaining a precise characterisation of the convergence rates of Lipschitz interpolation methods, we make a non-standard noise assumption (Assumption 8) utilising the concept of "non-regular" noise (Ibragimov and Has' Minskii [2013]) which describes the behaviour of the tails of the noise distribution in proximity of assumed error bounds. This type of assumption has been used in recent research on non-parametric boundary regression (see Hall and Van Keilegom [2009], Jirak et al. [2014] and ensuing works) and allows for a better comparison between the convergence rates of Lipschitz interpolation derived in this chapter and the ones guaranteed by Gaussian process regression and other kernel methods. In fact, the convergence rate bounds obtained in this chapter provide an explicit condition on the tail behaviour of the noise that indicates when the Lipschitz interpolation should be expected to asymptotically outperform or underperform other non-parametric approaches.

4.2 Lipschitz Interpolation: Set-up & Assumptions

Given an input space $\mathcal{X} \subset \mathbb{R}^d$ endowed with a metric $\mathfrak{d} : \mathcal{X}^2 \to \mathbb{R}_{\geq 0}$ and an output space[1] $\mathcal{Y} \subset \mathbb{R}$ endowed with a metric $\mathfrak{d}_{\mathcal{Y}} : \mathcal{Y}^2 \to \mathbb{R}_{\geq 0}$, we consider the non-parametric regression problem of estimating an unknown target function f described in equation (2.1) of Chapter 2.

In order to learn f, we assume that a sequence of sample sets $(D_n)_{n\in\mathbb{N}} := (G_n^{\mathcal{X}}, G_n^{\mathcal{Y}})_{n\in\mathbb{N}}$ defined such that $D_n \subset D_{n+1}$ for $n \in \mathbb{N}$ is available, where $G_n^{\mathcal{X}} := \{s_i | i = 1, ..., N_n\} \subset \mathcal{X}$ represents a set of sample inputs that can be either deterministically or randomly queried and $G_n^{\mathcal{Y}} := \{\tilde{f}_i | i = 1, ..., N_n\} \subset \mathcal{Y}$ denotes the set of noise-corrupted values of the target function f associated with the inputs in $G_n^{\mathcal{X}}$. Unless stated otherwise, we will also assume that elements of $G_n^{\mathcal{Y}}$ are of the form $\tilde{f}_k = f(s_k) + e_k$ where $(e_k)_{k\in\mathbb{N}}$ is a collection of random variables denoting the additive observational noise.

In this chapter, we will make the following assumption on the noise:

Assumption 7 *(General assumptions on noise) 0.The noise variables $(e_k)_{k\in\mathbb{N}}$ are assumed to be independent and identically[2] distributed random variables with compact support: $\exists \bar{\mathfrak{e}} > 0$ such that $\forall k \in \mathbb{N} : \mathbb{P}(e_k \in [-\bar{\mathfrak{e}}, \bar{\mathfrak{e}}]) = 1$. Furthermore, we assume that the bounds of the support are tight in the following sense:*
$\forall k \in \mathbb{N}, \epsilon \in (0, \bar{\epsilon})$:

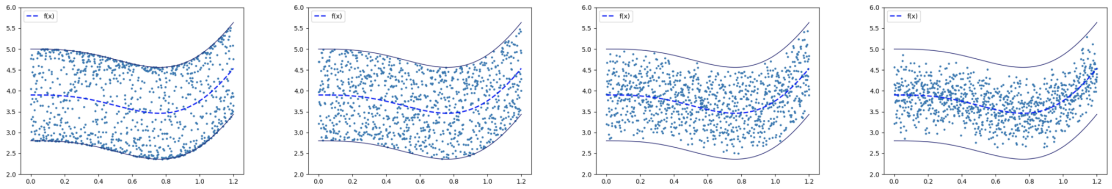
$$\mathbb{P}(e_k > \bar{\mathfrak{e}} - \epsilon) > 0 \quad \text{and} \quad \mathbb{P}(e_k < -\bar{\mathfrak{e}} + \epsilon) > 0$$

In order to derive precise upper bounds on the convergence rates, we will sometimes make an additional noise assumption which describes the behaviour of the noise at the boundary of its support. This assumption is given formally as follows:

Assumption 8 *(Assumptions on the boundary behaviour of the noise) Assume that*

[1] Here, it is possible to extend the analysis done in this chapter to a vector output space, i.e. $\mathcal{Y} \subset \mathbb{R}^m$ for $m \in \mathbb{N}$, by applying the obtained results in a component-wise fashion.

[2] The identically distributed assumption is made to alleviate notation and is not technically needed in our derivations.



(a) $\eta = 0.5$ (b) $\eta = 1$ (c) $\eta = 2$ (d) $\eta = 3$

Figure 4.21: Illustration of Assumption 8 for various η. The target function is given by $f(x) = -\sin(3x)x^2 + 5$ and the noise distribution is defined as a mixture of two truncated Weibull distributions. The solid lines define the error bounds of the observed data, i.e. (with abuse of notation) $f \pm \bar{\mathfrak{e}}$.

Assumption 7 holds. We assume that the behaviour of the noise near the bounds of the support can be characterised in the following sense:

$\exists \bar{\epsilon}, \gamma, \eta > 0 \, \forall k \in \mathbb{N}, \epsilon \in (0, \bar{\epsilon})$*:*

$$\mathbb{P}(e_k > \bar{\mathfrak{e}} - \epsilon) > \gamma \epsilon^\eta \quad \textit{and} \quad \mathbb{P}(e_k < -\bar{\mathfrak{e}} + \epsilon) > \gamma \epsilon^\eta.$$

Example 4.2.1 *(Error distributions) For $\eta = 1$, commonly used bounded error distributions such as the uniform or the truncated Gaussian distributions satisfy Assumption 8. More generally, any noise distribution for which the density can be bounded away from zero on a bounded symmetric support satisfies the assumption with $\eta = 1$.*

The assumption of boundedness of the error distribution given in Assumption 7 is standard in the Lipschitz interpolation literature (e.g. see Milanese and Novara [2004], Calliess et al. [2020]) as it ensures that the functions $\mathfrak{u}_n, \mathfrak{l}_n$ defined in Definition 2.2.1 are generally well-behaved. By contrast, as noted in the introduction of this chapter, the assumption on the tail of the noise distribution stated in Assumption 8 is non-standard in the literature. While this assumption will be not needed to ensure the asymptotic consistency of Lipschitz interpolation frameworks, the precise characterisation of the bounded tail of the noise distribution as a function of γ and η given in Assumption 8 makes it possible to derive a more refined convergence rate result that depends on η.

Remark 4.2.2 *The results of this chapter will be derived for the general Lipschitz interpolation framework defined in Definition 2.2.1. The same results can be shown to hold for the alternative framework proposed in Definition 2.2.3 with a slight modification of the proofs stated in this chapter. Furthermore, in this case, Assumptions 7 and 8 can be weakened by considering asymmetric error bounds, i.e.* $e \in [\bar{\mathfrak{e}}_1, \bar{\mathfrak{e}}_2]$ *with probability* 1 *where* $\bar{\mathfrak{e}}_1 < 0 < \bar{\mathfrak{e}}_2 \in \mathbb{R}$.

As the description of the input and output metrics has been general so far, we make the following simplifying assumption on the output metric in order to obtain our theoretical results.

Assumption 9 *(Assumption on* $\mathfrak{d}_{\mathcal{Y}}$*). In this chapter we will restrict ourselves to the case,* $\mathfrak{d}_{\mathcal{Y}}(y, y') = \|y - y'\|_{\mathcal{Y}}$, $\forall y, y' \in \mathcal{Y}$ *where* $\|.\|_{\mathcal{Y}}$ *is a norm on* $\mathcal{Y}$*. It will therefore be sufficient to derive our asymptotic results for the case:* $\|.\|_{\mathcal{Y}} = |.|$ *as discussed below.*

As the norms on $\mathcal{Y} \subset \mathbb{R}$ are of the form $\|y-y'\| = c|y-y'|$ $\forall y, y' \in \mathcal{Y}$ for some $c > 0$, it is sufficient to consider the case $\|y-y'\|_{\mathcal{Y}} = |y-y'|$, $\forall y, y' \in \mathcal{Y}$ in order to achieve our theoretical results. Assumption 9 is necessary in order to ensure that for arbitrary $x, x' \in \mathcal{X}$, the relations: $f(x) \leq f(x') + \frac{L}{c}\,\mathfrak{d}(x, x')$ and $f(x) - \frac{L}{c}\,\mathfrak{d}(x, x') \leq f(x')$ hold. In particular, for any sub-linear metric $\mathfrak{d}_{\mathcal{Y}}$, these inequalities no longer hold. We note however that no restrictions are made on the input metric $\mathfrak{d}$.

4.3 Asymptotic Consistency & Convergence Rates

In order for our consistency results to hold for both random and deterministic sampling approaches, we recall *Definition 8*: *"Becoming dense, rates,* $\overset{r}{\longrightarrow}, \overset{r}{\rightsquigarrow}, \overset{r}{\twoheadrightarrow}$*"* of (Calliess et al. [2020]) to define general sampling conditions for $(G_n^{\mathcal{X}})_{n\in\mathbb{N}}$.

Definition 4.3.1 *(Uniformly dense sampling) We say that the sequence of sets of sample inputs* $(G_n^{\mathcal{X}})_{n\in\mathbb{N}}$ *becomes uniformly dense relative to* $\mathcal{X}$ *at a rate* r *(denoted by* $(G_n^{\mathcal{X}}) \overset{r}{\twoheadrightarrow} \mathcal{X}$*) if* $\exists r : \mathbb{N} \to \mathbb{R}_+$ *such that* $\lim_{n\to\infty} r(n) = 0$ *and* $\forall n \in \mathbb{N}$,

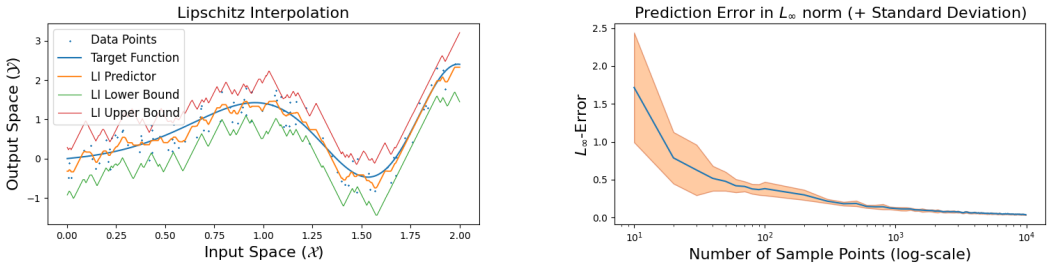


Figure 4.31: Illustration of of the consistency of Lipschitz interpolation for the target function: $f(x) = \sqrt{x}\sin(2x^2) + 0.5x$ on the input space $\mathcal{X} = [0, 2]$, with uniform sampling on $\mathcal{X}$ and with independent uniform noise: $U([-0.5, 0.5])$ on the observations ($\eta = 1$). The Lipschitz interpolation plotted in the lefthand figure utilised 100 samples and assumed that a bound of ($\bar{\mathfrak{e}}' = 0.7$) on the noise bound ($\bar{\mathfrak{e}} = 0.5$) was known in order to compute the lower and upper bounds (the LI predictors can be constructed without knowledge of $\bar{\mathfrak{e}}'$). The convergence rate and standard deviation plotted on the righthand figure were obtained by running the experiment independently 20 times. Both plots assumed that access to a bound on the best Lipschitz constant was known in order to apply the Lipschitz interpolation framework.

$\sup_{x\in\mathcal{X}} \inf_{s_n \in G_n^{\mathcal{X}}} \mathfrak{d}(s_n, x) \leq r(n)$.

Using this definition, we can provide the following asymptotic guarantee for the general Lipschitz interpolation method.

Theorem 4.3.2 *Suppose Assumptions 7 and 9 hold, $\mathcal{X}$ is bounded and the target function $f \in Lip(L^*, \mathfrak{d})$[3] with best Lipschitz constant $L^* \in \mathbb{R}_+$. If the sampling set sequence $(D_n)_{n\in\mathbb{N}}$ has sample inputs $(G_n^{\mathcal{X}})_{n\in\mathbb{N}}$ such that $\exists r \in o(1) : (G_n^{\mathcal{X}}) \overset{r}{\twoheadrightarrow} \mathcal{X}$ and the sequence of predictors $(\hat{f}_n)_{n\in\mathbb{N}}$ are computed by a general Lipschitz interpolation framework with a hyperparameter $L \in \mathbb{R}_{\geq 0}$ set such that $L \geq L^*$ then we have:*

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\left(\sup_{x\in\mathcal{X}} \mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x), f(x)) > \epsilon\right) = 0.$$

Before providing the proof of Theorem 4.3.2, we recall the notion of ϵ-covering that will be used in multiple proofs of this chapter.

Definition 4.3.3 *(ϵ-Cover) Let $d \in \mathbb{N}$, $\epsilon > 0$ and consider a set $\mathcal{X} \subset \mathbb{R}^d$ and a metric $\mathfrak{d}$ on $\mathbb{R}^d$. Denoting $B_\epsilon(x)$ the ball of radius ϵ centred in $x \in \mathcal{X}$ with respect to $\mathfrak{d}$, we define an ϵ-cover of $\mathcal{X}$ as a discrete subset $Cov(\epsilon) \subset \mathbb{R}^d$ such that*

[3] unless specified otherwise, Lipschitz continuity will be assumed to be w.r.t. the metrics $\mathfrak{d}, \mathfrak{d}_{\mathcal{Y}}$ on the spaces $\mathcal{X}, \mathcal{Y}$.

$\mathcal{X} \subset \bigcup_{x \in Cov(\epsilon)} B_\epsilon(x)$ *and the associated set of balls as* $\mathcal{B} := \{B_\epsilon(x) | x \in Cov(\epsilon)\}$. *We say furthermore that* $Cov(\epsilon)$ *is a* ϵ*-minimal cover of* $\mathcal{X}$ *if* $|Cov(\epsilon)| = min\{n : \exists \epsilon\text{-covering over } \mathcal{X} \text{ of size } n\}$.

Proof We begin by establishing a general bound on $\mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x), f(x))$, $\forall n \in \mathbb{N}$, $\forall x \in \mathcal{X}$. For any $x \in \mathcal{X}$ we have:

$$\begin{aligned}
&\mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x), f(x)) = |\hat{f}_n(x) - f(x)| \\
&= \left| \frac{1}{2} \min_{i=1,\dots,N_n} \{\tilde{f}_i + L\,\mathfrak{d}(x, s_i)\} + \frac{1}{2} \max_{i=1,\dots,N_n} \{\tilde{f}_i - L\,\mathfrak{d}(x, s_i)\} - f(x) \right| \\
&= \left| \frac{1}{2} \min_{i=1,\dots,N_n} \{\tilde{f}_i - f(x) + L\,\mathfrak{d}(x, s_i)\} + \frac{1}{2} \max_{i=1,\dots,N_n} \{\tilde{f}_i - f(x) - L\,\mathfrak{d}(x, s_i)\} \right|.
\end{aligned}$$

Using the Lipschitz continuity of f , we obtain the following set of inequality relations for the two terms stated above:

$$(1) \min_{i=1,\dots,N_n} \{e_i\} \leq \min_{i=1,\dots,N_n} \left\{ \tilde{f}_i - f(x) + L\,\mathfrak{d}(x, s_i) \right\} \leq \min_{i=1,\dots,N_n} \{e_i + (L^* + L)\,\mathfrak{d}(x, s_i)\}.$$
$$(2) \max_{i=1,\dots,N_n} \{e_i - (L^* + L)\,\mathfrak{d}(x, s_i)\} \leq \max_{i=1,\dots,N_n} \left\{ \tilde{f}_i - f(x) - L\,\mathfrak{d}(x, s_i) \right\} \leq \max_{i=1,\dots,N_n} \{e_i\}.$$

In combination, we see that

$$\begin{aligned}
&|\hat{f}_n(x) - f(x)| \\
&\leq \frac{1}{2} \max \Big\{ \underbrace{\max_{i=1,\dots,N_n} \{e_i\} + \min_{i=1,\dots,N_n} \{e_i + (L^* + L)\,\mathfrak{d}(x, s_i)\}}_{(I)}, \\
&\underbrace{- \min_{i=1,\dots,N_n} \{e_i\} - \max_{i=1,\dots,N_n} \{e_i - (L^* + L)\,\mathfrak{d}(x, s_i)\}}_{(II)} \Big\}.
\end{aligned}$$

$(I), (II)$ can then be bounded using the assumption of uniform convergence of the grid (see Definition 4.3.1). Define $R := \frac{\epsilon}{4(L^* + L)}$ and consider the minimal covering of $\mathcal{X}$ of radius R with respect to $\mathfrak{d}$ that we denote $Cov(R)$ and the associated set of hyperballs $\mathcal{B}$. By uniform convergence of the sample inputs, there exists $M \in \mathbb{N}$ such that $\forall n > M$: $\forall B \in \mathcal{B}$, $|B \cap G_n^{\mathcal{X}}| > 0$. Then, the following upper bound holds

for (I) with $n > M$ ((II) can be bounded in a similar way):

$$\begin{aligned}
&\max_{i=1,\dots,N_n}\{e_i\} + \min_{i=1,\dots,N_n}\{e_i + (L^* + L)\,\mathfrak{d}(x, s_i)\} \\
&\leq \max_{i=1,\dots,N_n}\{e_i\} + \min_{s_i \in B^x \cap G_n^{\mathcal{X}}}\{e(s_i) + (L^* + L)\,\mathfrak{d}(x, s_i)\} \\
&\leq \max_{i=1,\dots,N_n}\{e_i\} + \min_{s_i \in B^x \cap G_n^{\mathcal{X}}}\{e(s_i) + 2(L^* + L)R\}
\end{aligned}$$

where with abuse of notation, $e(s_i)$ denotes the noise term associated with the input s_i and B^x denotes a hyperball $B \in \mathcal{B}$ such that $x \in B$. Similarly for (II), we obtain

$$(II) \leq \max_{i=1,\dots,N_n}\{-e_i\} + \min_{s_i \in B^x \cap G_n^{\mathcal{X}}}\{-e(s_i) + 2(L^* + L)R\}.$$

Let $\epsilon > 0$. Utilising these bounds, $\forall n > M$, we obtain

$$\begin{aligned}
&\mathbb{P}(\sup_{x \in \mathcal{X}} \mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x), f(x)) > \epsilon) \\
&\leq \mathbb{P}\Big(2(L^* + L)R + \frac{1}{2}\sup_{x \in \mathcal{X}} \max\Big\{ \max_{i=1,\dots,N_n}\{e_i\} + \min_{s_i \in B^x \cap G_n^{\mathcal{X}}}\{e(s_i)\}, \\
&- \min_{i=1,\dots,N_n}\{e_i\} - \max_{s_i \in B^x \cap G_n^{\mathcal{X}}}\{e(s_i)\}\Big\} > \epsilon\Big) \\
&\leq \mathbb{P}\Big(\frac{1}{2}\max_{B \in \mathcal{B}} \max\Big\{ \max_{i=1,\dots,N_n}\{e_i\} + \min_{s_i \in B \cap G_n^{\mathcal{X}}}\{e(s_i)\}, \\
&- \min_{i=1,\dots,N_n}\{e_i\} - \max_{s_i \in B \cap G_n^{\mathcal{X}}}\{e(s_i)\}\Big\} > \epsilon - 2(L^* + L)R\Big) \\
&\leq \mathbb{P}\left(\max_{B \in \mathcal{B}} \max_{i=1,\dots,N_n}\{e_i\} + \min_{s_i \in B \cap G_n^{\mathcal{X}}}\{e(s_i)\} > \frac{\epsilon}{2}\right) \\
&+ \mathbb{P}\left(\max_{B \in \mathcal{B}} \max_{i=1,\dots,N_n}\{-e_i\} + \min_{s_i \in B \cap G_n^{\mathcal{X}}}\{-e(s_i)\}\} > \frac{\epsilon}{2}\right).
\end{aligned}$$

where the last inequality follows by definition of R. Both probability terms stated above can be shown to converge to 0 as follows:

$$\mathbb{P}\left(\max_{B \in \mathcal{B}} \max_{i=1,\dots,N_n}\{e_i\} + \min_{s_i \in B \cap G_n^{\mathcal{X}}}\{e(s_i)\} > \frac{\epsilon}{2}\right)$$

$$= 1 - \mathbb{P}\left(\max_{B \in \mathcal{B}} \max_{i=1,\dots,N_n}\{e_i\} + \min_{s_i \in B \cap G_n^{\mathcal{X}}}\{e(s_i)\} \leq \frac{\epsilon}{2}\right)$$

$$= 1 - \mathbb{P}\left(\forall B \in \mathcal{B}, \max_{i=1,\dots,N_n}\{e_i\} + \min_{s_i \in B \cap G_n^{\mathcal{X}}}\{e(s_i)\} \leq \frac{\epsilon}{2}\right)$$

$$\leq 1 - \mathbb{P}\left(\forall B \in \mathcal{B} : \max_{i=1,\dots,N_n}\{e_i\} \in I_1, \min_{s_i \in B \cap G_n^{\mathcal{X}}}\{e(s_i)\} \in I_2\right)$$

where $I_1 := [\bar{\mathfrak{e}} - \frac{\epsilon}{4}, \bar{\mathfrak{e}}]$ and $I_2 := [-\bar{\mathfrak{e}}, -\bar{\mathfrak{e}} + \frac{\epsilon}{4}]$. Applying a similar argument to the one given in $(\star\star)$ in the proof of Theorem 4.3.5, we have that the last term is upper bounded by

$$1 - \prod_{B \in \mathcal{B}} \mathbb{P}\left(\max_{i=1,\dots,N_n}\{e_i\} \in I_1, \min_{s_i \in B \cap G_n^{\mathcal{X}}}\{e(s_i)\} \in I_2\right)$$

$$\leq 1 - \mathbb{P}\left(\max_{i=1,\dots,N_n}\{e_i\} \in I_1, \min_{i=1,\dots,L_n}\{e_i\} \in I_2\right)^{|\mathcal{B}|}$$

where $L_n := \min_{B \in \mathcal{B}} |B \cap G_n^{\mathcal{X}}|$ and $|.|$ is used to denote the cardinality operator for finite sets. By the uniformity of the convergence of the sample inputs, we have that $\lim_{n\to\infty} L_n = \lim_{n\to\infty} N_n = +\infty$. Using basic identities of probability theory and applying Assumption 7, we have that

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{i=1,\dots,N_n}\{e_i\} \in I_1, \min_{i=1,\dots,L_n}\{e_i\} \in I_2\right) = 1.$$

which implies that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{x \in \mathcal{X}} \mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x), f(x)) > \epsilon\right) = 0$$

and concludes the proof.

■

Theorem 4.3.2 ensures that the classical Lipschitz interpolation method is asymptotically consistent for a general selection of input metrics. Furthermore, a similar result for Lipschitz interpolation with a multi-dimensional output setting $\mathcal{Y} \subset R^m$ for $m \in \mathbb{N}$ follows naturally by applying Theorem 4.3.2 to each output component function (the noise assumption would need to be modified in this case; e.g. see Assumption 13).

In general, we are mostly interested in simple metric choices for $\mathfrak{d}$. In this case with additional assumptions on $\mathcal{X}$ and $\mathcal{Y}$, we can extend the result obtained in

Theorem 4.3.2 by deriving asymptotic rates of convergence for the general Lipschitz interpolation method. More precisely, we have the following definition (Györfi et al. [2002]).

Definition 4.3.4 *Consider a sequence of non-parametric predictors* $(\hat{f}_n)_{n\in\mathbb{N}}$ *and a class of functions* $\mathcal{C}$ *endowed with a norm* $\|.\|$*. Let* $(a_n)_{n\in\mathbb{N}}$ *be a sequence of positive constants in* $\mathbb{R}$*. We define* $(a_n)_{n\in\mathbb{N}}$ *as the rate of convergence of* $(\hat{f}_n)_{n\in\mathbb{N}}$ *on* $\mathcal{C}$ *with respect to* $\|.\|$ *if there exists* $c > 0$ *such that*

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{C}} \mathbb{E}\left[a_n^{-1}\|\hat{f}_n - f\|\right] = c < \infty.$$

In order to avoid extreme cases of compact spaces, the following general assumption provides a light geometric assumption on $\mathcal{X}$.

Assumption 10 *(Geometric Assumption on* $\mathcal{X}$*) Let* $\mathcal{X} \subset \mathbb{R}^d$ *be compact and convex. There exist two constants* $r_0 > 0, \theta \in (0, 1]$ *such that* $\forall x \in \mathcal{X}$*,* $r \in (0, r_0)$ *:*
$\text{vol}\left(B_r(x) \cap \mathcal{X}\right) \geq \theta \,\text{vol}\left(B_r(x)\right)$

Assumption 10 has been used in the learning theory literature (e.g. see Hu et al. [2020] Bachoc et al. [2021]) and ensures that for all $x \in \mathcal{X}$, a constant fraction of ball with a sufficiently small radius and centred in x is contained in $\mathcal{X}$. For example, if $\mathcal{X}$ is a the unit hypercube then Assumption 10 holds with $r_0 = 1, \theta = 2^{-d}$.

The additional assumptions on the sampling of the sample inputs $(D_n)_{n\in\mathbb{N}}$ and metric of the input space $\mathfrak{d}$ are relatively standard and are given as follows.

Assumption 11 *(Assumption on Sampling)* $(G_n^{\mathcal{X}})_{n\in\mathbb{N}}$ *is a randomly sampled sequence on* $\mathcal{X}$ *with a sampling distribution density that is bounded away from zero on* $\mathcal{X}$*.*

Assumption 12 *(Hölder Condition) We restrict the input space metrics under consideration to be of the form* $\mathfrak{d}(x, y) = \|x - y\|_p^\alpha$ *where* $\alpha \in (0, 1]$ *and* $\|.\|_p$ *denotes the usual p-norm on* $\mathbb{R}^d$ *with* $p \in \mathbb{N} \cup \{+\infty\}$*.*

Theorem 4.3.5 *Consider an input space $\mathcal{X} \subset \mathbb{R}^d$ that satisfies Assumption 10, an output space $\mathcal{Y} \subset \mathbb{R}$ and the function space $\mathcal{C} = Lip(L^*, \mathfrak{d})$ with $L^* \in \mathbb{R}_{\geq 0}$ endowed with the sup-norm: $\|h\|_\infty = \sup_{x \in \mathcal{X}} \|h(x)\|_\mathcal{Y}$. Assume that Assumptions 7, 8, 11, 12 with $\alpha \in (0, 1]$, $p \in \mathbb{N}$ hold. Then, any sequence of predictors $(\hat{f}_n)_{n \in \mathbb{N}}$ generated by the general Lipschitz interpolation framework with a hyperparameter $L \geq L^*$ achieves a rate of convergence of at least $(a_n)_{n \in \mathbb{N}} := \left((n^{-1} log(n))^{\frac{\alpha}{d+\eta\alpha}}\right)_{n \in \mathbb{N}}$ with respect to $\|.\|_\infty$, i.e.*

$$\limsup_{n \to \infty} \sup_{f \in Lip(L^*, \mathfrak{d})} \mathbb{E}\left[a_n^{-1} \|\hat{f}_n - f\|_\infty\right] < \infty.$$

Proof See appendix 4.B.

■

Convergence lower bounds do not exist for the exact setting considered in this chapter signifying that we cannot directly compare the rates stated in Theorem 4.3.5 to a theoretically optimal convergence rate. Instead, we can note that the convergence rate of Lipschitz interpolation is in line with several known optimal rates in related settings (see Table 4.31), i.e. non-parametric regression on the Lipschitz continuous function space endowed with an L_2 or L_∞ norm. In particular, we note that the exponent of the convergence rate derived for Lipschitz interpolation exactly matches the exponent of the convergence rate derived in Tsybakov [2004] in the case where the noise distribution is assumed to be uniform (i.e. $\eta = 1$). Our convergence rate is however larger by a log-factor due to a difference in norm.

Furthermore, by varying η in Assumption 8, we can compare our rate of convergence: $\mathrm{O}((n^{-1} log(n))^{\frac{\alpha}{d+\eta\alpha}})_{n \in \mathbb{N}}$ to classical non-parametric convergence rates. More precisely, we observe that

- For $\eta < 2$: the derived convergence rates for Lipschitz interpolation are better than the known optimal convergence rates obtained under a Gaussian tail noise assumption on the error distribution: $(n^{-1} log(n))^{\frac{\alpha}{2\alpha+d}}$ (Stone [1982]) which are attained[4] by Gaussian process regression (Yang et al. [2017]) and

[4]Note that these methods can be shown to converge at this rate under the simple assumption

Algorithm/Type	Convergence Rate	Noise Assumption	Norm
LI (Upper Bound)	$O(n^{-1}log(n))^{\frac{\alpha}{d+\eta\alpha}}$	Bounded	L_∞.
Optimal (Tsybakov [2004])	$\Theta\left(n^{-1}\right)^{\frac{\alpha}{d+\alpha}}$	Uniform ($\eta = 1$)	L_2
Optimal (Stone [1982])	$\Theta\left(n^{-1}log(n)\right)^{\frac{\alpha}{d+2\alpha}}$	Gaussian[5]	L_∞
Optimal (Stone [1982])	$\Theta\left(n^{-1}\right)^{\frac{\alpha}{d+2\alpha}}$	Gaussian[4]	L_2
Optimal (Jirak et al. [2014])	$\Theta\left(n^{-1}\right)^{\frac{\alpha}{1+\eta\alpha}}$ (d=1)	Boundary Regr.	L_2
Upper Bound (Selk et al. [2022])	$O(n^{-1}log(n))^{\frac{\alpha}{d+\eta\alpha}}$	Boundary Regr.	L_∞

Table 4.31: Comparison of the convergence rate derived in Theorem 4.3.5 with optimal rates of convergence rates in similar settings and discussion given in this section.

other kernel-based non-parametric methods such as local polynomial regression (Stone [1982]) or the Nadaraja-Watson estimator (Tsybakov [2004], Müller and Wefelmeyer [2010]).

- For $\eta > 2$: the opposite becomes true and these alternative non-parametric methods can be expected to converge quicker asymptotically than Lipschitz interpolation.

This "η-condition" provides a theoretical tool for comparing the expected long-run performance of Lipschitz interpolation relative to alternative non-parametric methods and can help guide the choice of the system identification approach if information on the non-regularity of the noise distribution is obtainable. We note that the convergence rates of the kernel-based non-parametric methods stated in Table 4.31 hold under general noise assumptions (see footnotes 4 and 5 below) and that, aside from the Nadaraja-Watson estimator, no formal derivation of improved convergence rates in the bounded noise setting considered in this chapter currently exists[6] for these methods. As these kernel-based non-parametric frameworks generally rely on local averaging of the noise in order to prove convergence, it is expected that their convergence rates do not improve with respect to their classical convergence rates (stated in Table 4.31) under Assumption 7 and Assumption 8. This has been formally shown to be true for the Nadaraja-Watson estimator by Müller and Wefelmeyer [2010] and a more general discussion on the topic can be found in Meister and Reiß [2013].

of bounded variance (Györfi et al. [2002]).

[5]Various generalisations of this noise assumption exist, see (Stone [1982]).

[6]To the extent of our knowledge.

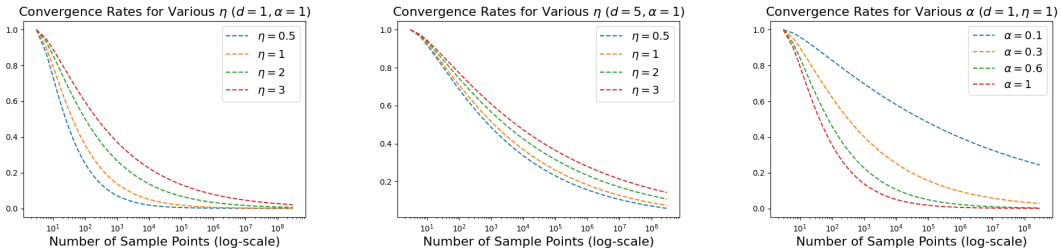


Figure 4.32: Illustration of the behaviour of the convergence rates derived in Theorem 4.3.5 for various values of (d, α, η).

As discussed above, the convergence rates obtained in Theorem 4.3.5 under the bounded noise assumptions are better than the classical optimal convergence rates derived by Stone [1982]. This is possible as the lower bounds of these optimal convergence rates are generally derived under the condition that the noise has a positive density with respect to the Lebesgue measure on $\mathbb{R}$ which does not hold for the noise assumptions of this chapter. As a consequence, $O\left(n^{-1}log(n)\right)^{\frac{\alpha}{d+\eta\alpha}}$ provides a new general upper bound on the non-parametric regression problem in the bounded noise setting and future work can be done on deriving lower bounds to match these results. We expect the lower bounds to be tight given recent results by Jirak et al. [2014] on the optimal convergence rates of the related non-parametric boundary regression problem (see below for a more detailed discussion).

The optimality results of Theorem 2 of Milanese and Novara [2004] show that $(\mathfrak{u}_n^2)_{n\in\mathbb{N}}$, $(\mathfrak{l}_n^2)_{n\in\mathbb{N}}$ (see Remark 2.2.3 and Theorem 2.2.4) are exactly equal to the upper and lower bounds of the *feasible systems set*, i.e. the set of all data-consistent Lipschitz continuous systems and therefore provide worst-case error prediction bounds. With little modification to the proof of Theorem 4.3.5, both error bounds can be shown to converge to f at the same rate as $(\hat{f}_n)_{n\in\mathbb{N}}$ as stated in the following Corollary.

Corollary 4.3.6 *Assume that the setting and assumptions of Theorem 4.3.5 holds. The worst-case prediction guarantees $(\mathfrak{u}_n^2)_{n\in\mathbb{N}}$, $(\mathfrak{l}_n^2)_{n\in\mathbb{N}}$ defined in Remark 2.2.3 with second hyperparameter: $\bar{\mathfrak{e}}' = \bar{\mathfrak{e}}$, converge uniformly to any target function $f \in Lip(L^*, \mathfrak{d})$ at a rate of at least $((n^{-1}log(n))^{\frac{\alpha}{d+\eta\alpha}})_{n\in\mathbb{N}}$.*

Proof Follows from the proof of Theorem 4.3.5.

■

A connection between our convergence results and recent work on non-parametric boundary regression (see Hall and Van Keilegom [2009] and ensuing works) can be made. More precisely, consider the predictive functions $(\tilde{\mathfrak{u}}_n)_{n\in\mathbb{N}}$, $(\tilde{\mathfrak{l}}_n)_{n\in\mathbb{N}}$ defined for all $n \in \mathbb{N}$ as $\tilde{\mathfrak{u}}_n : \mathcal{X} \to \mathcal{Y}$, $x \mapsto \mathfrak{u}_n(x) + \bar{\mathfrak{e}}_1 + \bar{\mathfrak{e}}_2$ and $\tilde{\mathfrak{l}}_n : \mathcal{X} \to \mathcal{Y}$, $x \mapsto \mathfrak{l}_n(x) - \bar{\mathfrak{e}}_1 - \bar{\mathfrak{e}}_2$ where $\bar{\mathfrak{e}}_1, \bar{\mathfrak{e}}_2$ are tight asymmetric bounds on the error. $(\tilde{\mathfrak{u}}_n)_{n\in\mathbb{N}}$, $(\tilde{\mathfrak{l}}_n)_{n\in\mathbb{N}}$ can be interpreted as conservative non-parametric boundary regression methods. Therefore, in the context of bounded noise, the two problems are equivalent and we can again slightly modify the proof of Theorem 4.3.5 to obtain the same uniform asymptotic convergence rates of $O(n^{-1}log(n))^{\frac{\alpha}{\eta d+\alpha}})$ as $(\hat{f}_n)_{n\in\mathbb{N}}$. These rates exactly match the recently derived best convergence rates in the multivariate boundary regression problem (Selk et al. [2022]) and have the same exponent[7] as the optimal rates derived with respect to the L_2 norm (Jirak et al. [2014]). In order to properly define $(\tilde{\mathfrak{u}}_n)_{n\in\mathbb{N}}$, $(\tilde{\mathfrak{l}}_n)_{n\in\mathbb{N}}$, prior knowledge of an upper bound on the Lipschitz constant $(L \geq L^*)$ as well as the Hölder exponent (α) and of tight bounds of the noise $(\bar{\mathfrak{e}}_1, \bar{\mathfrak{e}}_2)$ are needed. However in contrast to the proposed "best" non-parametric estimators that attain the optimal rates, we do not require prior knowledge of the degree of "non-regularity" of the noise (η, defined in Assumption 8) which is usually required in order to define an optimal bandwidth hyperparameter (Drees et al. [2019], Selk et al. [2022]). In the bounded noise setting, our assumption is therefore arguably more natural and simpler to verify in practice as (η) is generally hard to determine precisely.

4.4 Online Learning: Asymptotics

A set-up not yet explicitly considered in this chapter but relevant to control applications is when the output variables can be used as input variables. More specifically,

[7] They differ by a log-factor which is usual when considering the L_∞ norm instead of the L_2 norm.

we consider the case where f models the dynamics of a semi-autoregressive stochastic system as described by equation (2.2) stated in Chapter 2:

$$y_n = f(x_n) + e_n$$

where $x_n = (y_{n-d_y}, ..., y_{n-1}, u_{n-d_u}, ..., u_n)$ with $y_i \in \mathcal{Y} \subset \mathbb{R}$ and $u_i \in \mathcal{U} \subset \mathbb{R}^s$ for $d_y, d_u, s \in \mathbb{N}$ and $e_n \in \mathbb{R}$ is a noise variable that satisfies Assumption 7. Here, y_i denotes the autoregressive inputs and u_i denotes vectors of past and current control inputs. In this setting, we will therefore consider $\mathcal{X} = \mathbb{R}^{d_y} \times \mathcal{U}^{d_u+1} \subset \mathbb{R}^{d_y+(d_u+1)(s)}, \mathcal{Y} = \mathbb{R}$. If the dynamics and control inputs are such that the underlying dynamical system is ergodic, then Theorem 4.3.2 can be applied and a weaker version of Theorem 4.3.5 can be derived. However, in general, this cannot be guaranteed and the following result on the asymptotic point-wise convergence of the general Lipschitz interpolation framework is needed.

Corollary 4.4.1 *Consider $\mathcal{X}, \mathcal{Y}, (x_n)_{n\in\mathbb{N}}, (u_n)_{n\in\mathbb{N}}, (y_n)_{n\in\mathbb{N}}$ as defined above, $L^* \geq 0$ and $(\hat{f}_n)_{n\in\mathbb{N}}$ as defined in Definition 2.2.1 with $L \geq L^*$ and $(\mathcal{D}_n)_{n\in\mathbb{N}} = (x_n, y_n)_{n\in\mathbb{N}}$. Assume that Assumptions 7, 9 hold. Assume furthermore that $\mathcal{U} \subset \mathbb{R}^s$ is bounded. Then $\forall p \in \mathbb{N}$, $M^* \in \mathbb{R}^+$*

$$\lim_{n\to\infty} \sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \mathbb{E}[\|f(x_{n+1}) - \hat{f}_n(x_{n+1})\|_{\mathcal{Y}}^p] \to 0$$

where $\overline{Lip}(L^*, \mathfrak{d}, M^*)$ denotes the set containing all functions in $Lip(L^*, \mathfrak{d})$ that are bounded by M^*, i.e. $|f(x)| \leq M^*$.

Proof As in the proof of Theorem 4.3.5, we have that for all $n \geq 1$ and any sampling procedure $\mathcal{D}_n$, $\sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \|\hat{f}_n - f\|_\infty$ is uniformly bounded with probability 1. This follows from (1) the existence of a bounded set $\tilde{\mathcal{X}} \subset \mathcal{X}$ such that $(x_n)_{n\in\mathbb{N}} \subset \tilde{\mathcal{X}}$ (with probability 1) which is due to the boundedness of $\overline{Lip}(L^*, \mathfrak{d}, M^*)$, the compactness of $\mathcal{U}$ and Assumption 7 (which implies that the noise is bounded), (2) $f \in Lip(L^*, \mathfrak{d})$ and (3) by construction of the Lipschitz interpolation framework. More precisely, we have $\forall n \in \mathbb{N}$, $\sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \|\hat{f}_n - f\|_\infty \leq 2\bar{\mathfrak{e}} + 2L\delta_{\mathfrak{d}}(\tilde{\mathcal{X}})$ where

$\delta_{\mathfrak{d}}(\tilde{\mathcal{X}}) := \sup_{x,y\in\tilde{\mathcal{X}}} \mathfrak{d}(x,y)$. Using Lemma 4.C.1, it is therefore sufficient to show convergence in probability, i.e.

$$\forall \epsilon > 0 : \lim_{n\to\infty} \sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \mathbb{P}(|\hat{f}_n(x_{n+1}) - f(x_{n+1})| > \epsilon) = 0$$

which can be done through a modified proof of Theorem 4.3.2 as follows.

Fix $\epsilon > 0$ and consider the minimal covering of $\bar{\mathcal{X}}$ by balls of radius $r < \frac{\epsilon}{4(L^*+L)}$ which we denote $cov(r)$ and the associated set of hyperballs $\mathcal{B}$ (the existence of a finite covering is guaranteed by the compactness of $\tilde{\mathcal{X}}$). Let $\delta > 0$ be arbitrary, we show that for sufficiently large n,

$$\sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \mathbb{P}(|\hat{f}_n(x_{n+1}) - f(x_{n+1})| > \epsilon) < \delta.$$

By the finiteness of $\mathcal{B}$, we have

$$\lim_{N\to\infty} \mathbb{P}(\forall B \in \mathcal{B} : |(x_n)_{n\geq N} \cap B| \in \{0, +\infty\}) = 1.$$

Therefore, there exists $N_1 := N_1(\delta) \in \mathbb{N}$ such that the event:

$$E(N_1) := \{\forall B \in \mathcal{B} : |(x_n)_{n\geq N_1} \cap B| \in \{0, +\infty\}\}$$

holds with probability of at least $1 - \frac{\delta}{2}$. Then, denoting by $\tilde{\mathcal{B}} \subset \mathcal{B}$ the subset of $\mathcal{B}$ consisting of hyperballs that contain an infinite number of elements of $(x_n)_{n\geq N_1}$, we can proceed as in the proof of Theorem 4.3.2.

Let $f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)$ be arbitrary. For $n > \tilde{N}_1 > N_1$ with $\tilde{N}_1$ sufficiently large (such that there is at least one sample input in each hyperball of $\tilde{\mathcal{B}}$), applying the same arguments as in the proof of Theorem 4.3.2:

$$\begin{aligned}
&\mathbb{P}\left(\mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x_{n+1}), f(x_{n+1})) > \epsilon\right) \\
&\leq \mathbb{P}\left(\mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x_{n+1}), f(x_{n+1})) > \epsilon | E(N_1)\right) \mathbb{P}\left(E(N_1)\right) + \mathbb{P}\left(E(N_1)^c\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(E(N_1)\right)\left(\mathbb{P}\left(\max_{B\in\tilde{\mathcal{B}}}|\max_{i=1,\dots,n}\{e_i\}+\min_{x_i\in B}\{e(x_i)\}|>\frac{\epsilon}{2}\middle|E(N_1)\right)\right.\\
&\left.+\mathbb{P}\left(\max_{B\in\tilde{\mathcal{B}}}|\min_{i=1,\dots,n}\{e_i\}+\max_{x_i\in B}\{e(x_i)\}|>\frac{\epsilon}{2}\middle|E(N_1)\right)\right)+\frac{\delta}{2}\\
&\leq \mathbb{P}\left(\max_{B\in\tilde{\mathcal{B}}}|\max_{i=1,\dots,n}\{e_i\}+\min_{x_i\in B}\{e(x_i)\}|>\frac{\epsilon}{2}\right)\\
&+\mathbb{P}\left(\max_{B\in\tilde{\mathcal{B}}}|\min_{i=1,\dots,n}\{e_i\}+\max_{x_i\in B}\{e(x_i)\}|>\frac{\epsilon}{2}\right)+\frac{\delta}{2}.
\end{aligned}
$$

As the choice of $f\in\overline{Lip}(L^*,\mathfrak{d},M^*)$ was arbitrary and the upper bound expressed above does not depend on f, we have that the sum of the first two terms of this upper bound can be treated with the same approach as the one used to conclude the proof of Theorem 4.3.2 and upper bounded by $\frac{\delta}{2}$ for all $n>N_2$ with an appropriately selectetd $N_2\in\mathbb{N}$. This implies for all $n>\max(\tilde{N}_1,N_2)$:

$$
\begin{aligned}
&\mathbb{P}\left(|\hat{f}_n(x_{n+1})-f(x_{n+1})|>\epsilon\right)\\
&\leq \lim_{n\to\infty}\mathbb{P}\left(\max_{B\in\tilde{\mathcal{B}}}|\max_{i=1,\dots,n}\{e_i\}+\min_{x_i\in B}\{e(x_i)\}|>\frac{\epsilon}{2}\right)\\
&+\lim_{n\to\infty}\mathbb{P}\left(\max_{B\in\tilde{\mathcal{B}}}|\min_{i=1,\dots,n}\{e_i\}+\max_{x_i\in B}\{e(x_i)\}|>\frac{\epsilon}{2}\right)+\frac{\delta}{2}\\
&\leq \frac{\delta}{2}+\frac{\delta}{2}=\delta
\end{aligned}
$$

which concludes the proof.

■

The setting considered in Corollary 4.4.1 is the same as the one considered in Milanese and Novara [2004] and in ensuing applications of the Lipschitz interpolation framework in the context of MPC (see Canale et al. [2014], Manzano et al. [2020]). As in Corollary 4.3.6, the worst-case prediction guarantees $(\mathfrak{u}_n^2)_{n\in\mathbb{N}}$, $(\mathfrak{l}_n^2)_{n\in\mathbb{N}}$ can be shown to provide similar guarantees to the one proposed Corollary 4.4.1 which provides a theoretical guarantee that even conservative adaptive controllers relying on worst-case bounds of Lipschitz interpolation methods will consider the true underlying dynamics in the long run. In Section 4.6, a slight modification of Corollary 4.6.3 that considers dynamics with multidimensional outputs $(y_n)_{n\in\mathbb{N}}$ is given. This exten-

sion is then applied in the context of tracking control in order to obtain closed-loop stability guarantees for a simple online-learning based controller

Remark 4.4.2 *The assumptions of Corollary 4.4.1 are weaker then the those of similar results that can be found in the literature. This is due to the fact that the conclusion of the corollary is also weaker as only a type of "point-wise convergence" is established.*

To conclude this section, we remark that if an additional assumption is made on the sequence of inputs $(x_n)_{n\in\mathbb{N}}$, then the convergence rate derived in Theorem 4.3.5 holds in the online learning setting. This assumption is given using the following definition on the "regularity of the sampling" of $(x_n)_{n\in\mathbb{N}}$.

Definition 4.4.3 *(Regularity Assumption for $(x_n)_{n\in\mathbb{N}}$) We say that $(x_n)_{n\in\mathbb{N}}$ is regularly sampled on a set $\bar{\mathcal{X}} \subset \mathcal{X}$ if $\exists N \in \mathbb{N}$, $(x_n)_{n\in\mathbb{N}_{\geq N}} \subset \bar{\mathcal{X}}$ and $\exists M \in \mathbb{N}$ such that $\forall n > N$ and $\forall A \subset \bar{\mathcal{X}}$,*

$$\mathbb{P}(x_{n+M} \in A | x_n) > C\mu(A)$$

where $\mu(A)$ denotes the Lebesgue measure of A and $C > 0$ is an arbitrary constant.

In essence, Definition 2.2.1 states that $(x_n)_{n\in\mathbb{N}}$ is regularly sampled on a given set $\bar{\mathcal{X}} \subset \mathcal{X}$ if $(x_n)_{n\in\mathbb{N}}$ will eventually be contained in $\bar{\mathcal{X}}$ and will continue to visit all of $\bar{\mathcal{X}}$ with non-zero probability. The existence of such a set depends implicitly on the target function and the defined control inputs.

Corollary 4.4.4 *Assume that the setting and assumptions of Corollary 4.4.1 hold and consider $f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)$. If Assumption 8 holds and the stochastic control law $u_{n+1} := u(x_n, \hat{f}_n, \mathcal{D}_n)$ is defined such that $(x_n)_{n\in\mathbb{N}}$ is regularly sampled on a bounded set $\bar{\mathcal{X}} \subset \mathcal{X}$ that satisfies Assumption 10, then*

$$\limsup_{n\to\infty} \mathbb{E}[a_n^{-1} \| f(x_{n+1}) - \hat{f}_n(x_{n+1}) \|_{\mathcal{Y}}] < \infty.$$

where $(a_n)_{n\in\mathbb{N}} := ((n^{-1}log(n))^{\frac{\alpha}{d+\eta\alpha}})_{n\in\mathbb{N}}$.

Remark 4.4.5 *From the proof of Corollary 4.4.1, we have that if $\mathcal{U}$ is bounded and $f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)$, then there exists a bounded $\tilde{\mathcal{X}} \subset \mathcal{X}$ that contains $(x_n)_{n\in\mathbb{N}}$ with probability 1. Therefore, only the second part of Definition 4.4.3 and the geometric shape of $\bar{\mathcal{X}}$ need to be checked in order for Corollary 4.4.4 to hold.*

Proof The proof of Corollary 4.4.4 follows from Theorem 4.3.5. More precisely:

By assumption, we have that there exists $M, N \in \mathbb{N}$ and a bounded set $\bar{\mathcal{X}} \subset \mathcal{X}$ such that Definition 4.4.3 and Assumption 10 hold. Consider the sequence $(x_n)_{n\in\mathbb{N}_{\geq N}} \subset \bar{\mathcal{X}}$ and the subsequence $(\tilde{x}_n)_{n\in\mathbb{N}} \subset (x_n)_{n\in\mathbb{N}_{\geq N}}$ defined such that $\tilde{x}_n = x_{Mn+N}$ for all $n \in \mathbb{N}$. From Definition 4.4.3, we have that for all $n \in \mathbb{N}$, $\tilde{x}_n$ is sampled on $\bar{\mathcal{X}}$ with a probability distribution whose density is bounded away from zero on all of $\bar{\mathcal{X}}$.

Then, defining $(\hat{f}_n^M)_{n\in\mathbb{N}}$ as the predictors of the Lipschitz interpolation framework with hyperparameter L and sample inputs $(\tilde{x}_n)_{n\in\mathbb{N}}$, we can apply Theorem 4.3.5 to $(\hat{f}_n^M)_{n\in\mathbb{N}}$. This implies that $(\hat{f}_n^M)_{n\in\mathbb{N}}$ converges uniformly on $\bar{\mathcal{X}}$ to f at a rate that is upper bounded by $(a_{\lfloor \frac{n}{M} \rfloor})_{n\in\mathbb{N}} = \tilde{c}(a_n)_{n\in\mathbb{N}}$ for some $\tilde{c} >$ that depends on M and where $n \in \mathbb{N}$ denotes the index of the original sequence: $(x_n)_{n\in\mathbb{N}}$. As the asymptotic convergence rate of $(\hat{f}_n)_{n\in\mathbb{N}}$ is at least as fast as the convergence rate of $(\hat{f}_n^M)_{n\in\mathbb{N}}$ due to the fact that the input samples utilised by $(\hat{f}_n^M)_{n\in\mathbb{N}}$ are also utilised by $(\hat{f}_n)_{n\in\mathbb{N}}$, we have that $(\hat{f}_n)_{n\in\mathbb{N}}$ achieves the same uniform convergence rate on $\bar{\mathcal{X}}$. Finally, as $(x_n)_{n\in\mathbb{N}_{\geq N}} \subset \bar{\mathcal{X}}$, the same converge rate holds for the pointwise asymptotic convergence of $(x_n)_{n\in\mathbb{N}}$, i.e.

$$\limsup_{n\to\infty} \mathbb{E}[a_n^{-1} \| f(x_{n+1}) - \hat{f}_n(x_{n+1}) \|_{\mathcal{Y}}]$$

with $(a_n)_{n\in\mathbb{N}} := ((n^{-1}log(n))^{\frac{\alpha}{d+\eta\alpha}})_{n\in\mathbb{N}}$.

■

While Corollary 4.4.4 provides an interesting extension to Theorem 4.3.5, the characterisation of the regularly sampling set $\bar{X}$ and the necessity of ensuring that Assumption 10 holds for $\bar{\mathcal{X}}$ can be difficult to do in practice. Therefore, in comparison to Corollary 4.4.1 which can be directly utilised in various control applications,

Corollary 4.4.4 is essentially a theoretical result.

4.5 Removing the Lipschitz Constant Assumption

As discussed in the introduction and studied in Chapter 3, the main difficulty of the Lipschitz interpolation framework is obtaining a suitable hyper-parameter that properly estimates the Lipschitz constant of the unknown target function. In cases where prior knowledge of the Lipschitz constant of f is not obtainable, an additional step is therefore needed. While one solution would be to compute this estimate offline beforehand, this approach is problematic when considering a stream of data. Instead, one can consider the approach developed by Novara et al. [2013] and applied in the context of Lipschitz interpolation by Calliess et al. [2020] which utilises a modified version of Strongin's Lipschitz constant estimator (Strongin [1973]) to $(D_n)_{n\in\mathbb{N}}$ to obtain a sequence $(L(n))_{n\in\mathbb{N}}$ of approximations of L^*. These estimates can be continuously updated with the arrival of new data and are defined formally in the following definition.

Definition 4.5.1 *(LACKI rule) The Lazily Adapted Lipschitz Constant Kinky Inference (LACKI) rule computes a Lipschitz interpolation predictor $\hat{\mathfrak{f}}_{\mathfrak{n}}$ as per Definition 2.2.1, but where L depends on $(D_n)_{n\in\mathbb{N}}$ and is computed as follows:*

$$L(n) := \max\Big\{0, \max_{(s,s')\in U_n} \frac{\mathfrak{d}_{\mathcal{Y}}(\tilde{f}(s), \tilde{f}(s')) - \lambda}{\mathfrak{d}(s, s')}\Big\}, \tag{4.1}$$

where $U_n = \{(g_1, g_2) \in G_n^{\mathcal{X}} \times G_n^{\mathcal{X}} | \, \mathfrak{d}(g_1, g_2) > 0\}$ and λ is a hyperparameter.

The errors are estimated correctly if the λ hyper-parameter of the LACKI rule is set to $2\bar{\mathfrak{e}}$. Calliess et al. [2020] provides worst-case prediction bounds even when the errors are not correctly estimated. In this chapter, we focus on the case where the error bounds are known and λ can be correctly specified. We note that the Lipschitz estimator $L(n)$ given by LACKI is the smallest Lipschitz constant that is consistent with the data. In other words, it reduces the hypothesis space of Lipschitz continuous functions $Lip(L(n), \mathfrak{d})$ that the target function f could belong to.

We start by showing that the LACKI rule proposed in Definition 4.5.1 converges asymptotically to the best Lipschitz constant of the unknown target function.

Lemma 4.5.2 *If the assumptions of Theorem 4.3.2 hold, then :*

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}(|L(n) - L^*| > \epsilon) = 0$$

Proof Fix an arbitrary $\epsilon > 0$. We start by defining an auxiliary function F:

$$\begin{aligned} F : Dom(F) := \mathcal{X} \times \mathcal{X} - \{(x,x)|x \in \mathcal{X}\} &\longrightarrow \mathbb{R}_{\leq 0} \\ (x,y) &\longmapsto \frac{\mathfrak{d}_{\mathcal{Y}}(f(x), f(y))}{\mathfrak{d}(x,y)} \end{aligned}$$

By construction, $L^* = \sup_{(x,y)\in Dom(F)} F(x,y)$ and there exists $(x_1, x_2) \in Dom(F)$ such that $L^* - \frac{\epsilon}{2} \leq F(x_1, x_2) \leq L^*$. Hence,

$$\begin{aligned} &\mathbb{P}\left(|L(n) - L^*| > \epsilon\right) \\ &\leq \mathbb{P}\left(|L(n) - F(x_1,x_2)| + |F(x_1,x_2) - L^*| > \epsilon\right) \\ &= \mathbb{P}\left(|F(x_1,x_2) - L(n)| > \frac{\epsilon}{2}\right) = \mathbb{P}\left(F(x_1,x_2) - L(n) > \frac{\epsilon}{2}\right). \end{aligned}$$

Since F is continuous on its domain, we have that $\exists \delta_1 > 0$ such that $\forall (x,y) \in B_{\delta_1}((x_1,x_2))$[8] $\cap\, Dom(F)$, $|F(x_1,x_2) - F(x,y)| < \frac{\epsilon}{2}$. Defining $0 < \delta_2 < \min\{\frac{\delta_1}{2}, \frac{\mathfrak{d}(x_1,x_2)}{2}\}$, we consider the two hyperballs $B_1 := B_{\delta_2}(x_1)$, $B_2 := B_{\delta_2}(x_1)$. Then

$$\begin{aligned} &F(x_1,x_2) - L(n) \\ &= F(x_1,x_2) - \max_{(s,s')\in U_n} \frac{|\tilde{f}(s) - \tilde{f}(s')| - \lambda}{\mathfrak{d}(s,s')} \\ &\leq F(x_1,x_2) - \max_{s_i\in B_1, s_j \in B_2} \frac{|\tilde{f}(s_1) - \tilde{f}(s_j)| - \lambda}{\mathfrak{d}(s_i,s_j)} \\ &\leq F(x_1,x_2) - \max_{\substack{s_i\in B_1, s_j \in B_2 \\ cond(s_i,s_j)}} \frac{|f(s_i) - f(s_j)| + |e(s_i) - e(s_j)| - \lambda}{\mathfrak{d}(s_i,s_j)} \end{aligned}$$

[8] Here, $B_\delta((x_1,x_2))$ denotes the ball centered in (x_1,x_2) of radius δ_1 with respect to $\mathfrak{d}_{\mathcal{X}\times\mathcal{X}}$ defined such that $\mathfrak{d}_{\mathcal{X}\times\mathcal{X}}((x_1,x_2),(x_1',x_2')) = \mathfrak{d}(x_1,x_1') + \mathfrak{d}(x_2,x_2')$

$$\leq F(x_1, x_2) - \min_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} \frac{|f(s_i) - f(s_j)|}{\mathfrak{d}(s_i, s_j)} - \max_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} \frac{|e(s_i) - e(s_j)| - \lambda}{\mathfrak{d}(s_i, s_j)}.$$

where $cond(x, y) := \{sgn\,(f(s_i) - f(s_j)) = sgn\,(e(s_i) - e(s_j))\}$ and with abuse of notation, $\tilde{f}(s_i)\, e(s_i)$ denote the noise term associated with the input s_i. By definition of B_1, B_2, we have

$$F(x_1, x_2) - \min_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} \frac{|f(s_i) - f(s_j)|}{\mathfrak{d}(s_i, s_j)} = F(x_1, x_2) - \min_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} F(s_i, s_j) \leq \frac{\epsilon}{4}.$$

Substituting this value into the initial expression, we can obtain the upper bound

$$\begin{aligned} &\frac{\epsilon}{4} - \max_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} \frac{|e(s_i) - e(s_j)| - \lambda}{\mathfrak{d}(s_i, s_j)} \\ \leq &\frac{\epsilon}{4} + \min_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} \frac{\lambda - |e(s_i) - e(s_j)|}{\mathfrak{d}(s_i, s_j)} \\ \leq &\frac{\epsilon}{4} + \min_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} \frac{\lambda - |e(s_i) - e(s_j)|}{\mathfrak{d}(x_1, x_2) - 2\delta_2}. \end{aligned}$$

By the assumption of uniformly dense sampling, there exists $M \in \mathbb{N}$ such that $r(M) < \delta_2$. Therefore, for $n > M$,

$$\mathbb{P}\left(F(x_1, x_2) - L(n) > \frac{\epsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\min_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} \frac{\lambda - |e(s_i) - e(s_j)|}{\mathfrak{d}(x_1, x_2) - 2\delta_2} > \frac{\epsilon}{4}\right)$$

$$\leq \mathbb{P}\left(\min_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} \{\lambda - |e(s_i) - e(s_j)|\} > \frac{\epsilon}{4}(\,\mathfrak{d}(x_1, x_2) - 2\delta_2)\right)$$

$$= \mathbb{P}\left(\max_{\substack{s_i \in B_1, s_j \in B_2 \\ cond(s_i, s_j)}} |e(s_i) - e(s_j)| < \lambda - \frac{\epsilon}{4}(\mathfrak{d}(x_1, x_2) - 2\delta_2) \right).$$

As $\lambda = 2\bar{\mathfrak{e}}$ and $\mathfrak{d}(x_1, x_2) > 2\delta_2$, the last expression can be shown to converge to 0 as n goes to ∞ by a similar argument to the one used in the proof of Theorem 4.3.2.

■

Lemma 4.5.2 proves that the modified version of Strongin's estimate defined in Definition 4.5.1 is a consistent Lipschitz constant estimator under bounded noise. It is therefore of interest for applications outside the one considered in this chapter, e.g. see in particular global optimisation methods that depend explicitly on the Lipschitz constant (see for example Malherbe and Vayatis [2017]). One main drawback however is that none of the finite sample estimates generated by the LACKI rule upper bound the true Lipschitz constant. This is discussed in more detail after Theorem 4.5.3.

Using Theorem 4.3.2 and Lemma 4.5.2, we can now show that the sequence of LACKI predictors $(\hat{f}_n)_{n\in\mathbb{N}}$ converges uniformly and in probability to the target function f.

Theorem 4.5.3 *If the assumptions of Theorem 4.3.2 hold, then the sequence of LACKI predictors $(\hat{f}_n)_{n\in\mathbb{N}}$ with $\lambda = 2\bar{\mathfrak{e}}$ converges to f uniformly and in probability:*

$$\forall \epsilon > 0, \lim_{n\to\infty} \mathbb{P}\left(\sup_{x\in\mathcal{X}} \mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x), f(x)) > \epsilon \right) = 0$$

Proof The proof of Theorem 4.5.3 follows from Theorem 4.3.2 and Lemma 4.5.2. Fix an arbitrary $\epsilon > 0$, we have

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}} \mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x), f(x)) > \epsilon \right)$$

$$\leq \mathbb{P}\left(\sup_{x\in\mathcal{X}} \mathfrak{d}_{\mathcal{Y}}(\hat{f}_n(x), \hat{f}^*_n(x)) > \frac{\epsilon}{2} \right) + \mathbb{P}\left(\sup_{x\in\mathcal{X}} \mathfrak{d}_{\mathcal{Y}}(\hat{f}^*_n(x), f(x)) > \frac{\epsilon}{2} \right)$$

where $(\hat{f}^*_n)_{n\in\mathbb{N}}$ denotes the general Lipschitz interpolation framework with a hyper-parameter equal to the best Lipschitz constant L^* of f. The second term of the upper bound given above converges to 0 as $n \to \infty$ by Theorem 4.3.2. For the first

term, we have

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}} \mathfrak{d}_{\mathcal{Y}}(\hat{f}_n^*(x), \hat{f}_n(x)) > \frac{\epsilon}{2}\right)$$

$$\leq \mathbb{P}\left(\sup_{x\in\mathcal{X}} \frac{1}{2}|\min_{i=1,\dots,N_n}\{\tilde{f}_i + L(n)\,\mathfrak{d}(x,s_i)\} - \min_{i=1,\dots,N_n}\{\tilde{f}_i + L^*\,\mathfrak{d}(x,s_i)\}| > \frac{\epsilon}{4}\right)$$

$$+\mathbb{P}\left(\sup_{x\in\mathcal{X}} \frac{1}{2}|\max_{i=1,\dots,N_n}\{\tilde{f}_i - L(n)\,\mathfrak{d}(x,s_i)\} - \max_{i=1,\dots,N_n}\{\tilde{f}_i - L^*\,\mathfrak{d}(x,s_i)\}| > \frac{\epsilon}{4}\right)$$

$$\leq \mathbb{P}\left(\sup_{x\in\mathcal{X}} \mathfrak{d}(x,s_i^*)|L^* - L(n)| > \frac{\epsilon}{4}\right) + \mathbb{P}\left(\sup_{x\in\mathcal{X}} \mathfrak{d}(x,s_k^*)|L^* - L(n)| > \frac{\epsilon}{4}\right)$$

$$\leq 2\mathbb{P}\left(\delta_{\mathfrak{d}}(\mathcal{X})|L^* - L(n)| > \frac{\epsilon}{4}\right) \tag{4.2}$$

where $s_i^* := \text{argmin}_{i=1,\dots,N_n}\{\tilde{f}_i + L(n)\,\mathfrak{d}(x,s_i)\}$ and $s_k^* := \text{argmax}_{i=1,\dots,N_n}\{\tilde{f}_i - L(n)\,\mathfrak{d}(x,s_i)\}$. As $\delta_{\mathfrak{d}}(\mathcal{X})$ is finite by assumption, Lemma 4.5.2 can be applied to show that $\mathbb{P}(\delta_{\mathfrak{d}}(\mathcal{X})|L^* - L(n)| > \frac{\epsilon}{4})$ converges to 0.

■

In general, it suffices for the sequence of Lipschitz constant estimates to converge to a value that is bigger than the best Lipschitz constant in order for the consistency guarantees given in Theorem 4.5.3 to hold. This follows from the fact that Lemma 4.5.2 holds for any Lipschitz interpolation framework with $L \geq L^*$. Furthermore, if the Lipschitz constant estimate can be guaranteed to be feasible[9] in a finite number of queries and is asymptotically bounded, then the rate of convergence of the adaptive Lipschitz interpolation method matches the one derived in Theorem 4.3.5. Unfortunately, as remarked above, the LACKI rule proposed in Definition 4.5.1 is not feasible for any finite number of sample points but converges only asymptotically to the true best Lipschitz constant. One approach to remedying this problem would be to include a multiplicative factor $\kappa \geq 1$ (similar to the original approach proposed by Strongin [1973] in the noiseless sampling setting) in the LACKI rule. However, developing a principled approach to setting κ is non-trivial and depends on second order partial derivatives of the unknown target function.

Furthermore, in contrast to the general Lipschitz interpolation approach, the LACKI

[9] i.e. $\hat{L}(n) \geq L^*$.

estimator is also not necessarily asymptotically consistent in the setting of a non-linear discrete-time dynamic system. This is due to the dependency on the sampling sequence. More specifically, without uniform sampling on $\mathcal{X}$, it is possible for the Lipschitz constant estimate generated by the LACKI rule to never become sufficiently large in order to ensure that the relations (1) and (2) derived and utilised in the proof of Lemma 4.5.2 hold. In other words, the Lipschitz interpolation framework of LACKI might utilise samples drawn from the initial steps of the sampling sequence which belong to a subset of $\mathcal{X}$ that is never revisited, leading to a possible underestimation of the Lipschitz constant. This issue could potentially be fixed by including a "memory hyper-parameter" that limits the number of past observations considered in the $\mathfrak{u}_n$, $\mathfrak{l}_n$ functions. This extension will be investigated in future work.

In essence, while the general Lipschitz interpolation framework can be shown to perform well as a non-parametric estimation method, the additional difficulty of Lipschitz constant estimation implies that many of the desirable asymptotic properties become difficult to obtain for a fully adaptive version of the framework. A detailed discussion on this issue can be found in Huang et al. [2023], see Chapter 3, where optimal convergence rates are given for the Lipschitz constant estimation problem and a feasible asymptotically consistent estimation method is developed.

4.6 Connections to Online Learning and Control

We conclude this chapter by providing a theoretical illustration of how the results derived in previous sections can be utilised in the context of control-related applications. More precisely, we slightly modify the online consistency results of the general Lipschitz interpolation stated in Section 4.4 in order to obtain closed-loop stability of a class of online learning-based trajectory tracking controllers discussed in Sanner and Slotine [1991], Åström and Wittenmark [2013], Chowdhary et al. [2013], Calliess et al. [2020].

We briefly recall the setting of the trajectory tracking control problem considered by Calliess et al. [2020]. The goal is to ensure that a sequence of states $(y_n)_{n\in\mathbb{N}}$ follows

a given reference trajectory $(\xi_n)_{n\in\mathbb{N}}$. In order to do so, it is assumed that the states $(y_n)_{n\in\mathbb{N}}$ satisfy a multivariate recurrence relation described as follows:

$$y_n = f(x_n)$$

where $x_n = (y_{n-d_y}, ..., y_{n-1}, u_{n-d_u}, ..., u_n)$ with $y_i \in \mathcal{Y} \subset \mathbb{R}^l$ denoting the past autoregressive inputs, $u_i \in \mathcal{U} \subset \mathbb{R}^s$ denoting a vector of past or current control inputs for $d_y, d_u, s, l \in \mathbb{N}$. In this setting, we will therefore consider $\mathcal{X} = \mathbb{R}^{(l)(d_y)} \times \mathcal{U}^{d_u+1} \subset \mathbb{R}^{(l)(d_y)+(s)(d_u+1)}$, $\mathcal{Y} = \mathbb{R}^l$. Note that in contrast to the setting considered in Section 4.4 the noise does not impact the state and will only be assumed to be observational: we assume that the Lipschitz interpolation framework has access to noisy samples of function values $f(x_i)$ at each time step $i < n$: $D_n = \{(x_i, \tilde{f}_i) | i < n\}$.

Under this assumption on the system dynamics, the problem becomes equivalent to defining a control law that ensures that the tracking error $(\zeta_n)_{n\in\mathbb{N}}$, $\zeta_n = \xi_n - y_n$ becomes stable: obtaining, in an ideal scenario, a closed-loop recurrence relation

$$\zeta_{n+1} = \phi(\zeta_n)$$

where ϕ is a contraction with a desirable fixed point ζ_*.

This type of stability is well-known to be achievable when the dynamics of the states $(y_n)_{n\in\mathbb{N}}$ are known and sufficiently well-behaved (Åström and Wittenmark [2013]) or when f is assumed unknown but well approximated by linear learning-based methods (Limanond and Tsakllis [2000]). Obtaining such guarantees in the setting where f is assumed both unknown and non-linear is less clear although significant research has been conducted with the use of non-parametric regression methods (Sanner and Slotine [1991], Chowdhary et al. [2013], Calliess et al. [2020]).

Under a general assumption on the control law, the online-learning guarantees of the Lipschitz interpolation method (Corollary 4.4.1 and Lemma 4.6.3) derived in this chapter can be shown to directly imply the convergence of the tracking error to a fixed point, therefore ensuring the asymptotic stability of the controller.

To do so, we begin by formally extending the online guarantees of the Lipschitz interpolation stated in Corollary 4.4.1 to the multi-dimensional online setting described above. In this case, the Lipschitz interpolation framework is applied component-wise as follows:

Definition 4.6.1 *(Multi-dimensional Lipschitz interpolation) Let $L \in \mathbb{R}_{\geq 0}$ be a selected hyper-parameter. Using the set-up defined above, we define the sequence of predictors $(\hat{f}_n)_{n\in\mathbb{N}}$, $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$ associated to $(D_n)_{n\in\mathbb{N}}$, as*

$$\forall j \in \{1, ..., l\}, \quad \hat{\mathfrak{f}}^i_{\mathfrak{n}}(x) := \frac{1}{2}\mathfrak{u}^j_n(x) + \frac{1}{2}\mathfrak{l}^j_n(x),$$

where $\mathfrak{u}^j_n, \mathfrak{l}^j_n : \mathcal{X} \to \mathbb{R}$ are defined as

$$\mathfrak{u}^j_n(x) = \min_{i=1,...,N_n} \tilde{f}^j_{n,i} + L\,\mathfrak{d}(x, s_i)$$
$$\mathfrak{l}^j_n(x) = \max_{i=1,...,N_n} \tilde{f}^j_{n,i} - L\,\mathfrak{d}(x, s_i)$$

for all $j \in \{1, ..., l\}$.

We note that under Assumption 14 provided below, it is relatively straightforward to observe that each component of the target function is also Lipschitz continuous with the same Lipschitz constant. This implies that the properties utilised in the previous sections hold component-wise for the multi-dimensional Lipschitz interpolation framework.

In order to derive the desired online guarantee for the Lipschitz interpolation framework described in Definition 4.6.1, we first extend the assumptions of the previous sections to the multi-dimensional output setting.

Assumption 13 *(Assumption on multi-dimensional noise) The noise variables $(e_k)_{k\in\mathbb{N}}$, $e_k \in \mathbb{R}^d$ are assumed to be independent and identically[10] distributed random variables such that $\exists \bar{\mathfrak{e}} \in \mathbb{R}^d_+$, such that $\forall k \in \mathbb{N}$ $j \in \{1, ..., d\}$, $e^j_k \in [-\bar{\mathfrak{e}}^j, \bar{\mathfrak{e}}^j]$ with probability 1. We assume further that the bounds of the support are tight, i.e. $\forall \epsilon > 0$,*

[10]The identically distributed assumption is made to alleviate notation and is not technically needed in our derivations.

$\forall j \in \{1, ..., d\}$,

$$\mathbb{P}(e_k^j \in [-\bar{\mathfrak{e}}^j + \epsilon, \bar{\mathfrak{e}}^j]), \mathbb{P}(e_k^j \in [-\bar{\mathfrak{e}}^j, \bar{\mathfrak{e}}^j - \epsilon]) > 0.$$

Assumption 14 *(Assumption on $\mathfrak{d}_{\mathcal{Y}}$). In this section, we will restrict ourselves to the case, $\mathfrak{d}_{\mathcal{Y}}(y, y') = \|y - y'\|_1$, $\forall y, y' \in \mathcal{Y}$ where $\|.\|_1$ denotes the usual 1-norm.*

Remark 4.6.2 *By the strong equivalence of norms on $\mathbb{R}^l$, it is sufficient to show the results of this section for $\|.\|_{\mathcal{Y}} = \|.\|_1$. Additionally, we note that if a Lipschitz constant of the target function is known for a given norm on $\mathbb{R}^l$, then it is straightforward to compute a feasible Lipschitz constant for any other norm on $\mathbb{R}^l$.*

Corollary 4.6.3 *(Multidimensional Online Learning) Consider the multidimensional setting described above[11], $L^*, M^* \in \mathbb{R}^+$ and $(\hat{f}_n)_{n\in\mathbb{N}}$ as defined in Definition 4.6.1 with $L \geq L^*$ and $(\mathcal{D}_n)_{n\in\mathbb{N}} = (x_n, y_n)_{n\in\mathbb{N}}$. Assume that Assumptions 13 and 14 hold. Assume furthermore that $\mathcal{U}$ is compact. Then*

$$\lim_{n\to\infty} \sup_{f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)} \mathbb{E}[\|f(x_{n+1}) - \hat{f}_n(x_{n+1})\|_{\mathcal{Y}}] \to 0$$

where we recall that $\overline{Lip}(L^*, \mathfrak{d}, M^*)$ denotes the set containing all functions in $Lip(L^*, \mathfrak{d})$ that are bounded by M^*, i.e. $\|f(x)\|_{\mathcal{Y}} \leq M^*$.

Proof By the strong equivalence of norms on $\mathbb{R}^l$, it is sufficient to show Lemma 4.6.3 for $\|.\|_{\mathcal{Y}} = \|.\|_1$:

$$\lim_{n\to\infty} \sup_{f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)} \mathbb{E}[\|f(x_{n+1}) - \hat{f}_n(x_{n+1})\|_1] \to 0.$$

This is implied if $\forall j \in \{1, ..., l\}$:

$$\lim_{n\to\infty} \sup_{f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)} \mathbb{E}[|f^j(x_{n+1}) - \hat{f}_n^j(x_{n+1})|] \to 0$$

where $f_n^j, \hat{f}_n^j$ denote the j-th component functions of $f_n, \hat{f}_n$. This statement can

[11] With the same arguments, one can show that the same result holds for the multidimensional version of the dynamical system described in Section 4.4.

be derived using the same arguments as the ones given in the proof of Corollary 4.4.1 as, under Assumption 14, f is component-wise Lipschitz continuous with Lipschitz constant L^* and by construction, the multi-dimensional Lipschitz interpolation framework can be considered component-wise.

■

Utilising Lemma 4.6.3, we can now state the closed-loop guarantees of an online controller based on the Lipschitz interpolation framework.

Theorem 4.6.4 *Assume the settings described above. Assume that reference trajectory $(\xi_n)_{n\in\mathbb{N}}$ is bounded and that the recursive plant dynamics satisfy: $f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)$ for some $L^*, M^* > 0$. Let $(\hat{f}_n)_{n\in\mathbb{N}}$ be the predictors generated by the Lipschitz interpolation framework with hyperparameter $L \geq L^*$ and $(\mathcal{D}_n)_{n\in\mathbb{N}} = (x_n, \tilde{f}_i)_{n\in\mathbb{N}}$. If there exists a bounded control law $u_{n+1} := u(x_n, \hat{f}_n, \mathcal{D}_n)$ such that the closed loop dynamics are given by:*

$$\zeta_{n+1} = \phi(\zeta_n) + d_n$$

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

$$\lim_{n\to\infty} \mathbb{E}[\|\zeta_n - \zeta_*\|_{\mathcal{Y}}] = 0.$$

Proof The proof follows a modified version of the proof Theorem 17 of Calliess et al. [2020] and an application of Corollary 4.6.4.

Define the *nominal reference error* $(\bar{\zeta}_n)_{n\in\mathbb{N}}$, $\bar{\zeta}_0 = \zeta_0$, $\bar{\zeta}_{n+1} = \phi(\bar{\zeta}_n)$ for $n \in \mathbb{N}$. Fix an arbitrary $\epsilon > 0$. The proof of Theorem 4.6.4 follows from the following sequence of steps.

By Banach fixed point Theorem, $\exists n_0 \in \mathbb{N}$ such that $\forall n \geq n_0$, $\|\bar{\zeta}_n - \zeta_*\|_{\mathcal{Y}} < \frac{\epsilon}{3}$.

Inductively, one can show that $\forall n, k \in \mathbb{N}$,

$$\mathbb{E}[\|\zeta_k - \bar{\zeta}_k\|_{\mathcal{Y}}]$$

$$\leq \lambda_\phi^n \mathbb{E}[\|\zeta_k - \bar{\zeta}_k\|_{\mathcal{Y}}] + \sum_{i=0}^{n-1} \lambda_\phi^{n-1-i} \mathbb{E}[\|d_{k+i}\|_{\mathcal{Y}}]$$

$$\leq \lambda_\phi^n \mathbb{E}[\|\zeta_k - \bar{\zeta}_k\|_{\mathcal{Y}}] + \frac{1}{1-\lambda_\phi} \max_{i=0,\dots,n-1} \mathbb{E}[\|d_{k+i}\|_{\mathcal{Y}}].$$

By Lemma 4.6.3, we have that $\mathbb{E}[\|d_n\|_{\mathcal{Y}}]$ converges to 0 as n goes to infinity. Therefore, $\exists k_0 \in \mathbb{N}$ such that $\forall k \geq k_0$,

$$\frac{1}{1-\lambda_\phi} \max_{i=0,\dots,n-1} \mathbb{E}[\|d_{k+i}\|_{\mathcal{Y}}] \leq \frac{\epsilon}{3}.$$

Let $m_0 := \max\{n_0, k_0\}$. Since $\lambda_\phi^{q_0} < 1$, there exists $q_0 \in \mathbb{N}$ such that:

$$\lambda_\phi^{q_0} \mathbb{E}[\|\zeta_{m_0} - \bar{\zeta}_{m_0}\|_{\mathcal{Y}}] < \frac{\epsilon}{3}.$$

Let $M := m_0 + q_0$. Combining the above steps, we have for any $m > M$ there exists $n \geq q_0$ such that $m = m_0 + n$. This implies

$$\mathbb{E}[\|\zeta_m - \zeta_*\|_{\mathcal{Y}}] \leq \|\zeta_* - \bar{\zeta}_m\|_{\mathcal{Y}} + \mathbb{E}[\|\zeta_m - \bar{\zeta}_m\|_{\mathcal{Y}}]$$

$$\leq \frac{\epsilon}{3} + \lambda_\phi^{q_0} \mathbb{E}[\|\zeta_{m_0} - \bar{\zeta}_{m_0}\|_{\mathcal{Y}}] + \frac{1}{1-\lambda_\phi} \max_{i=0,\dots,n-1} \mathbb{E}[\|d_{m_0+i}\|_{\mathcal{Y}}]$$

$$\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

As the choice of ϵ was arbitrary, this concludes the proof.

■

Theorem 4.6.4 provides a theoretical result on the global stability in expectation of a general class of control problems where both the system dynamics and the error dynamics are assumed to be both unknown and non-linear. An extension of Theorem 4.6.4 which states that the convergence rates of the Lipschitz constant estimator derived in Corollary 4.4.4 hold for the tracking error $(\zeta_n)_{n\in\mathbb{N}}$ can also be obtained. However, this result would be contingent on the difficult-to-verify "regularity of the sampling" assumption of $(x_n)_{n\in\mathbb{N}}$ (as defined in Definition 4.4.3) and is therefore of limited interest. We provide it in the appendix for completeness.

Unfortunately, for numerous applications, the contraction assumption on dynamics of the tracking error: ϕ is too stringent to be observed in practice. To alleviate this issue, Theorem 4.6.4 can be extended to consider the more general assumption that ϕ is an eventually contracting function if ϕ is also assumed to be a linear. More formally, we define an eventually contracting function as follows.

Definition 4.6.5 *(Eventually Contracting Function) Let $l \in \mathbb{N}$. A continuous function $h : \mathbb{R}^l \to \mathbb{R}^l$ is said to be eventually contracting if there exists $N \in \mathbb{N}$ and $\lambda \in [0, 1)$ such that $\forall x, y \in \mathbb{R}^l$:*

$$\mathfrak{d}(h^N(x), h^N(y)) \leq \lambda\, \mathfrak{d}(x, y).$$

As with the contracting functions considered above, eventually contracting functions can be shown to admit a unique fixed point ξ^*. Additionally, it is well-known that a linear function $h : \mathbb{R}^l \to \mathbb{R}^l$ defined as $h(x) = Mx$ for some matrix $M \in \mathbb{R}^{l \times l}$ is eventually contracting if and only if the spectral radius of M is strictly smaller than 1: $\rho(M) < 1$.

The assumption of the existence of a control law $u_{n+1} := u(x_n, \hat{f}_n, \mathcal{D}_n)$ such that the closed loop dynamics are given by: $\zeta_{n+1} = \phi(\zeta_n) + d_n$and ϕ is eventually contracting can be observed in applications such as the removal of wing rock during the landing of modern fighter aircraft (Monahemi and Krstic [1996], Chowdhary et al. [2013]). Therefore, if Theorem 4.6.4 can be extended to hold under these assumptions, then, conditional on the existence of feasible Lipschitz constant estimate, the online learning-based reference tracking controllers utilising a Lipschitz interpolation can be ensured to be globally asymptotically stable in expectation in these settings.

This alternative result is stated in the following corollary.

Corollary 4.6.6 *Assume the setting and initial assumptions of Theorem 4.6.4. If there exists a bounded control law $u_{n+1} := u(x_n, \hat{f}_n, \mathcal{D}_n)$ such that the closed loop dynamics are given by:*

$$\zeta_{n+1} = \phi(\zeta_n) + d_n$$

where $d_n := f(x_n) - \hat{f}(x_n)$ *is the one-step prediction error and* $\phi : \mathbb{R}^l \to \mathbb{R}^l$, $\phi(\zeta) = M\zeta$ *for a matrix* $M \in \mathbb{R}^{l\times l}$ *that is a stable, i.e.* $\rho(M) < 1$. *Then*

$$\lim_{n\to\infty} \mathbb{E}[\|\zeta_n\|_{\mathcal{Y}}] = 0.$$

Proof The proof of Corollary 4.6.6 is similar to the on given for Theorem 4.6.4.

Define the nominal reference error $(\bar{\zeta}_n)_{n\in\mathbb{N}}$, $\bar{\zeta}_0 = \zeta_0$, $\bar{\zeta}_{n+1} = \phi(\bar{\zeta}_n)$ for $n \in \mathbb{N}$. Fix an arbitrary $\epsilon > 0$.

As $\rho(M) < 1$, we have that $\lim_{n\to\infty} \bar{\zeta}_n = 0$ (Hasselblatt and Katok [2003]). This implies that $\exists n_0 \in \mathbb{N}$ such that $\forall n \geq n_0$, $\|\bar{\zeta}_n\|_{\mathcal{Y}} < \frac{\epsilon}{3}$.

Inductively, one can show that $\forall n, k \in \mathbb{N}$,

$$\mathbb{E}[\|\zeta_{n+k} - \bar{\zeta}_{n+k}\|_{\mathcal{Y}}]$$

$$\leq \|M^n\|_{\mathcal{Y}}\mathbb{E}\left[\|(\zeta_k - \bar{\zeta}_k)\|_{\mathcal{Y}}\right] + \sum_{i=0}^{n-1} \|M^{n-1-i}\|_{\mathcal{Y}}\mathbb{E}[\|d_{k+i}\|_{\mathcal{Y}}].$$

By Gelfand's formula we have $\lim_{n\to\infty} \|M^k\|^{\frac{1}{k}} = \rho(M) < 1$ for any matrix norm $\|.\|$. This implies that there exists n_1 such that for all $n \geq n_1$: $\|M^n\|_{\mathcal{Y}} \leq \|M^n\|_{\mathcal{Y}}^{\frac{1}{n}} < \lambda_\phi < 1$ for some $\lambda_\phi \in (0, 1)$. Utilising this relation and matrix-vector inequalities, we obtain the following inequalities: let $n \geq n_1$, there exists $n_2 \in \mathbb{N} \cup \{0\}$ such that $n = n_1 \left\lfloor \frac{n}{n_1} \right\rfloor + n_2$:

$$\|M^n\|_{\mathcal{Y}} = \|M^{n_1(\left\lfloor \frac{n}{n_1} \right\rfloor - 1)} M^{n_1+n_2}\|_{\mathcal{Y}} \leq \|M^{n_1}\|_{\mathcal{Y}}^{(\left\lfloor \frac{n}{n_1} \right\rfloor - 1)} \|M^{n_1+n_2}\|_{\mathcal{Y}} \leq \lambda_\phi^{(\left\lfloor \frac{n}{n_1} \right\rfloor - 1)} \lambda_\phi = \lambda_\phi^{\left\lfloor \frac{n}{n_1} \right\rfloor}.$$

Substituting this inequality into the bound given above, we obtain:

$$\|M^n\|_{\mathcal{Y}}\mathbb{E}\left[\|(\zeta_k - \bar{\zeta}_k)\|_{\mathcal{Y}}\right] + \sum_{i=0}^{n-1} \|M^{n-1-i}\|_{\mathcal{Y}}\mathbb{E}[\|d_{k+i}\|_{\mathcal{Y}}]$$

$$\leq \lambda_\phi^{\left\lfloor \frac{n}{n_1} \right\rfloor}\mathbb{E}[\|(\zeta_k - \bar{\zeta}_k)\|_{\mathcal{Y}}] + \max_{i=0,\ldots,n-1} \mathbb{E}[\|d_{k+i}\|_{\mathcal{Y}}] \left(K_{n_1} + \sum_{i=1}^{\left\lceil \frac{n-1}{n_1} \right\rceil} n_1 \lambda_\phi^i \right)$$

$$\leq \lambda_\phi^{\lfloor \frac{n}{n_1} \rfloor} \mathbb{E}[\|(\zeta_k - \bar{\zeta}_k)\|_{\mathcal{Y}}] + \max_{i=0,\dots,n-1} \mathbb{E}[\|d_{k+i}\|_{\mathcal{Y}}] \left(K_{n_1} + \frac{n_1}{1-\lambda_\phi} \right)$$

where $K_{n_1} := \sum_{i=0}^{n_1-1} \|M^i\|_{\mathcal{Y}}$ By Lemma 4.6.3, we have that $\mathbb{E}[\|d_n\|_{\mathcal{Y}}]$ converges to 0 as n goes to infinity. Therefore, $\exists k_0 \in \mathbb{N}$ such that $\forall k \geq k_0$,

$$\left(K_{n_1} + \frac{n_1}{1-\lambda_\phi} \right) \max_{i=0,\dots,n-1} \mathbb{E}[\|d_{k+i}\|_{\mathcal{Y}}] \leq \frac{\epsilon}{3}.$$

Let $m_0 := \max\{n_0, k_0\}$. There exists $q_0 \in \mathbb{N}$ such that

$$\lambda_\phi^{\lfloor \frac{q_0}{n_1} \rfloor} \mathbb{E}[\|\zeta_{m_0} - \bar{\zeta}_{m_0}\|_{\mathcal{Y}}] < \frac{\epsilon}{3}.$$

Let $M := m_0 + q_0$. Combining the above steps, we have that for all $m > M$, there exists $n \geq q_0$ such that $m = m_0 + n$. This implies

$$\mathbb{E}[\|\zeta_m\|_{\mathcal{Y}}] \leq \|\bar{\zeta}_m\|_{\mathcal{Y}} + \mathbb{E}[\|\zeta_m - \bar{\zeta}_m\|_{\mathcal{Y}}]$$

$$\leq \frac{\epsilon}{3} + \lambda_\phi^{\lfloor \frac{q_0}{n_1} \rfloor} \mathbb{E}[\|\zeta_{m_0} - \bar{\zeta}_{m_0}\|_{\mathcal{Y}}] + \left(K_{n_1} + \frac{n_1}{1-\lambda_\phi} \right) \max_{i=0,\dots,n-1} \mathbb{E}[\|d_{m_0+i}\|_{\mathcal{Y}}]$$

$$\leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

As the choice of ϵ was arbitrary, this concludes the proof.

■

4.6.1 Example - model-reference adaptive control of a single pendulum

As a simple illustration, we replicate a modification of the model-reference adaptive control example in Calliess et al. [2020]. Here, we control an Euler discretisation of a torque-actuated single pendulum in set-point control mode:

We consider a torque controlled pendulum where forces u can be applied to the joint of the pendulum. The angle of the pendulum is called q. We define a state $x = [q\dot{q}]$. In continuous-time, it's dynamics are given by the ODE $\ddot{q} = f(x) + u$

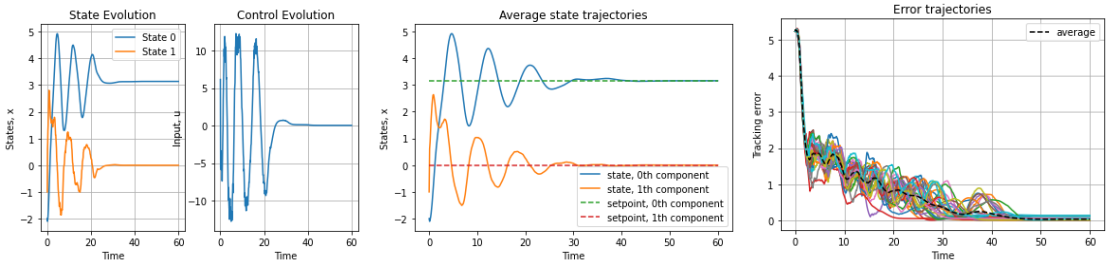


Figure 4.61: Illustration of the pendulum control example. A single run is depicted in the leftmost figure showing how the controller learns to drive the state to the set-point in spite of the noise and initially uncertain dynamics. Note, "Time" is simulation time $t = \Delta n$ [sec.] for discrete time steps $n = 0, 1, \ldots$. The second figure shows how the mean trajectories, averaged over 30 repetitions (each with new draws from the noise distribution), converge to the set-point. An illustration of our theory, predicting vanishing tracking errors in the mean, is depicted in the rightmost figure: For each repetition of the experiment, the colored lines show the error trajectories $(\|\xi(\Delta n) - x(\Delta n)\|)_{n=0,1,2,\ldots}$ as well as their empirical mean (black dashed line).

where $f(x) = -\sin(q) - \dot{q}$ may be uncertain a priori and hence, needs to be learned online while we control.

As explained in Section 4 of Calliess et al. [2020], using online learning of noisy measurements of angular accelerations (but assuming full state observability) we can use Lipschitz interpolation to learn a model $\hat{f}_n$ at time step n define a control law $u(x) = -\hat{f}_n(x) - K_1 x_1 - K_2 x_2$ with gains $K_1, K_2 > 0$ such that the closed-loop error dynamics becomes:

$$\zeta_{n+1} = \phi(\zeta_n) + \Delta d_n \tag{4.3}$$

$$= \underbrace{M\zeta_n}_{=:\phi(\zeta_n)} + \Delta d_n \tag{4.4}$$

where $\Delta = 0.1$ is a sampling period for the time discretisation, $d_n = f(x_n) - \hat{f}_n(x_n)$ is the tracking error and $M = \begin{pmatrix} 1 & \Delta \\ -\Delta K_1 & 1 - \Delta K_2 \end{pmatrix}$ is a matrix, where the gain parameters $K_1, K_2 = 1$ were chosen to render M stable (i.e. such that its spectral radius $\rho(M) < 1$). This renders the closed-loop tracking error dynamics consistent with the one considered in Corollary 4.6.6 and as previously discussed, implies that the error dynamics are an eventual contraction.

For this experiment, we chose Lipschitz interpolation with a fixed Lipschitz constant $L = 11$ which is a Lipschitz constant of the true dynamics. To learn f the Lipschitz interpolator had access to acceleration corrupted by uniformly distributed noise drawn i.i.d. from the interval [-2,2] given a performance example in a relatively low signal-to-noise ration setting. Starting in initial state $x_0 = [-2., -1.]$, the controller was given a set-point reference $\xi_n = [2\pi, 0], \forall n \in \mathbb{N}$. An example run of the controller as well as empirical measurements of the tracking dynamics across 30 trials of the experiments for different noise realisations are given in Figure 4.61. Note, consistent with our theory, the plots show how the tracking error appears to vanish in the mean (and in fact for all realisations).

Before we conclude, we will need to point out some limitations to the online control application given in this section:

Remark 4.6.7 *Firstly, our example assumed knowledge of the Lipschitz constant. While not an uncommon assumption, we recognise that having to know the true Lipschitz constant is a practical limitation. Therefore methods such as LACKI (Calliess et al. [2020]) or POKI (Calliess [2017]) could also be employed to incorporate full black-box learning. Our example however, merely serves as a simple illustration of our currently developed theory. Secondly, we assumed that states were observable but that only accelerations were noisy.*

Extending our results to learning-based control settings that involve Lipschitz constant parameter estimation and extending the theory to realistic settings of noisy state observations would, in our opinion, be an interesting direction to investigate in future work.

4.7 Conclusion

In conclusion, this chapter has provided a comprehensive investigation into the asymptotic convergence properties of general Lipschitz interpolation methods in the presence of bounded stochastic noise. Through our analysis, we have established

probabilistic consistency guarantees for the classical approach within a broad context. Furthermore, by deriving upper bounds on the uniform convergence rates, we have aligned these bounds with the well-established optimal rates of non-parametric regression observed in similar settings. These rates provide a precise characterisation of the impact of the behaviour of the noise at the boundary of its support on the non-parametric uniform convergence rates and are, as far as the authors of this chapter. are aware, novel to the literature.

These established bounds can also serve as useful tools for the comparative asymptotic assessment of Lipschitz interpolation against alternative non-parametric regression techniques determining the circumstances under which Lipschitz interpolation frameworks can be anticipated to be asymptotically better or worse. In particular, an explicit condition on the noise's behaviour at the boundary of its support can be utilised to predict this out-or under-performance.

Extending our work, we have expanded our asymptotic results to consider online learning in discrete-time stochastic systems. The additional consistency guarantees we provide in this context carry practical significance, as we show how they can be utilised to establish closed-loop stability assurances for a simple online learning-based controller in the setting of model reference adaptive control. We note that these asymptotic results also hold for the worst-case upper and lower bounds provided by the ceiling and floor predictors of the Lipschitz interpolation framework. This implies that even the most conservative adaptive controllers relying on worst-case bounds of Lipschitz interpolation methods will consider the true underlying dynamics in the long run.

Finally, we have provided a brief theoretical study of the fully data-driven LACKI framework (Calliess et al. [2020]) which extends classical Lipschitz interpolation by incorporating a Lipschitz constant estimation mechanism into the algorithm. We show asymptotic consistency of both the Lipschiz constant estimation method and the extended framework which can serve to further theoretically motivate the use of the LACKI in practice.

Appendices

Contents

Appendix 4.A Additional Results (Convergence rate of tracking error)

We provide the theoretical convergence rates obtained for the tracking error in the application to online learning-based control. As noted in Section 4.6, this result is not generally applicable as verification of the "regular sampling" condition defined in Definition 4.4.3 is difficult to do in practice.

Corollary 4.A.1 *Assume that the setting and assumptions of Corollary 4.6.3 hold. Assume furthermore that the stochastic control law* $u_{n+1} := u(x_n, \hat{f}_n, \mathcal{D}_n)$ *is defined such that* $(x_n)_{n\in\mathbb{N}}$ *is regularly sampled on a set* $\bar{\mathcal{X}} \subset \mathcal{X}$ *that satisfies Assumption 10 and that the noise vectors* $(\epsilon_n)_{n\in\mathbb{N}}$ *are component-wise independent. Then,*

$$\limsup_{n\to\infty} \mathbb{E}[a_n^{-1} \| f(x_{n+1}) - \hat{f}_n(x_{n+1}) \| y] < \infty.$$

where $(a_n)_{n\in\mathbb{N}} := ((n^{-1}log(n))^{\frac{\alpha}{d+\eta\alpha}})_{n\in\mathbb{N}}$.

Proof (sketch) Applying the same arguments as in the proof of Lemma 4.6.3, it is sufficient to consider: $\forall j \in \{1, ..., l\}$

$$\limsup_{n\to\infty} \mathbb{E}[a_n^{-1}|f^j(x_{n+1}) - \hat{f}_n^j(x_{n+1})|].$$

Since the noise is component-wise independent, we can apply the arguments utilised in the proof of Corollary 4.4.4 for all $j \in \{1, ..., l\}$ and conclude the proof.

■

Theorem 4.A.2 *Assume the settings and assumptions of Theorem 4.6.4 hold. If the stochastic control law* $u_{n+1} := u(x_n, \hat{f}_n, \mathcal{D}_n)$ *is such that* $(x_n)_{n\in\mathbb{N}}$ *is regularly sampled on a set* $\bar{\mathcal{X}} \subset \mathcal{X}$ *that satisfies Assumption 10 and the noise vectors* $(\epsilon_n)_{n\in\mathbb{N}}$ *are component-wise independent. Then,*

$$\limsup_{n\to\infty} \mathbb{E}[a_n^{-1}\|e_n - e_*\|_{\mathcal{Y}}] < \infty$$

where $(a_n)_{n\in\mathbb{N}} := ((n^{-1}log(n))^{\frac{\alpha}{d+\eta\alpha}})_{n\in\mathbb{N}}$.

Proof (sketch) Follows from applying the proof of Theorem 4.6.4 and noting that the slowest converging term at the end of the proof is given by

$$\frac{1}{1-\lambda_\phi} \max_{i=0,...,n-1} \mathbb{E}[\|d_{m_0+i}\|_{\mathcal{Y}}].$$

This term can be upper bounded by applying Corollary 4.A.1 which therefore provides the convergence rate and concludes the proof.

■

Appendix 4.B Proof of Theorem 4.3.5

Proof From the proof of Theorem 4.3.2, we have $\forall f \in Lip(L^*, \mathfrak{d}), x \in \mathcal{X}$,

$$|\hat{f}_n(x) - f(x)|$$

$$\leq \max\Big\{ \min_{i=1,\dots,N_n} \{\frac{e_i}{2} + \frac{L^* + L}{2}\mathfrak{d}(x, s_i)\} + \max_{i=1,\dots,N_n} \{\frac{e_i}{2}\},$$

$$- \min_{i=1,\dots,N_n} \{\frac{e_i}{2}\} - \max_{i=1,\dots,N_n} \{\frac{e_i}{2} - \frac{L^* + L}{2}\mathfrak{d}(x, s_i)\}\Big\}.$$

Consider the minimal covering of $\mathcal{X}$ of radius $R_n := a_n = (n^{-1}log(n))^{\frac{\alpha}{d+\eta\alpha}}$ with respect to $\mathfrak{d}$ denoted $Cov(R_n)$ and the associated set of hyperballs; $\mathcal{B}_n$. Assuming that n is large enough such that every hyperball in $\mathcal{B}_n$ contains at least one input of $G_n^{\mathcal{X}}$, we have that the following upper bound holds:

$$|\hat{f}_n(x) - f(x)|$$

$$\leq \max\left\{ \min_{s_i \in B^x \cap G_n^{\mathcal{X}}} \{\frac{e(s_i)}{2}\}, \min_{s_i \in B^x \cap G_n^{\mathcal{X}}} \{-\frac{e(s_i)}{2}\} \right\} + \frac{\bar{\mathfrak{e}}}{2} + (L^* + L)R_n$$

where with abuse of notation, $e(s_i)$ denotes the noise variable associated with the input s_i and $B^x \in \mathcal{B}_n$ such that $x \in B$. For all $n \in \mathbb{N}$, we define the following random variable and event

$$A_n := \max_{B \in \mathcal{B}_n} \max\left\{ \min_{s_i \in B \cap G_n^{\mathcal{X}}} \{\frac{e(s_i)}{2}\}, - \max_{s_i \in B \cap G_n^{\mathcal{X}}} \{\frac{e(s_i)}{2}\} \right\} + \frac{\bar{\mathfrak{e}}}{2}$$

$$E_n := \{\forall B \in \mathcal{B}_n, |\{i \in [n]\}, s_i \in B\}| > 0\}\,.$$

Then, from $(\star)$ given at the end of the proof we have that it is sufficient to consider

$$\sup_{f \in Lip(L^*, \mathfrak{d})} \mathbb{E}\left[a_n^{-1} \|\hat{f}_n - f\|_\infty \Big| E_n \right] \mathbb{P}(E_n)$$

in order for Theorem 4.3.5 to hold. For $n \in \mathbb{N}$ sufficiently large such that $\mathbb{P}(E_n) > 0$ (see $(\star)$), we can apply the upper bound on $|\hat{f}_n(x) - f(x)|$ derived above:

$$\sup_{f \in Lip(L^*, \mathfrak{d})} \mathbb{E}\left[a_n^{-1} \|\hat{f}_n - f\|_\infty \Big| E_n \right] = \sup_{f \in Lip(L^*, \mathfrak{d})} \mathbb{E}\left[a_n^{-1} \sup_{x \in \mathcal{X}} |f_n(x) - f(x)| \Big| E_n \right]$$

$$\leq a_n^{-1} \sup_{f \in Lip(L^*, \mathfrak{d})} \mathbb{E}\left[\sup_{x \in \mathcal{X}} \max\left\{\min_{s_i \in B^x \cap G_n^{\mathcal{X}}} \{\frac{e(s_i)}{2}\}, \min_{s_i \in B^x \cap G_n^{\mathcal{X}}} \{-\frac{e(s_i)}{2}\}\right\} + \frac{\bar{\mathfrak{e}}}{2} + (L^* + L)R_n \middle| E_n\right]$$

$$= a_n^{-1}(L^* + L)R_n + \mathbb{E}\left[a_n^{-1} \max_{B \in \mathcal{B}_n} \max\left\{\min_{s_i \in B \cap G_n^{\mathcal{X}}} \{\frac{e(s_i)}{2}\}, \min_{s_i \in B \cap G_n^{\mathcal{X}}} \{-\frac{e(s_i)}{2}\}\right\} + \frac{\bar{\mathfrak{e}}}{2} \middle| E_n\right]$$

$$= (L^* + L)a_n^{-1}R_n + \mathbb{E}\left[a_n^{-1}A_n \middle| E_n\right].$$

By definition of R_n, the first term: $(L^* + L)a_n^{-1}R_n = (L^* + L)$ is bounded for all $n \in \mathbb{N}$. We can therefore focus on upper bounding the second term:

$$\mathbb{E}\left[a_n^{-1}A_n \middle| E_n\right] \mathbb{P}(E_n) = \mathbb{E}\left[a_n^{-1}A_n\right].$$

Using $0 \leq A_n \leq 2\bar{\mathfrak{e}}$ with probability 1, we have $\forall C_0 > 0$,

$$a_n^{-1}A_n \leq C_0 1_{\{a_n^{-1}A_n \leq C_0\}} + 2\bar{\mathfrak{e}}a_n^{-1}1_{\{a_n^{-1}A_n > C_0\}}$$

with probability 1. This implies that

$$\mathbb{E}[a_n^{-1}A_n] \leq C_0 + 2\bar{\mathfrak{e}}a_n^{-1}\mathbb{P}(A_n > C_0 a_n).$$

It is therefore sufficient to show that $\exists C_0 > 0$ such that

$$\limsup_{n \to \infty} \sup_{f \in Lip(L^*, \mathfrak{d})} a_n^{-1}\mathbb{P}(A_n > C_0 a_n) < \infty.$$

We have

$$\mathbb{P}(A_n > C_0 a_n) = 1 - \mathbb{P}(\forall B \in \mathcal{B}_n : \min_{s_i \in B \cap G_n^{\mathcal{X}}} e(s_i) \in I_1, \max_{s_i \in B \cap G_n^{\mathcal{X}}} e(s_i) \in I_2)$$

$$\overset{(\star\star)}{\leq} 1 - \prod_{B \in \mathcal{B}_n} \mathbb{P}\left(\min_{s_i \in B \cap G_n^{\mathcal{X}}} e(s_i) \in I_1, \max_{s_i \in B \cap G_n^{\mathcal{X}}} e(s_i) \in I_2\right)$$

$$\leq 1 - \mathbb{P}\left(\min_{i \in 1,\ldots,N_{\mathcal{B}_n}} e_i \in I_1, \max_{i \in 1,\ldots,N_{\mathcal{B}_n}} e_i \in I_2\right)^{|\mathcal{B}_n|}$$

where $I_1 := [-\bar{\mathfrak{e}}, -\bar{\mathfrak{e}} + 2C_0 a_n)$, $I_2 := (\bar{\mathfrak{e}} - 2C_0 a_n, \bar{\mathfrak{e}}]$ and $N_{\mathcal{B}_n} := \min_{B \in \mathcal{B}_n} |B \cap G_n^{\mathcal{X}}|$. The second to last inequality follows from arguments given in $(\star\star)$ provided at the

end of the proof. For n large enough such that $2C_0a_n < \bar{\epsilon}$, we can apply Assumption 8 to simplify the left hand expression as follows;

$$\mathbb{P}\left(\min_{i\in 1,\dots,N_{\mathcal{B}_n}} e_i \in I_1, \max_{i\in 1,\dots,N_{\mathcal{B}_n}} e_i \in I_2\right)$$

$$\geq \mathbb{P}\left(\min_{i\in 1,\dots,N_{\mathcal{B}_n}} e_i \in I_1\right) \cdot \mathbb{P}\left(\max_{i\in 1,\dots,N_{\mathcal{B}_n}} e_i \in I_2 | \min_{i\in 1,\dots,N_{\mathcal{B}_n}} e_i \in I_1\right)$$

$$\geq \left(1-(1-\gamma(2C_0a_n)^\eta)^{N_{\mathcal{B}_n}}\right)\left(1-(1-\gamma(2C_0a_n)^\eta)^{N_{\mathcal{B}_n}-1}\right)$$

$$\geq \left(1-2\left(1-\gamma(2C_0a_n)^\eta\right)^{N_{\mathcal{B}_n}-1}\right)^2.$$

Therefore, we have

$$\sup_{f\in Lip(L^*,\mathfrak{d})} a_n^{-1}\mathbb{P}(A_n > C_0a_n)$$

$$\leq a_n^{-1}\left(1-\left(1-2(1-\gamma(2C_0a_n)^\eta)^{N_{\mathcal{B}_n}-1}\right)^{2|\mathcal{B}_n|}\right).$$

$$\leq a_n^{-1}\left(1-\left(1-2(1-\gamma(2C_0a_n)^\eta)^{N_{\mathcal{B}_n}-1}\right)^{\frac{C_1}{a_n{}^{\frac{d}{\alpha}}}}\right).$$

where we used the fact that there exists a constant $C_1 > 0$ (that can depend on d) such that $2|\mathcal{B}_n| \leq \frac{C_1}{R_n{}^{\frac{d}{\alpha}}} = \frac{C_1}{a_n{}^{\frac{d}{\alpha}}}$ which is a modification of (Wu [2017], Theorem 14.2) that follows from the assumed convexity of $\mathcal{X}$. By Lemma 4.C.2, in order for the above expression to be bounded, it is sufficient that $2(1-\gamma(2C_0a_n)^\eta)^{N_{\mathcal{B}_n}-1}$ behaves like $C_2'a_n{}^{(\frac{d}{\alpha}+1)}$ for an arbitrary $C_2' > 0$ as n goes to infinity. More precisely, let $C_2' = 1$, it is sufficient to have:

$$2(1-\gamma(2C_0a_n)^\eta)^{N_{\mathcal{B}_n}-1} \leq a_n{}^{(\frac{d}{\alpha}+1)}$$

$$\iff N_{\mathcal{B}_n} \geq 1 + (\frac{d}{\alpha}+1)\frac{\log(a_n)}{\log(1-\gamma(2C_0a_n)^\eta)}$$

as n goes to infinity. The right-hand expression can be re-expressed as the series expansion:

$$\frac{d+\alpha}{\alpha\gamma(2C_0)^\eta}\frac{1}{a_n{}^\eta}\log(\frac{1}{a_n}) + O(\log(\frac{1}{a_n}))$$

as a_n goes to 0. Therefore, for any $C_2 > \frac{d+\alpha}{\alpha\gamma(2C_0)^\eta}$ and $n > 0$ large enough, we have $1 + (\frac{d}{\alpha} + 1)\frac{\log(a_n)}{\log(1-\gamma(2C_0a_n)^\eta)} < C_2 \log(\frac{1}{a_n})\frac{1}{{a_n}^\eta}$ and it suffices to have

$$N_{\mathcal{B}_n} \geq C_2 \log(\frac{1}{a_n})\frac{1}{{a_n}^\eta}$$

as n goes to infinity in order for $\lim_{n\to\infty} \sup_{f\in Lip(L^*,\mathfrak{d})} a_n^{-1}\mathbb{P}(A_n > C_0a_n)$ to be bounded:
If $N_{\mathcal{B}_n} \geq C_2 \log(\frac{1}{a_n})\frac{1}{{a_n}^\eta}$,

$$\limsup_{n\to\infty} \sup_{f\in Lip(L^*,\mathfrak{d})} a_n^{-1}\mathbb{P}(A_n > C_0a_n) \leq a_n^{-1}\left(1 - \left(1 - {a_n}^{(\frac{d}{\alpha}+1)}\right)^{\frac{C_1}{{a_n}^{\frac{d}{\alpha}}}}\right) \leq 2C_1.$$

where the last inequality follows from Lemma 4.C.2.

Therefore, the final step of the proof is to show that $N_{\mathcal{B}_n} \geq C_2 \log(\frac{1}{a_n})\frac{1}{{a_n}^\eta}$ occurs with a probability that converges to 1 at a rate of a_n as n goes to infinity.

More precisely, let $n \in \mathbb{N}$ and fix an arbitrary constant $C_2 > \frac{d+\alpha}{\alpha\gamma(2C_0)^\eta}$ based on the condition given above (note that $C_0 > 0$ can be set arbitrarily large).

$$S_n := \left\{\forall B \in \mathcal{B}_n : \ |\{i \in [n]\}, s_i \in B\}| > C_2 \log(\frac{1}{a_n})\frac{1}{{a_n}^\eta}\right\}$$

i.e. the event that there is more than $C_2 \log(\frac{1}{a_n})\frac{1}{{a_n}^\eta}$ queries in each hyperball in $\mathcal{B}_n$. Utilising the asymptotic bound developed in the first part of the proof, we have by the law of total probability:

$$\limsup_{n\to\infty} \mathbb{E}[a_n^{-1}A_n] \leq C_0 + 2\bar{\mathfrak{e}} \limsup_{n\to\infty} a_n^{-1}\mathbb{P}(A_n > C_0a_n)$$

$$\leq C_0 + 2\bar{\mathfrak{e}} \limsup_{n\to\infty} a_n^{-1}\left(\mathbb{P}(A_n > C_0a_n|S_n) + \mathbb{P}(S_n^c)\right) \leq C_0 + 4\bar{\mathfrak{e}}C_1 + \limsup_{n\to\infty} a_n^{-1}\mathbb{P}(S_n^c).$$

where the last inequality can be obtained by applying Lemma 4.C.2.

To conclude the proof, we need to show that $\limsup_{n\to\infty} a_n^{-1}\mathbb{P}(S_n^c)$ is bounded. We have (denoting $b_n := \log(\frac{1}{a_n})\frac{1}{{a_n}^\eta}$ to alleviate notation):

$$\mathbb{P}(S_n) = \mathbb{P}\left(\forall B \in \mathcal{B}_n : |\{i \in [n]\}, s_i \in B\}| > C_2b_n\right)$$

$$\geq \prod_{B\in\mathcal{B}_n} \mathbb{P}\left(|\{i \in [\left\lfloor \frac{n}{2} \right\rfloor]\}, s_i \in B\}| > C_2 b_n\right)$$

$$\geq \prod_{B\in\mathcal{B}_n} \left(1 - \mathbb{P}\left(|\{i \in [\left\lfloor \frac{n}{2} \right\rfloor]\}, s_i \in B\}| \leq C_2 b_n\right)\right)$$

where the first inequality stated above follows from $(\star\star\star)$ (shown at the end of the proof) for $n \in \mathbb{N}$ if C_2 satisfies the condition: $C_2 \leq \frac{d+\eta\alpha}{2C_1\alpha}$ where C_1, d, α, η are constants.

Then, Assumption 10 on $\mathcal{X}$ implies that there $r_0 > 0, \theta \in (0,1]$ such that $\forall x \in \mathcal{X}$, $r \in (0, r_0)$, $\mathrm{vol}\left(B_r(x) \cap \mathcal{X}\right) \geq \theta\, \mathrm{vol}\left(B_r(x)\right)$. Therefore, for all $n \in \mathbb{N}$ such that $R_n < r_0$, we have that Assumption 10 can be applied to $B \in \mathcal{B}_n$. Using Assumption 11, we have that the random variable defined by $|\{i \in [\left\lfloor \frac{n}{2} \right\rfloor]\}, s_i \in B\}|$ follows a binomial distribution with a success probability p that can be lower bounded by $C_3' \frac{vol(B)}{vol(\mathcal{X})} = C_3'' a_n^{\frac{d}{\alpha}}$ for $C_3', C_3'' > 0$ and with expectation:

$$\mathbb{E}\left[|\{i \in [\left\lfloor \frac{n}{2} \right\rfloor]\}, s_i \in B\}|\right] = \left\lfloor \frac{n}{2} \right\rfloor p \geq \frac{C_3''}{3} n^{\frac{\eta\alpha}{\eta\alpha+d}} \log(n)^{\frac{d}{d+\eta\alpha}} = C_3 n^{\frac{\eta\alpha}{\eta\alpha+d}} \log(n)^{\frac{d}{d+\eta\alpha}}$$

where $C_3 := \frac{C_3''}{3}$. (Note: the "3" denominator was arbitrarily selected in order to remove the ceiling operator in the above equation).

As the only condition on C_2 is given by the bound $C_2 > \frac{d+\alpha\xi}{\alpha\gamma(2C_0)^\eta}$, C_2 can be set arbitrarily small as C_0 can be set arbitrarily large. Therefore, there exists $C_0 > 0$, $C_2 > 0$ such that $C_2 \leq \min\{\frac{d+\eta\alpha}{2C_1\alpha}, \frac{C_3(d+\eta\alpha)}{\alpha}\}$ which implies that $(\star\star\star)$ holds and that:

$$C_2 b_n = C_2 \left(\frac{\alpha}{d+\eta\alpha}\right) \log\left(\frac{n}{\log(n)}\right) \left(\frac{n}{\log(n)}\right)^{\frac{\eta\alpha}{\eta\alpha+d}}$$

$$\leq C_2 \left(\frac{\alpha}{d+\eta\alpha}\right) n^{\frac{\eta\alpha}{\eta\alpha+d}} \log(n)^{\frac{d}{d+\eta\alpha}} \leq C_3 n^{\frac{\eta\alpha}{\eta\alpha+d}} \log(n)^{\frac{d}{d+\eta\alpha}} \leq \frac{\mathbb{E}[|\{i \in [n]\}, s_i \in B\}|]}{2}.$$

This last relation implies that we can apply Lemma 1 of Stone [1982] to obtain the upper bound:

$$\mathbb{P}\left(|\{i \in [n]\}, s_i \in B\}| \leq C_2 b_n\right) \leq \mathbb{P}\left(|\{i \in [n]\}, s_i \in B\}| \leq \frac{\mathbb{E}\left[|\{i \in [n]\}, s_i \in B\}|\right]}{2}\right)$$

$$\leq \left(\frac{2}{e}\right)^{\frac{\mathbb{E}[|\{i\in[n]\},s_i\in B\}|]}{2}}$$

which in turn implies

$$(1-\mathbb{P}\left(|\{i\in[n]\},s_i\in B\}|\leq C_2 b_n\right))^{|\mathcal{B}_n|}\geq\left(1-\left(\frac{2}{e}\right)^{\frac{\mathbb{E}[|\{i\in[n]\},s_i\in B\}|]}{2}}\right)^{|\mathcal{B}_n|}.$$

Plugging this expression back into $\limsup_{n\to\infty} a_n^{-1}\mathbb{P}(S_n^c)$, we obtain

$$a_n^{-1}\mathbb{P}\left(S_n^c\right)\leq a_n^{-1}\left(1-\left(1-(\frac{2}{e})^{\frac{\mathbb{E}[|\{i\in[n]\},s_i\in B\}|]}{2}}\right)^{|\mathcal{B}_n|}\right).$$

$$\leq a_n^{-1}\left(1-\left(1-(\frac{2}{e})^{\frac{C_3}{2}n^{\frac{\eta\alpha}{\eta\alpha+d}}\log(n)^{\frac{d}{d+\eta\alpha}}}\right)^{C_1 a_n{}^{\frac{d}{\alpha}}}\right)$$

which converges to 0 as n goes infinity and concludes the proof (follows from the exponential speed of convergence of $(\frac{2}{e})^{\frac{C_3}{2}n^{\frac{\eta\alpha}{\eta\alpha+d}}\log(n)^{\frac{d}{d+\eta\alpha}}}$).

($\star$) For completeness we revisit the assumption made in the proof. Recall for all $n\in\mathbb{N}$,

$$E_n:=\{\forall B\in\mathcal{B}_n,|\{i\in[n]\},s_i\in B\}|>0\}.$$

Then, by law of total expectation, we have $\forall f\in Lip(L^*,\mathfrak{d})$, $n\in\mathbb{N}$ sufficiently large such that $\mathbb{P}(E_n)>0$ (which exists since $\mathbb{P}(E_n)>\mathbb{P}(S_n)\overset{n\to\infty}{\to}1$);

$$\mathbb{E}[a_n^{-1}\|\hat{f}_n-f\|_\infty]$$

$$=a_n^{-1}(\mathbb{E}\left[\|\hat{f}_n-f\|_\infty\middle|E_n\right]\mathbb{P}(E_n)+\mathbb{E}\left[\|\hat{f}_n-f\|_\infty\middle|E_n^c\right]\mathbb{P}(E_n^c)).$$

For all $n\geq 1$, $f\in Lip(L^*,\mathfrak{d})$ and any sampling procedure $\mathcal{D}_n$, $\|\hat{f}_n-f\|_\infty$ is uniformly bounded with probability 1. More precisely, we have $\sup_{f\in Lip(L^*,\mathfrak{d})}\|\hat{f}_n-f\|_\infty\leq 2\overline{\mathfrak{e}}+2L\delta_{\mathfrak{d}}(\mathcal{X})$ where $\delta_{\mathfrak{d}}(\mathcal{X}):=\sup_{x,x'\in\mathcal{X}}\mathfrak{d}(x,x')$ with probability 1 which follows from $f\in Lip(L^*,\mathfrak{d})$ and by the design of the Lipschitz interpolation framework. This bound is finite by the assumed compactness of $\mathcal{X}$. Therefore, denoting $K:=$

$2\bar{\mathfrak{e}} + 2L\delta_{\mathfrak{d}}(\mathcal{X})$, we have that the above statement is upper bounded by

$$\mathbb{E}\left[a_n^{-1}\|\hat{f}_n - f\|_\infty \middle| E_n\right]\mathbb{P}(E_n) + a_n^{-1}K\mathbb{P}(E_n^c).$$

The first term is equal to the simplified expression assumed earlier in the proof and the second term converges to 0 since $\mathbb{P}(E_n^c) \leq \mathbb{P}(S_n^c)$ and $\lim_{n\to\infty} a_n^{-1}\mathbb{P}(S_n^c) = 0$ as shown above.

($\star\star$) For all $B \in \mathcal{B}_n$, let $E(B)$ denote the event

$$E(B) := \left\{ \min_{s_i \in B \cap G_n^{\mathcal{X}}} e_i \in I_1, \max_{s_i \in B \cap G_n^{\mathcal{X}}} e_i \in I_2 \right\}.$$

Then, imposing an arbitrary ordering of the hyperballs in $\mathcal{B}_n$, we have

$$\mathbb{P}\left(\forall B \in \mathcal{B}_n, \min_{s_i \in B \cap G_n^{\mathcal{X}}} e_i \in I_1, \max_{s_i \in B \cap G_n^{\mathcal{X}}} e_i \in I_2\right)$$

$$= \mathbb{P}\left(E(B_1)\right) \prod_{i=2}^{|\mathcal{B}_n|} \mathbb{P}\left(E(B_i)|E(B_{i-1}), ..., E(B_1)\right).$$

For all $i \in \{1, ..., |\mathcal{B}_n|\}$, we observe that either there exists $j \in \{1, ..., i-1\}$ such that $B_i \cap B_j \cap G_n^{\mathcal{X}} \neq 0$ in which case

$$\mathbb{P}\left(E(B_i)|E(B_{i-1}), ..., E(B_1)\right) > \mathbb{P}(E(B_i))$$

or no such j exists, in which case

$$\mathbb{P}(E(B_i)|E(B_{i-1}), ..., E(B_1)) = \mathbb{P}(E(B_i)).$$

Therefore, we have that

$$\mathbb{P}(\forall B \in \mathcal{B}_n, \min_{s_i \in B \cap G_n^{\mathcal{X}}} e_i \in I_1, \max_{s_i \in B \cap G_n^{\mathcal{X}}} e_i \in I_2)$$

$$\geq \prod_{B \in \mathcal{B}_n} \mathbb{P}(\min_{s_i \in B \cap G_n^{\mathcal{X}}} e_i \in I_1, \max_{s_i \in B \cap G_n^{\mathcal{X}}} e_i \in I_2).$$

$(\star\star\star)$ In order to alleviate notation, for all $B \in \mathcal{B}_n$, we define the following event:

$$\mathcal{E}_B(n) := \{|\{i \in [n]\}, s_i \in B\}| > C_2 b_n\}$$

It is trivial to see that for all $B \in \mathcal{B}_n$, $\mathcal{E}_B(n)$ is increasing in n (when b_n is kept fixed). Utilising the arbitrary numbering of $\mathcal{B}_n$ defined above in $(\star\star)$, we have

$$\mathbb{P}(S_n) = \mathbb{P}(\forall B \in \mathcal{B}_n : \mathcal{E}_B(n)) = \mathbb{P}\left(\mathcal{E}_{B_2}(n)\right) \prod_{i=1}^{|\mathcal{B}_n|} \mathbb{P}\left(\mathcal{E}_{B_i}(n) | \mathcal{E}_{B_{i-1}}(n), ..., \mathcal{E}_{B_1}(n)\right)$$

$$\geq \prod_{i=1}^{|\mathcal{B}_n|} \mathbb{P}\left(\mathcal{E}_{B_i}(n - (i-1)C_2 b_n) \geq \prod_{B \in \mathcal{B}_n} \mathbb{P}\left(\mathcal{E}_B(n - |\mathcal{B}_n| C_2 b_n)\right.\right.$$

where the second to last inequality holds due to the independence of the input sampling. Computing $|\mathcal{B}_n| C_2 b_n$, we obtain

$$|\mathcal{B}_n| C_2 b_n \leq C_2 \frac{C_1}{a_n^{\frac{d}{\alpha}}} \log(\frac{1}{a_n}) \frac{1}{a_n^{\eta}} = C_1 C_2 \frac{\log(\frac{1}{a_n})}{a_n^{\frac{d+\eta\alpha}{\alpha}}}$$

$$= nC_1C_2 \frac{\alpha}{d+\eta\alpha} (1 - \frac{\log(\log(n))}{\log(n)}) \leq nC_1C_2 \frac{\alpha}{d+\eta\alpha}.$$

Therefore, setting the condition $C_2 \leq \frac{d+\eta\alpha}{2C_1\alpha}$, we have

$$\mathbb{P}(S_n) \geq \prod_{B \in \mathcal{B}_n} \mathbb{P}\left(\mathcal{E}_B\left(n(1 - C_1C_2 \frac{\alpha}{d+\eta\alpha})\right)\right) \geq \prod_{B \in \mathcal{B}_n} \mathbb{P}\left(\mathcal{E}_B(\frac{n}{2})\right).$$

■

Appendix 4.C Technical Lemmas

Lemma 4.C.1 *Assume that the settings and assumptions of Theorem 4.3.5 hold. Let $(g_n)_{n\in\mathbb{N}}$ denote a sequence of non-parametric predictors and $(b_n)_{n\in\mathbb{N}}$ denote a convergence rate sequence (that converges to 0). If $\exists K > 0$ such that $\sup_{f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)} \|\hat{g}_n -$*

$f\|_\infty < K$ $\forall n \in \mathbb{N}$ *with probability 1, then*

$$\lim_{n\to\infty} \sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \mathbb{E}[(f(x_{n+1}) - \hat{g}_n(x_{n+1}))^p] \to 0 \tag{4.5}$$

if and only if $\forall \epsilon > 0$

$$\lim_{n\to\infty} \sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \mathbb{P}(|f(x_{n+1}) - \hat{g}_n(x_{n+1})| > \epsilon) = 0 \tag{4.6}$$

where $\overline{Lip}(L^*, \mathfrak{d}, M^*)$ *is as defined in Corollary 4.4.1.*

Proof " $\Longrightarrow$ " can be trivially obtained by applying Markov's inequality. We show the " $\Longleftarrow$ " statement. Fix $\epsilon > 0$ and consider an arbitrary $f \in \overline{Lip}(L^*, \mathfrak{d}, M^*)$. Define $A_n := |f(x_{n+1}) - \hat{g}_n(x_{n+1})| > \epsilon$, we have

$$(f(x_{n+1}) - \hat{g}_n(x_{n+1}))^p \leq \epsilon^p 1_{A_n^c} + K^p 1_{A_n}$$

with probability 1. This implies that

$$\sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \mathbb{E}[(f(x_{n+1}) - \hat{g}_n(x_{n+1}))^p]$$

$$\leq \epsilon^p + K^p \sup_{f\in\overline{Lip}(L^*,\mathfrak{d},M^*)} \mathbb{P}(|f(x_{n+1}) - \hat{g}_n(x_{n+1})| > \epsilon).$$

$$\overset{n\to 0}{\leq} \epsilon^p$$

As the choice of ϵ was arbitrary, (4.5) holds.

■

Lemma 4.C.2 $\forall p, c > 0$, *we have*

$$\limsup_{x\to\infty} x\left(1 - (1 - \frac{1}{x^{p+1}})^{cx^p}\right) \leq 2c$$

Proof Lemma 4.C.2 can be shown as follows.

$$x\left(1-(1-\frac{1}{x^{p+1}})^{cx^p}\right)=x\left(1-e^{cx^p\log(1-\frac{1}{x^{p+1}})}\right).$$

Expanding the exponent based on the power series expression of $\log(1+x)$, we obtain

$$cx^p\log(1-\frac{1}{x^{p+1}})=-cx^p\sum_{m=1}^{\infty}\frac{1}{mx^{m(p+1)}}=-\frac{cx^p}{x^{p+1}}\sum_{m=0}^{\infty}\frac{1}{(m+1)x^{m(p+1)}}$$

$$\geq-\frac{c}{x}\sum_{m=0}^{\infty}\frac{1}{x^{m(p+1)}}=-\frac{c}{x}\frac{x^{(p+1)}}{x^{(p+1)}-1}\geq-\frac{2c}{x}$$

for sufficiently large x. Substituting this equation back into the initial bound, we obtain:

$$\limsup_{x\to\infty}x\left(1-e^{cx^p\log(1-\frac{1}{x^{p+1}})}\right)\leq\limsup_{x\to\infty}x\left(1-e^{-\frac{2c}{x}}\right)\overset{x\to\infty}{\to}2c.$$

■

5 | Non-linear Mean Reversion

Contents

5.1 Introduction

In this chapter, we consider the second objective of the thesis, that is the development of practical theoretical tools that leverage Lipschitz regularity properties for

machine learning applications, in the context of time series analysis. Our goal is to derive theoretical results, of use to learning-based models, regarding the mean reversion properties of a stochastic process. More precisely, we are interested in examining the necessary conditions for the presence of mean reversion and in deriving an estimation of the mean reversion speed. We focus on this particular set of theoretical properties as the majority of existing results relating to mean reversion rely on linear or prescribed model assumptions (see Taylor et al. [2001] and discussion therein) and due to the fact that mean reversion is a fundamental topic in econometrics and finance; a concrete motivation for the theoretical results stated in this chapter is to enhance trading decision rules in statistical arbitrage frameworks. By using our proposed approach, we aim to leverage the flexible structural assumptions of machine learning models used in modern system identification methods[1] and Lipschitz regularity properties of these models to obtain a more precise theoretical characterisation of mean reversion than can be obtained by classical time series models.

Defining mean reversion in the context of non-linear time series models is not straightforward. As mentioned above, mean reversion has generally been studied under the assumption that the underlying dynamics are linear or with strong assumptions on the functional form of the underlying autoregressive model, see for example (Taylor et al. [2001], Mukherji [2011], Krauss [2017]). In this case, a characterisation is relatively clear and practical; directly relying on the model parameters to verify unit-root stationarity and first hitting time properties. Unfortunately, in the case of the non-linear dynamics we consider, such a characterisation is not directly available as the parameter space is significantly more complex. As an alternative, we utilise the Lipschitz-type regularity assumptions on the dynamics in order to infer properties related to the ergodicity and stationarity of the data-generating process which imply mean-reverting behaviour (as noted by Domowitz and El-Gamal [2001], Geman [2007], Fouque et al. [2011]) as well as first hitting time guarantees which serve as a proxy for mean reversion speed. As we explain in Section 5.2.3, this choice of assumption regarding the dynamics is key as they can be transformed

[1] Also referred to as a time series modelling method in the context of this chapter.

into conditions on the $\|.\|_1$ norm of the gradient of the time series dynamics and can therefore be efficiently verified for some of the most popular machine learning models. In particular, for modelling approaches that rely on neural networks, this quantity can be computed with an automatic differentiation approach by slightly modifying the commonly used back-propagation algorithm in order to obtain the input-output partial derivatives.

We would like to emphasize that the theoretical results obtained in this chapter are also of interest in their own right, not just in the context of time series modelling using machine learning. We highlight the following two theoretical contributions.

- We extend the class of general non-linear autoregressive processes that are geometrically ergodic (and hence stationary if certain initial conditions are satisfied). This result can therefore be seen as a continuation of the results obtained in several publications on the subject (see Section 5.1.1 for a more in-depth discussion).
- We derive tight probabilistic bounds on the first hitting times of general classes of partially contractive non-linear autoregressive processes. While these bounds are not given in closed form as they contain non-analytic definite integrals, they can be computed numerically offline, and solutions stored in a look-up table. To the best of our knowledge, no other first hitting time guarantees have been derived that cover the general class of non-linear stochastic processes that we consider.

As a practical illustration of our proposed approach, we consider the trading of mean-reverting assets which arises in pairs trading and statistical arbitrage strategies in finance. This application is particularly interesting in the context of our chapter as mean-reverting financial time series exhibit several stylised facts, such as asymmetric mean-reversion or varying mean-reversion speed across the input space that cannot be captured by most classical time series models. We apply our theoretical results in this context in two ways. First, we show how the first hitting time bounds derived in this chapter can be directly translated to probabilistic bounds on the return of a mean reversion trading strategy, provided the underlying autoregressive model of

the mean reverting time series satisfies the required contractive Lipschitz conditions. Second, we propose a set of trading decision rules based on the application of the mean reversion properties developed in this chapter and a neural network based forecasting model. This approach is benchmarked against standard trading decision rules that can be found in the Pairs Trading literature in an empirical experiment using real and artificial data. We find that our proposed trading decision rules identifies better trading opportunities in terms of holding time, average return per holding time, return volatility and Sharpe ratio of held positions. This approach is, to the best of our knowledge, the first to incorporate modern machine learning mechanisms into the trading decision making process of mean reversion.

5.1.1 Related Literature

The investigation of theoretical properties relating to the ergodicity and stationarity of non-linear autoregressive processes is a classical topic in econometrics which has been carried out for a variety of different choices of assumptions on the autoregressive dynamics. For an exposition of relevant theoretical definitions and results related to the ergodicity of non-linear processes, we refer the reader to Nummelin [2004] and for a more general overview of theoretical properties of non-linear autoregressive processes we refer to Meyn and Tweedie [2012].

A foundational paper by Tweedie [1976] first established an equivalence between the study of stability in deterministic dynamical systems using Lyapunov functions and the study of geometric ergodicity in non-linear processes. This result, and the Tweedie criterion given in the same paper, have provided a basis for an entire stream of research; e.g. Tjøstheim [1990]; Bhattacharya and Lee [1995]; Lu [1998] that look to relax the assumptions on the autoregressive function of the non-linear process. In particular, we note the theoretical results of An and Huang [1996] and Cline and Pu [1999] which prove the geometric ergodicity of the underlying data generating process when the autoregressive function satisfies a relaxed Lipschitz-type assumption and are the most comparable to the results on ergodicity stated in this chapter. We also mention Liebscher [2005] who improves on the results of An and Huang [1996]

and Cline and Pu [1999], explores the relationship between geometric ergodicity and mixing properties and investigates the geometric ergodicity of the EXPARCH and threshold autoregressive models. More recently, several papers; Boussama et al. [2011]; Chen et al. [2018]; Kheifets and Saikkonen [2020], have explored the ergodic properties of various non-linear autoregressive models that have a specific structure such as vector star models or generalised exponential autoregressive models. Finally, while little work has been done with a direct application to machine learning based time series modelling in mind, we highlight the work of Trapletti et al. [2000] which provides conditions for the ergodicity and stationarity of autoregressive neural network processes. The goal of this work overlaps with ours as it also focuses on non-linear processes that arise in the context of machine learning based time series modelling and the paper proposes several interesting results, however it suffers from a major drawback in that it requires the boundedness of all activation functions of the neural network which excludes commonly used choices such as ReLU.

The results on geometric ergodicity derived in this chapter extend the results of An and Huang [1996] and Cline and Pu [1999] by setting similar but different Lipschitz-type conditions that are better suited for application in the context of machine learning based time series modelling. Our conditions, applied to neural network frameworks, also avoid imposing the boundedness assumption on the activation function made in Trapletti et al. [2000]. We recall that our goal is not to derive the most general class of non-linear ergodic processes but rather, as discussed in the introduction, to focus on a dynamical model class that can be defined by easily verifiable yet sufficiently general conditions, so that they can be directly accessible for application in conjunction with non-linear time series modelling methods, especially those pertaining to modern machine learning.

In contrast to the literature on the geometric ergodicity of non-linear processes, the existing literature on the first hitting times of contractive processes is spread across a diverse range of contexts. For discrete time series, usual approaches assume that the underlying dynamics are linear and stationary, i.e. follow a stationary AR(p) model.

The first hitting time probabilities and expectation are then computed numerically (Basak and Ho [2004], Di Nardo et al. [2008]) or can be lower bounded analytically in the case of the AR(1) model (Novikov and Kordzakhia [2008]). This approach has been explored in various domains; in statistical arbitrage and quantitative finance for optimal thresholds setting (Krauss [2017], Puspaningrum et al. [2010]), for predicting population extinction and time to extinction in ecology (Ferguson and Ponciano [2014]), signal detection and surveillance analysis(Frisén and Sonesson [2006]) or structural health monitoring (Mollineaux and Rajagopal [2015], Noh et al. [2009]). For continuous time price series, dynamics are usually assumed to follow Ornstein-Uhlenbeck (OU) dynamics in which case the first hitting time probabilities can be obtained semi-analytically (Lipton and Kaushansky [2018], Martin et al. [2019] and references therein) under some additional assumptions. Applications are numerous and involve, for example, hydrology (Fisher et al. [2014]), neuroscience (Lánskỳ and Smith [1989]) or quantitative finance (Bertram [2010]), (Zeng and Lee [2014]). Note that, even though in the aforementioned works specific forms of dynamic models were presupposed, the computation of the first hitting time probabilities had to rely on numerical approximation. Simplifying this computation is difficult even for simple autoregressive functions and remains an open question for both Ornstein-Uhlenbeck models and AR(p) models.

In this chapter, we aim to extend the existing literature as it only provides an understanding of first hitting times for time series whose dynamics conform to a specific (linear) structure. In particular, when those functions are identified by black-box machine learning algorithms, existing results are not applicable and additional theoretical guarantees are needed. Providing sufficient criteria for mean-reversion type behaviour of non-linear autoregressive processes and the provision of first-hitting time bounds for such processes is one of the core contributions of our chapter.

5.1.2 Outline

The first part of the chapter presents the main theoretical results and is divided into three subsections. Section 5.2.1 states the theoretical results on general mean

reversion for non-linear processes in the form of geometric ergodicity and stationarity properties. Section 5.2.2 provides theoretical results pertaining to mean reversion speed defined by first hitting time guarantees of the stochastic process's return to the mean. Finally, Section 5.2.3 discusses how these results can be applied in the context of machine learning based time series modelling in order to more accurately characterise non-linear mean reversion.

The second part of the chapter provides an application of the theoretical results in the context of financial trading. In order to do so, we first provide a short discussion that shows how the first hitting time guarantees can be directly translated into guarantees on trading profits when precise estimates of the partial derivatives of the underlying dynamics are known. We then utilise the derived non-linear mean reversion results to define trading decision rules in a statistical arbitrage framework. We illustrate the performance of the proposed trading decision rules empirically by conducting a benchmark experiment on a small set of (pairwise)-mean reverting stocks and a series of artificially generated data. A short conclusion and discussion of future work is given at the end and all major proofs can be found in the Appendix.

5.2 Theoretical Results

In this chapter, we will consider a sequence of random variables $(y_t)_{t\in\mathbb{N}}$ generated by the nonlinear autoregressive structure[2] described in equation (2.2) in Chapter 2. More precisely, we assume the following: let $d \in \mathbb{N}$

$$y_{t+1} := \begin{cases} f(y_t, ..., y_{t-(d-2)}, y_{t-(d-1)}) + \epsilon_{t+1} \text{ for } t \geq 0 \\ a_{-t}, \quad \text{for } t \in \{-1, ..., -d\}. \end{cases} \tag{5.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ denotes the time series dynamics (or autoregressive function), $a \in \mathbb{R}^d$ represents the initial conditions and $(\epsilon_t)_{t\in\mathbb{N}}$ is a mean zero stochastic process.

In order to derive mean reversion properties for $(y_t)_{t\in\mathbb{N}}$, assumptions need to be made

[2]In this chapter, we do not consider control inputs.

on f and $(\epsilon_t)_{t\in\mathbb{N}}$. As discussed in the Chapter 2, we aim to make the assumption on f as general as possible but defined in such a way that it can be readily verified if f has been approximated by $\hat{f}$ obtained from a learning-based time series model (the impact of the approximation error between f and $\hat{f}$ is briefly discussed in Section 5.2.3). This is done by considering Lipschitz-type assumptions that, under additional regularity assumptions, can be transformed into assumptions on the gradient of f. As this chapter will utilise a notion of local-Lipschitz continuity, we rehearse the definition of the space of Lipschitz continuous functions provided in Chapter 2 in order to explicitly denote the subset of $\mathbb{R}^d$ on which the time series dynamics f is Lipschitz continuous.

Notation 5.2.1 *Let $\mathcal{D} \subseteq \mathbb{R}^d$, $L \in \mathbb{R}_+$ be a constant and consider a norm $\|\cdot\|$ on $\mathbb{R}^d$. We define the space of* L-Lipschitz continuous functions *with respect to $\|\cdot\|$ on $\mathcal{D}$ as*

$$Lip(L, \mathcal{D}, \|\cdot\|) := \{f : \mathcal{D} \to \mathbb{R} |\, |f(x) - f(x')| \leq L\|x - x'\|, \forall x, x' \in \mathcal{D}\}\,.$$

5.2.1 Geometric Ergodicity of Non-linear Processes

Geometric ergodicity can be used to imply mean reversion as it studies the long-term convergence of a process towards a stable probability measure. Furthermore, through the well-known Birkhoff-Khinchin theorem (Birkhoff [1931]), it provides long-term convergence guarantees for the statistical quantities of a time series by linking the time sample estimation to the space average computed using the invariant probability measure associated with the autoregressive model. In Appendix 5.A, a more precise discussion on the link between mean reversion and geometric ergodicity is provided.

In order to show that $(y_t)_{t\in\mathbb{N}}$ is geometrically ergodic, an assumption on the noise process $(\epsilon_t)_{t\in\mathbb{N}}$ must be made. This is given in the following statement:

Assumption 15 *$(\epsilon_t)_{t\in\mathbb{N}}$ is a sequence of independent and identically distributed random variables with positive density everywhere on $\mathbb{R}$ such that $\mathbb{E}[|\epsilon_1|] < \infty$.*

Assumption 15 is standard in the literature on geometric ergodicity of non-linear models and is necessary to apply Tweedie's criterion (Tweedie [1976]) in order to show geometric ergodicity. For example, a sequence of i.i.d. variables following a Gaussian noise distribution or a Student's t-distribution would satisfy this assumption. Using Assumption 15 as well as a global contraction condition with respect to $\|.\|_\infty$, we can obtain the following known result (Nummelin [2004]) which is fairly trivial to prove.

Lemma 5.2.2 *Let $\bar{\alpha} \in (0,1)$, $f \in Lip(\bar{\alpha}, \mathbb{R}^d, \|\cdot\|_\infty)$ and $a \in \mathbb{R}^d$. If the noise terms $(\epsilon_t)_{t\in\mathbb{N}}$ verify Assumption 15, then $(y_t)_{t\in\mathbb{N}}$ is geometrically ergodic.*

A global assumption on f, as given in Lemma 5.2.2 by $Lip(\bar{\alpha}, \mathbb{R}^d, \|\cdot\|_\infty)$, makes sense as a condition for processes that follow classical linear autoregressive models as in this case the same dependence structure holds for all inputs of the state space. However for machine learning frameworks that adaptively model the dependence structure over the state space, these global conditions are overly restrictive. In this context, we therefore aim to establish more local conditions for ensuring the geometric ergodicity of the process. This is formally addressed in the following theorem by restricting the contraction assumption on f to a subset of $\mathbb{R}^d$;

Theorem 5.2.3 *Let $\bar{\alpha} \in (0,1)$, $K \subset \mathbb{R}^d$ be compact and assume the noise terms $(\epsilon_t)_{t\in\mathbb{N}}$ verify Assumption 15. If $f \in Lip(\bar{\alpha}, \mathbb{R}^d \setminus K, \|\cdot\|_\infty)$ and is bounded on compact sets, then $(y_t)_{t\in\mathbb{N}}$ is geometrically ergodic.*

This theorem extends the class of geometrically ergodic non-linear processes described in An and Huang [1996]; Cline and Pu [1999]. For $(y_t)_{t\in\mathbb{N}}$ defined by (5.1), ergodic behaviour can therefore be observed if $\bar{\alpha} \in (0,1)$ and the dynamics of the time series are $\bar{\alpha}$-Lipschitz continuous with respect to $\|\cdot\|_\infty$ above (resp. below) a given upper (resp. lower) barrier. For example, it is possible for $(y_t)_{t\in\mathbb{N}}$ to be a non-continuous or a linear non-unit root stationary process when it stays between the two barriers and still exhibit mean reverting properties if the dynamics are contracting when the time series breaches the barriers.

We note that the result of Theorem 5.2.3 can be directly used to imply the following corollary on the stationarity of $(y_t)_{t\in\mathbb{N}}$.

Corollary 5.2.4 *Let $(y_t)_{t\in\mathbb{N}}$ be as in Theorem 5.2.3. Then $(y_t)_{t\in\mathbb{N}}$ is asymptotically stationary.*

Using two short lemmas (see the Appendix for the statement of the lemmas), we can rewrite the Lipschitz condition of Theorem 5.2.3 in terms of the gradient of f under additional regularity conditions.

Notation 5.2.5 *For $c \in \mathbb{R}^d, R \in \mathbb{R}_+$, we define the ball centered in c and with radius R as $B(0,R) = \{x \in \mathbb{R}^d : ||x||_\infty < R\}$.*

Proposition 5.2.6 *(Ergodicity for C^1 functions) Let $K = \overline{B_{\|\cdot\|_\infty}(c,R)}$ for $c \in \mathbb{R}^d$, $R \in \mathbb{R}_+$. Assume f is bounded on K, $f \in C^1(\mathbb{R}^d \setminus K)$ and $(\epsilon_t)_{t\in\mathbb{N}}$ satisfies Assumption 15. If there exists $\bar{\alpha} \in (0,1)$ such that $\forall x \in \mathbb{R}^d \setminus K$, $\|\nabla f(x)\|_1 \leq \bar{\alpha}$ then $(y_t)_{t\in\mathbb{N}}$ defined by equation (5.1) is geometrically ergodic.*

As will be discussed in Section 5.2.3, this proposition can be applied in the context of neural network frameworks as the input-output gradients of the neural network are easily computable.

5.2.2 First Hitting Time Guarantees for Contracting Non-linear Processes

From the previous subsection, we have that if a non-linear process $(y_t)_{t\in\mathbb{N}}$ satisfies the contraction assumption stated in Theorem 5.2.3, then the process is mean reverting. However, this characterisation does not include any information on the speed or behaviour of the mean reversion. In this section, we aim to compute these properties for $(y_t)_{t\in\mathbb{N}}$ in the form of first hitting time guarantees on the return to the mean. As this task is more difficult, additional assumptions are needed. We assume that the noise process $(\epsilon_t)_{t\in\mathbb{N}}$ satisfies additionally.

Assumption 16 $(\epsilon_t)_{t\in\mathbb{N}}$ *are independent and identically distributed random variables with bounded non-zero variance and a probability density function denoted by* f_ϵ.

We also replace the contraction condition on f by a slightly modified and weaker Lipschitz-type assumption: we assume f to be a contraction relative to a weighted norm denoted $\|\cdot\|_{\alpha^*}$ which is defined as: for $\alpha^* \in \mathbb{R}^d_{\geq 0}$, the α^*-norm $\|\cdot\|_{\alpha^*} : \mathbb{R}^d \to \mathbb{R}$ is the weighted l_1-norm

$$\forall x \in \mathbb{R}^d, \quad \|x\|_{\alpha^*} = \sum_{i=1}^{d} \alpha_i^* |x_i|.$$

Using this norm, we can define the concept of an α^*-*contracting process* on $\mathcal{D} \subseteq \mathbb{R}^d$ as an autoregressive process with a transition function f that is contained in $Lip(1, \mathcal{D}, \|\cdot\|_{\alpha^*})$ and $\alpha^* \in \triangle_+ := \{x \in \mathbb{R}^d_{\geq 0} | \sum_{i=1}^d x_i < 1\}$. Our Lipschitz-type assumption is then given by:

Assumption 17 *Our time series* $(y_t)_{t\in\mathbb{N}}$ *is an* α^* *contracting process , i.e.*

$$f \in \mathcal{L}^{\alpha^*}(\mathcal{D}) := Lip(1, \mathcal{D}, \|\cdot\|_{\alpha^*})$$

for some $\alpha^* \in \triangle_+$ *and* $\mathcal{D} = \mathbb{R}^d$.

One may wonder how this α^* contracting condition relates to a simpler Lipschitz assumption on f. Let $\mathcal{D} \subseteq \mathbb{R}^d$, $\alpha^* \in \triangle_+$ and $\bar{\alpha} := \sum_{i=1}^d \alpha_i$. We have the following relationship between Lipschitz function spaces: (1) $\mathcal{L}^{\alpha^*}(\mathcal{D}) \subseteq Lip(\bar{\alpha}, \mathcal{D}, \|\cdot\|_\infty)$ and (2) for $\delta \in (0,1)$, define $\alpha^* = (\frac{\delta}{d}, ..., \frac{\delta}{d})^\top$, then $Lip(\frac{\delta}{d}, \mathcal{D}, \|\cdot\|_1) \subseteq \mathcal{L}^{\alpha^*}(\mathcal{D})$. (Note that the first condition implies that the assumption used in Theorem 5.2.3 is weaker than Assumption 17).

Therefore, although the α^* condition is notationally heavy, it is useful as it provides a weaker assumption than an alternative Lipschitz condition based on the $\|\cdot\|_1$ norm. Perhaps more importantly, the α^* condition provides additional flexibility that allows for the dependence on previous time lags to be greater than $\frac{1}{d}$ as long as the sum of the α^* coefficients is smaller than unity. This feature is useful in

practice where models generally depend on recent time lags more. Finally, if the α^* coefficients are obtained by using a machine learning estimation of f then the input dimension of the estimation model can be greater than d with no explicit consequences on the α^* condition.

We define the following matrices that will be used express the first hitting time guarantees.

Notation 5.2.7 *(Relevant matrices) For any $\alpha^* \in \triangle_+ := \{x \in \mathbb{R}^d_{\geq 0} | \sum_{i=1}^d x_i < 1\}$ and $T \in \mathbb{N}$, we define the associated matrices $A(T) \in \mathbb{R}^{T \times T}$ and $B \in \mathbb{R}^{d \times d}$. Here, $A(T)$ is a lower triangular banded matrix and B is a sparse matrix whose entries are given by:*

$$A(T)_{ij} := \begin{cases} 1, & \text{if } i-j=0 \\ -\alpha^*_{(i-j)}, & \text{if } 0 < i-j \leq d \\ 0, & \text{otherwise.)} \end{cases} \tag{5.2}$$

$$B_{ij} := \begin{cases} 1, & \text{if } i-j=1 \\ \alpha^*_j, & \text{if } i=1 \text{ and } 1 \leq j \leq d \\ 0, & \text{otherwise)} \end{cases} \tag{5.3}$$

for all $i, j \in \{1, ..., T\}$.

We now state the bounds on first hitting times of a time series generated by an autoregressive model satisfying Assumption 17. Appealing to Banach's fixed point theorem one can show the existence of a unique fixed point for f: $y^* = f(y^*, \ldots, y^*)$. As noted in the previous section, the contractive properties of the time series result in a generalisation of mean-reverting behavior where the fixed point serves as the level to which the time series will tend to revert to in the long run after being exposed to a shock. More formally, we define the following.

For $a \in \mathbb{R}^d$ with $a_d > y^*$ and $\gamma \in [0, a_d - y^*)$, we define the upper first hitting time of $(y_t)_{t \in \mathbb{N}}$:

$$\tau^+_\gamma := \inf\{t \in \mathbb{N} | y_t - y^* < \gamma\} \ .$$

Similarly, for $a_d < y^*$ and $\gamma \in [a_d - y^*, 0)$, we define the lower first hitting time of $(y_t)_{t\in\mathbb{N}}$:

$$\tau_\gamma^- := \inf\{t \in \mathbb{N} | y_t - y^* > \gamma\} \ .$$

The initial value a_d can be seen as having resulted from a "shock" in the time series and γ as a return barrier that indicates proximity to the long-run "mean" y^*. The first hitting times τ_γ^+ and τ_γ^- are linked to the speed of mean reversion measured at various levels (γ). By conditioning on past hitting times and the last result of Wise [1955], one can show our following principal result:

Notation 5.2.8 *To alleviate notation, we denote the projection operator onto the $i-th$ component: $\pi_i : \mathbb{R}^d \to \mathbb{R}, (x_1, \ldots, x_d)^\top \mapsto x_i$ for $i \in \{1, ..., d\}$.*

Theorem 5.2.9 *For $T \in \mathbb{N}$, define*

$$\mathfrak{I}^+_{(\alpha^*,y^*)}(T) := \int_{-b_1}^{\infty} ... \int_{-b_T}^{\infty} f_{\epsilon_{1:T}}(A(T)x)dx$$

where $A(T)$ is defined in (5.2), $f_{\epsilon_{1:T}}$ is the joint probability density function of any finite sequence of consecutive noise variables: $\epsilon_{1:T} := (\epsilon_1, ..., \epsilon_T)$ defined according to Assumption 16 and $b_i := \pi_1\big(B^i(a - y^*\mathbf{1}_d)\big) - \gamma$ for $i = 1, \ldots, T$ where B is defined in (5.3). We have:

$$(i)\ \mathbb{P}\big(\tau_\gamma^+ > T\big) \leq \mathfrak{I}^+_{(\alpha^*,y^*)}(T) < 1 \text{ and}$$

$$(ii)\ \mathbb{E}[\tau_\gamma^+] \leq 1 + \sum_{T=1}^{\infty} \mathfrak{I}^+_{(\alpha^*,y^*)}(T).$$

Analogous bounds can be derived for $\mathbb{P}\big(\tau_\gamma^- > T\big)$ and $\mathbb{E}[\tau_\gamma^-]$.

Proof See appendix.

■

Theorem 5.2.9 provides a lower bound on the cumulative density function of the first hitting times of $(y_t)_{t\in\mathbb{N}}$ as it returns to the fixed point of its autoregressive model. By varying the choice of barrier γ, the bound given in Theorem 5.2.9 can be used as

a theoretical guarantee on the speed of the mean reversion $(y_t)_{t\in\mathbb{N}}$. It is important to note that contrary to the results on ergodicity given in the previous subsection, these bounds do not provide any information on the long-run convergence of the data generating process.

The multi-dimensional integral expression stated in $\mathfrak{I}^+$ corresponds to the computation of the orthant probabilities of a T-dimensional random vector. This computation can be done using quadrature, sparse grids or Monte-Carlo methods and dedicated software libraries exist (Hahn [2005]). Furthermore, this computation can be done offline and a look-up table can be created. In the case where $(\epsilon_t)_{t\in\mathbb{N}}$ is i.i.d. Gaussian, the T-dimensional random vector is distributed as $\mathcal{N}(b, V^{-1})$. Extensive research has been done to optimise the numerical evaluation of this type of expression and fast quasi Monte Carlo methods can be used for accurate computation for $T < 100$ (Genz and Bretz [2009]).

Some comments on the behaviour $\mathfrak{I}^+_{(\alpha^*,y^*)}(T)$:

Proposition 5.2.10 *Consider the same setup as in Theorem 5.2.9. Then* $\mathfrak{I}^+_{(\alpha^*,y^*)}$ *satisfies the following properties*

1. $\forall T \in \mathbb{N}$, $\mathfrak{I}^+_{(\alpha^*,y^*)}(T)$ *is decreasing in* γ.
2. $\forall T \in \mathbb{N}_{>d}$, *if* $\forall\, i$, $\alpha_i^* \leq \beta_i^*$ *and* $\exists\, j$ *s.t.* $\alpha_j^* < \beta_j^*$ *then* $\mathfrak{I}^+_{(\alpha^*,y^*)}(T) < \mathfrak{I}^+_{(\beta^*,y^*)}(T)$.
3. $\forall T \in \mathbb{N}$: $\lim_{\|\alpha^*\|_1 \to 0} \mathfrak{I}^+_{(\alpha^*,y^*)}(T) = \frac{1}{2^T}$.

Proof See appendix.

■

As with the main result of the previous section on the ergodicity of $(y_t)_{t\in\mathbb{N}}$, a useful extension of Theorem 5.2.9 is to weaken Assumption 17 to take advantage of the flexibility of the non-linear autoregressive model. An improvement of this type is stated in the following Corollary.

Corollary 5.2.11 *Assume that Assumption 16 holds. Consider* $a \in \mathbb{R}^d$ *and* $\mathcal{D}^* = \prod_{i=1}^d \mathbb{R}_{\geq b_i}$ *for* $b \in \mathbb{R}^d$. *If* $\exists \alpha^*_{D^*} \in \mathbb{R}^d$ *satisfying* $f|_{\mathcal{D}^*} \in \mathcal{L}^{\alpha^*_{D^*}}(D^*)$ *and* f *does not*

have a fixed point in D^*, then, $\forall \gamma \in [0, a_d - \max_{i\in 1,\dots,d}\{b_i\}]$ and $\forall T \in \mathbb{N}$, the upper bounds stated in Theorem 5.2.9 hold with the α^* coefficients replaced by $\alpha^*_{D^*}$.

Proof The proof of Corollary 5.2.11 follows from the proof of Theorem 5.2.9 and Proposition 5.2.10.2.

■

Corollary 5.2.11 can also be used to tighten the upper bounds of Theorem 5.2.9. Indeed, fix $\gamma \in [0, a_d - y^*)$. Then, defining $\mathcal{D}^* = \prod_{i=1}^d \mathbb{R}_{\geq b_i}$ where $b_i = \min\{\gamma, \min_{j\leq i} a_j\}$, we have that Corollary 5.2.11 can be applied with $\alpha^*_{D^*}$ coefficients instead of Theorem 5.2.9 with the global α^*.
The following result states that under the classical Gaussian noise assumption, $\mathbb{E}[\tau^+_\gamma]$ is finite.

Lemma 5.2.12 *If, instead of Assumption 16, $(\epsilon_t)_{t\in\mathbb{N}}$ is assumed to be a Gaussian white noise process then $\mathbb{E}[\tau^+_\gamma] < \infty$.*

Proof Follows from proof of Theorem 5.2.9 and Basak and Ho [2004].

■

One may wonder how effective the bounds given in Theorem 5.2.9 are and whether they can be improved for the assumptions used in this chapter. The following result shows that the bounds stated in Theorem 5.2.9 cannot be improved for $\alpha^* \in \triangle_+$.

Lemma 5.2.13 *(Tightness) The upper bounds in Theorem 5.2.9 are tight for all $\alpha^* \in \triangle_+$.*

Proof Proposition 5.2.13 follows from the proof of Theorem 5.2.9.

■

The proof of Theorem 5.2.9 shows that the bounds are tight when the dynamics of $(y_t)_{t\in\mathbb{N}}$ can be represented by a linear autoregressive model (AR(p), $p \in \mathbb{N}$). In particular, for $\alpha^* \in \triangle_+$, this implies that any non-linear model that is Lipschitz continuous with respect to $\|\cdot\|_{\alpha^*}$ and has a fixed point y^*, will have its first hitting

time probabilities and expectation upper bounded by a linear AR process with coefficients α^* and intercept c (specified such that the mean of the AR process is y^*).

5.2.3 Link to Machine Learning-Based Models

In this subsection we explain how the theoretical results obtained the in this chapter can be utilised in the context of machine learning frameworks that are sufficiently smooth. We illustrate the practical performance of the first hitting time guarantees stated in Section 5.2.2 based on synthetic data and a neural time series model implementation. A commonly used AR(1)-based approach to estimating the first hitting times is used to benchmark our approach (see Novikov and Kordzakhia [2008] for the theoretical derivation and Sukparungsee and Novikov [2006] for a numerical application of the AR(1)-based approach).

Assume that the following time series data: $\mathcal{S}_N = \{y_t\}_{t\in\{1,\dots,N\}}$ can be observed for some $N \in \mathbb{N}$ and that an autoregressive machine learning forecasting model $\hat{f}$ has been fitted to the data. Then, using $\hat{f}$ as a replacement for the transition function f one can aim to estimate $\bar{\alpha}$ (as defined in Lemma 5.2.2 or Theorem 5.2.3) or the α^* coefficients. To do this, a main advantage of the theoretical results obtained so far in this chapter is the intuitive formulation of the Lipschitz type conditions that were used. In particular, if $\hat{f}$ is differentiable, we can utilise the partial derivatives of $\hat{f}$ to verify the conditions needed to apply the results. This has already been explicitly shown in Proposition 5.2.6 for the theoretical results on ergodicity where the assumptions are directly stated in terms of $\max_x \|\nabla f(x)\|_1$ and can be shown for the first hitting time guarantees given in Theorem 5.2.9 and Corollary 5.2.6 by applying the following lemma;

Lemma 5.2.14 *If the domain $\mathcal{D} \subseteq \mathbb{R}^d$ of f is convex and $f \in C^1(\mathcal{D})$ then $f \in \mathcal{L}^{\alpha^*}(\mathcal{D})$ with $\alpha_i^* = \max_{x\in\mathcal{D}} |\frac{\partial f}{\partial x_i}(x)|$.*

Proof (Trivial) Follows directly from an application of the multivariate version of the mean value theorem and the convexity of $\mathcal{D}$.

Algorithm 1 Characterising Mean Reversion with Neural Networks

Input: A time series data set $(y_t)_{t\in\{1,\dots,n\}}$

1: fit a Neural Network model $\hat{f}$ to $(y_t)_{t\in\{1,\dots,n\}}$
2: use automatic differentiation to compute $\nabla\hat{\psi}$
3: define a set $\mathcal{V}$ of the form given in Corollary 5 based on $\max_{x\in\mathcal{V}} \left\|\nabla\hat{\psi}(x)\right\|_1$ ▷ (Final choice of $\mathcal{V}$ depends on the application)
4: **if** $\mathcal{V}$ exists: **then**
5: assert that $(y_t)_{t\in\mathbb{N}}$ is mean reverting
6: compute the $\alpha^*_{\mathcal{V}}$ coefficients of $\hat{\psi}$
7: **if** $\alpha^*_{\mathcal{V}} \in \triangle_+$ **then**
8: estimate standard deviation of noise using residuals
9: compute $\mathfrak{I}^-_{(\alpha^*_{\mathcal{V}}, y^*_{\mathcal{V}})}(T)$*for various choices of thresholds and time steps T
10: use $\mathfrak{I}^-_{(\alpha^*_{\mathcal{V}}, y^*_{\mathcal{V}})}(T)$ to infer guarantees on first hitting times
11: **end if**
12: **end if**

Figure 5.21: (Algorithm) Characterising mean reversion with neural networks and input-output gradient computations.

■

From Lemma 5.2.14, we have that if there exists $\{\alpha^*_i\}_{i\in\{1,\dots,d\}}$ coefficients such that $\max_{x\in\mathcal{D}} |\frac{\partial \hat{f}}{\partial x_i}(x)| \leq \alpha^*_i$ for all $i \in \{1, ..., d\}$ and $\sum_{i=1}^d \alpha^*_i < 1$ then Theorem 5.2.9 and Proposition 5.2.12 can be applied. While the computation of Lipschitz constants of machine learning models can be computationally difficult (with the exception of some non-parametric frameworks, see Calliess et al. [2020] and Chapter 3), obtaining an approximation of the gradients is generally more straightforward. In particular, for non-linear autoregressive models that rely on neural networks, backpropagation can be used to compute partial derivatives and existing deep learning libraries (e.g. Pytorch or Tensorflow) can be utilised (see torch.autograd or tf.GradientTape). In Algorithm 1, we proceduralise this discussion and describe more precisely how neural networks can be applied in order to approximately verify the conditions of non-linear mean reversion derived in practice.

We note that the type of input-output partial derivative computation discussed in the previous paragraph and utilised in Algorithm 1 has been extensively used in computer vision and explainable AI for input sensitivity analysis, see Baehrens et al. [2010]; Simonyan et al. [2013] and has started to expand to other application areas. Existing heuristics and techniques from these fields can therefore be leveraged in order to improve Steps 2 and 3 of Algorithm 1. Furthermore, for several non-

parametric machine learning model choices that utilise neural networks, it is possible to incorporate gradient learning directly into the model fitting process which would offer a more direct way of estimating $\max_x \|\nabla f(x)\|_1$ and the $\{\alpha_i^*\}_{i\in\{1,\dots,d\}}$ coefficients, see Ismail et al. [2021].

We also note that Step 2 of Algorithm 1 only provides a discretised approximation of $\|\nabla \hat{f}\|_1$ across the input space. If $\hat{f}$ is a neural network with ReLU activation functions, we can consider recent work by (Jordan and Dimakis [2020]) that proposes a Mixed-Integer Programming (MIP) approach for exactly computing local Lipschitz constants of ReLU networks in order to modify Algorithm 1 to avoid this discretisation. More precisely, if $\hat{f}$ is a ReLU network, Steps 2 and 3 of Algorithm 1 can be replaced by an alternative series of steps that involves solving a MIP problem on a set $\mathcal{V}$ which is given as input and then applying Theorem 5.2.3 and Corollary 5.2.11. This approach is however technically difficult as it suffers from the worst-case exponential time complexity of solving a MIP problem and the difficulty of defining MIP problems with non-convex input sets $\mathcal{V}$ of the type used in Theorem 5.2.3. Therefore, as an assumption on the accuracy of the discretisation approximation of Step 2 is weak relative to the one made on the accuracy of the approximation of ∇f by $\nabla \hat{f}$ and the fact that the computation of $\max_{x\in\mathcal{D}} |\frac{\partial f}{\partial x_i}(x)|$ can be done efficiently in low dimensions through grid search (Scaman and Virmaux [2018]), we consider the more approximate approach described by Algorithm 1. We note that this approach is consistent with the baseline method for Lipschitz constant estimation of neural networks proposed in the experimental section of Scaman and Virmaux [2018].

In Table 5.21, we illustrate how the upper bounds given in Theorem 5.2.9 perform in practice. Using a standard autoregressive model, we generate a time series of fixed length: 1000 timesteps and model the noise with a Gaussian distribution. Then, utilising a 2-layer Neural Network with sigmoid activation we estimate the α^* constants of the underlying function and compute $\mathfrak{I}^+_{(\alpha^*,y^*)}(T)$ for $1 \leq T \leq 40$ (We stop at 40 as the first hitting time probabilities are approximately equal to 0 for $T \geq 40$). These values are compared against the "true" first hitting probabilities stochastic process defined by the selected autoregression function. We compute

Method	$L_1(20)$	$L_1(40)$	$\mathbb{E}[\tau\|\tau \leq 40]$
AR(1): $x_{t+1} = 0.9x_t + \epsilon_{t+1}$, $\mathbb{E}[\tau\|\tau \leq 40] = 8.96$			
Non-linear Estim.	0.12 ± 0.067	0.06 ± 0.033	9.05 ± 1.23
AR(1) Estim.	0.087 ± 0.046	0.052 ± 0.027	7.95 ± 0.048
AR(3): $x_{t+1} = 0.7x_t + 0.15x_{t-1} + 0.05x_{t-2} + \epsilon_{t+1}$, $\mathbb{E}[\tau\|\tau \leq 40] = 7.66$			
Non-linear Estim.	0.102 ± 0.032	0.1 ± 0.08	11.86 ± 1.13
AR(1) Estim.	0.074 ± 0.042	0.038 ± 0.02	7.43 ± 0.58
ESTAR(1): $x_{t+1} = 0.4x_t + 0.3x_t(1 - e^{-\frac{x_t^2}{2}}) + \epsilon_{t+1}$, $\mathbb{E}[\tau\|\tau \leq 40] = 3.4$			
Non-linear Estim.	0.093 ± 0.021	0.04 ± 0.009	4.77 ± 0.61
AR(1) Estim.	0.085 ± 0.0095	0.038 ± 0.004	3.57 ± 0.04
Neur. Net.: $x_{t+1} = NN(x_t, x_{t-1}, x_{t-2}) + \epsilon_{t+1}$, $\mathbb{E}[\tau\|\tau \leq 40] = 7.36$			
Non-linear Estim.	0.17 ± 0.061	0.06 ± 0.03	10 ± 1.33
AR(1) Estim.	0.085 ± 0.037	0.08 ± 0.03	3.8 ± 1.2

Table 5.21: Performance of the first hitting time bounds in practice.

the (averaged) $L_1(20)$-error: $\frac{1}{20}\sum_{T=1}^{20} |\mathfrak{I}^+_{(\alpha^*,y^*)}(T) - \mathbb{P}\big(\tau_\gamma^+ > T\big)|$, (averaged) $L_1(40)$-error: $\frac{1}{40}\sum_{T=1}^{40} |\mathfrak{I}^+_{(\alpha^*,y^*)}(T) - \mathbb{P}\big(\tau_\gamma^+ > T\big)|$ and the value of the estimation of the conditional expectation $\mathbb{E}[\tau|\tau \leq 40]$. These values are averaged over 10 simulations and the standard deviation of the obtained results is also stated. As a benchmark, we estimate the first hitting time probabilities of the time series using an AR(1) model as is most commonly done in practice (see discussion in introduction).

As the AR(1) estimation approach aims to directly estimate the first hitting time of the stochastic process (instead of ensuring a lower bound), one could expect it to be more precise than the non-linear first hitting time estimation approach in terms of $L_1(20)$ and $L_1(40)$ metrics. While this can be observed for some values in Table 5.21, we have that in the majority of computed loss metrics the performance of our proposed approach is competitive with the results of the AR(1) first hitting time estimation method. The estimated values of $\mathbb{E}[\tau|\tau \leq 40]$ then illustrate the fact that the non-linear estimation method aims to ensure a lower bound on the first hitting times of the time series. We note that for each estimation, $\alpha^* \in \triangle_+$ on a subset of $\mathbb{R}^d$ of the form given in Algorithm 1. This implies that $\max_x \|\nabla f(x)\|_1 < 1$ and that the time series is geometrically ergodic.

One caveat to the discussion of this section is that the robustness of the estimation of the α^* coefficients can be difficult to obtain as it depends strongly on the precision of the system identification method. Some research on robust estimation of the

gradient/partial derivatives for neural network based approaches can be found (e.g. see Cardaliaguet and Euvrard [1992] Wang et al. [2019]), however the impact of the estimation error on the partial derivatives estimates and thereby on the α^* estimates remains an open question that we will explore in future work.

5.3 Trading Mean Reversion

In this section, we apply the theoretical results developed in Section 5.2.1 and Section 5.2.2 to inform financial trading decisions of statistical arbitrage strategies. In general terms, these strategies can be defined as trading frameworks that utilise inter-dependencies between the price time series of a set of financial assets to construct a portfolio containing these assets that generates consistent market neutral returns. Although the approach to detecting and leveraging the inter-dependencies can be quite varied (see Avellaneda and Lee [2010], Krauss [2017]), the end goal is generally the same: constructing a mean reverting time series from the underlying financial data that can be studied to obtain trading signals. To make this clearer, we provide the following example that considers the popular statistical arbitrage strategy of pairs trading:

Example 5.3.1 *One trades a synthetic asset whose price series Z is computed as the difference of two other assets X, Y. That is, one trades $Z_t = X_t - \beta y_t$.* Hedging coefficient β *is tuned to render Z mean reverting. A pairs trading strategy then aims to profit by leveraging the mean reverting behaviour of the synthetic asset. It enters a long trade whenever the price of the synthetic asset reaches a threshold level U_1 that is far below the mean. It closes the long trade whenever the asset price has reverted back to a level L_1 close to the mean by selling it. Conversely, the strategy goes short trade is initiated the price of Z reaches a level U_2 that is far above the mean by short-selling the synthetic asset and closes out the position upon reaching a level L_2 near the mean. This is illustrated in Figure 5.31[a]. Here we traded a simulated synthetic asset employing our strategy.*

The optimisation of the U, L thresholds described in Example 5.3.1 is a key component of creating a successful statistical arbitrage strategy as it provides the rules for systematically entering and exiting the underlying trading positions. The choice of these thresholds directly impacts the return, volatility and average holding time of the strategy, but is difficult to do in practice given the noisiness of financial data. In this section, we focus on the problem of threshold setting and propose a machine learning based approach which utilises the theoretical results developed in this chapter to derive optimal thresholds.

5.3.1 Existing Approaches

Existing academic literature on trading decision rules based on threshold setting can be separated into three broad categories (See Krauss [2017] for a recent overview of the statistical arbitrage literature).

- **Fixed model approach**: This approach assumes that the price series of the synthetic asset follows a discretised Ornstein-Uhlenbeck model (AR(1)) in order to derive optimal trading thresholds and policies that optimise standard trading metrics (see Elliott et al. [2005]; Bertram [2010]). Extensions for OU modelling with jump processes (Stübinger and Endres [2018]), stop-loss rules (Leung and Li [2015]) and regime-switching (Bai and Wu [2018]) have also been developed.

- **Heuristic approach**: A commonly used approach in practice is the one described in (Gatev et al. [2006]) which simply sets the entry thresholds(U) at two standard deviations and the exit threshold (L) at the long-term mean. Variations of this approach have also been implemented by changing the entry thresholds through heuristics based on the financial data considered.

- **Optimisation approach**: This approach sets the U, L thresholds by directly maximising returns based on past observations of the price time series of the synthetic asset Z (e.g. Vidyamurthy [2004]). While this type of framework may seem optimal, it is prone to data snooping and has been avoided in existing

literature.

Apart from a few exceptions (e.g. Dunis et al. [2008] Dunis et al. [2015]) which do not focus directly on threshold setting, most of the relevant research has ignored settings where the synthetic asset has been identified with a (non-linear) machine learning method. This is despite the fact that learning based approaches offer the flexibility to capture stylised facts such as the existence of asymmetrical mean reversion (Chen et al. [2011]) or the complicated non-linear nature of the autocorrelation structures of financial price time series. Note that these properties are not indicated by the statistical testing that is commonly used to detect mean reversion i.e. unit-root stationarity tests and can generally only be observed by a more precise modelling of the mean reversion of the time series of the constructed synthetic asset (see Choi and Moh [2007]).

Our approach. In Section 5.3.3, we provide a threshold setting approach that fits a neural network to the synthetic mean reverting time series and applies Algorithm 1 to obtain subsets $\mathcal{V} \in \mathbb{R}_+$ and corresponding $\alpha^*_{\mathcal{V}}$ coefficients that identify domains of the subspace where $(y_t)_{t\in\mathbb{N}}$ is quickly mean reverting. This is described more precisely in Algorithm 2. Before that, in Section 5.3.2, we briefly discuss how the first time guarantees developed in this chapter can be directly leveraged to obtain theoretical guarantees on the trade returns that depend on the choice of the U, L trading thresholds.

5.3.2 Statistical Arbitrage with Precise Knowledge of α^*

Using our theoretical results, we show how, having obtained precise estimates for the α^*-coefficients and the first hitting time bounds from Theorem 5.2.9, we can inform the selection of the entry and exit trading thresholds U, L such that we get a probabilistic guarantee on the trade time and expected return. An illustrative example with a simulated synthetic asset is given in Figure 1[b,c]. To tune U, L and understand the profitability properties of the trades of the strategy, we are interested in bounds involving the following variables:

Definition 5.3.2 (Informal definition of trading variables)

- $r(U, L, c)$*: return of a single trade at thresholds* (U, L) *and transaction cost* c*.*
- $S(U, L)$*: time taken to close positions once they have been opened (with threshold (U,L)).*
- $\mathcal{R}_{Trade}(U, L, c) := \frac{r(U,L,c)}{S(U,L)}$*: average return of a single trade per unit of time with thresholds* (U, L)*.*

Under common noise assumptions (e.g. Gaussian with finite standard deviation of σ), we can utilise Theorem 5.2.9 to obtain an upper bound $\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p)$ on $S(U, L)$ that holds with high probability $p \in [0, 1)$; $\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p) := \min\{T \in \mathbb{N} | \, 1 - \mathfrak{I}^+_{(\alpha^*,y^*)}(T) \geq p\}$ where $\mathfrak{I}^+_{(\alpha^*,y^*)}$ depends on the choice of U, L and σ. This upper bound can then be used to set a probabilistic guarantee on the average return per unit of time;

Proposition 5.3.3 *Let* $\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p)$ *be as defined above,*

$$\mathbb{P}\Big(\mathcal{R}_{Trade}(U, L) \geq \frac{r(U, L, c)}{\mathcal{T}_{(\alpha^*,\sigma)}(U, L, p)}\Big) \geq p \tag{5.4}$$

where $p \in [0, 1)$ *is a chosen confidence level. Furthermore, we have*

$$\mathbb{E}[\mathcal{R}_{Trade}(U, L)] \geq \frac{r(U, L, c)}{1 + \sum_{T=1}^{\infty} \mathfrak{I}^+_{(\alpha^*,y^*)}(T)}. \tag{5.5}$$

Proof This result follows directly from Theorem 5.2.9 and Jensen's inequality.

■

Proposition 5.3.3 is comparable to the semi-analytical trading guarantees derived under an AR(1) model assumption by Bertram [2010]. As a lower bound for $r(U, L, c)$ is generally easily obtainable by considering the difference in value of the underlying positions at U and L, (5.4) and (5.5) can be used to determine trading thresholds that guarantee in expectation or with high probability a sufficiently high average return per unit time.

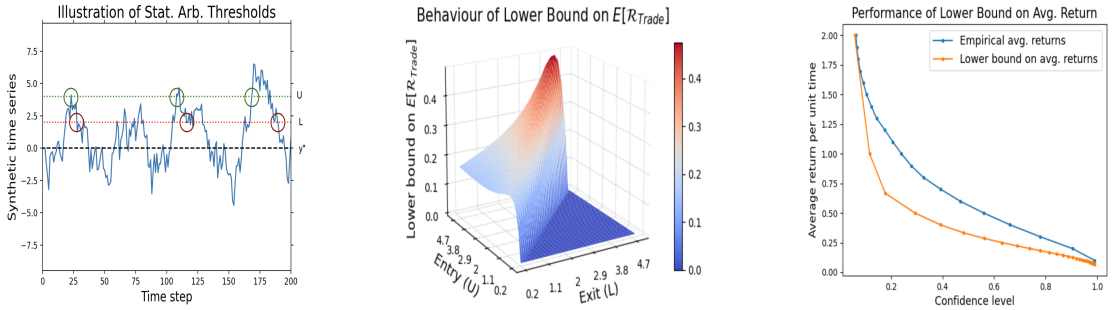


Figure 5.31: **[a]**: *Statistical arbitrage thresholds for the short position. Positions are opened (green circles) when the time series hits U (green line) and closed (red circles) when it subsequently hits L (red line). The dashed black line represents the "fixed point" y^*.* **[b]**: *Dependence of the expected return lower bound guarantee (Eq. 5.5) on the entry threshold (U) and exit threshold (L). Here, $\alpha^* = (0.7, 0.15, 0.05)$ and U, L are given in units of noise standard deviation.* **[c]**: *Setting thresholds $U = 4.4, L = 2.2$, the bound on $\mathcal{R}_{Trade}(U, L)$ given in Eq. 5.4 is illustrated empirically for various choices of confidence levels (p) by computing the empirical distribution of the returns for positions opened and closed at thresholds (U, L).*

.

The final optimisation of the trading thresholds will then also depend on the number of times the position entry threshold U is hit (i.e. the number of times a position in the underlying securities can be opened), the desired duration of the trade and the average return per unit of time of other trading opportunities in the portfolio. Figure 5.31[b] provide an illustration of the behaviour of the lower bound guarantees on $\mathbb{E}[\mathcal{R}_{Trade}(U, L)]$ stated in (5.5) for various values of U and L. These lower bounds were computed in the context of a simple case of a single mean reverting asset (implies $r(U, L, c) \geq U - L$) when the dynamics of the synthetic asset were assumed to be α^*-Lipschitz contracting with $\alpha^* = (0.7, 0.15, 0.05)$. For a specific choice of U, L, Figure 5.31[c] illustrates the lower bound stated in (5.4). To obtain the bound, the relation $r(U, L, c) \geq U - L$ was utilised. The experiments were run 5000 times by simulating from a neural network (4-layers, Relu activation, trained on real financial data) with $\alpha^* = (0.7, 0.15, 0.05)$ in order to obtain the illustrated empirical distribution. As expected, for each confidence level p the curve representing the lower bound given in (5.4) lies beneath the curve representing the empirically estimated $(1 - p)$-th quantile of the average return per unit of time.

Algorithm 2 α^* Threshold Setting Approach:

Input: A time series data set $(y_t)_{t\in\{1,\dots,n\}}$, Mean-reversion speed parameter: $\delta \in (0,1)$
Output: Trading thresholds U, L

1: Apply Algorithm 1 to find $\{l_1^1, ..., l_d^1\}$, $\{l_1^2, ..., l_d^2\}$ and a mean reverting subset $\mathcal{V}^* := \mathbb{R}\setminus\prod_{i=1}^d [l_i^1, l_i^2]$ such that $\bar{\alpha}^*_{\mathcal{V}_1^*}, \bar{\alpha}^*_{\mathcal{V}_2^*} \leq \delta$ where $\mathcal{V}_1^* := \prod_{i=1}^d [l_i^2, \infty)$ and $\mathcal{V}_2^* := \prod_{i=1}^d (-\infty, l_i^1]$
2: Estimate the past standard deviation σ of the synthetic asset
3: Set the lower threshold for position exit at $L_1 = l_1^1$ and the upper threshold for position exit at $L_2 = l_1^2$
4: Set the lower threshold for position entry at $U_1 = \{l_1^1 - \sigma, ..., l_d^1\}$ and the upper threshold for position entry at $U_2 = \{l_1^2 + \sigma, ..., l_d^2\}$

Figure 5.32: (Algorithm) α^* threshold setting approach for improving trading decision rules in statistical arbitrage strategies.

5.3.3 Statistical Arbitrage with Non-linear Mean Reversion

The theoretical trade return guarantees presented in the previous subsection depended on the assumption that precise estimates of the partial derivatives of f are readily available in order to infer the α^* coefficients. Unfortunately, this assumption does not generally hold in practice. In the case where only an approximation of the partial derivatives can be determined, we propose an alternative threshold setting rule that utilises a small neural network based time series model to estimate $\max_x \|\nabla f(x)\|_1$ and the α^* coefficients and then utilises these estimates to set the position exit threshold (i.e. L). The position entry threshold (i.e. U) is then set heuristically depending on the exit threshold, observed past standard deviation and transaction cost. The goal of this approach is not to directly optimise for higher returns but rather to utilise the functional flexibility of the neural network and the theory developed in this chapter (see in particular Propositions 5.2.6, 5.2.10 and Corollary 5.2.11) to decrease return volatility and asset holding time of each trade by identifying subsets of the state space that exhibit faster mean reversion and setting the trading thresholds accordingly. To this end, a hyper-parameter δ can be specified in Algorithm 2 to indicate the desired mean reversion speed in our trading decision rules. Ideally, trades based on these decision rules should be quick, with constant high daily returns and should therefore be well suited to trading environments with high transaction costs, when significant turnover is needed (i.e. due to drift, avoiding regime shifts, etc.) and/or when numerous trading strategies are being deployed simultaneously.

We note that our proposed approach focuses on setting optimal thresholds and therefore assumes that a stationary mean reverting time series has already been constructed. Mean reversion is however indirectly considered as the $\max_x \|\nabla f(x)\|_1$ will automatically increase in value as the time series stops mean reverting.

Outline of the experiment: To illustrate the performance of the proposed α^* threshold setting approach we perform a series of experiments on both real and artificial data. Since the purpose of this chapter is the improvement of trading decision rules on mean reverting assets, we focus on implementing and comparing various approaches at this level of a statistical arbitrage strategy and do not implement a full strategy i.e. constructing the mean reverting time series from underlying financial assets. The selected stocks were found in various academic texts and a *look-ahead bias* was used in order to ascertain the mean reversion of the synthetic asset; we verify that the synthetic time series has an ADF test with a p-value under 0.1 over the full time period. An overview of the data used:

- Real Data: Four pairs/mean reverting equity price time series are used; (V;MA), (EWA;EWC), (VNRX), (EURN) over the time period 2017-2021. We train the neural network model and apply Algorithm 2 on the data between 2017 and 2019. The performance of the strategy is then measured on the 2019-2021 period.

- Artificial Data: Two-layer neural networks with ReLu activation were trained on real data and then used to generate time series. In this case, the partial derivatives of f and therefore the "optimal" trading thresholds of our approach are assumed known apriori. The performance of the α^* threshold setting approach is tested on a "2 year" trading simulation" (500 data points) . This experiment is conducted 200 times and averaged.

The performance of the α^* threshold setting approach is benchmarked against two standard trading decision rules; (1) a *fixed model* approach that assumes that the synthetic asset follows an AR(1) model in order to utilise analytical formulae given by Bertram [2010] to set the upper and lower thresholds. These serve as both position entry and exit thresholds. (2) A *heuristic approach* used in Gatev et al.

[2006] which sets the upper and lower entry thresholds at 2 standard deviations above/below the empirical mean and the position exit threshold at the mean.

To measure the performance of the proposed threshold setting approach a series of well known financial measures is used; total return,Ssharpe ratio, maximum drawdown and average holding time. Additionally, we add two measures;

1. Sharpe ratio 2: Sharpe ratio computed only on days where an underlying position in the assets is open. This quantity provides a better understanding of the trade return volatility and makes sense as in practice one would trade on a number of pairs concurrently.

2. Trade return average: discussed in the previous subsection and is the average daily return of an opened position. This measure gives an understanding of the relation between return and holding time of the asset.

The results of the empirical analysis are presented in Table 5.41 and Table 5.42 (see Appendix) for varying levels of transaction costs. In addition to the financial measures discussed above we also report the entry and exit thresholds of each trading approach.

Threshold Comparison: In the conducted experiments, the thresholds set by the AR(1) approach are closely positioned around the mean. This makes the strategy dependent on a precise estimation of the mean or the drift of the synthetic asset and might lead to heavy losses when the time series no longer mean reverts to a point between the two thresholds. In contrast, GGR thresholds are generally set much further away from the empirical mean of the time series mitigating the risk of misspecified mean reversion outside the position entry thresholds. The GGR thresholds perform optimally when mean crossing happens frequently i.e. when strong "global" mean reversion is observed. Finally, the thresholds set by the proposed α^* approach are non-symmetrical and the specified exit thresholds are significantly more conservative than the other two approaches. While certain trading opportunities may be missed in this case, additional flexibility/error in the specification of the empirical mean or drift of the synthetic time series is obtained. An illustration of the

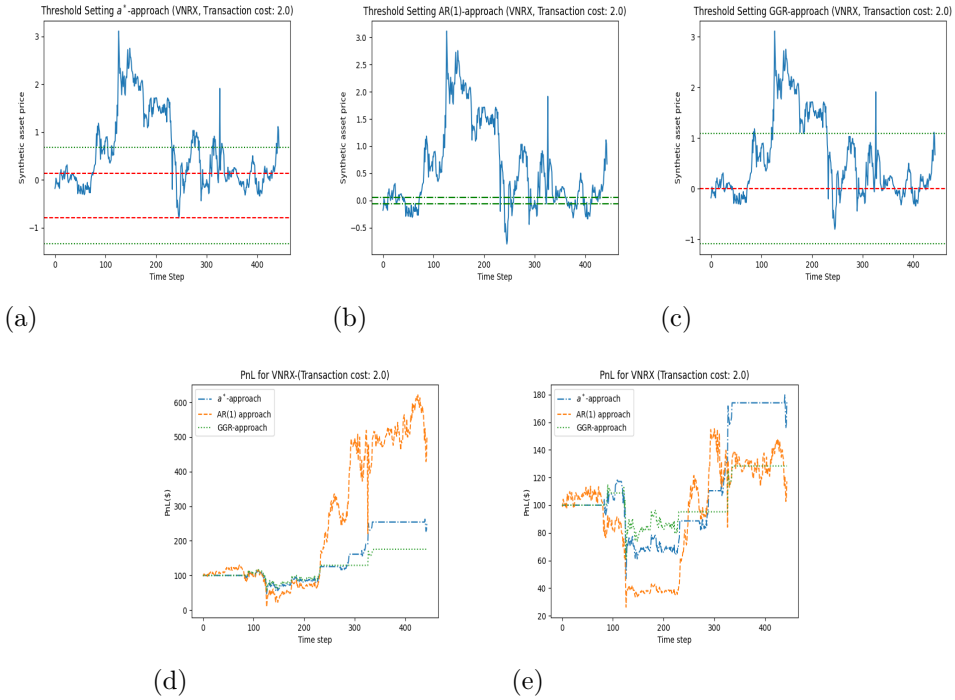


Figure 5.33: **[a,b,c]** Illustration of the α^*, AR(1) and GGR threshold setting approaches that define a trading signal based on the mean reversion of VNRX. The red lines represent the position exit thresholds and the green lines the position entry thresholds. For **[b]** the green lines represent both entry and exit thresholds. **[d]** PnL of trading decision rules based on the threshold setting approaches with 2% transaction cost. **[e]** PnL of trading decision rules based on the threshold setting approaches with 2% transaction cost with additional max constraint set on holding time (two weeks).

differences in threshold setting is given in Figures 5.33[a], 5.33[b] and 5.33[c].

Trading Performance: The performance of the trading decision rules based on the thresholds described in our empirical analysis offers insight on the advantages and disadvantages of each approach. When the dynamics of the underlying model are captured well by an AR(1) model, the AR(1) trading decision rules offer a way of maximising returns by trading quickly and frequently. However, the return volatility of held positions by this approach as measured by Sharpe Ratio 2 tends to be considerably higher than other approaches which implies that significant risk is taken to attain the higher returns. Furthermore, in cases where the noise model is not well specified which inevitably happens in practice, e.g. un-modeled exogenous variables, the maximum drawdown can be heavy (For VNRX the maximum drawdown reaches -91.30%!) and the maximum holding time can last an extended period of time.

Alternatively, the GGR trading decision rules tend to mitigate some of these issues as observed by the smaller empirical Sharpe Ratio 2 and maximum drawdown values in Table 5.41. However, this trading approach often results in high average holding times which are not sufficiently compensated by the increased trade return (i.e. difference between entry and exit thresholds) as described by the empirical trade return average. High holding time can be particularly problematic in practice as it leads to higher transaction costs and increased risk of being caught in a regime switch before the position is exited.

The observed performance of the proposed α^* trading decision rules in our experiments shows an approach that has a low average holding time and a high trade return average comparable to the values obtained by the AR(1) decision rules. In addition, the α^* approach maintains a low return volatility, relatively low maximum drawdown and a high Sharpe Ratio 2 which consistently outperforms both of the other approaches. The total returns of the α^* trading decision rules are generally smaller than the total returns of the other two threshold setting approaches as our proposed method trades less frequently than the AR(1) decision rules and for a smaller potential trade return than the trades specified by the GGR decision rules. The differences in total returns are however often not substantial.

In essence, the α^* trading decision rules sacrifices trading opportunities and trade returns in order to focus on high quality trades with low return volatility and holding time. As noted above, this is an expected consequence of our theoretical results as the flexibility of neural network modelling and the selection of a mean-reversion speed hyper-parameter δ in Algorithm 2 can be used to identify subsets of the state space where the time series mean reverts faster in terms of first hitting time (see Proposition 5.2.10). It is important to note that we have assumed that no portfolio adjustements due to drift or hedging were required in our experiments. If these conditions are included in the trading environment it should further improve the total return performance of the α^* trading decision rules relative to the other approaches. This is illustrated by Figure 5.33[e] where an additional constraint of total position turnover was imposed after 10 time steps.

5.4 Conclusions

This chapter is, to the best of our knowledge, the first to model mean reversion in the context of machine learning- based time series modelling. The chapter derives theoretical properties of non-linear autoregressive processes specifically intended for this goal and describes how these results can be leveraged in practice with popular machine learning frameworks. By doing so, the proposed approach takes advantage of the structural flexibility of a machine learning based modelling approach in order to obtain a more precise characterisation of mean reversion.

Specifically, we present theoretical results extending the class of general non-linear geometrical ergodic processes and derive novel first hitting time bounds for non-linear time series that can be interpreted as a measure of mean reversion speed. These results rest on contraction conditions that can be transformed into assumptions on the partial derivatives of an underlying autoregressive model and therefore be readily verified by neural network based approaches through automatic differentiation (cf. Algorithm 1). We provide a brief experiment on synthetic data to showcase how our approach can improve on existing mean reversion modelling frameworks by considering the first hitting times of contracting non-linear time series.

As an application we show how this chapter's theoretical results can be utilised in practice to define trading decision rules of statistical arbitrage strategies. The trading strategy constructed in this chapter shows how a more precise modelling of mean reversion can be utilised to improve performance compared to commonly used trading decision rules in terms of low trade return volatility and holding time, while minimally affecting trade return and Sharpe ratio.

Table 2[a]: α^* approach (Transaction Cost: 0.2%)

Stock	Upper In/Exit	Lower In/Exit	Return	Sharpe	Sharpe 2	Return Avg.	Drawdown	Avg. Holding
V-MA	17.51/11.19	-19.83/-13.51	0.29	1.32	2.69	n 1.15	-7.10	15.43
EWA-EWC	1.02/0.37	-1.31/-0.65	0.10	0.96	4.39	0.61	-4.70	7.33
VNRX-	0.68/0.13	-1.35/-0.8	2.14	1.30	1.97	3.80	-56.90	38.80
EURN-	1.15/0.59	-0.98/-0.42	0.98	1.00	1.42	1.79	-27.50	26.14
Simulation	14.89/7.6	-20.29/-13.01	1.47	3.09	13.17	3.39	-0.03	3.40

Table 2[b]: GGR approach (Transaction Cost: 0.2%)

Stock	Upper In/Exit	Lower In/Exit	Return	Sharpe	Sharpe 2	Return Avg.	Drawdown	Avg. Holding
V-MA	12.26/0.0	-12.26/0.0	0.37	1.29	1.60	0.57	-9.5	41.00
EWA-EWC	1.27/0.0	-1.27/0.0	0.04	0.37	1.34	0.13	-4.7	34.00
VNRX-	1.1/0.0	-1.1/0.0	0.96	0.93	1.55	2.12	-53.4	80.00
EURN-	1.11/0.0	-1.11/0.0	1.21	1.07	1.40	0.97	-28.6	37.00
Simulation	14.96/0.0	-14.96/0.0	1.52	1.99	2.30	1.18	-0.1	30.62

Table 2[c]: AR(1) model approach (Transaction Cost 0.2%)

Stock	Upper In/Exit	Lower In/Exit	Return	Sharpe	Sharpe 2	Return Avg.	Drawdown	Avg. Holding
V-MA	0.29/-0.29	-0.29/0.29	0.23	0.77	0.77	0.30	-10.70	18.08
EWA-EWC	0.03/-0.03	-0.03/0.03	0.21	1.17	1.17	0.19	-4.70	14.43
VNRX-	0.01/-0.01	-0.01/0.01	7.72	0.95	0.96	10.89	-91.30	11.51
EURN-	0.02/-0.02	-0.02/0.02	1.14	0.77	0.78	2.08	-46.50	14.96
Simulation	0.31/-0.31	-0.31/0.31	2.37	1.75	1.57	2.24	-0.11	10.82

Table 5.41: Performance of the α^*, AR(1) and GGR threshold setting approaches in a low transaction cost environment based on the empirical analysis described in Section 5.3.3

Table 3[a]: α^* approach (Transaction Cost: 2.0%)

Stocks	Upper In/Exit	Lower In/Exit	Returns	Sharpe	Sharpe 2	Trade Ret.	Drawdown	Avg. Holding
V-MA	20.96/11.19	-23.28/-13.51	0.11	0.58	1.35	0.52	-7.40	20.50
EWA-EWC	1.4/0.37	-1.68/-0.65	0.08	1.01	5.43	0.51	-2.00	8.00
VNRX-	0.75/0.13	-1.42/-0.8	1.55	1.11	1.70	3.09	-57.70	46.75
EURN-	1.32/0.59	-1.15/-0.42	0.85	0.89	1.27	1.54	-29.20	25.86
Simulation	14.89/7.6	-20.29/-13.01	0.83	2.49	9.40	2.22	-0.03	3.40

Table 3[b]: GGR approach (Transaction Cost: 2.0%)

Stocks	Upper In/Exit	Lower In/Exit	Returns	Sharpe	Sharpe 2	Trade Ret.	Drawdown	Avg. Holding
V-MA	15.71/0.0	-15.71/0.0	0.16	0.60	0.90	0.31	-10.8	39.20
EWA-EWC	1.65/0.0	-1.65/0.0	0.07	0.63	2.36	0.22	-4.2	31.00
VNRX-	1.17/0.0	-1.17/0.0	0.76	0.77	1.30	2.04	-52.9	79.00
EURN-	1.28/0.0	-1.28/0.0	1.23	1.05	1.42	1.37	-29.2	34.67
Simulation	18.56/0.0	-18.56/0.0	1.10	1.57	1.91	0.91	-0.1	30.62

Table 3[c]: AR(1) model approach (Transaction Cost 2.0%)

Stocks	Upper In/Exit	Lower In/Exit	Returns	Sharpe	Sharpe 2	Trade Ret.	Drawdown	Avg. Holding
V-MA	2.87/-2.87	-2.87/2.87	0.07	0.24	0.24	0.05	-12.10	40.80
EWA-EWC	0.31/-0.31	-0.31/0.31	0.03	0.15	0.15	0.02	-5.70	49.88
VNRX-	0.06/-0.06	-0.06/0.06	4.04	0.61	0.62	7.26	-93.40	24.35
EURN-	0.14/-0.14	-0.14/0.14	0.72	0.56	0.57	1.27	-48.50	24.38
Simulation	3.0/-3.0	-3.0/3.0	0.72	0.74	0.92	0.88	-0.12	22.61

Table 5.42: Performance of the α^*, AR(1) and GGR threshold setting approaches in a high transaction cost environment. Note that only the AR(1) approach depends on the transaction cost to set trading thresholds.

Appendices

Contents

Appendix 5.A Geometric Ergodicity & Mean Reversion.

As noted by (Domowitz and El-Gamal [2001], Geman [2007], Fouque et al. [2011]), geometric ergodicity and mean reversion are closely related concepts. In this section, we recall the definition of geometric ergodicity and briefly discuss this connection.

In order to do so, we consider a d-dimensional reformulation of the $(y_t)_{t\in\mathbb{N}}$ process defined in Section 5.2. More formally, $(Y_t)_{t\in\mathbb{N}}$ is a $\mathbb{R}^d$-valued 1-step Markov process that satisfies:

$$Y_{t+1} = \Phi(Y_t) + \epsilon_t e_1 \tag{5.6}$$

where $\Phi : \mathbb{R}^d \to \mathbb{R}^d$, $\Phi(x_1, ..., x_d) = (f(x_1, ..., x_d), x_1, ..., x_{d-1})^\top$ and e_1 is the first unit vector of $\mathbb{R}^d$. We also specify the state space: $(\mathbb{R}^d, \mathcal{B}_d, \mu_d)$ of $(Y_t)_{t\in\mathbb{N}}$ where $\mathcal{B}_d$ is the class of Borel sets on $\mathbb{R}^d$ and μ_d is the Lesbegue measure. The geometric

ergodicity of $(Y_t)_{t\in\mathbb{N}}$ can then be defined as follows.

We first say that $(Y_t)_{t\in\mathbb{N}}$ is ϕ-irreducible if there exists a non-trivial probability measure ϕ on $\mathbb{R}^d$ such that for every set $S \in \mathcal{B}_d$ of non ϕ-measure 0,

$$\sum_{n=1}^{\infty} p^n(S|a) > 0,$$

where $p^t(S|a)$ is defined by $p^t(S|a) := \mathbb{P}(Y_t \in S|Y_{(-1:-t)} = a)$. Assuming that $(Y_t)_{t\in\mathbb{N}}$ is ϕ-irreducible, we can then define geometric ergodicity as follows; $(Y_t)_{t\in\mathbb{N}}$ is geometrically ergodic if there exists an invariant probability measure π (defined below) and a constant $\rho > 1$ such that $\forall x \in \mathbb{R}$,

$$\lim_{t\to\infty} \frac{1}{\rho^t} \|p^t(dS|a) - \pi(dS)\| \to 0 \tag{5.7}$$

where $\|.\|$ denotes the variation norm on the space of signed measures on $(\mathbb{R}^d, \mathcal{B}_d)$. Furthermore, if (5.7) holds, then π is the unique invariant probability measure for $(Y_t)_{t\in\mathbb{N}}$ (Nummelin [2004]). We recall that a measure π defined on $(\mathbb{R}^d, \mathcal{B}_d)$ is said to be an invariant measure for $(Y_t)_{t\in\mathbb{N}}$ if $\forall S \in \mathcal{B}_d$,

$$\pi(S) = \int_{\mathbb{R}^d} p^n(S|a)\pi(da).$$

The geometric ergodicity property described in (5.7) implies that the long-run probabilistic behaviour of $(Y_t)_{t\in\mathbb{N}}$ and therefore of $(y_t)_{t\in\mathbb{N}}$ will be stable, converging to a fixed invariant probability measure. This notion can be more explicitly connected to mean-reversion through the Birkhoff-Khinchin theorem (Birkhoff [1931]) which implies that if (5.7) holds, then for all measurable[3] functions $g : \mathbb{R}^d \to \mathbb{R}$, the long-run time average of $(g(Y_t))_{t\in\mathbb{N}}$ converges almost surely to the deterministic average with respect to the invariant probability measure π of $(Y_t)_{t\in\mathbb{N}}$:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} g(Y_t) = \int_{\mathbb{R}^d} g(Y)\pi(dY).$$

[3]With respect to the π.

In particular, if g is the projection of vectors $x \in \mathbb{R}^d$ onto their first component x_1, then:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} y_t = \mathbb{E}_\pi[Y]$$

where Y is a random variable on $(\mathbb{R}, \mathcal{B})$ with distribution π. This result states that the long-run average of c will converge to a fixed constant determined by the invariant measure π which implies that $(y_t)_{t\in\mathbb{N}}$ will either satisfy $\lim_{T\to\infty} y_t = \mathbb{E}_\pi[Y]$ or will cross-over $\mathbb{E}_\pi[Y]$ infinitely many times. In the context of the trading application considered in Section 5.3, both of these cases provide pertinent information on mean reverting characteristics of $(y_t)_{t\in\mathbb{N}}$ which can be used to inform decision rules regarding trading threshold setting.

Appendix 5.B Proof of Ergodicity & Stationarity Results

In order to utilise past results (see Tweedie [1976], Chan and Tong [1985], Tjøstheim [1990]) to show the geometrical ergodicity of $(y_t)_{t\in\mathbb{N}}$, we will consider the reformulation of $(y_t)_{t\in\mathbb{N}}$ into a $\mathbb{R}^d$-valued 1-step Markov process: $(Y_t)_{t\in\mathbb{N}}$ in this chapter.

Proof of Lemma 5.2.2.

Follows from Banach's fixed point theorem and Theorem 5.1 of Chan and Tong [1985].

■

Notation: For $c \in \mathbb{R}^d, R \in \mathbb{R}_+$, we define the ball centered in c and with radius R as $B(0, R) = \{x \in \mathbb{R}^d : ||x||_\infty < R\}$

Proof of Theorem 5.2.3. Without loss of generality, we suppose $K := \overline{B(0, R)}$ for $R \in \mathbb{R}_+$.

Let $(Y_t)_{t\in\mathbb{N}}$ be the equivalent one step Markov chain in $\mathbb{R}^d$ defined in (5.6). The proof of Theorem 5.2.3 follows from the two following results;

Lemma 5.B.1 *(Section 2 of Chan and Tong [1985]) Let f be as defined in Theorem 5.2.3 and assume that the noise assumption defined in Assumption 15 holds. Then, $(Y_t)_{t\in\mathbb{N}}$ is μ_d-irreducible and aperiodic. Furthermore, any compact set $K \subset \mathbb{R}^d$ is small.*

Theorem 5.B.2 *(Extension of Tweedie's Criterion given in Tjøstheim [1990]) Let $(Y_t)_{t\in\mathbb{N}}$ be μ_d-irreducible and aperiodic. Suppose that there exists a non-negative measurable function $g : \mathbb{R}^d \to \mathbb{R}$, positive constants $c_1, c_2 \in \mathbb{R}_+$, $\rho \in (0,1)$,a small set $\tilde{K}$ and a positive integer $h \in \mathbb{N}$, such that*

$$\mathbb{E}[g(Y_{t+h})|Y_t = a] \leq \rho g(a) - c_1, \quad \forall a \notin \tilde{K} \tag{5.8}$$

$$\mathbb{E}[g(Y_{t+h})|Y_t = a] \leq c_2, \quad \forall a \in \tilde{K} \tag{5.9}$$

then $(Y_t)_{t\in\mathbb{N}}$ is geometrically ergodic.

In order to show the geometric ergodicity of $(y_t)_{t\in\mathbb{N}}$, it is therefore sufficient to find $g : \mathbb{R}^d \to \mathbb{R}$, $c_1, c_2 \in \mathbb{R}_+$, $\rho \in (0,1)$ and a small set $\tilde{K}$ such that equations (5.8) and (5.9) are verified. In order to do so, we choose $g : \mathbb{R}^d \to \mathbb{R}$, $g(x) = \|x\|_\infty$ and $h = d$. We now show that there exists a positive constant $c_1 \in \mathbb{R}_+$, $\rho \in (0,1)$ and a small set $\tilde{K}$ such that our choices of g and h satisfy the drift condition (5.8).
Fix an arbitrary $a \notin K$, we consider two cases;

1. $\{Y_{t+j}\}_{j\in\{1,\dots,d\}} \subseteq \mathbb{R}^d \setminus K$.
2. There exists $j \in \{1,\dots,d\}$ such that $Y_{t+j} \in K$.

(1.): Consider $j \in \{1,\dots,d\}$, and fix an arbitrary point $z \in \mathbb{R}^d \setminus K$. Then, for $Y_{t+j} \notin K$ and initial conditions $Y_t = a$ (Note: $Y_j = (y_j, \dots, y_{j-d+1})^\top$), we denote: $\forall i \in \{0,\dots,d-1\}$; $\bar{y}_{t+j-i} := \begin{cases} f(Y_{t+j-i-1}) \text{ if } j > i \\ a_{i-j+1} \text{ otherwise.} \end{cases}$, $\bar{\epsilon}_{t+j-i} := \begin{cases} \epsilon_{t+j-i} \text{ if } j > i \\ 0 \text{ otherwise.} \end{cases}$
and observe the following relation:

$$\begin{aligned} &\|Y_{t+j}\|_\infty \\ &= \max_{i\in\{0,\dots,d-1\}} |y_{t+j-i}| \leq \max_{i\in\{0,\dots,d-1\}} |\bar{y}_{t+j-i} + \bar{\epsilon}_{t+j-i}| \end{aligned}$$

$$\leq \max_{i\in\{0,\dots,d-1\}} |\bar{y}_{t+j-i} - f(z)| + \max_{i\in\{1,\dots,d\}} |\epsilon_i| + |f(z)|$$
$$\leq \bar{\alpha} \max_{i\in\{1,\dots,d\}} \|Y_{t+j-i}\|_\infty + \max_{i\in\{1,\dots,d\}} |\epsilon_i| + \bar{\alpha}\|z\|_\infty + |f(z)|$$
$$\leq \bar{\alpha} \max\{\|Y_{t+j-1}\|_\infty, \|a\|_\infty\} + \max_{i\in\{1,\dots,d\}} |\epsilon_i| + \bar{\alpha}\|z\|_\infty + |f(z)|.$$

As only the first term of the last equation depends on a, we can iterate the above relation to obtain;

$$\|Y_{t+j}\|_\infty \leq \bar{\alpha}\|a\|_\infty + C_j = \bar{\alpha}g(a) + C_j$$

where C_j is random variable that only depends on $\bar{\alpha}$, $\{\epsilon_1, ..., \epsilon_d\}$ and z. Therefore,

$$\mathbb{E}\left[\|Y_{t+d})\|_\infty \middle| Y_t = a, \{Y_{t+j}\}_{j\in\{1,\dots,d\}} \subseteq \mathbb{R}^d \setminus K\right] \leq \bar{\alpha}g(a) + \mathbb{E}[C_d]$$

where $\mathbb{E}[C_d] \in \mathbb{R}$ is a function of the constants $\mathbb{E}[\max_{i\in\{1,\dots,d\}} |\epsilon_i|]$, $\bar{\alpha}\|z\|_\infty$ and $|f(z)|$ and does not depend on a. Set $\tilde{R}' \in \mathbb{R}_+$, $\bar{\alpha}' \in (\bar{\alpha}, 1)$ and $c_1' > 0$, such that $K \subseteq B(0, \tilde{R}')$

$$(\bar{\alpha}' - \bar{\alpha})\tilde{R}' - c_1' > \mathbb{E}[C_d]$$

As the choice of $z \in \mathbb{R}^d \setminus K$ can be same the for all a, this implies that for all $a \in \mathbb{R}^d \setminus \overline{B(0, \tilde{R}')}$, we have

$$\mathbb{E}\left[\|Y_{t+d}\|_\infty \middle| Y_t = a, \{Y_{t+j}\}_{j\in\{1,\dots,d\}} \subseteq \mathbb{R}^d \setminus K\right] \leq \bar{\alpha}g(a) + \mathbb{E}[C_d]$$

$$\leq \bar{\alpha}'g(a) + (\bar{\alpha} - \bar{\alpha}')\tilde{R}' + \mathbb{E}[C_d] \leq \bar{\alpha}'g(a) - c_1'.$$

which verifies equation (5.8) of of Theorem 5.B.2.

(2.): Since there exists $j \in \{1, ..., d\}$ such that $Y_{t+j} \in K$, we cannot use the same approach as above. Define $M := \max\{\max_{x\in K} |f(x)|, \max_{x\in K} \|x\|_\infty\}$ and $j^* := \max\{j \in \{1, ..., d\} \text{ s.t. } Y_{t+j} \in K\}$ which both exist by assumption. We have that either $j^* = d$ and therefore $\|Y_{t+d}\|_\infty \leq M$ or $j^* < d$ in which case $\|Y_{t+j^*}\|_\infty \leq M$. Denoting $\tilde{O}_{j^*+i} := |f(z)| + \bar{\alpha}\|z\|_\infty + |\epsilon_{t+j^*+i}|$ where $z \in \mathbb{R}^d \setminus K$ is the arbitrarily selected point of the first part of the proof, we can use a similar approach to the one

used in **(1.)**, to obtain the following relation:

$$
\begin{aligned}
|y_{t+j^*+1}| &\leq |f(z)| + \bar{\alpha}\|z\|_\infty + |\epsilon_{t+j^*+1}| + \bar{\alpha}\|Y_{t+j^*}\|_\infty \leq \tilde{O}_{j^*+1} + \bar{\alpha}M, \\
|y_{t+j^*+2}| &\leq \tilde{O}_{j^*+1} + \bar{\alpha}\|Y_{t+j^*+1}\|_\infty \leq \tilde{O}_{j^*+2} + \bar{\alpha}\tilde{O}_{j^*+1} + \bar{\alpha}M, \\
&\ldots \\
|y_{t+d}| &\leq \sum_{i=0}^{d-(j^*+1)} \bar{\alpha}^{d-(j^*+1)-i}\tilde{O}_{j^*+i+1} + \bar{\alpha}M.
\end{aligned}
$$

Since $\{\epsilon_i\}_{i\in\{1,\ldots,d\}}$ are i.i.d., we have that $\forall i \neq j$, $\mathbb{E}[\tilde{O}_{t+i}] = \mathbb{E}[\tilde{O}_{t+j}] =: \mathbb{E}[\tilde{O}]$ and therefore

$$
\mathbb{E}\left[\|Y_{t+d}\|_\infty | Y_t = a, \exists j \in \{1,\ldots,d\} \text{ s.t. } Y_{t+j} \in K\right] \leq \sum_{i=1}^{d-1} \bar{\alpha}^{(d-1)-i}\mathbb{E}[\tilde{O}] + \bar{\alpha}M.
$$

As the term on the right-hand side of the above equation is constant we can set $\tilde{R}'' > M$, $\bar{\alpha}'' \in (0,1)$ and $c_1'' > 0$ such that $\sum_{i=1}^{d-1} \bar{\alpha}^{(d-1)-i}\mathbb{E}[\tilde{O}] + \bar{\alpha}M \leq \bar{\alpha}''\tilde{R}'' - c_1''$. For all $a \in \mathbb{R}^d \setminus \overline{B(0,\tilde{R}'')}$,

$$
\mathbb{E}\left[\|Y_{t+d}\|_\infty | Y_t = a, \exists j \in \{1,\ldots,d\} \text{ s.t. } Y_{t+j} \in K\right] \leq \bar{\alpha}''\tilde{R}'' - c_1''.
$$

Finally, let $\tilde{R}''' \in \mathbb{R}_+$ be such that $K \subseteq \overline{B(0,\tilde{R}''')}$ and set $\tilde{R} := \max\{\tilde{R}', \tilde{R}'', \tilde{R}'''\}$, $\alpha := \max\{\bar{\alpha}', \bar{\alpha}''\}$ and $c_1 := \min\{c_1', c_1''\}$. Combining **(1.)** and **(2.)** together, we obtain for all $a \in \mathbb{R}^d \setminus \overline{B(0,\tilde{R})}$,

$$
\begin{aligned}
&\mathbb{E}[\|Y_{t+d}\|_\infty \big| Y_t = a] \\
&\leq \mathbb{P}\left(\{Y_{t+j}\}_{j\in\{1,\ldots,d\}} \subseteq \mathbb{R}^d \setminus K\right) \mathbb{E}\left[\|Y_{t+d}\|_\infty \big| Y_t = a, \{Y_{t+j}\}_{j\in\{1,\ldots,d\}} \subseteq \mathbb{R}^d \setminus K\right] \\
&+\mathbb{P}\left(\exists j \in \{1,\ldots,d\} \text{ s.t. } Y_{t+j} \in K\right) \mathbb{E}\left[\|Y_{t+d}\|_\infty | Y_t = a, \exists j \in \{1,\ldots,d\} \text{ s.t. } Y_{t+j} \in K\right] \\
&\leq \left(\mathbb{P}\left(\{Y_{t+j}\}_{j\in\{1,\ldots,d\}} \subseteq \mathbb{R}^d \setminus K\right) + \mathbb{P}\left(\exists j \in \{1,\ldots,d\} \text{ s.t. } Y_{t+j} \in K\right)\right)\left(\alpha\tilde{R} - c_1\right) \\
&= \alpha\tilde{R} - c_1 \leq \alpha g(a) - c_1
\end{aligned}
$$

which verifies condition (5.8) given in Theorem 5.B.2 is satisfied. The existence of

a positive constant $c_2 > 0$ such that condition (5.9) holds for $\tilde{K} = \overline{B(0, \tilde{R})}$ follows directly from the assumption that f is bounded on compact sets. Therefore, the second condition of Theorem 5.B.2 is satisfied and $(y_t)_{t\in\mathbb{N}}$ is geometrically ergodic.

■

Proof of Proposition 5.2.6. Consider $K = \overline{B_{\|\cdot\|_\infty}(c, R)}$ with $c \in \mathbb{R}^d$ and $R \in \mathbb{R}_+$. We define the following projection function;

$$p_K : \mathbb{R}^d \setminus K \to \mathbb{R}^d \setminus K;$$

$$x \mapsto p_K(x) = \begin{cases} p_K(x)_i = c_i + r, \text{ if } x_i > c_i + r \\ p_K(x)_i = c_i - r, \text{ if } x_i < c_i - r \\ p_K(x)_i = 0 \quad \text{otherwise.} \end{cases}$$

where the image set of p_K is denoted by $P_K := Im(p_K)$. It is trivial to see that P_K is countable and $|P_K| < \infty$.

In order to prove Proposition 5.2.6, we first state and prove the following lemma that considers the projection functions: p_K.

Lemma 5.B.3 *Let $\bar{\alpha} \in (0, 1)$ and $K = \overline{B(c, R)}$ for $c \in \mathbb{R}^d$, $R \in \mathbb{R}_+$ and assume the noise terms $(\epsilon_t)_{t\in\mathbb{N}}$ verify Assumption 15. If f is bounded on compact sets and such that*

$$\forall x \in \mathbb{R}^d \setminus K, \ |f(x) - f(p_K(x))| \leq \bar{\alpha}\|x - p_K(x)\|_\infty$$

then $(y_t)_{t\in\mathbb{N}}$ is geometrically ergodic.

Proof of Lemma 5.B.3. Without loss of generality, we can assume $c = 0$. The lemma then follows from applying the same approach as in the proof of Theorem 5.2.3 replacing the arbitrarily selected point $z \in \mathbb{R}\setminus K$ with $p_K(y_{t+j})$ for $j \in \{1, ..., d\}$ at each iteration of the upper bounds obtained when $y_{t+j} \notin K$. Note that since $p_K(\mathbb{R}^d \setminus K) \subseteq P_k$ which is finite, we have that the implicit dependence[4] of $p_K(y_{t+j})$

[4]In the proof of Theorem 5.2.3, the choice of $z \in \mathbb{R}\setminus K$ was arbitrary and did not depend on a while in the proof of Proposition 5.B.3, by construction, the selected points are given by $p_K(y_{t+j})$ $j \in \{1, ..., d\}$ which depends, through y_{t+j} $j \in \{1, ..., d\}$, on the initial conditions a.

on initial conditions $a \notin K$ can be removed by taking the maximum over the vertex set. This implies that comparable constants to the C_d and $\tilde{O}$ terms defined in the proof of Theorem 5.2.3 can be defined to hold for all a in the context of this proof.

■

Before continuing with the proof of Proposition 5.2.6, we also state the following useful lemma.

Lemma 5.B.4 *(Basic Lemma) Let O be an arbitrary set and $f \in C^1(O)$. If $L := \sup_{x \in O} \|\nabla f(x)\|_1 < \infty$, then $\forall x, y \in O$ such that $v := \{z \in \mathbb{R}^d | \exists \lambda, x = \lambda x + (1 - \lambda)y\} \subseteq O$, $|f(x) - f(y)| \leq L\|x - y\|_\infty$.*

The proof of Proposition 5.2.6 then follows from combining Lemma 5.B.3, Lemma 5.B.4 and the fact that $\forall x \in \mathbb{R}^d \setminus K$, $\{z \in \mathbb{R}^d | \exists \lambda, z = \lambda x + (1 - \lambda)p_K(x)\} \subseteq \mathbb{R}^d \setminus K$.

■

Appendix 5.C Proof of First Hitting Time Guarantees

Proof of Theorem 5.2.9. The proof of Theorem 5.2.9 is given as follows. From Assumption 17, we have that the first hitting time τ_γ^+ can be upper bounded by the first hitting time $\tau_\gamma^z := \inf\{t \in \mathbb{N} | z_t < \gamma\}$ of a linear AR(d) $(z_t)_{t \in \mathbb{N}}$ process with coefficients equal to the α^* vector, initial conditions $(a_1 - y^1, ..., a_d - y^*) \in \mathbb{R}^d$ and same noise process $(\epsilon_t)_{t \in \mathbb{N}}$ as $(y_t)_{t \in \mathbb{N}}$. Then, for an arbitrary $T \in \mathbb{N}$:

$$\mathbb{P}\big(\tau_\gamma^z > T\big) = \mathbb{P}\big(\min_{t \in \{1,...,T\}} z_t > \gamma\big) = \mathbb{P}\big(\bigcap_{t=1}^{T} \{z_t > \gamma\}\big).$$

For every time step $t \in \{1, ..., T\}$, by iterating backwards from timestep t, we can re-express z_t as

$$z_t = \sum_{i=1}^{t} \beta(\alpha^*, t, i)\epsilon_i + \pi_1(B^t(a - y^* \mathbf{1}_d)). \tag{5.10}$$

where $\beta(\alpha^*, t, i)$ is a constant that depends on α^*, t and i. Define $\sigma^2 := var(\epsilon_1)$ which is finite and non-zero by Assumption 16. The independence and identical distributions of the noise variables imply $\forall s \leq t \in \mathbb{N}$,

$$\frac{cov(z_s, z_t)}{\sigma^2} = \sum_{i=1}^{s} \beta(\alpha^*, s, i)\beta(\alpha^*, t, i) = \langle M(T)_s, M(T)_t \rangle$$

where $M(T)_i$ denotes the i-th row of of a matrix $M \in \mathbb{R}^{T\times T}$. Therefore, the covariance matrix $V_T \in \mathbb{R}^{T\times T}$ of $(z_t)_{t\in\{1,\dots,T\}}$ is given by $V_T = \sigma^2 M(T)M(T)^\top$. From (35) in Wise [1955], we have that the sample covariance of $(z_t)_{t\in\{1,\dots,T\}}$ is known and given by $V_T^{-1} = \frac{1}{\sigma^2} A_{\alpha^*}(T)A_{\alpha^*}(T)^\top$ with $A_{\alpha^*}(T)$ is defined in (5.2). By uniqueness of the square root of a matrix and of the inverse matrix we obtain an explicit expression for $M(T)$: $M(T)^{-1} = A_{\alpha^*}(T)^\top$.

Using the above relation, equation (5.10) and $\det(A_{\alpha^*}(T)) = 1$, we obtain the bound:

$$\begin{aligned}
\mathbb{P}\big(\bigcap_{t=1}^{T}\{z_t > \gamma\}\big) &\leq \mathbb{P}\big(M(T)\epsilon_{1:T} > -b\big) \\
&= \int_{-b_1}^{\infty} \dots \int_{-b_T}^{\infty} \frac{f_{\epsilon_{1:T}}(A_{\alpha^*}(T)\begin{pmatrix} x_1 \\ \vdots \\ x_T \end{pmatrix})}{|\det(A_{\alpha^*}(T)^{-1})|} dx_1 \dots dx_T \\
&= \mathfrak{I}^+_{(\alpha^*, y^*)}(T).
\end{aligned}$$

The second statement of 5.2.9 follows almost immediately. For $N \in \mathbb{N}$, define E_N as the partial sum $E_N := \sum_{T=1}^{N} T\mathbb{P}\big(\tau_\gamma^+ = T\big)$ then

$$\begin{aligned}
E_N &= \sum_{T=1}^{N} T\mathbb{P}\big(\tau_\gamma^+ = T\big) = \sum_{T=1}^{N}\sum_{t=1}^{T} \mathbb{P}\big(\tau_\gamma^+ = T\big) \\
&= \sum_{t=1}^{N}\sum_{T=t}^{N} \mathbb{P}\big(\tau_\gamma^+ = T\big) = \sum_{T=1}^{N} \mathbb{P}\big(\tau_\gamma^+ \geq T\big) \\
&= 1 + \sum_{T=1}^{N} \mathbb{P}\big(\tau_\gamma^+ > T\big) \leq 1 + \sum_{T=1}^{N} \mathfrak{I}^+_{(\alpha^*, y^*)}(T).
\end{aligned}$$

Taking limits on both sides of the inequality yields

$$\mathbb{E}[\tau_\gamma^+] = \sum_{T=1}^{\infty} T\mathbb{P}\big(\tau_\gamma^+ = T\big) \leq 1 + \sum_{T=1}^{\infty} \mathfrak{I}^+_{(\alpha^*, y^*)}(T).$$

■

Proof of Proposition 5.2.10. In this proof, we modify previous notation to emphasize dependence on the parameters studied in Remark 5.2.10 and we define the following notation: for every $T \in \mathbb{N}$, we denote by $J_T \subseteq \mathbb{R}^T$ the set $J_T := \prod_{i=1}^{T}[-b_i, \infty)$ where $b_i := \left(B^i_{\alpha^*}|a - y^*\mathbf{1}_d|\right)_1 - \gamma$. (i): Consider $\gamma_1, \gamma_2 \in [0, a_d - y^*)$ with $\gamma_1 \leq \gamma_2$. We have $\forall i = 1, ..., T$, $b_i(\gamma_1) \geq b_i(\gamma_2)$ which implies that $J_T(\gamma_1) \supseteq J_T(\gamma_2)$ and subsequently

$$\begin{aligned}\mathfrak{I}^+_{(\alpha^*,y^*)}(T,\gamma_1) &= \int_{J_T(\gamma_1)} f_{\epsilon_{1:T}}(A_{\alpha^*}(T)x)dx \\ &\geq \int_{J_T(\gamma_2)} f_{\epsilon_{1:T}}(A_{\alpha^*}(T)x)dx = \mathfrak{I}^+_{(\alpha^*,y^*)}(T,\gamma_2).\end{aligned}$$

(ii): Consider $\alpha^*, \beta^* \in \triangle_+$ with $\forall\, i$, $\alpha^*_i \leq \beta^*_i$ and such that $\exists\, j$ with $\alpha^*_j < \beta^*_j$. It follows that $\forall t \in \mathbb{N}_{>d}$, $z_t(\alpha^*) < z_t(\beta^*))$ (where $z_t(\alpha^*)$ ($z(\beta^*)$ denotes a linear AR(d) process defined with coefficients equal to α^* (β^*)) and noise process $(\epsilon_t)_{t\in\mathbb{N}}$. Then, $\forall T \in \mathbb{N}_{>d}$,

$$\begin{aligned}\mathfrak{I}^+_{(\alpha^*,y^*)}(T) &= \mathbb{P}\big(\bigcap_{t=1}^{T}\{z_t(\alpha^*) > \gamma\}\big) \leq \mathbb{P}\big(\bigcap_{t=1}^{T}\{z_t(\beta^*) > \gamma\}\big) \\ &= \mathfrak{I}^+_{(\beta^*,y^*)}(T).\end{aligned}$$

(iii): Let $J_T^{(-1)}$ denote $\{x \in \mathbb{R}^T | A_{\alpha^*}(T)^{-1}x \in J_T\}$, then

$$\begin{aligned}\mathfrak{I}^+_{(\alpha^*,y^*)}(T) &= \mathbb{P}\big(M(T)\epsilon(T) > -\mathbf{b}\big) = \int_{J_T} f_{\epsilon_{1:T}}(A_{\alpha^*}(T)x)dx \\ &= \int_{J_T^{(-1)}} f_{\epsilon_{1:T}}(x)dx = \int_{\mathbb{R}^T} f_{\epsilon_{1:T}}(x)\mathbf{1}_{J_T^{(-1)}}(x)dx.\end{aligned}$$

where $\mathbf{1}_A(x)$ denotes the indicator of a subset A. Define $g(x, \alpha^*) = f_{\epsilon_{1:T}}(x)\mathbf{1}_{J_T^{(-1)}}$. Since g verifies all the conditions of Theorem 5.6 of Elstrodt [1996] for $\alpha^*_0 := 0$, we have $\int_{\mathbb{R}^T} g(x, \alpha^*)dx$ is continuous in α^*_0. This implies that $\mathfrak{I}^+_{(\alpha^*,y^*)}(T)$ is continuous in $\alpha^* = 0$ and therefore $\lim_{\|\alpha^*\|_1 \to 0} \mathfrak{I}^+_{(\alpha^*,y^*)}(T) = \mathfrak{I}^+_{(0,y^*)}(T) = \frac{1}{2^T}$.

■

6 | Conclusion

Contents

6.1 Summary of contributions

This chapter concludes this thesis by providing an overview of our main research contributions and examining potential avenues for future work that build upon the various results established in previous chapters.

In Chapter 3, we investigated the fundamental problem of estimating the Lipschitz constant of an unknown target function under minimal parametric assumptions. As a first theoretical contribution, we derived novel lower bounds on the sample complexity of this problem for both noise-free and noisy sampling settings under mild assumptions. We then proposed a simple Lipschitz learning algorithm (LCLS[1]) which we showed to be asymptotically consistent under general noise assumptions. We established finite sample guarantees for LCLS thereby deriving upper bounds on the sample complexity of the Lipschitz learning problem. Our analysis shows that the sample complexity rates derived in this chapter are optimal in both the noise-free setting and in the noisy setting when the noise is assumed to follow a

[1] *Lipschitz Constant Estimation by Least Squares Regression.*

Gaussian distribution and that LCLS is a sample-optimal algorithm in both cases. Finally, we showed that, by design, the LCLS algorithm is computationally faster than existing theoretically consistent methods, and can be readily adapted to various noise assumptions with little to no prior knowledge of the target function properties or noise distribution.

In Chapter 4, we examined the asymptotic convergence properties of general Lipschitz interpolation methods under bounded stochastic noise assumptions. We established probabilistic consistency guarantees of the classical approach in a general context and derived upper bounds on the uniform convergence rates consistent with well-known optimal rates of non-parametric regression in related settings. Our bounds contribute to the literature in two ways: Firstly, from a theoretical perspective, they provide a novel characterisation of the non-parametric uniform convergence rate in the bounded noise setting. Secondly, they can serve as a theoretical tool for comparing Lipschitz interpolation to alternative non-parametric regression methods and provide an explicit condition on the behavior of the noise at the boundary of its support, indicating when Lipschitz interpolation should be expected to asymptotically outperform or underperform other approaches. In the second part of this chapter, we extended these results to include additional consistency guarantees for online learning of system dynamics in discrete-time stochastic systems and demonstrated their usefulness in deriving closed-loop stability guarantees for a basic online-learning-based controller. We also derived asymptotic consistency for the fully data-driven LACKI framework (Calliess et al. [2020]) when the assumption of prior knowledge of the Lipschitz constant is removed.

Finally, in Chapter 5, we investigated the theoretical properties, in relation to mean reversion, of general classes of non-linear autoregressive processes that arise in the context of machine learning-based time series modelling. More precisely, we utilised relaxed Lipschitz-type regularity assumptions on the dynamics of these processes to obtain geometric ergodicity/stationarity and to provide tight probabilistic upper bounds on the first hitting times of the process as it reverts back to the mean. As a practical case study, we demonstrated how our theoretical results can be applied

to neural network time series models to define trading decision rules for statistical arbitrage strategies and to provide probabilistic guarantees on the PnL.

6.2 Future Work

While this thesis has provided a comprehensive theoretical study of a variety of theoretical problems pertaining to Lipschitz continuous machine learning, many challenges and open problems persist. In particular, in the long run, it would be of great interest to explore the generalisability of these types of findings beyond the Lipschitz continuous case and to observe how the conclusions reached in this work can be improved under strengthened functional assumptions on the target function, particularly through the consideration of specific modern machine learning models.

For the shorter term, throughout this thesis, we have provided insight into potential extensions and avenues for future exploration. Presented below is a concise compilation of both long term goals and promising extensions stemming from the results developed in this thesis, which we believe would be of interest for future work.

- The LCLS algorithm proposed in Chapter 3 generates global Lipschitz constants. In practice, it is relatively common that local Lipschitz constants are utilised instead to improve performance. A natural extension would therefore be to modify the LCLS algorithm to recursively compute local Lipschitz constants and to obtain theoretical guarantees for this extension. Some initial results in this direction have already been derived and are relatively promising.
- Extending the lower bounds on the sample complexity in the noisy setting derived in Chapter 3 under stronger assumptions on the target function, a more restricted class of Lipschitz learning algorithms and/or weaker noise assumptions is another interesting potential research direction. In particular, it would be of interest to consider the bounded noise setting of Chapter 4 which is generally assumed when applying Lipschitz interpolation frameworks.
- With respect to the asymptotic analysis done in Chapter 4, a potential re-

search avenue would be the derivation of lower bounds on the non-parametric convergence rates under the same settings and assumptions. These would, ideally but not unexpectedly, given existing results on the optimal convergence rates of non-parametric boundary regression by Jirak et al. [2014], demonstrate the optimality of the upper bounds on the non-parametric convergence rates developed in that chapter.

- The extension of the convergence rate upper bounds of Chapter 4 to the more practical and fully data-driven extensions of Lipschitz interpolation such as LACKI (Calliess et al. [2020]), POKI Calliess [2017] or LCLS-KI (Chapter 3) would also be of interest. These new bounds would most likely be probabilistic in nature and could potentially be obtained by combining the results of Chapter 3 and Chapter 4.

- The first hitting guarantees derived in Chapter 5 make nominal model assumptions and do not consider the impact of estimation error. A potential research direction could therefore incorporate this additional uncertainty into the first hitting time bounds to obtain a more practical result.

- A more practical research direction, could consider a more comprehensive investigation into the empirical performance of the neural network-based trading decision rules proposed in Chapter 5 when applied to enhance statistical arbitrage strategies.

Bibliography

Ahsan S Alvi, Binxin Ru, Jan Calliess, Stephen J Roberts, and Michael A Osborne. Asynchronous batch Bayesian optimisation with improved local penalisation. *arXiv preprint arXiv:1901.10452*, 2019.

HZ An and FC Huang. The Geometrical Ergodicity of Nonlinear Autoregressive Models. *Statistica Sinica*, pages 943–956, 1996.

Karl J Åström and Björn Wittenmark. *Adaptive control*. Courier Corporation, 2013.

Marco Avellaneda and Jeong-Hyun Lee. Statistical Arbitrage in the US Equities Market. *Quantitative Finance*, 10(7):761–782, 2010.

François Bachoc, Tom Cesari, and Sébastien Gerchinovitz. Instance-dependent bounds for zeroth-order Lipschitz optimization with error certificates. *Advances in Neural Information Processing Systems*, 34:24180–24192, 2021.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to Explain Individual Classification Decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

Yang Bai and Lan Wu. Analytic Value Function for Optimal Regime-Switching Pairs Trading Rules. *Quantitative Finance*, 18(4):637–654, 2018.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Gopal K Basak and Kwok-Wah Remus Ho. Level-crossing Probabilities and First-Passage Times for Linear Processes. *Advances in applied probability*, pages 643–666, 2004.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Gleb Beliakov. Monotonicity preserving approximation of multivariate scattered data. *BIT numerical mathematics*, 45(4):653–677, 2005.

Gleb Beliakov. Interpolation of Lipschitz functions. *Journal of computational and applied mathematics*, 196(1):20–44, 2006.

Felix Berkenkamp and Angela P Schoellig. Safe and robust learning control with Gaussian processes. In *2015 European Control Conference (ECC)*, pages 2496–2501. IEEE, 2015.

Felix Berkenkamp, Riccardo Moriconi, Angela P Schoellig, and Andreas Krause. Safe learning of regions of attraction for uncertain, nonlinear systems with Gaussian processes. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4661–4666. IEEE, 2016.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-

based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30, 2017.

William K Bertram. Analytic Solutions for Optimal Statistical Arbitrage Trading. *Physica A: Statistical Mechanics and its Applications*, 389(11):2234–2243, 2010.

Rabi Bhattacharya and Chanho Lee. On Geometric Ergodicity of Nonlinear Autoregressive Models. *Statistics & Probability Letters*, 22(4):311–315, 1995.

George D Birkhoff. Proof of the Ergodic Theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660, 1931.

Arno Blaas, Jose Maria Manzano, Daniel Limon, and Jan Calliess. Localised kinky inference. In *2019 18th European Control Conference (ECC)*, pages 985–992. IEEE, 2019.

Farid Boussama, Florian Fuchs, and Robert Stelzer. Stationarity and Geometric Ergodicity of BEKK; Multivariate GARCH models. *Stochastic Processes and their Applications*, 121(10):2331–2360, 2011.

Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5:411–444, 2022.

Sébastien Bubeck, Gilles Stoltz, and Jia Yuan Yu. Lipschitz bandits without the Lipschitz constant. In *International Conference on Algorithmic Learning Theory*, pages 144–158. Springer, 2011.

Adam D Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.

David Burt, Carl Edward Rasmussen, and Mark Van Der Wilk. Rates of convergence for sparse variational Gaussian process regression. In *International Conference on Machine Learning*, pages 862–871. PMLR, 2019.

Jan-Peter Calliess. Lipschitz optimisation for Lipschitz interpolation. In *2017 American Control Conference (ACC)*, pages 3141–3146. IEEE, 2017.

Jan-Peter Calliess, Stephen J Roberts, Carl Edward Rasmussen, and Jan Maciejowski. Lazily adapted constant kinky inference for nonparametric regression and model-reference adaptive control. *Automatica*, 122:109216, 2020.

M Canale, L Fagiano, and MC Signorile. Nonlinear model predictive control from data: a set membership approach. *International Journal of Robust and Nonlinear Control*, 24 (1):123–139, 2014.

Massimo Canale, Lorenzo Fagiano, and Mario Milanese. Set membership approximation theory for fast implementation of model predictive control laws. *Automatica*, 45(1): 45–54, 2009.

Pierre Cardaliaguet and Guillaume Euvrard. Approximation of a Function and its Derivative with a Neural Network. *Neural Networks*, 5(2):207–220, 1992.

Ankush Chakrabarty, Devesh K Jha, and Yebin Wang. Data-driven control policies for partially known systems via kernelized Lipschitz learning. In *2019 American Control Conference (ACC)*, pages 4192–4197. IEEE, 2019.

Ankush Chakrabarty, Devesh K Jha, Gregery T Buzzard, Yebin Wang, and Kyriakos G Vamvoudakis. Safe approximate dynamic programming via kernelized Lipschitz estimation. *IEEE transactions on neural networks and learning systems*, 32(1):405–419, 2020.

Kung S Chan and Howell Tong. On the Use of the Deterministic Lyapunov Function for the Ergodicity of Stochastic Difference Equations. *Advances in applied probability*, 17 (3):666–678, 1985.

Cathy WS Chen, Feng-Chi Liu, and Mike KP So. A Review of Threshold Time Series Models in Finance. *Statistics and its Interface*, 4(2):167–181, 2011.

Guang-yong Chen, Min Gan, and Guo-long Chen. Generalized Exponential Autoregressive Models for Nonlinear Time Series: Stationarity, Estimation and Applications. *Information Sciences*, 438:46–57, 2018.

Xiaohong Chen and Timothy M Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, 188(2):447–465, 2015.

Chi-Young Choi and Young-Kyu Moh. How Useful are Tests for Unit-Root in Distinguishing Unit-Root Processes from Stationary but Non-Linear Processes? *The Econometrics Journal*, 10(1):82–112, 2007.

Girish Chowdhary, Hassan A Kingravi, Jonathan P How, and Patricio A Vela. A Bayesian nonparametric approach to adaptive control using Gaussian processes. In *52nd IEEE Conference on Decision and Control*, pages 874–879. IEEE, 2013.

Kim Christensen, Mathias Siggaard, and Bezirgen Veliyev. A Machine Learning Approach to Volatility Forecasting. *Journal of Financial Econometrics*, nbac020, 2022.

Daren BH Cline and Huay-min H Pu. Geometric Ergodicity of Nonlinear Time Series. *Statistica Sinica*, pages 1103–1118, 1999.

Danny D'Agostino. An efficient global optimization algorithm with adaptive estimates of the local Lipschitz constants. *arXiv preprint arXiv:2211.04129*, 2022.

Elvira Di Nardo et al. On the first passage time for autoregressive processes. *Scientiae Mathematicae Japonicae*, 67(2):137–152, 2008.

Matthew F Dixon, Igor Halperin, and Paul Bilokon. *Machine learning in finance*, volume 1170. Springer, 2020.

Ian Domowitz and Mahmoud A El-Gamal. A consistent nonparametric test of ergodicity for time series with applications. *Journal of Econometrics*, 102(2):365–398, 2001.

Holger Drees, Natalie Neumeyer, and Leonie Selk. Estimation and hypotheses testing in boundary regression models. *Bernoulli*, 25(1):424 – 463, 2019. 10.3150/17-BEJ992. URL https://doi.org/10.3150/17-BEJ992.

Christian L Dunis, Jason Laws, and Ben Evans. Trading Futures Spread Portfolios: Applications of Higher Order and Recurrent Networks. *The European Journal of Finance*, 14(6):503–521, 2008.

Christian L Dunis, Jason Laws, Peter W Middleton, and Andreas Karathanasopoulos. Trading and Hedging the Corn/Ethanol Crush Spread using Time-Varying Leverage and Nonlinear Models. *The European Journal of Finance*, 21(4):352–375, 2015.

Robert J Elliott, John Van Der Hoek*, and William P Malcolm. Pairs Trading. *Quantitative Finance*, 5(3):271–276, 2005.

Jürgen Elstrodt. *Maß-und Integrationstheorie*, volume 7. Springer, 1996.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. *arXiv preprint arXiv:1906.04893*, 2019.

K Ferentios. On Tcebycheff's type inequalities. *Trabajos de Estadistica y de Investigacion Operativa*, 33(1):125, 1982.

Jake M Ferguson and José M Ponciano. Predicting the Process of Extinction in Experimental Microcosms and Accounting for Interspecific Interactions in Single-Species Time Series. *Ecology letters*, 17(2):251–259, 2014.

Aiden J Fisher, David A Green, Andrew V Metcalfe, and Kunle Akande. First-Passage Time Criteria for the Operation of Reservoirs. *Journal of hydrology*, 519:1836–1847, 2014.

Jean-Pierre Fouque, George Papanicolaou, Ronnie Sircar, and Knut Sølna. *Multiscale stochastic volatility for equity, interest rate, and credit derivatives*. Cambridge University Press, 2011.

Marianne Frisén and Christian Sonesson. Optimal Surveillance based on Exponentially Weighted Moving Averages. *Sequential Analysis*, 25(4):379–403, 2006.

Evan Gatev, William N Goetzmann, and K Geert Rouwenhorst. Pairs Trading: Performance of a Relative-Value Arbitrage Rule. *The Review of Financial Studies*, 19(3): 797–827, 2006.

Hélyette Geman. Mean reversion versus random walk in oil and natural gas prices. *Advances in Mathematical finance*, pages 219–228, 2007.

Alan Genz and Frank Bretz. *Computation of Multivariate Normal and t Probabilities*, volume 195. Springer Science & Business Media, 2009.

Javier González, Zhenwen Dai, Philipp Hennig, and Neil Lawrence. Batch Bayesian optimization via local penalization. In *Artificial intelligence and statistics*, pages 648–657. PMLR, 2016.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

László Györfi, Michael Köhler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Thomas Hahn. Cuba, A Library for Multidimensional Numerical Integration. *Computer Physics Communications*, 168(2):78–95, 2005.

Peter Hall and Ingrid Van Keilegom. Nonparametric regression when errors are positioned at end-points. *Bernoulli*, 15(3):614–633, 2009.

Qiyang Han and Jon A. Wellner. Convergence rates of least squares regression estimators with heavy-tailed errors. *The Annals of Statistics*, 47(4):pp. 2286–2319, 2019. ISSN 00905364, 21688966. URL https://www.jstor.org/stable/26754231.

Bruce E Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748, 2008.

Boris Hasselblatt and Anatole Katok. *A first course in dynamics: with a panorama of recent developments*. Cambridge University Press, 2003.

Lukas Hewing, Juraj Kabzan, and Melanie N Zeilinger. Cautious model predictive control using Gaussian process regression. *IEEE Transactions on Control Systems Technology*, 28(6):2736–2743, 2019.

Lukas Hewing, Kim P Wabersich, Marcel Menner, and Melanie N Zeilinger. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:269–296, 2020.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Ming-Wei Hsu, Stefan Lessmann, Ming-Chien Sung, Tiejun Ma, and Johnnie EV Johnson. Bridging the Divide in Financial Market Forecasting: Machine Learners vs. Financial Economists. *Expert Systems with Applications*, 61:215–234, 2016.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In *Conference on Learning Theory*, pages 2007–2010. PMLR, 2020.

Julien Huang, Stephen Roberts, and Jan-Peter Calliess. On the sample complexity of Lipschitz learning algorithms. *Working Paper*, 2023.

Ildar Abdulovich Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical estimation: asymptotic theory*, volume 16. Springer Science & Business Media, 2013.

Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving Deep Learning Interpretability by Saliency Guided Training. *Advances in Neural Information Processing Systems*, 34, 2021.

Moritz Jirak, Alexander Meister, and Markus Reiss. Adaptive function estimation in nonparametric regression with one-sided errors. *The Annals of Statistics*, pages 1970–2002, 2014.

Donald R Jones, Cary D Perttunen, and Bruce E Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of optimization Theory and Applications*, 79(1): 157–181, 1993.

Matt Jordan and Alexandros G Dimakis. Exactly computing the local Lipschitz constant of relu networks. *Advances in Neural Information Processing Systems*, 33:7344–7353, 2020.

Mohammad Khajenejad, Zeyuan Jin, and Sze Zheng Yong. State and unknown terrain estimation for planetary rovers via interval observers. *Advanced Intelligent Systems*, page 2100040, 2021.

Igor L Kheifets and Pentti J Saikkonen. Stationarity and Ergodicity of Vector STAR Models. *Econometric Reviews*, 39(4):407–414, 2020.

Craig Knuth, Glen Chou, Necmiye Ozay, and Dmitry Berenson. Planning with learned dynamics: Probabilistic guarantees on safety and reachability via Lipschitz constants. *IEEE Robotics and Automation Letters*, 6(3):5129–5136, 2021.

Christopher Krauss. Statistical Arbitrage Pairs Trading Strategies: Review and Outlook. *Journal of Economic Surveys*, 31(2):513–545, 2017.

Dmitri E Kvasov and Yaroslav D Sergeyev. Lipschitz gradients for global optimization in a one-point-based partitioning scheme. *Journal of Computational and Applied Mathematics*, 236(16):4042–4054, 2012.

Petr Lánskỳ and Charles E Smith. The Effect of a Random Initial Value in Neural First-Passage-Time Models. *Mathematical biosciences*, 93(2):191–215, 1989.

Tim Leung and Xin Li. Optimal Mean Reversion Trading with Transaction Costs and Stop-Loss Exit. *International Journal of Theoretical and Applied Finance*, 18(03):1550020, 2015.

Eckhard Liebscher. Towards a Unified Approach for Proving Geometric Ergodicity and Mixing Properties of Nonlinear Autoregressive Processes. *Journal of Time Series Analysis*, 26(5):669–689, 2005.

S. Limanond and K.S. Tsakllis. Model reference adaptive and nonadaptive control of linear time-varying plants. *IEEE Transactions on Automatic Control*, 45(7):1290–1300, 2000. 10.1109/9.867022.

Daniel Limon, J Calliess, and Jan Marian Maciejowski. Learning-based nonlinear model predictive control. *IFAC-PapersOnLine*, 50(1):7769–7776, 2017.

Alexander Lipton and Vadim Kaushansky. On the First Hitting Time Density of an Ornstein-Uhlenbeck process. *arXiv preprint arXiv:1810.02390*, 2018.

Lennart Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1): 1–12, 2010.

Zudi Lu. On the Geometric Ergodicity of a Non-Linear Autoregressive Model with an Autoregressive Conditional Heteroscedastic Term. *Statistica Sinica*, pages 1205–1217, 1998.

Emilio T Maddalena and Colin N Jones. Learning non-parametric models with guarantees: A smooth Lipschitz regression approach. *IFAC-PapersOnLine*, 53(2):965–970, 2020a.

Emilio T Maddalena and Colin N Jones. Nsm converges to a k-nn regressor under loose Lipschitz estimates. *IEEE Control Systems Letters*, 4(4):880–885, 2020b.

Stefan Magureanu, Richard Combes, and Alexandre Proutiere. Lipschitz bandits: Regret lower bound and optimal algorithms. In *Conference on Learning Theory*, pages 975–999. PMLR, 2014.

Cédric Malherbe and Nicolas Vayatis. Global optimization of Lipschitz functions. In *International Conference on Machine Learning*, pages 2314–2323. PMLR, 2017.

José María Manzano, Daniel Limon, David Muñoz de la Peña, and Jan-Peter Calliess. Robust learning-based MPC for nonlinear constrained systems. *Automatica*, 117:108948, 2020.

José María Manzano, David Munoz de la Pena, Jan-Peter Calliess, and Daniel Limon. Componentwise hölder inference for robust learning-based mpc. *IEEE Transactions on Automatic Control*, 66(11):5577–5583, 2021.

Jose Maria Manzano, David Muñoz de la Peña, and Daniel Limon. Input-to-state stable predictive control based on continuous projected kinky inference. *International Journal of Robust and Nonlinear Control*, 2022.

RJ Martin, MJ Kearney, and RV Craster. Long-and Short-Time Asymptotics of the First-passage Time of the Ornstein–Uhlenbeck and other Mean-Reverting Processes. *Journal of Physics A: Mathematical and Theoretical*, 52(13):134001, 2019.

Elias Masry. Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–599, 1996.

Alexander Meister and Markus Reiß. Asymptotic equivalence for nonparametric regression with non-regular errors. *Probability Theory and Related Fields*, 155:201–229, 2013.

Ali Mesbah, Kim P Wabersich, Angela P Schoellig, Melanie N Zeilinger, Sergio Lucia, Thomas A Badgwell, and Joel A Paulson. Fusion of machine learning and MPC under uncertainty: What advances are on the horizon? In *2022 American Control Conference (ACC)*, pages 342–357. IEEE, 2022.

Sean P Meyn and Richard L Tweedie. *Markov Chains and Stochastic Stability*. Springer Science & Business Media, 2012.

Mario Milanese and Carlo Novara. Set membership identification of nonlinear systems. *Automatica*, 40(6):957–975, 2004.

Regina Hunter Mladineo. An algorithm for finding the global maximum of a multimodal, multivariate function. *Mathematical Programming*, 34(2):188–200, 1986.

M Mollineaux and R Rajagopal. Structural Health Monitoring of Progressive Damage. *Earthquake Engineering & Structural Dynamics*, 44(4):583–600, 2015.

Mogen M Monahemi and Miroslav Krstic. Control of wing rock motion using adaptive feedback linearization. *Journal of guidance, control, and dynamics*, 19(4):905–912, 1996.

Sandip Mukherji. Are Stock Returns Still Mean-Reverting? *Review of Financial Economics*, 20(1):22–27, 2011.

Ursula U Müller and Wolfgang Wefelmeyer. Estimation in nonparametric regression with non-regular errors. *Communications in Statistics-Theory and Methods*, 39(8-9):1619–1629, 2010.

Elisa Negrini, Giovanna Citti, and Luca Capogna. System identification through Lipschitz regularized deep neural networks. *Journal of Computational Physics*, 444:110549, 2021.

AS Nemirovskij, Boris Polyak, and AB Tsybakov. Rate of convergence of nonparametric estimates of maximum-likelihood type. *Problems of Information Transmission*, 21(4): 258–272, 1985.

Hae Young Noh, K Krishnan Nair, Anne S Kiremidjian, and CH Loh. Application of Time Series based Damage Detection Algorithms to the Benchmark Experiment at the

National Center for Research on Earthquake Engineering (NCREE) in Taipei, Taiwan. *Smart Structures and Systems*, 5(1):95–117, 2009.

Carlo Novara, Lorenzo Fagiano, and Mario Milanese. Direct feedback control design for nonlinear systems. *Automatica*, 49(4):849–860, 2013.

Alexander Novikov and Nino Kordzakhia. Martingales and First Passage Times of AR (1) Sequences. *Stochastics: An International Journal of Probability and Stochastics Processes*, 80(2-3):197–210, 2008.

Esa Nummelin. *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge University Press, 2004.

Patricia Pauli, Anne Koch, Julian Berberich, Paul Kohler, and Frank Allgöwer. Training robust neural networks using Lipschitz bounds. *IEEE Control Systems Letters*, 6:121–126, 2021.

Iosif F Pinelis and Aleksandr Ivanovich Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148, 1986.

Heni Puspaningrum, Yan-Xia Lin, and Chandra M Gulati. Finding the Optimal Pre-Set Boundaries for Pairs Trading Strategy based on Cointegration Technique. *Journal of Statistical Theory and Practice*, 4(3):391–419, 2010.

Robert M Sanner and Jean-Jacques E Slotine. Gaussian networks for direct adaptive control. In *1991 American control conference*, pages 2153–2159. IEEE, 1991.

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *arXiv preprint arXiv:1805.10965*, 2018.

Johaness Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Matthias W Seeger, Sham M Kakade, and Dean P Foster. Information consistency of nonparametric Gaussian process methods. *IEEE Transactions on Information Theory*, 54(5):2376–2382, 2008.

Leonie Selk, Charles Tillier, and Orlando Marigliano. Multivariate boundary regression models. *Scandinavian Journal of Statistics*, 49(1):400–426, 2022.

Yaroslav D Sergeyev. An information global optimization algorithm with local tuning. *SIAM Journal on Optimization*, 5(4):858–870, 1995.

Yaroslav D Sergeyev, Antonio Candelieri, Dmitri E Kvasov, and Riccardo Perego. Safe global optimization of expensive noisy black-box functions in the δ-Lipschitz framework. *Soft Computing*, 24(23):17715–17735, 2020.

Bruno O Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9(3):379–388, 1972.

Amit K Shukla, Manvendra Janmaijaya, Ajith Abraham, and Pranab K Muhuri. Engineering applications of artificial intelligence: A bibliometric analysis of 30 years (1988–2018). *Engineering Applications of Artificial Intelligence*, 85:517–532, 2019.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034*, 2013.

Ingo Steinwart, Don R Hush, Clint Scovel, et al. Optimal rates for regularized least squares regression. In *COLT*, pages 79–93, 2009.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

RG Strongin. On the convergence of an algorithm for finding a global extremum. *Eng. Cybernetics*, 11:549–555, 1973.

Roman Strongin, Konstantin Barkalov, and Semen Bevzuk. Acceleration of global search by implementing dual estimates for Lipschitz constant. In *International Conference on Numerical Computations: Theory and Algorithms*, pages 478–486. Springer, 2019.

Johannes Stübinger and Sylvia Endres. Pairs Trading with a Mean-Reverting Jump–Diffusion Model on High-Frequency Data. *Quantitative Finance*, 18(10):1735–1751, 2018.

Saowanit Sukparungsee and Alexander Novikov. On ewma procedure for detection of a change in observations via martingale approach. *Current Applied Science and Technology*, 6(2a):373–380, 2006.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Mark P Taylor, David A Peel, and Lucio Sarno. Nonlinear Mean-Reversion in Real Exchange Rates: Toward a Solution to the Purchasing Power Parity Puzzles. *International economic review*, 42(4):1015–1042, 2001.

Alexander J. Smola Thomas Hofmann, Bernhard Scholkopf. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

Dag Tjøstheim. Non-Linear Time Series and Markov Chains. *Advances in Applied Probability*, pages 587–611, 1990.

Adrian Trapletti, Friedrich Leisch, and Kurt Hornik. Stationary and integrated autoregressive neural network processes. *Neural Computation*, 12(10):2427–2450, 2000.

Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

RL Tweedie. Criteria for Classifying General Markov Chains. *Advances in Applied Probability*, pages 737–771, 1976.

Aad Van Der Vaart and Harry Van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12(6), 2011.

AW van der Vaart and JH van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

Ganapathy Vidyamurthy. *Pairs Trading: Quantitative Methods and Analysis*, volume 217. John Wiley & Sons, 2004.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

WenWu Wang, Ping Yu, Lu Lin, and Tiejun Tong. Robust Estimation of Derivatives using Locally Weighted Least Absolute Deviation Regression. *The Journal of Machine Learning Research*, 20(1):2157–2205, 2019.

Yining Wang, Sivaraman Balakrishnan, and Aarti Singh. Optimization of smooth functions with noisy observations: Local minimax rates. *Advances in Neural Information Processing Systems*, 31, 2018.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

J Wise. The Autocorrelation Function and the Spectral Density Function. *Biometrika*, 42 (1/2):151–159, 1955.

GR Wood and BP Zhang. Estimation of the Lipschitz constant of a function. *Journal of Global Optimization*, 8(1):91–103, 1996.

Yihong Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW (UIUC)*, 16, 2017.

George Wynne, François-Xavier Briol, and Mark Girolami. Convergence guarantees for Gaussian process means with misspecified likelihoods and smoothness. *The Journal of Machine Learning Research*, 22(1):5468–5507, 2021.

Yun Yang, Anirban Bhattacharya, and Debdeep Pati. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.

Zhengqin Zeng and Chi-Guhn Lee. Pairs Trading: Optimal Thresholds and Profitability. *Quantitative Finance*, 14(11):1881–1893, 2014.

Ruikun Zhou, Thanin Quartz, Hans De Sterck, and Jun Liu. Neural Lyapunov control of unknown nonlinear systems with stability guarantees. *Advances in Neural Information Processing Systems*, 35:29113–29125, 2022.