Contributions to Kernel Methods in Systems and Control

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Christian Martin Fiedler
aus Kronach, Deutschland

Berichter: Univ.-Prof. Dr. sc. Sebastian Trimpe
Univ.-Prof. Dr. rer. nat. Michael Matthias Herty

Tag der mündlichen Prüfung: 25. März 2025

Abstract

Machine learning is increasingly used in systems and control, which is motivated by increasingly challenging control, simulation and analysis problems, abundant data and computing resources, as well as impressive theoretical and methodological advances in machine learning. The established class of kernel methods is of particular interest in this context, due to their rich theory, efficient and reliable algorithms, and modularity, and indeed kernel methods are increasingly used in systems and control. This thesis contributes to this flourishing field, focusing on two exemplary and complementary topics.

First, many learning-based control approaches are based on combining uncertainty bounds for Gaussian process (GP) regression with robust control methods. We revisit the foundations of this domain by consolidating, improving, and carefully evaluating the required uncertainty bounds. As an application, we demonstrate how they can be combined with modern robust controller synthesis, leading to learning-enhanced robust control with rigorous control-theoretic and statistical guarantees. We furthermore discuss a severe practical limitation of these approaches, the a priori knowledge of an upper bound on the reproducing kernel Hilbert space (RKHS) norm of the target function, and propose to combine geometric assumptions together with kernel machines as a promising alternative.

Second, we initiate a new research direction by combining kernels with mean field limits as appearing in kinetic theory. Motivated by learning problems on large-scale multiagent systems, we introduce mean field limits of kernels, and provide an extensive theory for the resulting RKHSs. This is used in turn in the analysis of kernel-based statistical learning in the mean field limit, which not only is a novel form of large-scale limit in theoretical machine learning, but provides also a solid foundation for applications in kinetic theory. Finally, using the theory of reproducing kernels, we establish the first existence result for the mean field limit of very general discrete-time multiagent systems, and use this in mean field optimal control.

In summary, in this thesis we improve and refine existing uses of kernel methods in systems and control, helping to consolidate the area of learning-based control and pushing it further towards practical applications, and we introduce novel uses of kernels and their theory in systems and control, with many interesting directions for future work.

Kurzzusammenfassung

Maschinelles Lernen wird zunehmend in der System- und Regelungstechnik eingesetzt, was durch herausfordernde Kontroll-, Simulations- und Analyseprobleme, große Daten- und Rechenressourcen, sowie beeindruckende theoretische und methodische Fortschritte des maschinellen Lernens motiviert ist. In diesem Kontext ist die etablierte Klasse der Kernmethoden von besonderem Interesse, da diese über eine reichhaltige Theorie, effiziente Algorithmen sowie hohe Modularität verfügen, und in der Tat werden diese Methoden zunehmend in der System- und Regelungstechnik eingesetzt. Die vorliegende Dissertation trägt zu diesem sehr aktiven Feld bei, wobei der Fokus auf zwei exemplarischen und komplementären Themen liegt.

Viele lernbasierte Regelungsmethoden basieren auf der Kombination von Unsicherheitsschranken für Regression mit Gaußschen Prozessen (GP) mit Methoden der robusten Regelung. Wir betrachten die Grundlagen dieser Methoden, indem wir die benötigten Unsicherheitsschranken konsolideren, verbessern und ausführlich empirisch untersuchen. Als eine Anwendung kombinieren wir diese mit modernen Synthesemethoden der robusten Regelung, was zu lern-bereicherter robusten Regelung mit rigorosen kontroll-theoretischen und statistischen Garantien führt. Des Weiteren diskutieren wir eine schwerwiegende praktische Einschränkung dieser Ansätze, nämlich das Erfordernis einer a-priori Schranke für die Norm der Zielfunktion in einem reproduzierenden Kern-Hilberraum (RKHS), und wir schlagen als Alternative die Verwendung von geometrischen Vorwissen zusammen mit Kernmaschinen vor.

Zweitens starten wir eine neue Forschungsrichtung durch die Kombination von Kernen und dem Mean Field Limit, wie sie in der kinetischen Theorie vorkommen. Motiviert durch Lernprobleme auf großen Multiagentensystem führen das Mean Field Limit von Kernen ein, und studieren sehr ausführlich die entsprechenden RKHSs. Dies wiederum wird in der Analyse von kernbasierten Lernmethoden im Mean Field Limit genutzt, was nicht nur ein neues Limit-Konzept in der Theorie des maschinellen Lernens darstellt, sondern auch eine solide Grundlage für Anwendungen in der kinetischen Theorie liefert. Schließlich verwenden wir die Theorie reproduzierender Kerne, um das erste Existenzresult für das Mean Field Limit von sehr allgemeinen zeitdiskreten Multiagentensystemen zu zeigen, was anschließend in der Optimalsteuerung im Mean Field Limit angewandt wird.

Zusammenfassend verbessern und verfeinern wir existierende Anwendungen von Kernmethoden in der System- und Regelungstechnik, was zur Konsoliderung der lernbasierten Regelung beiträgt und einen weiteren Schritt Richtung praktischer Anwendungen erlaubt, und zudem führen wir eine neuartige Anwendungen von reproduzierenden Kernen und ihrer Theorie in diesem Kontext ein, mit vielfältigen Anknüpfungspunkten für weitergehende Arbeiten.

Preface

This thesis is based on work that the author has conducted mainly at the Institute for Data Science in Mechanical Engineering (DSME) at RWTH Aachen University (since January 2021), with some parts of the work done while the author was affiliated with the Max Planck Institute for Intelligent Systems Stuttgart and the Chair for Mathematical Systems Theory at the University of Stuttgart. The work was performed under the supervision of Univ.-Prof. Dr. sc. Sebastian Trimpe, and the author was additionally advised by Prof. Dr. Michael Herty (at RWTH Aachen University) and Prof. Dr. Carsten W. Scherer (at the University of Stuttgart).

Part of this work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016, and by the Cyber Valley Initiative. Support by the Stuttgart Center for Simulation Science (SimTech) is additionally acknowledged, as well as by the International Max Planck Research School for Intelligent Systems (IMPRS-IS). Part of this work was also funded by the Excellence Strategy of the Federal Government and the Länder (EXS-SF-SFDdM035), and by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) through the project EEMotion.

Parts of this thesis have appeared in articles published by IEEE, which requires the inclusion of the following disclaimer.[1]

Chapter 9 contains material published by AIMS in the journal Kinetic and Related Methods[2] and material published by Cambridge University Press in the European

[1] As required by IEEE, this note has been taken verbatim from `https://journals.ieeeauthorcenter.ieee.org/choose-a-publishing-agreement/avoid-infringement-upon-ieee-copyright/`, retrieved 07 April 2025.

[2] The material can be included in this thesis according to `https://www.aimsciences.org/index/list/FAQ`, retrieved 07 April 2025.

Journal of Applied Mathematics.[3]

Finally, I would like to add some personal acknowledgements. First of all, I would like to thank Univ.-Prof. Dr. sc. Sebastian Trimpe for accepting me as a PhD student, his constant support, advice and transparency, and the many opportunities and the great freedom that allowed me to pursue and develop my academic interests. I am also particularly grateful for the flexibility that was very helpful during difficult times.

I would also like to thank Prof. Dr. Michael Herty, not only for being the second examiner of this thesis, but also for his advice and the many interesting discussions, and reigniting my interest in multiagent systems and introducing me to kinetic theory.

Additionally, I would like to thank Prof. Dr. Carsten W. Scherer for his advice and the many discussions during the initial stages of this work, as well as for introducing me to modern robust control and sharing many of his insights into this field.

I would also like to thank Prof. Dr. sc. Bastian Leibe for agreeing to be an additional examiner, and Prof. Dr. Martin Grohe for chairing the examination committee.

Furthermore, I would like to thank Prof. Dr. Lorenzo Rosasco for hosting me at the University of Genoa and our ongoing (and fun) collaboration, as well as Prof. Dr. Raffaello Camoriano for his support and inviting me to Torino.

During the work on this thesis I had the chance for exchange with many researchers, and I would like to thank in particular Dr. Johannes Köhler (for insightful discussions on MPC, and with whom I co-supervised a master thesis) and Prof. Dr. Matthias A. Müller.

Furthermore, I benefited from the great teaching and support of many academics, and I would like to thank in particular Prof. Dr. Lars Grüne, Prof. Dr. Thomas Kriecherbauer, Prof. Dr. Christina Kuttler, Prof. Dr. Stephen Eglen, and Prof. Dr. Massimo Fornasier.

I would also like to thank the institute staff in Stuttgart and Aachen, especially Jutta Hess, Elisabeth Schaettgen, Claudia and Kurt Capellmann, and Hannah Franken.

A big thank you also to my current and former colleagues, especially (in some-

what chronological order) Dr. Friedrich Solowjow and Dr. Steve Heim (not only for many interesting discussions, but also for helping me out in Stuttgart and Aachen logistically), Dr. Andreas René Geist, Prof. Dr. Dominik Baumann, Alexander von Rohr (again not only for many discussions and chats, but also for helping me out in Aachen), Dr. Mona Fenet Ménou, Pierre-François Massiani (for many, many interesting and fun discussions and chats, as well as his constant support and, well, helping me out in Aachen), Katharina Ensinger, Dr. Andres Posada Moreno, my former officemate and collaborator Lukas Kreisköther, Paul Brunzema, Alexander Gräfe, Bernd Frauenknecht, Emma Cramer, Henrik Hose (especially for his tips for Aachen and Cologne, among many other things), Johanna Menn (who I collaborated with for some parts of this thesis), my great officemate Antonia Holzapfel, Noel Brindise, Antoine Moncho, Dr. David Stenger, and many more.

Finally, a very big and grateful thank you to my family and friends – there are too many things to thank them for to even attempt to list them, so I just say Danke, thank you, merci, $ευχαριστω$, gracias.

Contents

1. Introduction

We start by providing background and context, in particular, motivating the focus on kernel methods in the context of systems and control. We then describe on a high level the two main topics of this thesis, before giving a detailed outline of this work and its contributions.

1.1. Background and motivation

The field of systems and control – and related disciplines like dynamical systems theory, control engineering, process engineering, operations research – has recently experienced a surge in interaction with machine learning and related disciplines like statistics and data science. While insights and methods from systems and control have permeated into parts of machine learning, e.g., in the context of optimization algorithms [178, 143, 84], or structured state space models [8], the influence of learning for and in systems and control is arguably far greater. Of course, learning (and data) has played an important role in systems and control almost since the inception of this field, most notably in systems identification [121] and adaptive control [97]. However, with increasing availability of data and compute, and considerable progress in machine learning on both theoretical and methodological levels, learning and data have become a central focus of research in systems and control, often subsumed under the term *learning-based control*. This term might also encompass related activities, like data-driven control [132], distribution-free approaches like the scenario approach [46], methods based on the Koopman approach [133], and even reinforcement learning [193].

While the use of machine learning in systems and control appears to be very promising, it is often highly non-trivial due to specific challenges arising from the nature of dynamical systems and the requirements of control engineering. From a theoretical perspective, data generated by dynamical systems have inherent de-

pendencies which can lead to statistical complications, and despite recent progress, even the linear case still poses open problems [202]. Similarly, learning a dynamical system might require targeted exploration [40, 129] in contrast to passive sampling. Furthermore, in many applications stringent guarantees are required of a control method, including stability, constraint satisfaction, or even performance guarantees [15, 87], often under external disturbances and imprecise knowledge of the system to be controlled. Finally, since systems and control is in many cases concerned with *physical* systems, considerations of sample efficiency, safety, and use of prior knowledge become relevant.

The present thesis contributes to the flourishing field of machine learning for systems and control. In light of the challenges outlined above, we focus on a particular class of learning methods – *kernel methods*. Roughly speaking, this is a class of learning algorithms that rely on a mathematical object known as a (reproducing) kernel, which is associated with a specific function space, called a reproducing kernel Hilbert space (RKHS). Kernel methods are a very established class of learning methods, with a very mature theory, efficient algorithms, and a large variety of models and learning approaches available [181, 189]. Furthermore, they have strong connections to other fields, including scattered data approximation [217] and statistical methods for stochastic processes [32], which increases their applicability, and allows for profiting from developments in these related fields. Kernel methods are particularly attractive for use in systems and control due to the following characteristics:

1. They have a very well-developed theory that is often easy to use in downstream applications, which is in contrast to other popular approaches like *Deep Learning*[1].

2. There exist systematic ways to include prior knowledge, cf. also Chapter 4.

3. Kernel methods tend to be sample efficient and reliable, e.g., by avoiding nonconvex optimization problems [166, 181].

Finally, kernel methods are already popular and successful in the context of systems

[1] There has been tremendous progress in the understanding of deep learning in recent years, but the theory is still not comparable to what has been achieved for kernel methods. Ironically, a lot of progress in the theory of deep learning has been achieved by reduction to kernel methods and their theory, e.g., [98].

and control, both on a theoretical (e.g, [129]) and practical level, e.g., in system identification [159].

1.2. Goals of this thesis

The discussion in the preceding sections motivates the following question, which will form the starting point for this thesis:

What can kernel methods do for systems and control?

While we start with this rather broad question, the present thesis focuses on two representative and complementary topics, which demonstrate the power of kernel methods and their theory in the context of systems and control. In the following, we will give a high-level overview of those two topics and state our goals. For conciseness, we will provide a discussion of the relevant literature in the respective upcoming chapters.

Uncertainty sets and learning-based control Modern control methods often rely on various types of models, usually of the system to be controlled, but also of external signals or even uncertainties in the system behaviour. In general, a better model leads to better control performance. Since for many real-world systems first-principles modelling becomes challenging, it is therefore tempting to use machine learning to improve the models (or even generate them in the first place), and this is indeed a strong focus of learning-based control. However, as already touched upon above, the use of learning in the context of control comes with unique challenges, and we would like to elaborate on two of these. First, since control is a very mature field, with a deep theory and considerable experience by its practitioners, it is highly advisable to avoid ab initio learning, and instead try to include prior knowledge. For many applications, rigorous control-theoretic guarantees are required, and these can indeed be provided for many modern control methodologies. Retaining such guarantees even when a learning component is involved is therefore a second interesting challenge. As suggested by the preceding introductory remarks, this suggests the use of kernel methods, and indeed they are very popular in the context of learning-based control, especially Gaussian process (GP) regression, for which uncertainty bounds are available. However, control applications put very specific

and demanding requirements on these uncertainty sets. For this reason, we *revisit uncertainty bound for kernel methods in the context of control*. We will investigate rigorous, yet practical uncertainty bounds for kernel methods suitable for control applications, with a particular emphasis on reasonable assumptions and their practical ramifications, and exemplary control applications. The overarching goal is to achieve learning-enhanced control with rigorous guarantees that are meaningful in practice.

Kernels and the mean field limit The field of kinetic theory is concerned with the modelling, analysis, simulation, and control of large-scale system consisting of interacting components, with gas dynamics as a prime example. A very important tool in this area is the *mean field limit*, which is one way to go from a microscopic perspecive (considering individual, discrete entities) to a mesoscopic level (working with the distribution of entitites). Curiously, the use of kernel methods in this area appears to be almost completely unexplored. We therefore start to pursue this direction with two completely novel applications of kernels. First, we investigate the mean field limit of kernels and their RKHSs, as well as associated statistical learning problems. This is motivated by certain learning problems in the context of interacting particle systems, but it is also interesting from a purely theoretical perspective. Second, we use theoretical tools from kernel methods in the context of discrete-time multiagent systems. More precisely, we use kernel mean embeddings (KMEs) to established an existence result for the mean field limit of such systems, and to the best of our knowledge this is also the first such result. We therefore demonstrate that kernel methods and their theory are a promising avenue in the context of kinetic theory.

Finally, we would like to stress that the nature of the two exemplary topics of this thesis are complementary. While the first one is an established and active topic, and our focus is on carefully revisiting its foundations and their practical ramifications, our second topic forms a novel and innovative domain, with many interesting open questions and promising avenues for future work.

1.3. Contributions and outline

We now outline the remainder of this thesis, motivating our approaches and pointing out our contributions. The thesis is structured in three parts. In Part I, we provide technical background on kernels and reproducing kernel Hilbert spaces, since these form the theoretical foundation for much of following developments. Parts II and III correspond to the two concrete areas this thesis contributes to, kernel methods with uncertainty sets and their use in learning-based control, and kernels in the context of mean field limits. Since the present thesis touches upon several research fields and requires tools from a range of disciplines, each part contains an introductory chapter, and we have placed the discussion of related work and relevant background literature in the corresponding chapters. This improves the reading flow, and makes the thesis accessible to a broader audience.

1.3.1. Part I: Foundations

The primary technical tool for the remainder of the thesis are kernels and their associated RKHSs. We therefore provide ample background on this topic after this introductory chapter.

While there are many good introductions to RKHSs (see the next chapter for some pointers to the literature), we identified an unfortunate gap in the literature: most introductions choose a particular perspective, which leads to some concepts and results appearing unnatural, unless viewed from a different, more appropriate perspective. This hinders a quick comprehension by beginners of this field and slows down a deeper understanding. In Chapter 2 we therefore present a gentle introduction to kernels and RKHSs, choosing for each concept and result the most natural perspective to present it, contributing a novel and complementary exposition to the kernel literature.

In later chapters, geometric properties and regularity of RKHS functions will play a role, in particular, Lipschitz continuity thereof. Unfortunately, to the best of our knowledge, there is no systematic exposition and investigation of Lipschitz (or more generally, Hölder) continuity of RKHS functions, so we devote Chapter 3 to this topic. In this way, we provide the first comprehensive survey of Lipschitz and Hölder continuity in RKHSs, which collects and refines many existing results, and presents some new ones.

1.3.2. Part II: Uncertainty bounds and learning-based control

In Part II, we commence with our first main objective of this thesis: investigating and improving uncertainty bounds for kernel methods in the context of learning-based control, and applications therein. Since for learning-based control regression is most important learning setup, we restrict us to such problems in this part of the thesis.

In the introductory Chapter 4, we start with an exposition of uncertainty sets in learning-based control. This will be discussed in very general framework, which appears to go beyond the existing literature on learning-based control. We provide also background on GP regression and kernel ridge regression, as these are among the most common kernel methods in learning-based control.

The investigation of uncertainty bounds for these kernel methods starts in Chapter 5. We describe some simple uncertainty bounds based on concentration inequalities for quadratic forms, and we review the state-of-the-art bounds based on self-normalization. By interpreting the kernel methods as instances of regularized least-squares in Hilbert spaces, we can give a particularly transparent presentation. In addition, we provide an elementary derivation of the self-normalization results that actually explains how these results arise. We also describe our results providing uncertainty bounds that are robust to model misspecifications, which are the first results of this kind.

In Chapter 6, we carefully evaluate the uncertainty bounds using numerical experiments. Based on both the theoretical and practical insights, we then apply them to a learning-based control application, for which we focus on learning-enhanced robust controller synthesis with statistical and control-theoretic guarantees. To the best of our knowledge, this is the first application of kernel methods with frequentist uncertainty bounds in the context of modern robust controller synthesis that provides rigorous statistical and control-theoretical guarantees.

Unfortunately, the uncertainty bounds have a subtle, but severe practical issue – the need for an a priori bound on the RKHS norm of the target function. In Chapter 7 we argue that this forms a severe obstacle to their applicability in control, and actually prohibits their use in many relevant safety-critical scenarios. To overcome this issue, we propose to instead use geometric assumptions on the target function, which can be connected to established prior knowledge. We implement and evaluate

this strategy in the context of hard shape constraint kernel machines, and evaluate it using numerical experiments.

Summarizing, uncertainty bounds for kernel methods in the context of learning-based control is a rather active subject with considerable activity in the last decade. On a high level, our contribution to this area lies in carefully and critically revisiting its foundation, in particular, the assumptions used and the way the bounds are used, and proposing ways to make the theory more practical.

1.3.3. Part III: Mean field limits and kernels

In Part III of the thesis, we turn to the subject of kernels and mean field limits. We start with some background on the mean field limit in Chapter 8, and provide motivation for the study of the mean field limit of kernels.

In Chapter 9, we then investigate in detail the mean field limit of kernels and their RKHSs. This entails an existence result of the mean field limit of kernels as well as large classes of kernels that allow such a limit. Furthermore, we provide an essentially complete characterization of the involved RKHSs under the mean field limit. This seems to be a completely new area of study, to which we contribute also a rather complete theory.

In Chapter 10, we then turn to investigating statistical learning with kernels in the mean field. We start with some results on the approximation capabilities of kernels in this setting, and formulate an appropriate variant of the representer theorem, which is the main ingredient that makes learning with kernels numerically feasible. After providing a concise introduction of the standard framework of statistical learning theory, we then present convergence results for learning problems involving kernels in the mean field limit. To the best of our knowledge, the setting of this chapter is also novel, and we contribute substantial theoretical results, relying also on novel techniques like applying Γ-convergence arguments to kernels and RKHSs.

Finally, in Chapter 11 we turn to discrete-time multiagent systems. Using kernel mean embeddings, we provide the first existence result for the mean field limit of a very general class of such systems. As an application, we consider corresponding optimal control problems, and prove that the relaxed dynamical programming principle, as used in the analysis of nonlinear model predictive control without terminal constraints, also holds in the mean field limit.

We conclude the thesis in Chapter 12 with a summary and an outline of some interesting directions for future work. In the appendix we have included a chapter on statistical learning theory for kernel-based methods on distributional inputs, presenting new oracle inequalities and stability-based generalization bounds. While not directly related to learning for systems and control, many of the techniques appearing throughout this thesis will be used there, and as in Chapters 9 and 10, kernels on probability distributions will play an important role.

1.4. Publications

The present thesis is based on several publications, preprints and manuscripts. At the beginning of every chapter, we will state from which publication or manuscript material has been used for the respective chapter. Furthermore, we end each chapter (apart from this introductory Chapter 1 and the concluding Chapter 12) with a section describing the relation to existing work, and additional details on the present author's contributions.

For the reader's convenience, we now provide an overview of works involving the author that are part of this thesis, or are related in some way to it. Unless mentioned otherwise, the present author's position in the author list reflects his contribution. In particular, the first author is usually the main author in terms of scientific contributions and writing. If the ordering of the authors is alphabetical, or the first authorship is shared, this will be explicitly pointed out in the remainder of this section.

Regarding Part I, Chapter 2 is based on an earlier version of

> **Christian Fiedler** and Sebastian Trimpe. *A panoramic introduction to reproducing kernel Hilbert spaces.* Manuscript in preparation, 2024.

and Chapter 3 is taken mostly verbatim from the preprint

> **Christian Fiedler**. *Lipschitz and Hölder continuity in reproducing kernel Hilbert spaces.* arXiv preprint, 2023.

In Part II, Chapters 4, 5, and 6 build heavily on the following publications, with some parts taken verbatim,

Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. *Practical and rigorous uncertainty bounds for Gaussian process regression.* AAAI Conference on Artificial Intelligence (AAAI), 2021. Updated version with corrections as preprint arXiv:2105.02796v2

Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. *Learning-enhanced robust controller synthesis with rigorous statistical and control-theoretic guarantees.* 60th IEEE Conference on Decision and Control (CDC), 2021.© 2021 IEEE.

Most of Chapter 7 is from the articles

Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. *Learning functions and uncertainty sets using geometrically constrained kernel regression.* 61st IEEE Conference on Decision and Control (CDC), 2022. © 2022 IEEE.

Christian Fiedler[2], Johanna Menn, Lukas Kreisköther, and Sebastian Trimpe. *On safety in safe Bayesian optimization.* Transactions on Machine Learning Research (TMLR), 2024.

In addition, the findings from this last work have been disseminated in the following extended abstract, which has been presented as a poster at the corresponding symposium,

Christian Fiedler[3], Johanna Menn, and Sebastian Trimpe. Safety in safe Bayesian optimization and its ramifications for control. Extended abstract, Symposium on Systems Theory in Data and Optimization, 2024. Available as preprint arXiv:2501.13697

Most of Part III has appeared in the following articles

Christian Fiedler[4], Michael Herty, Michael Rom, Chiara Segala, and Sebastian Trimpe. *Reproducing kernel Hilbert spaces in the mean field limit.* Kinetic and Related Models, 2023. Published by American Institute of Mathematical Sciences.

[2]Joint first authorship with J. Menn. The two first authors are ordered alphabetically.
[3]Joint first authorship with J. Menn.
[4]First and main author, the remaining authors are ordered alphabetically.

> **Christian Fiedler**, Michael Herty, and Sebastian Trimpe. *On kernel-based statistical learning theory in the mean field limit.* Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS), 2023.
>
> **Christian Fiedler**[5], Michael Herty, Chiara Segala, and Sebastian Trimpe. *Recent kernel methods for interacting particle systems: first numerical results.* European Journal of Applied Mathematics, 2024.
>
> **Christian Fiedler**, Michael Herty, and Sebastian Trimpe. *Mean field limits for discrete-time dynamical systems via kernel mean embeddings.* IEEE Control Systems Letters, 2023. © 2023 IEEE.

The last article was also accepted at the American Control Conference (ACC) 2024 and presented there, and the author was also invited to present a poster on this article at the 4th Symposium on Machine Learning and Dynamical Systems at the Fields Institute, Toronto, Canada. Furthermore, the findings from [CF3, CF6, CF5] were disseminated in the following extended abstract, which was selected for an oral presentation at the corresponding workshop,

> **Christian Fiedler**, Sebastian Trimpe, and Michael Herty. *Reproducing kernels in and for the mean field limit.* International Workshop on Deep Learning and Kernel Machines (DEEPK), 2024.

In an appendix, we have also included the following article,

> **Christian Fiedler**, Pierre-François Massiani, Friedrich Solowjow, and Sebastian Trimpe. *Towards statistical learning theory with distributional inputs.* International Conference on Machine Learning (ICML), 2024.

The topic of this work is not related to the main goals of this thesis, but many of the techniques used in the former play also an important role in the latter, so we have decided to include it.

Finally, several works of the author appeared during the work on this thesis, but are not included. In particular, the following two articles contribute to the field of nonlinear discrete-time control,

[5] Alphabetical ordering, main authorship shared with C. Segala.

Christian Fiedler and Sebastian Trimpe. *Revisiting the derivation of stage costs in infinite horizon discrete-time optimal control.* 30th Mediterranean Conference on Control and Automation (MED), 2022.

Christian Fiedler and Sebastian Trimpe. *Analysis of EMPC schemes without terminal constraints via local incremental stabilizability.* European Control Conference (ECC), 2024.

The following article is based on work from the author's master thesis, and the article was finalized during the preparation of this thesis,

Christian Fiedler[6], Massimo Fornasier, Timo Klock, and Michael Rauchensteiner. *Stable recovery of entangled weights: Towards robust identification of deep neural networks from minimal samples.* Applied and Computational Harmonic Analysis, 2023.

The author was co-supervisor of the master thesis of A. Tokmak, which resulted in the following preprint, which has been accepted for publication in IEEE Transactions on Automatic Control,

Abdullah Tokmak, **Christian Fiedler**, Melanie N. Zeilinger, Sebastian Trimpe, and Johannes Köhler. *Automatic nonlinear mpc approximation with closed-loop guarantees.* arXiv preprint, 2023.

Finally, the author participated in the research leading to the following preprint, a revision of which is about to be submitted to the Transactions on Machine Learning Research,

Friedrich Solowjow, Dominik Baumann, **Christian Fiedler**, Andreas Jocham, Thomas Seel, and Sebastian Trimpe. *A Kernel Two-sample Test for Dynamical Systems.* arXiv preprint, 2022.

[6] Alphabetical ordering.

Part I.

Foundations

2. Introduction to reproducing kernel Hilbert spaces

Reproducing Kernel Hilbert Spaces (RKHSs) play a central role in many areas of mathematical sciences and engineering, from machine learning [181, 182, 189], statistics and probability theory [32], numerical approximation methods [217] and numerical methods for partial differential equations [71], to mathematical physics [7] and pure mathematics [152]. In particular, RKHSs form the theoretical foundation for most kernel methods, and therefore later parts of the thesis will heavily rely upon these function spaces. In the following, we provide a self-contained introduction to RKHSs and related concepts like reproducing kernels. This serves two purposes: On the one hand, this provides necessary background for the remainder of this thesis. On the other hand, we hope to close a gap in the literature. We observed that many excellent introductions to RKHSs like [189, Chapter 4], [32], or [218], choose a particular perspective, usually motivated by the respective application domain. However, a concept can appear unnatural or difficult to comprehend when introduced in one context, but it can become very natural or intuitive when viewed from a different perspective. For example, the definition of a kernel in terms of feature space-feature map pairs might appear somewhat arbitrary in the context of RKHSs, but becomes very natural in the context of the kernel trick as used in support vector machines (SVMs), cf. Section 2.3. Similarly, the motivation of native spaces might be unclear in the context of SVMs, but becomes obvious when viewed from the perspective of scattered data approximation, as explained in Section 2.5. What appears to be missing is a *perspective-agnostic introduction* to the RKHS framework that presents each of the core concept in its most natural setting, and thereby allowing an easier understanding as well as a comprehension of the bigger picture by the learner. With the present introduction, we hope to close this gap in the literature.

Since specialization to the real case, which is the relevant setting for this thesis,

leads to no significant simplification (apart from saving some conjugate signs in the notation), we cover both the case $\mathbb{K} = \mathbb{R}$ and $\mathbb{C}$ simultaneously. The contribution of this chapter lies in the presentation of the material, while all technical results and examples are well-known. In case no reference is provided, it is because the corresponding result is folklore.

This chapter is based on, and in large parts taken verbatim from, an early version of [CF14]. Detailed comments on the author's contributions are provided in Section 2.7

2.1. Reproducing Kernel Hilbert Spaces: Continuous function evaluation

We start by introducing the concept of a reproducing kernel Hilbert space. The idea is based on [32, Section 1.1 (Introduction)], though our presentation is considerably more detailed. Let $\mathcal{X} \neq \emptyset$ be an arbitrary nonempty set. A common task in the mathematical sciences is to add additional structure to such a set. Among the most convenient and richest structures are $\mathbb{K}$-Hilbert spaces, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. We would like to have a general recipe for adding such a $\mathbb{K}$-Hilbert space structure to an *arbitrary* set $\mathcal{X}$. For this, we need the following.

1. A $\mathbb{K}$-vectorspace H that is derived from $\mathcal{X}$
2. Reasonable assumptions on an inner product $\langle \cdot, \cdot \rangle$ on H
3. A way to embed elements of $\mathcal{X}$ into H

We will now give a general recipe that naturally motivates the definition of an RKHS, as will be formalized in Definition 2.1.1.

First, we need a $\mathbb{K}$-vectorspace. Since $\mathcal{X}$ is arbitrary, the most basic choice is to use a subspace of $V^{\mathcal{X}}$, the space of functions from $\mathcal{X}$ into V, for an arbitrary $\mathbb{K}$-vectorspace V (with the usual pointwise addition and scalar multiplication). The simplest choice of V is of course $V = \mathbb{K}$, so we continue with a vector subspace $H \subseteq \mathbb{K}^{\mathcal{X}}$. We leave the concrete choice of H open for now.

Next, we turn $H \subseteq \mathbb{K}^{\mathcal{X}}$ into a $\mathbb{K}$-Hilbert space by adding a scalar product $\langle \cdot, \cdot \rangle$ such that the induced metric space is complete. What are reasonable requirements

on $\langle \cdot, \cdot \rangle$? Of course, $(H, \langle \cdot, \cdot \rangle)$ should be connected to $\mathcal{X}$ in some meaningful way, otherwise the whole procedure is pointless. Since we want to work with arbitrary sets $\mathcal{X}$, the only connection we can make between H and $\mathcal{X}$ without imposing additional assumptions on $\mathcal{X}$ is the *evaluation of functions* $f \in H$. Furthermore, it is very likely that we want to do analysis in H, and the following situation frequently appears in analytical constructions and proofs.

In order to construct a function with certain properties, one starts with a sequence of functions $(f_n)_n \subseteq H$ having said properties (and maybe approximating something). If $(f_n)_n$ is a Cauchy sequence in H with respect to (w.r.t.) the metric induced by $\langle \cdot, \cdot \rangle$, then this sequence converges to a unique $f \in H$ (since H is a Hilbert space and hence complete), usually inheriting the properties of interest from the Cauchy sequence. This is a very flexible approach to construct functions in H with prescribed properties.

We would like to have this technique available in our Hilbert space $(H, \langle \cdot, \cdot \rangle)$. Since the only connection from H back to $\mathcal{X}$ is via function evaluation, and since in this technique we use a limit to construct a function f, we arrive at the following requirement: For all $x \in \mathcal{X}$, we need $f_n(x) \to f(x)$. Since we want a generic method, we require this for all sequences $(f_n)_n \subseteq H$ with $f_n \xrightarrow{\|\cdot\|_H} f \in H$ (this is equivalent to requiring it for all Cauchy sequences). But since in Hilbert spaces continuity is equivalent to sequential continuity, this is equivalent to the following: For all $x \in \mathcal{X}$, the evaluation functionals $H \ni f \mapsto f(x) \in \mathbb{K}$ are continuous, i.e., they are in the (topological) dual space of H. This leads to the following central definition.

Definition 2.1.1. Let $\mathcal{X} \neq \emptyset$ be an arbitrary set and $(H, \langle \cdot, \cdot \rangle_H)$ a function $\mathbb{K}$-Hilbert space, i.e., $H \subseteq \mathbb{K}^{\mathcal{X}}$. We call H a *reproducing kernel Hilbert space (RKHS)* if for all $x \in \mathcal{X}$ the corresponding evaluation functionals[1] $\delta_x \colon H \to \mathbb{K}$, $\delta_x(f) = f(x)$, are continuous, i.e., for all $x \in \mathcal{X}$ we have $\delta_x \in L(H, \mathbb{K}) = H'$.

This choice of terminology will be clarified in the next section. Finally, the question arises how to actually embed $\mathcal{X}$ into H. In other words, for each $x \in \mathcal{X}$, we need an element $f_x \in H$. Since we want a universal recipe, we should not impose any additional assumptions. But it turns out that with the developments so far, there is a universal solution to this task. Since we are in a Hilbert space setting,

[1] These are also called Dirac functionals.

by the Riesz Representation Theorem, we have a bijective correspondence between continuous linear functionals on H and elements from H,

$$L \in H' \longleftrightarrow f = IL \in H,$$

where $I\colon H' \to H$ is the map assigning to each continuous functional on $L \in H'$ its unique Riesz representer $IL \in H$, so that $Lh = \langle h, IL\rangle_H$ for all $h \in H$. Since in an RKHS, for all $x \in \mathcal{X}$, we have $\delta_x \in H'$, we can embed $\mathcal{X}$ into H via

$$\mathcal{X} \ni x \mapsto f_x = I\delta_x \in H$$

In Section 2.2, we will make this embedding more concrete, and in Section 2.3, we will put it into the bigger context of the RKHS framework. Before presenting concrete examples of RKHSs, the following (well-known) example demonstrates that it is not trivial to fulfill the definition of an RKHS.

Example 2.1.2. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, and consider the Lebesgue space $L^2(\Omega, \mathcal{A}, \mu)$. Recall that this is the Hilbert space consisting of μ-almost everywhere equivalence classes of measurable and square-integrable (w.r.t. μ) $\mathbb{K}$-valued functions on Ω. For example, if $\mathcal{B}([0,1])$ is the Borel σ-algebra on $[0,1]$ and λ the Lebesgue measure on $[0,1]$, then $L^2([0,1], \mathcal{B}([0,1]), \lambda)$ is the space of Borel-measurable functions $f : [0,1] \to \mathbb{K}$, that are square-integrable (so $\int_{[0,1]} |f(x)|^2 \mathrm{d}\lambda(x) < \infty$), with inner product $\langle f, g\rangle_{L^2} = \int_{[0,1]} f(x)\overline{g(x)}\mathrm{d}\lambda(x)$, and identifying functions that are almost everywhere equal.

These Lebesgue spaces $L^2(\Omega, \mathcal{A}, \mu)$ are some of the most important examples of Hilbert spaces, but *they are not RKHSs*. The reason is simple: By definition, an RKHS is a Hilbert space *of functions*, but these Lebesgue spaces consist of *equivalence classes of functions*. In particular, evaluation of elements of $L^2(\Omega, \mathcal{A}, \mu)$ is not defined at individual inputs $x \in \Omega$ (it is only defined μ-almost everywhere).

For more discussion of this classic non-example, we refer to [152, Section 1.2.2].

It is time for concrete examples of Hilbert spaces. Our first example is classical and appears already in the seminal work [14], with our presentation loosely inspired by [152, Section 1.2.1].

Example 2.1.3. Let $\mathcal{X} \neq \emptyset$ be a set, let $H \subseteq \mathbb{K}^{\mathcal{X}}$ be a finite-dimensional Hilbert space of functions on $\mathcal{X}$, and choose some orthonormal basis (ONB) $e_1, \ldots, e_M$ of

H. Let $x \in \mathcal{X}$ be arbitrary. Observe that for all $g \in H$ we have

$$g(x) = \left(\sum_{m=1}^{M} \langle g, b_M \rangle_H b_m \right)(x) = \sum_{m=1}^{M} \langle g, b_M \rangle_H b_m(x), \tag{2.1}$$

where we first used that the $e_1, \ldots, e_M$ form an ONB, and then the definition of addition and multiplication in function spaces. Now, consider $f, f_n \in H$, $n \in \mathbb{N}_+$, with $\lim_{n\to\infty} f_n = f$. We then have

$$\begin{aligned} \lim_{n\to\infty} f_n(x) &= \lim_{n\to\infty} \sum_{m=1}^{M} \langle f_n, b_m \rangle_H b_m(x) = \sum_{m=1}^{M} \lim_{n\to\infty} \langle f_n, b_m \rangle_H b_m(x) \\ &= \sum_{m=1}^{M} \left\langle \lim_{n\to\infty} f_n, b_m \right\rangle_H b_m(x) = \sum_{m=1}^{M} \langle f, b_m \rangle_H b_m(x) = f(x), \end{aligned}$$

where we used in the first equality our observation (2.1), then the finiteness of the sum, in the third equality the continuity of the scalar product, and in the last equality again our observation (2.1). Altogether, we found that all evaluation functionals are continuous in H.

Since H was arbitrary, we arrive at the following result: *All finite-dimensional Hilbert spaces of functions are RKHSs.*

The next class of RKHSs is an important example from complex analysis. Our presentation follows roughly [152, Section 1.4.1].

Example 2.1.4. Denote by $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$ the open unit disc of complex numbers, and define $\ell_2(\mathbb{N}_0, \mathbb{C}) = \{(a_n)_{n\in\mathbb{N}_0} \in \mathbb{C}^{\mathbb{N}_0} \mid \sum_{n=0}^{\infty} |a_n|^2 < \infty\}$, the set of square-summable complex sequences. Recall that the latter becomes a Hilbert space when introducing the inner product $\langle (a_n)_n, (b_n)_n \rangle_{\ell_2} = \sum_{n=0}^{\infty} a_n \overline{b_n}$.

We consider functions on $\mathbb{D}$ that are represented by power series, i.e., $f(z) = \sum_{n=0}^{\infty} a_n z^n$, such that $(a_n)_{n\in\mathbb{N}_0} \in \ell_2(\mathbb{N}_0, \mathbb{C})$. Denote the set of all of these functions by $H^2(\mathbb{D})$. It is a complex vector space, and

$$\langle f, g \rangle_{H^2} = \sum_{n=0}^{\infty} a_n \overline{b_n},$$

where $f, g \in H^2(\mathbb{D})$ with representations $f(z) = \sum_{n=0}^{\infty} a_n z^n$, $g(z) = \sum_{n=0}^{\infty} b_n z^n$, defines an inner product on $H^2(\mathbb{D})$. Observe that $I_{H^2} : H^2(\mathbb{D}) \to \ell^2(\mathbb{N}_0, \mathbb{C})$,

$\sum_{n=0}^{\infty} a_n z^n \mapsto (a_n)_{n\in\mathbb{N}_0}$ is a well-defined isometric isomorphism, which implies that $H^2(\mathbb{D})$ is a Hilbert space, which is called a *Hardy space*. In particular, for $f \in H^2(\mathbb{D})$ with $f(z) = \sum_{n=0}^{\infty} a_n z^n$, we have $\|f\|_{H^2} = \|(a_n)_n\|_{\ell_2}$.

It turns out that $H^2(\mathbb{D})$ is even an RKHS on $\mathbb{D}$. To verify this, we have to show that all evaluation functionals are continuous. Let $z \in \mathbb{D}$ be arbitrary. For $f \in H^2(\mathbb{D})$ with representation $f(z) = \sum_{n=0}^{\infty} a_n z^n$, we have

$$
\begin{aligned}
|\delta_z(f)| = |f(z)| &= \left|\sum_{n=0}^{\infty} a_n z^n\right| \\
&\le \sum_{n=0}^{\infty} |a_n z^n| && \textit{Triangle inequality} \\
&\le \sqrt{\sum_{n=0}^{\infty} |a_n|^2}\sqrt{\sum_{n=0}^{\infty} |z|^n} && \textit{Cauchy-Schwarz inequality (in } \ell_2(\mathbb{N}_0,\mathbb{C})\textit{)} \\
&= \frac{1}{1-|z|}\|(a_n)_n\|_{\ell_2} && \textit{Geometric series, } |z|<1 \\
&= \frac{1}{1-|z|}\|f\|_{H^2},
\end{aligned}
$$

which shows that δ_z is continuous.

2.2. Reproducing Kernels: Evaluation by scalar products

Our next goal is to introduce the concept of a reproducing kernel. The presentation in this section (and all the results) appears to be folklore, and similar expositions can be found in [152, Chapter 2] and [189, Section 4.2]. Let $\mathcal{X} \neq \emptyset$ be an arbitrary nonempty set and $H \subseteq \mathbb{K}^{\mathcal{X}}$ a $\mathbb{K}$-Hilbert function space. We have two fundamental operations in this context: *Evaluation* of a function $f \in H$ at an input $x \in \mathcal{X}$, i.e., $f \mapsto f(x)$, and the *inner product* of two functions $f_1, f_2 \in H$, i.e., $\langle f_1, f_2\rangle_H$. We now want to connect these two concepts. One motivation is that working with inner products can be very convenient, both in terms of theory as well as computation.

Here is one way to achieve this connection. Assume that there exists a family of functions in H, $(f_x)_{x\in\mathcal{X}}$, such that for all $f \in H$ and $x \in \mathcal{X}$ we have $f(x) = \langle f, f_x\rangle_H$. In other words, we can replace function evaluation by inner products. How does such a family of functions look like? They always have the following symmetry property.

Lemma 2.2.1. Let $(f_x)_{x\in\mathcal{X}}$, $f_x \in H$, be such that such that for all $f \in H$ and $x \in \mathcal{X}$, we have $f(x) = \langle f, f_x\rangle_H$. Then for all $x, x' \in \mathcal{X}$, we have $f_x(x') = \overline{f_{x'}(x)}$.

Note that for $\mathbb{K} = \mathbb{R}$ we even have $f_x(x') = f_{x'}(x)$ for all $x, x' \in \mathcal{X}$.

Proof. Let $x, x' \in \mathcal{X}$ be arbitrary, then we have $f_x(x') = \langle f_x, f_{x'}\rangle_H = \overline{\langle f_{x'}, f_x\rangle_H} = \overline{f_{x'}(x)}$. □

This result indicates that the index x of f_x, and an input x' on which f_x is evaluated, are somehow on equal footing. To make this notationally clear, we define $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by $k(x, x') = f_{x'}(x)$. This leads us to the following concept.

Definition 2.2.2. Let $\mathcal{X} \neq \emptyset$ be a set and $H \subseteq \mathbb{K}^{\mathcal{X}}$ a $\mathbb{K}$-Hilbert space of functions. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel (RK)* of H if

1. For all $x \in \mathcal{X}$, $k(\cdot, x) \in H$
2. For all $f \in H$ and $x \in \mathcal{X}$, $f(x) = \langle f, k(\cdot, x)\rangle_H$

The second property in the preceding definition is usually called the *reproducing property*, since the scalar product with the reproducing kernel *reproduces* the value of a function from H. Furthermore, note that the proof of Lemma 2.2.1 shows that if $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ is a reproducing kernel, then for all $x, x' \in \mathcal{X}$ we have $k(x, x') = \overline{k(x', x)}$. Finally, Definition 2.2.2 is also meaningful if H is just a $\mathbb{K}$-pre Hilbert space of functions.

Example 2.2.3. Let $\mathcal{X} \neq \emptyset$ be some set. We can interpret a function $f : \mathcal{X} \to \mathbb{K}$ as a (scalar) *signal*. For example, if $\mathcal{X} = \mathbb{N}_0$, f could be a discrete-time measurement of some quantity, if $\mathcal{X} = \mathbb{R}$, f could be a physical quantity that changes over time, and if $\mathcal{X} = \{0, \dots, 255\} \times \{0, \dots, 255\}$, f could be a (gray-scale) square image of size 256x256.

An important task in *signal processing* is *sampling* of a signal. Essentially, this is the task of reconstructing a signal from a certain set of measurements (the samples) of the signal. Here is an idealized formalization of this task. Let $H \subseteq \mathbb{K}^{\mathcal{X}}$ be a space of signals, then we would like to find a sequence of inputs $x_n \in \mathcal{X}$ and functions $b_n \in H$, $n \in \mathbb{N}_+$, such that

$$f(x) = \sum_{n=1}^{\infty} f(x_n) b_n(x)$$

for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$. Such a representation of f is called a *sampling expansion*. If $\mathcal{H}$ is a Hilbert space with reproducing kernel k, we can rewrite this as

$$f(x) = \sum_{n=1}^{\infty} \langle f, k(\cdot, x_n) \rangle_H b_n(x).$$

This means that the sampling problem becomes amenable to Hilbert space methods. Since sampling methods do not play a role in the remainder of the thesis, we do not present detailed results in this direction, but rather refer to the excellent introductory article [77].

Let us now have a closer look at reproducing kernels. First, if a reproducing kernel of a Hilbert space of functions exists, it is unique.

Lemma 2.2.4. A Hilbert space of functions $H \subseteq \mathbb{K}^{\mathcal{X}}$ can have at most one reproducing kernel.

Proof. Let $k, \tilde{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ be two reproducing kernels for H. Let now $x, x' \in \mathcal{X}$ be arbitrary, then we have

$$k(x, x') = \langle k(\cdot, x'), \tilde{k}(\cdot, x) \rangle_H = \overline{\langle \tilde{k}(\cdot, x), k(\cdot, x') \rangle_H} = \overline{\tilde{k}(x', x)} = \tilde{k}(x, x'),$$

where we applied the reproducing property of $\tilde{k}(\cdot, x)$ to $k(\cdot, x') \in H$ in the first equality, then we applied the reproducing property of $k(\cdot, x')$ to $\tilde{k}(\cdot, x)$ in the third inequality, and finally we used the symmetry of a reproducing kernel. □

The next example is also classical, and our presentation is based on [152, Section 1.3.2].

Example 2.2.5. Let $T > 0$ and consider $L^2([-T, T]) = L^2([-T, T], \mathcal{B}([-T, T]), \lambda)$, the space of (almost-everywhere equivalence classes of) square-integrable functions on $[-T, T]$, which is a Hilbert space. For $f \in L^2([-T, T])$, we denote its Fourier transform by

$$\hat{f}(\omega) = \int_{-T}^{T} f(t) e^{-2\pi i \omega t} \mathrm{d}t.$$

Define now

$$PW_T = \{\hat{f} \mid f \in L^2([-T, T])\}$$

and observe that PW_T is a space of functions, and not just equivalence classes of functions. For each $F \in PW_T$, there exists a unique $f \in L^2([-T,T])$ with $F = \hat{f}$, so we can define $\|F\|_{PW_T} = \|f\|_{L^2([-T,T])}$. Since $L^2([-T,T])$ is a Hilbert space, this turns also PW_T into a Hilbert space, which is called a *Paley-Wiener space*. These spaces play an important role in signal processing and complex analysis.

It turns out that PW_T has a reproducing kernel k. To find it, let $F \in PW_T$ and $x \in \mathbb{R}$ be arbitrary, and let us try to identify a function $k(\cdot, x) \in PW_T$ such that $F(x) = \langle F, k(\cdot, x)\rangle_{PW_T}$. By definition of PW_T, there exists $f \in L^2([-T,T])$ with $F = \hat{f}$. We therefore get

$$\begin{aligned} F(x) &= \int_{-T}^{T} f(t) e^{-2\pi i x t} \mathrm{d}t && F = \hat{f} \\ &= \langle f, e^{2\pi i x \cdot}\rangle_{L^2([-T,T])} && \textit{Definition } \langle\cdot,\cdot\rangle_{L^2([-T,T])} \\ &= \langle f, g\rangle_{L^2([-T,T])} && \textit{Definition } g \\ &= \langle F, \hat{g}\rangle_{PW_T} && \textit{Definition of } PW_T \end{aligned}$$

where we defined $g : [-T,T] \to \mathbb{C}$, $g(t) = e^{2\pi i x t}$. Since F was arbitrary and there is only one reproducing kernel, we find that $k(\cdot, x) = \hat{g}$, hence

$$k(x, x') = \int_{-T}^{T} e^{2\pi i x' t} e^{2\pi i x t} \mathrm{d}t = \begin{cases} \frac{1}{\pi} \frac{\sin(2\pi T(x - x'))}{x - x'} & \text{if } x \neq x' \\ 2T & \text{otherwise} \end{cases}$$

Next, a function $f : \mathcal{X} \to \mathbb{K}$ is completely described by its values $f(x)$, $x \in \mathcal{X}$, i.e., by function evaluations. If H has a reproducing kernel, all function evaluations are described by this reproducing kernel. Intuitively, the reproducing kernel should then completely describe H. The next result confirms this intuition.

Proposition 2.2.6. If H has a reproducing kernel k, then the set $\{k(\cdot, x) \mid x \in \mathcal{X}\}$ is total in H, i.e., $\operatorname{span}\{k(\cdot, x) \mid x \in \mathcal{X}\}$ is dense in H.

Proof. Define $H_0 = \{k(\cdot, x) \mid x \in \mathcal{X}\}$ and let $f \in H_0^\perp$, i.e., for all $x \in \mathcal{X}$ we have $\langle f, k(\cdot, x)\rangle_H = 0$. For an arbitrary $x \in \mathcal{X}$ this then leads to

$$f(x) = \langle f, k(\cdot, x)\rangle_H = 0,$$

i.e., $f = 0_H$, which implies that H_0 is total in H. □

Actually, a reproducing kernel completely determines the corresponding Hilbert space of functions in the following sense.

Proposition 2.2.7. Let $(H_i, \langle\cdot,\cdot\rangle_i)$, $i = 1, 2$, be two $\mathbb{K}$-Hilbert spaces of functions on $\mathbb{K}^{\mathcal{X}}$, i.e., $H_i \subseteq \mathbb{K}^{\mathcal{X}}$, having reproducing kernels $k_i : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$. If $k_1 \equiv k_2$ (i.e., $k_1(x, x') = k_2(x, x')$ for all $x, x' \in \mathcal{X}$, so they are equal as maps), then $H_1 = H_2$ as sets and $\langle\cdot,\cdot\rangle_1 \equiv \langle\cdot,\cdot\rangle_2$.

In other words, a map $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ can be a reproducing kernel of at most one Hilbert function space. This result complements Lemma 2.2.4, which states that a Hilbert function space can have at most one reproducing kernel.

Proof. We show $H_1 \subseteq H_2$, the reverse inclusion then follows by symmetry. By definition of a reproducing kernel, $k_1(\cdot, x) \in H_1$ for all $x \in \mathcal{X}$, and since H_1 is a vector space, we also have $H_0 = \operatorname{span}\{k_1(\cdot, x) \mid x \in \mathcal{X}\}$. But since $k_1 \equiv k_2$, we then have for all $x \in \mathcal{X}$ that $k_1(\cdot, x) = k_2(\cdot, x) \in H_2$, and since also H_2 is a vector space, we find that $H_0 = \operatorname{span}\{k_2(\cdot, x) \mid x \in \mathcal{X}\} \subseteq H_2$.

Next, let $f = \sum_{i=1}^{N} \alpha_i k_1(\cdot, x_i)$ and $g = \sum_{j=1}^{M} \beta_j k_1(\cdot, y_j)$ be two functions from H_0, hence $f, g \in H_1$ and $f, g \in H_2$, then

$$\begin{aligned}\langle f, g\rangle_1 &= \left\langle \sum_{i=1}^{N} \alpha_i k_1(\cdot, x_i), \sum_{j=1}^{M} \beta_j k_1(\cdot, y_j) \right\rangle_1 = \sum_{i=1}^{N}\sum_{j=1}^{M} \alpha_i \bar{\beta}_j k_1(y_j, x_i) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{M} \alpha_i \bar{\beta}_j k_2(y_j, x_i) = \left\langle \sum_{i=1}^{N} \alpha_i k_2(\cdot, x_i), \sum_{j=1}^{M} \beta_j k_2(\cdot, y_j) \right\rangle_2 = \langle f, g\rangle_2.\end{aligned}$$

This shows that $\langle\cdot,\cdot\rangle_1|_{H_0} \equiv \langle\cdot,\cdot\rangle_2|_{H_0}$.

Let now $f \in H_1$ be arbitrary. According to Proposition 2.2.6, H_0 is dense in H_1, so there exists a sequence $(f_n)_n \subseteq H_0$ such that $f_n \xrightarrow{\|\cdot\|_1} f$. Since $(f_n)_n$ is convergent, it is a Cauchy sequence w.r.t. $\|\cdot\|_1$. Recall that $H_0 \subseteq H_2$, so we have $f_n \in H_2$ for all $n \in \mathbb{N}$, and since $\langle\cdot,\cdot\rangle_1|_{H_0} \equiv \langle\cdot,\cdot\rangle_2|_{H_0}$, $(f_n)_n$ is also a Cauchy sequence in H_2. Because H_2 is complete, there exists $g \in H_2$ such that $f_n \xrightarrow{\|\cdot\|_2} g$.

We now show that $f = g$. For this, let $x \in \mathcal{X}$ be arbitrary, then

$$\begin{aligned}
f(x) &= \langle f, k_1(\cdot, x)\rangle_1 && \textit{Reproducing property (in } H_1\textit{)} \\
&= \langle \lim_n f_n, k_1(\cdot, x)\rangle_1 && f_n \to f \textit{ in } H_1 \\
&= \lim_n \langle f_n, k_1(\cdot, x)\rangle_1 && \textit{Continuity of } \langle\cdot,\cdot\rangle_1 \\
&= \lim_n f_n(x) && \textit{Reproducing property in } H_1 \\
&= \lim_n \langle f_n, k_2(\cdot, x)\rangle_2 && \textit{Reproducing property in } H_2 \\
&= \langle \lim_n f_n, k_2(\cdot, x)\rangle_2 && \textit{Continuity of } \langle\cdot,\cdot\rangle_2 \\
&= \langle g, k_2(\cdot, x)\rangle_2 && f_n \to g \textit{ in } H_2 \\
&= g(x).
\end{aligned}$$

Altogether, we found that $H_1 \subseteq H_2$.

Finally, since $\|\cdot\|_1 = \|\cdot\|_2$ on the dense set H_0, we get equality of the norms and hence the scalar products. □

Finally, the existence of a reproducing kernel implies continuity of evaluation.

Lemma 2.2.8. If H has a reproducing kernel, then each evaluation functional δ_x is continuous.

Proof. Let $f \in H$ and $x \in \mathcal{X}$ be arbitrary, then

$$|\delta_x(f)| = |f(x)| = \langle f, k(\cdot, x)\rangle_H| \leq \|k(\cdot, x)\|_H \|f\|_H,$$

i.e., $\|\delta_x\|_{H'} \leq \|k(\cdot, x)\|_H$, hence δ_x is continuous. □

Remark 2.2.9. Inspecting the proofs of Lemma 2.2.1, 2.2.4 and 2.2.8 reveals that these results also hold if H is a pre Hilbert space.

We can now connect RKHSs with reproducing kernels: Let H be a $\mathbb{K}$-Hilbert space. If it has a reproducing kernel, then according to Lemma 2.2.8, it is an RKHS as formalized in Definition 2.1.1. Recalling the developments in Section 2.1, we even have a converse. Let H be an RKHS and let I be the Riesz representer map. Define for $x \in \mathcal{X}$ the function $k(\cdot, x) = I\delta_x \in H$, then by definition, k is a

reproducing kernel for H. Summarizing, we have the following result that explains the terminology RKHS.

Theorem 2.2.10. Let $\mathcal{X} \neq \emptyset$ be a nonempty set and let $H \subseteq \mathbb{K}^{\mathcal{X}}$ be a Hilbert space of functions. H is an RKHS if and only if it has a reproducing kernel.

Example 2.2.11. Consider the situation of Example 2.1.4, where we introduced the Hardy space $H^2(\mathbb{D})$. We know already that it is an RKHS, so it has a reproducing kernel $k : \mathbb{D} \times \mathbb{D} \to \mathbb{C}$ according to Theorem 2.2.10, which is unique according to Lemma 2.2.4. Let us try to find this reproducing kernel k.

Let $f \in H^2(\mathbb{D})$ with representation $f(z) = \sum_{n=0}^{\infty} a_n z^n$, and $z_2 \in \mathbb{D}$. Since $k(\cdot, z_2) \in H^2(\mathbb{D})$, there exists $(b_n)_n \in \ell_2(\mathbb{N}_0, \mathbb{C})$ with $k(z_1, z_2) = \sum_{n=0}^{\infty} b_n z_1^n$ for all $z_1 \in \mathbb{D}$. Furthermore, we have

$$\begin{aligned} f(z_2) &= \sum_{n=0}^{\infty} a_n z_2^n && \textit{Series representation of } f \\ &= \langle f, k(\cdot, z_2) \rangle_{H^2(\mathbb{D})} && \textit{Reproducing property of } k \\ &= \sum_{n=0}^{\infty} a_n \overline{b_n}, && \textit{Definition } \langle \cdot, \cdot \rangle_{H^2(\mathbb{D})} \end{aligned}$$

so equating coefficients we find that $b_n = \overline{z_2}^n$ for all $n \in \mathbb{N}_0$. This means that for all $z_1, z_2 \in \mathbb{D}$ we have

$$k(z_1, z_2) = \sum_{n=0}^{\infty} z_1^n \overline{z_2}^n = \frac{1}{1 - z_1 \overline{z_2}},$$

and it is the unique reproducing kernel k of $H^2(\mathbb{D})$. It is called the *Szegö kernel*.

The preceding example illustrates a typical technique for finding the reproducing kernel of an RKHS: Solve the equation $f(x) = \langle f, k(\cdot, x) \rangle_k$ for all $f \in H_k$ and $x \in \mathcal{X}$, to get $k(\cdot, x) \in H_k$. The resulting function k is then the unique reproducing kernel.

2.3. Kernels: Feature spaces and the kernel trick

The perspective in this section is well-known in machine learning, see for example [189] and [181] for classic textbook accounts. In computational sciences, especially in machine learning, tasks can often be made easier by *lifting* them to some new spaces. For example, in machine learning and statistics, $\mathcal{X}$ might be the space from which

the (input) samples of a data set come from. To such samples $x \in \mathcal{X}$, we can then assign features $\Phi(x) \in \mathcal{H}$ in some new space $\mathcal{H}$ by using a given map $\Phi : \mathcal{X} \to \mathcal{H}$. Quite frequently, $\mathcal{H}$ is a Hilbert space, so one can apply linear algorithms on the lifted data points $\Phi(x)$. In this manner, one can easily "non-linearize" a linear method, making it more powerful.

Example 2.3.1. One important task in machine learning is *binary classification*. Given a description $x \in \mathcal{X}$ of an object (typically, $\mathcal{X} \subseteq \mathbb{R}^d$, where $d \in \mathbb{N}_+$ can be very large), the goal is to assign the object to one of two classes c_1 and c_2. Formally, we want a map $c : \mathcal{X} \to \{c_1, c_2\}$. For example, x could be an image of a part at the end of a manufacturing process, and based on this image an algorithm should decide whether the part is acceptable (say, class c_1) or faulty (class c_2).

A particularly simple situation is *linear separability*[2], where $\mathcal{X}$ is a vector space, and there exists some linear function $f : \mathcal{X} \to \mathbb{R}$ and some constant $b \in \mathbb{R}$ such that for an object with description $x \in \mathcal{X}$,

$$c(x) = \begin{cases} c_1 & \text{if } f(x) \geq b \\ c_2 & \text{otherwise} \end{cases}$$

is the correct class of this object. The subset $\{x \in \mathcal{X} \mid f(x) = b\}$ is then called the *decision boundary*, and corresponds in this case to an affine-linear space. In most practical scenarios, linear separation will not be possible. Put differently, a linear decision boundary is not complex enough to separate the two classes.

However, while frequently linear separation in $\mathcal{X}$ is not possible, it is in another space $\mathcal{H}$. Here, $\mathcal{X} = \mathbb{R}^2$, and the two classes cannot be separated by a straight line (left panel). However, by transforming the input $x = (x_1, x_2)$ into $(x_1, x_2, \sqrt{x_1^2 + x_2^2})$, linear separation is easily possible, though now in $\mathbb{R}^3$ (middle panel). Formally, let $\mathcal{H} = \mathbb{R}^3$, define $\Phi : \mathcal{X} \to \mathcal{H}$ by $\Phi(x) = (x_1, x_2, \sqrt{x_1^2 + x_2^2})$, and let $h : \mathcal{H} \to \mathbb{R}$, $b \in \mathbb{R}$ such that $h(\Phi(x)) \geq b$ if the object with description x belongs to class c_1, and $h(\Phi(x)) < b$ otherwise. Note that the resulting classifier

$$c(x) = \begin{cases} c_1 & \text{if } h(\Phi(x)) \geq b \\ c_2 & \text{otherwise} \end{cases}$$

[2] Affine-linear separation might be slightly more accurate.

is now *nonlinear* in the orignal input space $\mathcal{X}$, i.e., $\{x \in \mathcal{X} \mid h(\Phi(x)) = b\}$ is not an affine-linear subspace anymore.

In particular, if an algorithm interacts with its inputs only via scalar products, then such an algorithm can be automatically "non-linearized" by replacing an input x by $\Phi(x)$, and the original scalar product by the one in $\mathcal{H}$. Actually, in this situation, we do not even need to work with objects in $\mathcal{H}$ directly, only with their scalar products. The next classic example illustrates this.

Example 2.3.2. In *linear regression*, one models a linear function $f : \mathbb{R}^d \to \mathbb{R}$ by the ansatz $h(\cdot, w) : \mathbb{R}^d \to \mathbb{R}$, $h(x, w) = x^\top w = \langle x, w \rangle_{\mathbb{R}^d}$, where $w \in \mathbb{R}^d$ are the *weights* parametrizing the modeling function $h(\cdot, w)$. Let $((x_1, y_1), \ldots, (x_N, y_N))$ be a data set, where we interpret $x_n \in \mathbb{R}^d$ as an input, and $y_n \in \mathbb{R}$ as the corresponding (usually noisy) output of the unknown function f. In *ridge regression* we determine weights from this data set by solving the optimization problem

$$\min_{w \in \mathbb{R}^d} \sum_{n=1}^{N} (y_n - h(x_n, w))^2 + \lambda \|w\|^2, \tag{2.2}$$

where $\lambda \in \mathbb{R}_{>0}$ is the *regularization parameter*. Including the term $\lambda\|w\|^2$ in the optimization problem encourages small weights, and the strength of this penalization is controlled by λ. The optimization problem has a unique solution given explicitly by

$$w_\lambda^* = (X^\top X + \lambda I_d)^{-1} X^\top y, \tag{2.3}$$

where we defined

$$X = \begin{pmatrix} x_1 & \cdots & x_N \end{pmatrix}^\top, \quad y = \begin{pmatrix} y_1 & \cdots & y_N \end{pmatrix}^\top. \tag{2.4}$$

We can use this linear method to model also nonlinear functions. Consider a map $\Phi : \mathbb{R}^d \to \mathbb{R}^M$, where $M \in \mathbb{N}_+$ might be rather large, and define $h(\cdot, w) : \mathbb{R}^d \to \mathbb{R}$ by $h(x, w) = \Phi(x)^\top w = \langle \Phi(x), w \rangle_{\mathbb{R}^M}$. The optimization problem

$$\min_{w \in \mathbb{R}^M} \sum_{n=1}^{N} (y_n - h(x_n, w))^2 + \lambda \|w\|^2, \tag{2.5}$$

has still a unique solution, now given by

$$\tilde{w}_\lambda^* = (\tilde{X}^\top \tilde{X} + \lambda I_M)^{-1} \tilde{X}^\top y, \tag{2.6}$$

where we defined

$$\tilde{X} = \begin{pmatrix} \Phi(x_1)^\top \\ \vdots \\ \Phi(x_N)^\top \end{pmatrix}. \tag{2.7}$$

Let us have a look at the output of this linear method, the function

$$h(x, \tilde{w}_\lambda^*) = \langle \Phi(x), \tilde{w}_\lambda^* \rangle_{\mathbb{R}^M} = \Phi(x)^\top (\tilde{X}^\top \tilde{X} + \lambda I_M)^{-1} \tilde{X}^\top y. \tag{2.8}$$

Using the push-through identity, we have $(\tilde{X}^\top \tilde{X} + \lambda I_M)^{-1} \tilde{X}^\top = \tilde{X}^\top (\tilde{X}\tilde{X}^\top + \lambda I_N)^{-1}$, so we get

$$h(x, \tilde{w}_\lambda^*) = \Phi(x)^\top \tilde{X}^\top (\tilde{X}\tilde{X}^\top + \lambda I_N)^{-1} y. \tag{2.9}$$

Inspecting the subexpressions $\Phi(x)^\top \tilde{X}^\top$ and $\tilde{X}\tilde{X}^\top$,

$$\Phi(x)^\top \tilde{X}^\top = \Phi(x)^\top \begin{pmatrix} \Phi(x_1) & \cdots & \Phi(x_N) \end{pmatrix} = \begin{pmatrix} \Phi(x)^\top \Phi(x_1) & \cdots & \Phi(x)^\top \Phi(x_N) \end{pmatrix}$$

$$\tilde{X}\tilde{X}^\top = \begin{pmatrix} \Phi(x_1)^\top \\ \vdots \\ \Phi(x_N)^\top \end{pmatrix} \begin{pmatrix} \Phi(x_1) & \cdots & \Phi(x_N) \end{pmatrix} = \begin{pmatrix} \Phi(x_1)^\top \Phi(x_1) & \cdots & \Phi(x_1)^\top \Phi(x_N) \\ \vdots & \ldots & \vdots \\ \Phi(x_N)^\top \Phi(x_1) & \cdots & \Phi(x_N)^\top \Phi(x_N) \end{pmatrix}$$

we see that $h(x, \tilde{w}_\lambda^*)$ does not contain any element from $\mathbb{R}^M$ in isolation, but only in scalar products. To make this more explicit, define $k(x, x') = \langle \Phi(x'), \Phi(x) \rangle_{\mathbb{R}^M}$, then we can rewrite the ridge regression solution as

$$h(x, \tilde{w}_\lambda^*) = (k(x, x_n))_{n=1,\ldots,N}^\top \left((k(x_i, x_j))_{i,j=1,\ldots,N} + \lambda I_N \right)^\top y. \tag{2.10}$$

In order to compute it, we only have to be able to evaluate k.

The technique described here, *replacing all expression of the form* $\langle x, x' \rangle_{\mathbb{R}^d}$ *by* $\langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}$ *for some Hilbert space* $\mathcal{H}$ *and some* $\mathcal{H}$*-valued map* Φ, is called the *kernel trick*. Note that the new space $\mathcal{H}$ does not explicitly appear at all, so it can be implicitly defined, or be even infinite-dimensional. All we need is a way to evaluate for given x, x' the expression $\langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}$ in some efficient manner. The

next classic example (from [181, Chapter 2]) illustrates this.

Example 2.3.3. Let $\mathcal{X} = \mathbb{R}^d$, $d \in \mathbb{N}_+$, and consider $\Phi(x) = \left(x_{i_1} \cdot \ldots \cdot x_{i_m}\right)_{1 \leq i_1, \cdots, i_m \leq d}$, $1 \leq m \leq d$. This means that $\mathcal{H}$ is the set of all ordered mth monomials of vectors $x \in \mathbb{R}^d$, and we assign the usual scalar product to $\mathcal{H}$. It is clear that the dimension of $\mathcal{H}$ rapidly increases with d and m. As a consequence, computing $\langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}$ by first computing $\Phi(x), \Phi(x')$, and subsequently computing the scalar product becomes infeasible even for moderate d and m. Let us have a closer look at this scalar product,

$$\begin{aligned}
\langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}} &= \sum_{1 \leq i_1, \ldots, i_m \leq d} (x'_{i_1} \cdot \ldots \cdot x'_{i_m}) \cdot (x_{i_1} \cdot \ldots \cdot x_{i_m}) \\
&= \sum_{i_1=1}^{d} \cdots \sum_{i_m=1}^{d} (x'_{i_1} \cdot \ldots \cdot x'_{i_m}) \cdot (x_{i_1} \cdot \ldots \cdot x_{i_m}) \\
&= \left(\sum_{i_1=1}^{d} x'_{i_1} \cdot x_{i_1} \right) \cdot \ldots \cdot \left(\sum_{i_m=1}^{d} x'_{i_m} \cdot x_{i_m} \right) \\
&= \left(\sum_{i=1}^{d} x_i \cdot x'_i \right)^m .
\end{aligned}$$

We find that $\langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}} = k(x, x')$, where we defined $k(x, x') = \left(\sum_{i=1}^{d} x_i \cdot x'_i \right)^m$. Note that $k(x, x')$ can be very efficiently computed ($(d-1)$ additions and $(m-1) \cdot d$ multiplications), and that neither $\mathcal{H}$ nor Φ appear explicitly in the definition of k.

The preceding considerations lead immediately to the next concept.

Definition 2.3.4. Let $\mathcal{X} \neq \emptyset$ be a set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ some function. We say that k is a ($\mathbb{K}$-)kernel (on $\mathcal{X}$) if there exist a $\mathbb{K}$-Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that

$$\forall x, x' \in \mathcal{X} : k(x, x') = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}. \tag{2.11}$$

In the situation of Definition 2.3.4, $\mathcal{H}$ is called a *feature space* and Φ a *feature map* of k. Note that in general $\mathcal{H}$ and Φ are not unique. Furthermore, in the case $\mathbb{K} = \mathbb{R}$, we have that (2.11) is equivalent to $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$, which is the form often found in the machine learning literature.

The following result provides a first connection between Definition 2.3.4 and the concepts from the previous section.

Proposition 2.3.5. Let $H \subseteq \mathbb{K}^{\mathcal{X}}$ be an RKHS with reproducing kernel k. Then k is a kernel in the sense of Definition 2.3.4 with feature space H and feature map $\Phi_k : \mathcal{X} \to H$, $\Phi_k(x) = k(\cdot, x)$, called the *canonical feature map* of k.

Proof. By the definition of a reproducing kernel, for all $x \in \mathcal{X}$ we have $\Phi_k(x) = k(\cdot, x) \in H$, so Φ_k is well-defined, and by the reproducing property of k we get $k(x, x') = \langle k(\cdot, x'), k(\cdot, x)\rangle_H = \langle \Phi_k(x'), \Phi_k(x)\rangle_H$ for all $x, x' \in \mathcal{X}$. □

Ensuring that Proposition 2.3.5 holds is one reason why in Definition 2.3.4 we have $k(x, x') = \langle \Phi(x'), \Phi(x)\rangle_{\mathcal{H}}$ (instead of $k(x, x') = \langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}}$).

Example 2.3.6. Consider again Example 2.1.4. For all $z_1, z_2 \in \mathbb{D}$, we have

$$k(z_1, z_2) = \frac{1}{1 - z_1\overline{z_2}} = \sum_{n=0}^{\infty} z_1^n \overline{z_2^n} = \sum_{n=0}^{\infty} \overline{z_2^n}\,\overline{\overline{z_1^n}} = \langle (\overline{z_2^n})_{n\in\mathbb{N}_0}, (\overline{z_1^n})_{n\in\mathbb{N}_0}\rangle_{\ell_2},$$

so $\ell_2(\mathbb{N}_0, \mathbb{C})$ is a feature space, and $\Phi : \mathbb{D} \to \ell_2(\mathbb{N}_0, \mathbb{C})$, $\Phi(z) = (\overline{z^n})_{n\in\mathbb{N}_0}$ is a feature map for the Szegö kernel k.

We saw that every reproducing kernel (for some Hilbert space of functions) is a kernel. But what about the converse? Given a kernel k, does there exist a Hilbert space of functions such that k is the reproducing kernel for it? This is indeed the case, as described in Theorem 2.3.7. The approach originated probably with [173], and our presentation is based on [189, Theorem 4.21].

Theorem 2.3.7. Let $\mathcal{X} \neq \emptyset$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ be a kernel with feature space $\mathcal{H}$ and feature map $\Phi : \mathcal{X} \to \mathcal{H}$. Define

$$H = \{x \mapsto \langle h, \Phi(x)\rangle_{\mathcal{H}} \mid h \in \mathcal{H}\} = \mathrm{im}_{\mathcal{H}}\Psi,$$

where $\Psi : \mathcal{H} \to \mathbb{K}^{\mathcal{X}}$, $\Psi(h) = \langle h, \Phi(\cdot)\rangle_{\mathcal{H}}$, and for each $f \in H$ define

$$\|f\|_H = \inf_{h\in\Psi^{-1}(f)} \|h\|_{\mathcal{H}}.$$

Then $(H, \|\cdot\|_H)$ is a $\mathbb{K}$-Hilbert space of functions with reproducing kernel k. In particular, H is an RKHS.

Proof. **Step 1** Since Ψ is linear, H is a vector space of functions. Define $\mathcal{H}_0 = \text{null}(\Psi)$, then $\mathcal{H}_0$ is closed: Let $(h_n)_n \subseteq \mathcal{H}_0$ with $h_n \to h \in \mathcal{H}$, then for all $x \in \mathcal{X}$

$$(\Psi(h))(x) = \langle h, \Phi(x)\rangle_{\mathcal{H}} = \langle \lim_n h_n, \Phi(x)\rangle_{\mathcal{H}} = \lim_n \langle h_n, \Phi(x)\rangle_{\mathcal{H}} = \lim_n (\Psi(h_n))(x) = 0,$$

where we used the continuity of the scalar product in the third equality and $h_n \in \text{null}(\Psi)$ in the last inequality. We therefore find $\Psi(h) \equiv 0$ and hence $h \in \mathcal{H}_0$, establishing that $\mathcal{H}_0$ is closed. We can now decompose[3] $\mathcal{H} = \mathcal{H}_0 \oplus_\perp \mathcal{H}_1$ with $\mathcal{H}_1 = \mathcal{H}_0^\perp$, and let us define $\bar{\Psi} = \Psi|_{\mathcal{H}_1}$. By construction, $\bar{\Psi}$ is injective. It is also surjective. To show this, let $f \in H = \text{im}_{\mathcal{H}}\Psi$, then there exists $h = h_0 \oplus_\perp h_1$, where $h_i \in \mathcal{H}_i$ for $i = 0, 1$, such that $f = \Psi(h) = \Psi(h_0 + h_1) = \Psi(h_0) + \Psi(h_1) = \bar{\Psi}(h_1)$, where we used that $\Psi(h_0) = 0$ (since $h_0 \in \mathcal{H}_0 = \text{null}(\Psi)$), and $\Psi(h_1) = \bar{\Psi}(h_1)$. This shows that f is indeed in the image of $\bar{\Psi}$, and since f was an arbitrary element from H, we established that $\bar{\Psi}$ is surjective.

Step 2 We can now show that H is a Hilbert space. Let $f \in H$, then

$$\begin{aligned}
\|f\|_H^2 &= \inf_{h \in \bar{\Psi}^{-1}(f)} \|h\|_{\mathcal{H}}^2 \\
&= \inf_{\substack{h_0 \in \mathcal{H}_0, h_1 \in \mathcal{H}_1 \\ f = \Psi(h_0 + h_1)}} \|h_0 + h_1\|_{\mathcal{H}}^2 \\
&= \inf_{\substack{h_1 \in \bar{\Psi}^{-1}(f) \\ h_0 \in \mathcal{H}_0}} \|h_0\|_{\mathcal{H}}^2 + \|h_1\|_{\mathcal{H}}^2 \\
&= \|\bar{\Psi}^{-1}(f)\|_{\mathcal{H}}^2
\end{aligned}$$

This implies that $\|\cdot\|_H$ is a Hilbert space norm and $\bar{\Psi}$ is an isometric isomorphism between $\mathcal{H}_1$ and H. In particular, for $f, g \in H$ we have $\langle f, g\rangle_H = \langle \bar{\Psi}^{-1} f, \bar{\Psi}^{-1} g\rangle_{\mathcal{H}}$.

Step 3 Finally, we can establish that H is an RKHS with reproducing kernel k. Let $x \in \mathcal{X}$, then

$$k(\cdot, x) = \langle \Phi(x), \Phi(\cdot)\rangle_{\mathcal{H}} = \Psi(\Phi(x)) \in H.$$

[3] Here we need that $\mathcal{H}_0$ is closed.

Furthermore, for $f \in H$ we have

$$f(x) = \langle \bar{\Psi}^{-1} f, \Phi(x) \rangle_{\mathcal{H}} = \langle \bar{\Psi}^{-1} f, \bar{\Psi}^{-1}(k(\cdot, x)) \rangle_{\mathcal{H}} = \langle f, k(\cdot, x) \rangle_H,$$

where we used in the second equality that $\Phi(x) = \bar{\Psi}^{-1}(k(\cdot, x))$. This follows since $k(\cdot, x) = \Psi(\Phi(x))$ and for all $h_0 \in \mathcal{H}_0$ we have $\langle h_0, \Phi(x) \rangle_{\mathcal{H}} = \Psi(h_0)(x) = 0$, i.e., $\Phi(x) \perp \mathcal{H}_0$, hence $\Phi(x) \in \mathcal{H}_1$ and therefore $k(\cdot, x) = \bar{\Psi}(\Phi(x))$. Altogether, k is a reproducing kernel for H, implying that H is an RKHS. □

Before moving on, we would like to point out two observations which are probably well-known, but rarely explicitly stated in the literature.

Remark 2.3.8. Consider the situation and notations of Theorem 2.3.7 and its proof.

1. For all $f \in H$, there exists exactly one $h \in H$ with $f = \Psi(h)$ and $\|f\|_H = \|h\|_{\mathcal{H}}$. Intuitively, this means that the RKHS induced by a kernel can by identified with the feature space. Put differently, *any* feature space of a kernel acts as a representation of the RKHS induced by the kernel.

   To *prove* this claim, it is enough to observe that $\bar{\Psi}$ is an isometric isomorphism between $\mathcal{H}_1$ and H, and to take the definition $\bar{\Psi}$ into account.

2. The map Ψ is an isometry if and only if $\mathrm{im}_{\mathcal{X}}(\Phi) = \{\Phi(x) \mid x \in \mathcal{X}\}$ is total in $\mathcal{H}$, i.e., if $\overline{\mathrm{span}\{\Phi(x) \mid x \in \mathcal{X}\}}^{\|\cdot\|_{\mathcal{H}}} = \mathcal{H}$.

   To *show* this, note that by definition Ψ is an isometry if and only if $\mathrm{null}(\Psi) = \{0\}$. Furthermore, for all $h \in \mathcal{H}$ we have $h \in \mathrm{null}(\Psi)$ if and only if $\Psi(h_0)(x) = \langle h_0, \Phi(x) \rangle_{\mathcal{H}}$ for all $x \in \mathcal{X}$, i.e., if and only if $h_0 \in \mathrm{im}_{\mathcal{X}} \Phi^{\perp}$. But $\mathrm{im}_{\mathcal{X}}(\Phi)$ is total if and only if $\mathrm{im}_{\mathcal{X}}(\Phi)^{\perp} = \{0\}$.

If k is a kernel, then there exists by definition a corresponding feature space-feature map pair $(\mathcal{H}, \Phi)$, and Theorem 2.3.7 ensures the existence of an RKHS H such that k is the reproducing kernel for this RKHS. According to Proposition 2.2.7, this is then the unique RKHS with reproducing kernel k. Conversely, every reproducing kernel of an RKHS is a kernel according to Proposition 2.3.5. Let us summarize all of this.

Theorem 2.3.9. Let $\mathcal{X} \neq \emptyset$ be a set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a function. Then k is a ($\mathbb{K}$-)kernel if and only if it is the reproducing kernel of a ($\mathbb{K}$-)RKHS, and the latter is unique for a given kernel.

Let us build some additional intuition on the relation between kernels in the sense of Definition 2.3.4 and RKHSs. If k is a kernel, then by definition it has a feature space-feature map pair, and Theorem 2.3.7 constructs an RKHS from this. In other words, for a kernel we can always use the associated RKHS as a feature space and the canonical feature map Φ_k as a feature map. Moreover, the next result (which appears as part of [189, Theorem 4.21]) shows that the RKHS is a rather special feature space.

Lemma 2.3.10. Consider the situation of Theorem 2.3.7. For all $h \in \mathcal{H}$ we have $\|\Psi(h)\|_H \leq \|h\|_{\mathcal{H}}$[4]. Additionally, let $\mathcal{B}_1^{\mathcal{H}}$ and $\mathcal{B}_1^H$ be the open unit ball of $\mathcal{H}$ and H, respectively, then $\Psi(\mathcal{B}_1^{\mathcal{H}}) = \mathcal{B}_1^H$, i.e., Ψ is a metric surjection.

Proof. Decompose $h \in \mathcal{H}$ as $h = h_0 \oplus_\perp h_1$, $h_i \in \mathcal{H}_i$ for $i = 0, 1$, then

$$\|\Psi(h)\|_H = \|\Psi(h_0 + h_1)\|_H = \|\Psi(h_1)\|_H = \|\bar{\Psi}(h_1)\|_H = \|h_1\|_{\mathcal{H}} \leq \|h\|_{\mathcal{H}},$$

showing that $\Psi(\mathcal{B}_1^{\mathcal{H}}) \subseteq \mathcal{B}_1^H$. Let $\mathcal{B}_1^{\mathcal{H}_1}$ be the open unit ball in $\mathcal{H}_1$, then we have $\mathcal{B}_1^{\mathcal{H}_1} \subseteq \mathcal{B}_1^{\mathcal{H}}$ and since $\bar{\Psi}$ is an isometric isomorphism between $\mathcal{H}_1$ and H we even get $\Psi(\mathcal{B}_1^{\mathcal{H}}) = \mathcal{B}_1^H$. □

Note that in Theorem 2.3.7 we can use *any* feature space-feature map pair of a given kernel. Lemma 2.3.10 then says that the corresponding RKHS is the *smallest* possible feature space for the kernel. Intuitively, the map Ψ marginalizes out any superfluous parts of a given feature space for a kernel.

2.4. Positive semidefiniteness: From matrices to kernels

Let $M \in \mathbb{K}^{n \times n}$ and recall that M is called *positive semidefinite* if for all $v \in \mathbb{C}^n$ we have $v^* M v \geq 0$. Note that this implicitly includes the requirement that $v^* M v \in \mathbb{R}$.

A matrix $M \in \mathbb{K}^{n \times n}$ can also be interpreted as a bivariate function: Let $\mathcal{X} = \{1, \ldots, n\}$ and define $m : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ by $m(i, j) = M_{ij}$ for all $i, j \in \mathcal{X}$. We can now characterize the positive semidefiniteness of the matrix M using the bivariate function m. The following result is well-known, but curiously it is rarely stated explicitly (and proved) in expositions of RKHSs and kernels.

[4] Since Ψ is linear, this is equivalent to 1-Lipschitz continuity, which is also called *nonexpanding*.

Lemma 2.4.1. The matrix $M \in \mathbb{K}^{n\times n}$ is positive semidefinite if and only if for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$, and all $\alpha_1, \ldots, \alpha_N \in \mathbb{C}$ we have

$$\sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j m(x_j, x_i) \geq 0. \tag{2.12}$$

Proof. Assume (2.12) holds and let $v \in \mathbb{C}^n$ be arbitrary. Set $N = n$ and define $x_i = i$ and $\alpha_i = v_i$ for $i = 1, \ldots, n$, then we get from (2.12) that

$$v^* M v = \sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j m(x_j, x_i) \geq 0,$$

showing that M is positive semidefinite.

Conversely, assume that M is positive semidefinite and let $N \in \mathbb{N}$ (we can w.l.o.g. assume that $N > 0$), $x_1, \ldots, x_N \in \mathcal{X}$ and $\alpha_1, \ldots, \alpha_N \in \mathbb{C}$ be arbitrary. For $j = 1, \ldots, n$ define $N_j = \sharp\{k = 1, \ldots, N \mid x_k = j\}$ and (if $N_j > 0$) for $k = 1, \ldots, N_j$ define $i_k^{(j)} \in \mathcal{X}$ such that $x_{i_k^{(j)}} = j$. Furthermore, for $j = 1, \ldots, n$ define $v_j = \sum_{k=1}^{N_j} \alpha_{i_k^{(j)}}$ if $N_j > 0$ and $v_j = 0$ otherwise. We then have

$$\begin{aligned}
\sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j m(x_j, x_i) &= \sum_{i,j=1}^{n} \sum_{k=1}^{N_i} \sum_{\ell=1}^{N_j} \alpha_{i_k^{(i)}} \bar{\alpha}_{i_\ell^{(j)}} m(j, i) \\
&= \sum_{i,j=1}^{n} \left(\sum_{k=1}^{N_i} \alpha_{i_k^{(i)}} \right) \left(\sum_{k\ell=1}^{N_j} \alpha_{i_\ell^{(j)}} \right) m(j, i) \\
&= \sum_{i,j=1}^{n} v_i \bar{v}_j M_{ji} \\
&= v^* M v \geq 0,
\end{aligned}$$

establishing (2.12). □

Remark 2.4.2. Let $M \in \mathbb{R}^{n\times n}$, then it is well-known that M is positive semidefinite if and only if M is symmetric (so $M_{ij} = M_{ji}$ for all $i, j = 1, \ldots, n$), and that for all $v \in \mathbb{R}^n$ we have $v^\top M v \geq 0$, cf. [32, Lemma 1.3.4]. We also have a real equivalent to Lemma 2.4.1: M is positive semidefinite if and only if m is symmetric, (so for all $x, y \in \mathcal{X}$ we have $m(x, y) = m(y, x)$), and for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$, and all $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$ again (2.12) holds.

Lemma 2.4.1 motivates the following extension of positive semidefiniteness from $n \times n$ matrices to arbitrary bivariate functions.

Definition 2.4.3. Let $\mathcal{X} \neq \emptyset$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$. We call k positive semidefinite if for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$, and all $\alpha_1, \ldots, \alpha_N \in \mathbb{C}$ we have

$$\sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j k(x_j, x_i) \geq 0. \tag{2.13}$$

We can rephrase Definition 2.4.3 as follows: A bivariate function k is positive semidefinite if for all choices $x_1, \ldots, x_N \in \mathcal{X}$ the matrix $(k(x_j, x_i))_{i,j=1,\ldots,N}$ is positive semidefinite.

Remark 2.4.4.

1. Let $\mathbb{K} = \mathbb{R}$, then an equivalent variant of Definition 2.4.3 is as follows: We call $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ positive semidefinite if k is symmetric, i.e., for all $x, x' \in \mathcal{X}$ we have $k(x, x') = k(x', x)$, and the condition in Definition 2.4.3 holds for all $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$.

2. Recall that $M \in \mathbb{K}^{n \times n}$ is called positive definite if for all $v \in \mathbb{C}^n \setminus \{0\}$ we have $v^* M v > 0$. We can formulate a corresponding variant of Lemma 2.4.1: Define again $\mathcal{X} = \{1, \ldots, n\}$ and $m(i, j) = M_{ij}$, then M is positive definite if and only if for all $N \in \mathbb{N}$, pairwise distinct $x_1, \ldots, x_N \in \mathcal{X}$ (note that this implies that $N \leq n$) and $\alpha_1, \ldots, \alpha_N \in \mathbb{C}$ not all zero we have $\sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j m(x_j, x_i) > 0$.

3. The preceding item motivates the following definition: Let $\mathcal{X} \neq \emptyset$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$. We call k positive definite if for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$ pairwise distinct, and all $\alpha_1, \ldots, \alpha_N \in \mathbb{C}$ not all zero we have $\sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j m(x_j, x_i) > 0$. Similarly to Definition 2.4.3, this is equivalent to all matrices $(k(x_j, x_i))_{i,j=1,\ldots,N}$ being positive definite.

4. If $\mathbb{K} = \mathbb{R}$, we have again an equivalent definition of positive definiteness: The function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if and only if k is symmetric, and for all $N \in \mathbb{N}$, $x_1, \ldots, x_N \in \mathcal{X}$ pairwise distinct, and all $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$ not all zero we have $\sum_{i,j=1}^{N} \alpha_i \alpha_j m(x_j, x_i) > 0$.

5. The terminology in the literature is not uniform: Positive semidefiniteness in the sense of Definition 2.4.3 is sometimes called *positive definiteness* (or

sometimes *of positive type*) and positive definiteness is called *strict positive definiteness*.

In Section 2.5 we will see another motivation for the concept of a positive semidefinite (and positive definite) bivariate function.

Let us now build a connection to the developments of the preceding sections with the following result, which is folklore.

Proposition 2.4.5. Let $\mathcal{X} \neq \emptyset$ be a set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a kernel in the sense of Definition 2.3.4. Then k is positive semidefinite in the sense of Definition 2.4.3. In particular, if k is the reproducing kernel of an RKHS, then it is positive semidefinite.

Proof. Let $\mathcal{H}$ be a feature space and $\Phi : \mathcal{X} \to \mathcal{H}$ a feature map for k. Let $N \in \mathbb{N}$ (w.l.o.g. $N > 0$), $x_1, \dots, x_N \in \mathcal{X}$ and $\alpha_1, \dots, \alpha_N \in \mathbb{C}$, then

$$\begin{aligned}\sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j k(x_j, x_i) &= \sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{N} \alpha_i \Phi(x_i), \sum_{j=1}^{N} \alpha_j \Phi(x_j) \right\rangle_{\mathcal{H}} \\ &= \left\| \sum_{i=1}^{N} \alpha_i \Phi(x_i) \right\|_{\mathcal{H}} \geq 0.\end{aligned}$$

The last claim is clear since a reproducing kernel is a kernel according to Proposition 2.3.5. □

The proof of Proposition 2.4.5 reveals that for $x_1, \dots, x_N \in \mathcal{X}$, the matrix $(k(x_j, x_i))_{i,j}$ can be interpreted as a Gram matrix. On the one hand, this provides an intuitive explanation as to why a kernel is positive semidefinite. On the other hand, it suggests that positive semidefinite functions should be kernels. This is indeed the case and this result appears naturally in the next section in a different context.

2.5. Native spaces: RKHSs as natural approximation spaces

We now turn to another perspective on RKHSs, which is motivated by the scattered data approximation literature [217]. Our presentation is inspired by [175, Chapter 3] and [32, Chapter 1]. An important task in mathematics and numerics is function

interpolation. Let $\mathcal{X} \neq \emptyset$ be some set, $F \subseteq \mathbb{K}^{\mathcal{X}}$ a set of functions, and $x_1, \ldots, x_N \in \mathcal{X}$, $N \in \mathbb{N}$, as well as $y_1, \ldots, y_N \in \mathbb{K}$. The goal is to find some $f \in F$ such that $f(x_n) = y_n$ for all $n = 1, \ldots, N$. Note that ideally we want to use the same F for all such problem instances and even get a unique solution $f \in F$. A simple approach is to take F as a vector space of functions, fix a linearly independent family of functions $(\phi_n)_n$ spanning F, and for a concrete problem instance use the ansatz

$$\sum_{n=1}^{N} \alpha_n \phi_n(x_m) = y_m, \qquad m = 1, \ldots, N. \tag{2.14}$$

Note that the ϕ_n are independent of the problem data, i.e., they do not depend on the inputs x_n or interpolated values y_n. Due to the Mairhuber-Curtis Theorem (see [218, Chapter 2] for a thorough discussion), using such a data-independent approach is in many cases not possible. As a simple remedy, we can make the functions in the ansatz (2.14) data-dependent. For this, let us replace $\phi_1, \ldots, \phi_N$ by $\phi(\cdot, x_1), \ldots, \phi(\cdot, x_N)$ for some function $\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$, leading to the new system of linear equations

$$\sum_{n=1}^{N} \alpha_n \phi(x_m, x_n) = y_m, \qquad m = 1, \ldots, N, \tag{2.15}$$

which can be conveniently written as $\Phi\alpha = y$ by setting

$$\Phi = \begin{pmatrix} \phi(x_1, x_1) & \cdots & \phi(x_1, x_N) \\ \vdots & & \vdots \\ \phi(x_N, x_1) & \cdots & \phi(x_N, x_N) \end{pmatrix} \quad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}.$$

The interpolation problem has now been replaced by the problem of solving a simple linear equation system. From a numerical perspective particularly benign are linear equations with a positive semidefinite Φ. Since we want to use the same ϕ for all possible interpolation problems (on input set $\mathcal{X}$), this leads to the requirement that all resulting matrices Φ (in this context called interpolation matrices) are positive semidefinite. But recalling the developments in Section 2.4, we see that this is equivalent to the function ϕ being positive semidefinite in the sense of Definition 2.4.3. In other words, if we want to use the ansatz (2.15) and end up with a positive

semidefinite interpolation matrix Φ, we have to use a positive semidefinite function ϕ. This provides another motivation for the concepts in Section 2.4.

Remark 2.5.1. Ideally, we want that Φ is also invertible so that ansatz (2.15) leads to the existence of a unique solution. It is clear that for this we need that $x_1, \dots, x_N$ are pairwise distinct. Similarly to the arguments above, this means that we want some ϕ such that for all pairwise distinct $x_1, \dots, x_N \in \mathcal{X}$ the interpolation matrix is positive definite. But this is equivalent to ϕ being positive definite in the sense of Section 2.4. However, since this will not play a role in the general theory of RKHS, we do not elaborate on it and instead refer to [218], for example, for more details.

Let now $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ be a positive semidefinite function (in the sense of Definition 2.4.3). Using the approach outlined so far in this section, we are working with the vector space of functions

$$H_0 = \left\{ \sum_{n=1}^{N} \alpha_n k(\cdot, x_n) \mid N \in \mathbb{N},\ x_n \in \mathcal{X},\ \alpha_n \in \mathbb{K} \right\} = \operatorname{span}\{k(\cdot, x) \mid x \in \mathcal{X}\}. \quad (2.16)$$

What happens if we want to use H_0 to approximate some function by going to the limit? To use an analogy, step functions on a compact set are dense w.r.t. the supremum norm in the set of continuous functions, i.e., we can use step functions to approximate continuous functions arbitrarily well. What is a natural space of functions that can be approximated well by H_0? In order to answer this, we need a suitable topology (to allow limits), and since k is positive semidefinite, it is natural to look for a scalar product (inducing a topology). What is a reasonable choice? Since we want to approximate *functions*, norm convergence should imply pointwise convergence. But this means that we look for an RKHS H with $H_0 \subseteq H$, cf. our discussion in Section 2.1. From Section 2.2 we know that this RKHS has a unique reproducing kernel. A natural choice is then to use k as the reproducing kernel. In particular, we need that for all $x, x' \in \mathcal{X}$ we have $\langle k(\cdot, x'), k(\cdot, x) \rangle_H = k(x, x')$. Extending this relation sesquilinearly to all of H_0 then essentially gives a scalar product on H_0. This is formalized in the next well-known result.

Proposition 2.5.2. Define

$$\langle \cdot, \cdot \rangle_0 : H_0 \times H_0 \to \mathbb{K}, \langle f, g \rangle_0 = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \bar{\beta}_j k(y_j, x_i), \quad (2.17)$$

where $f = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{M} \beta_j k(\cdot, y_j)$ are two arbitrary representations of $f, g \in H_0$. Then $\langle \cdot, \cdot \rangle_0$ is a well-defined scalar product on H_0.

Proof. Let $f, g \in H_0$ and choose two representations $f = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i)$ and $g = \sum_{j=1}^{M} \beta_j k(\cdot, y_j)$. Then

$$\langle f, g \rangle_0 = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \bar{\beta}_j k(y_j, x_i) = \sum_{i=1}^{N} \alpha_i \left(\sum_{j=1}^{M} \bar{\beta}_j \overline{k(x_i, y_y)} \right) = \sum_{i=1}^{N} \alpha_i \overline{g(x_i)}$$

shows that $\langle f, g \rangle_0$ is independent of the representation of g (since only its function evaluations are used), and

$$\langle f, g \rangle_0 = \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \bar{\beta}_j k(y_j, x_i) = \sum_{i=j}^{M} \bar{\beta}_j \left(\sum_{i=1}^{N} \alpha_i k(y_j, x_i) \right) = \sum_{j=1}^{M} \bar{\beta}_j f(y_j)$$

shows that it is also independent of the representation of f, hence $\langle \cdot, \cdot \rangle_0$ is well-defined. Furthermore, it is clear that $\langle \cdot, \cdot \rangle_0$ is sequilinear (bilinear for $\mathbb{K} = \mathbb{R}$).

Next, let $f = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i) \in H_0$ be arbitrary, then the positive semidefiniteness of k shows

$$\left\langle \sum_{i=1}^{N} \alpha_i k(\cdot, x_i), \sum_{j=1}^{N} \alpha_j k(\cdot, x_j) \right\rangle_0 = \sum_{i,j=1}^{N} \alpha_i \bar{\alpha}_j k(x_j, x_i) \geq 0,$$

establishing the positive semidefiniteness of $\langle \cdot, \cdot \rangle_0$. We now need to show that $\langle \cdot, \cdot \rangle_0$ is even positive definite. By the usual argument, $\langle \cdot, \cdot \rangle_0$ fulfills the Cauchy-Schwarz inequality. Let $f = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i) \in H_0$ such that $\langle f, f \rangle_0 = 0$. For all $x \in \mathcal{X}$ we then find that

$$\begin{aligned} |f(x)| &= \left| \sum_{i=1}^{N} \alpha_i k(x, x_i) \right| = \left| \left\langle \sum_{i=1}^{N} \alpha_i k(\cdot, x_i), k(\cdot, x) \right\rangle_0 \right| \\ &\leq \left\| \sum_{i=1}^{N} \alpha_i k(\cdot, x_i) \right\|_0 \, \|k(\cdot, x)\|_0 = 0, \end{aligned}$$

where we used Cauchy-Schwarz in the inequality and $\|f\|_0 = \sqrt{\langle f, f \rangle_0} = 0$ in the last equality. This shows that $f \equiv 0$ and hence the positive definiteness of $\langle \cdot, \cdot \rangle_0$. Altogether, $\langle \cdot, \cdot \rangle_0$ is a well-defined scalar product on H_0. □

Before turning to the question of limits in H_0, we make the following observation: The definition of $\langle \cdot, \cdot \rangle_0$ implies that for all $f \in H_0$ and $x \in \mathcal{X}$ we have $f(x) = \langle f, k(\cdot, x) \rangle_0$ (this has been used in the proof of Proposition 2.5.2), i.e., k is a reproducing kernel for the pre Hilbert space H_0. In particular, we have for all $x, x' \in \mathcal{X}$ that $k(x, x') = \langle k(\cdot, x'), k(\cdot, x) \rangle_0$. This almost shows that k is a kernel, but we need completeness of the feature space (and H_0 is in general not complete, hence it cannot be used as a feature space here). Let $\mathcal{H}$ be a completion of H_0 and let $I : H_0 \to \mathcal{H}$ be the corresponding isometric embedding. We can then define $\Phi : \mathcal{X} \to \mathcal{H}$ by $\Phi(x) = I[k(\cdot, x)]$ and get

$$k(x, x') = \langle k(\cdot, x'), k(\cdot, x) \rangle_0 = \langle I[k(\cdot, x')], I[k(\cdot, x)] \rangle_{\mathcal{H}} = \langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}},$$

showing that every positive semidefinite function is a kernel. Combining this with Proposition 2.4.5 we get the next important result, see e.g. [189, Theorem 4.16].

Theorem 2.5.3. Let $\mathcal{X} \neq \emptyset$ be arbitrary and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ some function. Then k is a kernel if and only if it is positive semidefinite.

We know from Theorem 2.3.7 that every kernel is the reproducing kernel of an RKHS that can be built from any feature space-feature map pair. Applying this to the completion $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ just introduced establishes the following central result, known as the Moore-Aronszajn Theorem.

Theorem 2.5.4. Let $\mathcal{X} \neq \emptyset$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ be some function. Then k is positive semidefinite if and only if it is the reproducing kernel of a (uniquely determined) RKHS H.

Since we applied Theorem 2.3.7 to the abstract completion $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$, it is not quite clear at this point how the corresponding RKHS actually looks like. To build more intuition and add another perspective, let us return to H_0 and to the question of limits of functions in this space. Let $(f_n)_n$ be a Cauchy sequence in $(H_0, \| \cdot \|_0)$. For $x \in \mathcal{X}$ and arbitrary $n, m \in \mathbb{N}$ we have

$$|f_n(x) - f_m(x)| = |\langle f_n - f_m, k(\cdot, x) \rangle_0| \leq \|f_n - f_m\|_0 \|k(\cdot, x)\|_0,$$

showing that $(f_n(x))_n$ is a Cauchy sequence in $\mathbb{K}$ and hence convergent. For each Cauchy sequence $(f_n)_n$ in $(H_0, \| \cdot \|_0)$ we can therefore define a function $f : \mathcal{X} \to \mathbb{K}$

by $f(x) = \lim_n f_n(x)$. Let H be the set of all such functions, then H is a $\mathbb{K}$-vector space that contains H_0. Next, we need to extend $\langle \cdot, \cdot \rangle_0$ to this space. This is done in the next result, which is from the proof of [32, Theorem 1.3.2].

Lemma 2.5.5. For $f, g \in H$ let $(f_n)_n$, $(g_n)_n$ be two Cauchy sequences such that $f_n(x) \to f(x)$ and $g_n(x) \to g(x)$ for all $x \in \mathcal{X}$. Then $\lim_n \langle f_n, g_n \rangle_0$ exists and is independent of the choice of $(f_n)_n, (g_n)_n$ and

$$\langle \cdot, \cdot \rangle_H : H \times H \to \mathbb{K},\ \langle f, g \rangle_H = \lim_n \langle f_n, g_n \rangle_0 \tag{2.18}$$

defines a scalar product on H. Furthermore, $\langle \cdot, \cdot \rangle_H|_{H_0} = \langle \cdot, \cdot \rangle_0$.

Proof. **Step 1** For $n, m \in \mathbb{N}$ we have

$$\begin{aligned} |\langle f_n, g_n \rangle_0 - \langle f_m, g_m \rangle_0| &= |\langle f_n - f_m, g_n \rangle_0 + \langle f_m, g_n - g_m \rangle_0| \\ &\leq \|f_n - f_m\|_0 \|g_n\|_0 + \|f_m\|_0 \|g_n - g_m\|_0, \end{aligned}$$

where we used the triangle and Cauchy-Schwarz inqualities. Since Cauchy sequences are bounded, this shows that $(\langle f_n, g_n \rangle_0)_n$ is a Cauchy sequence in $\mathbb{K}$ and hence convergent.

Step 2 We need the following auxiliary result: Let $(h_n)_n$ be a Cauchy sequence in $(H_0, \langle \cdot, \cdot \rangle_0)$ s.t. $h_n(x) \to 0$ for all $x \in \mathcal{X}$, then $\|h_n\|_0 \to 0$. To see this, let $\epsilon > 0$ be arbitrary, let $B > 0$ be a bound on $\|h_n\|_0$ (exists since Cauchy sequences are bounded) and choose $N_\epsilon \in \mathbb{N}$ such that $\|h_n - h_{N_\epsilon}\|_0 \leq \frac{\epsilon}{B}$ for all $n \geq N_\epsilon$. Let

$$h_{N_\epsilon} = \sum_{m=1}^{M} \alpha_m k(\cdot, x_m)$$

be some representation of $h_{N_\epsilon} \in H_0$. We then get

$$\begin{aligned} \|h_n\|_0^2 = \langle h_n, h_n \rangle_0 &= \langle h_n - h_{N_\epsilon}, h_n \rangle + \langle h_{N_\epsilon}, h_n \rangle_0 \\ &\leq \|h_n - h_{N_\epsilon}\|_0 \|h_n\|_0 + \sum_{m=1}^{M} \alpha_m h_n(x_m) \\ &\leq \epsilon + \sum_{m=1}^{M} \alpha_m h_n(x_m), \end{aligned}$$

where we used Cauchy-Schwarz and the reproducing property of k in the first inequality and the choice of N_ϵ in the second inequality. Since N_ϵ is constant and h_n converges pointwise to 0, we get $\limsup_n \|h_n\|_0^2 \leq \epsilon$. Since $\epsilon > 0$ was arbitrary, we find $\|h_n\|_0 \to 0$.

Step 3 Let $(\tilde{f}_n)_n, (\tilde{g}_n)_n$ be another two Cauchy sequences with $\tilde{f}_n(x) \to f(x)$ and $\tilde{g}_n(x) \to g(x)$ for all $x \in \mathcal{X}$. Then $(f_n - \tilde{f}_n)_n$ and $(g_n - \tilde{g}_n)_n$ are Cauchy sequences that converge pointwise to 0, hence according to Step 2 also in norm, and

$$\begin{aligned} |\langle f_n, g_n\rangle_0 - \langle \tilde{f}_n, \tilde{g}_n\rangle_0| &= |\langle f_n - \tilde{f}_n, g_n\rangle_0 - \langle \tilde{f}_n, \tilde{g}_n - g_n\rangle_0| \\ &\leq \|f_n - \tilde{f}_n\|_0 \|g_n\|_0 + \|\tilde{f}_n\|_0 \|g_n - \tilde{g}_n\|_0 \end{aligned}$$

shows that $\lim_n \langle f_n, g_n\rangle_0 = \lim_n \langle \tilde{f}_n, \tilde{g}_n\rangle_0$. Summarizing, (2.18) is independent of the chosen Cauchy sequences, hence well-defined.

Step 4 It is clear that $\langle \cdot, \cdot \rangle_H$ is bilinear, Hermitian (symmetric for $\mathbb{K} = \mathbb{R}$) and positive semidefinite (these properties are induced by $\langle \cdot, \cdot \rangle_0$ and the linearity of the limit). To establish even positive definiteness of $\langle \cdot, \cdot \rangle_H$, let $f \in H$ such that $\langle f, f\rangle_H = 0$ and choose any Cauchy sequence $(f_n)_n$ in $(H_0, \langle \cdot, \cdot \rangle_0)$ with $f_n(x) \to f(x)$ for all $x \in \mathcal{X}$. By definition of $\langle \cdot, \cdot \rangle$ we then get $\lim_n \|f_n\|_0 = \|f\|_H$ and

$$f(x) = \lim_n f_n(x) = \lim_n \langle f_n, k(\cdot, x)\rangle_0 \leq \|k(\cdot, x)\|_0 \cdot \lim_n \|f_n\|_0 = 0,$$

where we used the reproducing property of k in H_0 in the second inequality. We therefore find that $f \equiv 0$, showing that $\langle \cdot, \cdot \rangle_H$ is positive definite.

Altogether, $\langle \cdot, \cdot \rangle_H$ is a well-defined scalar product on H and by construction it is equal to $\langle \cdot, \cdot \rangle_0$ on H_0. □

It turns out that we arrived at an RKHS with kernel k.

Theorem 2.5.6. $(H, \langle \cdot, \cdot \rangle_H)$ is an RKHS with reproducing kernel k and H_0 is dense in H.

The following proof is based on the one of [32, Theorem 1.3.2].

Proof. **Step 1** We first need the following result: Let $f \in H$, $(f_n)_n$ a Cauchy sequence in $(H_0, \langle \cdot, \cdot \rangle_0)$ that converges pointwise to f, then f_n converges also w.r.t. $\|\cdot\|_H$ to f, i.e., $\|f_n - f\|_H \to 0$. To see this, let $\epsilon > 0$ be arbitrary and choose $N_\epsilon \in \mathbb{N}$

such that for all $n, m \geq N_\epsilon$ we have $\|f_n - f_m\|_0 \leq \epsilon$. Observe that $(f_m - f_{N_\epsilon})_m$ is a Cauchy sequence in H_0 that converges pointwise to $f - f_{N_\epsilon}$, hence $\|f - f_{N_\epsilon}\|_0 = \lim_m \|f_m - f_{N_\epsilon}\|_0 \leq \epsilon$. Since $\epsilon > 0$ was arbitrary, this shows that $\|f - f_n\|_H \to 0$.

Note that this step shows also the density of H_0 in H.

Step 2 Let $f \in H$ and $x \in \mathcal{X}$ be arbitrary. Choose a Cauchy sequence $(f_n)_n$ in H_0 that converges pointwise to f, then

$$f(x) = \lim_n f_n(x) = \lim_n \langle f_n, k(\cdot, x)\rangle_0 = \langle f, k(\cdot, x)\rangle_H.$$

Step 3 H is complete: Let $(f_n)_n$ be a Cauchy sequence in $(H, \langle\cdot,\cdot\rangle_H)$. For each $x \in \mathcal{X}$ and $n, m \in \mathbb{N}$ we have

$$|f_n(x) - f_m(x)| = |\langle f_n - f_m, k(\cdot, x)\rangle_H| \leq \|f_n - f_m\|_H \|k(\cdot, x)\|_H,$$

where we have used Step 2 in the first equality. This shows that $(f_n(x))_n$ is a Cauchy sequence in $\mathbb{K}$, hence convergent, so we can define the function $f \in \mathbb{K}^{\mathcal{X}}$ by $f(x) = \lim_n f_n(x)$. We have to show that $f \in H$ and that $f_n \to f$ in $(H, \|\cdot\|_H)$. For this, we construct a Cauchy sequence in H_0 that converges pointwise to f (showing by definition of H that $f \in H$) and then use it also to show $\|f - f_n\|_H \to 0$.

Let $(\epsilon_n)_n$ some sequence with $\epsilon_n > 0$ and $\epsilon_n \to 0$. For each $n \in \mathbb{N}$ let $(h_m^{(n)})_m$ be a Cauchy sequence in H_0 that converges pointwise to f_n. According to Step 1, $\|h_m^{(n)} - f_n\|_H \to 0$, so we can choose $M_n \in \mathbb{N}$ such that $\|h_{M_n}^{(n)} - f_n\|_H \leq \epsilon_n$ and then set $g_n = h_{M_n}^{(n)}$. We show that $(g_n)_n$ is a Cauchy sequence in H_0: Let $\epsilon > 0$ be arbitary. Choose $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$ we have $\epsilon_n \leq \frac{\epsilon}{3}$. Choose $N_2 \in \mathbb{N}$ such that for all $n, m \geq N_2$ we have $\|f_n - f_m\|_H \leq \frac{\epsilon}{3}$ (exists since $(f_n)_n$ is a Cauchy sequence in H). Then for all $n, m \geq \max\{N_1, N_2\}$ we have

$$\begin{aligned}\|g_n - g_m\|_0 &\leq \|g_n - f_n\|_H + \|f_n - f_m\|_H + \|f_m - g_m\|_H \\ &\leq \epsilon_n + \frac{\epsilon}{3} + \epsilon_m \leq \epsilon,\end{aligned}$$

establishing that $(g_n)_n$ is a Cauchy sequence. Next, this sequence converges pointwise to f: Let $x \in \mathcal{X}$ and $\epsilon > 0$ be arbitrary. Choose $N_1 \in \mathbb{N}$ such that $|f(x) - f_n(x)| \leq \frac{\epsilon}{2}$ for all $n \geq N_1$ (recall that f_n converges pointwise to f) and choose $N_2 \in \mathbb{N}$ such that for all $n \geq N_2$ we have $\epsilon_n \|k(\cdot, x)\|_H \leq \frac{\epsilon}{2}$. Then for all $n \geq \max\{N_1, N_2\}$

we have

$$\begin{aligned}|f(x) - g_n(x)| &\leq |f(x) - f_n(x)| + |f_n(x) - g_n(x)| \\ &\leq \frac{\epsilon}{2} + |\langle f_n - g_n, k(\cdot, x)\rangle_H| \\ &\leq \frac{\epsilon}{2} + \|f_n - g_n\|_H \|k(\cdot, x)\|_H \leq \epsilon.\end{aligned}$$

Summarizing, $(g_n)_n$ is a Cauchy sequence in H_0 that converges pointwise to f, showing that $f \in H$. Finally, we use this to show that $\|f - f_n\|_H \to 0$. Let $\epsilon > 0$ be arbitrary. From Step 1 we get that $\|f - g_n\|_H \to 0$, so we can choose $N_1 \in \mathbb{N}$ such that $\|f - g_n\|_H \leq \epsilon/2$. Furthermore, choose $N_2 \in \mathbb{N}$ such that for all $n \geq N_2$ we have $\epsilon_n \leq \epsilon/2$. We then get for all $n \geq \max\{N_1, N_2\}$ that

$$\|f - f_n\|_H \leq \|f - g_n\|_H + \|g_n - f_n\|_H \leq \epsilon.$$

Altogether, this establishes completeness of H.

Finally, since H is a Hilbert space with a reproducing kernel, it is an RKHS. □

A very important aspect of the developments in this section is that we constructed a completion of the pre Hilbert space (of functions) H_0 that is again a space of functions instead of just an abstract Hilbert space. One says that H is a *functional completion* of H_0.

Remark 2.5.7. It turns out that Theorem 2.5.6 is a special case of the following more general result. Let $\mathcal{X} \neq \emptyset$ be some set and $(H_0, \langle\cdot,\cdot\rangle_0)$ be a pre Hilbert space of functions on $\mathcal{X}$. Then the following two statements are equivalent.

1. There exists a Hilbert space $(H, \langle\cdot,\cdot\rangle_H)$ with a reproducing kernel such that $H_0 \subseteq H$ (as a sub pre Hilbert space)

2. For all $x \in \mathcal{X}$ the evaluation functionals $\delta_x : H_0 \ni f \mapsto f(x)$ are continuous in $(H_0, \langle\cdot,\cdot\rangle_0)$, and if a Cauchy sequence $(f_n)_n$ in H_0 converges pointwise to 0, then it also converges to 0 in norm.

For a proof see for example [32, Theorem 2].

2.6. Stochastic processes and RKHSs

We now turn to some connections between stochastic processes and RKHSs.

2.6.1. Kernels and covariance functions

In the following, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space in the background. Given some set $\mathcal{X} \neq \emptyset$, a *stochastic process with index set* $\mathcal{X}$ (or *on* $\mathcal{X}$) is a collection of random variables $\mathbb{Y} = (Y_x)_{x \in \mathcal{X}}$. We call $\mathbb{Y}$ a *locally square-integrable* or *second-order process* if Y_x is square-integrable for all $x \in \mathcal{X}$, i.e., $Y_x \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$. In this case, we can define the *mean function*

$$m_{\mathbb{Y}} : \mathcal{X} \to \mathbb{R}, \quad m_{\mathbb{Y}}(x) = \mathbb{E}[Y_x]$$

and the (centered) *covariance function*

$$k_{\mathbb{Y}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad k_{\mathbb{Y}}(x, x') = \operatorname{Cov}(Y_{x'}, Y_x) = \mathbb{E}[(Y_x - m(x))\overline{(Y_{x'} - m(x'))}]$$

of the stochastic process $\mathbb{Y}$.

Let $x_1, \ldots, x_N \in \mathcal{X}$, then the stochastic process $\mathbb{Y}$ induces a random vector $Y = \begin{pmatrix} Y_{x_1} & \cdots & Y_{x_N} \end{pmatrix}^\top$ with mean vector and covariance matrix (sometimes called variance matrix) given by

$$\mathbb{E}[Y] = \begin{pmatrix} m_{\mathbb{Y}}(x_1) & \cdots & m_{\mathbb{Y}}(x_N) \end{pmatrix}^\top$$

$$\operatorname{Var}[Y] = \begin{pmatrix} \operatorname{Cov}(Y_{x_1}, Y_{x_1}) & \cdots & \operatorname{Cov}(Y_{x_1}, Y_{x_N}) \\ \vdots & & \vdots \\ \operatorname{Cov}(Y_{x_N}, Y_{x_1}) & \cdots & \operatorname{Cov}(Y_{x_N}, Y_{x_N}) \end{pmatrix} = \begin{pmatrix} k_{\mathbb{Y}}(x_1, x_1) & \cdots & k_{\mathbb{Y}}(x_N, x_1) \\ \vdots & & \vdots \\ k_{\mathbb{Y}}(x_1, x_N) & \cdots & k_{\mathbb{Y}}(x_N, x_N) \end{pmatrix}.$$

As is well-known, a covariance matrix of a random vector is always positive semidefinite. But this means that $k_{\mathbb{Y}}$ is positive semidefinite according to Definition 2.4.3, so it is a kernel according to Theorem 2.5.3.

Let now $m : \mathcal{X} \to \mathbb{K}$ be any function and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a positive semidefinite function according to Definition 2.4.3, so a kernel. It is well-known that one can construct a stochastic process $\mathbb{Y}$ with index set $\mathcal{X}$, mean function $m_{\mathbb{Y}} = m$ and covariance function $k_{\mathbb{Y}} = k$. We can summarize all of this as follows.

Theorem 2.6.1. Let $\mathcal{X} \neq \emptyset$ be some set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a function. k is a kernel if and only if it is the covariance function of a second-order stochastic process with index set $\mathcal{X}$.

This result also justifies why in the context of second-order stochastic processes the terms *covariance function* and *kernel* are used interchangably, in particular, in machine learning [166].

2.6.2. RKHSs generated by stochastic processes

Let $\mathbb{Y} = (Y_x)_{x \in \mathcal{X}}$ be a second-order stochastic process. Instead of a collection of random variables, we can also interpret it as a random function. Formally, the random variables Y_x are (measurable) mappings $Y_x : \Omega \to \mathbb{R}$, and a realization $\omega \in \Omega$ of the underlying random experiment (described by $(\Omega, \mathcal{A}, \mathbb{P})$) induces a realization $x \mapsto Y_x(\omega)$ of the random function modeled by the stochastic process $\mathbb{Y}$. We can interpret $\mathcal{Y} = \{Y_x \mid x \in \mathcal{X}\}$ as the set of all "measurement points" of this random function. In signal processing, statistics and related fields, one often uses a *linear measurement model*, in which different "measurement points" are combined in a linear combination. The resulting linear measurements (or "linear measurement devices") are given by $\mathcal{H}_0 = \mathrm{span}\mathcal{Y}$. To work conveniently with the linear measurements (for example, for approximation arguments), it would be good if this set were closed under limits (in an appropriate sense). Since $\mathbb{Y}$ is a second-order process, we have by definition $\mathcal{Y} \subseteq \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, and a fortiori $\mathcal{H}_0 \subseteq \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, so it seems reasonable to turn $\mathcal{H}_0$ even into a Hilbert space. This is our next goal.

If $\mathcal{G}$ is a $\mathbb{K}$-Hilbert space and $\emptyset \neq \mathcal{F}_0 \subseteq \mathcal{G}$ a subset, we can turn the latter into a Hilbert space by setting $\mathcal{F} = \overline{\mathrm{span}\mathcal{F}_0}^{\|\cdot\|_{\mathcal{G}}}$. While we have $\mathcal{H}_0 \subseteq \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, the set of square-integrable random variables $\mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$ is *not* a Hilbert space. The reason is that in general if $f \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, then $\|f\|_{\mathcal{L}^2} = 0$ does not imply $f = 0$, but only $f = 0$ $\mathbb{P}$-a.s. As is well-known, this problem can be circumvented by going to the quotient space, a procedure we now recall. Let $\mathcal{N} = \{f : \Omega \to \mathbb{K} \mid f \text{ measurable}, f = 0\, \mathbb{P} - \text{a.s.}\}$, and define $L^2(\Omega, \mathcal{A}, \mathbb{P}) = \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})/\mathcal{N}$. Note that the elements of $L^2(\Omega, \mathcal{A}, \mathbb{P})$ are not functions, but rather equivalence classes of functions. For $f \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathbb{P})$, we write $[f]$ for the equivalence class of f, and by setting $\|[f]\|_{L^2(\Omega,\mathcal{A},\mathbb{P})} = \|f\|_{\mathcal{L}^2(\Omega,\mathcal{A},\mathbb{P})}$ we get a well-defined Hilbert space norm on $L^2(\Omega, \mathcal{A}, \mathbb{P})$.

We can now derive a Hilbert space from $\mathcal{Y}$. Define $H_0 = \operatorname{span}\{[Y_x] \mid x \in \mathcal{X}\}$, which is a subspace of the Hilbert space $L^2(\Omega, \mathcal{A}, \mathbb{P})$, so the closure $H = \overline{H_0}^{\|\cdot\|_{L^2(\Omega,\mathcal{A},\mathbb{P})}}$ is again a Hilbert space. By identifying Y_x with $[Y_x]$ for all $x \in \mathcal{X}$, we can interpret H as the set of all linear measurements and their limits (w.r.t. the topology of L^2) of the random function $\mathbb{Y}$. For this reason, the space H is called the *Hilbert space generated by the stochastic process* $\mathbb{Y}$.

Let us have a closer look at this space, in particular, its inner product. For simplicity, consider from now on a mean-zero stochastic process, i.e., $m_{\mathbb{Y}} \equiv 0$, and in order to reduce notational overhead, denote the covariance function of $\mathbb{Y}$ by k. Observe now that we have for all $x, x' \in \mathcal{X}$ that

$$\langle [Y_x], [Y_{x'}] \rangle_H = \langle Y_x, Y_{x'} \rangle_{\mathcal{L}^2} = \mathbb{E}[Y_x \overline{Y_{x'}}] = \operatorname{Cov}(Y_x, Y_{x'}) = k(x, x') = \overline{k(x', x)}$$

so H is *almost* a feature space, and $x \mapsto [Y_x]$ is *almost* a feature map for k, we only need to get rid of the complex conjugation. Tracing back our construction shows that we need to introduce a complex conjugation early on, so let us *redefine* $\mathcal{Y} = \{Y_x \mid x \in \mathcal{X}\}$ and $\mathcal{H}_0 = \operatorname{span}\mathcal{Y}$, as well as $H_0 = \operatorname{span}\{[\overline{Y_x}] \mid x \in \mathcal{X}\}$ and $H = \overline{H_0}^{\|\cdot\|_{L^2(\Omega,\mathcal{A},\mathbb{P})}}$. With this modification, we now have for all $x, x' \in \mathcal{X}$ that

$$\langle [\overline{Y_x}], [\overline{Y_{x'}}] \rangle_H = \langle \overline{Y_x}, \overline{Y_{x'}} \rangle_{\mathcal{L}^2} = \mathbb{E}\left[\overline{Y_x}\,\overline{\overline{Y_{x'}}}\right] = \mathbb{E}[Y_{x'} \overline{Y_x}] = \operatorname{Cov}(Y_{x'}, Y_x) = k(x', x),$$

so H is indeed a feature space of the kernel k, and $\Phi : \mathcal{X} \to H$, $\Phi(x) = [\overline{Y_x}]$ is a corresponding feature map. Recall from Theorem 2.3.7 that this implies $H_k = \operatorname{im}_H \Psi$, where $\Psi : H \to H_k$, $\Psi([Y]) = \langle [Y], \Phi(\cdot) \rangle_H$. This means that $f \in H_k$ if and only if there exists $[Y] \in H$ such that for all $x \in \mathcal{X}$

$$f(x) = \langle [Y], \Phi(x) \rangle_H = \langle [Y], [\overline{Y}_x] \rangle_{L^2} = E\left[Y \overline{\overline{Y_x}}\right] = \mathbb{E}[Y_x Y].$$

The Hilbert space generated by $\mathbb{Y}$ therefore allows a rather concrete description of the RKHS H_k. However, the connection between these two spaces is even stronger.

The canonical feature map Φ_k represents an input element $x \in \mathcal{X}$ by the RKHS function $k(\cdot, x)$. Similarly, the feature map Φ represents an input element x by $[\overline{Y_x}]$. We can therefore embed $\mathcal{Y}$ into H_k^{pre} by defining $I_{\mathcal{Y}} : \mathcal{Y} \to H_k^{\text{pre}}$, $I_{\mathcal{Y}}(\overline{Y_x}) = k(\cdot, x)$. In turn, this map can be extended to all of $\mathcal{H}_0$.

Lemma 2.6.2. Setting

$$I_{\mathcal{H}_0}\left(\sum_{n=1}^{N}\alpha_n\overline{Y_{x_n}}\right)=\sum_{n=1}^{N}\alpha_n k(\cdot,x_n)$$

for all $\alpha_1,\dots,\alpha_N\in\mathbb{K}$, $x_1,\dots,x_N\in\mathcal{X}$, leads to a well-defined linear and surjective map $I_{\mathcal{H}_0}:\mathcal{H}_0\to H_k^{\mathrm{pre}}$. Furthermore, for all $Y\in\mathcal{H}_0$ we have $\|I_{\mathcal{H}_0}(Y)\|_k=\|Y\|_{\mathcal{L}^2}$.

Proof. Let $\alpha_1,\dots,\alpha_N\in\mathbb{K}$, $x_1,\dots,x_N\in\mathcal{X}$ be arbitrary and observe that

$$\begin{aligned}
\left\|\sum_{n=1}^{N}\alpha_n\overline{Y_{x_n}}\right\|_{\mathcal{L}^2}^2 &= \mathbb{E}\left[\left(\sum_{n=1}^{N}\alpha_n\overline{Y_{x_n}}\right)\overline{\left(\sum_{n=1}^{N}\alpha_n\overline{Y_{x_n}}\right)}\right]\\
&= \sum_{i,j=1}^{N}\alpha_i\bar{\alpha}_j\mathbb{E}[\overline{Y_{x_i}}Y_{x_j}]\\
&= \sum_{i,j=1}^{N}\alpha_i\bar{\alpha}_j\mathrm{Cov}(Y_{x_j},Y_{x_i})=\sum_{i,j=1}^{N}\alpha_i\bar{\alpha}_j k(x_j,x_i)\\
&= \sum_{i,j=1}^{N}\alpha_i\bar{\alpha}_j\langle k(\cdot,x_i),k(\cdot,x_j)\rangle_k\\
&= \left\langle\sum_{i=1}^{N}\alpha_i k(\cdot,x_i),\sum_{j=1}^{N}\alpha_j k(\cdot,x_j)\right\rangle_k\\
&= \left\|\sum_{n=1}^{N}\alpha_n k(\cdot,x_n)\right\|_k^2.
\end{aligned}$$

Given $Y\in\mathcal{H}_0$ and two representations

$$Y=\sum_{n=1}^{N}\alpha_n\overline{Y_{x_n}}=\sum_{m=1}^{M}\beta_m\overline{Y_{x'_m}},$$

we therefore get

$$\begin{aligned}
\left\|\sum_{n=1}^{N}\alpha_n k(\cdot,x_n)-\sum_{m=1}^{M}\beta_m k(\cdot,x'_m)\right\|_k &= \left\|\sum_{n=1}^{N}\alpha_n\overline{Y_{x_n}}-\sum_{m=1}^{M}\beta_m\overline{Y_{x'_m}}\right\|_{\mathcal{L}^2}\\
&= \|Y-Y\|_{\mathcal{L}^2}=0,
\end{aligned}$$

which shows that $I_{\mathcal{H}_0}$ is well-defined, linear and $\|I_{\mathcal{H}_0}(Y)\|_k=\|Y\|_{\mathcal{L}^2}$ for all $Y\in$

$\mathcal{H}_0$. To show that $I_{\mathcal{H}_0}$ is surjective, let $f \in H_k^{\text{pre}}$ be arbitrary, and choose any representation $f = \sum_{n=1}^N \alpha_n k(\cdot, x_n)$. Defining $Y = \sum_{n=1}^N \alpha_n Y_{x_n} \in \mathcal{H}_0$, we then find $I_{\mathcal{H}_0}(Y) = f$. □

We now have a correspondence between $\mathcal{H}_0$ and H_k^{pre}. We can turn this into a map between H_0 and H_k^{pre} by defining $I_{H_0}([Y]) = I_{\mathcal{H}_0}(Y)$, where $Y \in \mathcal{H}_0$ is an element from the corresponding equivalence class.

Lemma 2.6.3. $I_{H_0} : H_0 \to H_k^{\text{pre}}$ is a well-defined, linear and bijective map. Furthermore, $\|I_{H_0}([Y])\|_k = \|[Y]\|_{H_0}$ for all $Y \in \mathcal{H}_0$, i.e., I_{H_0} is an isometry.

Proof. Let $Y_1, Y_2 \in \mathcal{H}_0$ with $[Y_1] = [Y_2]$, i.e., $Y_1 = Y_2$ $\mathbb{P}$-a.s., which implies $\|Y_1 - Y_2\|_{\mathcal{L}^2} = 0$. We therefore get

$$\|I_{H_0}([Y_1]) - I_{H_0}([Y_2])\|_k = \|I_{\mathcal{H}_0}(Y_1) - I_{\mathcal{H}_0}(Y_1)\|_k = \|Y_1 - Y_2\|_{\mathcal{L}^2} = 0,$$

which shows that I_{H_0} is well-defined and $\|I_{H_0}([Y])\|_k = \|Y\|_{\mathcal{L}^2} = \|[Y]\|_{L^2}$ for all $Y \in \mathcal{H}_0$. The linearity and surjectivity is inherited from $I_{\mathcal{H}_0}$, and since I_{H_0} is isometric and L^2 is a Hilbert space, the former is also injective. □

Since I_{H_0} is a linear isometry, it is continuous, and since H_0 is by definition dense in H, we can extend I_{H_0} uniquely to a linear isometry $I_H : H \to H_k$. In particular, I_H is injective and for all $[Y_1], [Y_2] \in H$ we have $\langle I_H([Y_1]), I_H([Y_2])\rangle_k = \langle [Y_1], [Y_2]\rangle_H$. Furthermore, for all $f \in H_k$ there exists $[Y] \in H$ with $f = \langle [Y], \Phi(\cdot)\rangle_H$, which implies that for all $x \in \mathcal{X}$

$$\begin{aligned} I_H([Y])(x) &= \langle I_H([Y]), k(\cdot, x)\rangle_k = \langle I_H([Y]), I_H([\overline{Y_x}])\rangle_k \\ &= \langle [Y], [\overline{Y_x}]\rangle_H = \langle [Y], \Phi(x)\rangle_H = f(x), \end{aligned}$$

i.e., $I_H([Y]) = f$. This shows that I_H is also surjective. Altogether, we arrived at the following result, which is known as *Loeve's representation theorem*, and the map I_H is sometimes called the *canonical isomorphism*.

Theorem 2.6.4. The Hilbert spaces H generated by the stochastic process $\mathbb{Y}$ is isometrically isomorphic to the RKHS H_k. In particular, for all $f \in H_k$ there exists $[Y] \in H$ such that $f(x) = \mathbb{E}[YY_x]$ for all $x \in \mathcal{X}$.

We can therefore identify the Hilbert generated by a zero-mean second-order stochastic process with the RKHS corresponding to the covariance function of this stochastic process. This also means that by considering the space of all linear measurements (and their limits) of such a stochastic process, we arrived again at the concept of an RKHS.

2.7. Comments

As stated in the beginning of this chapter, no new results or examples are contained in our presentation, and we heavily relied on existing expositions of RKHSs and related concepts, in particular, [32], [152], and [189, Chapter 4]. However, we are not aware of a similar perspective-agnostic introduction to RKHSs. This chapter, which is based on an early version of the manuscript [CF14], has been conceived and written by the author of the present thesis, with some editorial input by the first supervisor S. Trimpe.

3. Lipschitz and Hölder continuity in RKHSs

RKHSs are functions spaces that are generated by a special bivariate function, their reproducing kernel, cf. Section 2.2. As is well-known, properties of the functions in an RKHS and properties of the reproducing kernel of an RKHS are closely connected, and these connections have been thoroughly investigated, with a good overview provided in [189, Chapter 4]. This connection is important since an RKHS is generated by its reproducing kernel, and the latter is user-defined in most applications of RKHSs. By choosing or constructing an appropriate reproducing kernel, tailored function spaces can be created, which can then be used in interpolation, approximation, optimization and related problems.

Particularily relevant for many applications are regularity properties of function spaces. In the case of RKHSs, continuity and differentiability of functions is fully determined by the corresponding reproducing kernel, cf. [189, Lemma 4.29, Corollary 4.36]. Furthermore, there is a close connection between certain Sobolev spaces and RKHSs, cf. [217, 71]. Another important regularity notion, which is in between mere continuity and differentiability, is Lipschitz continuity, or more generally Hölder continuity. Recall that if $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ are two metric spaces, and $f : \mathcal{X} \to \mathcal{Y}$ a function, we call f *Lipschitz continuous* if there exists $L \in \mathbb{R}_{\geq 0}$ such that for all $x, x' \in \mathcal{X}$ we have $d_\mathcal{Y}(f(x), f(x')) \leq L d_\mathcal{X}(x, x')$. Each such $L \in \mathbb{R}_{\geq 0}$ is called a *Lipschitz constant* for f, we sometimes we say that f is L-Lipschitz continuous. Similarly, if there exists $\alpha \in \mathbb{R}_{>0}$ and $L_\alpha \in \mathbb{R}_{\geq 0}$ such that for all $x, x' \in \mathcal{X}$ we have $d_\mathcal{Y}(f(x), f(x')) \leq L_\alpha d_\mathcal{X}(x, x')^\alpha$, then f is called *α-Hölder continuous*, and each such L_α is called a *Hölder constant* for f. In particular, 1-Hölder continuity is Lipschitz continuity.

Lipschitz and Hölder continuity are classic notions that appear prominently for example in the theory of ordinary differential equations [11] and partial differential

equations [69], respectively. Hölder continuity is also frequently used in the theory of nonparametric statistics [203, 80]. Moreover, there is now a considerable and well-developed theory of spaces of Lipschitz continuous functions, cf. [56]. In addition, Lipschitz continuity (and to a lesser extent also Hölder continuity) is used as the foundation of practical algorithms. For example, Lipschitz continuity (and a known Lipschitz constant) is a core assumption in many global optimization approaches [161]. Lipschitz continuity also forms the basis for many non-stochastic learning algorithms, especially in the context of systems identification [139, 44]. Finally, Lipschitz continuity (and related properties) will also play a role in the latter parts of the present thesis, in the context of safe learning and reasonable practical prior knowledge.

All of this forms a strong motivation to investigate Lipschitz and Hölder continuity in RKHSs. In particular, a central question is how (if at all) the Lipschitz or Hölder continuity of the reproducing kernel of an RKHS influences the corresponding continuity properties of RKHS functions. To the best of our knowledge, there is no systematic investigation into these questions, despite the importance of RKHSs and Lipschitz and Hölder continuity, respectively, and the considerable effort that went into investigating the connection between kernel properties and RKHS function properties. That RKHS functions are always Lipschitz continuous w.r.t. the kernel metric, as reviewed in Section 3.2, is well-known. The more interesting question of Lipschitz and Hölder continuity w.r.t. an arbitrary metric seems to have been barely covered in the literature, and the only previous work we are aware of that explicitly addresses this question, is [72]. In this chapter, close this gap in the literature, and also provide context and background for the Lipschitz-based methods presented in later chapters.

Apart from minor modifications, this chapter corresponds to the preprint [CF1]. Detailed comments on the author's contribution are provided in Section 3.6.

3.1. Preliminaries and background

We cover the real and complex case simultaneously, using the symbol $\mathbb{K}$ for $\mathbb{R}$ or $\mathbb{C}$. Unless noted otherwise, $\mathcal{X}$ will be a non-empty set. We call $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ *Hermitian* if for all $x, x' \in \mathcal{X}$, we have $\kappa(x, x') = \overline{\kappa(x', x)}$. Note that if κ is Hermitian, then $\kappa(x, x) \in \mathbb{R}$ for all $x \in \mathcal{X}$. If $\mathbb{K} = \mathbb{R}$, then κ is Hermitian if and only if it is

symmetric in its two arguments.

Since we state several results for bounded kernels or bounded RKHS functions, we recall the following characterization of boundedness in RKHSs.

Lemma 3.1.1. Let $\mathcal{X} \neq \emptyset$ be some set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a kernel on $\mathcal{X}$. The following statements are equivalent.

1. k is bounded
2. $\|k\|_\infty := \sup_{x\in\mathcal{X}} \sqrt{k(x,x)} < \infty$
3. There exists a feature space-feature map pair $(\mathcal{H}, \Phi)$ such that Φ is bounded
4. For all feature space-feature map pairs $(\mathcal{H}, \Phi)$, Φ is bounded
5. All $f \in H_k$ are bounded

If any of the statements is true, then for all feature space-feature map pairs $(\mathcal{H}, \Phi)$, we have $\|k\|_\infty = \sup_{x\in\mathcal{X}} \|\Phi(x)\|_\mathcal{H}$, and $|f(x)| \leq \|f\|_k \|k\|_\infty$, for all $f \in H_k$ and $x \in \mathcal{X}$.

Proof. Let $(\mathcal{H}, \Phi)$ be any feature space-feature map. For $x, x' \in \mathcal{X}$ we have

$$|k(x,x')| = |\langle \Phi(x'), \Phi(x)\rangle_\mathcal{H}| \leq \|\Phi(x')\|_\mathcal{H} \|\Phi(x)\|_\mathcal{H} = \sqrt{k(x',x')}\sqrt{k(x,x)},$$

and the equivalence of the first four items is now clear. The equivalence between the first and last item is provided by [189, Lemma 4.23].

Finally, since for any feature space-feature map pair $(\mathcal{H}, \Phi)$, and all $x \in \mathcal{X}$, we have $\sqrt{k(x,x)} = \|\Phi(x)\|_\mathcal{H}$, and for all $f \in H_k$ we have $|f(x)| = |\langle f, k(\cdot, x)\rangle_k| \leq \|f\|_k \sqrt{k(x,x)}$, the last assertion follows. □

Finally, we recall the following result on Parseval frames in an RKHS, which corresponds to [152, Theorem 2.10, Exercise 3.7], and is called *Papadakis Theorem* there.

Theorem 3.1.2. Let $\mathcal{X} \neq \emptyset$ be a set and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a kernel on $\mathcal{X}$.

1. If $(f_i)_{i\in I}$ is a Parseval frame in H_k, then for all $x, x' \in \mathcal{X}$

$$k(x,x') = \sum_{i\in I} f_i(x)\overline{f_i(x')}, \tag{3.1}$$

where the convergence is pointwise.

2. Consider a family of functions $(f_i)_{i\in I}$, where $f_i \in \mathbb{K}^{\mathcal{X}}$ for all $i \in I$, such that

$$k(x,x') = \sum_{i\in I} f_i(x)\overline{f_i(x')} \tag{3.2}$$

for all $x, x' \in \mathcal{X}$, where the convergence is pointwise. Then $f_i \in H_k$ for all $i \in I$, and $(f_i)_{i\in I}$ is a Parseval frame in H_k.

3.2. Lipschitz continuity and the kernel metric

Let k be a kernel on $\mathcal{X}$, and $(\mathcal{H}, \Phi)$ a corresponding feature space-feature map pair, then

$$d_\Phi : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_{\geq 0}, \quad d_\Phi(x,x') = \|\Phi(x) - \Phi(x')\|_{\mathcal{H}} \tag{3.3}$$

is a semimetric on $\mathcal{X}$. If $(\mathcal{H},\Phi) = (H_k, \Phi_k)$, we set $d_k = d_{\Phi_k}$ and call this the *kernel (semi)metric*. Note that this holds for *any* set $\mathcal{X}$, no matter whether it has additional structure on it or not.

The next result is well-known, but rarely explicitly stated.

Lemma 3.2.1. Let $k : \mathcal{X}\times\mathcal{X} \to \mathbb{K}$ be a kernel on $\mathcal{X} \neq \emptyset$. Then for all feature space-feature map pairs $(\mathcal{H},\Phi)$, we have $d_\Phi = d_k$.

When working with d_k, this result allows us to work with d_Φ instead, where Φ is any feature map, and vice versa.

Proof. Let $(\mathcal{H},\Phi)$ be a feature space-feature map pair, and $x, x' \in \mathcal{X}$ be arbitrary. We then have

$$\begin{aligned} d_\Phi(x,x') &= \|\Phi(x) - \Phi(x')\|_{\mathcal{H}} = \sqrt{\langle \Phi(x) - \Phi(x'), \Phi(x) - \Phi(x')\rangle_{\mathcal{H}}} \\ &= \sqrt{\langle\Phi(x),\Phi(x)\rangle_{\mathcal{H}} + \langle\Phi(x),\Phi(x')\rangle_{\mathcal{H}} + \langle\Phi(x'),\Phi(x)\rangle_{\mathcal{H}} + \langle\Phi(x'),\Phi(x')\rangle_{\mathcal{H}}} \\ &= \sqrt{k(x,x) + k(x,x') + k(x',x) + k(x',x')} \\ &= \sqrt{\langle k(\cdot,x) - k(\cdot,x'), k(\cdot,x) - k(\cdot,x')\rangle_k} = \|\Phi_k(x) - \Phi_k(x')\|_k = d_k(x,x'), \end{aligned}$$

establishing the claim. □

It is therefore natural to investigate Lipschitz continuity of RKHS functions w.r.t. the kernel metric. We start with the following classic result, which seems to be folklore.

Proposition 3.2.2. Let $\mathcal{X} \neq \emptyset$ be some set, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a kernel on $\mathcal{X}$, and d_k the corresponding kernel (semi)metric. For all $f \in H_k$, we have that f is Lipschitz continuous w.r.t. d_k with Lipschitz constant $\|f\|_k$.

In other words, RKHS functions are always Lipschitz continuous w.r.t. the kernel (semi)metric, and their RKHS norm is a Lipschitz constant. This reinforces the intuition that the RKHS norm is a measure of complexity or smoothness of an RKHS function w.r.t. a kernel: The smaller the RKHS norm, the smaller the Lipschitz bound of an RKHS function w.r.t. to the kernel (semi)metric.

Proof. Let $f \in H_k$, then we have

$$|f(x) - f(x')| = |\langle f, k(\cdot, x) - k(\cdot, x')\rangle_k| \leq \|f\|_k \|k(\cdot, x) - k(\cdot, x')\|_k = \|f\|_k d_k(x, x')$$

for all $x, x' \in \mathcal{X}$. □

The next result seems to be less well-known. Parts of it can be found for example in [9, Proposition 2.4].

Proposition 3.2.3. Let $\mathcal{X} \neq \emptyset$ be some set, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a kernel on $\mathcal{X}$.

1. The function $k(\cdot, x) \in H_k$ is Lipschitz continuous w.r.t. d_k with Lipschitz constant $\sqrt{k(x, x)}$, for all $x \in \mathcal{X}$.
2. For all $x_1, x_1', x_2, x_2' \in \mathcal{X}$, $|k(x_1, x_2) - k(x_1', x_2')| \leq B(d_k(x_1, x_1') + d_k(x_2, x_2'))$ with

   $$B = \min\left\{\max\left\{\sqrt{k(x_2, x_2)}, \sqrt{k(x_1', x_1')}\right\}, \max\left\{\sqrt{k(x_1, x_1)}, \sqrt{k(x_2', x_2')}\right\}\right\}$$

   If k is bounded, then it is Lipschitz continuous w.r.t. the product metric on $\mathcal{X} \times \mathcal{X}$ with Lipschitz constant $\|k\|_\infty$.
3. For all $x, x' \in \mathcal{X}$,

   $$|k(x, x) - k(x', x')| \leq 2 \max\{\sqrt{k(x, x)}, \sqrt{k(x', x')}\} d_k(x, x'). \tag{3.4}$$

If k is bounded, then $x \mapsto k(x,x)$ is Lipschitz continuous w.r.t. d_k with Lipschitz constant $2\|k\|_\infty$.

4. The function $x \mapsto \sqrt{k(x,x)}$ is Lipschitz continuous w.r.t. d_k and 1 is a Lipschitz constant.

5. If $(\mathcal{H}, \Phi)$ is any feature space-feature map-pair, then Φ is Lipschitz continuous w.r.t. d_k with Lipschitz constant 1.

Proof. The first item follows immediately from Proposition 3.2.2 clear since $\|k(\cdot,x)\|_k = \sqrt{k(x,x)}$.

To show the second item, let $x_1, x_1', x_2, x_2' \in \mathcal{X}$, then

$$\begin{aligned}
|k(x_1,x_2) - k(x_1',x_2')| &\leq ||k(x_1,x_2) - k(x_1',x_2)| + |k(x_1',x_2) - k(x_1',x_2')| \\
&= |k(x_1,x_2) - k(x_1',x_2)| + |k(x_2,x_1') - k(x_2',x_1')| \\
&\leq \sqrt{k(x_2,x_2)} d_k(x_1,x_1') + \sqrt{k(x_1',x_1')} d_k(x_2,x_2') \\
&\leq \max\left\{\sqrt{k(x_2,x_2)}, \sqrt{k(x_1',x_1')}\right\} (d_k(x_1,x_1') + d_k(x_2,x_2')).
\end{aligned}$$

Repeating this computation with x_1, x_2' instead of x_2, x_1' establishes the claim.

The next item is now an immediate consequence.

For the second to last item, let $x, x' \in \mathcal{X}$, then the converse triangle inequality (in H_k) leads to

$$|\sqrt{k(x,x)} - \sqrt{k(x',x')}| = |\|k(\cdot,x)\|_k - \|k(\cdot,x')\|_k| \leq \|k(\cdot,x) - k(\cdot,x')\| = d_k(x,x'),$$

so $x \mapsto \sqrt{k(x,x)}$ is indeed 1-Lipschitz w.r.t. d_k.

The last item is clear. □

3.3. Lipschitz and Hölder continuity on metric spaces

As we recalled in the preceding section, RKHS functions are always Lipschitz continuous w.r.t. the kernel (semi)metric. However, this metric is in general independent of any additional structure on the input set. In particular, if the input set is already a metric space, then this structure is essentially ignored by the kernel (semi)metric. In many applications, we are given a metric space as input set, and we would like to

have Lipschitz or Hölder continuity of RKHS functions w.r.t. to the existing metric on the input space. We will now investigate this question in depth.

3.3.1. Preliminaries

Since kernels are special bivariate functions, we present some preliminary material on Hölder and Lipschitz continuity of general functions of two variables. Everything in this subsection is elementary and probably known, but we could not locate explicit references, hence we provide all the details.

Let $(\mathcal{X}, d_\mathcal{X})$ be a metric space and $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ some function.

Lemma 3.3.1. Assume that there exist a constant $\alpha \in \mathbb{R}_{>0}$, some function $L_\alpha : \mathcal{X} \to \mathbb{R}_{\geq 0}$, and for all $x \in \mathcal{X}$ a set $U_x \subseteq \mathcal{X}$ with $x \in U_x$, such that for all $x_1, x_1', x_2, x_2' \in \mathcal{X}$ we have

$$|\kappa(x_1, x_2) - \kappa(x_1', x_2')| \leq L_\alpha(x)(d_\mathcal{X}(x_1, x_1')^\alpha + d_\mathcal{X}(x_2, x_2')^\alpha). \tag{3.5}$$

1. For all $x_2 \in \mathcal{X}$ and all $x_1, x_1' \in U_{x_2}$, we have that

$$|\kappa(x_1, x_2) - \kappa(x_1', x_2)| \leq L_\alpha(x) d_\mathcal{X}(x_1, x_1')^\alpha. \tag{3.6}$$

2. Assume furthermore that κ is Hermitian. We then have for all $x \in \mathcal{X}$ and $x' \in U_x$ with $x \in U_{x'}$ that

$$|\kappa(x) - \kappa(x')| \leq (L_\alpha(x) + L_\alpha(x')) d_\mathcal{X}(x, x')^\alpha, \tag{3.7}$$

   where we defined $\kappa(x) := \kappa(x, x)$.

Proof. The first claim is trivial. For the second, let $x \in \mathcal{X}$ and $x' \in U_x$ be arbitrary, then we have

$$\begin{aligned}
|\kappa(x) - \kappa(x')| &= |\kappa(x, x) - \kappa(x', x')| \\
&\leq |\kappa(x, x) - \kappa(x', x)| + |\kappa(x', x) - \kappa(x', x')| \\
&= |\kappa(x, x) - \kappa(x', x)| + |\kappa(x, x') - \kappa(x', x')| \\
&\leq (L_\alpha(x) + L_\alpha(x')) d_\mathcal{X}(x, x')^\alpha,
\end{aligned}$$

where we used $|\kappa(x', x) - \kappa(x', x')| = |\overline{\kappa(x, x')} - \overline{\kappa(x', x')}| = |\kappa(x, x') - \kappa(x', x')|$ in the second equality. □

Lemma 3.3.2. Assume that there exist a constant $\alpha \in \mathbb{R}_{>0}$, some function $L_\alpha : \mathcal{X} \to \mathbb{R}_{\geq 0}$, and for all $x \in \mathcal{X}$ a set $U_x \subseteq \mathcal{X}$ with $x \in U_x$, such that for all $x_1, x_1' \in \mathcal{X}$ we have

$$|\kappa(x_1, x) - \kappa(x_1', x)| \leq L_\alpha(x) d_{\mathcal{X}}(x_1, x_1')^\alpha. \tag{3.8}$$

If κ is Hermitian, then we have for all $x_1, x_1', x_2, x_2' \in \mathcal{X}$ with $x_1, x_1' \in U_{x_2}$ and $x_2, x_2' \in U_{x_1'}$ that

$$|\kappa(x_1, x_2) - \kappa(x_1', x_2')| \leq L_\alpha(x_2) d_{\mathcal{X}}(x_1, x_1')^\alpha + L_\alpha(x_1') d_{\mathcal{X}}(x_2, x_2')^\alpha. \tag{3.9}$$

Proof. Let $x_1, x_1', x_2, x_2' \in \mathcal{X}$ such that $x_1, x_1' \in U_{x_2}$ and $x_2, x_2' \in U_{x_1'}$, then we get

$$\begin{aligned}
|\kappa(x_1, x_2) - \kappa(x_1', x_2')| &\leq |\kappa(x_1, x_2) - \kappa(x_1', x_2)| + |\kappa(x_1', x_2) - \kappa(x_1', x_2')| \\
&= |\kappa(x_1, x_2) - \kappa(x_1', x_2)| + |\overline{\kappa(x_2, x_1')} - \overline{\kappa(x_2', x_1')}| \\
&= |\kappa(x_1, x_2) - \kappa(x_1', x_2)| + |\kappa(x_2, x_1') - \kappa(x_2', x_1')| \\
&\leq L_\alpha(x_2) d_{\mathcal{X}}(x_1, x_1')^\alpha + L_\alpha(x_1') d_{\mathcal{X}}(x_2, x_2')^\alpha.
\end{aligned}$$

□

We now consider the special case of Lipschitz continuity, corresponding to $\alpha = 1$ in the preceding results.

Definition 3.3.3. We call κ *Lipschitz continuous in the first argument with Lipschitz constant* $L \in \mathbb{R}_{\geq 0}$, or *$L$-Lipschitz continuous in the first argument*, if for all $x_1, x_1', x_2 \in \mathcal{X}$ we have

$$|\kappa(x_1, x_2) - \kappa(x_1', x_2)| \leq L d_{\mathcal{X}}(x_1, x_1'). \tag{3.10}$$

Similarly, we define L-Lipschitz-continuity in the second argument. Finally, we call κ *separately L-Lipschitz continuous* if it is L-Lipschitz continuous in the first and the second coordinate.

Proposition 3.3.4. Let κ be Hermitian, then the following statements are equivalent.

1. κ is L-Lipschitz continuous (w.r.t. the product metric on $\mathcal{X} \times \mathcal{X}$)
2. κ is L-Lipschitz continuous in the first argument
3. κ is L-Lipschitz continuous in the second argument
4. κ is separately L-Lipschitz continuous

Proof. By definition, if κ is separately L-Lipschitz continuous, it is L-Lipschitz continuous in the first and second argument. Since κ is Hermitian, the equivalence of items 2 and 3 are clear, so any one these two items implies the fourth item. Lemma 3.3.1 shows that item 1 implies item 4. Finally, Lemma 3.3.2 shows that item 2 implies item 1. □

Since kernels are always Hermitian, Proposition 3.3.4 immediately leads to the following result.

Corollary 3.3.5. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ be a kernel, and $L \in \mathbb{R}_{\geq 0}$. k is L-Lipschitz continuous if and only if it is separately L-Lipschitz continuous.

Why is Corollary 3.3.5 interesting? Let $\mathcal{X}$ be a topological space and k a kernel on $\mathcal{X}$. It is well-known that k is continuous if and only if it is separately continuous, i.e., $k(\cdot, x)$ is continuous for all $x \in \mathcal{X}$, and $x \mapsto k(x, x)$ is continuous, cf. [189, Lemma 4.29]. In particular, separate continuity of k is not enough for k to be continuous. For example, there exists a kernel on $\mathcal{X} = [-1, 1]$ that is bounded and separately continuous, but not continuous, cf. [118]. Corollary 3.3.5 asserts that in contrast to continuity, *Lipschitz continuity* is equivalent to separate Lipschitz continuity for kernels.

3.3.2. RKHS functions of Hölder-continuous kernels

We now investigate how Hölder continuity of the kernel induces Hölder continuity of RKHS functions. We start with the following very general result, which covers essentially all potentially relevant forms of Lipschitz and Hölder continuity. It is a generalization of [72, Proposition 5.2].

Theorem 3.3.6. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a kernel. Let $\alpha \in \mathbb{R}_{>0}$ and assume that there exist a function $L_\alpha : \mathcal{X} \to \mathbb{R}_{\geq 0}$ and for each $x \in \mathcal{X}$

a set $U_x \subseteq \mathcal{X}$ with $x \in U_x$, such that for all $x_1, x_1' \in U_x$ we have

$$|k(x_1, x) - k(x_1', x)| \leq L_\alpha(x) d_\mathcal{X}(x_1, x_1')^\alpha. \tag{3.11}$$

1. Let $(\mathcal{H}, \Phi)$ be an arbitrary feature space-feature map-pair for k. For all $x, x' \in \mathcal{X}$ with $x' \in U_x$ we have

$$\|\Phi(x) - \Phi(x')\|_\mathcal{H} \leq \sqrt{2L_\alpha(x)} d_\mathcal{X}(x, x')^{\frac{\alpha}{2}}. \tag{3.12}$$

2. For all $f \in H_k$ and $x, x' \in \mathcal{X}$ with $x' \in U_x$ we have

$$|f(x) - f(x')| \leq \sqrt{2L_\alpha(x)} \|f\|_k d_\mathcal{X}(x, x')^{\frac{\alpha}{2}}. \tag{3.13}$$

Proof. Let $x, x' \in \mathcal{X}$ with $x' \in U_x$ be arbitrary. If $(\mathcal{H}, \Phi)$ is a feature space-feature map-pair for k, then we get

$$\begin{aligned}
\|\Phi(x) - \Phi(x')\|_\mathcal{H} &= d_\Phi(x, x') = d_k(x, x') \\
&= \sqrt{k(x,x) + k(x',x') - k(x,x') - k(x',x)} \\
&\leq \sqrt{|k(x,x) - k(x',x)| + |k(x,x') - k(x',x')|} \\
&\leq \sqrt{2L_\alpha(x) d_\mathcal{X}(x,x')^\alpha},
\end{aligned}$$

where we used in the last inequality that $x' \in U_x$.

Let now $f \in H_k$, then we have

$$\begin{aligned}
|f(x) - f(x')| &\leq \|f\|_k \|k(\cdot, x) - k(\cdot, x')\|_k \\
&\leq \sqrt{2L_\alpha(x)} \|f\|_k d_\mathcal{X}(x, x')^{\frac{\alpha}{2}},
\end{aligned}$$

where we used that (H_k, Φ_k) is a feature space-feature map-pair for k. □

For convenience, we record the following special case.

Corollary 3.3.7. Let $(\mathcal{X}, d_\mathcal{X})$ be a metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a kernel that is separately L-Lipschitz continuous, then for every $f \in H_k$ and $x, x' \in \mathcal{X}$ we have

$$|f(x) - f(x')| \leq \sqrt{2L} \sqrt{d_\mathcal{X}(x, x')}. \tag{3.14}$$

Remark 3.3.8. Consider the situation of Theorem 3.3.6.

1. If $\alpha \in (0,1)$, $\delta \in \mathbb{R}_{>0}$, $U_x = \mathcal{B}_\delta(x)$ and $L_\alpha \equiv L_k$ for some $L_k \in \mathbb{R}_{\geq 0}$, then we recover [72, Proposition 5.2].

2. If $\alpha \in (0,1)$, $U_x = \mathcal{X}$ for all $x \in \mathcal{X}$, $L_\alpha \equiv L_k$ for some $L_k \in \mathbb{R}_{\geq 0}$, then we get that for $f \in H_k$ and $x, x' \in \mathcal{X}$ that

$$|f(x) - f(x')| \leq \sqrt{2L_k}\|f\|_k d_{\mathcal{X}}(x,x')^{\frac{\alpha}{2}}$$

   We can describe this as "A separately α-Hölder continuous kernel leads to RKHS functions that are $\alpha/2$-Hölder continuous".

3.3.3. Converse results

In Section 3.2 we saw that every RKHS function $f \in H_k$ is Lipschitz continuous w.r.t. d_k with Lipschitz constant $\|f\|_k$. Furthermore, in Section 3.3 results were presented that ensure that RKHS functions are Hölder continuous w.r.t. a given metric on the input set, if the kernel fulfills a certain continuity condition. But what about the converse? Assume we have a Hilbert function space H such that all $f \in H$ are Lipschitz continuous (or Hölder continous) w.r.t. a given metric and Lipschitz (or Hölder) constant $\|f\|_H$. What can we say about H? And if H is an RKHS, what can we say about the kernel? To the best of our knowledge, these questions have not been addressed so far.

In this subsection, let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and $H \subseteq \mathbb{K}^{\mathcal{X}}$ a Hilbert space of functions.

Assumption 3.3.9. There exists $\alpha \in \mathbb{R}_{>0}$ such that all $f \in H$ are α-Hölder continuous with Hölder constant $\|f\|_H$.

Proposition 3.3.10. Suppose Assumption 3.3.9 holds, and that H is an RKHS. Furthermore, let k be the uniquely determined kernel with $H_k = H$.

1. For all $x \in \mathcal{X}$, $k(\cdot, x) \in H$ is α-Hölder continuous with Hölder constant $\sqrt{k(x,x)}$. If k is bounded, then $k(\cdot, x)$ is α-Hölder continuous with Hölder constant $\|k\|_\infty$, for all $x \in \mathcal{X}$.

2. For all $x_1, x_1', x_2, x_2' \in \mathcal{X}$, $|k(x_1, x_2) - k(x_1', x_2')| \leq B(d_{\mathcal{X}}(x_1, x_1')^\alpha + d_{\mathcal{X}}(x_2, x_2')^\alpha)$ with

$$B = \min\left\{\max\{\sqrt{k(x_2, x_2)}, \sqrt{k(x_1', x_1')}\}, \max\{\sqrt{k(x_1, x_1)}, \sqrt{k(x_2', x_2')}\}, \right\}$$

If k is bounded, then

$$|k(x_1, x_2) - k(x_1', x_2')| \leq \|k\|_\infty (d_{\mathcal{X}}(x_1, x_1')^\alpha + d_{\mathcal{X}}(x_2, x_2')^\alpha) \tag{3.15}$$

for all $x_1, x_1', x_2, x_2' \in \mathcal{X}$.

3. For all $x, x' \in \mathcal{X}$,

$$d_k(x, x') \leq \sqrt{\sqrt{k(x, x)} + \sqrt{k(x', x')}} d(x, x')^{\frac{\alpha}{2}}. \tag{3.16}$$

If k is bounded, then

$$d_k(x, x') \leq \sqrt{2\|k\|_\infty} d(x, x')^{\frac{\alpha}{2}}. \tag{3.17}$$

4. If $(\mathcal{H}, \Phi)$ is any feature space-feature map-pair, and k is bounded, then Φ is $\frac{\alpha}{2}$-Hölder continuous with Hölder constant $\sqrt{2\|k\|_\infty}$.

Proof. The first claim follows immediately from Assumption 3.3.9 and the fact that $\|k(\cdot, x)\|_k = \sqrt{k(x, x)}$ for all $x \in \mathcal{X}$, and the definition of $\|k\|_\infty$.

Let $x_1, x_1', x_2, x_2' \in \mathcal{X}$ be arbitrary. Using Lemma 3.3.2 leads to

$$\begin{aligned} |k(x_1, x_2) - k(x_1', x_2')| &\leq \sqrt{k(x_2, x_2)} d_{\mathcal{X}}(x_1, x_1') + \sqrt{k(x_1', x_1')} d_{\mathcal{X}}(x_2, x_2') \\ &\leq \max\left\{\sqrt{k(x_2, x_2)}, \sqrt{k(x_1', x_1')}\right\}, \end{aligned}$$

and repeating this computing with x_1, x_2' instead of x_2, x_1' establishes the second

assertion. Additionally,

$$\begin{aligned} d_k(x,x') &= \sqrt{k(x,x) - k(x,x') - k(x',x) + k(x',x')} \\ &\leq \sqrt{|k(x,x) - k(x',x)| + |k(x,x') - k(x',x')|} \\ &\leq \sqrt{\sqrt{k(x,x)} + \sqrt{k(x',x')} d_{\mathcal{X}}(x,x')^{\alpha}}, \end{aligned}$$

showing the third claim. This also establishes the last assertion, since for any feature space-feature map pair $(\mathcal{H}, \Phi)$ and all $x, x' \in \mathcal{X}$ we have $\|\Phi(x) - \Phi(x')\|_{\mathcal{H}} = d_k(x,x')$. □

Corollary 3.3.11. Assume that all $f \in H$ are Lipschitz continuous with Lipschitz constant $\|f\|_H$, that H is an RKHS, and that the uniquely determined kernel k with $H_k = H$ is bounded. Then k is Lipschitz continuous with Lipschitz constant $\|k\|_\infty$.

The following result provides a simple condition for H to be an RKHS, if H fulfills Assumption 3.3.9.

Proposition 3.3.12. Suppose Assumption 3.3.9 holds, and that there exists $x_0 \in \mathcal{X}$ such that $f(x_0) = 0$ for all $f \in H$. In this case, H is an RKHS. Furthermore, $\sqrt{k(x,x)} \leq d_{\mathcal{X}}(x, x_0)$ for all $x \in \mathcal{X}$, where k is the uniquely determined reproducing kernel of H.

Proof. Let $x \in \mathcal{X}$ and consider the corresponding evaluation functional $\delta_x : H \to \mathbb{K}$, $\delta_x f = f(x)$. We then have for all $f \in \mathcal{X}$ that

$$|\delta_x f| = |f(x)| = |f(x) - f(x_0)| \leq \|f\|_H d_{\mathcal{X}}(x, x_0),$$

which shows that δ_x is continuous, and $\|\delta_x\| \leq d_{\mathcal{X}}(x, x_0)$. Therefore, H is an RKHS. Let k be its uniquely determined reproducing kernel, then

$$\sqrt{k(x,x)} = \|k(\cdot, x)\|_H = \|\delta_x\| \leq d(x, x_0),$$

since $k(\cdot, x)$ is the uniquely determined Riesz representer of δ_x in H. □

Combining Proposition 3.3.12 with Lemma 3.1.1 leads to the following result.

Corollary 3.3.13. Assume that all $f \in H$ are bounded and Lipschitz continuous with Lispchitz constant $\|f\|_H$. Then H is an RKHS with a bounded and Lipschitz continuous kernel k having Lipschitz constant $\|k\|_\infty$.

In RKHSs, Assumption 3.3.9 can be relaxed.

Lemma 3.3.14. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ be a kernel and H_k its RKHS. Let $D \subseteq H_k$ be dense, and assume that there exists $\alpha \in \mathbb{R}_{>0}$ such that all $f \in D$ are α-Hölder continuous w.r.t. $d_\mathcal{X}$ with Hölder bound $\|f\|_k$. Then all $f \in H_k$ are α-Hölder continuous with Hölder bound $\|f\|_k$.

Proof. Let $f \in H_k$ and $x, x' \in \mathcal{X}$ be arbitrary. Since D is dense in H_k, there exists $(f_n)_{n\in\mathbb{N}_+} \subseteq D$ such that $f_n \to f$ (in H_k). We then have

$$\begin{aligned}
|f(x) - f(x')| &= |\langle f, k(\cdot, x) - k(\cdot, x')\rangle_k| = |\langle \lim_{n\to\infty} f_n, k(\cdot, x) - k(\cdot, x')\rangle_k| \\
&= \lim_{n\to\infty} |\langle f_n, k(\cdot, x) - k(\cdot, x')\rangle_k| = \lim_{n\to\infty} |f_n(x) - f_n(x')| \\
&\leq \lim_{n\to\infty} \|f_n\|_k d(x, x')^\alpha = \|f\|_k d(x, x')^\alpha.
\end{aligned}$$

□

Finally, under an additional assumption on $d_\mathcal{X}$, Assumption 3.3.9 implies the existence of an RKHS on H. The construction is classical, cf. [17, Chapter I], but has not been used in this context before.

Suppose that Assumption 3.3.9 holds and that $d_\mathcal{X}$ is a *Hilbertian metric*, i.e., there exists a $\mathbb{K}$-Hilbert space $\mathcal{H}$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$, such that $d_\mathcal{X}(x, x') = \|\Phi(x) - \Phi(x')\|_\mathcal{H}$.

Define $\mathcal{H}_0 = \{\Phi(x) \mid x \in \mathcal{X}\} \subseteq \mathcal{H}$, and for $f \in H$ set $\ell_f : \mathcal{H}_0 \to \mathbb{K}$ by $\ell_f(\Phi(x)) = f(x)$.

Lemma 3.3.15. For all $f \in H$, ℓ_f as above is a well-defined, linear and continuous map.

Proof. Let $f \in H$ be arbitrary. In order to show that ℓ_f is well-defined, let $x, x'\mathcal{X}$ such that $\Phi(x) = \Phi(x')$. We then have

$$|\ell_f(\Phi(x)) - \ell_f(\Phi(x'))| = |f(x) - f(x')| \leq \|f\|_H d_\mathcal{X}(x, x')^\alpha = \|f\|_H \|\Phi(x) - \Phi(x')\|_\mathcal{H}^\alpha = 0,$$

so $\ell_f(\Phi(x)) = \ell_f(\Phi(x'))$, and ℓ_f is indeed well-defined. Linearity and continuity are now clear. □

Given $f \in H$, we can now extend ℓ_f linearly to $\tilde{\ell_f} : \mathrm{span}\mathcal{H}_0 \to \mathbb{K}$, and the resulting map is still well-defined, linear and continuous. Define now $\mathcal{H}_\mathcal{X} = \overline{\mathrm{span}\mathcal{H}_0}^{\|\cdot\|_\mathcal{H}}$, then by construction $\mathcal{H}_0$ is dense in $\mathcal{H}_\mathcal{X}$. This means that for all $f \in H$, there exists a unique linear and continuous extension $\overline{\ell_f} : \mathcal{H}_\mathcal{X} \to \mathbb{K}$ of $\tilde{\ell_f}$. Note that this means that for all $f \in H$, $\overline{\ell_f} \in \mathcal{H}'_\mathcal{X}$ (the topological dual of $\mathcal{H}_\mathcal{X}$). Since $\mathcal{H}_\mathcal{X}$ is itself a Hilbert space (because it is a closed subset of a Hilbert space), for each $f \in H$, there exists a unique Riesz representer $R(\ell_f) \in \mathcal{H}_\mathcal{X}$. Define for all $f_1, f_2 \in H$

$$k(f_1, f_2) = \langle R(\ell_{f_2}), R(\ell_{f_1}) \rangle_{\mathcal{H}_\mathcal{X}}, \tag{3.18}$$

then k is a kernel on H with feature space $\mathcal{H}_\mathcal{X}$ and feature map $H \ni f \mapsto R(\ell_f) \in \mathcal{H}_\mathcal{X}$. The corresponding RKHS of k is given by

$$H_k = \{f \mapsto \ell_f h \mid h \in \mathcal{H}_\mathcal{X}\}, \tag{3.19}$$

cf. [189, Theorem 6.21].

3.4. Lipschitz and Hölder continuity inducing kernels

Essentially, the results in Section 3.3 ensure that RKHS functions of α-Hölder continuous kernels are $\alpha/2$-Hölder continuous. In particular, these results do not guarantee that RKHS functions of Lipschitz continuous kernels are themselves Lipschitz continuous. However, for many applications the regularity properties (here Lipschitz and Hölder continuity) of RKHS functions matter most, and a kernel should be chosen that enforces the desired regularity properties for the induced RKHS functions. This motivates the investigation of kernels that *induce* prescribed Hölder continuity of its RKHS functions.

3.4.1. Series expansions

We start by characterizing *all kernels* on a given metric space that have RKHS functions with prescribed Hölder continuity. To the best of our knowledge, this result is new.

Theorem 3.4.1. Let $(\mathcal{X}, d_\mathcal{X})$ be a metric space, k a kernel on $\mathcal{X}$, and $\alpha \in \mathbb{R}_{>0}$. The following statements are equivalent.

1. There exists $C \in \mathbb{R}_{>0}$ such that all $f \in H_k$ are α-Hölder continuous with Hölder constant $C\|f\|_k$.

2. There exists a Parseval frame $(f_i)_{i\in I}$ in H_k, such that for all $i \in I$, f_i is α-Hölder continuous with Hölder constant $L_i \in \mathbb{R}_{\geq 0}$, and $\sup_{i\in I} L_i < \infty$.

3. There exists a family of functions $(f_i)_{i\in I}$, $f_i : \mathcal{X} \to \mathbb{K}$, such that for all $i \in I$, f_i is α-Hölder continuous with Hölder constant $L_i \in \mathbb{R}_{\geq 0}$, and $\sup_{i\in I} L_i < \infty$, and for all $x, x' \in \mathcal{X}$
$$k(x, x') = \sum_{i\in I} f_i(x)\overline{f_i(x')}, \tag{3.20}$$
where the convergence is pointwise.

Proof. 2 ⇒ 1 Let $(f_i)_{i\in I}$ be a Parseval frame in H_k, such that for all $i \in I$, f_i is α-Hölder continuous with Hölder constant $L_i \in \mathbb{R}_{\geq 0}$, and $\sup_{i\in I} L_i < \infty$. Let $f \in H_k$ and $x, x' \in \mathcal{X}$ be arbitrary, then we have

$$\begin{aligned}
|f(x) - f(x')| &= \left|\sum_{i\in I}\langle f, f_i\rangle_k f_i(x) - \sum_{i\in I}\langle f, f_i\rangle_k f_i(x')\right| = \left|\sum_{i\in I}\langle f, f_i\rangle_k (f_i(x) - f_i(x'))\right| \\
&\leq \sum_{i\in I} |\langle f, f_i\rangle_k||f_i(x) - f_i(x')| \\
&\leq \sum_{i\in I} |\langle f, f_i\rangle_k| L_i d_\mathcal{X}(x, x')^\alpha \\
&\leq \left(\sum_{i\in I} |\langle f, f_i\rangle_k|\right)\left(\sup_{i\in I} L_i\right) d_\mathcal{X}(x, x')^\alpha \\
&\leq \sqrt{\sum_{i\in I} |\langle f, f_i\rangle_k|^2}\left(\sup_{i\in I} L_i\right) d_\mathcal{X}(x, x')^\alpha = \|f\|_k \left(\sup_{i\in I} L_i\right) d_\mathcal{X}(x, x')^\alpha.
\end{aligned}$$

In the first equality we used that $(f_i)_{i\in I}$ is a Parseval frame, and that norm convergence (in H_k) implies pointwise convergence. For the first inequality, we used the triangle inequality, and for the second inequality we used the assumption that f_i is

α-Hölder continuous with Hölder constant L_i. In the last inequality, we used

$$\sum_{i\in I} |\langle f, f_i\rangle_k| = \|(\langle f, f_i\rangle_k)_{i\in I}\|_{\ell_1(I)} \leq \|(\langle f, f_i\rangle_k)_{i\in I}\|_{\ell_2(I)} = \sqrt{\sum_{i\in I} |\langle f, f_i\rangle_k|^2}.$$

2 ⇒ 1 Let $(e_i)_{i\in I}$ be an ONB of H_k, so $\|e_i\|_k = 1$ for all $i \in I$. By assumption, all e_i are α-Hölder continuous with Hölder constant 1, and since an ONB is a Parseval frame, the claim follows.

2 ⇒ 3 This implication follows immediately from Theorem 3.1.2.

3 ⇒ 2 Let $(f_i)_{i\in I}$ be a family of function as given in the third item. By Theorem 3.1.2, $f_i \in H_k$ for all $i \in I$, and $(f_i)_{i\in I}$ forms a Parseval frame, so this family of functions fulfills the conditions in the second item. □

Since orthonormal bases (ONBs) are Parseval frames, we get immediately the following result.

Corollary 3.4.2. Let $(\mathcal{X}, d_\mathcal{X})$ be a metric space, k a kernel on $\mathcal{X}$, and $\alpha \in \mathbb{R}_{>0}$. The following statements are equivalent.

1. All $f \in H_k$ are α-Hölder continuous with Hölder constant $\|f\|_k$.
2. There exists an ONB $(e_i)_{i\in I}$ in H_k such that for all $i \in I$, e_i is α-Hölder continuous with Hölder constant 1.
3. For all ONB $(e_i)_{i\in I}$ in H_k, and all $i \in I$, e_i is α-Hölder continuous with Hölder constant 1.
4. For all $x, x' \in \mathcal{X}$,
$$k(x, x') = \sum_{i\in I} e_i(x)\overline{e_i(x')}, \tag{3.21}$$
where the convergence is pointwise, and $(e_i)_{i\in I}$ is an ONB $(e_i)_{i\in I}$ in H_k such that for all $i \in I$, e_i is α-Hölder-continuous with Hölder constant 1.

3.4.2. Ranges of integral operators

It is well-known that there is a close connection between the theory of RKHSs and integral operators. For example, for RKHSs defined on measure spaces and under suitable technical assumptions, Mercer's theorem allows a spectral decomposition of

the reproducing kernel, and an explicit description of the RKHS in terms of eigenfunctions of a related integral operator. For details, we refer to [189, Section 4.5]. Moreover, integral operators defined using the reproducing kernel of an RKHS can have ranges contained in the RKHS under suitable assumptions, cf. [189, Theorem 6.26]. This motivates the study of Hölder continuity properties for functions in the image set of integral operators.

A general result Before embarking on this task, we present a result for rather general integral maps. It is essentially a direct generalization of [72, Theorem 5.1].

Proposition 3.4.3. Let $(\mathcal{Y}, \mathcal{A}, \mu)$ be a measure space, $(\mathcal{X}, d_{\mathcal{X}})$ a metric space, $1 < p, q < \infty$ with $1/p + 1/q = 1$, and $k : \mathcal{X} \times \mathcal{Y} \to \mathbb{K}$ a function such that the following holds.

1. For all $x \in \mathcal{X}$, the function $k(x, \cdot)$ is measurable.
2. For all $g \in L^q(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{K})$ and all $x \in \mathcal{X}$, $k(x, \cdot) \cdot g \in L^1(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{K})$.
3. There exists $\alpha \in \mathbb{R}_{>0}$, $L_\alpha \in \mathcal{L}^p(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{R}_{\geq 0})$, such that for μ-almost all $y \in \mathcal{Y}$, the function $k(\cdot, y)$ is α-Hölder continuous with Hölder constant $L_\alpha(y)$.

In this case,

$$S_k : L^q(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{K}) \to \mathbb{K}^{\mathcal{X}}, \quad (S_k g)(x) = \int_{\mathcal{Y}} k(x, y) g(y) \mathrm{d}\mu(y) \tag{3.22}$$

is a well-defined linear mapping, and for all $g \in L^q(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{K})$, the function $f = S_k g$ is α-Hölder continuous with Hölder constant $\|L_\alpha\|_{\mathcal{L}^p} \|g\|_{L^q}$.

Proof. Since for all $g \in L^q(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{K})$ and all $x \in \mathcal{X}$ the function $k(x, \cdot)g \in L^1(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{K})$, the mapping S_k is well-defined. The linearity is now clear.

Let $g \in L^q(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{K})$, define $f = S_k g$, and let $x, x' \in \mathcal{X}$ be arbitrary, then

$$\begin{aligned} |f(x) - f(x')| &= \left| \int_{\mathcal{Y}} (k(x, y) - k(x', y)) g(y) \mathrm{d}\mu(y) \right| \leq \int_{\mathcal{Y}} |k(x, y) - k(x', y)| |g(y)| \mathrm{d}\mu(y) \\ &\leq \int_{\mathcal{Y}} L_\alpha(y) |g(y)| \mathrm{d}\mu(y) d_{\mathcal{X}}(x, x') \leq \|L_\alpha\|_{\mathcal{L}^p} \|g\|_{L^q} d_{\mathcal{X}}(x, x'), \end{aligned}$$

so f is indeed α-Hölder continuous with Hölder constant $\|L_\alpha\|_{\mathcal{L}^p} \|g\|_{L^q}$. □

Example To illustrate Proposition 3.4.3, we consider the rather general class of integral operators described in [216, Abschnitt 6.3]. Let $(\mathcal{X}, \mathcal{A}_\mathcal{X}, \mu)$ and $(\mathcal{Y}, \mathcal{A}_\mathcal{Y}, \nu)$ be measure spaces, $1 < p, q < \infty$ with $1/p + 1/q = 1$, and $k : \mathcal{X} \times \mathcal{Y} \to \mathbb{K}$ be measurable. Assume that for all $g \in L^q(\mathcal{Y}, \mathcal{A}_\mathcal{Y}, \nu)$ and μ-almost all $x \in \mathcal{X}$, $k(x, \cdot)g \in L^1(\mathcal{Y}, \mathcal{A}_\mathcal{Y}, \nu)$, and that by defining ($\mu$-almost all) $x \in \mathcal{X}$

$$(T_k g)(x) = \int_\mathcal{Y} k(x, y) g(y) \mathrm{d}\nu(y) \tag{3.23}$$

we get $T_k g \in L^p(\mathcal{X}, \mathcal{A}_\mathcal{X}, \mu)$. Under these conditions, $T_k : L^q(\mathcal{Y}, \mathcal{A}_\mathcal{Y}, \nu) \to L^p(\mathcal{X}, \mathcal{A}_\mathcal{X}, \mu)$ is a well-defined, linear and bounded operator.

Assume furthermore that $(\mathcal{X}, d_\mathcal{X})$ is a metric space, and that there exists $\alpha \in \mathbb{R}_{>0}$ and $L_\alpha \in \mathcal{L}^p(\mathcal{Y}, \mathcal{A}, \mu, \mathbb{R}_{\geq 0})$, such that for μ-almost all $y \in \mathcal{Y}$, the function $k(\cdot, y)$ is α-Hölder continuous with Hölder constant $L_\alpha(y)$. Let $g \in L^q(\mathcal{Y}, \mathcal{A}_\mathcal{Y}, \nu)$, then there exists a μ-nullset $\mathcal{N}_g$ such that (setting for brevity $\mathcal{X}_g = \mathcal{X} \setminus \mathcal{N}_g$) $f : \mathcal{X}_g \to \mathbb{K}$, $f(x) = (T_k g)(x)$ is well-defined. Proposition 3.4.3 now ensures that f is α-Hölder continuous with Hölder constant $\|L_\alpha\|_{\mathcal{L}^p} \|g\|_{L^q}$, though f is only defined on the restricted metric space $(\mathcal{X}_g, d_\mathcal{X}|_{\mathcal{X}_g \times \mathcal{X}_g})$.

In particular, each element[1] of the image set of T_k contains a μ-almost everywhere defined function that is α-Hölder continuous.

We can strengthen this result. Let $\mathcal{A}_\mathcal{X}$ be the Borel σ-algebra on $\mathcal{X}$, and assume that $\mu(U) > 0$ for all open nonempty $U \subseteq \mathcal{X}$. In this case, $\mathcal{X}_g$ is dense in $\mathcal{X}$, since otherwise $\mathcal{N}_g$ contains a nonempty open set U, and hence $\mu(\mathcal{N}_g) \geq \mu(U) > 0$, a contradiction to the fact that $\mathcal{N}_g$ is a μ-nullset. Since f is defined on a dense subsetset of $\mathcal{X}$, and it is continuous (since it is α-Hölder continuous on $\mathcal{X}_g$), there exists a unique extension $\bar{f} : \mathcal{X} \to \mathbb{K}$ that is also α-Hölder continuous. Defining $\bar{T}_k g := \bar{f}$, we thus arrived at a linear operator from $L^q(\mathcal{Y}, \mathcal{A}_\mathcal{Y}, \nu)$ into $\mathcal{L}(\mathcal{X}, \mathcal{A}_\mathcal{X}, \mu)$ with its range space consisting of α-Hölder continuous functions.

Integral operators into RKHSs Let us return to the setting of RKHSs. If an RKHS is defined on a measure space, and the kernel fulfills an integrability condition, then the RKHS consists of integrable functions, and the kernel allows the definition of a related integral operator with range contained in the RKHS. The next result provides a sufficient condition for Hölder continuity of RKHS functions in the range of this

[1] Recall that this is an equivalence class of functions on $\mathcal{X}$.

integral operator.

Proposition 3.4.4. Let $(\mathcal{X}, d_\mathcal{X})$ be a metric space, $(\mathcal{X}, \mathcal{A}, \mu)$ a σ-finite measure space,[2] $1 < p, q < \infty$ with $1/p + 1/q = 1$, and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ a measurable kernel such that H_k is separable and

$$\|k\|_{L^p} = \left(\int (k(x,x))^{\frac{p}{2}} \mathrm{d}\mu(x) \right)^{\frac{1}{p}} < \infty. \tag{3.24}$$

Assume that there exist $\alpha \in \mathbb{R}_{>0}$, $L_\alpha \in \mathcal{L}^p(\mathcal{X}, \mathcal{A}, \mu, \mathbb{R}_{\geq 0})$ such that for μ-almost all $x \in \mathcal{X}$ the function $k(\cdot, x)$ is α-Hölder continuous with Hölder constant $L_\alpha(x)$.

Under these conditions,

$$S_k : L^q(\mathcal{X}, \mathcal{A}, \mu, \mathbb{K}) \to H_k, \quad (S_k g)(x) = \int_\mathcal{X} k(x, x') g(x') \mathrm{d}\mu(x') \tag{3.25}$$

is a well-defined, bounded linear operator, and for all $g \in L^q(\mathcal{X}, \mathcal{A}, \mu, \mathbb{K})$, the function $f = S_k g \in H_k$ is α-Hölder continuous with Hölder constant $\|L_\alpha\|_{\mathcal{L}^p} \|g\|_{L^q}$.

Finally, all functions in H_k are p-integrable,[3] and if the inclusion $\mathrm{id} : H_k \to L^p(\mathcal{X}, \mathcal{A}, \mu, \mathbb{K})$ is injective, then the image of S_k is dense in H_k.

Proof. That S_k is well-defined, linear and bounded, follows from [189, Theorem 6.26]. The statement on the Hölder continuity of the functions in the images of S_k is a direct consequence of Proposition 3.4.3. The last claim follows again from [189, Theorem 6.26]. □

3.4.3. Feature mixture kernels

Theorem 3.4.1 characterizes Hölder continuity inducing kernels via series expansion. However, these might be difficult to work with, so an alternative description of such kernels can be useful. The next result presents a very general construction which is based on a mixture of feature maps. It vastly generalizes a method apparently introduced in [220].

Theorem 3.4.5. Let $(\Omega, \mathcal{A})$ be a measurable space, μ a finite nonnegative measure on $(\Omega, \mathcal{A})$, $(\mathcal{X}, d_\mathcal{X})$ a metric space, and $\mathcal{H}$ a $\mathbb{K}$-Hilbert space. Furthermore, let

[2] $\mathcal{A}$ can, but does not have to be the Borel σ-algebra on the metric space $\mathcal{X}$.
[3] This means that for all $f \in H_k$, $\int_\mathcal{X} |f(x)|^p \mathrm{d}\mu(x) < \infty$.

$\Phi(x,\cdot) \in \mathcal{L}^2(\Omega, \mathcal{A}, \mu, \mathcal{H})$ for all $x \in \mathcal{X}$. Finally, assume that there exist α, $L_\Phi \in \mathbb{R}_{>0}$ such that for μ-almost all $\omega \in \Omega$, $\Phi(\cdot,\omega)$ is α-Hölder continuous with Hölder constant L_Φ. Then

$$k(x,x') = \int_\Omega \langle \Phi(x',\omega), \Phi(x,\omega)\rangle_{\mathcal{H}} \mathrm{d}\mu(\omega) \tag{3.26}$$

is a well-defined kernel on $\mathcal{X}$, and all $f \in H_k$ are α-Hölder continuous with Hölder constant $L_\Phi \sqrt{\mu(\Omega)} \|f\|_k$.

Proof. First, we show that k is well-defined. Let $x, x' \in \mathcal{X}$, then $\|\Phi(x,\cdot)\|_{\mathcal{H}}, \|\Phi(x',\cdot)\|_{\mathcal{H}}$ are square-integrable, so we get

$$\begin{aligned}\int_\Omega |\langle \Phi(x',\omega), \Phi(x,\omega)\rangle_{\mathcal{H}}| \mathrm{d}\mu(\omega) &\leq \int_\Omega \|\Phi(x,\omega)\|_{\mathcal{H}} \|\Phi(x',\omega)\|_{\mathcal{H}} \mathrm{d}\mu(\omega) \\ &\leq \left(\int_\Omega \|\Phi(x,\omega)\|_{\mathcal{H}}^2 \mathrm{d}\mu(\omega)\right)^{\frac{1}{2}} \left(\int_\Omega \|\Phi(x',\omega)\|_{\mathcal{H}}^2 \mathrm{d}\mu(\omega)\right)^{\frac{1}{2}} < \infty,\end{aligned}$$

where we used Cauchy-Schwarz first in $\mathcal{H}$, then in $\mathcal{L}^2$. Next, we show that k is kernel by verifying that it is positive semidefinite. Let $x_1, \ldots, x_N \in \mathcal{X}$ and $c_1, \ldots, c_N \in \mathbb{C}$ be arbitrary, then

$$\begin{aligned}\sum_{i,j=1}^N c_i \overline{c_j} k(x_j, x_i) &= \int_\Omega \sum_{i,j=1}^N c_i \overline{c_j} \langle \Phi(x_j,\omega), \Phi(x_i,\omega)\rangle_{\mathcal{H}} \mathrm{d}\mu(\omega) \\ &= \int_\Omega \left\langle \sum_{i=1}^N c_i \Phi(x_i,\omega), \sum_{j=1}^N c_j \Phi(x_j,) \right\rangle_{\mathcal{H}} \mathrm{d}\mu(\omega) \\ &= \int_\Omega \left\| \sum_{i=1}^N c_i \Phi(x_i,\omega) \right\|_{\mathcal{H}}^2 \mathrm{d}\mu(\omega) \geq 0,\end{aligned}$$

so k is indeed positive semidefinite. Finally, let $f \in H_k$ and $x, x' \in \mathcal{X}$ be arbitrary, then $|f(x) - f(x')| \leq \|f\|_k d_k(x,x')$. Observe now that

$$\begin{aligned}d_k(x,x')^2 &= k(x,x) + k(x,x') + k(x',x) + k(x',x') \\ &= \int_\Omega \langle \Phi(x,\omega), \Phi(x,\omega)\rangle_{\mathcal{H}} + \langle \Phi(x,\omega), \Phi(x',\omega)\rangle_{\mathcal{H}} \\ &\quad + \langle \Phi(x',\omega), \Phi(x,\omega)\rangle_{\mathcal{H}} + \langle \Phi(x',\omega), \Phi(x',\omega)\rangle_{\mathcal{H}} \mathrm{d}\mu(\omega) \\ &= \int_\Omega \langle \Phi(x,\omega) - \Phi(x',\omega), \Phi(x,\omega) - \Phi(x',\omega)\rangle_{\mathcal{H}} \mathrm{d}\mu(\omega),\end{aligned}$$

hence

$$\begin{aligned} d_k(x,x')^2 &= \int_\Omega \|\Phi(x,\omega) - \Phi(x',\omega)\|_{\mathcal{H}} \mathrm{d}\mu(\omega) \\ &\leq \int_\Omega L_\Phi^2 d_{\mathcal{X}}(x,x')^{2\alpha} \mathrm{d}\mu(\omega) = L_\Phi^2 \mu(\Omega) d_{\mathcal{X}}(x,x')^{2\alpha}, \end{aligned}$$

so we get

$$|f(x) - f(x')| \leq \|f\|_k d_k(x,x') \leq L_\Phi \sqrt{\mu(\Omega)} \|f\|_k d_{\mathcal{X}}(x,x')^\alpha.$$

□

If the nonnegative measure in the preceding result is a probability measure, we get the following result as a special case.

Corollary 3.4.6. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space, $\mathcal{H}$ a $\mathbb{K}$-Hilbert space, and $(\Phi(x))_{x\in\mathcal{X}}$ a family of square-integrable Φ-valued random variables. Assume that there exist $\alpha, L_\Phi \in \mathbb{R}_{>0}$ such that Φ is almost surely α-Hölder continuous with Hölder constant L_Φ. Then

$$k(x,x') = \mathbb{E}[\langle \Phi(x'), \Phi(x) \rangle_{\mathcal{H}}] \tag{3.27}$$

is a well-defined kernel on $\mathcal{X}$, and all $f \in H_k$ are α-Hölder continuous with Hölder constant $L_\Phi \|f\|_k$.

The importance of this result is the fact that the kernel k described there is a *random feature kernel* in the sense of [164]. In particular, in practice $k(x,x')$ can be approximated by sampling from the random variables $\Phi(x), \Phi(x')$.

Finally, we can formulate another special case, which recovers the approach from [220].

Proposition 3.4.7. Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space, P a Borel probability measure on $\mathcal{X}$, $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{K}$ an α-Hölder-continuous function with Hölder-constant L_φ, and define $\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ by $\phi(x,z) = \varphi(d_{\mathcal{X}}(x,z))$. If $\phi(x,\cdot) \in \mathcal{L}^2(\mathcal{X}, P)$ for all $x \in \mathcal{X}$, then

$$k(x,x') = \int_{\mathcal{X}} \phi(x',z)\overline{\phi(x,z)} \mathrm{d}P(z) \tag{3.28}$$

is a well-defined kernel on $\mathcal{X}$, and all $f \in H_k$ are α-Hölder continuous with Hölder constant $L_\varphi \|f\|_k$.

Proof. We show that for all $z \in \mathcal{X}$, the function $\phi(\cdot, z)$ is α-Hölder continuous with Hölder constant L_φ. For this, let $x, x' \in \mathcal{X}$ be arbitrary, then

$$\begin{aligned}|\phi(x,z) - \phi(x',z)| &= |\varphi(d_\mathcal{X}(x,z)) - \varphi(d_\mathcal{X}(x',z))| \\ &\leq L_\varphi |d_\mathcal{X}(x,z)^\alpha - d_\mathcal{X}(x',z)^\alpha| \leq L_\varphi d_\mathcal{X}(x,x')^\alpha,\end{aligned}$$

where we used the inverse triangle inequality for the metric $(x, x') \mapsto d_\mathcal{X}(x, x')^\alpha$ in the last step. The result follows now from Theorem 3.4.5 by choosing $\Omega = \mathcal{X}$, $\mu = P$, $\mathcal{H} = \mathbb{K}$, and $\Phi = \phi$, and the fact that $P(\mathcal{X}) = 1$. □

3.5. Discussion

We presented a comprehensive discussion of Lipschitz and Hölder continuity of RKHS functions. Starting with the well-known Lipschitz continuity w.r.t. the kernel (semi)metric, we then investigated Hölder-continuity w.r.t. a given metric, including converse results, i.e., consequences of Hölder continuity in function spaces related to RKHSs. Finally, we provided characterizations as well as sufficient conditions for kernels inducing prescribed Lipschitz and Hölder continuity of their RKHS functions w.r.t. a given metric, an important aspect for applications.

The results presented here can be used to construct tailored kernels ensuring Lipschitz or Hölder continuous RKHS functions, or to check that existing kernels have such RKHS functions. Furthermore, because the results are *quantitative*, they can be used in numerical methods.

Finally, we would like to point out three interesting questions for future work.

First, the Lipschitz and Hölder continuity in RKHS that we have been concerned with here, are of a strong *uniform* nature, since the corresponding Lipschitz or Hölder constants are proportional to the RKHS function of the respective function, cf. the developments in Section 3.3. It would be interesting to investigate whether there exist kernels that enforce weaker, nonuniform Lipschitz or continuity properties.

Second, we investigated sufficient conditions for Lipschitz and Hölder continuity of RKHS functions via integral operators. However, all statements are restricted to the range space of the involved integral operators. Under some conditions, these range spaces are dense in RKHSs, so it would be interesting to investigate whether the Lipschitz and Hölder continuity properties transfers to the whole RKHS. Note

that this is not trivial since in the Hölder constant in Proposition 3.4.3 involves the L^q-norm of the preimage function, not the RKHS norm of the image function.

Finally, the results in Section 3.3.2 provide Lipschitz or Hölder constants involving the RKHS norm. However, it is unclear how conservative these results are, i.e., how much larger the Lipschitz or Hölder constants are compared to the best possible constants. Intuitively, it is clear that for generic RKHS functions there will be some conservatism. It would be interesting to investigate how big this conservatism is, and how it depends on properties of the kernel.

3.6. Comments

Apart from very minor changes, this chapter corresponds verbatim to the manuscript [CF1]. The author of this thesis is the sole author of the aforementioned manuscript.

Part II.

Uncertainty bounds and learning-based control

4. Uncertainty in learning-based control

In the present chapter, we set the stage for our work on uncertainty bounds in the context of learning-based control. First, we provide some background and context, which will motivate our specific approaches later on. We then give a concise introduction to Gaussian process regression and kernel ridge ression, which are the primary tools in this part of the thesis. Finally, we conclude with a discussion of uncertainty bounds for these methods.

4.1. Introduction

We start with some considerations on the role of uncertainty in learning-based control. First, we outline on an abstract level how uncertainty and robust control interact in the context of learning based control. We choose a very general perspective to emphasize the fundamental character of these issues, which are independent of any particular learning or control method. This is particularly relevant given the vast amount of literature on learning-based control and related fields. Based on this exposition, we will then describe our focus for the remainder of this part, and argue why it is attractive from a conceptual perspective.

The considerations here are well-known, but rarely described precisely in the learning-based control literature. Our exposition is similar to the literature on behavioural systems theory [131] and non-falsified control [172], but we are not aware of a suitable presentation in the context of learning-based control.

4.1.1. Uncertainty, learning and robust control

Let $\mathcal{M}$ be the set of all models of interest, and assume that a particular model $\theta_* \in \mathcal{M}$ fully describes the true system (or process or phenomenon) of interest. On an abstract level, in control θ_* describes the system that needs to be controlled, in

this context often called *plant*. Let $\mathcal{C}$ be the set of possible controllers, and let us formalize the control goal by a predicate P, so we say that $\mathrm{P}(\theta, c)$ holds or is true, if a controller $c \in \mathcal{C}$ achieves the control goal for system $\theta \in \mathcal{M}$. For example, the goal might be stabilization of an equilibrium with a static state feedback controller. If the control goal is non-trivial, the controller needs to be suitable for the given system, but the latter is in general not fully known. In other words, we want to achieve a control goal for θ_* with some $c \in \mathcal{C}$ without knowing θ_* exactly.

In *robust control*, the following strategy is used to deal with this problem [65]. A set $\mathcal{U} \subseteq \mathcal{M}$, often called an *uncertainty set*, is determined with $\theta_* \in \mathcal{U}$, and a controller $c_\mathrm{R} \in \mathcal{C}$ is built such that $\mathrm{P}(\theta, c_\mathrm{R})$ holds for *all* $\theta \in \mathcal{U}$, i.e., the controller achieves the control goal for all possible systems. Since this includes the actual system θ_*, the controller actually achieves the goal. Defining

$$\mathcal{P}_\mathrm{P}(c) = \{\theta \in \mathcal{M} \mid \mathrm{P}(\theta, c) \text{ holds}\}, \tag{4.1}$$

this means $\mathcal{U} \subseteq \mathcal{P}_\mathrm{P}(c_\mathrm{R})$, though the latter set will be usually larger than $\mathcal{U}$. It is clear that the larger $\mathcal{U}$ is, the more difficult it will be to ensure $\mathcal{U} \subseteq \mathcal{P}_\mathrm{P}(c)$ for a single controller $c \in \mathcal{C}$. The worst case is of course $\{c \in \mathcal{C} \mid \mathcal{U} \subseteq \mathcal{P}_\mathrm{P}(c)\} = \emptyset$, so no controller can fulfill the goal for all $\theta \in \mathcal{U}$. In many control tasks, one is also interested in a quantitative *performance measure* of the controller, and insisting on $\mathcal{U} \subseteq \mathcal{P}_\mathrm{P}(c)$ makes the optimization (or tuning) of $c \in \mathcal{C}$ w.r.t. a performance measure more difficult. Intuitively, a larger $\mathcal{U}$ requires a more *conservative* controller, leading to worse performance in general.

The preceding discussion motivates the search for a small $\mathcal{U}$. Using prior knowledge, one might get some $\mathcal{U}_0 \subseteq \mathcal{M}$ with $\theta_* \in \mathcal{U}_0$, but this uncertainty set might be quite large. In learning-based control, the situation can be improved with *data* from the actual system. Let $\mathcal{Y}$ be the set of all possible measurement results, $\mathcal{N}$ the set of all possible external noise realizations, and $\mathcal{F} : \mathcal{M} \times \mathcal{N} \to \mathcal{Y}$ the map describing the measurement process. Given a measurement $y \in \mathcal{Y}$, define

$$\mathcal{V}(y) = \{\theta \in \mathcal{M} \mid \exists \eta \in \mathcal{N} :\ y = \mathcal{F}(\theta, \eta)\}. \tag{4.2}$$

In words, $\mathcal{V}(y)$ is the set of all systems that could have led to the observation y. This is essentially an abstract, simplified version of spaces appearing in the *unfalsified*

control approach [172], and can be interpreted as a noisy variant of the *version space* appearing in (classic) artificial intelligence and machine learning [171]. Importantly, by definition we have for all $\theta \in \mathcal{M}$ that

$$\theta \in \mathcal{V}(\mathcal{F}(\theta, \eta)) \quad \forall \eta \in \mathcal{N}. \tag{4.3}$$

In words, an arbitrary, but fixed target $\theta \in \mathcal{M}$ will always be contained in the induced uncertainty set $\mathcal{V}(\mathcal{F}(\theta, \eta))$, no matter which noise realization appears in the data generating process. Consider now $y = \mathcal{F}(\theta_*, \eta)$ for some in general unknown $\eta \in \mathcal{N}$, and define $\mathcal{U} = \mathcal{V}(y) \cap \mathcal{U}_0$. By construction, $\theta_* \in \mathcal{U}$, and for any $c_{\mathrm{R}} \in \mathcal{C}$ with $\mathcal{U} \subseteq \mathcal{P}_{\mathrm{P}}(c_{\mathrm{R}})$, $\mathrm{P}(\theta_*, c_{\mathrm{R}})$ holds, so the control goal is achieved by c_{R}, without knowing the exact θ_* and regardless of which unknown noise realization η corrupted the data. In other words, data has been used to reduce the set of possible systems, and the remaining uncertainty is dealt with a robust control strategy. On an abstract level, this is how most (if not all) learning-based control approaches with guarantees work.

Sometimes $\mathcal{U}$ is already small enough so that a satisfying $c_{\mathrm{R}} \in \mathcal{C}$ with $\mathcal{U} \subseteq \mathcal{P}_{\mathrm{P}}(c_{\mathrm{R}})$ can be identified. However, in general $\mathcal{V}(y)$ will be rather large, and we might even have $\mathcal{V}(y) = \mathcal{M}$ (so we did not gain anything from data). In order to improve on this situation, we can essentially reduce the *confidence*[1] in our inference about θ_*.

Worst-case approach In some situations, it *might* be reasonable to assume a reduced set of noise values $\bar{\mathcal{N}} \subseteq \mathcal{N}$, leading to

$$\bar{\mathcal{V}}(y) = \{\theta \in \mathcal{M} \mid \exists \eta \in \bar{\mathcal{N}} : \; y = \mathcal{F}(\theta, \eta)\} \subseteq \mathcal{V}(y), \tag{4.4}$$

and $\bar{\mathcal{U}} = \bar{\mathcal{V}}(y) \cap \mathcal{U}_0$ might be sufficiently small for the downstream tasks, in our case finding an appropriate $c_{\mathrm{R}} \in \mathcal{C}$ with $\bar{\mathcal{U}} \subseteq \mathcal{P}_{\mathrm{P}}(c_{\mathrm{R}})$. Specific instantiations of this approach are ubiquitous in systems and control, from nonlinear set membership estimation [139] to recent data-driven robust control approaches [25].

Note that we can interpret this as a *worst case* approach: No matter which noise realization $\eta \in \bar{\mathcal{N}}$ interfered with the data generating process, $\theta_* \in \bar{\mathcal{U}}$ will hold. In particular, no additional modelling assumptions are necessary, and also adversarial disturbance models are covered.

[1] Here it is meant in an intuitive, colloquial way, not in a formal sense as e.g. in statistics.

In order to achieve a reasonably small $\bar{\mathcal{U}}$, the reduced noise set $\bar{\mathcal{N}}$ needs to be small enough. While this can be reasonable in some situation, on the one hand it excludes certain noise models. For example, if one has an additive noise model, then in general it requires the assumption of bounded noise, excluding for example the common Gaussian noise assumption. On the other hand, it is not possible to include additional prior knowledge in the form of stochastic assumptions.

Frequentist uncertainty sets The preceding drawbacks motivate the following approach. Let $\bar{\mathcal{P}}$ be a set of probability distributions[2] over $\mathcal{N}$ and construct a family of maps $\mathcal{U}_\delta : \mathcal{Y} \to 2^{\mathcal{U}_0}$, $\delta \in (0,1)$, such that for all $\theta \in \mathcal{U}_0$, $P \in \bar{\mathcal{P}}$ and $\delta \in (0,1)$ we have

$$\mathbb{P}_{\eta \sim P}[\theta \in \mathcal{U}_\delta(\mathcal{F}(\theta, \eta))] \geq 1 - \delta. \tag{4.5}$$

If $y = \mathcal{F}(\theta_*, \eta)$ with $\eta \sim P$ for some $P \in \bar{\mathcal{P}}$, then for a user-defined $\delta \in (0,1)$ we have

$$\mathbb{P}[\theta_* \in \mathcal{U}_\delta(y)] \geq 1 - \delta. \tag{4.6}$$

This is the setting of *frequentist statistics*. We have a fixed, i.e., deterministic or constant ground truth θ_*, and randomness only enters through the data generating process (in the present formalism, through the now random η). In particular, $\mathcal{U}_\delta(y)$ is a *(frequentist) confidence set of level* $1 - \delta$, which is a random set.

If we construct now $c_{\mathrm{R}} \in \mathcal{C}$ such that $\mathcal{U}_\delta(y) \subseteq \mathcal{P}_{\mathrm{P}}(c_{\mathrm{R}})$ for a user-defined $\delta \in (0,1)$, then $\mathbb{P}[\mathrm{P}(\theta_*, c_{\mathrm{R}}) \text{ holds}] \geq 1 - \delta$. In words, the (robust) controller c_{R} achieves the control task (on the actual, unknown system θ_*) with probability at least $1 - \delta$. This also shows why it is important to have a *user-defined* $\delta \in (0,1)$, since the final probability of success is determined by the user of the controller. For example, for a safety-critical application, an extremely small δ might be desirable, whereas for a non-safety-critical scenario a moderate δ can be enough. Since a smaller δ leads to a potentially larger uncertainty set $\mathcal{U}_\delta(y)$, we have in a general a reliability-performance (or safety-performance) tradeoff.

Finally, note that the frequentist approach fits very nicely together with the standard robust control approach. In particular, no modifications to the latter have to be done. Furthermore, if a probabilistic controller synthesis procedure is used, this

[2] We ignore measurability issues here. In the concrete settings considered later on, they do not pose any problem.

can be taken into account with a simple union bound: if $\delta \in (0,1)$ is the final desired probability of control success, then we can use $\mathcal{U}_{\delta/2}(y)$ as the uncertainty set and require a confidence of $\delta/2$ in the probabilistic controller synthesis algorithm. The union bound then assures that the final controller achieves the control task on θ_* with probability at least $1-\delta$.

This approach, or rather concrete instantiations thereof, is very popular in modern learning-based control, cf. e.g. [29, 31, 111] for typical examples, as well as Chapter 6.

Bayesian or probabilistic approach Let now Q be a distribution on $\mathcal{U}_0$, let $\bar{\mathcal{P}}$ be a set of probability distributions over $\mathcal{N}$ and construct a family of maps $\mathcal{U}_\delta : \mathcal{Y} \to 2^{\mathcal{U}_0}$ such that for all $P \in \bar{\mathcal{P}}$ and $\delta \in (0,1)$ we have

$$\mathbb{P}_{(\theta,\eta)\sim Q\otimes P}[\theta \in \mathcal{U}_\delta(\mathcal{F}(\theta,\eta))] \geq 1-\delta. \tag{4.7}$$

Assume now that $\theta_* \sim Q$, $\eta \sim P$ for some $P \in \bar{\mathcal{P}}$, and that θ_* and η are independent. Defining $y = \mathcal{F}(\theta_*, \eta)$, we then have for all $\delta \in (0,1)$ that

$$\mathbb{P}[\theta_* \in \mathcal{U}_\delta(y)] \geq 1-\delta.$$

Furthermore, if we construct $c_{\mathrm{R}} \in \mathcal{C}$ such that $\mathcal{U}_\delta(y) \subseteq \mathcal{P}_{\mathrm{P}}(c_{\mathrm{R}})$ for a user-defined $\delta \in (0,1)$, then $\mathbb{P}[\mathrm{P}(\theta_*, c_{\mathrm{R}}) \text{ holds}] \geq 1-\delta$.

We can interpret this in two ways. In the *Bayesian* approach, Q corresponds to a *prior* over all possible models $\mathcal{U}_0 \subseteq \mathcal{M}$ (including the prior knowledge that we can restrict us to $\mathcal{U}_0$). The idea is that Q encodes all of our prior knowledge about the unknown θ_*, and for both practical as well as formal (and even philosophical) reasons it is advisable to express this in the form of a probability distribution, cf. [99]. Furthermore, $P \in \bar{\mathcal{P}}$ together with the map $\mathcal{F}$ can be interpreted as a likelihood model. In this context, $\mathcal{U}_\delta$ becomes a *Bayesian confidence set* or *belief set* at level $1-\delta$.

We can also interpret it from a more stochastic (or probabilistic) perspective. Assume that we are actually not interested in *one specific* θ_*, but rather many instances that are distributed according to some Q. Consequently, we care about the overall success over many instances. In particular, the probability of success,

i.e., $\mathbb{P}[\theta_* \in \mathcal{U}_\delta(y)]$, is w.r.t. to the distribution over θ_* and the noise influencing the data generating process. But this is again exactly the formal setting in the Bayesian approach just outlined, though with a different practical interpretation of the probabilistic nature. For a concrete example of this setting in the context of learning-based control, we refer to [168], which contains additional discussions of probabilistic notions in robust control.

Before moving on, let us briefly summarize our discussion of frequentist and probabilistic uncertainty sets. In both cases we have a stochastic noise model, and the resulting uncertainty bounds are formulated in a stochastic sense. In particular, *formally* both types of bounds look similar. However, from a practical perspective, the *interpretation* of these bounds is rather different. Whereas a frequentist uncertainty bound holds for an unknown, but fixed ground truth, and the stochasticity enters only through noise, a probabilistic or Bayesian uncertainty bound holds w.r.t. a prior distribution over the object of interest, instead of a fixed, unknown ground truth. In the next section, we will discuss these differences in the context of learning-based control.

4.1.2. Discussion

It is clear that it depends on the context which of the aforementioned three approaches is most appropriate for a given application in learning-based control. Relevant aspects include inter alia the type of prior knowledge (e.g., qualitative or quantitative, stochastic or not), formal requirements of the downstream control methodology (e.g., parametric or nonparametric uncertainty sets), and the level of reliability or safety of the overall control scheme (e.g., deterministic or stochastic guarantees, asymptotic or non-asympotic setting). In practical scenarios, these factors might still leave some room for choice. In this part of the thesis, we will focus primarily on the second approach, and in Chapter 7 we will also consider the third approach. We will now briefly discuss these choices, and how all of this will be made concrete.

Our starting point is the overall strategy outlined above. We would like to stress that this is in no way original, but rather the standard approach in learning-based control, at least in the context of works providing theoretical guarantees. However, we put special emphasis on achieving *rigorous (statistical and control-theoretic)*

guarantees on the overall methods and ensuring that only *reasonable, practical assumptions* are made. We would like to argue that this strongly suggests to rely on the second approach from above (frequentist uncertainty sets), and if necessary switching to the first approach (worst-case uncertainty sets).

Ideally, we achieve guarantees at the end that are meaningful for practitioners, which in the present context means giving guarantees in a form that are useful or accepted in control engineering. In particular, any uncertainty should be of a form that is acceptable in practice or has an accepted interpretation. Worst-case guarantees (so that success is guaranteed for any possible noise realization from a predefined set) are certainly acceptable in robust control, and appear to be the standard form of guarantees, cf. e.g. [167]. Furthermore, frequentist guarantees (i.e., statistical guarantees from stochastic assumptions about the noise, but not the underlying system) are conceptually also well-suited in the context of robust control. In addition, a frequentist setup allows more practical noise models, cf. the discussion above, and in general the uncertainty will reduce (in a stochastic sense) with increasing data, as is investigated in-depth in the theory of machine learning [189].

However, a Bayesian or probabilistic approach appears to be problematic in this context. In general, we are interested in giving guarantees for a specific, but unknown instance θ_*, and even more importantly, it is unclear what the probabilistic guarantees mean in the context of robust control. Note that this is also related to deep foundational issues in probabilistic robust control, which are beyond the scope of the present thesis, cf. e.g. [199].

Finally, our focus on rigorous statistical and control-theoretic guarantees, under practically meaningful assumptions, also entails the following consequences.

1. The learning method has to be able to provide reasonably small uncertainty sets that are suitable for downstream control tasks.

2. We need to be able to include prior knowledge, e.g., from first-principles modelling or engineering experience, ideally in a systematic fashion.

3. All components used (learning and control algorithms) have to come with appropriate theoretical guarantees.

The second and third item are a strong motivation to focus on kernel methods, cf.

also Chapter 1, and the first item suggests to start with GP regression, since it comes with a natural measure of uncertainty, cf. the following section.

4.2. Kernel and Gaussian process regression

We now provide a concise introduction to Gaussian process regression, and the closely related kernel ridge regression. Furthermore, we discuss some practical aspects of these methods. Our exposition is based on standard machine learning references on these topics, in particular, [166] and [102].

4.2.1. Gaussian Process regression

In this section, we fix a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ in the background. Furthermore, we use a slightly different notation compared to Section 2.6 for consistency with the machine learning literature.

Gaussian process regression is a nonparametric regression method based on Bayesian principles. As such, we need a *prior*, a *likelihood* or *measurement model*, and a way to work with resulting *posterior* (which is formally given by Bayes theorem, but might be intractable). Furthermore, it is *nonparametric* since it works directly with functions instead of finite-dimensional representations thereof[3]. In particular, the prior will be a distribution over functions, which formally can be interpreted as a stochastic process. We will now introduce the specific class of stochastic processes that will be used as priors in the following.

Let $\mathcal{X} \neq \emptyset$ be some set and $(f_x)_x$ a real-valued stochastic process with index set $\mathcal{X}$. In the following, we will also use the notation $f(x) = f_x$, $x \in \mathcal{X}$, as well as $f = (f(x))_{x \in \mathcal{X}}$. We call f a *Gaussian process* (GP) if for all pairwise distinct $x_1, \dots, x_N \in \mathcal{X}$ the N-dimensional random variable $\begin{pmatrix} f(x_1) & \cdots & f(x_N) \end{pmatrix}^\top$ has a Gaussian distribution. In this case, we call

$$\mu : \mathcal{X} \to \mathbb{R}, \quad \mu(x) = \mathbb{E}[f(x)] \tag{4.8}$$

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{4.9}$$

[3] Of course, if the input space is a finite set, then one can interpret Gaussian process again as a parametric method.

the *mean* and *covariance function* of f, and we write $f \sim \mathcal{GP}_{\mathcal{X}}(\mu, k)$, or $f \sim \mathcal{GP}(\mu, k)$ if the index set $\mathcal{X}$ is clear from the context. Since a Gaussian distribution is fully described by its mean and covariance matrix, and the finite-dimensional marginal distributions uniquely determine the law of a stochastic process, a Gaussian process is fully characterized by its mean function and its covariance function. Recall from Section 2.6 that k is a positive semidefinite function, hence k is often called the *kernel (function)*[4] of f. Conversely, for any function $\mu : \mathcal{X} \to \mathbb{R}$ and any symmetric and positive semidefinite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ there exists a Gaussian process with mean function μ and covariance function k.

Remark 4.2.1. As is well-known, in order to have a density w.r.t. the Lebesgue measure, a multivariate Gaussian distribution needs to have a positive definite covariance matrix. However, by requiring that the covariance function is only positive semidefinite (which is enough to ensure that it is a kernel, cf. Chapter 2), it might happen that some of the induced covariance matrices are only positive semidefinite. This is not a problem, since in general a multivariate Gaussian distribution is defined in a weak form, cf. [116, Chapter 1], and indeed, this issue is usually not explicitly dealt with in the machine learning literature. Furthermore, when the covariance function is positive definite and all inputs are pairwise distinct, then this situation does not occur in the first place.

If $f \sim \mathcal{GP}(\mu, k)$, then $g \sim \mathcal{GP}(0, k)$, where we defined $g(x) = f(x) - \mu(x)$. Conversely, if $f \sim \mathcal{GP}(0, k)$ and $\mu : \mathcal{X} \to \mathbb{R}$ is any function, then $g \sim \mathcal{GP}(\mu, k)$, where $g(x) = f(x) + \mu(x)$. This suggests that we can work with a zero mean function, which is what is done usually in the stochastic process and machine learning literature. So, consider a prior $f \sim \mathcal{GP}_{\mathcal{X}}(0, k)$.

Remark 4.2.2. Technically, a stochastic process cannot be a prior, since the latter has to be a distribution (over the objects of interest), and the former is a random object, which has a distribution (its law). However, from a practical perspective, and for the following theoretical developments, this does not cause any problems since we work only with the stochastic process and its finite-dimensional marginals instead of the induced probability measures. For more technical background on this issue, we refer to the literature on nonparametric statistics, e.g., [80].

[4] This terminology is problematic, since there exists a related, but formally different object called *kernel of a second-order stochastic process*.

Next, we need a likelihood model. Given a data set $\mathcal{D} = ((x_1, y_1), \dots, (x_N, y_N)) \in (\mathcal{X} \times \mathbb{R})^N$, assume that $y_n = f(x_n) + \eta_n$, $n = 1, \dots, N$, with noise $\eta_1, \dots, \eta_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \lambda)$, and we also assume independence of the noise from f.

If we now condition on the data, then the posterior process $f|_{\mathcal{D}}$ is still a Gaussian process, and the (characterizing) mean and covariance functions are given in closed form. In short, we have $f|_{\mathcal{D}} \sim \mathcal{GP}_{\mathcal{X}}(\mu_{\mathcal{D}}, k_{\mathcal{D}})$ with

$$\mu_{\mathcal{D}}(x) = k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{y} \tag{4.10}$$

$$k_{\mathcal{D}}(x) = k(x, x') - k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x') \tag{4.11}$$

where we defined

$$k_{\mathcal{D}}(x) = \begin{pmatrix} k(x, x_1) \\ \vdots \\ k(x, x_N) \end{pmatrix}, \quad K_{\mathcal{D}} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}.$$

The matrix $K_{\mathcal{D}}$ is usually called *kernel matrix* or *Gram matrix*. An explanation for the latter terminology can be found in Section 5.2. Finally, we also define the posterior variance (function)

$$\sigma^2_{\mathcal{D}}(x) = k_{\mathcal{D}}(x, x) = k(x, x) - k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x). \tag{4.12}$$

The posterior mean $\mu_{\mathcal{D}}$ is usually interpreted as a nominal prediction of the unknown target function generating the data, and the posterior standard deviation $\sigma_{\mathcal{D}}$ is interpreted as a measure of uncertainty.

4.2.2. Kernel ridge regression

We still consider the problem of nonparametric regression, but now from a very different perspective. Instead of following a Bayesian paradigm, we fix a *hypothesis space*, and then search for a candidate hypothesis (here a function) by using an optimization problem depending on the given data.

Let $\mathcal{X}$ be some input set and $\mathcal{D} = ((x_1, y_1), \dots, (x_N, y_N)) \in (\mathcal{X} \times \mathbb{R})^N$ a data set. As the hypothesis space we use an RKHS H_k, where k is a kernel on $\mathcal{X}$, and we

search for a suitable hypothesis with the following optimization problem

$$\min_{f \in H_k} \sum_{n=1}^{N} (y_n - f(x_n))^2 + \lambda \|f\|_k^2, \tag{4.13}$$

where $\lambda \in \mathbb{R}_{>0}$ is a regularization parameter. This approach is called *kernel ridge regression*, since it is a kernelized version of *ridge regression* [89, Section 3.4]. It is a form of regularized empirical risk minimization, using the square loss and the (squared) RKHS norm as regularizer [189]. Note that H_k is often infinite-dimensional, and hence the preceding minimization problem is an *infinite-dimensional optimization problem*. However, it turns out that it is equivalent to a finite-dimensional optimization problem, and the unique solution has the closed form[5]

$$f_\lambda(x) = k_\mathcal{D}(x)^\top \left(K_\mathcal{D} + \lambda I\right)^{-1} \boldsymbol{y}, \tag{4.14}$$

with $k_\mathcal{D}, K_\mathcal{D}$, and $\boldsymbol{y}$ as defined before. But we immediately recognize that $f_\lambda = \mu_\mathcal{D}$, the posterior mean of GP regression, if we choose a zero mean GP prior with covariance function k, and assume additive i.i.d. $\mathcal{N}(0, \lambda)$ noise. For more details on kernel ridge regression and the connection to GP regression, we refer to [102, Section 3.3]. Furthermore, we will revisit kernel ridge regression in Section 5.2 from a different perspective, that will be particularly helpful for the connection with GP regression.

4.2.3. Further aspects and extensions

In the following, we elaborate on some practical aspects of GP regression and kernel ridge regression, as well as generalizations and the connection to related methods.

Prior knowledge and choice of kernel A key advantage of kernel methods, including Gaussian process regression and kernel ridge regression, is the systematic inclusion of prior knowledge. More precisely, if certain prior knowledge about the target function is known, then in many cases there exists systematic ways to include this in the kernel-based learning method by *enforcing* corresponding properties. This becomes particularly transparent for kernel ridge regression, which searches for a

[5] This follows from the representer theorem for kernel machines [180, 102].

suitable hypothesis (and estimate of the target function) in an RKHS, which in turn is generated from its reproducing kernel, cf. Chapter 2. In particular, properties of the RKHS functions are to a large extent determined by the chosen kernel. For example, regularity properties of the kernel, like continuity, Lipschitz continuity or differentiability, are inherited by the RKHS functions, cf. [189, Section 4.3], Chapter 3, and [217, Chapter 10]. Put differently, by choosing an appropriately regular kernel, the outcome of kernel ridge regression will fulfill the corresponding regularity constraint.

In principle, the same connection holds between the covariance function and sample paths of a GP, cf. [3, Chapter 1], but since a GP is a stochastic process, this situation is more delicate, cf. also [102, Section 4].

The previously mentioned properties are mostly *qualitative*, but inclusion of more quantitative properties is also possible. For example, invariances can be enforced by the kernel. In multioutput setting, cf. also the remarks below, linear differential constraints can be easily included [101], and using computer algebra this can be even automated [115].

Since all of this pertains to the kernels (for kernel ridge regression) and covariance functions (for GP regression), but not the learning method itself, we work in the following with generic kernels as much as possible, since this makes all of the aforementioned approaches immediately applicable. Furthermore, another advantage of kernel methods is given by the separation of the input set and the learning method – any structure (or lack thereof) of the input set enters the learning method only through the kernel, which makes kernel methods also applicable to graphs, texts or sets as inputs, as long as appropriate kernel are available, cf. [182] for many such examples.

Finally, in GP regression there is an additional way to introduce prior knowledge – by choice of the prior mean function. This is commonly used in two ways [166, Chapter 2]. On the one hand, appropriate prior knowledge can be model by a nominal function, which is then used as the GP prior mean, or equivalently, substracted from the data, so that the GP models only the difference. On the other hand, the prior mean can also be modelled in a parametric manner, which turns GP regression into a semiparametric method. As common in the literature, we do not consider these extensions, and work with a zero mean prior in the following. With a view towards the developments in Chapter 5, we would like to mention that

frequentist uncertainty bounds would be an interesting extension and to the best of our knowledge no results in this direction exist so far.

Hyperparameters Even if a specific class of covariance functions has been selected (from prior knowledge, via a model selection procedure, or even just convenience), in general this still leaves open the choice of a certain set of parameters of the covariance function or kernel, which in this context are called *hyperparameters*. For example, many translation invariant covariance functions $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (assuming that $\mathcal{X}$ has a vector space structure) are of the form

$$k(x, x') = \psi\left(\gamma^{-1}(x - x')\right), \tag{4.15}$$

where $\gamma \in \mathbb{R}_{>0}$ is called the *length scale*.

Remark 4.2.3. Note that the literature is not uniform w.r.t. the parameterization. For example, in the GP regression literature, e.g. [166, Section 2.3], the well-known SE kernel (with $\mathcal{X} \subseteq \mathbb{R}^d$) is often written as

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right) \tag{4.16}$$

and $\ell \in \mathbb{R}_{>0}$ is called its length scale, whereas in the kernel methods literature, e.g. [189, Proposition 4.10], one often finds

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\gamma^2}\right) \tag{4.17}$$

and $\gamma \in \mathbb{R}_{>0}$ is called its width.

Similarly, it is often assumed that $\psi(0) = 1$ and one considers the form

$$k(x, x') = \sigma_f^2 \psi\left(\gamma^{-1}(x - x')\right), \tag{4.18}$$

where σ_f^2 is called the *signal variance* in the GP regression literature [166, Section 2.3]. This terminology is explained by the fact that if $f \sim \mathcal{GP}_{\mathcal{X}}(\mu, k)$, then for all $x \in \mathcal{X}$, $\mathrm{Var}[f(x)] = \mathrm{Var}[Z]$ for $Z \sim \mathcal{N}(\mu(x), k(x, x))$, so $\mathrm{Var}[f(x)] = k(x, x) = \sigma_f^2 \psi(0) = \sigma_f^2$.

In practice, two systematic approaches for selecting hyperparameters are common [166, Section 2.3]. The most principled way is to apply a Bayesian approach and put a prior on the hyperparameters, in this context called a *hyperprior*. This prior has its own hyperparameter, however, it is argued that the influence of the latter is reduced due to the hierarchical structure. However, inference in this setting is much more challenging and in general one looses the advantage of explicit expressions for the posterior. Therefore, most often one an *empirial Bayes*, also called *type-II likelihood*, approach. The idea is to maximize the likelihood of the observed data by varying the hyperparameters.

While these methods work well in practice, they considerably complicate the theoretical treatment, and in particular, the development of uncertainty bounds. For example, in many works on kernelized bandits which rely on GP regression, the question of hyperparameter choice or misspecification thereof is not dealt with in the theoretical developments, cf. [186, 54, 219]. In fact, the theoretical properties of hyperparameter choice in GP regression are still under active investigation, with first results in [197, 103].

In the remainder of this chapter, we follow the common approach in the literature and do not consider any hyperparameter selection method per se. If the choice of hyperparameter does play a role, e.g. when considering a target function from the RKHS generated by the covariance function, we either assume that all hyperparameters are known, or we derive results that are robust to certain misspecifications, cf. Section 5.5.

Scalability to large data sets Inspecting the explicit expressions for the posterior in GP regression, and the closed form solution of kernel ridge regression, shows that a symmetric and positive definite $N \times N$ matrix, where N are the number of data points, needs to be inverted, or equivalently a corresponding equation system needs to be solved. This means that using GP regression or kernel ridge regression needs $\mathcal{O}(N^2)$ space and (roughly) $\mathcal{O}(N^3)$ time. At least in an offline-learning setting, the typical data set sizes in control do not pose a problem here. However, in modern machine learning, N will be so large that the naive approach of matrix inversion (or solving the equation system) becomes infeasible, and hence approximation methods are necessary. For an overview of classic methods, we refer to [166, Chapter 8] and [144, Section 18.5]. In the context of kernel ridge regression, the *Nystrom method*

has been successfully applied to scale to billions of data points, e.g., [169, 170, 135]. Another approach is to use *random (Fourier) features*, and we refer to [120] for an extensive survey. In the context of GP regression, sparse GPs can be used in a large-scale setting, cf. e.g. [184, 163]. However, using these techniques in the context of uncertainty bounds for GPs (and kernel ridge regression) is problematic due to *variance starvation*, which leads to an underestimate of the uncertainty, and the approximation has to be taken into account for the uncertainty bounds, cf. e.g. [43]. Since we are anyway interested in control-applications with moderate N, and to keep the following developments to a reasonable length, we do not consider uncertainty bounds in the context of approximation methods for large-scale applications.

Multiple outputs and more general problems So far, we only considered learning a scalar function $f_* : \mathcal{X} \to \mathbb{R}$. However, it is clear that in practical scenarios, target functions with multiple outputs appear, or even more general output spaces. At least in the case of multiple scalar outputs, three common approaches can be used.

1. If the target function is of the form $f_* : \mathcal{X} \to \mathbb{R}^m$, then one can apply any method for learning a scalar function to the m scalar functions $f_1, \dots, f_m : \mathcal{X} \to \mathbb{R}$, defined by $f_i(x) = f_*(x)_i$, $i = 1, \dots, m$. While this is straightforward, it is clear that the learning method cannot profit from any known connection between the m scalar functions.

2. Consider the same situation as above. We cannot interpret f_* as a scalar function on an extended input set, $\tilde{f} : \{1, \dots, m\} \times \mathcal{X} \to \mathbb{R}$, defined by $g_*(i, x) = f_*(x)_i$ for $i \in \{1, \dots, m\}$ and $x \in \mathcal{X}$. Any data set generated by f_* can be transformed into a data set generated by g_*, to which a learning method for scalar functions can be applied. A learned function $\hat{g}$ can then be transformed into an estimate of f_* by setting $\hat{f} : \mathcal{X} \to \mathbb{R}^m$, $\hat{f}(x) = \begin{pmatrix} \hat{g}(1, x) & \cdots & \hat{g}(m, x) \end{pmatrix}^\top$.

3. There exist a generalization of GP regression to multiple outputs, cf. [10], and by considering a vector-RKHS [153, 138], also kernel ridge regression can be directly generalized to multiple output (actually arbitrary Hilbert-space valued outputs).

Note that 1. can be interpreted as a special case of 2., which in turn can be interpreted as a special case of 3.

In light of item 1. above, and to keep the following developments to a reasonable length, we will only consider learning scalar-valued functions. The (mostly immediate) generalization to multiple outputs and general vector-valued functions is left for future work.

Measurement functionals and connection to inverse problems So far, we have considered the measurement model $y = f_*(x) + \eta$, where f_* is the unknown target function, x some input, η unknown noise, and y the actual measurement value. This means that we access the target f_* through *evaluation functions*. In many application scenarios different measurement functionals appears. For example, suppose that $f_* : (0,1) \to \mathbb{R}$ is differentiable, then we might receive derivative measurements, corresponding to the measurement model $y = f'(x)+\eta$. Similarly, one could consider integral functionals (say, a convolution with a test function). As is well-known, GP regression supports a variety of such generalized measurement functionals, cf. e.g. [166, Section 9.4], although making this setup precise requires some care. The situation is easier for kernel ridge regression. As is clear from the exposition in Section 5.2, kernel ridge regression is a special case of regularized least-squares in Hilbert spaces with bounded linear measurement functionals. In the case of standard kernel ridge regression, these measurement functionals are just evaluations, which are continuous due to the RKHS assumption. As long as the measurement functional has a Riesz representer which can be easily evaluated, one can adapt kernel ridge regression and its associated theory. For example, if a kernel on $\mathbb{R}^d$ is sufficiently regular, then it has a derivative reproducing property, cf. [189, Lemma 4.34., Corollary 4.36], and hence linear partial differential measurement operators can be used with kernel ridge regression, similar to hard shape constrained kernel machines [20]. In particular, most of the developments in the next chapter generalize immediately to this setting.

Furthermore, this also indicates a close relation to inverse problems [213]. In inverse problems, one has a forward map $A : X \to Y$, an unknown element $x \in X$, and potentially noisy measurements from the forward map, $y = Ax + \eta$. The goal is then to recover x from the measured data, i.e., *invert* the data generating process. In general, this is challenging since information is often lost in the forward map or it cannot be inverted in a stable manner. For example, x might be a parameter set for a system of ordinary or partial differential equations, and A corresponds to the solu-

tion operator of a corresponding initial value or boundary value problem, potentially combined with an additional measurement or sampling operator. The inverse problem is therefore to recover the parameters from measurements of the evolved system. From this perspective, we can interpret our setting as a specific setup of an inverse problem, and indeed kernel ridge regression corresponds to Tikhonov regularization from the inverse problem literature. As a consequence, the following developments can be interesting for applications in uncertainty quantification in inverse problems, but due to our focus on applications to systems and control, we do not follow this direction in the present thesis.

Related methods In statistics, GP regression is often called *kriging*. More precisely, GP regression with a zero mean prior is known as *simple kriging*, GP regression with a constant, but unknown mean (which is jointly inferred, so we are in a semiparametric situation) is called *ordinary kriging*, and GP regression with a mean modelled by a linear (in the weights) model is called *universal kriging*, cf. e.g. [215].

Furthermore, GP regression also works under the assumption of no measurements noise (requiring in general a positive definite kernel), cf. [102, Section 3.1]. This situation appears when GPs are used as *surrogate models* [83], which happens in particular in the context of *computer experiments* [174]. In this case, the posterior mean coincides with the minimum norm interpolator over the RKHS induced by the covariance function [102, Section 3], as considered in scattered data approximation [217]. Furthermore, in the latter field the basic error bounds involve the *power function*, which (potentially up to some normalization) corresponds to the square of the posterior covariance function, cf. [217, Chapter 11].

4.3. Uncertainty bounds for GP regression

We are now ready to introduce the type of uncertainty bounds we need. Our primary motivation from the application side is learning-enhanced robust controller synthesis. As opposed to adaptive control or online controller tuning, the learning part happens here in an offline setting.[6] Furthermore, we consider the challenging setting were the unknown target is a function, as oppposed to a mere finite collection of parameters.

[6] Curiously, many of the bounds considered later on are anytime valid and can hence be even used in an online context.

As discussed in Section 4.2.3, we restrict us to scalar-valued functions. In addition, we assume access to noise-free inputs and noise-corrupted outputs. While the case of noisy inputs, or even non-observable input-output relations (requiring e.g. state estimation techniques), is very interesting, it is beyond the scope of the present thesis and left for future work. Finally, motivated by the discussion in Section 4.1, we aim at frequentist uncertainty bounds, and with a view towards applications in robust controller synthesis, they should be uniform in the inputs.

Let us formalize all of this. The unknown ground truth is a function $f_* : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is some input set (usually $\mathbb{R}^d$), for which we might have additional prior knowledge, e.g. in the form of regularity properties, or membership in a certain function space. We assume access to a data set $\mathcal{D} = ((x_1, y_1), \dots, (x_N, y_N))$ with $y_n = f_*(x_n) + \eta_n$, where $\eta_1, \dots, \eta_N$ is stochastic noise with some independence assumption. We make no assumptions on the inputs $x_1, \dots, x_N \in \mathcal{X}$, in particular, they might not be stochastic at all. This is important since in the context of learning based control, the inputs might have been generated by the underlying dynamical system, or an active exploration algorithm. We apply GP regression to this data set, assuming in general a zero mean prior with covariance function k and noise variance $\lambda \in \mathbb{R}_{>0}$ (or the corresponding kernel ridge regression formulation), leading as outlined in Section 4.2 to the posterior mean and variance functions $\mu_{\mathcal{D}}$ and $\sigma^2_{\mathcal{D}}$. Consider now a function $\eta_{\mathcal{D}} : (0, 1) \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, which can only depend on $\mathcal{D}$ (possible through $\sigma^2_{\mathcal{D}}$) and reasonable assumptions about f_*, but not on noise realizations or even the target function f_*. The overall goal is the construction of a suitable $\eta_{\mathcal{D}}$ such that for all $\delta \in (0, 1)$ we have

$$\mathbb{P}[|f_*(x) - \mu_{\mathcal{D}}(x)| \leq \eta_{\mathcal{D}}(\delta, x) \quad \forall x \in \mathcal{X}] \geq 1 - \delta. \tag{4.19}$$

In words, we want a frequentist uncertainty bound (since the ground truth f_* is deterministic, and the randomness only enters through the noisy data generating process) that is uniform in the inputs. Slightly imprecisely, we call $\eta_{\mathcal{D}}$ itself a frequentist uncertainty bound.

4.4. Comments

The present chapter has been written from scratch by the author of this thesis. The overview in Section 4.1 has profited from discussions of the author with S. Trimpe and C.W. Scherer, though the present formalization has not appeared before. Section 4.2 is a standard introduction to GP and kernel ridge regression, and it has been specifically written for this thesis for the reader's convenience. The setup in Section 4.3 is standard, cf. the kernelized bandit literature [186, 54], and very frequently used in learning-based control, cf. e.g. [31].

5. Frequentist uncertainty bounds for kernel and GP regression: Theory

In this chapter, we consider theoretical aspects of frequentist uncertainty bounds for GP regression as outlined in Section 4.3. We start in Section 5.1 with simple uncertainty bounds that follow immediately from the closed form solution of the GP posterior mean (corresponding to the kernel ridge regression estimate) by separating terms involving the target function and terms involving only the noise, and applying standard concentration inequalities to the latter. To prepare the discussion of more advanced bounds, in Section 5.2 we interpret kernel ridge regression as a special case of regularized least-squares in Hilbert spaces. Furthermore, state-of-the-art uncertainty bounds are based on self-normalized concentration inequalities, which is by now an established technique in the theoretical machine learning community. However, while the application of this technique to regularized least-squares is straightforward, from the existing literature it is unclear how one can come up with these arguments. To remedy this problem, in Section 5.3 we present an elementary and self-contained derivation of uncertainty bounds based on self-normalization. The results are formulated in the context of regularized least-squares in Hilbert spaces, and in Section 5.4 we translate them into the GP regression setting. Finally, in Section 5.5 we consider uncertainty bounds under model misspecification. Relation to prior work, and the contributions of the author, are discussed in Section 5.6.

This chapter is based on, with some parts taken verbatim from, the work [CF12].

5.1. Simple frequentist uncertainty bounds

Consider the setting outlined in Section 4.2. Observe that for all $x \in \mathcal{X}$

$$|f(x) - \mu_{\mathcal{D}}(x)| \leq \left| f(x) - k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{f} \right| + \left| k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\eta} \right|, \tag{5.1}$$

so we can separate in the error term the influence of f and the noise. For the first term, one can use the following well-known result. It appears for example in the first part of the proof of [54, Theorem 2].

Lemma 5.1.1. Let k be a kernel on $\mathcal{X}$, $f \in H_k$, $x_1, \ldots, x_N \in \mathcal{X}$, and $\lambda \in \mathbb{R}_{>0}$. For all $x \in \mathcal{X}$ we have

$$|f(x) - k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{f}| \leq \|f\|_k \sqrt{k(x,x) - k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x)}, \tag{5.2}$$

where

$$k_{\mathcal{D}}(x) = \begin{pmatrix} k(x, x_1) \\ \vdots \\ k(x, x_N) \end{pmatrix}, \quad K_{\mathcal{D}} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix}, \quad \boldsymbol{f} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{pmatrix}.$$

If $K_{\mathcal{D}}$ is invertible, the result also holds for $\lambda = 0$.

The case of a positive definite kernel and $\lambda = 0$ forms the foundation for most error bounds in the scattered data approximation literature, cf. [71]. In this setting, with the Golub-Weinberger bound an even stronger result is available. Since a proof of Lemma 5.1.1 is less easily found in the literature, we provide one here.

Proof. Using the reproducing property of k, we have $f(x) = \langle f, k(\cdot, x) \rangle_k$ and, defining for brevity $\alpha(x) = (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x)$,

$$\begin{aligned} k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{f} &= \sum_{n=1}^{N} \alpha_n(x) f(x_n) = \sum_{n=1}^{N} \alpha_n(x) \langle f, k(\cdot, x_n) \rangle_k \\ &= \left\langle f, \sum_{n=1}^{N} \alpha_n(x) k(\cdot, x_n) \right\rangle_k, \end{aligned}$$

so we get by Cauchy-Schwarz

$$\begin{aligned}|f(x) - k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{f}| &= \left| \langle f, k(\cdot, x) - \sum_{n=1}^{N} \alpha_n(x) k(\cdot, x_n) \rangle \right| \\ &\leq \|f\|_k \left\| k(\cdot, x) - \sum_{n=1}^{N} \alpha_n(x) k(\cdot, x_n) \right\|_k .\end{aligned}$$

The claim now follows from

$$\begin{aligned}&\left\| k(\cdot, x) - \sum_{n=1}^{N} \alpha_n(x) k(\cdot, x_n) \right\|_k^2 \\ &= \langle k(\cdot, x), k(\cdot, x) \rangle_k - 2 \left\langle k(\cdot, x), \sum_{n=1}^{N} \alpha_n(x) k(\cdot, x_n) \right\rangle_k \\ &\quad + \left\langle \sum_{i=1}^{N} \alpha_i(x) k(\cdot, x_i), \sum_{j=1}^{N} \alpha_j(x) k(\cdot, x_j) \right\rangle_k \\ &= k(x, x) - 2 \sum_{n=1}^{N} \alpha_n(x) k(x, x_n) + \sum_{i,j=1}^{N} \alpha_i(x) \alpha_j(x) k(x_j, x_i) \\ &= k(x, x) - 2 k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x) \\ &\quad + k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} K_{\mathcal{D}} (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x) \\ &\leq k(x, x) - 2 k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x) \\ &\quad + k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} (K_{\mathcal{D}} + \lambda I) (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x) \\ &= k(x, x) - 2 k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x) + k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x),\end{aligned}$$

where we used in the inequality that $K_{\mathcal{D}} \preceq K_{\mathcal{D}} + \lambda I$ (here, $\preceq$ refers as usual to the semidefinite ordering of symmetric matrices). □

As a simple consequence, if we know a bound $B \in \mathbb{R}_{\geq 0}$ on the RKHS norm of f, i.e., $\|f\|_k \leq B$, then we have a uniform (in the inputs) uncertainty bound (dealing with the error arising from sampling f at only finitely many points). In the context of GP regression, the bound then takes the form $B\sigma_{\mathcal{D}}(x)\ \forall x \in \mathcal{X}$.

If we are interested in a *pointwise* uncertainty bound, and the noise terms $\eta_1, \ldots, \eta_N$ are independent of the inputs, then it is easy to derive such an uncertainty bound if appropriate concentration results are available for the noise terms. As an example,

here is a pointwise uncertainty bound for independent subgaussian noise terms. We are not aware of this particular result and its elementary proof, but it would not be suprising if it already appeared in the vast literature on kernel ridge regression and GP regression.

Proposition 5.1.2. Let $f \in H_k$ and $B \in \mathbb{R}_{\geq 0}$ with $\|f\|_k \leq B$, and consider data $\mathcal{D} = ((x_1, y_1), \dots, (x_N, y_N))$ with $y_n = f(x_n) + \eta_n$, $n = 1, \dots, N$, where $\eta_1, \dots, \eta_N$ are independent R-subgaussian random variables. For all $\delta \in (0, 1)$ and all $x \in \mathcal{X}$, we have

$$\mathbb{P}\left[|f(x) - \mu_{\mathcal{D}}(x)| \leq B\sigma_{\mathcal{D}}(x) + 2R\|(K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x)\| \sqrt{\ln(1/\delta)}\right] \geq 1 - \delta.$$

Proof. We start by using (5.1), and bound the first term with Lemma 5.1.1. For the second term, define $\alpha(x) = (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x)$, then we have

$$\left|k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\eta}\right| = \left|\sum_{n=1}^{N} \alpha_n(x)\eta_n\right|,$$

and the Hoeffding inequality for subgaussian random variables (e.g, [76, Theorem 7.27]) shows that for all $t \geq 0$ we have

$$\mathbb{P}\left[\left|\sum_{n=1}^{N} \alpha_n(x)\eta_n\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2R^2\|\alpha\|^2}\right).$$

Solving $2\exp\left(-\frac{t^2}{2R^2\|\alpha\|^2}\right) = \delta$ for t establishes the claim. □

As is clear from the proof, a corresponding result holds mutatis mutandis for subexponential noise, or more generally zero-mean noise variables with an appropriate bound on the moment generating function (MGF). This remark applies also to the remaining uncertainty bounds in this section.

For many applications, in particular in learning-based control, it is important to have uncertainty bounds that are uniform in the inputs. The most straightforward approach is to separate any input-dependent term from the noise terms in (5.1). For

example, we could use Cauchy-Schwarz to get

$$\begin{aligned}\left|k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}}+\lambda I)^{-1}\boldsymbol{\eta}\right| &= \left|k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}}+\lambda I)^{-\frac{1}{2}} (K_{\mathcal{D}}+\lambda I)^{-\frac{1}{2}}\boldsymbol{\eta}\right| \\ &\leq \|(K_{\mathcal{D}}+\lambda I)^{-\frac{1}{2}} k_{\mathcal{D}}(x)\| \|(K_{\mathcal{D}}+\lambda I)^{-\frac{1}{2}}\boldsymbol{\eta}\| \\ &= \sqrt{k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}}+\lambda I)^{-1} k_{\mathcal{D}}(x)}\sqrt{\boldsymbol{\eta}^\top (K_{\mathcal{D}}+\lambda I)^{-1}\boldsymbol{\eta}} \\ &= \sqrt{k(x,x)-\sigma_{\mathcal{D}}^2(x)}\sqrt{\boldsymbol{\eta}^\top (K_{\mathcal{D}}+\lambda I)^{-1}\boldsymbol{\eta}}.\end{aligned}$$

If we have a concentration result for the random quadratic form appearing on the right, then this leads immediately to an input-uniform uncertainty bound. The next result implements this approach. It appeared in slightly different form as [CF12, Proposition 2].

Proposition 5.1.3. In the situation of Theorem 5.1.2, for all $\delta \in (0,1)$ we have

$$\mathbb{P}[\forall x \in \mathcal{X}: \quad |f(x)-\mu_{\mathcal{D}}(x)| \leq B\sigma_{\mathcal{D}}(x) + R\sqrt{k(x,x)-\sigma_{\mathcal{D}}^2(x)}\, b(\delta)] \geq 1-\delta \tag{5.3}$$

with

$$b(\delta) = \sqrt{\operatorname{tr}(Q) + 2\sqrt{\operatorname{tr}(Q)\ln(1/\delta)} + 2\|Q\|\ln(1/\delta)} \tag{5.4}$$

$$Q = (K_{\mathcal{D}}+\lambda I)^{-1} \tag{5.5}$$

Proof. Combining (5.1), Lemma 5.1.1, and the bound we just derived leads to

$$|f(x)-\mu_{\mathcal{D}}(x)| \leq B\sigma_{\mathcal{D}}(x) + \sqrt{k(x,x)-\sigma_{\mathcal{D}}^2(x)}\sqrt{\boldsymbol{\eta}^\top (K_{\mathcal{D}}+\lambda I)^{-1}\boldsymbol{\eta}}.$$

Since by assumption $\eta_1, \ldots, \eta_N$ are independent R-subgaussian, we have for all $\nu \in \mathbb{R}^N$

$$\mathbb{E}\left[\exp\left(\sum_{n=1}^N \nu_n\eta_n\right)\right] = \prod_{n=1}^N \mathbb{E}\left[\exp(\nu_n\eta_n)\right] \leq \prod_{n=1}^N \exp\left(\frac{\nu_n^2R^2}{2}\right) = \exp\left(\frac{\|\nu\|^2R^2}{2}\right),$$

so [96, Theorem 2.1] is applicable, which states that for all $t \in \mathbb{R}_{>0}$ we have

$$\mathbb{P}\left[\boldsymbol{\eta}^\top Q\boldsymbol{\eta} \geq R^2\left(\operatorname{tr}(Q) + 2\sqrt{\operatorname{tr}(Q)\,t} + 2t\|Q\|\right)\right] \leq e^{-t},$$

where we defined $Q = (K_{\mathcal{D}} + \lambda I)^{-1}$ for brevity. Setting $e^{-t} = \delta$ and applying the square root to both sides establishes the result. □

Note that the form of the uncertainty bound is not satisfying. The part involving the noise arises from

$$\sqrt{k(x,x) - \sigma_{\mathcal{D}}^2(x)}\sqrt{\boldsymbol{\eta}^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\eta}},$$

and intuitively, with more data $\sigma_{\mathcal{D}}^2(x)$ should get smaller, so $\sqrt{k(x,x) - \sigma_{\mathcal{D}}^2(x)}$ gets larger. This effect might not be counteracted by the second factor. Assume that $K_{\mathcal{D}} = I$ (this happens for example if $\mathcal{X}$ is an inner product space, $k(x,x') = \langle x, x' \rangle$ is the linear kernel, and the covariates form an orthonormal system), then

$$\sqrt{\boldsymbol{\eta}^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\eta}} = \sqrt{\boldsymbol{\eta}^\top ((1+\lambda) I)^{-1} \boldsymbol{\eta}} = \sqrt{\sum_{n=1}^{N} \frac{1}{1+\lambda} \eta_n^2},$$

so in this case (roughly speaking) the second factor is even growing with more data.

5.2. Interlude: Regularized least-squares in Hilbert spaces and kernel ridge regression

First, we show that the ℓ_2-regularized least-squares approach for linear regression works in arbitrary Hilbert spaces, using an elementary derivation. We then rewrite the results in a particularly insightful form, which is again well-known, but unfortunately rarely utilized in the exposition of regularized least-squares in Hilbert spaces, and derive kernel ridge regression as a special case. Finally, we connect these developments back to uncertainty bounds.

Regularized least-squares in Hilbert spaces Let H be a real Hilbert space[1] and consider a data set $\mathcal{D}_N = ((h_1, y_1), \ldots, (h_N, y_N)) \in (H \times \mathbb{R})^N$. Given a *regularization parameter* $\lambda \in \mathbb{R}_{>0}$, consider the optimization problem

$$\min_{h \in H} \sum_{n=1}^{N} (y_n - \langle h, h_n \rangle)^2 + \lambda \|h\|^2, \tag{5.6}$$

[1] Everything would also work for complex Hilbert spaces.

which is obviously convex. Observe now that for all $h \in H$

$$\begin{aligned}\sum_{n=1}^{N}(y_n - \langle h, h_n\rangle)^2 + \lambda\|h\|^2 &= \sum_{n=1}^{N}\langle h, h_n\rangle\langle h, h_n\rangle + \lambda\langle h, h\rangle - 2\sum_{n=1}^{N} y_n\langle h, h_n\rangle + \sum_{n=1}^{N} y_n^2 \\ &= \left\langle h, \sum_{n=1}^{N}\langle h, h_n\rangle h_n \right\rangle + \langle h, (\lambda\mathrm{id})h\rangle - 2\left\langle h, \sum_{n=1}^{N} y_n h_n \right\rangle + \sum_{n=1}^{N} y_n^2 \\ &= \left\langle h, \left(\sum_{n=1}^{N} h_n \otimes h_n + \lambda\mathrm{id}\right) h \right\rangle - 2\left\langle h, \sum_{n=1}^{N} y_n h_n \right\rangle + \sum_{n=1}^{N} y_n^2,\end{aligned}$$

where we used the common notation $u \otimes v := h \mapsto \langle h, u\rangle v$ in the last step.

Furthermore, if Q is a self-adjoint, invertible operator on H, and R some element of H, then we have for all $h \in H$ that

$$\langle h - R, Q(h - R)\rangle = \langle h, Qh\rangle - 2\langle h, QR\rangle + \langle R, QR\rangle,$$

so by identifying

$$Q = \sum_{n=1}^{N} h_n \otimes h_n + \lambda\mathrm{id}, \qquad h_\lambda = \left(\sum_{n=1}^{N} h_n \otimes h_n + \lambda\mathrm{id}\right)^{-1} \sum_{n=1}^{N} y_n h_n$$

(note that this Q is indeed self-adjoint and invertible), we can complete the square and find that

$$\begin{aligned}&\left\langle h, \left(\sum_{n=1}^{N} h_n \otimes h_n + \lambda\mathrm{id}\right) h \right\rangle - 2\left\langle h, \sum_{n=1}^{N} y_n h_n \right\rangle + \sum_{n=1}^{N} y_n^2 \\ &\quad = \left\langle h - h_\lambda, \left(\sum_{n=1}^{N} h_n \otimes h_n + \lambda\mathrm{id}\right)(h - h_\lambda) \right\rangle + \text{Terms without } h.\end{aligned}$$

This shows that h_λ is the unique solution of the optimization problem (5.6).

A different perspective We now provide a different perspective on the regularized least-squares solution h_λ by rewriting it. The developments in this section are probably well-known, and the following computations are similar to the first part of the proof of [54, Theorem 2], but we could not locate a reference that presents this in the context of regularized least-squares in Hilbert spaces.

Define two linear maps

$$S : \mathbb{R}^N \to H, \quad S\alpha = \sum_{n=1}^{N} \alpha_n h_n \tag{5.7}$$

$$A : H \to \mathbb{R}^N, \quad Ah = \begin{pmatrix} \langle h, h_1 \rangle \\ \vdots \\ \langle h, h_N \rangle \end{pmatrix}. \tag{5.8}$$

In the context of signal processing and harmonic analsyis, these maps are known as the *synthesis* and *analysis operators*, respectively [211]. As is well-known, $S^* = A$, which can be seen from

$$\langle S\alpha, h \rangle_H = \left\langle \sum_{n=1}^{N} \alpha_n h_n, h \right\rangle_H = \sum_{n=1}^{N} \alpha_n \langle h_n, h \rangle_H = \alpha^\top \begin{pmatrix} \langle h, h_1 \rangle_H \\ \vdots \\ \langle h, h_N \rangle_H \end{pmatrix} = \langle \alpha, Ah \rangle_{\mathbb{R}^N}.$$

Furthermore,

$$\left(\sum_{n=1}^{N} h_n \otimes h_n \right) h = \sum_{n=1}^{N} \langle h, h_n \rangle h_n = S \left(\begin{pmatrix} \langle h, h_1 \rangle \\ \vdots \\ \langle h, h_N \rangle \end{pmatrix} \right) = (SA)(h)$$

$$\sum_{n=1}^{N} y_n h_n = S\boldsymbol{y},$$

where we defined $\boldsymbol{y} = \begin{pmatrix} y_1 & \cdots & y_N \end{pmatrix}^\top$. Finally, since for the $\mathbb{R}^N$-standard ONB vectors e_i, e_j

$$\langle e_i, (AS)e_j \rangle_{\mathbb{R}^N} = \langle e_i, A(Se_j) \rangle_{\mathbb{R}^N} = \langle e_i, A(h_j) \rangle_{\mathbb{R}^N} = \langle h_j, h_i \rangle_H,$$

we can identify the linear map $AS : \mathbb{R}^N \to \mathbb{R}^N$ with the Gram matrix $H =$

$\left(\langle h_j, h_i\rangle_H\right)_{i,j=1,\dots,N}$. With these preparations, we have

$$\begin{aligned} h_\lambda &= \left(\sum_{n=1}^N h_n \otimes h_n + \lambda \mathrm{id}\right)^{-1} \sum_{n=1}^N y_n h_n = (SA + \lambda \mathrm{id}_H)^{-1} S\boldsymbol{y} \\ &= S(AS + \lambda \mathrm{id}_{\mathbb{R}^N})^{-1}\boldsymbol{y} = \begin{pmatrix} h_1 & \cdots & h_N \end{pmatrix} (H + \lambda I)^{-1}\boldsymbol{y}, \end{aligned}$$

where we used the well-known pull-through identity in the third equality.

Defining $\alpha = (H + \lambda I)^{-1}\boldsymbol{y}$, we can rewrite this as

$$h_\lambda = \sum_{n=1}^N \alpha_n h_n, \tag{5.9}$$

which says that the regularized least-squares solution is a linear combination of the inputs (or covariates) in the data set. In the context of linear regression in $\mathbb{R}^d$, this is discussed in standard textbooks, for example [89, Chapter 3]. In our setting, it has the additional consequence that even if H is infinite-dimensional, h_λ is contained in the finite-dimensional span of the inputs from the data set.

Recovering kernel ridge regression Suppose now that we are not interested in h_λ per se, but rather in one or more linear functionals of it. Restricting us to bounded linear functionals and using the Riesz representation theorem, we can focus on $\langle h, h_\lambda\rangle$, where h is the Riesz representer of the functional of interest. The developments in the preceding section now lead to

$$\begin{aligned} \langle h, h_\lambda\rangle &= \langle h, S(AS + \lambda I)^{-1}\boldsymbol{y}\rangle = \left\langle h, \sum_{n=1}^N \alpha_n h_n \right\rangle = \sum_{n=1}^N \alpha_n \langle h, h_n\rangle \\ &= \begin{pmatrix} \langle h, h_1\rangle & \cdots & \langle h, h_N\rangle \end{pmatrix} (AS + \lambda I)^{-1}\boldsymbol{y} = (Ah)^\top (AS + \lambda \mathrm{id}_{\mathbb{R}^N})^{-1}\boldsymbol{y}. \end{aligned}$$

Observe that in this expression of $\langle h, h_\lambda\rangle$ no element from H appears on its own anymore, but only in inner products. This is reminiscent of the kernel perspective from Section 2.3.

Let now k be a kernel on some set $\mathcal{X}$, and consider the situation $H = H_k$, $h_n = k(\cdot, x_n)$, and $h = k(\cdot, x)$. This means that the linear functionals of interest are evaluations of elements from H_k, and since H_k is an RKHSs, these are continuous

and can hence be represented by scalar products with elements from H_k, cf. Sections 2.1 and 2.2. The regularized least-squares problem now becomes

$$\min_{h \in H} \sum_{n=1}^{N} (y_n - \langle h, h_n \rangle)^2 + \lambda \|h\|_H^2 = \min_{f \in H_k} \sum_{n=1}^{N} (y_n - f(x))^2 + \lambda \|f\|_k^2,$$

and we get

$$H = \Big(\langle h_j, h_i \rangle \Big)_{i,j=1,\dots,N} = \Big(\langle k(\cdot, x_j), k(\cdot, x_i) \rangle_k \Big)_{i,j=1,\dots,N} = \Big(k(x_i, x_j) \Big)_{i,j=1,\dots,N}$$

and

$$\begin{aligned} h_\lambda(x) &= \langle k(\cdot, x), h_\lambda \rangle_k = \Big(\langle k(\cdot, x), k(\cdot, x_1) \rangle \quad \cdots \quad \langle k(\cdot, x), k(\cdot, x_N) \rangle \Big) (H + \lambda I)^{-1} \boldsymbol{y} \\ &= k_{\mathcal{D}}(x)(K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{y}, \end{aligned}$$

so we recovered kernel ridge regression as a special case of regularized least-squares in Hilbert spaces.

Note that in contrast to more conventional expositions of kernel ridge regression, like [102, Section 3], we did not use the general representer theorem, nor any existence or uniqueness results for solutions of kernel machines.

Back to uncertainty bounds As a simple application of the preceding developments, let us go back to Proposition 5.1.3. Recall that we were not satisfied with the form of the uncertainty bound, mostly due to the factor $\sqrt{k(x,x) - \sigma_{\mathcal{D}}^2(x)}$, which in turn arises from the usage of Cauchy-Schwarz. Using the interpretation of kernel ridge regression as regularized least-squares in a Hilbert space, we can try to apply

Cauchy-Schwarz to a different inner product,

$$\begin{aligned}
\left|k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\eta}\right| &= \left|A(k(\cdot, x))^\top (AS + \lambda \mathrm{id}_{\mathbb{R}^N})^{-1} \boldsymbol{\eta}\right| \\
&= \left|\langle Ak(\cdot, x), (AS + \lambda \mathrm{id}_{\mathbb{R}^N})^{-1} \boldsymbol{\eta}\rangle_{\mathbb{R}^N}\right| \\
&= \left|\langle k(\cdot, x), S(AS + \lambda \mathrm{id}_{\mathbb{R}^N})^{-1} \boldsymbol{\eta}\rangle_k\right| \\
&= \left|\langle k(\cdot, x), (SA + \lambda \mathrm{id}_H)^{-1} S\boldsymbol{\eta}\rangle_k\right| \\
&= \left|\langle (SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} k(\cdot, x), (SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} S\boldsymbol{\eta}\rangle_k\right| \\
&\leq \|(SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} k(\cdot, x)\|_k \|(SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} S\boldsymbol{\eta}\|_k.
\end{aligned}$$

Since

$$\begin{aligned}
\|(SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} S\boldsymbol{\eta}\|_k &= \sqrt{\langle (SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} S\boldsymbol{\eta}, (SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} S\boldsymbol{\eta}\rangle_k} \\
&= \sqrt{\langle \eta, A(SA + \lambda \mathrm{id}_H)^{-1} S\boldsymbol{\eta}\rangle_{\mathbb{R}^N}} \\
&= \sqrt{\langle \eta, AS(AS + \lambda \mathrm{id}_H)^{-1} \boldsymbol{\eta}\rangle_k} \\
&= \sqrt{\langle \eta, K_{\mathcal{D}}(K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\eta}\rangle_{\mathbb{R}^N}}
\end{aligned}$$

and $A(AS + \lambda \mathrm{id}_H)^{-1} S : \mathbb{R}^N \to \mathbb{R}^N$ is a self-adjoint, positive semidefinite map that can be identified with a positive definite matrix, the second term $\|(SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} S\boldsymbol{\eta}\|_k$ can be expressed with (finite) linear algebra expressions, and it is the square root of a random quadratic form, so the strategy from the proof of Proposition 5.1.3 is still applicable.

To deal with the first factor $\|(SA+\lambda \mathrm{id}_H)^{-\frac{1}{2}} k(\cdot, x)\|_k$, we recognize that our current bounding strategy is the same as in the proof of [54, Theorem 2], so we can use a

calculation which appears there,

$$\begin{aligned}
&\langle (SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} k(\cdot, x), (SA + \lambda \mathrm{id}_H)^{-\frac{1}{2}} k(\cdot, x) \rangle_k \\
&\quad = \langle k(\cdot, x), (SA + \lambda \mathrm{id}_H)^{-1} k(\cdot, x) \rangle_k \\
&\quad = \frac{1}{\lambda} \langle k(\cdot, x), (SA + \lambda \mathrm{id}_H)^{-1} (\lambda \mathrm{id}_H) k(\cdot, x) \rangle_k \\
&\quad = \frac{1}{\lambda} \langle k(\cdot, x), (SA + \lambda \mathrm{id}_H)^{-1} (SA + \lambda \mathrm{id}_H) k(\cdot, x) - (SA + \lambda \mathrm{id}_H)^{-1} SA k(\cdot, x) \rangle_k \\
&\quad = \frac{1}{\lambda} \Big(\langle k(\cdot, x), k(\cdot, x) \rangle_k - \langle k(\cdot, x), (SA + \lambda \mathrm{id}_H)^{-1} SA k(\cdot, x) \rangle_k \Big) \\
&\quad = \frac{1}{\lambda} \Big(k(x, x) - \langle k(\cdot, x), S(AS + \lambda I)^{-1} A k(\cdot, x) \rangle_k \Big) \\
&\quad = \frac{1}{\lambda} \Big(k(x, x) - \langle Ak(\cdot, x), (AS + \lambda I)^{-1} A k(\cdot, x) \rangle_{\mathbb{R}^N} \Big) \\
&\quad = \frac{1}{\lambda} \Big(k(x, x) - k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} k_{\mathcal{D}}(x) \Big) \\
&\quad = \frac{1}{\lambda} \sigma_{\mathcal{D}}^2(x).
\end{aligned}$$

Altogether we get

$$\left| k_{\mathcal{D}}(x)^\top (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\eta} \right| \leq \sqrt{\eta^\top K_{\mathcal{D}} (K_{\mathcal{D}} + \lambda I)^{-1} \boldsymbol{\eta}} \frac{\sigma_{\mathcal{D}}(x)}{\sqrt{\lambda}}.$$

A variant of Proposition 5.1.3 for this bounding strategy then becomes the following.

Proposition 5.2.1. In the situation of Theorem 5.1.2, for all $\delta \in (0, 1)$ we have

$$\mathbb{P}\left[\forall x \in \mathcal{X} : \quad |f(x) - \mu_{\mathcal{D}}(x)| \leq \left(B + \frac{R}{\sqrt{\lambda}} b(\delta) \right) \sigma_{\mathcal{D}}(x) \right] \geq 1 - \delta \tag{5.10}$$

with

$$b(\delta) = \sqrt{\mathrm{tr}\,(Q) + 2\sqrt{\mathrm{tr}\,(Q) \ln(1/\delta)} + 2\|Q\| \ln(1/\delta)} \tag{5.11}$$

and

$$Q = K_{\mathcal{D}} (K_{\mathcal{D}} + \lambda I)^{-1}. \tag{5.12}$$

The proof is completely analogous to the one of Proposition 5.1.3. Before moving on, we would like to record some remarks on this result.

Remark 5.2.2. 1. This result follows by changing the choice of concentration inequality in the proof of [54, Theorem 2], and keeping the data dependent terms

(instead of trying to bound them further with theoretically more amenable terms), and as such is in the same spirit as [CF12, Theorem 1].

2. The form of the uncertainty bound is $\beta \cdot \sigma_{\mathcal{D}}$ for some appropriate (input-independent) β, and as such is already of the standard form for uncertainty bounds for GP regression.

3. Furthermore, a very similar result, based on a completely different proof strategy is [205, Theorem 1]. Under essentially the same assumptions the righthand side in the bound becomes

$$\left(B + \frac{R}{\sqrt{\lambda}}\sqrt{2\ln(1/\delta)}\right)\sigma_{\mathcal{D}}(x). \tag{5.13}$$

5.3. An elementary derivation of frequentist uncertainty bounds based on self-normalization

The frequentist uncertainty bounds for GP regression (or equivalently, kernel ridge regression) most frequently used in learning-based control are [186, Theorem 6] and [54, Theorem 2] (as well as the results in [CF12] based on the latter reference) have been developed in the context of kernelized bandits. The current state-of-the-art for such uncertainty bounds has been achieved via *self-normalization*, first in [1, Chapter 3] as a direct generalization of the finite-dimensional case from [2], and then recently rediscovered by [219][2]. Self-normalized random variables and stochastic processes, and corresponding tail and concentration inequalities, form a very rich subject, with roots going back to the Student t-statistic [157]. The results in [2, 1, 219] are based on a particular technique known as *pseudomaximization* or *the method of mixtures*, sometimes also called *Laplace's method* [128]. In the context of self-normalization this technique goes back to [156], with the generalization to the multivariate case in [155]. Using this technique it is relatively straightforward to derive the relevant concentration inequalities in [2] and the corresponding generalizations. While pseudomaximization is motivated by a somewhat transparent intuition

[2]Similar to the results in [1, Chapter 3], in [128] frequentist bounds for GP regression have been derived as a generalization of the results in [2]. However, the latter reference relied on a Mercer expansion, and the resulting bound is slightly less general than the results in [1, Chapter 3] and [219].

[154], it appears to be very unclear how it leads relatively directly to the state-of-the-art uncertainty bounds. To the best of our knowledge, there is no exposition or derivation available that *explains* why or how this particular instance of pseudomaximization appears in the analysis of regularized least-squares (and eventually kernel ridge regression and GP regression).

To close this gap, in the following we provide a self-contained, elementary derivation. Every step will be intuitive and clear to anyone who has a basic background knowledge of probability theory and exponential concentration inequalities via Chernoff's method, at the level of [63, Chapter 1]. In particular, *the particular method of mixtures will be "rediscovered" or derived along the way.* In other words, our derivation shows one way how one can come up with the relevant arguments.

Before we start with our derivation, a final comment on the context of these results. The main motivation of works like [2] (and the predecessor [186]) is to provide *time-uniform frequentist uncertainty bounds.* This means that the data is generated sequentially, so the data set grows in discrete time steps, and the uncertainty bounds should hold in all time-steps (uniform in time)[3]. Interestingly, the following derivation *does not aim at time uniformity*, but the resulting bounds are easily transformed into time-uniform bounds in the end. In particular, this shows that the self-normalization structure arises naturally in the context of regularized least-squares, and the time uniformity appears to be secondary.

Setup Consider the following situation. Let $\mathcal{X} \neq \emptyset$ be some input set and $f_* : \mathcal{X} \to \mathbb{R}$ an unknown target function, which is accessible through noisy evaluations. Let $\mathcal{D}_N = ((x_1, y_1), \dots, (x_N, y_N))$ be a data set, assuming the noise model $y_n = f(x_n) + \eta_n$. Furthermore, $\mathcal{D}_N$ might have been generated interactively, i.e., input $x_n \in \mathcal{X}$ might depend on $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$. Let $\hat{f}_N$ be the outcome of some learning algorithm. We are interested in a probabilistic uncertainty bound *uniform in the input*, i.e., we want some $B_N : (0,1) \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ such that

$$\mathbb{P}\left[|\hat{f}_N(x) - f_*(x)| \leq B_N(\delta, x) \quad \forall x \in \mathcal{X}\right] \geq 1 - \delta \tag{5.14}$$

[3] The original literature on modern self-normalization and pseudomaximization also considered the continuous-time case in the context of sufficiently regular continuous-time martingales, cf. [157] for an overview and many pointers to the relevant literature.

holds for all $\delta \in (0,1)$. Importantly, B_N must only depend on $\mathcal{D}_N$ and known reasonable properties of f_* and the noise, but not directly on f_* or $\boldsymbol{\eta}_N = \begin{pmatrix} \eta_1 & \cdots & \eta_N \end{pmatrix}$, both of which are unknown.

We restrict us now to the model class $f_\theta : H \to \mathbb{R}$, $f_\theta(h) = \langle h, \theta \rangle$, parametrized by $\theta \in H$, where H is a Hilbert space (so for now we assume that $\mathcal{X} = H$). Furthermore, we use regularized least-squares, so $\hat{f} = \hat{f}_{N,\rho}$ with $\hat{f}_{N,\rho}(h) = \langle h, \hat{\theta}_{N,\rho} \rangle$, where $\hat{\theta}_{N,\rho}$ is the unique solution to the optimization problem

$$\min_{\theta \in H} \sum_{n=1}^{N} (y_n - f_\theta(x_n))^2 + \rho \|\theta\|^2. \tag{5.15}$$

Note that in contrast to Section 5.2, we use ρ instead of λ for the regularization parameter to conform with the literature on sequential least-squares and self-normalized concentration inequalities.

Since working with Hilbert spaces can be technically demanding, we start with $H = \mathbb{R}^d$. In this case, the solution to the regularized least-squares problem is given by

$$\hat{\theta}_{N,\rho} = (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{y}_N, \tag{5.16}$$

where we defined

$$\boldsymbol{X}_N = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix}, \quad \boldsymbol{y}_N = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}. \tag{5.17}$$

Recall that our goal for now is an uncertainty bound uniform in the inputs. Probably the most simple strategy to achieve this is to remove any input-dependent term[4] from the terms involving the noise, just as we did in Proposition 5.1.3. For all $x \in \mathbb{R}^d$ we have

$$|\hat{f}_{N,\rho}(x) - f_*(x)| = |\langle x, \hat{\theta}_{N,\rho} \rangle - \langle x, \theta_* \rangle| = |\langle x, \theta_{N,\rho} - \theta_* \rangle| \leq \|x\| \|\theta_{N,\rho} - \theta_*\|,$$

[4]Here we mean the test input after learning, not the inputs in the data set.

which suggests that we should look for a bound on $\|\theta_{N,\rho} - \theta_*\|$. Using

$$\begin{aligned}\|\theta_{N,\rho} - \theta_*\| &= \|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{y}_N - \theta_*\| \\ &= \|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top (\boldsymbol{X}_N \theta_* + \boldsymbol{\eta}_N) - \theta_*\| \\ &= \|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{X}_N \theta_* + (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N - \theta_*\| \\ &\leq \|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| + \left\| \left((\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{X}_N - I \right) \theta_* \right\|\end{aligned}$$

we can separate the noise term from the ground truth θ_*.

Probabilistic bound for $d = 1$ Our first goal is a probabilistic bound on the first term above. To simplify things, we start with the case $d = 1$, so we want to upper bound

$$\mathbb{P}\left[\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| \geq \beta \right] = \mathbb{P}\left[\left| \frac{A}{B + \rho} \right| \geq \beta \right],$$

where we defined for brevity $A = \boldsymbol{X}_N^\top \boldsymbol{\eta}_N$ and $B = \boldsymbol{X}_N^\top \boldsymbol{X}_N$. Here, $\beta \in \mathbb{R}_{\geq 0}$ is an appropriate uncertainty bound, i.e., a scalar which can only depend on $\mathcal{D}_N$ (which is known), but not $\boldsymbol{\eta}_N$ (which is never known in this setup). Ideally, we have an exponential bound, so we will try the well-known Chernoff technique [210, Chapter 2].

First, we should get rid of the absolute value inside the event, which can complicate things. To do so, we square both sides,

$$\mathbb{P}\left[\left| \frac{A}{B + \rho} \right| \geq \beta \right] = \mathbb{P}\left[\frac{A^2}{(B + \rho)^2} \geq \beta^2 \right].$$

Next, in order to apply the (generalized) Markov inequality, we have to ensure that on the righthand side nothing random is left anymore. It is reasonable to assume that our uncertainty bound has the form $\beta^2 = RD$, where R contains all randomness[5], and D is a purely deterministic term (containing e.g. the confidence level δ). We can interpret the term R as a modulator, that adjusts a fixed bound D based on random, but known information. Using the usual Chernoff technique leads to

$$\mathbb{P}\left[\frac{A^2}{(B + \rho)^2} \geq RD \right] = \mathbb{P}\left[\exp\left(\frac{A^2}{R(B + \rho)^2} \right) \geq \exp(D) \right] \leq \exp(-D) \mathbb{E}\left[\exp\left(\frac{A^2}{R(B + \rho)^2} \right) \right].$$

[5] However, this term needs to be computable without knowledge of $\boldsymbol{\eta}$.

However, in the following derivation a second random term will appear, inside the expectation, but outside the exponential in the expectation, and this second term needs to be removed in the end. This means that we should use the ansatz $\beta^2 = R_1(R_2 + D)$ instead, with R_1, R_2 containing all randomness, and D deterministic. This leads to

$$\begin{aligned}
\mathbb{P}\left[\frac{A^2}{(B+\rho)^2} \geq \beta^2\right] &= \mathbb{P}\left[\frac{A^2}{(B+\rho)^2} \geq R_1(R_2 + D)\right] \\
&= \mathbb{P}\left[\exp\left(\frac{A^2}{R_1(B+\rho)^2}\right) \geq \exp(D + R_2)\right] \\
&= \mathbb{P}\left[\exp\left(\frac{A^2}{R_1(B+\rho)^2}\right)\frac{1}{\exp(R_2)} \geq \exp(D)\right] \\
&\leq \exp(-D)\mathbb{E}\left[\exp\left(\frac{A^2}{R_1(B+\rho)^2}\right)\frac{1}{\exp(R_2)}\right].
\end{aligned}$$

We now need to upper bound the expectation, using properties of the random variables A and B. However, these are not independent, and they even appear in a fraction. To make progress, we have to get rid of the latter. As a first step, we need additional freedom inside the expectation. A classic technique is to put a 1 inside it,

$$\mathbb{E}\left[\exp\left(\frac{A^2}{R_1(B+\rho)^2}\right)\frac{1}{\exp(R_2)}\right] = \mathbb{E}\left[1 \cdot \exp\left(\frac{A^2}{R_1(B+\rho)^2}\right)\frac{1}{\exp(R_2)}\right],$$

and then expand the 1. One classic option is $1 = C/C$ with a constant C. Another option is $1 = \int 1 \mathrm{d}\mu$ with a freely chosen probability measure μ. Since we are anyway inside an expectation, we choose this latter option,

$$\mathbb{E}\left[1 \cdot \exp\left(\frac{A^2}{R_1(B+\rho)^2}\right)\frac{1}{\exp(R_2)}\right] = \mathbb{E}\left[\int 1 \mathrm{d}\mu \cdot \exp\left(\frac{A^2}{R_1(B+\rho)^2}\right)\frac{1}{\exp(R_2)}\right],$$

and since we need to access the exponential, we use a Borel measure of the form

$\mu(A) = C_\mu \int_A \exp(H(\lambda))\mathrm{d}\lambda$, where C_μ is a normalization constant, so we are at

$$\begin{aligned}
&\mathbb{E}\left[1 \cdot \exp\left(\frac{A^2}{R_1(B+\rho)^2}\right)\frac{1}{\exp(R_2)}\right] \\
&\quad = \mathbb{E}\left[C_\mu \int \exp(H(\lambda))\mathrm{d}\lambda \exp\left(\frac{A^2}{R_1(B+\rho)^2}\right)\frac{1}{\exp(R_2)}\right] \\
&\quad = \mathbb{E}\left[C_\mu \int \exp\left(H(\lambda) + \frac{A^2}{R_1(B+\rho)^2}\right)\mathrm{d}\lambda \frac{1}{\exp(R_2)}\right].
\end{aligned}$$

We can now use the term $H(\lambda)$ inside the exponential to get rid of the problematic fraction. A classic way to do so is to use *completing the square*. If $\gamma, \delta \in \mathbb{R}$ are constants, then we have for all $\lambda \in \mathbb{R}$ that

$$\begin{aligned}
-\gamma\lambda^2 + \delta\lambda &= -\gamma\left(\lambda^2 - 2\lambda\frac{\delta}{2\gamma} + \left(\frac{\delta}{2\gamma}\right)^2\right) - (-\gamma)\left(\frac{\delta}{2\gamma}\right)^2 \\
&= -\gamma\left(\lambda - \frac{\delta}{2\gamma}\right)^2 + \frac{\delta^2}{4\gamma}
\end{aligned}$$

as long as $\gamma \neq 0$. To ensure

$$\frac{\delta^2}{4\gamma} = \frac{A^2}{R_1(B+\rho)^2}$$

we can set

$$\delta = A, \quad 4\gamma = R_1(B+\rho)^2$$

as long as $R_1 > 0$, and by choosing

$$H(\lambda) = -\gamma\left(\lambda - \frac{\delta}{2\gamma}\right)^2 = -\frac{1}{4}(4\gamma)\left(\lambda - 2\frac{\delta}{4\gamma}\right)^2 = -\frac{1}{4}R_1(B+\rho)^2\left(\lambda - 2\frac{A}{R_1(B+\rho)^2}\right)^2$$

we get

$$H(\lambda) + \frac{A^2}{R_1(B+\rho)^2} = -\gamma\lambda^2 + \delta\lambda = -\frac{1}{4}(4\gamma)\lambda^2 + \delta\lambda = -\frac{1}{4}R_1(B+\rho)^2\lambda^2 + A\lambda.$$

Furthermore, we immediately recognize that μ is a Gaussian measure[6], and

$$\begin{aligned} H(\lambda) &= -\frac{1}{4}R_1(B+\rho)^2\left(\lambda - 2\frac{A}{R_1(B+\rho)^2}\right)^2 \\ &= -\frac{1}{2}\frac{1}{\left(\frac{R_1}{2}(B+\rho)^2\right)^{-1}}\left(\lambda - 2\frac{A}{R_1(B+\rho)^2}\right)^2 \end{aligned}$$

shows that

$$\mu = \mathcal{N}\left(2\frac{A}{R_1(B+\rho)^2}, \left(\frac{R_1}{2}(B+\rho)^2\right)^{-1}\right),$$

so

$$C_\mu = \frac{1}{\sqrt{2\pi}}\left(\left(\frac{R_1}{2}(B+\rho)^2\right)^{-1}\right)^{-\frac{1}{2}} = \frac{1}{\sqrt{2\pi}}\left(\frac{R_1}{2}(B+\rho)^2\right)^{\frac{1}{2}}.$$

By setting $R_1 = 2R_1'$, we can simplify the terms (by getting rid of the various 2s), and end up with

$$\begin{aligned} &\mathbb{E}\left[C_\mu \int \exp\left(H(\lambda) + \frac{A^2}{R_1(B+\rho)^2}\right) \mathrm{d}\lambda \frac{1}{\exp(R_2)}\right] \\ &\quad = \mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\sqrt{R_1'(B+\rho)^2}\int \exp\left(-\frac{1}{2}R_1'(B+\rho)^2\lambda^2 + A\lambda\right)\mathrm{d}\lambda\frac{1}{\exp(R_2)}\right]. \end{aligned}$$

Unfortunately, at this stage B and ρ are entangled in the term $(B+\rho)^2$, and a condition on A, B that involves the regularization parameter ρ appears to be unnatural. The problem stems from the squaring of $B+\rho$, but we can get rid of this by setting $R_1' = (B+\rho)^{-1}$, which leads to

$$\mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\sqrt{B+\rho}\int \exp\left(-\frac{1}{2}(B+\rho)\lambda^2 + A\lambda\right)\mathrm{d}\lambda\frac{1}{\exp(R_2)}\right]$$

To make the integral more tractable, let us try to turn it into a Gaussian integral

[6] Note that μ is random since it depends on A and B, but since overall we are inside an expectation, this is not a problem.

again,

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\sqrt{B+\rho}\int \exp\left(-\frac{1}{2}(B+\rho)\lambda^2+A\lambda\right)\mathrm{d}\lambda\frac{1}{\exp(R_2)}\right]\\
&\quad=\mathbb{E}\left[\frac{1}{\sqrt{2\pi}}\sqrt{B+\rho}\int \exp\left(-\frac{1}{2}\rho\lambda^2\right)\exp\left(-\frac{1}{2}B\lambda^2+A\lambda\right)\mathrm{d}\lambda\frac{1}{\exp(R_2)}\right]\\
&\quad=\mathbb{E}\left[\sqrt{B+\rho}\sqrt{\rho^{-1}}\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\rho^{-1}}}\int \exp\left(-\frac{1}{2\rho^{-1}}\lambda^2\right)\exp\left(-\frac{1}{2}B\lambda^2+A\lambda\right)\mathrm{d}\lambda\frac{1}{\exp(R_2)}\right]\\
&\quad=\mathbb{E}\left[\sqrt{\frac{B+\rho}{\rho}}\frac{1}{\exp(R_2)}\int \mathcal{N}(\lambda\mid 0,\rho^{-1})\exp\left(-\frac{1}{2}B\lambda^2+A\lambda\right)\mathrm{d}\lambda\right].
\end{aligned}$$

Observe now that if nothing random outside the integral is left, then we could interchange the integral with the expectation, and we end up with an MGF-type[7] bound. But we can achieve this easily by setting

$$R_2=\ln\left(\sqrt{\frac{B+\rho}{\rho}}\right),$$

so we get

$$\mathbb{E}\left[\int \mathcal{N}(\lambda\mid 0,\rho^{-1})\exp\left(-\frac{1}{2}B\lambda^2+A\lambda\right)\mathrm{d}\lambda\right]=\int \mathbb{E}\left[\exp\left(-\frac{1}{2}B\lambda^2+A\lambda\right)\right]\mathcal{N}(\lambda\mid 0,\rho^{-1})\mathrm{d}\lambda.$$

Assuming now an MGF-type bound

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}B\lambda^2+A\lambda\right)\right]\leq b(\lambda), \tag{5.18}$$

we end up with

$$\int \mathbb{E}\left[\exp\left(-\frac{1}{2}B\lambda^2+A\lambda\right)\right]\mathcal{N}(\lambda\mid 0,\rho^{-1})\mathrm{d}\lambda\leq\int b(\lambda)\mathcal{N}(\lambda\mid 0,\rho^{-1})\mathrm{d}\lambda=:C_\lambda.$$

Summing up, we have

$$\beta=\sqrt{R_1(R_2+D)}=\sqrt{\frac{2}{B+\rho}\ln\left(\sqrt{\frac{B+\rho}{\rho}}\right)},$$

[7] Moment generating function (MGF)

and we find that for all $D \geq 0$ we have

$$\mathbb{P}\left[\left|\frac{A}{B+\rho}\right| \geq \sqrt{\frac{2}{B+\rho}\ln\left(\sqrt{\frac{B+\rho}{\rho}}+D\right)}\right] \leq C_\lambda \exp(-D).$$

Assuming $C_\lambda > 0$, we can rewrite this in confidence form. Let $\delta \in (0,1)$, then

$$\delta = C_\lambda \exp(-D) \quad \Leftrightarrow \quad D = -\ln\left(\frac{\delta}{C_\lambda}\right) = \ln\left(\frac{C_\lambda}{\delta}\right),$$

which results in

$$\begin{aligned}
&\mathbb{P}\left[\left|\frac{A}{B+\rho}\right| \geq \sqrt{\frac{2}{B+\rho}\ln\left(\sqrt{\frac{B+\rho}{\rho}}+D\right)}\right] \\
&\quad = \mathbb{P}\left[\left|\frac{A}{B+\rho}\right| \geq \sqrt{\frac{2}{B+\rho}\ln\left(\sqrt{\frac{B+\rho}{\rho}}+\ln\left(\frac{C_\lambda}{\delta}\right)\right)}\right] \\
&\quad = \mathbb{P}\left[\left|\frac{A}{B+\rho}\right| \geq \sqrt{\frac{2}{B+\rho}\ln\left(\sqrt{\frac{C_\lambda}{\delta}\frac{B+\rho}{\rho}}\right)}\right] \\
&\quad \leq \delta.
\end{aligned}$$

Observe that we did not use the particular forms $A = \boldsymbol{X}_N \boldsymbol{\eta}_N$ and $B = \sum_{n=1}^N x_n^2$, but only that B is nonnegative. For convenience, let us summarize this result.

Lemma 5.3.1. Let A be a scalar random variable and B a nonnegative random variable. Assume that there exists a measurable $b : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}B\lambda^2 + A\lambda\right)\right] \leq b(\lambda) \quad \forall \lambda \in \mathbb{R}, \tag{5.19}$$

then for all $\rho \in \mathbb{R}_{>0}$ such that $0 < C_\lambda < \infty$, where

$$C_\lambda = \int b(\lambda)\mathcal{N}(\lambda \mid 0, 1/\rho)\mathrm{d}\lambda, \tag{5.20}$$

we have for all $\delta \in (0,1)$ that

$$\mathbb{P}\left[\left|\frac{A}{B+\rho}\right| \geq \sqrt{\frac{2}{B+\rho}\ln\left(\sqrt{\frac{C_\lambda}{\delta}}\frac{B+\rho}{\rho}\right)}\right] \leq \delta. \tag{5.21}$$

An MGF-type bound for $d = 1$ Next, let us try to find appropriate conditions ensuring (5.18). To make things easy, let us start with $N = 1$, so we need to bound

$$\begin{aligned}\mathbb{E}\left[\exp\left(-\frac{1}{2}B\lambda^2 + A\lambda\right)\right] &= \mathbb{E}\left[\exp\left(-\frac{1}{2}x_1^2\lambda^2 + \eta_1 x_1\lambda\right)\right] \\ &= \mathbb{E}\left[\exp\left(-\frac{1}{2}x_1^2\lambda^2\right)\exp\left(\eta_1 x_1\lambda\right)\right].\end{aligned}$$

Observe that the first factor looks like the classic MGF-bound of a subgaussian random variable (just with a minus), and the second term looks like the MGF of η_1 (at $x_1\lambda$). If x_1 were not random, then we could impose a subgaussianity assumption on η_1 and get a simple bound. In the present setup, x_1 might be random, but we can make it practically non-random by conditioning on it inside the expectation, using the tower property of the conditional expectation,

$$\begin{aligned}\mathbb{E}\left[\exp\left(-\frac{1}{2}x_1^2\lambda^2\right)\exp\left(\eta_1 x_1\lambda\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(-\frac{1}{2}x_1^2\lambda^2\right)\exp\left(\eta_1 x_1\lambda\right) \mid x_1\right]\right] \\ &= \mathbb{E}\left[\exp\left(-\frac{1}{2}x_1^2\lambda^2\right)\mathbb{E}\left[\exp\left(\eta_1 x_1\lambda\right) \mid x_1\right]\right] \\ &\leq \mathbb{E}\left[\exp\left(-\frac{1}{2}x_1^2\lambda^2\right)\exp\left(\frac{R^2x_1^2\lambda^2}{2}\right)\right] \\ &= \mathbb{E}\left[\exp\left(-\frac{x_1^2\lambda^2}{2}(1-R^2)\right)\right],\end{aligned}$$

where in the inequality we imposed the assumption that η_1 is conditionally (on x_1) R-subgaussian for some $R \in \mathbb{R}_{\geq 0}$, i.e.,

$$\mathbb{E}[\exp(\nu\eta_1) \mid x_1] \leq \exp\left(\frac{R^2\nu^2}{2}\right) \quad \forall \nu \in \mathbb{R}.$$

The preceding statement holds only almost surely w.r.t. the underlying probability measure, but since this does not pose a problem in the present context, here and in

the following we will omit this qualification. Since we need a bound that holds for all $\lambda \in \mathbb{R}$, we should ensure $1 - R^2 \geq 0$ (i.e., $R \leq 1$), so that

$$\mathbb{E}\left[\exp\left(-\frac{x_1^2\lambda^2}{2}(1-R^2)\right)\right] \leq \mathbb{E}\left[\exp\left(-\frac{x_1^2\lambda^2}{2}\cdot 0\right)\right] = 1.$$

In other words, for the case $N = 1$, if η_1 is conditionally (on x_1) 1-subgaussian, then we can use the bound $b(\lambda) = 1$ in (5.18).

Let us consider the case of general N. Since we mastered the case $N = 1$, it is sensible to proceed inductively, for which we should turn matrix-vector and matrix-matrix products into sums. As is well-known, we have

$$A = \boldsymbol{X}_N\boldsymbol{\eta}_N = \sum_{n=1}^{N}\eta_n x_n, \qquad B = \boldsymbol{X}_N^\top\boldsymbol{X}_N = \sum_{n=1}^{N}x_n^2.$$

Therefore, our goal is to bound

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}B\lambda^2 + A\lambda\right)\right] = \mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2\sum_{n=1}^{N}x_n^2 + \lambda\sum_{n=1}^{N}\eta_n x_n\right)\right].$$

Let us apply the strategy from above to η_N,

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2\sum_{n=1}^{N}x_n^2 + \lambda\sum_{n=1}^{N}\eta_n x_n\right)\right] \\
&\quad= \mathbb{E}\left[\exp\left(\sum_{n=1}^{N}-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right)\right] \\
&\quad= \mathbb{E}\left[\prod_{n=1}^{N}\exp\left(-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right)\right] \\
&\quad= \mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2 x_N^2 + \lambda\eta_N x_N\right)\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right)\right] \\
&\quad= \mathbb{E}\left[\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2 x_N^2\right)\exp\left(\lambda\eta_N x_N\right)\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right) \mid \mathcal{F}_N\right]\right] \\
&\quad= \mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2 x_N^2\right)\mathbb{E}\left[\exp\left(\lambda\eta_N x_N\right) \mid \mathcal{F}_N\right]\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right)\right],
\end{aligned}$$

where we defined the filtration $\mathcal{F}_n = \sigma(x_1, \eta_1, x_2, \eta_2, \ldots, \eta_{n-1}, x_n)$. Assume now that η_N is conditionally (on $\mathcal{F}_N$) 1-subgaussian, i.e.,

$$\mathbb{E}[\exp(\nu\eta_N) \mid \mathcal{F}_N] \leq \exp\left(\frac{R^2\nu^2}{2}\right) \quad \forall \nu \in \mathbb{R}$$

holds for $R = 1$. We then get

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2 x_N^2\right)\mathbb{E}\left[\exp\left(\lambda\eta_N x_N\right) \mid \mathcal{F}_N\right]\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right)\right] \\
&\quad\leq \mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2 x_N^2\right)\exp\left(\frac{\lambda^2 x_N^2}{2}\right)\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right)\right] \\
&\quad= \mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2 x_N^2 + \frac{\lambda^2 x_N^2}{2}\right)\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right)\right] \\
&\quad= \mathbb{E}\left[\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}\lambda^2 x_n^2 + \lambda\eta_n x_n\right)\right],
\end{aligned}$$

and we can proceed inductively. To summarize: If for $n = 1, \ldots, N$ the noise variable η_n is 1-subgaussian conditional on $\mathcal{F}_n$, then

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}B\lambda^2 + A\lambda\right)\right] = \mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^2\sum_{n=1}^{N}x_n^2 + \lambda\sum_{n=1}^{N}\eta_n x_n\right)\right] \leq 1.$$

Probabilistic bound for general d Let us generalize our probabilistic bound to the case of general d, so we want to upper bound

$$\mathbb{P}\left[\|(\boldsymbol{X}_N^\top\boldsymbol{X}_N + \rho I)^{-1}\boldsymbol{X}_N^\top\boldsymbol{\eta}_N\| \geq \beta\right].$$

Inspecting our derivation for the case $d = 1$ shows that almost all steps can in principle be done also for general d (e.g., by using a multivariate Gaussian distribution instead of a scalar one), only our choice of $R_1 = 2(B + \rho)^{-1}$ is not permissible anymore. The reason is that R_1 has to be scalar, but $(B+\rho)^{-1}$ becomes the matrix-valued term $(\boldsymbol{X}_N^\top\boldsymbol{X}_N+\rho I)^{-1}$ for general d. The previous choice of R_1 was necessary to avoid the term $(B + \rho)^2$ (instead of $B + \rho$). However, the problematic choice of

R_1 is not necessary if we start with

$$\mathbb{P}\left[\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| \geq \beta\right] = \mathbb{P}\left[\left|\frac{A}{\sqrt{B+\rho}}\right| \geq \beta\right],$$

instead of

$$\mathbb{P}\left[\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| \geq \beta\right] = \mathbb{P}\left[\left|\frac{A}{B+\rho}\right| \geq \beta\right].$$

Translating to the case of general d, this means we should try to upper bound

$$\mathbb{P}\left[\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| \geq \beta\right] = \mathbb{P}\left[\|(B+\rho I)^{-\frac{1}{2}} A\| \geq \beta\right],$$

where we defined $A = \boldsymbol{X}_N^\top \boldsymbol{\eta}_N$ and $B = \boldsymbol{X}_N^\top \boldsymbol{X}_N$ for general d. We will now adapt our derivation from above to this case and deal with the missing $(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}}$ later on.

Using again the ansatz $\beta^2 = R_1(R_2 + D)$ (with R_1, R_2 potentially random and D deterministic) for the desired uncertainty bound β, we get

$$\begin{aligned}
\mathbb{P}\left[\|(B+\rho I)^{-\frac{1}{2}} A\| \geq \beta\right] &= \mathbb{P}\left[\|(B+\rho I)^{-\frac{1}{2}} A\|^2 \geq R_1(R_2 + D)\right] \\
&= \mathbb{P}\left[\exp\left(\frac{1}{R_1}\|(B+\rho I)^{-\frac{1}{2}} A\|^2\right)\frac{1}{\exp(R_2)} \geq D\right] \\
&\leq \exp(-D)\mathbb{E}\left[\exp\left(\frac{1}{R_1}\|(B+\rho I)^{-\frac{1}{2}} A\|^2\right)\frac{1}{\exp(R_2)}\right]
\end{aligned}$$

Using a probability distribution $\mu(A) = C_\mu \int_A \exp(H(\lambda))\mathrm{d}\lambda$, now on $\mathbb{R}^d$, we get

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(\frac{1}{R_1}\|(B+\rho I)^{-\frac{1}{2}} A\|^2\right)\frac{1}{\exp(R_2)}\right] \\
&\quad = \mathbb{E}\left[C_\mu \int \exp(H(\lambda))\mathrm{d}\lambda \exp\left(\frac{1}{R_1}\|(B+\rho I)^{-\frac{1}{2}} A\|^2\right)\frac{1}{\exp(R_2)}\right] \\
&\quad = \mathbb{E}\left[C_\mu \int \exp\left(H(\lambda) + \frac{1}{R_1}\|(B+\rho I)^{-\frac{1}{2}} A\|^2\right)\mathrm{d}\lambda \frac{1}{\exp(R_2)}\right].
\end{aligned}$$

We have

$$\begin{aligned}\frac{1}{R_1}\|(B+\rho I)^{-\frac{1}{2}}A\|^2 &= \frac{1}{R_1}\langle (B+\rho I)^{-\frac{1}{2}}A, (B+\rho I)^{-\frac{1}{2}}A\rangle \\ &= \frac{1}{R_1}A^\top (B+\rho I)^{-1}A = A^\top (R_1(B+\rho I))^{-1}A\end{aligned}$$

which suggests using again completing the square, this time in $\mathbb{R}^d$. For a symmetric matrix $\Gamma \in \mathbb{R}^{d\times d}$ and a vector $\Delta \in \mathbb{R}^d$, we find for all $\lambda \in \mathbb{R}^d$ that

$$\begin{aligned}-\lambda^\top\Gamma\lambda + \Delta^\top\lambda &= -\left(\lambda^\top\Gamma\lambda - 2\left(\frac{1}{2}\Gamma^{-1}\Delta\right)^\top\Gamma\lambda + \left(\frac{1}{2}\Gamma^{-1}\Delta\right)^\top\Gamma\left(\frac{1}{2}\Gamma^{-1}\Delta\right)\right) \\ &\quad + \left(\frac{1}{2}\Delta\Gamma^{-1}\right)^\top\Gamma\left(\frac{1}{2}\Delta\Gamma^{-1}\right) \\ &= -\left(\lambda - \left(\frac{1}{2}\Gamma^{-1}\Delta\right)\right)^\top\Gamma\left(\lambda - \left(\frac{1}{2}\Gamma^{-1}\Delta\right)\right) + \frac{1}{4}\Delta^\top\Gamma^{-1}\Delta \\ &= -\left(\lambda - 2\left(4\Gamma\right)^{-1}\Delta\right)^\top\Gamma\left(\lambda - 2\left(4\Gamma\right)^{-1}\Delta\right) + \Delta^\top\left(4\Gamma\right)^{-1}\Delta,\end{aligned}$$

as long as Γ is invertible. We can therefore choose

$$4\Gamma = R_1(B+\rho I), \quad \Delta = A,$$

and

$$\begin{aligned}H(\lambda) &= -\left(\lambda - 2\left(4\Gamma\right)^{-1}\Delta\right)^\top\Gamma\left(\lambda - 2\left(4\Gamma\right)^{-1}\Delta\right) \\ &= -\frac{1}{4}\left(\lambda - 2\left(R_1(B+\rho I)\right)^{-1}\Delta\right)^\top\left(R_1(B+\rho I)\right)\left(\lambda - 2\left(R_1(B+\rho I)\right)^{-1}\Delta\right),\end{aligned}$$

which leads to

$$\begin{aligned}H(\lambda) + A^\top\left(R_1(B+\rho I)\right)^{-1}A &= H(\lambda) + \Delta^\top\left(4\Gamma\right)^{-1}\Delta \\ &= -\lambda^\top\Gamma\lambda + \Delta\lambda \\ &= -\frac{1}{4}\lambda^\top(4\Gamma)\lambda + \Delta\lambda \\ &= -\frac{1}{4}\lambda^\top(R_1(B+\rho I))\lambda + A\lambda\end{aligned}$$

As in the scalar case, this turns μ into a Gaussian measure with quadratic term

$$\begin{aligned} H(\lambda) &= -\frac{1}{4}\left(\lambda - 2\left(R_1(B+\rho I)\right)^{-1}\Delta\right)^\top \left(R_1(B+\rho I)\right)\left(\lambda - 2\left(R_1(B+\rho I)\right)^{-1}\Delta\right) \\ &= -\frac{1}{2}\left(\lambda - \left(\frac{R_1}{2}(B+\rho I)\right)^{-1} A\right)^\top \left(\frac{2}{R_1}(B+\rho I)^{-1}\right)^{-1}\left(\lambda - \left(\frac{R_1}{2}(B+\rho I)\right)^{-1} A\right), \end{aligned}$$

and in contrast to our previous derivation, we can simplify it by directly setting $R_1 = 2$, leading to

$$H(\lambda) = -\frac{1}{2}\left(\lambda - (B+\rho I)^{-1} A\right)^\top \left((B+\rho I)^{-1}\right)^{-1}\left(\lambda - (B+\rho I)^{-1} A\right).$$

This shows that

$$\mu = \mathcal{N}\left((B+\rho I)^{-1} A, (B+\rho I)^{-1}\right),$$

so

$$C_\mu = (2\pi)^{-\frac{d}{2}} \det\left((B+\rho I)^{-1}\right)^{-\frac{1}{2}} = (2\pi)^{-\frac{d}{2}} \det\left(B+\rho I\right)^{\frac{1}{2}}.$$

Furthermore, with this choice of R_1 we get

$$H(\lambda) + A^\top \left(R_1(B+\rho I)\right)^{-1} A = -\frac{1}{2}\lambda^\top (B+\rho I)\lambda + A\lambda.$$

Altogether, we are at

$$\begin{aligned} &\mathbb{E}\left[C_\mu \int \exp\left(H(\lambda) + \frac{1}{R_1}\|(B+\rho I)^{-\frac{1}{2}} A\|^2\right) \mathrm{d}\lambda \frac{1}{\exp(R_2)}\right] \\ &\quad = \mathbb{E}\left[(2\pi)^{-\frac{d}{2}} \det\left(B+\rho I\right)^{\frac{1}{2}} \int \exp\left(-\frac{1}{2}\lambda^\top (B+\rho I)\lambda + A\lambda\right) \mathrm{d}\lambda \frac{1}{\exp(R_2)}\right]. \end{aligned}$$

As in the scalar case, we can turn this into a Gaussian integral again,

$$\begin{aligned}
&\mathbb{E}\left[(2\pi)^{-\frac{d}{2}}\det(B+\rho I)^{\frac{1}{2}}\int\exp\left(-\frac{1}{2}\lambda^\top(B+\rho I)\lambda+A\lambda\right)\mathrm{d}\lambda\frac{1}{\exp(R_2)}\right]\\
&\quad=\mathbb{E}\left[\frac{1}{\exp(R_2)}(2\pi)^{-\frac{d}{2}}\det(B+\rho I)^{\frac{1}{2}}\int\exp\left(-\frac{1}{2}\lambda^\top(\rho I)\lambda\right)\exp\left(-\frac{1}{2}\lambda^\top B\lambda+A\lambda\right)\mathrm{d}\lambda\right]\\
&\quad=\mathbb{E}\left[\frac{1}{\exp(R_2)}\det(B+\rho I)^{\frac{1}{2}}\det((\rho I)^{-1})^{\frac{1}{2}}(2\pi)^{-\frac{d}{2}}\det((\rho I)^{-1})^{-\frac{1}{2}}\right.\\
&\qquad\qquad\left.\times\int\exp\left(-\frac{1}{2}\lambda^\top\left((\rho I)^{-1}\right)^{-1}\lambda\right)\exp\left(-\frac{1}{2}\lambda^\top B\lambda+A\lambda\right)\mathrm{d}\lambda\right]\\
&\quad=\mathbb{E}\left[\frac{1}{\exp(R_2)}\sqrt{\det(B+\rho I)/\det(\rho I)}\int\exp\left(-\frac{1}{2}\lambda^\top B\lambda+A\lambda\right)\mathcal{N}(\lambda\mid 0,(\rho I)^{-1})\mathrm{d}\lambda\right].
\end{aligned}$$

To remove the random term outside the integral (which is necessary for interchanging integration and expectation), we choose

$$R_2=\ln\left(\sqrt{\frac{\det(B+\rho I)}{\det(\rho I)}}\right),$$

and we get

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{\exp(R_2)}\sqrt{\det(B+\rho I)/\det(\rho I)}\int\exp\left(-\frac{1}{2}\lambda^\top B\lambda+A\lambda\right)\mathcal{N}(\lambda\mid 0,(\rho I)^{-1})\mathrm{d}\lambda\right]\\
&\quad=\mathbb{E}\left[\int\exp\left(-\frac{1}{2}\lambda^\top B\lambda+A\lambda\right)\mathcal{N}(\lambda\mid 0,(\rho I)^{-1})\mathrm{d}\lambda\right]\\
&\quad=\int\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^\top B\lambda+A\lambda\right)\right]\mathcal{N}(\lambda\mid 0,(\rho I)^{-1})\mathrm{d}\lambda.
\end{aligned}$$

Assuming again a bound of the form

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^\top B\lambda+A\lambda\right)\right]\leq b(\lambda)\quad\forall\lambda\in\mathbb{R}^d,$$

we get

$$\int\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^\top B\lambda+A\lambda\right)\right]\mathcal{N}(\lambda\mid 0,(\rho I)^{-1})\mathrm{d}\lambda\leq\int b(\lambda)\mathcal{N}(\lambda\mid 0,(\rho I)^{-1})\mathrm{d}\lambda=:C_\lambda.$$

Altogether (since $\beta = \sqrt{R_1(R_2 + D)}$) we find that for all $D \in \mathbb{R}_{\geq 0}$ we have

$$\mathbb{P}\left[\|(B+\rho I)^{-\frac{1}{2}}A\| \geq \beta\right] = \mathbb{P}\left[\|(B+\rho I)^{-\frac{1}{2}}A\| \geq \sqrt{2\left(\ln\left(\sqrt{\frac{\det(B+\rho I)}{\det(\rho I)}}\right) + D\right)}\right] \leq C_\lambda \exp(-D).$$

Just as in the scalar case, we can turn this into confidence form. Finally, note that we did not use the particular structure of $A = \boldsymbol{X}_N^\top \boldsymbol{\eta}_N$ and $B = \boldsymbol{X}_N^\top \boldsymbol{X}_N$, but just the fact that B is positive semidefinite. Let us summarize this result.

Lemma 5.3.2. Let A be an $\mathbb{R}^d$-valued and B a symmetric and positive semidefinite random variable of dimension $d \times d$. Assume that there exists some measurable $b : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^\top B\lambda + A\lambda\right)\right] \leq b(\lambda) \quad \forall \lambda \in \mathbb{R}^d, \tag{5.22}$$

then for all $\rho \in \mathbb{R}_{>0}$ such that $0 < C_\lambda < \infty$, where

$$C_\lambda = \int b(\lambda)\mathcal{N}(\lambda \mid 0, (\rho I)^{-1})\mathrm{d}\lambda, \tag{5.23}$$

we have for all $\delta \in (0, 1)$ that

$$\mathbb{P}\left[\|(B+\rho I)^{-\frac{1}{2}}A\| \geq \sqrt{2\ln\left(\frac{C_\lambda}{\delta}\sqrt{\frac{\det(B+\rho I)}{\det(\rho I)}}\right)}\right] \leq \delta. \tag{5.24}$$

The MGF-type bound for general d Let us check whether we can find a suitable bound (5.22) for our case of $A = \boldsymbol{X}_N^\top \boldsymbol{\eta}_N$ and $B = \boldsymbol{X}_N^\top \boldsymbol{X}_N$. Motivated by our derivation in the scalar case, we rewrite A and B as

$$A = \boldsymbol{X}_N^\top \boldsymbol{\eta}_N = \sum_{n=1}^N \eta_n x_n$$
$$B = \boldsymbol{X}_N^\top \boldsymbol{X}_N = \sum_{n=1}^N x_n x_n^\top.$$

For an arbitrary $\lambda \in \mathbb{R}^d$ we find that

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^\top B\lambda + A\lambda\right)\right] &= \mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^\top\left(\sum_{n=1}^N x_n x_n^\top\right)\lambda + \lambda^\top \sum_{n=1}^N \eta_n x_n\right)\right] \\
&= \mathbb{E}\left[\exp\left(\sum_{n=1}^N -\frac{1}{2}\lambda^\top x_n x_n^\top + \lambda^\top(\eta_n x_n)\right)\right] \\
&= \mathbb{E}\left[\exp\left(\sum_{n=1}^N -\frac{1}{2}(\lambda^\top x_n)^2 + (\lambda^\top x_n)\eta_n\right)\right] \\
&= \mathbb{E}\left[\prod_{n=1}^N \exp\left(-\frac{1}{2}(\lambda^\top x_n)^2 + (\lambda^\top x_n)\eta_n\right)\right],
\end{aligned}$$

which shows that our argument for the scalar case applies also for $\mathbb{R}^d$. So, define again the filtration $\mathcal{F}_n = \sigma(x_1, \eta_1, x_2, \eta_2, \ldots, \eta_{n-1}, x_n)$, and assume that η_n is conditionally (on $\mathcal{F}_n$) 1-subgaussian, i.e.,

$$\mathbb{E}[\exp(\nu\eta_n) \mid \mathcal{F}_n] \leq \exp\left(\frac{R^2\nu^2}{2}\right) \quad \forall \nu \in \mathbb{R}$$

holds for $R = 1$. Let $\lambda \in \mathbb{R}^d$ be arbitrary, then we get

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^\top B\lambda + A\lambda\right)\right] = \mathbb{E}\left[\prod_{n=1}^N \exp\left(-\frac{1}{2}(\lambda^\top x_n)^2 + (\lambda^\top x_n)\eta_n\right)\right] \\
&= \mathbb{E}\left[\exp\left(-\frac{1}{2}(\lambda^\top x_N)^2\right)\exp\left((\lambda^\top x_N)\eta_N\right)\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}(\lambda^\top x_n)^2 + (\lambda^\top x_n)\eta_n\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\exp\left(-\frac{1}{2}(\lambda^\top x_N)^2\right)\exp\left((\lambda^\top x_N)\eta_N\right)\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}(\lambda^\top x_n)^2 + (\lambda^\top x_n)\eta_n\right) \mid \mathcal{F}_N\right]\right] \\
&= \mathbb{E}\left[\exp\left(-\frac{1}{2}(\lambda^\top x_N)^2\right)\mathbb{E}\left[\exp\left((\lambda^\top x_N)\eta_N\right) \mid \mathcal{F}_N\right]\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}(\lambda^\top x_n)^2 + (\lambda^\top x_n)\eta_n\right)\right] \\
&\leq \mathbb{E}\left[\exp\left(-\frac{1}{2}(\lambda^\top x_N)^2\right)\exp\left(\frac{(\lambda^\top x_N)^2}{2}\right)\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}(\lambda^\top x_n)^2 + (\lambda^\top x_n)\eta_n\right)\right] \\
&= \mathbb{E}\left[\prod_{n=1}^{N-1}\exp\left(-\frac{1}{2}(\lambda^\top x_n)^2 + (\lambda^\top x_n)\eta_n\right)\right] \\
&\leq \ldots \leq 1.
\end{aligned}$$

To summarize, we have established a bound like (5.22) for arbitrary $N \in \mathbb{N}_+$.

An uncertainty bound for regularized least-squares in $\mathbb{R}^d$ It is time to put everything together. Assuming that η_n is conditionally (on $\mathcal{F}_n$) 1-subgaussian for all $n = 1, \dots, N$, we find that for all $\delta \in (0, 1)$

$$\mathbb{P}\left[\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| \geq \sqrt{2 \ln\left(\frac{1}{\delta}\sqrt{\frac{\det\left(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I\right)}{\det(\rho I)}}\right)}\right] \leq \delta.$$

If the noise terms η_n are conditionally R-subgaussian (for $R \in \mathbb{R}_{>0}$), then η_n/R is conditionally 1-subgaussian, so in this case we get

$$\begin{aligned}\mathbb{P}&\left[\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top (\boldsymbol{\eta}_N/R)\| \geq \sqrt{2 \ln\left(\frac{1}{\delta}\sqrt{\frac{\det\left(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I\right)}{\det(\rho I)}}\right)}\right] \\ &= \mathbb{P}\left[\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| \geq R\sqrt{2 \ln\left(\frac{1}{\delta}\sqrt{\frac{\det\left(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I\right)}{\det(\rho I)}}\right)}\right] \leq \delta.\end{aligned}$$

Going back to our starting point

$$\begin{aligned}|\hat{f}_{N,\rho}(x) - f_*(x)| &= |\langle x, \theta_{N,\rho} - \theta_*\rangle| \leq \|x\|\|\theta_{N,\rho} - \theta_*\| \\ &\leq \|x\|\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| + \|x\| \left\|\left((\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{X}_N - I\right)\theta_*\right\|,\end{aligned}$$

we notice that we cannot use our probabilistic bound yet, since the latter works for $\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\|$ and not $\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\|$. However, this can

be easily repaired in the application of Cauchy-Schwarz,

$$\begin{aligned}
|\hat{f}_{N,\rho}(x) - f_*(x)| &= |\langle x, \theta_{N,\rho} - \theta_* \rangle| = \left| \left\langle x, (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \boldsymbol{X}_N^\top \boldsymbol{y}_N - \theta_* \right\rangle \right| \\
&= \left| \left\langle (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} x, (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{y}_N - (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{\frac{1}{2}} \theta_* \right\rangle \right| \\
&\leq \|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} x\| \|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{y}_N - (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{\frac{1}{2}} \theta_* \| \\
&\leq \|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} x\| \Big(\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| \\
&\qquad + \|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{X}_N \theta_* - (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{\frac{1}{2}} \theta_* \| \Big)
\end{aligned}$$

The last term can be simplified,

$$\begin{aligned}
&\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{X}_N \theta_* - (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{\frac{1}{2}} \theta_* \| \\
&\quad = \| \left((\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{X}_N - (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{\frac{1}{2}} \right) \theta_* \| \\
&\quad = \| \left((\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I) - \rho (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} - (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{\frac{1}{2}} \right) \theta_* \| \\
&\quad = \|\rho (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \theta_* \| = \rho \sqrt{\theta_*^\top (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \theta_*}.
\end{aligned}$$

Since θ_* is unknown, the last term might still be somewhat problematic. However, since $\boldsymbol{X}_N^\top \boldsymbol{X}_N$ is positive semidefinite, the spectrum of $\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I$ of lower-bounded by ρ, and hence the spectrum of $(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1}$ is upper bounded by ρ^{-1}, so we get

$$\rho \sqrt{\theta_*^\top (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} \theta_*} \leq \rho \sqrt{\rho^{-1} \|\theta_*\|^2} = \sqrt{\rho} \|\theta_*\|,$$

and the norm of θ_* is probably the most reasonable bound in this context. To summarize,

$$|\hat{f}_{N,\rho}(x) - f_*(x)| \leq \sqrt{x^\top (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} x} \left(\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| + \sqrt{\rho} \|\theta_*\| \right).$$

Since $(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1}$ is positive definite, $(h, h') \mapsto h^\top (\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-1} h'$ defines an inner product, whose induced norm we denote by $\|\cdot\|_{N,\rho}$. With this notation, we have

$$|\hat{f}_{N,\rho}(x) - f_*(x)| \leq \|x\|_{N,\rho} \left(\|(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I)^{-\frac{1}{2}} \boldsymbol{X}_N^\top \boldsymbol{\eta}_N\| + \sqrt{\rho} \|\theta_*\| \right).$$

Finally, since the input x does not appear in the probabilistic bound we derived above, altogether we find that for all $\delta \in (0,1)$, with probability at least $1-\delta$ we have for all $x \in \mathcal{X}$ that

$$|\hat{f}_{N,\rho}(x) - f_*(x)| \leq \|x\|_{N,\rho} \left(R \sqrt{2 \ln \left(\frac{1}{\delta} \sqrt{\frac{\det\left(\boldsymbol{X}_N^\top \boldsymbol{X}_N + \rho I\right)}{\det(\rho I)}} \right)} + \sqrt{\rho} \|\theta_*\| \right).$$

Back to Hilbert spaces Since ultimately we want uncertainty bounds for kernel ridge regression (and GP regression), we need to generalize the results so far to infinite-dimensional Hilbert spaces. As a first step, let us consider a finite-dimensional Hilbert space. Essentially, we have to make sure that everything we did so far works also in a coordinate free manner.

Let H be a finite-dimensional (real) Hilbert space, say of dimension D. The covariates are now $h_1, \ldots, h_N \in H$, but as outlined in Section 5.2, the overall setup goes through with only minor changes, in particular, we now have to set

$$T = \sum_{n=1}^N h_n \otimes h_n \quad (\text{replacing } B = \boldsymbol{X}_N^\top \boldsymbol{X}_N)$$
$$S = \sum_{n=1}^N \eta_n h_n \quad (\text{replacing } A = \boldsymbol{X}_N^\top \boldsymbol{\eta}_N),$$

with $S \in H$ and T a self-adjoint operator on H. Let now $b_1, \ldots, b_D$ be an orthonormal basis of H, and define $\Phi : \mathbb{R}^D \to H$, $\Phi(v) = \sum_{i=1}^D v_i b_i$. This map is invertible and $\Phi^\top = \Phi^{-1}$. Finally, define $A = \Phi^\top S \in \mathbb{R}^D$ and $B = \Phi^\top T \Phi \in \mathbb{R}^{D \times D}$. Going through our previous derivation, we notice that the first location which needs an adjustment is inside the expectation after the Markov inequality,

$$\begin{aligned}
\|(T + \rho \mathrm{id}_H)^{-\frac{1}{2}} S\|_H^2 &= \langle S, (T + \rho \mathrm{id}_H)^{-1} S \rangle_H \\
&= \langle \Phi A, (\Phi B \Phi^\top + \Phi(\rho I)\Phi^\top)^{-1} \Phi A \rangle_H \\
&= \langle A, \Phi^\top (\Phi^\top)^{-1} (B + \rho I)^{-1} \Phi^{-1} \Phi A \rangle_{\mathbb{R}^D} \\
&= \langle A, (B + \rho I)^{-1} A \rangle_{\mathbb{R}^d}.
\end{aligned}$$

Furthermore, the determinant of a linear map on a finite-dimensional vector space is defined as the determinant of any matrix representation, which in turn is in-

dependent of the chosen basis, so we also find $\det(T + \rho \mathrm{id}_H) = \det(B + \rho I)$ and $\det(\rho \mathrm{id}_H) = \det(\rho I)$. Finally, we also need to translate the condition (5.22). Since for all $\lambda \in \mathbb{R}^D$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(-\frac{1}{2}\lambda^\top B\lambda + A^\top\lambda\right)\right] &= \mathbb{E}\left[\exp\left(-\frac{1}{2}\langle\lambda, \Phi^\top T\Phi\lambda\rangle_{\mathbb{R}^D} + \langle\lambda, \Phi^\top S\rangle_{\mathbb{R}^D}\right)\right] \\ &= \mathbb{E}\left[\exp\left(-\frac{1}{2}\langle(\Phi\lambda), T(\Phi\lambda)\rangle_H + \langle\Phi\lambda, S\rangle_H\right)\right],\end{aligned}$$

a bound for T and S induces a corresponding bound for B and A. Furthermore, in our derivation of the particular bound $b(\lambda) \equiv 1$ we did not use the structure of $\mathbb{R}^d$. The rest of the derivation now goes through without change and we get the following uncertainty bound.

Proposition 5.3.3. Let H be a finite-dimensional Hilbert space and $\theta_* \in H$, let $h_1, \ldots, h_N$ be H-valued random variables and $\eta_1, \ldots, \eta_N$ real-valued random variables, and define $y_n = \langle h_n, \theta_*\rangle_H + \eta_n$ and $\mathcal{F}_n = \sigma(x_1, \eta_1, x_2, \eta_2, \ldots, \eta_{n-1}, x_n)$ for $n = 1, \ldots, N$. Assume that there exists $R \in \mathbb{R}_{>0}$ such that for all $n = 1, \ldots, N$ we have

$$\mathbb{E}[\exp(\nu\eta_n) \mid \mathcal{F}_n] \leq \exp\left(\frac{R^2\nu^2}{2}\right) \quad \forall \nu \in \mathbb{R}. \tag{5.25}$$

Finally, let $\rho \in \mathbb{R}_{>0}$ and denote by $\theta_{\rho,N}$ the regularized least-squares estimate of θ_* from data $(h_1, y_1), \ldots, (h_N, y_N)$ with regularization parameter ρ. For all $\delta \in (0, 1)$, it then holds that with probability at least $1 - \delta$ we have for all $h \in H$ that

$$|\langle h, \theta_{\rho,N}\rangle - \langle h, \theta_*\rangle| \leq \|h\|_{N,\rho}\left(R\sqrt{2\ln\left(\frac{1}{\delta}\sqrt{\det\left(\rho^{-1}\sum_{n=1}^{N} h_n \otimes h_n + \mathrm{id}_H\right)}\right)} + \sqrt{\rho}\|\theta_*\|\right), \tag{5.26}$$

where we defined

$$\|h\|_{N,\rho} = \sqrt{\left\langle h, \left(\sum_{n=1}^{N} h_n \otimes h_n + \rho\mathrm{id}_H\right)^{-1} h\right\rangle_H}. \tag{5.27}$$

In preparation of the next developments, we have used the simplification

$$\frac{\det\left(\sum_{n=1}^{N} h_n \otimes h_n + \rho \mathrm{id}_H\right)}{\det(\rho \mathrm{id})} = \det\left(\rho^{-1} \sum_{n=1}^{N} h_n \otimes h_n + \mathrm{id}_H\right).$$

The final bound for separable Hilbert spaces In order to get back to an uncertainty bound for kernel ridge regression (or a frequentist uncertainty bound for GP regression), we have to generalize the preceding developments to Hilbert spaces, with Lemma 5.3.2 as the most important component. If we consider a separable Hilbert space H, then we have a countable orthonormal basis $(b_n)_n$ at our disposal, and via an orthogonal projections onto the finite dimensional subspace $\mathrm{span}\{b_1, \ldots, b_N\}$, the result from the previous section applies. Finally, a convergence argument then lifts the uncertainty bound to all of H. This is the strategy suggested by [219]. The only conceptual difficulty is handling the determinant of a self-adjoint operator of the form $Q + \mathrm{id}_H$. The following intuitive derivation is folklore, but for completeness we include it here.

Let A be a symmetric positive semidefinite matrix with eigenvalue decomposition $A = U\mathrm{diag}(\lambda_1, \ldots, \lambda_D)U^\top$. In this case, we have

$$\begin{aligned}\det(A + I) &= \det\left(U\mathrm{diag}(\lambda_1, \ldots, \lambda_D)U^\top + UU^\top\right) \\ &= \det(U)\det(\mathrm{diag}(\lambda_1, \ldots, \lambda_D) + I)\det(U^\top) = \prod_{i=1}^{D}(\lambda_d + 1).\end{aligned}$$

This suggests that if T is a self-adjoint, positive semidefinite operator on H with a countable eigenvalue decomposition, then we should define

$$\det(T + \mathrm{id}_H) = \lim_{D\to\infty} \prod_{i=1}^{D}(\lambda_i + 1), \tag{5.28}$$

where $(\lambda_n)_n$ is the sequence of eigenvalues. If T is a trace-class operator, then the latter limit exists and is finite. First, note that $1 + \lambda_i \geq 1$ for all i, so

$$1 \leq \prod_{i=1}^{D}(\lambda_i + 1) \leq \prod_{i=1}^{D'}(\lambda_i + 1)$$

for all $1 \leq D \leq D'$. Furthermore, for all $D \in \mathbb{N}_+$

$$\begin{aligned}\prod_{i=1}^{D}(\lambda_i+1) &= \left[\left(\prod_{i=1}^{D}(\lambda_i+1)\right)^{\frac{1}{D}}\right]^D \leq \left(\frac{1}{D}\sum_{i=1}^{D}1+\lambda\right)^D \\ &\leq \exp\left(\frac{1}{D}\sum_{i=1}^{D}\lambda_i\right)^D = \exp\left(\sum_{i=1}^{D}\lambda_i\right) \\ &\leq \exp\left(\sum_{i=1}^{\infty}\lambda_i\right) < \infty,\end{aligned}$$

where we used the arithmetic mean-geometric mean inequality in the first step, the elementary inequality $1+x \leq e^x$ in the second inequality, and the fact that T is trace-class (with nonnegative eigenvalues) in the last step. This variant of the determinant of an operator is nothing else than a special case of the *Fredholm determinant*, and a standard concept in the context of Gaussian measures on separable Hilbert spaces [59, Chapter 1].

We now have all ingredients to generalize Proposition 5.3.3 to separable Hilbert spaces. Since the argument is routine and no conceptual difficulties arise, we do not provide the details here, but instead refere to [219]. The corresponding result reads *exactly* as Proposition 5.3.3 (just with H a separable Hilbert space, and the Fredholm determinant instead of the usual determinant of endomorphisms of finite-dimensional Hilbert spaces).

Before returning to kernel ridge regression and GP regression, it is time for an observation and a surprising consequence. Everything we did works for an arbitrary, but fixed $N \in \mathbb{N}_+$. The central ingredient to get a concrete bound is the condition (5.18) (and the corresponding generalizations). However, the righthand side in the present setup *is just the constant 1, independent of* N. A closer look at the derivation via the noise assumption reveals that actually we have constructed a supermartingale. All of this makes applicable a classic technique, often attributed to Freedman [2], that boosts the result for an arbitrary, but fixed N to a result *uniform in* N. It proceeds by three simple steps.

1. Replace the constant N by a stopping time τ. This is done by using the supermartingale structure and a standard convergence argument.

2. Choose as a specific stopping time the infimum over all N for which the desired bound does not hold.

3. Bound the probability for the uniform bound by the probability for the bound with the stopping time.

In the present context, this technique has been implemented first by [2] for $H = \mathbb{R}^d$ and then in [1, Chapter 3] for separable Hilbert spaces. Another example of this technique can be found in [54], and we refer to any of these references for the technical details. Recently, this approach has been put into a broader context in [94], where the explicit stopping-time construction has been replaced by Ville's inequality, which is also the perspective chosen in [219].

As a very surprising outcome, the resulting uniform (in N) bound *looks exactly like the bound for fixed N*. Here is the final result. It first appeared in [1, Section 3.4], and was recently rediscovered in [219]. For consistency with the literature, we phrase it using a given filtration, to which the various objects are adapted (instead of using the filtration generated by the underlying processes).

Proposition 5.3.4. Let H be a separable Hilbert space and $\mathbb{F} = (\mathcal{F}_n)_{n\in\mathbb{N}}$ a filtration. Let $(h_n)_n$ be an H-valued stochastic process that is predictable w.r.t. $\mathcal{F}_n$, and $(\eta_n)_n$ a real-valued stochastic process adapted to $\mathcal{F}_{n+1}$. Assume that there exists $R \in \mathbb{R}_{>0}$ such that for all $n \in \mathbb{N}_+$

$$\mathbb{E}[\exp(\nu\eta_n) \mid \mathcal{F}_n] \leq \exp\left(\frac{R^2\nu^2}{2}\right) \quad \forall \nu \in \mathbb{R}. \tag{5.29}$$

Furthermore, let $\theta_* \in H$ and define $y_n = \langle h_n, \theta_*\rangle_H + \eta_n$ for $n \in \mathbb{N}_+$. Finally, let $\rho \in \mathbb{R}_{>0}$ and denote by $\theta_{\rho,N}$ the regularized least-squares estimate of θ_* from data $(h_1, y_1), \ldots, (h_N, y_N)$ with regularization parameter ρ. For all $\delta \in (0,1)$, it then holds that with probability at least $1-\delta$ that for all $n \in \mathbb{N}_+$ and all $h \in H$

$$|\langle h, \theta_{\rho,N}\rangle - \langle h, \theta_*\rangle| \leq \|h\|_{N,\rho}\left(R\sqrt{2\ln\left(\frac{1}{\delta}\sqrt{\det\left(\rho^{-1}\sum_{n=1}^{N} h_n \otimes h_n + \mathrm{id}_H\right)}\right)} + \sqrt{\rho}\|\theta_*\|\right), \tag{5.30}$$

where we defined

$$\|h\|_{N,\rho} = \sqrt{\left\langle h, \left(\sum_{n=1}^{N} h_n \otimes h_n + \rho \mathrm{id}_H\right)^{-1} h \right\rangle_H}. \tag{5.31}$$

5.4. Uncertainty bounds based on self-normalization

In this section, we present frequentist uncertainty bounds for kernel ridge and GP regression that arise from self-normalization techniques. As explained in Section 5.2, they form a special case of regularized least-squares in Hilbert spaces, and can hence be derived from the results in the preceding section. Indeed, this derivation is suggested and sketched in [1, Remark 3.13], and it has also been used (in a slightly different setup) in [54] for the proof of Theorem 2 in this reference, cf. also Lemma 1 there. For the reader's convenience, we provide in the following all the details of the relevant derivation, cf. also [219], which rediscovered (a variant of) the self-normalized concentration inequalities from [1, Section 3.2].

From Fredholm to matrix determinants We start with the determinant in the uncertainty bound. Recall that $\sum_{n=1}^{N} h_n \otimes h_n = SA$, where S and A are the synthesis and analysis operators introduced in Section 5.2, and define $\boldsymbol{H}_N = AS$. We now show[8] that if $\boldsymbol{H}_N$ is invertible, then

$$\det\left(\rho^{-1} \sum_{n=1}^{N} h_n \otimes h_n + \mathrm{id}_H\right) = \det\left(\rho^{-1} \boldsymbol{H}_N + I\right). \tag{5.32}$$

If $\lambda \in \mathbb{R}$ is an eigenvalue of $\sum_{n=1}^{N} h_n \otimes h_n$ with eigenvector $h \in H$, then defining $v = Ah \in \mathbb{R}^N$, we have $\boldsymbol{H}_N v = (AS)Ah = A(SAh) = \lambda Ah = \lambda v$. Conversely, if $\lambda \in \mathbb{R}$ is an eigenvalue of $\boldsymbol{H}_N$ with eigenvector $v \in \mathbb{R}^N$, then defining $h = Sv \in H$ we have $\left(\sum_{n=1}^{N} h_n \otimes h_n\right) h = (SA)Sv = S(AS)v = \lambda Sv = \lambda h$. Assume now that $\boldsymbol{H}_N$ is invertible, then this shows that $\sum_{n=1}^{N} h_n \otimes h_n$ has the same eigenvalues as $\boldsymbol{H}_N$ (and is hence also invertible). Let $\lambda_1, \ldots, \lambda_N \in \mathbb{R}_{>0}$ be the eigenvalues of $\boldsymbol{H}_N$, and hence all the non-zero eigenvalues of $\sum_{n=1}^{N} h_n \otimes h_n$, then $\lambda_1/\rho + 1, \ldots, \lambda_N/\rho + 1$ are the

[8] Curiously, we could not locate the following argument in the literature, where usually the *matrix-determinant lemma* is invoked, e.g., [1, Remark 3.13]. However, technically this latter result is only applicable to matrices, and not operators on Hilbert spaces.

eigenvalues of $\rho^{-1}\boldsymbol{H}_N + I$ and the non-zero eigenvalues of $\rho^{-1}\sum_{n=1}^N h_n \otimes h_n + \mathrm{id}_H$. We now have

$$\det\left(\rho^{-1}\sum_{n=1}^N h_n \otimes h_n + \mathrm{id}_H\right) = \prod_{i=1}^N \left(\frac{\lambda_i}{\rho} + 1\right) = \det\left(\rho^{-1}\boldsymbol{H}_N + I\right),$$

where the first equality follows from the definition of the determinant for operators (of the specific type considered here), and the second equality follows from the fact that $\rho^{-1}\boldsymbol{H}_N + I$ is orthogonally diagonizable, and hence the usual determinant is just the product of the eigenvalues.

Back to kernel ridge and GP regression Let $\mathcal{X} \neq \emptyset$ be some set, and let k be a kernel on $\mathcal{X}$ such that H_k is separable. For a sufficient condition, cf. [189, Lemma 4.33], and for in-depth discussion we refer to [32, Section 1.5] and [150]. In the following, we consider the case $H = H_k$, so the fixed target is an RKHS function $f_* = \theta_* \in H_k$. The covariates become $h_n = k(\cdot, x_n)$, where $x_n \in \mathcal{X}$, and we consider test inputs $h = k(\cdot, x)$, for $x \in \mathcal{X}$. As shown in Section 5.2, this leads to

$$f_{N,\rho}(x) := \theta_{N,\rho}(x) = k_N(x)^\top \left(\boldsymbol{K}_N + \rho I\right)^{-1} \boldsymbol{y}_N, \tag{5.33}$$

where as usual

$$k_N(x) = \begin{pmatrix} k(x, x_1) \\ \vdots \\ k(x, x_N) \end{pmatrix} \quad \boldsymbol{K}_N = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{pmatrix}.$$

Furthermore, we get $\boldsymbol{H}_N = \boldsymbol{K}_N$. Finally, as shown in Section 5.2, for any $h = k(\cdot, x)$ we get

$$\begin{aligned} \|k(\cdot, x)\|_{N,\rho}^2 &= \left\langle k(\cdot, x), \left(\sum_{n=1}^N k(\cdot, x_n) \otimes k(\cdot, x_n) + \rho \mathrm{id}_{H_k}\right)^{-1} k(\cdot, x) \right\rangle_k \\ &= \frac{1}{\rho}\left(k(x, x) - k_N(x)^\top \left(\boldsymbol{K}_N + \rho I\right)^{-1} k_N(x)\right) \\ &= \frac{1}{\rho}\sigma_N^2(x), \end{aligned}$$

where σ_N^2 is the posterior variance from GP regression with a zero mean prior with covariance function k, nominal additive i.i.d. Gaussian noise with variance ρ, and data $(x_1, y_1), \ldots, (x_N, y_N)$. For convenience, we summarize this result. It can be reconstructed from Remark 3.13 in [1], and has recently been rediscovered as [219, Theorem 2].

Proposition 5.4.1. Let $\mathcal{X}$ be a measurable space, k a kernel on $\mathcal{X}$ such that H_k is separable, and let $f_* \in H_k$. Let $\mathbb{F} = (\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration, $(x_n)_{n\in\mathbb{N}_+}$ an $\mathcal{X}$-valued stochastic process that is predictable w.r.t. $\mathcal{F}_n$, $(\eta_n)_n$ a real-valued stochastic process adapted to $\mathcal{F}_{n+1}$, and define $y_n = f_*(x_n) + \eta_n$ for $n \in \mathbb{N}_+$. Assume that there exists $R \in \mathbb{R}_{>0}$ such that for all $n \in \mathbb{N}_+$

$$\mathbb{E}[\exp(\nu\eta_n) \mid \mathcal{F}_n] \leq \exp\left(\frac{R^2\nu^2}{2}\right) \quad \forall \nu \in \mathbb{R}. \tag{5.34}$$

Finally, let $\rho \in \mathbb{R}_{>0}$ and denote by μ_N and σ_N^2 the posterior mean variance of GP regression with a zero mean prior with covariance function k, additive i.i.d. $\mathcal{N}(0, \rho)$ noise in the likelihood, and data $(x_1, y_1), \ldots, (x_N, y_N)$, and corresponding kernel matrix $\boldsymbol{K}_N$, for $N \in \mathbb{N}_+$.

For all $\delta \in (0, 1)$, it then holds that with probability at least $1 - \delta$ that for all $N \in \mathbb{N}_+$ and all $x \in \mathcal{X}$

$$|f_*(x) - \mu_N(x)| \leq \sigma_N(x)\left(\frac{R}{\sqrt{\rho}}\sqrt{2\ln\left(\frac{1}{\delta}\sqrt{\det\left(\rho^{-1}\boldsymbol{K}_N + I_N\right)}\right)} + \|f_*\|_k\right). \tag{5.35}$$

For N time steps, or equivalently the case of a data set of N data points, the uncertainty bound from the preceding result has the form

$$|f_*(x) - \mu_N(x)| \leq \beta_N(\delta, B, R, \lambda)\sigma_N(x) \tag{5.36}$$

with

$$\beta_N(\delta, B, R, \lambda) = \|f_*\|_k + \frac{R}{\sqrt{\rho}}\sqrt{2\ln\left(\frac{1}{\delta}\sqrt{\det\left(\rho^{-1}\boldsymbol{K}_N + I_N\right)}\right)}, \tag{5.37}$$

Of course, the RKHS norm $\|f_*\|_k$ in the bound above can be replaced by any constant $B \in \mathbb{R}_{\geq 0}$ with $\|f_*\|_k \leq B$. This uncertainty bound has the form of a tube around

the posterior mean function (equivalently, the solution of kernel ridge regression with an equivalent regularization parameter), of (half-)width $\beta_N(\delta, B, R, \lambda)\sigma_N(x) = \beta_N\sigma_N(x)$. From the perspective of GP regression (which is a Bayesian method), we can interpret β_N as a correction scaling, that translates the pointwise (in input and time) Bayesian uncertainty measure σ_N into a frequentist uncertainty bound, uniform in input and time.

5.5. Robustness to model misspecification

As usual in the literature, the results so far use the same kernel k for generating the RKHS containing the ground truth and as a covariance function in GPR. Obviously, in practice it is unlikely that one gets the kernel of the ground truth exactly right. Therefore, we now investigate what happens if we have a *misspecified model* or misspecification of the kernel.

5.5.1. A case of benign misspecification

In some cases, model misspecification of the kernel is not a problem. The following simple result describes one such situation. To the best of our knowledge, in the context of frequentist uncertainty bounds it first appeared explicitly as [CF12, Proposition 3]. Note that we have slightly reformulated this result to make clearer the conditions under which it holds.

Proposition 5.5.1. Consider the situation of Proposition 5.4.1, but this time assume that the ground truth f_* is from another RKHS $H_{\tilde{k}}$. If $H_{\tilde{k}} \subseteq H_k$, and the inclusion $id : H_{\tilde{k}} \to H_k$ is continuous with operator norm at most 1, then the result holds true without any modification. Similarly, if $H_{\tilde{k}} \subseteq H_k$, then the result still holds if the RKHS norm term in the bound is replaced by a constant B such that $\|f_*\|_k \leq B$.

This simple result can easily be used to verify that for many common situations misspecification of the kernel is not a problem. As an example, we consider the case of the popular isotropic SE kernel. The following result appeared first as [CF12, Proposition 4].

Proposition 5.5.2. Consider the situation of Proposition 5.4.1, but now assume that $f_* \in H_{\tilde{k}}$, where $\tilde{H}$ is the RKHS corresponding to the SE kernel $\tilde{k}$ (on $\emptyset \neq \mathcal{X} \subseteq \mathbb{R}^d$) with length scale $0 < \tilde{\gamma}$. Use for the Gaussian Process Regression the SE kernel k with length-scale $0 < \gamma \leq \tilde{\gamma}$. Finally let B be an upper bound on $\|f_*\|_k \leq B$. Then Proposition 5.4.1 holds true with B instead of the RKHS norm in the bound.

Proof. Follows immediately from Proposition 5.5.1 and [189, Proposition 4.46]. □

These results tell us that it is not a problem if we do not get the hyperparameter of the isotropic SE kernel right, as long as we *underestimate* the length-scale of the actual kernel. This is intuitively clear since a smaller length-scale corresponds to more complex functions, and indeed similar results are known in related contexts, for example [194]. Note that Proposition 5.5.2 does not imply that one should choose a very small length-scale. The result makes a statement on the validity of the frequentist uncertainty bound from Proposition 5.4.1, but not on the size of the uncertainty set. For a similar discussion in a slightly different setting see [215].

Finally, for Proposition 5.5.1, we need an upper bound on the RKHS norm of the target function w.r.t. the *nominal* kernel, i.e., the kernel used for GP or kernel ridge regression. In the situation of Proposition 5.5.2, this can easily be achieved by using [189, Equ (4.44)] if we know an upper bound on the RKHS norm w.r.t. the *misspecified* kernel. Using the notation of Proposition 5.5.2, if $\|f\|_{\tilde{k}} \leq \tilde{B}$, then we can choose $B = \tilde{B} \cdot (\tilde{\gamma}/\gamma)^{d/2}$.

5.5.2. Robustness to unstructured, but bounded kernel misspecification

We now turn to the situation where misspecification of the kernel might be problematic, and adjustments to the uncertainty bounds itself are necessary. For conciseness, we will only consider the case of *arbitrary, but bounded* misspecification of the kernel. This is similar to common disturbance models in robust control, where no structure is imposed on the disturbance, but a bound in magnitude. The following result is essentially an updated variant of [CF12, Theorem 5].

Theorem 5.5.3. Let $\mathcal{X}$ be a measurable space, $k, \tilde{k}$ two kernels on $\mathcal{X}$ such that $H_{\tilde{k}}$ is separable, and assume there exists $\tilde{\epsilon} \in \mathbb{R}_{\geq 0}$ such that $|k(x,x') - \tilde{k}(x,x')| \leq \tilde{\epsilon}$ for all $x, x' \in \mathcal{X}$, Let $f \in H_{\tilde{k}}$ and $B \in \mathbb{R}_{\geq 0}$ a constant with $\|f\|_{\tilde{k}} \leq B$. Let $\mathbb{F} = (\mathcal{F}_n)_{n\in\mathbb{N}}$ be a filtration, $(x_n)_{n\in\mathbb{N}_+}$ an $\mathcal{X}$-valued stochastic process that is predictable w.r.t. $\mathcal{F}_n$,

$(\eta_n)_n$ a real-valued stochastic process adapted to $\mathcal{F}_{n+1}$, and define $y_n = f(x_n) + \eta_n$ for $n \in \mathbb{N}_+$, and $\boldsymbol{y}_N = \begin{pmatrix} y_1 & \cdots & y_N \end{pmatrix}^\top$, for $N \in \mathbb{N}_+$. Assume that there exists $R \in \mathbb{R}_{>0}$ such that for all $n \in \mathbb{N}_+$

$$\mathbb{E}[\exp(\nu\eta_n) \mid \mathcal{F}_n] \leq \exp\left(\frac{R^2\nu^2}{2}\right) \quad \forall \nu \in \mathbb{R}. \tag{5.38}$$

Finally, let $\lambda \in \mathbb{R}_{>0}$ and denote by μ_N and σ_N^2 the posterior mean variance of GP regression with a zero mean prior with covariance function k, additive i.i.d. $\mathcal{N}(0, \rho)$ noise in the likelihood, and data $(x_1, y_1), \ldots, (x_N, y_N)$, and corresponding kernel matrix $\boldsymbol{K}_N$ and kernel vector function $\boldsymbol{k}_N(x) = \Big(k(x, x_n)\Big)_{n=1,\ldots,N}$, for $N \in \mathbb{N}_+$. Similarly, denote by $\tilde{\mu}_N$ and $\tilde{\sigma}_N^2$ the posterior mean variance of GP regression with a zero mean prior with covariance function $\tilde{k}$, additive i.i.d. $\mathcal{N}(0, \lambda)$ noise in the likelihood, and data $(x_1, y_1), \ldots, (x_N, y_N)$, and corresponding kernel matrix $\tilde{\boldsymbol{K}}_N$ and kernel vector function $\tilde{\boldsymbol{k}}_N$, for $N \in \mathbb{N}_+$.

For all $\delta \in (0, 1)$, it then holds that with probability at least $1 - \delta$ that for all $N \in \mathbb{N}_+$ and all $x \in \mathcal{X}$

$$\mathbb{P}[|f_*(x) - \mu_N(x)| \leq B_N(x) \quad \forall x \in \mathcal{X},\ N \in \mathbb{N}_+] \geq 1 - \delta \tag{5.39}$$

where

$$\begin{aligned}
B_N(x) &= C_N(x)\|\boldsymbol{y}_N\| + \bar{\beta}_N\sqrt{\sigma_N^2(x) + S_N(x)^2} \\
\bar{\beta}_N &= B + \frac{R}{\sqrt{\lambda}}\sqrt{2\ln\left(\frac{1}{\delta}\sqrt{\det\left(\lambda^{-1}\boldsymbol{K}_N + (1 + \lambda^{-1}N\tilde{\epsilon})I_N\right)}\right)} \\
C_N(x) &= \left(\frac{1}{\lambda} + \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\|\right)(\|\boldsymbol{k}_N(x)\| + \sqrt{N\tilde{\epsilon}}) + \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\|\sqrt{N\tilde{\epsilon}} \\
S_N(x)^2 &= \tilde{\epsilon} + \sqrt{N\tilde{\epsilon}}\|(\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| + (\sqrt{N\tilde{\epsilon}} + \|\boldsymbol{k}_N(x)\|)C_N(x)
\end{aligned}$$

Before proceeding to the proof of this result, we would like to briefly discuss the form of the bound.

In Proposition 5.4.1, we have a tube around $\mu_N(x)$ of width

$$\beta_N(\delta, B, R, \lambda)\sigma_N(x) = \frac{R}{\sqrt{\rho}}\sqrt{2\ln\left(\frac{1}{\delta}\sqrt{\det\left(\rho^{-1}\boldsymbol{K}_N + I_N\right)}\right)} + \|f_*\|_k, \tag{5.40}$$

whereas in Theorem 5.5.3, we have a tube around $\mu_N(x)$ of width

$$\bar{\beta}_N \sqrt{\sigma_N^2(x) + S_N^2(x)} + C_N(x)\|\boldsymbol{y}_N\|.$$

Because of the uncertainty or disturbance in the kernel we have to increase the nominal noise variance (used in the nominal noise model in GPR), increase the nominal posterior standard deviation from $\sigma_N(x)$ to $\sqrt{\sigma_N^2(x) + S_N^2(x)}$ and add an offset to the width of the tube of $C_N(x)\|\boldsymbol{y}_N\|$. In particular, the uncertainty set now depends on the measured values $\boldsymbol{y}_N$. Note that $C_N(x)$ and $S_N(x)$ depend on the input, but if necessary this dependence can be easily removed by finding an upper bound on $\|\boldsymbol{k}_N(x)\|$. An interesting observation is that even if $\sigma_N(x) = 0$, the tube around $\mu_N(x)$ has nonzero width. Intuitively this is clear since in general it can happen that $f \notin H$, but $\mu_N \in H$ by construction.

Proof of Theorem 5.5.3. Denote by $\tilde{\mu}_N$, $\tilde{\sigma}_N^2$, $\tilde{\boldsymbol{K}}_N$ the posterior mean, posterior variance and Gram matrix of the GP, but with $\tilde{k}$ as covariance function, and analogously $\tilde{\boldsymbol{k}}_N$. Let $x \in \mathcal{X}$ be arbitrary, then we have

$$|f(x) - \mu_N(x)| \leq |f(x) - \tilde{\mu}_N(x)| + |\tilde{\mu}_N(x) - \mu_N(x)|.$$

Our strategy will be to bound the first term on the right-hand-side with Proposition 5.4.1, and then upper bound all resulting or remaining quantities by expressions involving only k instead of $\tilde{k}$. As a preparation we first derive some elementary bounds that are frequently used later on. By assumption,

$$\|\tilde{\boldsymbol{k}}_N(x) - \boldsymbol{k}_N(x)\| = \sqrt{\sum_{i=1}^{N} (\tilde{k}(x, x_i) - k(x, x_i))^2} \leq \sqrt{N}\tilde{\epsilon}, \tag{5.41}$$

and hence (using the triangle inequality)

$$\|\tilde{\boldsymbol{k}}_N(x)\| \leq \|\boldsymbol{k}_N(x)\| + \|\tilde{\boldsymbol{k}}_N(x) - \boldsymbol{k}_N(x)\| \leq \|\boldsymbol{k}_N(x)\| + \sqrt{N}\tilde{\epsilon}. \tag{5.42}$$

Furthermore, since $\tilde{\boldsymbol{K}}_N$ is positive semidefinite

$$\|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\| = \lambda_{\max}((\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}) = \frac{1}{\lambda_{\min}(\tilde{\boldsymbol{K}}_N + \lambda I_N)} \leq \frac{1}{\lambda}$$

and hence together with the triangle inequality

$$\|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1} - (\boldsymbol{K}_N + \lambda I_N)^{-1}\| \leq \frac{1}{\lambda} + \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\|. \tag{5.43}$$

Finally, using first the triangle inequality and then the submultiplicativity of the spectral norm we get

$$\begin{aligned}
&\|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x) - (\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| \\
&\quad \leq \| \left((\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1} - (\boldsymbol{K}_N + \lambda I_N)^{-1}\right) \tilde{\boldsymbol{k}}_N(x)\| + \|(\boldsymbol{K}_N + \lambda I_N)^{-1}(\tilde{\boldsymbol{k}}_N(x) - \boldsymbol{k}_N(x))\| \\
&\quad \leq \|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1} - (\boldsymbol{K}_N + \lambda I_N)^{-1}\|\|\tilde{\boldsymbol{k}}_N(x)\| + \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\|\|\tilde{\boldsymbol{k}}_N(x) - \boldsymbol{k}_N(x)\|,
\end{aligned}$$

and hence from (5.41), (5.42), (5.43)

$$\|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x) - (\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| \leq C_N(x) \tag{5.44}$$

with

$$C_N(x) = \left(\frac{1}{\lambda} + \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\|\right)(\|\boldsymbol{k}_N(x)\| + \sqrt{N}\tilde{\epsilon}) + \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\|\sqrt{N}\tilde{\epsilon}.$$

Now,

$$\begin{aligned}
|\tilde{\mu}_N(x) - \mu_N(x)| &= |\tilde{\boldsymbol{k}}_N(x)(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\boldsymbol{y}_N - \boldsymbol{k}_N(x)(\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{y}_N| \\
&\leq \|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x) + (\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\|\|\boldsymbol{y}_N\| \\
&\leq C_N(x)\|\boldsymbol{y}_N\|,
\end{aligned}$$

where we used Cauchy-Schwarz in the first inequality and (5.44) in the second. Using Proposition 5.4.1 we get that

$$\mathbb{P}[|\tilde{\mu}_N(x) - f(x)| \leq \tilde{\beta}_N\tilde{\sigma}_N(x)\,\forall N \in \mathbb{N}, x \in D] \geq 1 - \delta$$

where

$$\begin{aligned}\tilde{\beta}_N &= B + \frac{R}{\sqrt{\lambda}}\sqrt{2\ln\left(\frac{1}{\delta}\sqrt{\det(\lambda^{-1}\tilde{\boldsymbol{K}} + I_N)}\right)} \\ &= B + \frac{R}{\sqrt{\lambda}}\sqrt{\ln\left(\det(\lambda^{-1}\tilde{\boldsymbol{K}} + I_N\right) + 2\ln(1/\delta)}\end{aligned}$$

Let $\lambda_i(\tilde{\boldsymbol{K}}_N)$ be the i-th largest eigenvalue of $\tilde{\boldsymbol{K}}_N$, then we have

$$\begin{aligned}\lambda_i(\tilde{\boldsymbol{K}}_N) &\leq \lambda_i(\boldsymbol{K}_N) + \|\tilde{\boldsymbol{K}}_N - \boldsymbol{K}_N\| \\ &\leq \lambda_i(\boldsymbol{K}_N) + \|\tilde{\boldsymbol{K}}_N - \boldsymbol{K}_N\|_F \\ &\leq \lambda_i(\boldsymbol{K}_N) + N\tilde{\epsilon},\end{aligned}$$

where we first used Weyl's inequality, then the fact that the operator norm $\|\cdot\|$ is always less or equal the Frobenius norm $\|\cdot\|_F$, and finally the definition of the Frobenius norm together with the assumption on k and $\tilde{k}$. We now get

$$\begin{aligned}\ln\det\left(\lambda^{-1}\tilde{\boldsymbol{K}}_N + I_N\right) &= \ln\left(\prod_{i=1}^{N}\lambda_i\left(\lambda^{-1}\tilde{\boldsymbol{K}}_N + I_N\right)\right) = \ln\left(\prod_{i=1}^{N}\left(\lambda^{-1}\lambda_i(\tilde{\boldsymbol{K}}_N) + 1\right)\right) \\ &= \sum_{i=1}^{N}\ln\left(\lambda^{-1}\lambda_i(\tilde{\boldsymbol{K}}_N) + 1\right) \\ &\leq \sum_{i=1}^{N}\ln\left(\lambda^{-1}\lambda_i(\boldsymbol{K}_N) + \lambda^{-1}N\tilde{\epsilon} + 1\right) \\ &= \sum_{i=1}^{N}\ln\left(\lambda_i\left(\lambda^{-1}\boldsymbol{K}_N + (1+\lambda^{-1}N\tilde{\epsilon})I_N\right)\right) \\ &= \ln\left(\prod_{i=1}^{N}\lambda_i\left(\lambda^{-1}\boldsymbol{K}_N + (1+\lambda^{-1}N\tilde{\epsilon})I_N\right)\right) \\ &= \ln\det\left(\lambda^{-1}\boldsymbol{K}_N + (1+\lambda^{-1}N\tilde{\epsilon})I_N\right),\end{aligned}$$

which we can use for

$$\begin{aligned}\tilde{\beta}_N &= B + \frac{R}{\sqrt{\lambda}}\sqrt{\ln\det\left(\lambda^{-1}\tilde{\boldsymbol{K}} + I_N\right) + 2\ln(1/\delta)} \\ &\leq B + \frac{R}{\sqrt{\lambda}}\sqrt{\ln\det\left(\lambda^{-1}\boldsymbol{K}_N + (1+\lambda^{-1}N\tilde{\epsilon})I_N\right) + 2\ln(1/\delta)} \\ &= B + \frac{R}{\sqrt{\lambda}}\sqrt{2\ln\left(\frac{1}{\delta}\sqrt{\det\left(\lambda^{-1}\boldsymbol{K}_N + (1+\lambda^{-1}N\tilde{\epsilon})I_N\right)}\right)} \\ &=: \bar{\beta}_N.\end{aligned}$$

Turning to the posterior variance, we get from the triangle inequality

$$\tilde{\sigma}_N^2(x) \leq \sigma_N^2(x) + |\sigma_N^2(x) - \tilde{\sigma}_N^2(x)|.$$

We continue with

$$\begin{aligned}|\sigma_N^2(x) - \tilde{\sigma}_N^2(x)| &= |k(x,x) - \boldsymbol{k}_N(x)^T(\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x) \\ &\qquad - \tilde{k}(x,x) + \tilde{\boldsymbol{k}}_N(x)^T(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x)| \\ &\leq |k(x,x) - \tilde{k}(x,x)| + |(\tilde{\boldsymbol{k}}_N(x) - \boldsymbol{k}_N(x))^T(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x)| \\ &\qquad + |\boldsymbol{k}_N(x)^T((\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x) - (\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x))| \\ &\leq |k(x,x) - \tilde{k}(x,x)| + \|\tilde{\boldsymbol{k}}_N(x) - \boldsymbol{k}_N(x)\|\|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x)\| \\ &\qquad + \|\boldsymbol{k}_N(x)\|\|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x) - (\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| \\ &\leq \tilde{\epsilon} + \sqrt{N}\tilde{\epsilon}\|(\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| + (\sqrt{N}\tilde{\epsilon} + \|\boldsymbol{k}_N(x)\|)C_N(x) \\ &=: S_N^2(x),\end{aligned}$$

where we used the triangle inequality again in the first inequality, Cauchy-Schwarz in the second inequality and finally (5.41), (5.42), (5.43) together with

$$\begin{aligned}&\|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x)\| \\ &\quad\leq \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| + \|(\tilde{\boldsymbol{K}}_N + \lambda I_N)^{-1}\tilde{\boldsymbol{k}}_N(x) - (\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| \\ &\quad\leq \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| + C_N(x)\end{aligned}$$

Putting everything together, we find that with probability at least $1 - \delta$

$$|\mu_N(x) - f(x)| \leq C_N(x)\|\boldsymbol{y}_N\| + \tilde{\beta}_N \tilde{\sigma}_N(x)$$

and therefore, using the upper bounds on $\tilde{\beta}_N$ and $\tilde{\sigma}_N(x)$ derived above, that with probability at least $1 - \delta$

$$|\mu_N(x) - f(x)| \leq B_N(x)$$

where

$$B_N(x) = C_N(x)\|\boldsymbol{y}_N\| + \bar{\beta}_N \sqrt{\sigma_N^2(x) + S_N(x)^2}$$

□

5.6. Comments

The general approach in Section 5.1 (separating terms with noise from terms with the target object) is folklore, and has been used in the context of uncertainty bounds for GP regression for example in [54]. In turn, this work has been the starting point of [CF12]. The elementary results Proposition 5.1.2 and 5.1.3 are motivated by results in this latter work, but have been rederived for this thesis, and similarly for Proposition 5.2.1. While we could not locate these specific results in the literature, they might be known (probably as technical auxiliary results), especially given their connection to linear inverse problems, see [142] for a recent reference. As already mentioned in the main text, the approach from Section 5.2 is well-known, but we could not find a reference for this particular exposition. All of the results (or minor variations thereof) in Section 5.3 are known, but to the best of our knowledge, our particular derivation, and our explanation how one can discover these techniques, is new. Section 5.5 is based on and partially taken verbatimly from [CF12]. This latter work arose through discussions of the author with S. Trimpe and C.W. Scherer, with all theoretical results derived by the present author. Note that the overall presentation of the theoretical results in this chapter is significantly different from [CF12]. In Section 6.4, we provide a detailed discussion of the contributions of this latter work, and how it is connected to related and existing work in the literature. Finally, the present chapter has been written from scratch by the present author of this thesis, with the above mentioned exceptions.

6. Frequentist uncertainty bounds for kernel and GP regression: Experiments and applications

In the preceding Chapter 5, we have presented a variety of frequentist uncertainty bounds for GP regression and kernel ridge regression, and we will now investigate some of them empirically through numerical experiments, and apply them for learning-based control approaches. Recalling our discussion from Chapter 4, in the context of learning-based control it is important that these uncertainty bounds can be numerically evaluated, and are not too conservative. Our first goal in this chapter is therefore to carefully investigate how the bounds behave empirically, which we do in Section 6.1. Key issues will be conservatism as well as robustness to model misspecification. Our experiments indicate that the uncertainty bounds are tight enough for use in learning-based control, which we will validate here with two use cases. In particular, in Section 6.2 the uncertainty bounds will be used together with modern robust controller synthesis.

This chapter is based on, with some parts taken verbatim from, [CF12, CF13] and [CF10][1]. Detailed comments on the author's contribution and relation to existing work are provided in Section 6.4.

[1] © IEEE 2021. Reprinted, with permission, from Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. *Learning-enhanced robust controller synthesis with rigorous statistical and control-theoretic guarantees.* 60th IEEE Conference on Decision and Control (CDC), 2021.

6.1. Experimental evaluation of frequentist uncertainty bounds

We now turn to a numerical investigation of the uncertainty bounds. First, we provide some background and detailed explanation of the experimental setup. This is followed by the results of the experiments on both the nominal and robust uncertainty bounds. Finally, we apply the bounds to an illustrative example from control.

6.1.1. Background and setup

As explained in detail in Chapter 4, we are interested in frequentist uncertainty bounds. In particular, this suggests a *frequentist evaluation* of the bounds. This means that for a given ground truth, many data sets are generated by sampling multiple noise realizations, the uncertainty bounds are computed, and for each realization it is checked whether the uncertainty set contains the ground truth.

Let us make this more precise using the abstract setup from Chapter 4. Consider the frequentist setting as formalized there, and for simplicity assume that $\bar{\mathcal{P}} = \{P\}$. Suppose we want to evaluate an uncertainty set (estimator) $(\mathcal{U}_\delta)_{\delta\in(0,1)}$ from a frequentist perspective. For this, can fix a ground truth $\theta_* \in \Theta$, a desired confidence level $\delta \in (0,1)$, and a finite number of samples $N \in \mathbb{N}_+$. We then generate independent noise realizations $\eta_1, \dots, \eta_N \sim P$, compute $y_n = \mathcal{F}(\theta_*, \eta_n)$ and $U_n = \mathcal{U}_\delta(y_n)$ for $n = 1, \dots, N$, as well as $F = \sum_{n=1}^{N} \mathbb{I}_{\theta_* \in U_n}$ (where $\mathbb{I}_p$ is 1 if p holds, otherwise 0) and $\hat{\delta} = F/N$. If $(\mathcal{U}_\delta)_{\delta\in(0,1)}$ really leads to frequentist uncertainty sets, then $\hat{\delta} \leq \delta$ for large N. Furthermore, $\hat{\delta}$ as well as the average "size" of $U_1, \dots, U_N$ can be used as indicators of the conservatism of the uncertainty sets. Finally, since frequentist uncertainty sets should hold for a given, but arbitrary ground truth, this should be checked for all $\theta_* \in \Theta$. In general, this is not feasible, so the preceding experiment should be repeated for a variety of elements from Θ. It is clear that such an experimental setup can be realized only in numerical experiments with known ground truths, since many repetitions (large N) are required, and ideally this is repeated with different ground truths. In particular, this requires the generation of many elements from Θ.

Before turning to the concrete setting that is considered for the remainder of

this chapter, we would like to discuss some delicate points of this setup. First, the approach outlined above can give only empirical indications about the behaviour of the uncertainty sets, since only finitely many samples and ground truths can be considered in experiments. The latter fact can be particularly problematic when the ground truths are synthetically generated with a method that has a certain bias, e.g., a tendency to produce only elements from some (potentially very small subset) $\tilde{\Theta} \subseteq \Theta$. In this case, experiments following the method from above can result in potentially misleading results, and in our setting, this is indeed a risk, as described in the next section. Second, the behaviour of uncertainty sets in practice can be very different from what is observed in such synthetic experiments. For example, the set of possible ground truths Θ might be much larger than what is actually possible in a concrete application class, which in turn could lead to very conservative behaviour of the uncertainty sets in the synthetic experiments.

Generating functions from an RKHS For the numerical experiments we need to generate ground truths, i.e., we need to randomly generate functions belonging to the RKHS of a given kernel. A generic approach is to use the pre-RKHS of the kernel which is contained (even densely w.r.t. the kernel norm) in the actual RKHS, cf. Chapter 2. Let X be a set and k a kernel on X. For any $N \in \mathbb{N}$, $x_1, \dots, x_N \in X$ and $\boldsymbol{\alpha} \in \mathbb{R}^N$, the function $\sum_{n=1}^N \alpha_n k(\cdot, x_n)$ is contained in the (unique) RKHS corresponding to k and has RKHS norm $\sqrt{\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}}$, where $\boldsymbol{K} = (k(x_i, x_j))_{i,j=1,\dots,N}$ is the corresponding Gram matrix. It is hence possible to generate an RKHS function f of prescribed RKHS norm B by randomly sampling inputs $x_1, \dots, x_N \in X$ and coefficients $\tilde{\alpha} \in \mathbb{R}^N$ and setting

$$f(x) = \sum_{n=1}^{N} \alpha_n k(x_n, x) \tag{6.1}$$

where $\boldsymbol{\alpha} = \frac{B}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}}} \tilde{\boldsymbol{\alpha}}$. Of course, in practice f can only be evaluated at finitely many points $\tilde{X} \subseteq X$. More concretely, we fix a finite evaluation grid $\tilde{X} \subseteq X$, choose uniformely a number $N \in [N_{\min}, N_{\max}] \cap \mathbb{N}$, choose uniformely N pairwise different points $x_1, \dots, x_N \in \tilde{X}$, sample $\tilde{\alpha}_i \sim \mathcal{N}(0, \sigma_f^2)$ and apply the construction (6.1). For precise choices of the parameters are given below.

As explained above, we are concerned with a frequentist setting, so there is a

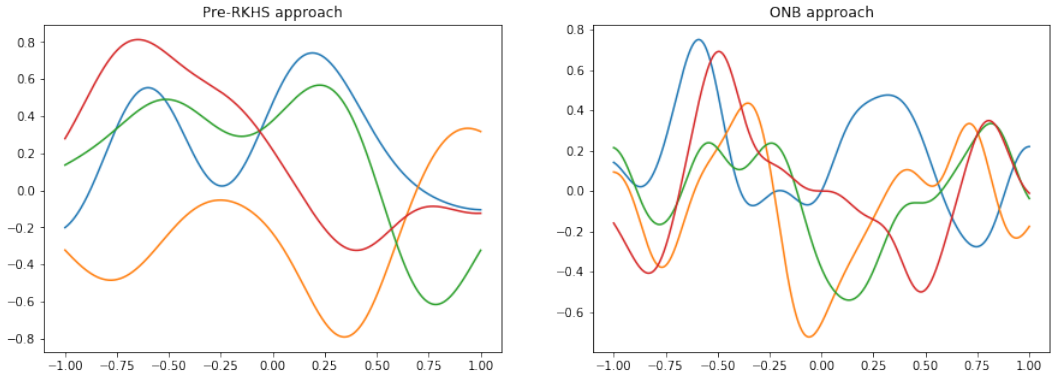

Figure 6.1.: Illustrating sampling from the Gaussian kernel with the pre-RKHS method (left) and with an explicit ONB (right). Details are provided in the text.

ground truth from a collection of possible ground truths and the results have to hold for each of these possible ground truths. In particular, even if a ground truth might be considered pathological, the results have to hold if they are to be considered *rigorous*. This aspect is important for numerical experiments, especially when trying to assess the conservatism of a result. In our setting it might happen that the results seem very conservative for functions that are randomly generated in a certain fashion, but there are RKHS functions (which might be difficult to generate) for which the results might be sharp. Let us illustrate this point with the Gaussian kernel. We use a uniform grid of 1000 points from $[-1, 1]$ together with the Gaussian kernel with length scale 0.2. For the pre-RKHS approach we use $N_{\min} = 5$ and $N_{\max} = 200$ and $\sigma_f^2 = 1$. As an alternative, we use the ONB described in [189, Section 4.4] and consider only the first 50 basis functions from [189, Equation (4.35)] for numerical reasons. We first select the number of basis functions N to use uniformely between 5 and 50 and then choose N such functions uniformely. As coefficients we sample $\alpha_i \sim \mathcal{N}(0, 1)$ i.i.d. for $i = 1, \dots, N$ and normalize (w.r.t. to ℓ_2-norm) and multiply by the targeted RKHS norm. For both the pre-RKHS approach and the ONB approach we use $\|\cdot\|_k = 2$ and sample 4 functions each. The result is shown in Figure 6.1. Clearly, the resulting functions have a different shape, despite having the same RKHS norm with respect to the same kernel. In particular, the functions generated using the ONB approach seem to make sharper turns. We like to stress

that this strongly suggests that in a frequentist setting one has to be careful with statements about the conservatism of a proposed bound or method that are based purely on empirical observations. It might be that the method for generating ground truths has a certain bias, i.e., has a tendency to produce only ground truths from a certain region of the space of all ground truths. To the best of our knowledge, this issue was discussed for the first time in [CF12].

Basic experimental setup Unless otherwise stated, we use $[-1, 1]$ as the input set and consider a uniform grid of 1000 points for function evaluations, and in each experiment we sample 50 RKHS functions as ground truth. In general, we use the pre-RKHS approach to generate functions from the corresponding RKHS, and additionally for the SE kernel we use the orthonormal basis (ONB) from [189, Section 4.4] with random, normally distributed coefficients. For each function we repeat the following learning instance 10000 times: We sample uniformly 50 inputs, evaluate the ground truth on these inputs and adding normal zero-mean i.i.d. noise with standard deviation (SD) 0.5. We then run GP regression on each of the training sets, compute the uncertainty sets and check on the equidistant grid of the input set whether the resulting uncertainty set contains the ground truth. We consider a learning instance a failure if the uncertainty set does not fully cover the ground truth at all 1000 evaluation points. Finally, instead of $\beta_N(\delta, B, R, \lambda)$ from Proposition 5.4.1, we use the slightly different

$$\beta_N = B + \frac{R}{\sqrt{\bar{\lambda}}}\sqrt{\ln\left(\det(\bar{\lambda}/\lambda \boldsymbol{K}_n + \bar{\lambda} I_n)\right) - 2\ln(\delta)} \tag{6.2}$$

with $\bar{\lambda} = \max\{1, \lambda\}$ (using λ instead of ρ for consistency with the GP literature). Similarly, for the robustness experiments, instead of B_N from Theorem 5.5.3, we use the slightly different

$$B_N = B + \frac{R}{\sqrt{\bar{\lambda}}}\sqrt{\ln\det\left(\frac{\bar{\lambda}}{\lambda}\boldsymbol{K}_N + \left(\frac{\bar{\lambda}}{\lambda}N\tilde{\epsilon} + \bar{\lambda}\right) I_N\right) - 2\ln(\delta)} \tag{6.3}$$

with

$$\begin{aligned} S_N^2(x) &= \tilde{\epsilon} + \sqrt{N}\tilde{\epsilon}\|(\boldsymbol{K}_N + \lambda I_N)^{-1}\boldsymbol{k}_N(x)\| \\ &\quad + (\sqrt{N}\tilde{\epsilon} + \|\boldsymbol{k}_N(x)\|)C_N(x) \\ C_N(x) &= \left(\frac{1}{\lambda} + \|(\boldsymbol{K}_N + \lambda I_N)^{-1}\|\right)(\|\boldsymbol{k}_N(x)\| + \sqrt{N}\tilde{\epsilon}) \\ &\quad + \|(\boldsymbol{K}_N - \lambda I_N)^{-1}\|\sqrt{N}\tilde{\epsilon} \end{aligned}$$

In Section 6.4, we discuss these differences. Finally, while frequentist uncertainty bounds for GP regression have been known since the seminal work [186], the first systematic empirical investigation from a frequentist perspective has been conducted in [CF12], to the best of our knowledge.

6.1.2. Experiments

We are now ready for the numerical experiments. For ease of reference, we label each experiment using the prefix *exp_*.

Nominal setting

Consider the case that the covariance function used in GP regression and the kernel corresponding to the RKHS of the target function coincide. First, we test the nominal bound (6.2) with SE and Matern kernels, respectively, for different values of δ. Here we use the SE kernel with length scale 0.2 (*exp_1_1_a*) and the Matern kernel with length scale 0.2 and $\nu = 1.5$ (*exp_1_1_b*). We generate RKHS functions of RKHS norm 2 using the pre-RKHS approach and use the same kernel for generating the ground truth and running GP regression. The nominal noise level of GP regression is set to $\lambda = 0.5$. The uncertainty set is generated using (6.2) with $B = 2$, $R = 0.5$ and $\delta = 0.1, 0.01, 0.001, 0.0001$. The mean of the scalings β_{50} (together with 1 SD, average is over all 50 RKHS functions and all learning instances) is shown in Table 6.1. As can be seen there, the scalings are reasonably small, roughly in the range of heuristics used in the literature. In particular, a violation of the uncertainty set was found in no instance (i.e. for all 50 RKHS functions and each of the 10000 learning instances). Furthermore, the graph of the target function is fully included in the uncertainty set in all repetitions. This is illustrated in Figure 6.2 (left), where

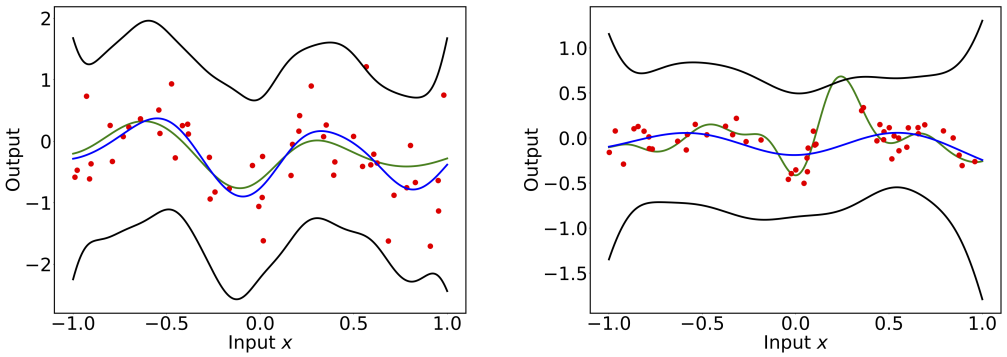

Figure 6.2.: Left (Nominal setting): Example function (green) from SE kernel with length-scale 0.5 and RKHS norm 2, learned from 50 samples (red). Shown is the posterior mean (blue) and the uncertainty set from (6.2) for $\delta = 0.01$. Right (Misspecified setting): Example function from SE kernel with length-scale 0.2, learned with GPR using SE covariance function with length-scale 0.5. The violation of the uncertainty set is clearly visible.

an example instance of this experiment is shown. As can be clearly seen there, the posterior mean (blue solid line) is well within the uncertainty set (with $\delta = 0.01$), which is not overly conservative.

Exploring conservatism

In order to explore the potential conservatism of the bound (6.2) we repeated the previous experiments with $\delta = 0.01$ and replaced β_{50} by 20 equidistant scalings between 2 and β_{50}. We used this changed setup for the SE kernel (*exp_1_2_a*), Matern kernel (*exp_1_2_b*) and SE kernel together with the ONB sampling approach (*exp_1_2_c*). Whereas for the Matern kernel also the heuristic $\beta = 2$ works for this example (still no uncertainty set violations), the situation is rather different for the SE kernel. As shown in Figure 6.3 for β close to 2 the frequency of uncertainty violations is much higher than 0.01, in particular for the case of sampling from the ONB.

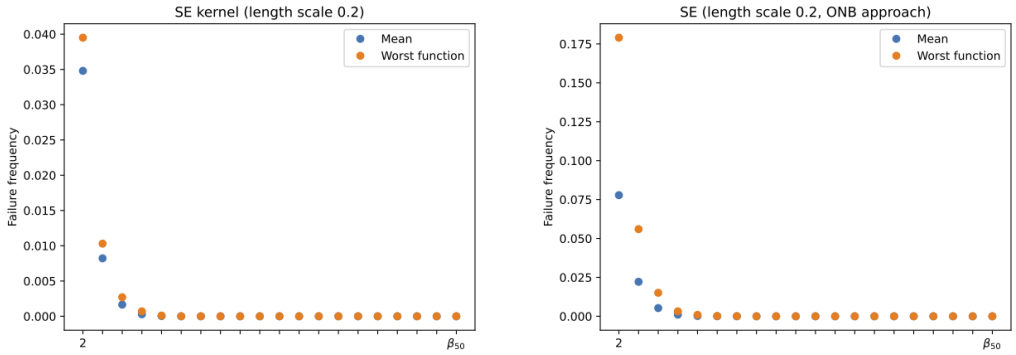

Figure 6.3.: Exploring conservatism of nominal bound (6.2). For each target function and learning instance 20 different uncertainty sets are tested. Each such uncertainty is generated from (6.2) by replacing β_{50} by $\beta = 2, \ldots, \beta_{50}$ (equidistant). Worst function means the fucntion with the highest failure frequency among all 50 target functions for the particular scaling. Ground truths sampled with pre-RKHS approach (left) and ONB approach (right).

Table 6.1.: β_{50} in nominal setting (mean $\pm$ standard deviation over all repetitions)

δ	0.1	0.01	0.001	0.0001
SE	6.95 ± 0.04	7.39 ± 0.04	7.80 ± 0.03	8.19 ± 0.03
Matern	7.36 ± 0.04	7.78 ± 0.04	8.16 ± 0.04	8.53 ± 0.04

Misspecified kernel, benign setting

We now consider misspecification of the kernel, i.e., different kernels are used for generating the ground truth and as prior covariance function in GP regression. As an example, we use the SE kernel with different length-scales and use the ONB from [189, Section 4.4] to generate the RKHS functions. We start with the *benign* setting from Proposition 5.5.2, where the RKHS corresponding to the covariance function used in GPR contains the RKHS of the target function. For this, identical settings as in our first experiment above are used, but now we generate functions from the SE kernel with length-scale 0.5 and use the SE kernel with length-scale 0.2 in GP regression (*exp_1_3_a*). As expected, we find the same results as above, i.e., the uncertainty set fully contains the ground truth in all repetitions. Furthermore, the scalings β_{50} are roughly of the same size as in the nominal case, cf. Table 6.2 (upper row). Note that we only report results for the ONB sampling approach, since no significant difference compared to the pre-RKHS approach arises.

Misspecified kernel, problematic setting

Next, we investigate the *problematic* setting where the RKHS corresponding to the covariance function used in GPR does not contain the RKHS of the target function anymore. As an example we use again the SE kernel, but now with length-scale 0.2 for generating RKHS functions and the SE kernel with length-scale 0.5 for GPR. We use both the pre-RKHS approach (*exp_1_4_a*) and the ONB approach (*exp_1_4_b*). The resulting scalings are reported in Table 6.2 (lower row), again only for the ONB sampling approach. For the ONB sampling we found that for 2, 6, 12 and 13 out of the 50 target functions the frequency of the uncertainty violation was higher than $\delta = 0.1, 0.01, 0.001, 0.0001$. Interestingly, when performing this experiment with the pre-RKHS approach, we did not find functions that violated the uncertainty bounds more often than prescribed. This reaffirmes our introductory remark that the method generating the test targets can lead to wrong conclusion from empirical evaluations of the theoretical results.

Table 6.2.: β_{50} in the misspecified setting (mean ± standard deviation over all repetitions)

δ	0.1	0.01	0.001	0.0001
Benign	6.53 ± 0.038	6.97 ± 0.035	7.39 ± 0.033	7.77 ± 0.031
Problematic	6.11 ± 0.03	6.64 ± 0.03	7.11 ± 0.02	7.54 ± 0.02

Table 6.3.: Width of robust uncertainty set (mean ± SD of average width)

δ	0.1	0.01	0.001	0.0001
Mean	71.68 ± 5.36	73.79 ± 5.36	75.64 ± 5.37	77.33 ± 5.37
SD	6.54 ± 1.73	6.73 ± 1.78	6.91 ± 1.82	7.06 ± 1.86

Robust result for misspecified setting

The results of the previous two experiments indicate that a model misspecification of the kernel can be a problem and a robust result like Theorem 5.5.3 is necessary. To investigate this, we repeat Experiment *exp_1_4_b* from the previous paragraph, but now using (6.3) instead of (6.2). We find no violation of the uncertainty set (over all 50 functions tested, all 10000 learning instances for each function and all δ tested). Since now the width of the uncertainty set is not a constant rescaling of the posterior standard deviation anymore, we report the mean (over all 50 functions and each of the 10000 learning instances) of the average width (over the input space) of the uncertainty sets (± SD w.r.t. averaging over all 50 functions and each of the 10000 learning instances) and the SD (w.r.t. to averaging over the input space), ± SD w.r.t. averaging over all 50 functions and each of the 10000 learning instances, in Table 6.3. An inspection of the average uncertainty set widths in Table 6.3 indicates some conservatism.

6.1.3. A first learning-based control example

We now illustrate the uncertainty bounds with a concrete, existing learning-based control method. As an example, we choose the algorithm from [185] which is a learning-based Robust Model Predictive Control (RMPC) approach that comes with rigorous control-theoretic guarantees.

Background

For convenience we now provide a cursory overview of background material. We can only provide a sketch and refer to standard textbooks for more details, e.g. [167] for a comprehensive introduction to MPC.

A common goal in control is feedback stabilization under state and input constraints. Consider a discrete-time dynamical system (or control system) described by

$$x_+ = f(x, u)$$

with state space X, input space U and transition function $f : X \times U \to X$. For simplicity assume that $X = \mathbb{R}^n$, $U = \mathbb{R}^m$ and that $f(0, 0) = 0$, i.e., $(0, 0)$ is an equilibrium. Furthermore, consider state constraints $\mathbb{X} \subseteq X$ and input constraints $\mathbb{U} \subseteq U$. In this setting, feedback stabilization amounts to finding a map $\mu : X \to U$ such that $x_* = 0$ is an asymptotically stable equilibrium for the resulting closed loop system described by

$$x_+ = f(x, \mu(x)),$$

and all resulting state-input trajectories are contained in the constraint set $\mathbb{X} \times \mathbb{U}$. Note that this requires restriction of the set of possible initial values.

In many applications not only stability, but also a form of optimality is required from the control system. For example, assume that being in state x and applying input u incurs a cost of $\ell(x, u)$. If the control system is run for a long time, then we would like a feedback μ that not only stabilizes the system, but also incurs a small infinite horizon cost

$$\sum_{n=0}^{\infty} \ell(x(n), \mu(x(n))),$$

where $x(n)$ is the resulting state trajectory. One common methodology for dealing with state constraints and optimal control tasks is MPC. If the system is in state x, MPC solves a finite horizon open loop problem, i.e., it determines a sequence $u(0), \ldots, u(N-1)$ of admissible input values that minimize some cost criterion. Only the first input $u(0)$ is applied to the system and this process is repeated at the next time instance. There is a comprehensive theory available on how to design the open loop optimal control problem solved in each instance in order to achieve desired closed loop properties. For details we refer to Chapters 1 and 2 in [167].

In many applications a control system has to deal with disturbances. Frequently the disturbances can be modelled in an additive manner, i.e., we have a control system of the form

$$x_+ = f(x, u) + w,$$

where $w \in \mathbb{W} \subseteq \mathbb{R}^n$ is an external disturbance. The feedback stabilization problem under constraints can now be adapted to this setting, resulting in robust feedback stabilization under constraints. The goal is now to find a feedback that ensures constraint satisfaction and stabilizes the origin in a relaxed sense (which depends on the size of the disturbance set $\mathbb{W}$). Furthermore, even in this more challenging situation one might have to deal with additional cost criterions.

MPC can be adapted to the setting with disturbances. The key idea of most approaches is to solve a constrained open loop optimal control problem where the constraints are tightened. The intuition is that even the worst case disturbance cannot throw the system out of the allowed state-input set. It is clear that this requires sufficiently small bounds on the size of the disturbances. For more details we refer to Chapter 3 in [167].

A concrete example

We follow [185] and consider the discrete-time system

$$\begin{bmatrix} x_1^+ \\ x_2^+ \end{bmatrix} = \begin{bmatrix} 0.995 & 0.095 \\ -0.095 & 0.900 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0.048 \\ 0.95 \end{bmatrix} u + \begin{bmatrix} 0 \\ -r(x_2) \end{bmatrix} \tag{6.4}$$

modelling a mass-spring-damper system with some nonlinearity r (this could be interpreted as a friction term). The goal is the stabilization of the origin subject to the state and control constraints $\mathbb{X} = [-10, 10] \times [-10, 10]$ and $\mathbb{U} = [-3, 3]$, as well as minimizing a quadratic cost.

The approach from [185] performs this task by interpreting (6.4) as a linear system with disturbance, given by the nonlinearity r, whose graph is a-priori known to lie in the set $\mathbb{W}_0 = [-10, 10] \times [-7, 7]$. The nonlinearity is assumed to be unknown and has to be learned from data. The RMPC algorithm requires as a disturbance sets $\mathbb{W}(x)$ such that $\begin{pmatrix} 0 & -r(x_2) \end{pmatrix}^\top \in \mathbb{W}(x)$ for all $x \in \mathbb{X}$, which are in turn used to generate tightened nominal constraints ensuring robust constraint satisfaction. Furthermore, the tighter the sets $\mathbb{W}(x)$ are, the better is the performance of the algorithm, cf. Chapter 3 in [167] for an in-depth discussion.

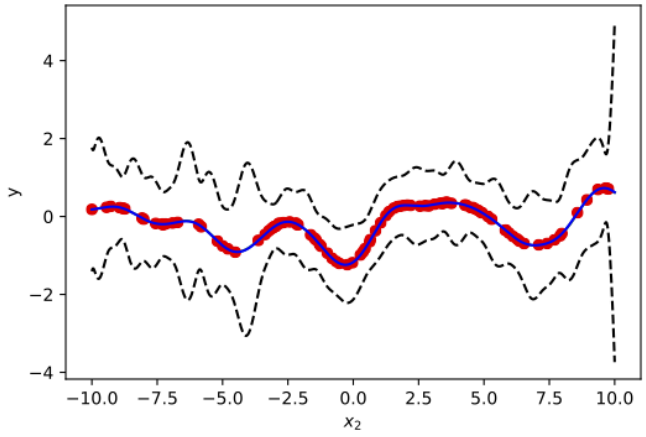

Figure 6.4.: Example nonlinearity. From the target function (blue solid line) 100 samples with noise (red dots) are sampled, which are used to get the uncertainty sets (dashed black lines).

For this experiment, we replaced the Stribeck friction curve used by [185] with a synthetic nonlinearity generated from a known RKHS. Furthermore, the nonlinearity is assumed to be unknown and has to be learned from data. More precisely, the nonlinearity r (which will be our ground truth) is sampled from the pre-RKHS of the scaled SE kernel

$$k(x, x') = 4 \exp\left(-\frac{(x - x')^2}{2 \times 0.8^2}\right)$$

with RKHS norm 2. Following [185], we uniformly sample 100 partial states $x_2 \in [-10, 10]$, evaluate r at these and add i.i.d. Gaussian noise with a standard deviation of 0.01 to it. The unknown function is then learned using GPR (using the nominal setting, i.e., with known k) from this data set. Using (6.2) then leads to an uncertainty set of the form $\mathbb{W}(x) = [\mu_{100}(x_2) - \beta_{100}\sigma_{100}(x_2), \mu_{100}(x_2) + \beta_{100}\sigma_{100}(x_2)]$, where we use $\delta = 0.001$. In particular, with probability at least $1 - \delta$ we can guarantee that $r(x_2) \in \mathbb{W}(x)$ holds for all $x \in \mathbb{X}$. The situation is displayed in Figure 6.4. In order to follow [185] as closely as possible, we exported the learning results and used the original Matlab script to compute $\mathbb{Z}_k$ (provided by R. Soloperto). In order to reduce computation time, we decided to use a 50×50 state space grid and an MPC horizon of 9.

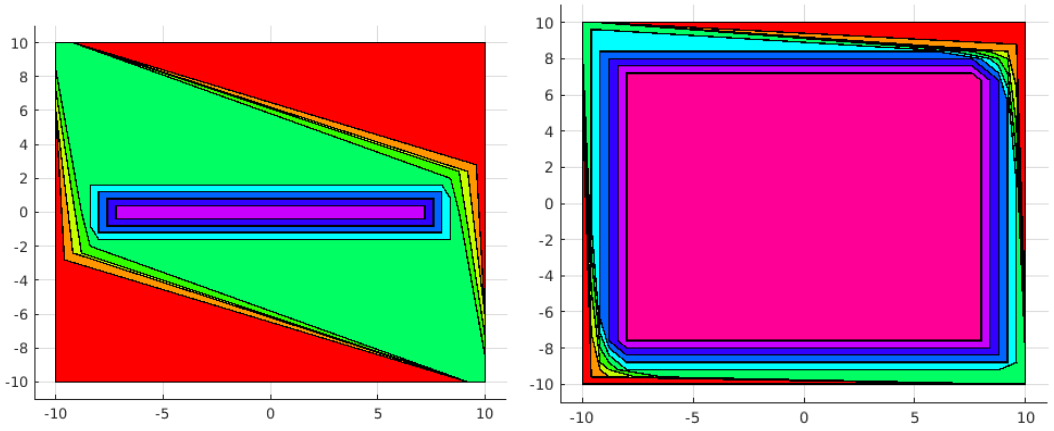

Figure 6.5.: Tightened state constraint sets $\mathbb{Z}_k$ for $k = 0, \ldots, 9$. Computed from a-priori uncertainty set $\mathbb{W}_0$ (LEFT) and learned uncertainty sets $\mathbb{W}(x)$ (RIGHT).

Result

Figure 6.5 shows the resulting tightened state constraints for an MPC horizon of 9. Clearly, the state constraint sets from the learned uncertainty sets are much larger. Furthermore, in contrast to previous work, we can guarantee that the RMPC controller using these tightened state constraints retains all control-theoretic guarantees with probability at least $1 - \delta$. In the present case we can ensure state and control constraint satisfaction, input-to-state stability and convergence to a neighborhood of the origin, with probability at least $1 - \delta$. This follows immediately from [185, Theorem 1], since the ground truth r is covered by the uncertainty sets $\mathbb{W}(x)$ with this probability, and the control-theoretic guarantees are deterministic, so they also hold with the same probability. Note that this is essentially a concrete instantiation of the abstract strategy discussed in Chapter 4.

6.2. Uncertainty bounds in learning-enhanced robust controller synthesis

We will now present an application of frequentist uncertainty bounds for GP regression in the context of robust controller synthesis. First, some background and context will be provided, then we describe our overall methodology, which we then

illustrate with a concrete example from robust control. To the best of our knowledge, this is the first combinatino of actual frequentist uncertainty bounds with modern robust controller synthesis.

6.2.1. Introduction and background

Many methods of modern control rely on accurate plant models [66]. Traditionally, models are obtained from first-principles modeling [16], harnessing extensive expert domain knowledge, or are derived experimentally using system identification [122]. In most cases, a combination of these strategies is used: First-principles models are enhanced with components derived from data. Therefore, recent advances in machine learning offer tremendous opportunities for control. By using advanced learning approaches, it is possible to improve models from real-world data sets, even in established areas of control engineering, cf. e.g. [160]. Furthermore, modern control systems are increasingly complex and learning-based approaches offer the chance to tackle this complexity.

As already eluded to in Chapter 4, control applications pose significant challenges for learning approaches. Beyond statistical and computational issues like such as potential non-independence of data-samples, unmeasured states and real-time feasibility issues, rigorous guarantees on the behavior of closed-loop control systems are of paramount importance in many applications, in particular, in safety-critical areas like aerospace control systems and human-robot interaction scenarios. Using model-based approaches, rigorous and practically relevant guarantees can often be given, mostly in the form of stability and constraint satisfaction guarantees. However, including learning-based components in control systems significantly complicates the situation, and giving theoretical guarantees on the overall system can be challenging. Additionally, for practical applicability of learning-based control it is important to leverage existing prior knowledge in a systematic manner. In many real-world scenarios, extensive domain and expert knowledge is available, often in the form of first-principles modeling utilizing disciplines like physics, chemistry or biology, as well as vast amounts of engineering experience and intuition. In order to make learning-based control approaches real-world feasible, reliable and data efficient, it is important to harness this prior knowledge.

Motivated by this, in this section we present a very general methodology for

learning-enhanced robust controller synthesis. Similarly to the example from Section 6.1.3, it will be an instance of the established strategy described on an abstract level in Chapter 4: We apply a machine learning method to unknown parts of the system, derive frequentist uncertainty bounds, and deal with the remaining uncertainty using robust control. However, we will put a particular emphasis on *using prior knowledge*, hence we prefer to call the methodology learning-*enhanced* instead of learning-*based*. In particular, we demonstrate how to use prior knowledge in a *systematic manner* in the learning process. We advocate for the use of the Linear Fractional Representation (LFR) [223, 177] and Integral Quadratic Constraint (IQC) [136, 209] framework to integrate the learning components with modern robust control approaches, in particular, robust controller synthesis with control-theoretic guarantees like robust stability and performance. In this way, we can give rigorous guarantees and can harness established engineering prior knowledge, cf. Figure 6.6 for an illustration.

Related work Combining uncertainty quantification for machine learning with robust control approaches is a common strategy in learning based control, in particular, in learning-based Model Predictive Control [92], cf. also our discussion in Chapter 4. Using frequentist uncertainty results for GPR in a robust control setting has been used for example in [112, 204, 90, 61], and the concept of δ-safety from [112] is conceptually very similar to our control-theoretic guarantees, cf. Theorem 6.2.8 below. However, due to the difficult applicability of earlier uncertainty bounds for GPR, only heuristics for the uncertainty sets are employed and, hence, all guarantees are lost in the end. Furthermore, often considerable prior knowledge about the system is not used, in particular, the fine structure of the uncertainties is not utilized in the control schemes. The LFR framework has been sporadically used in the context of learning-based control, cf. e.g. [28, 30]. However, its versatility and modularity has not been taken advantage of before. In particular, to the best of our knowledge it has not been used in the context of control-theoretic guarantees for learning-based control schemes. The recent works [26, 27] explore the usage of the LFR framework and modern robust controller synthesis in a learning context, however, they rely on a data-driven approach [93] and do not consider nonlinearities. As such these works can be seen as complementary to the present work.

Another alternative approach to learning-based controller synthesis uses a Bayesian framework, cf. e.g. [214], and can also be seen as complementary to the frequen-

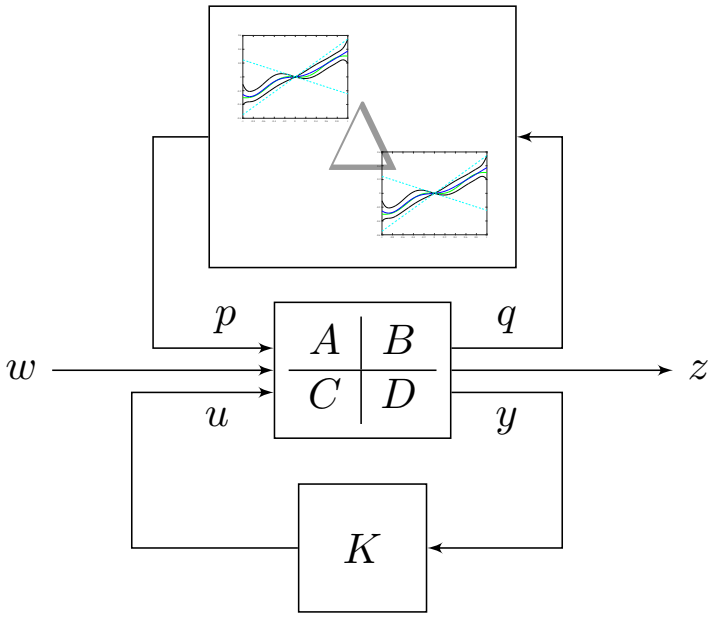

Figure 6.6.: Illustration of the LFR framework in the context of learning-enhanced robust controller synthesis. All uncertain and trouble-making parts are "pulled out" into the Δ-block. We use the framework to combine prior knowledge with components learned from data (in the Δ-block) in a systematic manner. © 2021 IEEE

tist perspective taken in this work. Furthermore, since our learning approach leads to uncertainty sets that can be interpreted as estimates of certain system-theoretic properties, our work is also related to recent methods for inferring such properties directly from data, cf. e.g. [110].

Notation We also need some additional notation. We indicate that a symmetric matrix $A \in \mathbb{S}^{n\times n}$ is positive (semi)-definite by $A(\succeq) \succ 0$. I is the identity matrix. We define $[N] := \{1, \dots, N\}$. Real-rational proper matrix functions without poles on the extended imaginary axis are denoted by $\mathcal{RL}_\infty^{n\times n}$, $\mathcal{L}_2$ is the usual Lebesgue space and $\langle \cdot, \cdot \rangle$ the corresponding inner product. The dimension of a signal x is denoted by n_x. We write $\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$ for the transfer matrix of an LTI system and

$A \star B$ for the usual (lower) LFT of two compatible systems.

Modeling uncertainty using IQCs and LFR In real-world applications, only parts of the plant are unknown and even the remaining uncertainties often carry a lot of structure. It is therefore important to be able to systematically combine known and unknown components of dynamical systems and to leverage sophisticated descriptions of the fine properties of uncertainties. An established framework in modern robust control, that is ideally suited for this task are Linear Fractional Representations (LFR). Consider the partially known system (using operator notation)

$$\begin{pmatrix} z \\ y \end{pmatrix} = P_* \begin{pmatrix} w \\ u \end{pmatrix}, \tag{6.5}$$

with control input u, measured output y, generalized disturbance w and controlled output z. In the LFR framework we model the known parts by a nominal LTI system

$$\begin{pmatrix} q \\ z \\ y \end{pmatrix} = \begin{pmatrix} G_{qp} & G_{qw} & G_{qu} \\ G_{zp} & G_{zw} & G_{zu} \\ G_{yp} & G_{yw} & G_{zu} \end{pmatrix} \begin{pmatrix} p \\ w \\ u \end{pmatrix} \tag{6.6}$$

in feedback connection with an uncertain system

$$p = \Delta(q), \tag{6.7}$$

where $\Delta \in \mathbf{\Delta}$, with the latter being the class of given uncertainties. In many cases substantial additional information about Δ is known, in particular, if machine learning methods are employed to reduce the epistemic uncertainty about the system. Here we propose to use the versatile and powerful framework of Integral Quadratic Constraints (IQCs) for this task. We say that an uncertainty Δ fulfills the IQC defined by the multiplier $\Pi \in \mathcal{RL}_\infty^{(n_q+n_p)\times(n_q+n_p)}$ if

$$\left\langle \begin{pmatrix} q \\ \Delta(q) \end{pmatrix}, \Pi \begin{pmatrix} q \\ \Delta(q) \end{pmatrix} \right\rangle \geq 0 \qquad \forall q \in \mathcal{L}_2. \tag{6.8}$$

For more details and a very comprehensive collection of multipliers we refer to [209]. Furthermore, IQCs can be used for robust stability and performance analysis: Con-

sider the system

$$\begin{pmatrix} q \\ z \end{pmatrix} = \begin{pmatrix} \mathcal{G}_{qp} & \mathcal{G}_{qw} \\ \mathcal{G}_{zp} & \mathcal{G}_{zw} \end{pmatrix} \begin{pmatrix} p \\ w \end{pmatrix} \tag{6.9}$$

resulting from connecting system (6.6) with a given LTI controller K. If an IQC description of the uncertainty is available, the well-known IQC stability theorem from [136] allows to conclude robust stability and even robust performance, cf. [209] for details. Using the KYP Lemma, cf. e.g. [165], leads to an LMI problem and makes the robust stability and performance analysis with IQCs computationally tractable, see again [209] for details.

6.2.2. Methodology

We are now ready to describe our proposed methodology for learning-enhanced robust controller synthesis. First, we precisely describe our problem setting. For concreteness, we will then focus on static nonlinearities, for which the uncertainty bounds described earlier are immediately applicable. Finally, we describe how the resulting uncertainty sets can be used for robust controller synthesis.

Problem setting

We propose to utilize the LFR framework described in Section 6.2.1 since it is an established and flexible framework for uncertain system modeling [223, 177]. In particular, it is well suited to integrate machine learning components into modern robust control: extensive prior system knowledge can be included into the nominal plant and all learning components are "pulled out" into the uncertain part. By using machine learning methods incorporating prior knowledge, it is possible to arrive at very data-efficient learning-enhanced control schemes. Furthermore, as we will demonstrate in the following, the LFR framework allows us to easily transfer statistical bounds into control-theoretic guarantees in a structured fashion.

For the rest of this section, we are concerned with an unknown plant (6.5) given in the LFR form (6.6), (6.7). The overall goal is to perform robust controller synthesis and to give robust stability and performance guarantees for the synthesized controller. For sake of concreteness, we assume a single objective optimization where possible performance weighting filters have been already included in the generalized

plant P and we are left with a standard robust (against the uncertainty in (6.7)) H_∞ synthesis problem.

Assume now that we are not satisfied with the achievable performance of the controller due to the uncertainty being too large. For this purpose, we will employ machine learning in order to reduce the epistemic uncertainty contained in (6.7), using a data set $\mathcal{D}$. Since the LFR framework is very modular, it causes no limitation to focus our attention to a single uncertain component that is to be learned. We would like to stress that it is easily possible to deal with multiple uncertainties containing learned components, even using different learning methods, and to combine these with uncertainties that classically emerge if capturing parametric model variations or unmodeled dynamics as resulting from a complexity reduction step, cf. [66].

Instead of describing an abstract general procedure, we opt to present the concrete case of a static (partially) unknown diagonal nonlinearity,

$$p = \phi(q), \tag{6.10}$$

with $n_p = n_q$ and $p_i(t) = \phi_i(q_i(t))$ for $i \in [n_p]$, $t \geq 0$. We will comment on possible generalizations and resulting complications during the description of our approach.

Learning static nonlinearities

Since the described procedure can be applied to all ϕ_i, $i \in [n_p]$ separately, we focus on a single scalar nonlinearity $\phi : \mathbb{R} \to \mathbb{R}$ in this and the next section. The corresponding data set $\mathcal{D} = ((x_n, y_n))_{n\in[N]}$ is of the form

$$y_n = \phi(x_n) + \epsilon_n, \tag{6.11}$$

where $\epsilon_1, \ldots, \epsilon_N$ is additive measurement noise. This means that the unknown part can be isolated and input-output samples can be collected. For concreteness, we make the following assumption on the noise.

Assumption 6.2.1. $\epsilon_1, \ldots, \epsilon_N$ are independent, R-subgaussian random variables with $R \geq 0$ the subgaussianity constant.

Note that this assumption not only entails that the noise is subgaussian, which

is a mild assumption in many engineering applications and covers many noise distributions encountered in practice, but also that we know a concrete subgaussian constant. Whether this is latter assumption is reasonable depends on the concrete application.

If we want to use GP regression and the frequentist uncertainty bounds from Chapter 5, we also need the following assumption.

Assumption 6.2.2. k is a kernel on $\mathbb{R}$ and $\phi \in H_k$ with $\|\phi\|_k \leq B$.

Note that this assumption is standard in the learning-based control literature, [112]. However, it turns out to be very problematic, as we will discuss in detail in the next chapter. For now, we follow learning-based control literature and accept this assumption. We can now use the uncertainty bounds from 5, and for consistency with the preceding sections, we use (6.2). Note that we use the notation $\beta_{\mathcal{D}}$ instead of β_N to stress that we are not in a sequential setup. For ease of reference, we record the resulting uncertainty bound here.

Lemma 6.2.3. If applying GP regression with covariance function k and nominal noise level $\lambda > 0$ to data set $\mathcal{D}$, we get with the notation introduced above, for any $\delta \in (0, 1)$ under Assumption 6.2.1 and 6.2.2 that

$$\mathbb{P}[|\phi(x) - \mu_{\mathcal{D}}(x)| \leq \beta_{\mathcal{D}}\sigma_{\mathcal{D}}(x)\, \forall x \in \mathbb{R}] \geq 1 - \delta.$$

This is illustrated in Figure 6.7 with the example from Section 6.2.4.

Additional prior knowledge on the static nonlinearity can be systematically included in GPR via the covariance function. As a concrete example, suppose we know that $\phi(0) = 0$, which is a requirement for a sector bounded nonlinearity. Given any kernel k with $k(0, 0) \neq 0$, following the procedure described in [100] leads to the new kernel

$$k_0(x, x') = k(x, x') - \frac{1}{k(0, 0)} k(x, 0)k(x', 0). \tag{6.12}$$

Note that in this particular example, the same effect can be achieved conditioning a GP prior on the noise free virtual data point $(0, 0)$. All functions contained in H_{k_0} as well as all samples from $\mathcal{GP}(0, k_0)$ are guaranteed to be zero in zero. Many other properties of functions can be encoded in this way, cf. e.g. [100, 79].

Remark 6.2.4. We would like to provide some remarks on the setup just described.

1. We would like to stress that the LFR framework is very general and supports a broad range of uncertainties. In particular, many other uncertainties and even other learning methods can be included in (6.7). For ease of presentation, we restrict ourselves to static nonlinearities and only one learning component, the indicated generalizations are then straightforward.

2. It is possible to deal with multivariate static nonlinearities with multi-output GPs, cf. the corresponding remarks in Chapter 4.

3. The independence assumption on the noise is reasonable in the setting used here, since we directly collect input-output samples, but of course the time-uniform uncertainty bounds from Chapter 5 allow for much more general noise assumptions.

4. Other machine learning methods can be used for this step, as long as numerical uncertainty bounds can be derived. For example, if we have magnitude bounds on the additive noise in (6.11) and no distributional assumptions, then the kernel methods and accompanying uncertainty bounds in [127] or the Nonlinear Set Membership framework [140] could as well be employed. We will come back to this point in Chapter 7.

Connection to robust control: High probability sector bounds

As common in modern robust control, the uncertainty set has to be transformed into a form that is amenable to controller synthesis [66]. As a concrete example, we illustrate this by deriving the tightest sector bounds compatible with the high probability uncertainty set from Section 6.2.2. Recall that $\varphi : \mathbb{R} \to \mathbb{R}$ belongs to a sector $[\kappa_1, \kappa_2]$ if

$$\kappa_1 x^2 \leq x\varphi(x) \leq \kappa_2 x^2 \qquad \forall x \in \mathbb{R}. \tag{6.13}$$

Generalizations to the multivariate settings are straightforward, cf. e.g. [106]. Lemma 6.2.3 leads immediately to the next result.

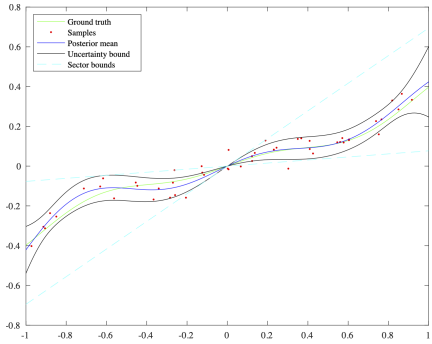

Figure 6.7.: The unknown nonlinearity is learned with GPR from samples, resulting in a high probability uncertainty set. This is used to derive sector bound estimates for the nonlinearity. © 2021 IEEE

Lemma 6.2.5. Consider the situation of Lemma 6.2.3 and assume that ϕ belongs to the sector $[\kappa_1, \kappa_2]$ with $\kappa_1 \leq \kappa_2$. If defining

$$\hat{\kappa}_1 := \min_{x \in [a,b] \setminus \{0\}} \frac{\min\{x \cdot (\mu_{\mathcal{D}}(x) - \beta_{\mathcal{D}}\sigma_{\mathcal{D}}(x)), x \cdot (\mu_{\mathcal{D}}(x) + \beta_{\mathcal{D}}\sigma_{\mathcal{D}}(x))\}}{x^2}$$
$$\hat{\kappa}_2 := \max_{x \in [a,b] \setminus \{0\}} \frac{\max\{x \cdot (\mu_{\mathcal{D}}(x) - \beta_{\mathcal{D}}\sigma_{\mathcal{D}}(x)), x \cdot (\mu_{\mathcal{D}}(x) + \beta_{\mathcal{D}}\sigma_{\mathcal{D}}(x))\}}{x^2},$$

then $[\hat{\kappa}_1, \hat{\kappa}_2]$ is an overapproximation of $[\kappa_1, \kappa_2]$, i.e. $\hat{\kappa}_1 \leq \kappa_1$ and $\hat{\kappa_2} \geq \kappa_2$, with probability at least $1 - \delta$.

This approach can be made rigorously computational for Lipschitz continuous functions by evaluating the uncertainty set from Lemma 6.2.3 on a fine grid and adapting Lemma 6.2.5 correspondingly.

Robust controller synthesis

We now tackle the robust controller synthesis problem with the learned uncertainty component using the versatile approach from [207] for the case of static IQC multipliers. Consider without loss of generality the following minimal realization of (6.6) (with possible performance weights already included in the generalized plant)

$$\begin{pmatrix} \dot{x} \\ \hline q \\ z \\ y \end{pmatrix} = \left(\begin{array}{c|ccc} A & B_p & B_w & B_u \\ \hline C_q & D_{qp} & D_{qw} & D_{qu} \\ C_z & D_{zp} & D_{zw} & D_{zu} \\ C_y & D_{yp} & D_{yw} & 0 \end{array}\right) \begin{pmatrix} x \\ p \\ w \\ u \end{pmatrix}. \tag{6.14}$$

Canceling the uncertainty channel $p \to q$ results in the system G_0. A standard H_∞ controller synthesis leads to an intial LTI controller K_0 with nominal performance level γ_0. Connecting K_0 to G on the control channel $u \to y$ results in

$$\mathcal{G}_0 = G \star K_0 = \left[\begin{array}{c|cc} \mathcal{A} & \mathcal{B}_p & \mathcal{B}_w \\ \hline \mathcal{C}_q & \mathcal{D}_{qp} & \mathcal{D}_{qw} \\ \mathcal{C}_z & \mathcal{D}_{zp} & \mathcal{D}_{zw} \end{array}\right]. \tag{6.15}$$

The corresponding realization can be derived by elementary manipulations and is omitted due to space constraints. Assume now that $\mathbf{\Delta}$ fulfills an IQC with multipliers $\Pi = P \in \mathcal{P}$, where $\mathcal{P}$ has an LMI description. This leads to an initial analysis LMI (6.16) with decision variables $\gamma > 0$, $\mathcal{X} = \mathcal{X}^\top$ and the LMI variable arising from $\mathcal{P}$.

$$\begin{pmatrix} \mathcal{X}\mathcal{A} + \mathcal{A}^\top\mathcal{X} & \mathcal{X}\mathcal{B}_p & \mathcal{X}\mathcal{B}_w & \mathcal{C}_z^\top \\ \mathcal{B}_p^\top\mathcal{X} & 0 & 0 & \mathcal{D}_{zp}^\top \\ \mathcal{B}_w^\top\mathcal{X} & 0 & -\gamma I & \mathcal{D}_{zw}^\top \\ \mathcal{C}_z & \mathcal{D}_{zp} & \mathcal{D}_{zw} & -\gamma I \end{pmatrix} + \begin{pmatrix} * \end{pmatrix}^\top P \begin{pmatrix} \mathcal{C}_q & \mathcal{D}_{qp} & \mathcal{D}_{qw} & 0 \\ 0 & I & 0 & 0 \end{pmatrix} \prec 0 \tag{6.16}$$

$$\begin{pmatrix} \mathcal{X}\mathcal{A} + \mathcal{A}^\top\mathcal{X} & \mathcal{X}\mathcal{B}_p\Psi_2^{-1} & \mathcal{X}\mathcal{B}_w & \mathcal{C}_z^\top \\ (\Psi_2^{-1})^\top\mathcal{B}_p^\top\mathcal{X} & 0 & 0 & (\Psi_2^{-1})^\top\mathcal{D}_{zp}^\top \\ \mathcal{B}_w^\top\mathcal{X} & 0 & -\gamma I & \mathcal{D}_{zw}^\top \\ \mathcal{C}_z & \mathcal{D}_{zp}\Psi_2^{-1} & \mathcal{D}_{zw} & -\gamma I \end{pmatrix} + \begin{pmatrix} * \end{pmatrix}^\top \hat{P} \begin{pmatrix} \Psi_1\mathcal{C}_q & (\Psi_1\mathcal{D}_{qp} + \Psi_3)\Psi_2^{-1} & \Psi_1\mathcal{D}_{qw} & 0 \\ 0 & I & 0 & 0 \end{pmatrix} \prec 0 \tag{6.17}$$

Minimizing γ leads to robust performance level $\tilde{\gamma}_0$ and corresponding multiplier P. Then consider the following factorization of the IQC multiplier,

$$P = \begin{pmatrix} * \end{pmatrix}^\top \hat{P} \begin{pmatrix} \Psi_1 & \Psi_3 \\ 0_{q\times p} & \Psi_2 \end{pmatrix} \tag{6.18}$$

with $\Psi_1 \in \mathbb{R}^{n_q \times n_q}$, $\Psi_3 \in \mathbb{R}^{n_q \times n_p}$, $\Psi_2 \in \mathbb{R}^{n_p \times n_p}$, Ψ_2 invertible and where $\hat{P}$ has the form

$$\begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix} \text{ or } \begin{pmatrix} I & 0 \\ 0 & -I \end{pmatrix}. \tag{6.19}$$

Note that this factorization is possible for all static multiplier classes of interest. Applying the factorization in (6.16), together with elementary manipulations and a congruence transformation with $\mathrm{diag}(I, \Psi_2^{-1}, I, I)$, results in (6.17). We recognize

that (6.17) is the initial analysis LMI (6.16) for the plant $G_1 \star K_0$, where

$$G_1 = \left[\begin{array}{c|cc} A & B_p\Psi_2^{-1} & \begin{matrix} B_w & B_u \end{matrix} \\ \hline \Psi_1 C_q & \Psi_1 D_{qp}\Psi_2^{-1} + \Psi_3\Psi_2^{-1} & \begin{matrix} \Psi_1 D_{qw} & \Psi_1 D_{qu} \end{matrix} \\ C_z & D_{zp}\Psi_2^{-1} & \begin{matrix} D_{zw} & D_{zu} \end{matrix} \\ C_y & D_{yp}\Psi_2^{-1} & \begin{matrix} D_{yw} & D_{yu} \end{matrix} \end{array}\right] \tag{6.20}$$

with new performance channels $w_1 \to z_1$ and $w \to z$. We can now run a standard controller synthesis on the generalized plant G_1 with the quadratic performance criterion

$$\left\langle \begin{pmatrix} z_1 \\ w_1 \end{pmatrix}, \hat{P} \begin{pmatrix} z_1 \\ w_1 \end{pmatrix} \right\rangle + \frac{1}{\gamma}\|w\|^2 - \gamma\|z\|^2 \leq -\epsilon(\|w_1\|^2 + \|w\|^2), \tag{6.21}$$

e.g. along the lines described in [179]. Minimization of γ leads to the robustified controller K_1 and performance level γ_1. These steps can now be iterated until no substantial improvement in γ is obtained.

Returning to the concrete setting of sector bounded nonlinearities, suppose that the method from Section 6.2.2 and 6.2.2 resulted in sector bound estimates $[\hat{\kappa}_1^{(i)}, \hat{\kappa}_2^{(i)}]$, $i \in [n_p]$. To proceed with the synthesis, we need another assumption.

Assumption 6.2.6. For all $i \in [n_p]$ we have $\hat{\kappa}_1^{(i)} \leq 0 \leq \hat{\kappa}_2^{(i)}$.

We can now express the estimated sector bounds with full block multipliers.

Lemma 6.2.7. Let $\delta \in (0,1)$ and run the learning method in Section 6.2.2 and 6.2.2 on each nonlinearity using independent data sets and replacing δ with δ/n_p. Define

$$\mathcal{P}_{\text{fb}} = \left\{ P \mid (\cdot)^\top P \begin{pmatrix} I \\ \Theta(\theta) \end{pmatrix} \succeq 0,\ (\cdot)^\top P \begin{pmatrix} 0 \\ I \end{pmatrix} \preceq 0,\ \theta \in \hat{\mathcal{K}} \right\}, \tag{6.22}$$

where $\hat{\mathcal{K}} = \left\{ (\theta_1, \ldots, \theta_{n_p}) \mid \theta_i \in \{\hat{\kappa}_1^{(i)}, \hat{\kappa}_2^{(i)}\},\ i \in [n_p] \right\}$ and for brevity $\Theta(\theta) = \operatorname{diag}(\theta)$. Then under Assumption 6.2.1, 6.2.2, 6.2.6, with probability at least $1-\delta$, (6.8) holds for all $\Pi \in \mathcal{P}_{\text{fb}}$.

Proof. Follows immediately from [207, Lemma 4.1], Lemma 6.2.5 and the union bound. □

Since the multipliers from (6.22) can be factorized according to (6.18), the synthesis-analysis procedure described above can be directly used.

Remarks. We would like to stress that the approach in [207] is much more general than the simplified setting used here for illustrative purposes. In particular, dynamic IQC multipliers can be used (involving non-trivial factorization results) which allow significantly more precise uncertainty descriptions leading to reduced conservatism. This is especially relevant for learning-based approaches, since the later might provide considerable additional information on the uncertainty. Furthermore, warm-start strategies are proposed in [207] in order to speed up the robust controller synthesis procedure. Finally, an approach for IQC-based robust synthesis in the context of gain scheduling is proposed in [206]. The extension of our approach to this setting is left for future work.

6.2.3. From statistical to control-theoretic guarantees

We can now give overall guarantees on the resulting controller.

Theorem 6.2.8. Let $\delta \in (0,1)$ and suppose that the learning procedure as outlined above is run with δ replaced by δ/n_p and $\mathcal{P}_{\text{fb}}$ is built according to (6.22). Under Assumption 6.2.1 to 6.2.6, if the controller synthesis algorithm described in Section 6.2.2 returns a controller K after $M \geq 1$ iterations with robust performance level $\tilde{\gamma}_{M+1}$, then with probability at least $1-\delta$, K will stabilize the true system and achieve robust performance level $\tilde{\gamma}_{M+1}$.

Proof. Follows immediate from Lemma 6.2.7 and the results in [207]. □

Let us discuss this result. For the learning component, we assume that there exists a fixed (but unknown) ground truth, here the diagonal nonlinearity (6.10). We derive from the learning component an uncertainty set, here described by the full block IQC multipliers (6.22), that contains the ground truth with given (high) probability $1-\delta$. This is combined with a robust method that comes with control-theoretic guarantees for this uncertainty set. Since the ground truth is contained in the uncertainty set with probability at least $1-\delta$, the guarantees hold with the same high probability.

Before turning to a concrete example, we would like to recall some of our arguments from Chapter 4. It is important to contrast this approach with randomized

methods in robust control [198], like the scenario approach [42]. In these methods, high-probability guarantees are given with respect to the randomization in the algorithms, while in our problem the randomness comes only from the data. Moreover, the results here are also different in nature if compared to work done in a Bayesian framework. In particular, we do not rely on a prior distribution which might be misspecified. We argue that in the modern robust control setting, the frequentist perspective is the most natural statistical setting since we assume a fixed ground truth, described by an uncertainty class.

6.2.4. A concrete example

To demonstrate the advantage of the learning component in robust control, we now compare the controller synthesis using a-priori sector bounds and the bounds learned using GPR. As a concrete example, we use the distillation column system from [207, Section 8.2], where we replace the original dead-zone nonlinearity with the unknown target function to be learned.

Setup

The situation is depicted as a blockdiagram in Figure 6.8. The plant model is given by

$$G(s) = \frac{1}{75s+1}\begin{pmatrix} 87.8 & -86.4 \\ 108.2 & -109.6 \end{pmatrix} \tag{6.23}$$

and we use the performance weights

$$W_e(s) = \frac{s+0.1}{2s+10^{-5}} I_2 \qquad W_u(s) = \frac{s+10}{s+100} I_2. \tag{6.24}$$

The synthesis objective is therefore to track the reference signal r at low frequencies and penalizing control action at high frequencies. For illustrative purposes we use the same unknown nonlinearity in both uncertainty channels, i.e., $p_i = \phi(q_i)$, $i = 1, 2$. For the experimental validation we need access to the ground truth, hence we use a function $\phi \in H_{k_0}$, where k_0 is given by (6.12) and for k we choose the popular Squared Exponential (SE) kernel $k(x, x') = \sigma_k^2 \exp\left(-\frac{(x-x')^2}{2\ell^2}\right)$ with lengthscale $\ell = 0.5$ and variance $\sigma_k^2 = 0.5$. The particular function ϕ is shown in Figure 6.7 and has RKHS norm 2.6053. Since we are interested in stabilization, we restrict attention to

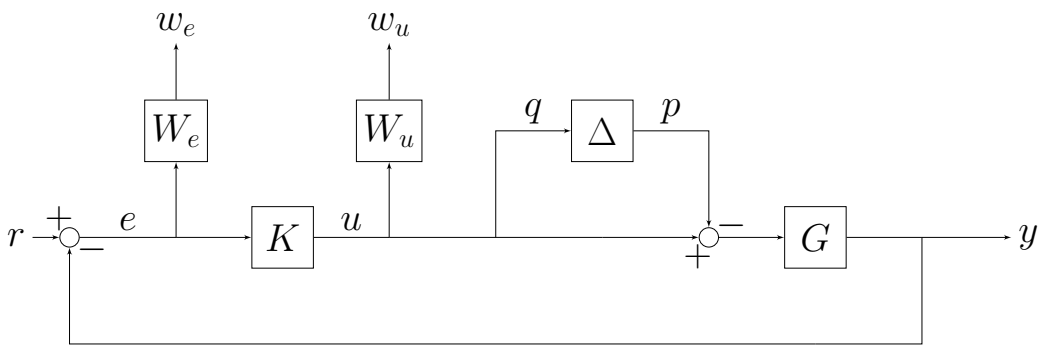

Figure 6.8.: Block diagram of the example. Here Δ is a repeated static nonlinearity. © 2021 IEEE

inputs to ϕ from $[-1, 1]$ and ϕ fulfills the sector condition $[-0.0572, 0.3575]$ on this interval. Note that it is possible to ensure containment of the relevant signals in such an interval with techniques from robust control, but due to space constraints we omit the details.

Using standard manipulations, the robust synthesis problem with performance weights from (6.24) and diagonal static nonlinearities can be framed as an LFR. Furthermore, we use the full block multipliers from Lemma 6.2.7. In the following experiments, we run the synthesis procedure outlined in Section 6.2.2 for 20 iterations.

Finally, to test the learning method we generate data sets $\mathcal{D} = ((x_n, y_n))_{n\in[N]}$ by sampling $N = 50$ inputs x_n uniformely from $[-1, 1]$, evaluate ϕ at these inputs and add independent $\mathcal{N}(0, 0.05)$ noise to get y_n.

Improving performance with learning

Assume we know a priori that ϕ belongs to the (rather wide) sector $[-0.9, 0.9]$. Running the robust controller synthesis leads to a controller with robust performance level 10.5575. This is now compared with the methodology described above. We use GP regression with a zero prior mean function and the kernel k_0 as covariance function as well as the true noise level 0.05. We follow previous works, e.g. [127], and assume an increased RKHS norm bound $B = 2\|\phi\|_{k_0}$ in the uncertainty bound. Setting $\delta = 0.001$ and following the procedure outlined in Section 6.2.2 leads to the sector bounds $[-0.2460, 0.5480]$. The situation is illustrated in Figure 6.7. Applying

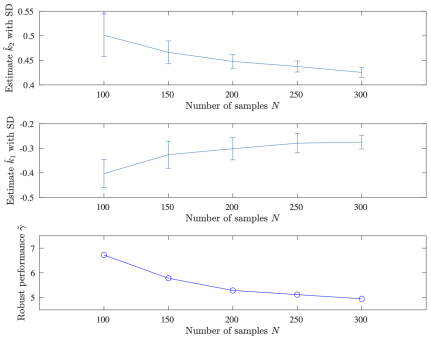

Figure 6.9.: Effect of increased data sets. In the upper two plots the estimated sector bounds for different data sizes are shown (averaged over 50 runs with corresponding empirical standard deviation), the lower plot shows the robust performance level of the resulting controllers. © 2022 IEEE

the controller synthesis method with this sector bound leads to a controller that has robust performance level 3.1581 at least with probability $1 - 0.001$.

Data-performance tradeoff

We now investigate the potential performance improvements with increasing size of the data set. For this we generate data sets with 100, 150, 200, 250, 300 data points and applied the learning method to these data sets, using the same procedures as in the last experiment. We repeated this 50 times. The resulting estimates of the sector bounds are shown in Figure 6.9 (upper two plots). It is clearly visible that more data is helpful for the learning method, since the sector bounds become tighter. Furthermore, running the robust controller synthesis on the estimated sector bounds (averaged over all 50 trials) leads to the results shown in Figure 6.9 (lower plot). Clearly, increased data helps to improve the robust performance. Note that the robust stability and performance of the controller is guaranteed with user-specified probability, while profiting from the additional information contained in the data set.

6.3. Conclusion

In Section 6.1, we conducted a thorough investigation of frequentist uncertainty bounds for GP regression from a frequentist perspective. To the best of our knowl-

edge, this was the first such systematic evaluation, cf. also the comments in the next section. We found that in a well-specified setting, variants of established uncertainty bounds are on the same order of magnitude as common heuristics. Furthermore, our empirical results strongly indicate that in some misspecified settings robust uncertainty bounds are required, and our results from Chapter 5 are suitable for some classes of misspecification. However, we also found some conservatism, and improved robust uncertainty bounds are an interesting subject for future work. Finally, as discussed above, numerical evidence in a frequentist setup has to be cautiously interpreted, and we tried to carefully describe and circumvent pitfalls, for example by using different generation methods for RKHS functions.

Motivated by the practically of the considered uncertainty bounds, in Section 6.2 we present a very general framework for learning-enhanced robust controller synthesis. By leveraging the establish LFR structure together with IQCs for describing uncertainties, we can easily include prior knowledge in the overall approach, and by relying on GP regression, we can include further prior knowledge in learning process and ensure rigorous statistical and control-theoretic guarantees. Tot he best of our knowledge, this is the first usage of actual frequentist uncertainty bounds together with modern robust controller synthesis.

Finally, assuming a well-specified modelling setup, all of the above relies on the knowledge of a subgaussianity constant for the noise and an upper bound on the RKHS norm of the target function. While the former is usually a rather mild assumption, the latter is very problematic in practice, as we will discuss in detail in the next chapter, where we will also propose some solutions to this problem.

6.4. Comments

Section 6.1 is based on and to a large extent taken verbatim from [CF12]. This latter work arose from discussions of the author with the S. Trimpe and C.W. Scherer. The author of this thesis designed and performed all experiments in that work, and wrote the manuscript, with editorial comments from C.W. Scherer and S. Trimpe. Section 6.2 is based on and to a large extent taken verbatim from [CF10]. The overall methodology has been suggested by C.W. Scherer and S. Trimpe, and the realization, including the formalization and all experiments, as well as the majority of the writing, has been done by the author of this thesis. The robust controller

synthesis in Section 6.2 has been performed with Matlab code from J. Veenman, cf. also [208].

Some comments on [CF12] The work on [CF12] was conducted in 2019 and 2020, and at that point, the two most popular results on frequentist uncertainty bounds for GP regression, especially in the context of learning-based control, were [186, Theorem 6] and [54, Theorem 2]. The former involves very large constants and requires bounded noise, and hence the latter result, involving smaller constants and requiring only conditionally subgaussian noise, should be preferred. The author of this thesis noticed that in the learning-based control literature, essentially all works using [186, Theorem 6] or [54, Theorem 2] replaced the scaling factors β_N from these results by some heuristic value, usually $\beta_N \equiv 2$ or 3, which breaks in general all guarantees of the overall algorithm. One reason for this could be that both of these results are formulated in terms of the maximum information gain, cf. [186] for background on this quantity, and in general, working numerically with this quantity is not convenient[2] and might lead to some conservatism. The main motivation of [CF12] was to provide bounds that can be computed and actually used in learning-based control algorithms. The author of this thesis noticed that the derivation of [54, Theorem 2] can be "stopped early", leading to an uncertainty bound that can be computed easily (if an RKHS norm bound and a subgaussian constant for the noise is known), and to the best of our knowledge, this fact was not used before in learning-based control, and hence seemed at that point to be a new contribution. However, the PhD thesis [1] contains essentially an uncertainty bound superior to [54, Theorem 2], but unfortunately this bound is not explicitly stated in [1] (though it is explained how to derive it in [1, Remark 3.4]), and it seems that these results from [1] have not been published apart from this thesis. During the preparation of this thesis, we also realized that similar results are contained in [128] and [67]. However, curiously these results have not been taken up in the learning-based control literature.

Since the corresponding result in [CF12] is based on [54, Theorem 2], it inherits the limitation of the latter to $\lambda \geq 1$ (and $\lambda > 1$ if the kernel is not positive definite).

[2] If a constant factor of conservatism and using a numerical approximation are deemed acceptable, then one can compute an upper bound on this quantity algorithmically, cf. [186] for details. We thank A. Krause for pointing this out.

To circumvent this limitation, a new scaling factor $\bar{\lambda}$ was introduced[3], which leads to (6.2) and (6.3). Proposition 5.4.1, essentially contained in [1] and rediscovered by [219], does not have this limitation, i.e., any $\lambda > 0$ is permitted. During the preparation of [CF8], the author noticed that in the setting of this latter result, for $0 < \lambda < 1$ and k positive definite, we have $\bar{\lambda} = \max\{1, \lambda\} = 1$ and hence

$$\begin{aligned}\frac{R}{\sqrt{\lambda}}\sqrt{2\ln\left(\frac{1}{\delta}\det\left(I_N + \frac{1}{\lambda}\boldsymbol{K}_N\right)\right)} &= \frac{R}{\sqrt{\lambda}}\sqrt{\ln\left(\det(1/\lambda\boldsymbol{K}_t + I_N)\right) - 2\ln(\delta)} \\ &= \frac{R}{\sqrt{\lambda}}\sqrt{\ln\left(\det(\bar{\lambda}/\lambda\boldsymbol{K}_N + \bar{\lambda}I_N)\right) - 2\ln(\delta)},\end{aligned}$$

so in this case [CF12, Theorem 1] (corresponding to (6.2)) reproduces the result Proposition 5.4.1, and similarly for (6.3). Additionally, since the only difference happens inside $\sqrt{\ln(\cdot)}$, any noticable difference between the two bounds will happen for $\lambda >> 1$, so any difference will be neglegible in practice. For this reason we decided to report the original experiments from [CF12] using (6.2) and (6.3).

Given this convoluted history of [CF12], we would like to explicitly summarize the contributions of this work from our current perspective.

- We pointed out the lack of convenient, computable frequentist uncertainty bounds for GP regression, at least in the context of learning-based control, and provided such bounds. This should be seen more as a conceptual contribution and *not* a significant technical contribution.[4]

- To the best of our knowledge, we provided the first frequentist uncertainty bounds for GP regression (in the form as usual used in learning-based control, cf. also our discussion on this point in Chapter 5) under model misspecification.

- We conducted the first systematic frequentist evaluation of such uncertainty bounds for GP regression, and pointing out some of the problems (like a bias in the randomly generated RKHS functions).

[3] Unfortunately, in [CF12] this was done incorrectly, and subsequently corrected [CF13]. We would like to thank L. Kreisköther and D. Baumann, who discovered this error. Note that the mistake leads to only minor quantitative changes in the results of [CF12], but not qualitative differences and no change in the conclusions and arguments of this work.

[4] In fact, even in the original work [CF12] we explicitly stated that our bounds are based directly on [54, Theorem 2].

- To the best of our knowledge, we applied computable frequentist uncertainty bounds to a (simple and existing) learning-based control application for the first time. This is *not* a technical innovation, since uncertainty bounds like [186, Theorem 6] and [54, Theorem 2] have been very popular in the context of learning-based control. However, to the best of our knowledge, *using the actual bounds in the algorithm instead of a heuristic* seems to have not been done before in learning-based control.

The simple bounds like [CF12, Proposition 2] could be interpreted as an additional minor contribution, however, the concrete form used in [CF12] is probably too conservative for applications, and the updated variants from Chapter 5 should be preferred. As a sidenote, an in-depth investigation and numerical evaluation of the latter could be an interesting avenue for future work.

7. Geometric prior knowledge and uncertainty bounds in learning-based control

In Chapter 6, we demonstrated through numerical experiments that frequentist uncertainty bounds as presented in Chapter 5 fulfill some of the desiderata arising from learning-based control applications, cf. also our discussion in Chapter 4. In particular, the uncertainty bounds can be evaluated numerically, and the empirical results indicate that the resulting uncertainty sets are tight enough for application in learning-based control scenarios. However, naturally these results need some assumptions, and in order to be useful in practice, these assumptions need to be reasonable from practicioner's perspective. In Section 7.1, we therefore carefully discuss the prior knowledge required for the uncertainty bounds presented and investigated in Chapters 5 and 6. In turns out that the requirement of a known bound on the RKHS norm of the target function is very problematic in the context of learning-based control, and we propose to use assumptions that are more geometric in nature as a replacement. In Section 7.2, we present several examples of such assumptions, and outline an approach that allows combining them with kernel methods.

This chapter is based on, with some parts taken verbatim from, [CF11][1] as well as [CF8]. Detailed comments on the author's contribution and the relation of this chapter to existing work are provided in Section 7.4.

[1] © IEEE 2022. Reprinted, with permission, from Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. *Learning functions and uncertainty sets using geometrically constrained kernel regression.* 61st IEEE Conference on Decision and Control (CDC), 2022.

7.1. The delicate question of quantitative prior knowledge

Let us go back to the high level picture outlined in Chapter 4, and recall the following popular strategy in learning-based control that we also use. A machine learning method is utilized to learn more about the underlying system, and the remaining uncertainty is quantified using frequentist uncertainty sets. Subsequently, these are transformed into a format suitable for a robust control method, which is finally used to control the system, and the control-theoretic guarantees then hold with the high confidence from the frequentist uncertainty sets. However, to get these end-to-end guarantees, we need concrete and valid frequentist uncertainty sets. The former means that the existence of such uncertainty sets is not enough, we need to actually compute them in a format suitable for the downstream robust control method, and the latter says that these uncertainty sets have to contain the ground truth with a prescribed high confidence. If no additional post-hoc verification of the resulting controller is performed, then we have to rely on the validity of the uncertainty sets. Since without some assumptions, it is impossible to get non-vacuous uncertainty sets in non-trivial situations, this means that all the guarantees ultimately rely on the prior knowledge used to derive the uncertainty sets. In particular, this strongly suggests that only *established prior knowledge that is deemed reasonable by the users* should be used to get these uncertainty sets.

To further illustrate the importance of this point, let us briefly introduce a particular variant of *safe Bayesian optimization*. The overall goal of Bayesian optimization (BO) is to optimize an unknown target function $f_* : \mathcal{X} \to \mathbb{R}$. The function is only accessible through noisy evaluations, i.e., the optimization algorithm can choose some input $x \in \mathcal{X}$, query f_* at this input, and receives a noise evaluation $y = f_*(x) + \eta$. Most BO algorithms maintain an internal model of the target function, usually via GP regression, which is used to decide which input to query next. Usually in BO it is expensive to query the target function, for example, because it corresponds to some physical experiment like running and evaluating a controller, and hence the goal is to be query-efficient. In safe BO, some inputs are unsafe and hence have to be avoided as queries during the optimization process. For example, in controller tuning with BO, these can correspond to controller parameters leading to instabilities or crashes of the plant or robot that is controlled. For the class of SafeOpt-type algorithms, as introduced by [192], the set of safe inputs is modelled by $S = \{s \in \mathcal{X} \mid g_*(x) \geq 0\}$,

where $g_* : \mathcal{X} \to \mathbb{R}$ is an unknown function that is also accessible through noisy queries. Starting from a known set $S_0 \subseteq S$, SafeOpt-type algorithm try to optimize f_* while only querying at inputs from S. This is achieved by maintaining an internal model of g_* and computing high probability uncertainty sets, which can be used to determine inputs that are safe with high probability. However, SafeOpt type algorithms are supposed to never query unsafe inputs with very high probability. This means that the uncertainty sets have to hold from the start and no tuning phase is allowed, so once again, the uncertainty sets must be derived from reasonable prior knowledge. For a more detailed discussion of these aspects, we refer to [CF8] and [CF9].

Of course, when using GP or kernel regression, frequentist uncertainty sets are available that can actually be computed, cf. Chapters 5 and 6. Similar bounds are also available for bounded noise without distributional assumptions, leading to worst-case uncertainty sets, cf. [127, 176]. However, all of these bounds assume that the target function is contained in an RKHS and need an upper bound on the RKHS norm of the target function. In the next section, we will critically investigate this assumptions.

7.1.1. The problem with the RKHS norm bound

As mentioned above, all frequentist-type uncertainty bounds for GP or kernel regression rely on the knowledge of an upper bound on the RKHS norm of the target function. More precisely, if we want to *compute* the uncertainty bounds, say in a learning-based control scheme, then we need to know a concrete bound on the RKHS norm of the target function. Note that this is a *much* stronger assumption than just membership of the target function in a known RKHS. As discussed above, for many applications in learning-based control it is important to get quantitative uncertainty bounds from reasonable prior knowledge, which in this context means that we need to get (among other ingredients) the aforementioned upper bound on the RKHS norm. Unfortunately, to the best of our knowledge, at present it is not possible to derive such a quantitative bound from established prior engineering knowledge in non-trivial situations. This is somewhat surprising given the extensive and user-friendly theory of kernel methods [189, 217], especially since the RKHS norm is in general a very well-understood object. In fact, many characterizations and explicit

representations for it are known, but they all appear to be not suited to connect it to *quantitative* prior engineering knowledge.

For an arbitrary kernel, one can use discretization-based variational characterizations of the RKHS norm (and RKHS functions), for example, by maximization over a family of lower bounds on the RKHS norm [CF6, Section B], [17, Chapter I], by minimization over certain bounds on function values at finitely many inputs [149, Theorem A.2.6], by minimization over finite interpolation problems [152, Theorem 3.11], or by minimization over certain matrix inequalities [152, Theorem 3.11]. For separable RKHSs, the RKHS norm can be expressed using a sampling expansion [113], or as the limit of norms of RKHSs over finite inputs [125, Lemma 4.6]. On the one hand, all of these variational problems have an explicit form and they work for *any* kernel (any kernel with separable RKHS, respectively). However, it is not at all clear how to relate these representations to common properties of functions that might be used as reliable prior knowledge to derive upper bounds on the RKHS norm. Furthermore, these variational problems generally cannot be used in numerical methods to estimate upper bounds on the RKHS norm, but only lower bounds, though they may be used for estimating bounds in heuristics [CF19]. Since these characterizations are based on discretizations of a given RKHS function, in particular, using the exact function values, they are not suitable in typical learning scenarios where the unknown target function is only accessible through noisy evaluations. If one considers more specific classes of kernels, other characterizations of the RKHS norm become available. For example, continuous kernels on a compact metric space equipped with a measure having full support (often called a Mercer kernel in this context) allow a description of the RKHS norm as a weighted ℓ_2-norm [189, Section 4.5], based on Mercer's theorem. This has a clear interpretation in the context of kernel methods, in particular, giving insight into the regularization behavior of the RKHS norm in optimization problems in kernel machines [89, Section 5.8], which in turn can be used to derive learning rates for various statistical learning problems [190]. More general forms of Mercer's theorem are available [191], which in turn lead to improved learning theory results [73]. While the RKHS norm representation for Mercer kernels is an important tool for statistical learning theory and provides intuition about the regularization behavior, it is again unclear how it can be used to derive *quantitative* RKHS norm bounds. Expressing the RKHS norm for Mercer kernels as a weighted ℓ_2-norm provides valuable *qualitative* intu-

ition about the corresponding RKHS norm, but we are not aware of any practically relevant example where this has been used to translate realistic prior knowledge into a concrete upper bound on the RKHS norm. Similarly, for sufficiently regular translation-invariant kernels, the RKHS norm can be expressed as a weighted integral over the Fourier transform of RKHS functions [217, Theorem 10.12]. This formulation allows an intuitive interpretation of the RKHS norm as a generalized energy, penalizing high-frequency behavior of RKHS functions (as determined by the Fourier transforms of the kernel). Several important function spaces are related to RKHSs, for example certain Sobolev spaces [217, Chapter 10] or Fock spaces [189, Section 4.4], which again have their own representations of the RKHS norm (potentially after some embedding). Again, all of these representations offer insights into the RKHS norm, and are important theoretical tools, but how this can be used to derive practically useful quantitative upper bounds on the RKHS norm remains unclear.

To summarize, while an extensive body of work on characterization and representation results for the RKHS norm is available, these results appear to be unsuited to derive numerical upper bounds on this norm using practically meaningful prior knowledge. Note that for many kernel methods and their theory this is not a problem, since most of them rely just on RKHS membership, but not the knowledge of an upper bound on the RKHS of the target function. A classic example are support vector machines (SVMs) and their theory. The algorithm itself needs no quantitative knowledge about the target function[2] to work, and the RKHS norm might appear in learning guarantees, but it just determines how good or fast the learning process goes, not whether it succeeds at all, cf. [189] for this theory.

Let us connect this back to the application in learning-based control. As discussed above, in general the uncertainty sets should be based only on reasonble and established prior knowledge. However, as the extensive discussion in this section shows, at present it appears to be not possible to derive a concrete RKHS upper norm bound from such knowledge, expressed as easily interpretable propreties of the target function. This means that at present there is an insurmountable gap between theoretically grounded kernel-based learning methods for control and their

[2]Note that this formulation is slightly imprecise. In statistical learning theory the notion of a target function does not appear implicitly, but functions attaining the Bayes risk (or the minimal risk in the corresponding hypothesis class) take on the role of the target function.

practical applicability. Note that the difficulties with the RKHS norm are occasionally acknowledge in the literature, cf. e.g. [117], but the severity of this problem is usually not discussed. Furthermore, this issue can actually lead to problems. As shown with numerical experiments in [CF8], underestimation of the RKHS norm can invalidate frequentist uncertainty sets (as is expected), and in turn this can lead to algorithmic failures, for example, safety violations in safe BO. Note that simply trying to use a conservative upper bound on the RKHS norm is in general not a viable strategy, since it is very difficult to decide in a concrete, non-trivial application what a "conservative" overstimate would be.

How to go forward from here? Trying to derive quantitative upper bounds on RKHS functions from established prior knowledge is one option and a very important research question. However, this appears to be an extremely hard problem[3] and an alternative approach seems necessary. In the next section, we propose one such direction.

7.1.2. An alternative approach: Geometric prior knowledge

Recall that in the setting relevant in this part of the thesis, we are concerned with prior knowledge about functions $f : \mathcal{X} \to \mathbb{R}$, cf. also the discussion on more general output sets in Chapter 4. One class of prior knowledge that is very broad, easily interpretable for practioners, and connected to many forms of established engineering prior knowledge, are *geometric properties of functions*. For example, if $\mathcal{X} \subseteq \mathbb{R}$, then one such property could be monotonicity of the target function, i.e., it is known that the function is nondecreasing or nonincreasing. It is clear that in many applications, this is a natural prior assumption on the target function. Similarly, we might know that the function belongs to a sector, cf. Section 6.2. If $\mathcal{X}$ is at least a metric space, then another such assumption is Lipschitz continuity with a known bound on the Lipschitz constant. This assumption has a very clear interpretation – a known bound on the rate of change of a function – and since it can be interpreted as knowledge about the sensitivity of the underlying problem, it is closely connection to established engineering notions. In particular, practitioners have a chance to judge whether a certain Lipschitz bound is reasonable in a given application, which is in stark

[3] For certain technical reasons, we suspect that it might be feasible in the context of kernel-based linear system identification [159], by connecting RKHS properties with frequency domain assumptions on the plant. However, pursuing this direction is beyond the scope of this thesis.

contrast to an RKHS norm bound. Furthermore, a considerable amount of system-theoretic knowledge, like invariance or stability properties, can be interpreted as geometric constraints, as used for example in learning dynamical systems under side information [5]. Importantly, when combined with a suitable measurement model, for example assuming bounded additive noise, then it is often easily possible to derive quantitative uncertainty bounds. One prominent example is the assumption of a Lipschitz bound combined with bounded measurement noise, for which explicit and tight uncertainty sets can be derived, which are often employed in systems and control [140], [45].

Summarizing, using geometric assumptions together with a suitable measurement model leads to explicit uncertainty sets, and the assumptions having a clear interpretation and can often be derived from prior domain knowledge. On the other hand, these approaches are often ad-hoc and they are rather disconnected from more established machine learning approaches like kernel methods. Furthermore the estimators and uncertainty sets have sometimes undesirable properties like non-differentiability, which can lead to problems for gradient-based methods [126].

This motivates us to combine the best of both worlds by using kernel methods together with geometric constraints. Results like those in Chapter 3 suggest that RKHSs with sufficiently regular kernels can be compatible with such constraints, but including such constraints in a kernel method can be nontrivial. In the following, we will present and evaluate one possible approach. In addition, in the recent work [CF8] we used a similar strategy in the context of safe Bayesian optimization, cf. Section 7.3.

7.2. Kernel regression and uncertainty bounds

In this section, we will use kernel methods on uncertainty sets derived from geometric constraints. This allows us to obtain nominal predictors with prescribed properties in a systematic manner, together with explicit uncertainty sets. Furthermore, we provide guaranteed overapproximations of the uncertainty set that have favorable properties like differentiability. We implement this strategy using the recently introduced Hard Shape Constrained Kernel Machines (HSKM) [19], which reduce the learning problems to standard convex optimization problems. To the best of our knowledge, this is the first work using kernel machines with guaranteed geometric

constraints in the context of uncertainty sets relevant for learning-based control. Furthermore, we retain the advantages of kernel methods coming with uncertainty sets, but do not need an upper bound on the RKHS norm. We illustrate the practical feasibility of the approach by means of numerical examples, including a simple control example.

7.2.1. Learning setting and goals

At the core, we consider the following regression problem. Let $f_* : \mathcal{X} \to \mathcal{Y}$ be a fixed, unknown static map, which is our target function or ground truth, where $\emptyset \neq \mathcal{X} \subseteq \mathbb{R}^d$ is open. Only for notational simplicity, we choose $\mathcal{Y} = \mathbb{R}$. Furthermore, let $\mathcal{D} = ((x_1, y_1), \ldots, (x_N, y_N))$ for $N \in \mathbb{N}_+$ be some data set satisfying $x_n \in \mathcal{X}$ and $y_n = f_*(x_n) + \epsilon_n$ for $n \in [N]$, with bounded additive noise $|\epsilon_n| \leq B_\epsilon$ for a known bound $B_\epsilon \in \mathbb{R}$. The goal is to find a nominal prediction or approximation $\hat{f} = \hat{f}_\mathcal{D}$ of f_* from the data set $\mathcal{D}$, together with an appropriate uncertainty set $\Delta_\mathcal{D} = \Delta(\ell_\mathcal{D}, u_\mathcal{D}) = \{f \in \mathcal{Y}^\mathcal{X} \mid \ell_\mathcal{D}(x) \leq f(x) \leq u_\mathcal{D}(x) \, \forall x \in \mathcal{X}\}$ with bounding functions $\ell_\mathcal{D}, u_\mathcal{D} \in \mathbb{R}^\mathcal{X}$. We take a worst-case perspective and require $f_* \in \Delta_\mathcal{D}$ for all noise realizations. Note that this learning scenario is yet another instantiation of the abstract approach outlined in Chapter 4, and it is very common in learning-based control. It appears for example when learning the dynamics for learning-based model predictive control [92] (where f_* corresponds to the unknown transition function or vector field), when approximating static nonlinearities for learning-based robust controller synthesis (where f_* is the unknown static map acting on the nominal plant) as in Section 6.2, or in safe learning-based controller tuning [31, 22] (where f_* corresponds to the unknown performance measure for given controller parameters).

Additionally, for some applications, bounding functions with certain prescribed properties like differentiability are desirable. For example, in the case of bounded Lipschitz constant, the tightest uncertainty sets are piecewise affine-linear [140], which can lead to problems for gradient based methods [126]. This leads to the problem of finding additional $\hat{\ell}_\mathcal{D}, \hat{u}_\mathcal{D} \in \mathbb{R}^\mathcal{X}$ with such desirable properties and still fulfilling $f_* \in \hat{\Delta}_\mathcal{D} = \Delta(\hat{\ell}_\mathcal{D}, \hat{u}_\mathcal{D})$. Note that simply applying a standard smoothing procedure to $\ell_\mathcal{D}, u_\mathcal{D}$ can be problematic since it might result in an invalid uncertainty set.

It is clear that some assumptions need to be imposed on the target function f_*

in order to be able to compute nontrivial uncertainty sets from the finite data set $\mathcal{D}$. Here, we use geometric constraints on f_* for this purpose, and we follow [19] and focus on constraints that can be formulated using inequalities involving linear differential operators with homogenous coefficients. To faciliate a concise exposition, we introduce some additional notation. Denote by $\mathbb{N}$ the nonnegative integers and by $\mathbb{N}_+$ the positive integers. Define for $N \in \mathbb{N}_+$ the set $[N] = \{1, \dots, N\}$. For $\boldsymbol{\alpha} \in \mathbb{N}^n$, define $|\boldsymbol{\alpha}| = \sum_{i=1}^n \alpha_i$ and for a sufficiently regular function $f \in \mathbb{R}^{\mathcal{X}}$ on some open $\mathcal{X} \subseteq \mathbb{R}^d$, define $\partial^{\boldsymbol{\alpha}} f = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} f$. Similarly, for $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ sufficiently regular, define $\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} k = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_1}}{\partial x_{1+d}^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} \frac{\partial^{\alpha_d}}{\partial x_{2d}^{\alpha_d}} k$ as well as $\partial_1^{\boldsymbol{\alpha}} k = \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}} k$ and $\partial_2^{\boldsymbol{\alpha}} k = \frac{\partial^{\alpha_1}}{\partial x_{1+d}^{\alpha_1}} \cdots \frac{\partial^{\alpha_d}}{\partial x_{2d}^{\alpha_d}} k$. Finally, $C^m(\mathcal{X}, \mathbb{R})$ denotes the set of m-times continuously differentiable functions.

Let us formalize now the geometric constraints we consider in the following. For $n \in [N_{\mathcal{C}}]$ with some $N_{\mathcal{C}} \in \mathbb{N}_+$, let

$$\mathcal{C}_n = \{f \in C^m(\mathcal{X}, \mathbb{R}) \mid L_n[f](x) \geq b_n \, \forall x \in \mathcal{K}_n\} \tag{7.1}$$

where $\mathcal{K}_n \subseteq \mathcal{X}$ is a non-empty compact set, $b_n \in \mathbb{R}$ and $L_n = \sum_{\ell=1}^{N_{L_n}} a_\ell^{(n)} \partial^{\boldsymbol{\alpha}_\ell^{(n)}}$, with $\boldsymbol{\alpha}_\ell^{(n)} \in \mathbb{N}^n$, $|\boldsymbol{\alpha}_\ell^{(n)}| \leq m$, and $a_\ell^{(n)} \in \mathbb{R}$. Note that this class includes many common geometric constraints, like slope-restrictions or convexity, cf. [19] for more examples. We continue with a simple, but instructive instance from this class.

Example 7.2.1 (Monotonicity)**.** Let $\mathcal{X} = (a, b)$, $[c, d] \subseteq (a, b)$, and $f_* \in \mathbb{R}^{\mathcal{X}}$ be a nondecreasing function that is bounded by $\ell \leq f_*(x) \leq u$ for all $x \in (a, b)$, where $\ell, u \in \mathbb{R}$. We can encode this in the form of (7.1) by setting $N_{\mathcal{C}} = 3$, $L_1 = \frac{\partial}{\partial x}$, $b_1 = 0$ and $L_2 = \mathrm{id}$, $b_2 = \ell$ and $L_3 = -\mathrm{id}$, $b_3 = -u$, and $\mathcal{K}_n = [c, d]$, $n = 1, 2, 3$.

Let us now summarize the learning problem we tackle: Given a data set $\mathcal{D}$ and prior knowledge about f_* including the constraints (7.1), find

1. a nominal prediction $\hat{f}$ that is reasonably close to f_*, fulfills all the constraints from prior knowledge and is guaranteed to be contained in a suitable uncertainty set;
2. an explicit uncertainty set $\Delta_{\mathcal{D}}$ that contains the ground truth f_*;
3. optionall an uncertainty set $\hat{\Delta}_{\mathcal{D}}$ that contains the ground truth and has desirable properties, e.g., being described by smooth bounding functions.

7.2.2. Related work

Before moving on to the proposed method, let us briefly discuss related work. In [117], the problematic nature of the RKHS norm bound assumption has been pointed out, and a GP based approach using Lipschitz continuity has been proposed to circumvent this problem. However, this approach is difficult to reconcile with robust control methods, because it relies on a probabilistic setting, cf. our discussion in Chapter 4. Uncertainty sets from geometric assumptions have been used for a while in the special case of Lipschitz bounds, for example in Nonlinear Set Membership estimation [140] and Kinky Inference [45], but these methods are very disconnected from established kernel-based approaches. Furthermore, while a variety of geometric and shape constraints are considered in statistics [86], explicit uncertainty sets based on finite data sets are usually not derived there [19]. A main aspect of the present section is the usage of realistic prior knowledge, in particular from the domain of systems and control. This leads to a natural connection to learning methods using side information, e.g. [5, 107] and the references therein. In fact, these works can be seen as complementary to our approach since the focus there is on good nominal prediction while here we focus on the containment in uncertainty sets. Closely related to our work is [126], where a scenario approach is proposed to get sufficiently regular predictors under bounded noise and a known Lipschitz constant. While this seems to be the first work making an explicit connection between kernel methods and geometric constraints in the context of uncertainty sets, only high-probability guarantees can be given owing to the scenario approach. Furthermore, the approach is tailored to the case of hard Lipschitz bounds. While enforcing constraints at finitely many inputs in kernel regression is easily possible with the Representer Theorem [180], enforcing geometric constraints is considerably harder since such constraints usually involve infinitely many inputs. Finally, a variety of constrained kernel methods are available, see [4] for a typical example, but these methods usually rely on relaxing the constraints, which invalidates any uncertainty sets and subsequent guarantees building on them. The latter aspects motivates us to use hard shape constrained kernel machines.

7.2.3. Hard shape constrained kernel machines

Consider a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. It is well-known (cf. e.g. [189, Corollary 4.36]) that if k has continuous partial derivatives $\partial^{\boldsymbol{\alpha},\boldsymbol{\alpha}} k$ for all $|\boldsymbol{\alpha}| \leq m$ for some $m \in \mathbb{N}_+$, then for all $f \in H_k$, we have $f \in C^m(\mathcal{X}, \mathbb{R})$ and for all $x \in \mathcal{X}$ and $|\boldsymbol{\alpha}| \leq m$, we have $\partial^{\boldsymbol{\alpha}} f(x) = \langle f, \partial^{\boldsymbol{\alpha}} \Phi_k(x) \rangle$, where the right-hand side is well-defined. From now on, we assume that k is sufficiently regular, which is only a mild assumption for many popular kernels.

We opt for optimization-based kernel methods and want to enforce the geometric constraints in the optimization problem, leading to the generic problem

$$\begin{aligned} &\min_{f \in H_k} \mathcal{L}(f(x_1), \ldots, f(x_M)) + \lambda \|f\|_k^2 \\ &\text{s.t. } f \in \mathcal{C}_n \; \forall n \in [N_{\mathcal{C}}] \end{aligned} \tag{7.2}$$

where $x_1, \ldots, x_M \in \mathcal{X}$ for some $M \in \mathbb{N}$ and $\mathcal{L} : \mathbb{R}^M \to \mathbb{R}$ is strictly convex and $\lambda \in \mathbb{R}_{>0}$ is a regularization parameter For example, in Section 7.2.4 we have $M = N$ and $\mathcal{L}$ will be a data-fit term. In general, this is an infinite-dimensional optimization problem, and in order to solve it without discarding or relaxing the geometric constraints, we use the tightening approach from [19]: For each $n \in [N_{\mathcal{C}}]$, let $\delta_n \in \mathbb{R}_{>0}$ be given and choose $x_m^{(n)} \in \mathcal{K}_n$ for $m \in [M_n]$ with some $M_n \in \mathbb{N}$, such that $(x_1^{(n)}, \ldots, x_{M_n}^{(n)})$ forms a δ_n-cover of $\mathcal{K}_n$, and choose $\eta_n \in \mathbb{R}_{>0}$ such that

$$\eta_n \geq \max_{m \in [M_n]} \sup_{\substack{x + x_m^{(n)} \in \mathcal{K}_n \\ \|x\| \leq \delta_n}} \|L_n[\Phi_k(x_m^{(n)})] - L_n[\Phi_k(x_m^{(n)} + x)]\|_k. \tag{7.3}$$

We can then formulate the new problem

$$\begin{aligned} &\min_{f \in H_k} \mathcal{L}(f(x_1), \ldots, f(x_M)) + \lambda \|f\|_k^2 \\ &\text{s.t. } L_n[f](x_m^{(n)}) \geq b_n + \eta_n \|L_n[f]\|_k \; \forall n \in [N_{\mathcal{C}}], \; m \in [M_n] \end{aligned} \tag{7.4}$$

and summarize its main properties in the next result, essentially a simplified variant of [19, Theorem 1].

Theorem 7.2.2. Every feasible f for problem (7.4) is also feasible for (7.2), i.e., the former is a tightening of the latter. Furthermore, every solution $\hat{f}$ of problem

(7.4) can be written in the form

$$\hat{f}(x) = \sum_{m=1}^{M} c_m k(x, x_m) + \sum_{(\boldsymbol{\alpha},\bar{x})\in\mathcal{G}} c_{(\boldsymbol{\alpha},\bar{x})} \partial_2^{\boldsymbol{\alpha}} k(x, \bar{x}), \tag{7.5}$$

where

$$\begin{aligned} \mathcal{G} \subseteq &\{\boldsymbol{\alpha}_\ell^{(n)} \mid n \in [N_{\mathcal{C}}],\, \ell \in [N_{L_n}]\} \\ &\times \{x_m^{(n)} \mid n \in [N_{\mathcal{C}}],\, m \in [M_n]\} \end{aligned} \tag{7.6}$$

and $c_m, c_{(\boldsymbol{\alpha},\bar{x})} \in \mathbb{R}$ for $m \in [M]$, $(\boldsymbol{\alpha}, \bar{x}) \in \mathcal{G}$.

For the proof and an explicit description of (7.5) and (7.6), see the supplementary material of [19]. The representation (7.5) is a variant of the Representer Theorem [180] and implies that problem (7.4) is equivalent to a finite-dimensional convex second order cone (SOC) optimization problem. The latter optimization problem can be explicitly formulated via standard arguments, cf. [19] for details. Note that if $\hat{f} \in H_k$ is feasible for (7.4), then $\hat{f} \in \mathcal{C}_n$ for all $n \in [N_{\mathcal{C}}]$, i.e., all geometric constraints are fulfilled.

Remark 7.2.3. The method from [19] supports the simultanous approximation of multiple functions which are are coupled by convex constraints. Furthermore, using a generalization, cf. [18], it is also possible to work with vector RKHS functions and semidefinite constraints. For conciseness, we do not use these more advanced capabilities in the present work, but we would like to stress that all of the following developments and results apply to these more sophisticated variants. Furthermore, using standard arguments, cf. [180, 222], additional constraints on the RKHS function values or derivatives at finitely many inputs can be enforced, which we will make use in the next section.

7.2.4. Geometrically constrained kernel regression with uncertainty sets

We now develop a geometrically constrained kernel regression approach to tackle the learning problem outlined in Section 7.2.1, considering each of the three subproblems in turn.

Nominal prediction

In order to find a nominal prediction $\hat{f}$ fulfilling all constraints from prior knowledge (in particular, the constraints (7.1)), while being contained in a suitable uncertainty set, we use a kernel optimization approach and enforce the geometric constraints from prior knowledge and the noise model in a HSKM. Note that this includes the implicit assumption that the RKHS used in the kernel machine includes a good approximation of the target function, which is a common and reasonable assumption [189]. First, we encode the prior knowledge as constraints of the form (7.1). This is demonstrated for a few illustrative examples below, with many more examples in [19, 18]. Next, in order to achieve a good nominal approximation of the target function, we use $\mathcal{L}(t_1, \ldots, t_N) = \sum_{n=1}^{N}(t_n - y_n)^2$, i.e., the sum of squares criterion as a data fit term. If the empirical mean of the noise terms is close to zero, we expect a good approximation of the target function for sufficiently large data sets.

Remark 7.2.4. Many different objectives are possible. For example, one can use sum of absolute values, which could add additional robustness to outliers. Furthermore, one can remove the data term completely, resulting in a Support Vector Regression-type problem, cf. e.g. [126]. Finally, adding an ℓ_1-penalty on the coefficients leads to sparsity, which becomes relevant when the resulting function has to be evaluated efficiently.

Finally, we apply the tightening procedure and the Representer Theorem described in Theorem 7.2.2. For convenience, we summarize this in the following result.

Theorem 7.2.5. Let $f_* \in \mathcal{C}_1 \cap \ldots \cap \mathcal{C}_{N_\mathcal{C}}$, $x_1, \ldots, x_N \in \mathcal{X}$ and, for $n \in [N]$, assume that $y_n = f_*(x_n) + \epsilon_n$, where $|\epsilon_n| \leq B_\epsilon$ for some known $B_\epsilon \in \mathbb{R}_{\geq 0}$. Consider the optimization problem

$$\begin{aligned} &\min_{f \in H_k} \sum_{n=1}^{N} (f(x_n) - y_n)^2 + \lambda \|f\|_k^2 \\ &\text{s.t.} |f(x_n) - y_n| \leq B_\epsilon \; \forall n \in [N] \\ &\quad L_n[f](x_m^{(n)}) \geq b_n + \eta_n \|L_n[f]\|_k \; \forall n \in [N_\mathcal{C}],\; m \in [M_n] \end{aligned} \tag{7.7}$$

Any feasible function for (7.7) fulfills all geometric constraints (7.1) and is contained in every valid uncertainty set that can be derived from the geometric constraints and

the noise assumption. Furthermore, every solution $\hat{f}$ of problem (7.4) can be written in the form (7.5).

Proof. The first and third claim are direct consequences of Theorem 7.2.2 and Remark 7.2.3. The second claim follows since such an uncertainty includes any function fulfilling all geometric constraints and is compatible with the data set and noise model, which is the case for any feasible function. □

The second claim implies that (7.7) is equivalent to a finite-dimensional convex SOC problem and, compared to the situation in Theorem 7.2.2, we only added $2N$ linear inequality constraints.

Uncertainty sets

In order to obtain explicit uncertainty sets, we combine the geometric constraints and bounded noise assumptions. In general, nontrivial uncertainty sets require a manual construction, taking the specific geometric constraints into account. We demonstrate this now with a straightforward, but instructive example.

Example 7.2.6 (continued)**.** Given a data set $\mathcal{D} = ((x_n, y_n))_{n\in[N]}$ as described in Section 7.2.1, elementary calculations show that the tightest uncertainty set in this situation is given by

$$\ell_{\mathcal{D}}(x) = \begin{cases} \ell & x \in [c, x_1) \\ \max\{\ell, y_1 - B_\epsilon, \dots, y_n - B_\epsilon\} & x \in [x_n, x_{n+1}) \\ \max\{\ell, y_1 - B_\epsilon, \dots, y_N - B_\epsilon\} & x \in [x_N, d] \end{cases}$$

$$u_{\mathcal{D}}(x) = \begin{cases} \min\{u, y_1 + B_\epsilon, \dots, y_N + B_\epsilon\} & x \in [c, x_1) \\ \min\{u, y_{n+1} + B_\epsilon, \dots, y_N + B_\epsilon\} & x \in [x_n, x_{n+1}) \\ u & x \in [x_N, d] \end{cases}$$

where we assumed without loss of generality that $c < x_1 < \ldots < x_N < d$ and $\ell < y_1 - B_\epsilon$, $y_N + B_\epsilon < u$.

Example 7.2.7 (Slope restriction)**.** Let $a \leq c < 0 < d \leq b$, $\underline{s} < 0 < \bar{s}$ and $f_* : (a, b) \to \mathbb{R}$ with $f_*(0) = 0$ and $f'_*(x) \geq \underline{s}$, $f'_*(x) \leq \bar{s}$ for all $x \in [c, d]$. Such functions appear frequently in control as slope-restricted nonlinearities. Given a data

set $\mathcal{D} = ((x_n, y_n))_{n\in[N]}$ as described in Section 7.2.1, this leads to an uncertainty set described by $u_\mathcal{D}(x) = \min\{b_n(x) \mid n = 0, \dots, N\}$, where we defined for $n \in [N]$ $b_n(x) = y_n + B_\epsilon + \bar{s}(x - x_n)$ if $x \geq x_n$, $b_n(x) = y_n + B_\epsilon + \underline{s}(x - x_n)$ otherwise, and $b_0(x) = \bar{s}x$ if $x \geq 0$ and $b_0(x) = \underline{s}x$ otherwise. Furthermore, $\ell_\mathcal{D}$ is given by an analogous construction. For an illustration on a concrete example, see Figure 7.2. These uncertainty sets are direct generalizations of the well-known tightest uncertainty sets from Lipschitz approaches like [140].

Remark 7.2.8. Note that, in general, bounding functions of an uncertainty set cannot be directly computed using geometrically constrained kernel regression. The latter searches over an RKHS which contains functions that are usually much smoother (in terms of regularity as well as complexity) than the class of all functions fulfilling the geometric constraints. While this is a desirable feature for nominal prediction [130, 126], it poses an obstacle for uncertainty sets that have to contain all possible function candidates fulfilling the geometric constraints and being compatible with the data.

Uncertainty sets with prescribed properties

Recall that for some applications an uncertainty set $\hat{\Delta}_\mathcal{D}$ described by bounding functions $\hat{\ell}_\mathcal{D}, \hat{u}_\mathcal{D}$ with prescribed properties, like differentiability, is desirable. To solve this task, we propose to start with a valid uncertainty $\Delta_\mathcal{D}$ and find an overapproximation $\hat{\Delta}_\mathcal{D} \supseteq \Delta_\mathcal{D}$ by using a HSKM to compute suitable bounding functions $\hat{\ell}_\mathcal{D}, \hat{u}_\mathcal{D}$. In particular, this leads to a valid uncertainty set since $\hat{\Delta}_\mathcal{D} \supseteq \Delta_\mathcal{D} \ni f_*$. The strategy is to minimize a suitable measure of overapproximation, subject to the constraint that the resulting functions still lead to an overapproximation of the given uncertainty set, as well as any additional constraints. For concreteness, consider the case of finding a smooth upper bounding function $\hat{u}_\mathcal{D}$, the case of $\hat{\ell}_\mathcal{D}$ being analogous. One option to implement this strategy is by solving

$$\begin{aligned} &\min_{u\in H_k} \sum_{m=1}^{M} (u(\hat{x}_m) - u_\mathcal{D}(\hat{x}_m))^2 + \lambda\|u\|_k^2 \\ &\text{s.t.} u(x) \geq u_\mathcal{D}(x)\, \forall x \in \mathcal{K} \end{aligned} \tag{7.8}$$

where $\hat{x}_1, \ldots, \hat{x}_M \in \mathcal{K}$ is a sufficiently fine grid and $\mathcal{K} \subseteq \mathcal{X}$ a compact subset of the input set, on which we want an overapproximation of the uncertainty set. One can now use the procedure outlined in Theorem 7.2.2. In particular, if the kernel k is sufficiently smooth, then the solutions $\hat{\ell}_{\mathcal{D}}, \hat{u}_{\mathcal{D}}$ of the resulting finite-dimensional convex problem are smooth bounding functions that are guaranteed to lead to a valid uncertainty set $\hat{\Delta}_{\mathcal{D}}$.

7.2.5. Examples

Illustrative examples

We now illustrate the methodology using Examples 7.2.1 and 7.2.7. For simplicity, we use the popular Squared Exponential kernel with length scale $\ell = 0.5$ and unit variance, albeit other kernel choices can readily be applied. For the tightening and determination of the relevant constants, we use the gridding approach from [19]. The resulting optimization problems are solved using MOSEK [13] and cvxpy [62]. Computation times were less than 2 min on a standard laptop.

Example 7.2.9 (continued)**.** As a concrete ground truth, choose $f_* : [0, 2] \to \mathbb{R}$, $f_*(x) = x + 0.3 \sin(2\pi x)$. We sample $N = 30$ inputs x_n uniformly from $[0, 2]$ and set $y_n = f_*(x_n) + \epsilon_n$, where ϵ_n are sampled independently and uniformly from $[-0.2, 0.2]$. The tightening in Theorem 7.2.2 is performed using an equidistant grid of 100 points from $[0, 2]$, following the procedure described in [19]. As an illustration of Remark 7.2.4, we removed the data fitting term from the objective. Furthermore, the uncertainty set described in Example 7.2.6 is computed and overapproximated using the approach outlined in Section 7.2.4. The result is shown in Figure 7.1. It can be seen that the uncertainty set is rather tight and, as indicated by Theorem 7.2.5, the nominal prediction is contained in this uncertainty set (as well as the ground truth). Furthermore, by the choice of the kernel, the overapproximation is smooth, but at the expense of conservatism.

Example 7.2.10 (continued)**.** Consider as a concrete example the piecewise affine-linear function $f_* : [-1.5, 1.5] \to \mathbb{R}$ defined as $f_*(x) = -I_{x<0} \cdot 0.5x + I_{x \geq 0.5} \cdot (x - 0.5)$, where I is the usual indicator function. Using $\underline{s} = -1$, $\bar{s} = 2$, as well as $N = 20$ and $B_\epsilon = 0.1$, leads to the results shown in Figure 7.2. Despite the nonsmoothness of

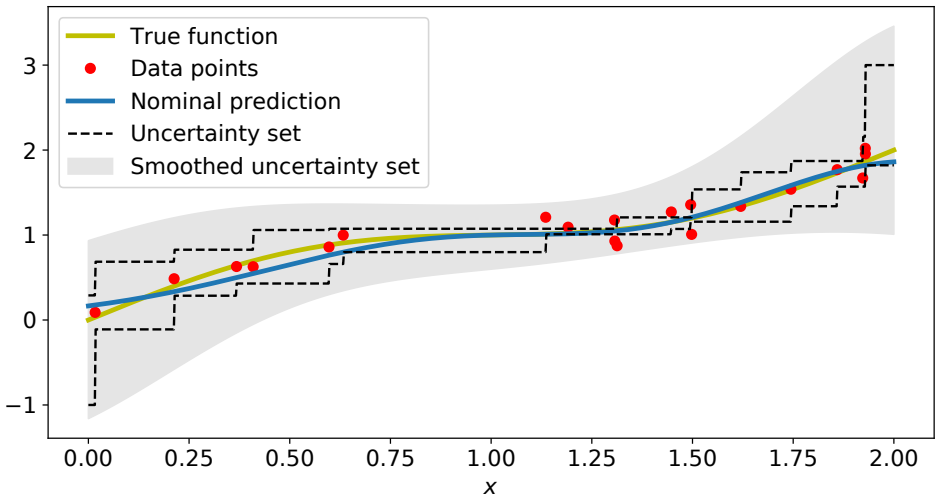

Figure 7.1.: Illustrating geometrically constrained regression on the example of a nondecreasing function. © 2022 IEEE

f_*, the nominal prediction is mostly close to it and the smoothed uncertainty sets are rather tight.

Control example

We now apply the methodology to an illustrative control example adapted from [117]. In particular, we use the nominal prediction from the proposed learning method for feedback linearization control for tracking a reference trajectory, while the uncertainty estimates are used for computing ultimate bounds for the tracking error.

Consider the 2d system $\dot{x}_1 = x_2$, $\dot{x}_2 = f(x) + u$, a reference input x_{ref} and define $x_{\mathrm{d}} = \begin{pmatrix} x_{\mathrm{ref}} & \dot{x}_{\mathrm{ref}} \end{pmatrix}$ and $e = x - x_{\mathrm{d}}$. Given a model $\hat{f}$, use the controller $u = -\hat{f}(x) + \ddot{x}_{\mathrm{ref}} - k_c r - \lambda e_2$ with parameters k_c, λ and filtered state $\dot{r} = f(x) - \hat{f}(x) - k_c r$. For concreteness [117, Section 5.1], let $f(x) = 1 - \sin(x_1) + 1/(1 + \exp(-x_2)) = 1 + f_1(x_1) + f_2(x_2)$ and $k_c = 5$, $\lambda = 1$. We take advantage of the additive structure of f and learn f_1, f_2 separately, using two data sets of the form $y_{i,n} = f_i(x_{i,n}) + \epsilon_{i,n}$, $i = 1, 2$, with $n = 1, \ldots, 50$, uniform additive noise with $|\epsilon_{i,n}| \leq B_\epsilon = 0.05$ and assume as prior knowledge a Lipschitz bound of 2 for both f_1 and f_2. We use the method from Section 7.2.4 to get $\hat{f}_1$ and $\hat{f}_2$, leading to the prediction model $\hat{f}(x) =$

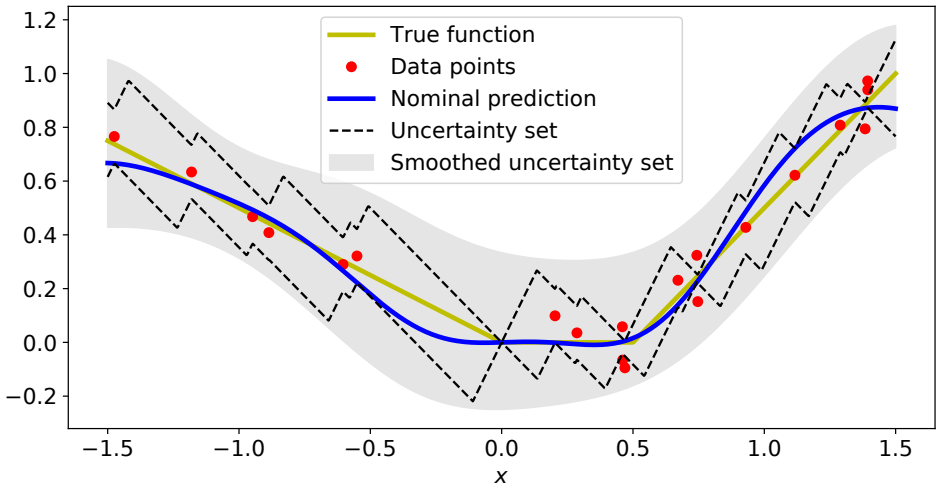

Figure 7.2.: Geometrically constrained kernel regression in the case of a Lipschitz continuous function. © 2022 IEEE

$1 + \hat{f}_1(x_1) + \hat{f}_2(x_2)$. Similar to Example 7.2.7, we also get an uncertainty set with bounding functions $\ell_{\mathcal{D}}, u_{\mathcal{D}}$ and smoothed bounding functions $\hat{\ell}_{\mathcal{D}}, \hat{u}_{\mathcal{D}}$ as described in Section 7.2.4. Figure 7.3 (top) shows the resulting closed loop trajectory for $x_{\mathrm{ref}}(t) = 2\sin(t)$, using $\hat{f}$ in the controller, as well as the local size of the uncertainty set $w_{\mathcal{D}}(x) = u_{\mathcal{D}}(x) - \ell_{\mathcal{D}}(x)$.

It has been shown in [117, Section 4] that e converges to a (time-varying) ball with radius $r_{\mathcal{D}}(x) = w_{\mathcal{D}}(x(t))/k_c\sqrt{\lambda^2 + 1}$, in particular, $r_{\mathcal{D}}$ is an ultimate bound for $\|e\|$. This bound can in turn be used for safety guarantees, cf. the discussion in [117]. Figure 7.3 (bottom) shows the norm of e, the radius $r_{\mathcal{D}}$ as well as the smooth overapproximation $\hat{r}_{\mathcal{D}}(x) = \hat{w}_{\mathcal{D}}(x)/k_c\sqrt{\lambda^2 + 1}$ with $\hat{w}_{\mathcal{D}}(x) = \hat{u}_{\mathcal{D}}(x) - \hat{\ell}_{\mathcal{D}}(x)$. As expected, $\hat{r}_{\mathcal{D}}$ changes more smoothly than $r_{\mathcal{D}}$ at the expense of conservatism, due to the smoother but more conservative uncertainty set.

Note that we did not need any RKHS norm estimate in this example. Furthermore, in contrast to [117], all guarantees hold in a worst-case sense and not only probabilistically.

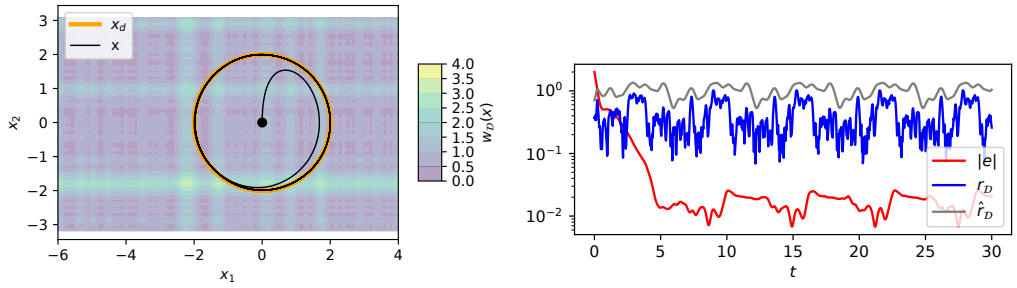

Figure 7.3.: Top: Reference and closed-loop trajectory using controller with nominal prediction model. Background color indicates local size of uncertainty set. Bottom: Tracking error norm $\|e\|$ and asymptotic bounds $r_{\mathcal{D}}$, $\hat{r}_{\mathcal{D}}$ over time. Note that after a short transient phase, the tracking error stays well below the asymptotic bound from [117, Section 4]. © 2021 IEEE

7.2.6. Discussion

We presented a versatile kernel based approach for the problem of learning functions together with uncertainty sets from noisy data. Relying only on geometric constraints and a bounded noise model, this leads to learned functions with guaranteed properties and explicit uncertainty sets. Since a multitude of prior engineering knowledge can be expressed as geometric constraints, this paves the way to practical and reliable uncertainty sets for learning-based control.

At the moment, three relevant limitations of our approach can be identified: First, HSKM rely on a gridding approach and hence suffer from the curse of dimensionality in the input space dimension, which is a limitation inherited by our methods. Second, the feasibility of the convex optimization problems is not guaranteed, though in practice this does not seem to pose a problem. In particular, if the problems are feasible, the geometric constraints are guaranteed to be fulfilled. Finally, the derivation of uncertainty sets from geometric assumptions currently has to be done manually on a case-by-case basis.

The methodology outlined here can be easily generalized and extended in many directions: Target functions with multiple outputs can be handled by approximating all coordinate functions separately, potentially coupled by convex SOC constraints [19] or, alternatively, approximated using a function from a vector RKHS [18]. Since the approach in Theorem 7.2.5 is also compatible with the generalizations developed

in [18], semidefinite constraints can be handled analogously. This is particularly relevant for learning-based control, since many relevant properties of dynamical systems can be expressed using Linear Matrix Inequalities, cf. e.g. [179]. Finally, the methodology outlined here is compatible with any constrained kernel regression method that supports geometric constraints, and Theorem 7.2.5 applies mutatis mutandis to any such method. As already hinted at in [19], linear integral functional constraints can also be used.

Ongoing work is concerned with more complex application scenarios, stochastic noise models, learning vector-valued functions, and combining the presented methodology with other modeling approaches.

7.3. Conclusion

In Section 7.1, we argued that uncertainty bounds in learning-based control should be only based on assumptions that can be clearly linked to reasonable prior knowledge accepted by practitioners. Furthermore, we argued that at present a concrete numerical upper bound on the RKHS norm does not fulfill this requirement, and we proposed to switch towards geometric properties of the ground truth as alternative assumptions, and combine this with kernel methods. In Section 7.2, we proposed one concrete framework for this approach, which showed promising initial results and many interesting avenues for future research in this area remain. However, as already pointed out there, automating the derivation of uncertainty sets from geometric assumptions, and broading the class of suitable measurement models, in particular, including also stochastic noise, is an important open problem. Furthermore, in the recent work [CF8] we used a similar approach in the context of safe BO, and by relying only on a Lipschitz bound for ensuring safety, we achieved excellent performance while guaranteeing safety using only reasonable and easily interpretable quantitative prior knowledge. Since this work is beyond the scope of this thesis, we refer to [CF8] for more details, and [CF9] for a concise overview. In summary, while interesting open problems and plenty of opportunities for extensions remain open, combining geometric assumptions for uncertainty sets with kernel methods appears as a very promising approach for learning-based control with rigorous guarantees, that can translate into practice.

7.4. Comments

Section 7.1 is based on [CF11] and [CF8], with some parts taken verbatim from these references. Section 7.2 is based on and to a large extent taken verbatim from the work [CF11]. The corresponding methods were developed, implemented and evaluated by the author of this thesis, who also wrote the article with editorial input from C.W. Scherer and S. Trimpe. The line of work in [CF8] was initiated by S. Trimpe, the main approach (LoSBO) was developed by the author of this thesis together with L. Kreisköther, who also performed all the initial experiments. The additional algorithm LoS-GP-UCB reported in [CF8] arose through discussions with J. Menn, P. Brunzema and A. von Rohr. All the experiments in [CF8] were conducted by J. Menn with support by the author of this thesis, who also wrote most of the manuscript, with editorial input from J. Menn and S. Trimpe., and some of the plots were prepared by the student assistant S. Azirar. The extended abstract [CF9] was written mostly by the author of this thesis, with editorial input by J. Menn.

Part III.

Kernels and the mean field limit

8. Introduction

We now turn to our second main contribution, investigating kernels in the context of mean field limits. The main concept in the next chapters are *mean field limits of functions*, which we motivate and formalize in detail in Section 8.2. To provide some context, in Section 8.3 we give a self-contained introduction to (formal) mean field limits as commonly used in kinetic theory, and link this to mean field limits of functions. Finally, in Section 8.4 we explain how all of this is connected to kernel methods, and motivate the developments in the upcoming chapters. For the reader's convenience, we also include some technical background on kernel mean embeddings in Section 8.5, since this will be used frequently later on.

This chapter is partially based on, and some parts have been taken verbatim from, [CF6] and [CF5]. Detailed comments on the author's contribution and relation to existing work are provided in Section 8.6.

8.1. Multiagent systems, kinetic theory and mean field limits

Models with many variables play an important role in many fields of mathematical and physical sciences. In this context, going to the limit of infinitely many variables is an important analysis and modeling approach. Multiagent systems (MAS), or synonymously, interacting particle systems (IPS) is a rich and thriving field at the intersection of systems and control, applied mathematics, computer science, and physics. This area has started in the statistical mechanics of many-particle systems, in particular, gas dynamics [52, 151]. In past decades, the field has expanded its investigation to many complex systems, both natural and engineered. Applications include animal movement (inter alia swarms of birds, schools of fish, colonies of microorganisms) [21, 104], social and political dynamics [200, 51], crowd modeling and control (pedestrian movement, gathering at large events like football games or

concerts) [68, 57, 6], swarms of robots [158, 148, 53] or vehicular traffic (in particular, traffic jams) [201]. There is now a vast literature on such applications, and we refer to the surveys [146, 212, 23, 24, 82] as starting points. Typical questions concern the long-term behavior of such systems, in particular, emergent phenomena like consensus or alignment [146].

Recently, machine learning has played an increasing role also in this area. While first-principles modeling has been very successful for interacting particle systems in physical domains, using this approach to model the interaction rules in complex domains like social and opinion dynamics, pedestrian and animal movement or vehicular traffic, can be problematic. Therefore, learning interaction rules from data has been recently intensively investigated, for example, in the pioneering works [35, 124]. The data consists typically of (sampled) trajectories of the particle states, potentially with measurement noise, and the goal is to learn a good approximation of the interaction rule ϕ.

In many relevant applications, the number of agents or particles is very large. For example, even small volumes of gases typically contain an enormous number of molecules, a microscopic modeling approach quickly becomes infeasible. In particular, simulation and control on the microscopic level, modelling every individual particle or agent, is not possible anymore in such applications. One way to deal with this difficult is to go from the microscopic level to the *mesoscopic* level, and considering only the *distribution* of the agents or particles, instead of every individual one. This forms one of the major subjects of *kinetic theory*. Several approaches for this transition are available, both formal and rigorous. For example, from a continuous-time microsopic model, one can derive a Boltzmann-type equation, and in an appropriate scaling-limit, one ends up on the mesoscopic level [151]. Alternatively, one can go directly to the mesoscopic level with the *mean field limit*. In Section 8.3, we provide a gentle introduction to the (formal) mean field limit in a typical setting of kinetic theory. However, our starting points are *mean field limits of functions* (or rather of a sequence of functions), which we motivate and formalize in detail in Section 8.2. It turns out that this has a direct connection to kernel methods, cf. Section 8.4.

8.2. Mean field limit of functions

Some notation and terminology For $M \in \mathbb{N}_+$, we denote by $\mathcal{S}_M$ the set of permutations of $\{1, \dots, M\}$. For some set $\mathcal{X}$, we will frequently use the notation $\vec{x} \in \mathcal{X}^M$, $M \in \mathbb{N}_+$, for tuples of elements of $\mathcal{X}$, and the notation x_m to refer to the m-th element of $\vec{x}$, $1 \leq m \leq M$. Furthermore, given $\vec{x} \in \mathcal{X}^M$ and $\sigma \in \mathcal{S}_M$, we write for brevity

$$\sigma\vec{x} = \begin{pmatrix} x_{\sigma(1)} & \cdots & x_{\sigma(M)} \end{pmatrix}. \tag{8.1}$$

For $x \in \mathcal{X}$, we denote by δ_x the Dirac probability measure with atom on x. Note that technically a measure is only defined on a measurable space, but since

$$\delta_x(A) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

holds for any $A \subseteq \mathcal{X}$, not specifying the underlying σ-algebra does not cause any problems. Furthermore, note that in earlier chapters of this thesis, δ_x has been used to denote the evaluation fucntional over a given function space, but since this notation is so established, and no confusion can arise, we deemed this ambiguity acceptable. For $\vec{x} \in \mathcal{X}^M$, we write

$$\hat{\mu}[\vec{x}] = \frac{1}{M} \sum_{i=1}^{M} \delta_{x_i} \tag{8.2}$$

for the *empirical measure with atoms* $x_1, \dots, x_M$. Note that the order of the atoms does not matter, so $\hat{\mu}[\sigma\vec{x}] = \hat{\mu}[\vec{x}]$ for all $\sigma \in \mathcal{S}_M$. Furthermore, we write $\mathcal{E}_M(\mathcal{X})$ for the set of empirical measures with M atoms (potentially with repetitions), and

$$\mathcal{E} = \bigcup_{M \in \mathbb{N}_+} \mathcal{E}_M(\mathcal{X}) \tag{8.3}$$

for the set of all empirical measure with a finite number of atoms.

Let $\mathcal{Y} \neq \emptyset$ be some additional set. We call a function $f : \mathcal{X}^M \to \mathcal{Y}$ *permutation-invariant* if for all $x_1, \dots, x_M \in \mathcal{X}$ and $\sigma \in \mathcal{S}_M$ we have

$$f(x_{\sigma(1)}, \dots, x_{\sigma(M)}) = f(x_1, \dots, x_M),$$

or more short $f(\sigma\vec{x}) = f(\vec{x})$. A permutation invariant function is sometimes also called *symmetric*.

Finally, there is a close connection between permutation invariant functions, and functions on empirical measures. More precisely, a function $F_M : \mathcal{E}_M(\mathcal{X}) \to \mathcal{Y}$ on empirical measures with M atoms is in a unique correspondence with a function $f_M : \mathcal{X}^M \to \mathcal{Y}$ that is permutation invariant. Given F_M, we can define f_M by setting

$$f_M(x_1, \ldots, x_M) = F_M(\hat{\mu}[(x_1, \ldots, x_M)]),$$

and we find for all $\vec{x} \in \mathcal{X}^M$ and $\sigma \in \mathcal{S}_M$ that

$$f_M(\sigma\vec{x}) = F_M(\hat{\mu}[\sigma\vec{x}]) = F_M(\hat{\mu}[\vec{x}]) = f_M(\vec{x}),$$

so f_M is indeed permutation invariant. Given f_M, we can define F_M by

$$F_M\left(\frac{1}{M}\sum_{i=1}^{M}\delta_{x_i}\right) = f_M(x_1, \ldots, x_M),$$

and this is well-defined since if $\vec{x}, \vec{x}' \in \mathcal{X}^M$ with $\hat{\mu}[\vec{x}] = \hat{\mu}[\vec{x}']$, we have $\vec{x} = \sigma\vec{x}'$ for some $\sigma \in \mathcal{S}_M$, so we get

$$f_M(\vec{x}) = f_M(\sigma\vec{x}') = f_M(\vec{x}').$$

We can therefore work interchangably with functions on empirical measures and permutation invariant functions.

Functions of empirical measures Consider now the following situation. We have collections of entities of the same type, and we are interested in a property of such collectives. Furthermore, we assume that the entities are indistinguishable, so if we have a collection of M entities, it is best modelled as an empirical measure with the M entities as atoms. At least for finitely many entities, the property of interest can then be modelled as scalar functions on empirical measures. Arbitrary collections of entities can be modelled by probability distributions, though the intuition is not as transparent anymore. In the context of multiagent systems, this situation occurs when we are interested in a certain *state-dependent functional or feature of the system.*

Formally, we have a set $\mathcal{X} \neq \emptyset$, which could be for example the state space of an individual agent of a multiagent system. Furthermore, we have maps $F_M : \mathcal{E}_M(\mathcal{X}) \to \mathbb{R}$, $M \in \mathbb{N}_+$, and given $\vec{x} \in \mathcal{X}^M$, we interpret $F_M(\hat{\mu}[\vec{x}])$ as the value of the property of interest for the collection of entities modelled by $\hat{\mu}[\vec{x}]$. Roughly speaking, this means that we can compute the property of interest for a finite number of entities. But what if we want to compute the property for arbitrary collective of entities? Ideally, we "extend" the family of maps $(F_M)_M$ from empirical measures to a map on all (or at least sufficiently regular) probability distributions over $\mathcal{X}$.

One way to do this is the *mean field limit of a sequence of functions*. Before moving on to this concept, we need some more background on probability distributions.

Technical background on probability measures Unless noted otherwise, from now on (X, d_X) denotes a compact metric space. Let $\mathcal{P}(X)$ be the set of Borel probability measures on X. We consider the usual weak convergence[1] of probability distributions, cf. e.g. [109, Chapter 13], so $(\mu_n)_n \subseteq \mathcal{P}(X)$ converges to $\mu \in \mathcal{P}(X)$ iff for all bounded and continuous $\phi : X \to \mathbb{R}$ we have

$$\lim_{n\to\infty} \int_X \phi(x) \mathrm{d}\mu_n(x) = \int_X \phi(x) \mathrm{d}\mu(x).$$

Of course, since X is compact, this is equivalent to requiring the convergence for all continuous ϕ. It is well-known, cf. [64, Section 11.8], that this convergence in $\mathcal{P}(X)$ can be metrized by the Kantorowich-Rubinstein distance d_{KR}, defined by

$$d_{\mathrm{KR}}(\mu_1, \mu_2) = \sup \left\{ \int_X \phi(x) \mathrm{d}(\mu_1 - \mu_2)(x) \mid \phi : X \to \mathbb{R} \text{ is 1-Lipschitz} \right\},$$

and $\mathcal{P}(X)$ is compact under this metric. Furthermore, the empirical measures are dense in $\mathcal{P}(X)$ under the given metric.

Defining the mean field limit of functions Let us return to the problem from above and restrict us to a setting with convenient analytical tools available. Instead of an arbitrary set $\mathcal{X}$, we consider the compact metric space (X, d_X), and suppose that for all $M \in \mathbb{N}_+$, we have $F_M : \mathcal{E}_M(X) \to \mathbb{R}$ given, and the goal is to find

[1] Note that in the literature, in particular in the context of optimal transport, it is also called *narrow convergence*, cf. e.g. [12, Chapter 6].

some $F : \mathcal{P}(X) \to \mathbb{R}$ that in an appropriate sense extends the F_M. Recalling our interpretation from above, using the F_M we can compute some property of interest for all empirical measures with a finite number of atoms, and F will allow us to compute this property for all Borel probability measures $\mu \in \mathcal{P}(X)$.

How should such a F look like? Since $\mathcal{E}_M(X) \subseteq \mathcal{P}(X)$, a natural requirement of an extension would be $F|_{\mathcal{E}_M(X)} = F_M$ for all $M \in \mathbb{N}_+$, however, we do *not* require this, since the F_M might not be consistent with each other. Instead, we start from the *denseness* of $\mathcal{E}(X)$ in $\mathcal{P}(X)$ w.r.t. d_{KR}. This means that for all $\mu \in \mathcal{P}(X)$ and $\epsilon > 0$, there exists some $\vec{x}_\epsilon \in X^{M_\epsilon}$ for some $M_\epsilon \in \mathbb{N}_+$, such that $d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_\epsilon], \mu) \leq \epsilon$. In general, a smaller ϵ requires a larger M_ϵ, and conversely, empirical measures in $\mathcal{E}_M(X)$ should become increasingly better approximators with larger M. But this means intuitively that for a large M, F_M is almost capable of computing the property of interest for any Borel probability measure, since $\mathcal{E}_M(X)$ can approximate arbitrary measures already quite well, and hence F_M and F should be somewhat close, i.e. $F_M(\hat{\mu}[\vec{x}]) \approx F(\hat{\mu}[\vec{x}])$. We can formalize this by requiring that

$$\sup_{\hat{\mu}[\vec{x}] \in \mathcal{E}_M(X)} |F_M(\hat{\mu}[\vec{x}]) - F(\hat{\mu}[\vec{x}])| \to 0 \quad M \to \infty. \tag{8.4}$$

This is almost the definition of the mean field limit of a sequence of functions, as found in the literature. The only difference is that it is customarily defined for permutation invariant functions.

To make this connection, recall that we can identify the functions $F_M : \mathcal{E}_M(X) \to \mathbb{R}$ with permutation invariant functions $f_M : X^M \to \mathbb{R}$, so (8.4) becomes

$$\sup_{\vec{x} \in X^M} |f_M(\vec{x}) - F(\hat{\mu}[\vec{x}])| \to 0 \quad M \to \infty.$$

If the preceding holds, we call F the *mean field limit* of $(f_M)_M$. This notion has been introduced by P.-L. Lions in the context of mean field games [49], and it is by now a very established notion in kinetic theory, where it appears for example in mean field (optimal) control [91, 74]. For convenience, we record this concept in the following formal definition.

Definition 8.2.1. Let $\mathcal{X}$ be a measurable space, and let $\mathcal{P}$ be a set of probability distributions on $\mathcal{X}$ that contains all empirical probability measures with finitely

many atoms. Consider a sequence of functions $f_M : \mathcal{X}^M \to \mathbb{R}$, $M \in \mathbb{N}_+$, and a function $f : \mathcal{P} \to \mathbb{R}$. If

$$\lim_{M\to\infty} \sup_{\vec{x}\in\mathcal{X}^M} |f_M(\vec{x}) - f(\hat{\mu}[\vec{x}])| = 0, \tag{8.5}$$

then we say that $(f_M)_M$ converges to f in mean field, or that f is the mean field limit of the f_M, and we write $f_M \xrightarrow{\mathcal{P}_1} F$.

A classic existence result In most situations, we have the functions f_M, $M \in \mathbb{N}_+$, and would like to get the corresponding mean field limit, so that we can deal with arbitrary (Borel) probability measures. In other words, what about the existence of mean field limits of functions?

For convenience, we recall a by now classic result, cf. [49, Theorem 2.1]. To formulate it, we define a *modulus of continuity* as a function $\omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that is continuous, non decreasing and with $\omega(0) = 0$. Later we use that for every $R \in \mathbb{R}_{>0}$ and every modulus of continuity ω, we can find a concave modulus of continuity $\tilde{\omega} : [0, R] \to \mathbb{R}_{\geq 0}$ such that $\omega(r) \leq \tilde{\omega}(r)$ for all $r \in [0, R]$.

Proposition 8.2.2. Assume the following about the $(f_M)_M$.

1. *(Symmetry in $\vec{x}$)* For all $M \in \mathbb{N}_+$, $\vec{x} \in X^M$ and permutations $\sigma \in \mathcal{S}_M$, we have

$$f_M(\sigma\vec{x}) := f(x_{\sigma(1)}, \ldots, x_{\sigma(M)}) = f(\vec{x})$$

2. *(Uniform boundedness)* There exists $C_f \in \mathbb{R}_{\geq 0}$ such that

$$\forall M \in \mathbb{N}_+, \vec{x} \in X^M : |f_M(\vec{x})| \leq C_f$$

3. *(Uniform continuity)* There exists a modulus of continuity $\omega_f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that for all $M \in \mathbb{N}_+$, $\vec{x}_1, \vec{x}_2 \in X^M$

$$|f_M(\vec{x}_1) - f_M(\vec{x}_2)| \leq \omega_f\left(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2])\right)$$

Then there exists a subsequence $(f_{M_\ell})_\ell$ and some $f : \mathcal{P}(X) \to \mathbb{R}$, such that

$$\lim_{\ell\to\infty} \sup_{\vec{x}\in X^{M_\ell}} |f_{M_\ell}(\vec{x}) - f(\hat{\mu}[\vec{x}])| = 0.$$

Furthermore, f is continuous as function on $\mathcal{P}(X)$ and (uniformly) bounded by C_f.

Note that in the assumptions of this result, the symmetry in $\vec{x}$ for f_M is actually implied by the uniform continuity, cf. [50, Remark 1.3]. Furthermore, this latter property also implies continuity with respect to the product metric on X^M.

8.3. Interlude: But where is the mean field?

While standard in the literature, the terminology *mean field* limit might appear unclear at this point. To provide intuition and more context, we include in the following a gentle and high level outline of a typical mean field limit argument in the context of multiagent systems. Our presentation is folklore, and follows expositions like [34, Chapter 4] and [81].

To make things more concrete, we consider first-order dynamics in continuous time, which includes alignment dynamics. A system with $M \in \mathbb{N}_+$ agents, where agent $i = 1, \dots, M$ has state $x_i(t) \in \mathbb{R}^d$ at time $t \geq 0$, can be described by

$$\dot{x}_i = \frac{1}{M} \sum_{j=1}^{M} \Psi(x_i, x_j)(x_j - x_i) \tag{8.6}$$

$$x_i(0) = x_i^0 \qquad i = 1, \dots, M \tag{8.7}$$

where $\Psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ models the interaction strength between two agents, and is often of the form $\Psi(x, x') = \psi(\|x - x'\|)$. We would like to lift the dynamics to the mesoscopic level, so if $f(t, x)$ is the density of agents at time $t \geq 0$ at state $x \in \mathbb{R}^d$, we want to describe the evolution of f over time. Here is the basic idea. Instead of describing the evolution of the density f directly, we describe the evolution of the average of some feature $\phi : \mathbb{R}^d \to \mathbb{R}$ of the agents, i.e.,

$$t \mapsto \int \phi(x) f(t, x) \mathrm{d}x.$$

For technical reasons, we now work with measures (the density f induces a measure $A \mapsto \int_A f(t, x)\mathrm{d}x$ for $A \subseteq \mathbb{R}^d$ measurable). Our goal is now as follows. Let $\mu(t)$ be the distribution ("density") of agents at time t, then we want a model for the evolution $t \mapsto \mu(t)$ starting with initial distribution ("density") $\mu(0) = \mu_0$. But as already remarked earlier, a (microscopic) state $\vec{x} = (x_1, \dots, x_M) \in (\mathbb{R}^d)^M$ corresponds to

the *empirical measure* $\hat{\mu}[\vec{x}]$. The average of some feature $\phi : \mathbb{R}^d \to \mathbb{R}$ of the agents in microscopic state $\vec{x}$ can hence be expressed as the integral

$$\int \phi(x) \mathrm{d}\hat{\mu}[\vec{x}](x) = \frac{1}{M} \sum_{m=1}^{M} \phi(x_m).$$

Consider now $\phi \in C_c^\infty(\mathbb{R}^d, \mathbb{R})$ (smooth functions with compact support, our "features" of interest) and let us derive an ODE for $\int \phi(x) \mathrm{d}\hat{\mu}[\vec{x}](x)$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \int \phi(x) \mathrm{d}\hat{\mu}[\vec{x}(t)](x) &= \frac{\mathrm{d}}{\mathrm{d}t} \frac{1}{M} \sum_{i=1}^{M} \phi(x_i(t)) \\
&= \frac{1}{M} \sum_{i=1}^{M} \frac{\mathrm{d}}{\mathrm{d}t} \phi(x_i(t)) \\
&= \frac{1}{M} \sum_{i=1}^{M} \nabla_x \phi(x_i(t))^\top \dot{x}_i(t) \\
&= \frac{1}{M} \sum_{i=1}^{M} \frac{1}{M} \sum_{j=1}^{M} \Psi(x_i, x_j) \nabla_x \phi(x_i(t))^\top (x_j(t) - x_i(t)) \\
&= \frac{1}{M} \sum_{i=1}^{M} \nabla_x \phi(x_i(t))^\top \left(\frac{1}{M} \sum_{j=1}^{M} \Psi(x_i, x_j)(x_j(t) - x_i(t)) \right) \\
&= \int \nabla_x \phi(x)^\top \underbrace{\left(\int \Psi(x, y)(y - x) \mathrm{d}\hat{\mu}[\vec{x}(t)](y) \right)}_{= F(x, \hat{\mu}[\vec{x}(t)])} \mathrm{d}\hat{\mu}[\vec{x}(t)](x).
\end{aligned}$$

We have at the moment

$$\frac{\mathrm{d}}{\mathrm{d}t} \int \phi(x) \mathrm{d}\hat{\mu}[\vec{x}(t)](x) = \int \nabla_x \phi(x_i(t))^\top F(x, \hat{\mu}[\vec{x}(t)]) \mathrm{d}\hat{\mu}[\vec{x}(t)](x).$$

Now *formally* apply partial integration (recall that ϕ has compact support) to get

$$\begin{aligned}
&\int \nabla_x \phi(x_i(t))^\top F(x, \hat{\mu}[\vec{x}(t)]) \mathrm{d}\hat{\mu}[\vec{x}(t)](x) \\
&\quad = \underbrace{[\phi(x_i(t))^\top F(x, \hat{\mu}[\vec{x}(t)])]}_{=0} + (-1)^d \int \phi(x_i(t))^\top \nabla_x F(x, \hat{\mu}[\vec{x}(t)]) \mathrm{d}\hat{\mu}[\vec{x}(t)](x),
\end{aligned}$$

so we arrive at

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \phi(x)\mathrm{d}\hat{\mu}[\vec{x}(t)](x) = (-1)^d \int \phi(x_i(t))^\top \nabla_x F(x, \hat{\mu}[\vec{x}(t)])\mathrm{d}\hat{\mu}[\vec{x}(t)](x).$$

This holds for all $M \in \mathbb{N}_+$, so we can *formally* consider $M \to \infty$, the *formal mean field limit*. Our previous derivations suggest the following. Let μ_0 be the initial distribution of agents, then $\mu(t)$ is the distribution of agents at time $t \geq 0$ according to the alignment dynamics if

1. $\mu(0) = \mu_0$

2. For all $\phi \in C_c^\infty(\mathbb{R}^d, \mathbb{R})$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \phi(x)\mathrm{d}\mu(t)(x) = \int \nabla_x \phi(x)^\top F(x, \mu(t))\mathrm{d}\mu(t)(x). \tag{8.8}$$

   where

$$F(x, \mu) = \int \Psi(x, y)(y - x)\mathrm{d}\mu(y) \tag{8.9}$$

Equation (8.9) explains the term *mean field*: $F(x, \mu)$ is the average "force" exerted by agents distributed according to μ and felt by an agent at state x. In other words, it is the *mean* of the *field* acting on a representative agent at a given state. Furthermore, if the interaction function is translation-invariant, i.e., $\Psi(x, y) = \psi(x - y)$ for some function $\psi : \mathbb{R}^d \to \mathbb{R}$, then

$$F(x, \mu) = \int \psi(x - y)(x - y)\mathrm{d}\mu(y) = \int K(x - y)\mathrm{d}\mu(y) = (K * \mu)(x),$$

i.e., $F(x, \mu)$ is the convolution of the function $K(z) = \psi(z)z$ with the measure μ. In many works on mean field dynamics, one can find this formulation, cf. e.g. [34, Chapter 4].

For additional insight, consider now the case $d = 1$. Applying *formally* partial integration to

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \phi(x)\mathrm{d}\mu(t)(x) = \int \nabla_x \phi(x) \cdot F(x, \mu(t))\mathrm{d}\mu(t)(x) = \int \frac{\partial}{\partial x}\phi(x) \cdot F(x, \mu(t))\mathrm{d}\mu(t)(x)$$

leads to (recall that ϕ has compact support)

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \phi(x)\mathrm{d}\mu(t)(x) = \underbrace{[\phi(x)F(x,\mu(t))]_{-\infty}^{\infty}}_{=0} - \int \phi(x)\frac{\partial}{\partial x}F(x,\mu(t))\mathrm{d}\mu(t)(x).$$

Assume now that there exists a suitable f with $\mu(t)(A) = \int_A f(t,x)\mathrm{d}x$, so $\mu(t)$ has the *density* $f(t,\cdot)$. We then get

$$\frac{\mathrm{d}}{\mathrm{d}t}\int \phi(x)f(t,x)\mathrm{d}x + \int \phi(x)\frac{\partial}{\partial x}F(x,f(t,\cdot))f(t,x)\mathrm{d}x = 0$$

with

$$F(x,f) = \int \Psi(x,y)(x-y)f(x)\mathrm{d}y,$$

motivating the notation

$$\partial_t f(t,x) + \partial_x \int \Psi(x,y)(x-y)f(t,y)\mathrm{d}y = 0,$$

a *kinetic PDE in strong form.*

To summarize, we have lifted microscopic dynamics to mesoscopic dynamics by interpreting a finite collection of agents as an empirical measure, and then formally replaced the empirical measure by an arbitrary measure. The latter can be interpreted as approximating an arbitrary probability measure by empirical measures with an increasining number of atoms, so that the given probability measure is the *limit* of the empirical measures. Furthermore, in the dynamics on measures the *mean* of the *field* acting on the agents appears, which suggests the terminology *mean field limit*. Finally, the preceding arguments can be made rigorous, cf. e.g. [81].

Let us connect all of this back to Section 8.2. From a physical perspective, Definition 8.2.1 is not a traditional mean field limit, however, it shares the motivation arising from approximating arbitrary probability measures by empirical measures. A more fitting terminology for Definition 8.2.1 might *distributional extension* or *asymptotic distributional extension* (as discussed in Section 8.2, the mean field limit does not have to be a proper extension), but the former terminology is standard in the literature. Furthermore, even in the context of traditional mean field limits of ODE dynamics, the need for the mean field limit of functions arises, for example, in mean field optimal control [74]. In the next section, another such application will

appear.

8.4. Kernels enter the picture

In the present part of the thesis, we will explore two applications of kernels and kernel methods to kinetic theory.

Kernels in the mean field limit Frequently, the state of such a complex multiagent system can be easily measured or estimated, e.g., by video recordings or image snapshots for bird swarms or schools of fish, and microscopy recordings for microorganism colonies; aerial imaging for human crowds (e.g., via quadcopters); and polling and social media analysis for opinion dynamics. However, some interesting features of the whole system might be more difficult to measure. For example, how a swarm of birds or a school of fish will react to an external stimulus (like an approaching predator), given the current state of the population. Such a reaction could be a change of density or spread of the population, or a change in mean velocity. Another example is given by features of a society in opinion dynamics (average happiness, aggression potential, susceptibility to adversarial interventions), given the current "opinion state". Measuring such features can be difficult, for example, due to a required intervention. Formally, if $\mathcal{X}$ is the state space of an individual agent (say, a metric space or just $\mathbb{R}^d$), such a feature is a functional $F_M : \mathcal{X}^M \to \mathbb{R}$ of the current state of the system, and since the state is often easy to measure, it would be useful to have an explicit mapping from state to feature of interest. However, since first principles modeling is unlikely to be successful in the domains considered here, it is promising to learn such a mapping from data. We can formalize this as a standard supervised learning task: The data set consists of $D_N^{[M]} = ((\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N))$, where $\vec{x}_n \in \mathcal{X}^M$ are snapshot measurements of the particle states (corresponding to the input of the functional) and $y_n \in \mathbb{R}$ is the value of the functional of interest, potentially with measurement noise, at snapshot state $\vec{x}_n$. Let us assume an additive noise model, i.e., $y_n = F_M(\vec{x}_n) + \epsilon_n$ for $n = 1, \dots, N$, where $\epsilon_1, \dots, \epsilon_N \in \mathbb{R}$ are noise variables. This is a regression problem that could be solved for example using a Support Vector Machine (SVM) [189]: Let $k_M : \mathcal{X}^M \times \mathcal{X}^M \to \mathbb{R}$ be a kernel on $\mathcal{X}^M$ with associated RKHS H_M, and $\ell_M : \mathcal{X}^M \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function,

then the resulting approximation of the functional F_M is given by

$$F_{\ell_M, D_N^{[M]}, \lambda} = \operatorname{argmin}_{f \in H_M} \frac{1}{N} \sum_{n=1}^{N} \ell_M(\vec{x}_n, y_n, f(\vec{x}_n)) + \lambda \|f\|_M^2, \tag{8.10}$$

where $\lambda \in \mathbb{R}_{>0}$ is the regularization parameter and $\|\cdot\|_M$ the RKHS norm.

Suppose now we are interested in systems with a very large number of agents or interacting particles. As outlined above, in this context it is reasonable to go to the mesoscopic level, so instead of trajectories of particle states of the form $[0,T] \ni t \mapsto \vec{x}(t) \in \mathcal{X}^M$, we then have trajectories of probability measures $[0,T] \ni t \mapsto \mu(t) \in \mathcal{P}(\mathcal{X})$. This immediately raises the question of whether the learning setup outlined above also allows a corresponding kinetic limit. More precisely, let $K \subseteq \mathcal{X}$ be compact and assume that all particles remain confined to this compactum, i.e., $x_i(t) \in K$ for all $i = 1, \dots, M$ and all $t \in [0,T]$ under the microscopic dynamics. If the underlying dynamics have a mean field limit, then it is reasonable to assume that the finite-input functionals $F_M : K^M \to \mathbb{R}$ converge also in mean field to some $F : \mathcal{P}(K) \to \mathbb{R}$ for $M \to \infty$. In turn, we can now formulate a corresponding learning problem on the mean field level: A data set is then given by $D_N = ((\mu_1, y_1), \dots, (\mu_N, y_N))$, where $\mu_n \in \mathcal{P}(K)$ are snapshots of the particle state distribution over time and $y_n \in \mathbb{R}$ are again potentially noisy measurements of the functional. Assuming an additive noise model, this corresponds to $y_n = F(\mu_n) + \epsilon_n$, $n = 1, \dots, N$. If we want to use an SVM on the kinetic level, we need a kernel $k : \mathcal{P}(K) \times \mathcal{P}(K) \to \mathbb{R}$ on probability distributions. There are several options available for this, see e.g. [55]. However, assuming that all ingredients of the learning problem arise as a mean field limit, this naturally leads to the question of whether a mean field limit of kernels exists, and what this means for the relation of the learning problems on the finite-input and kinetic level. This motivates us to investigate kernels and their RKHSs in the mean field in Chapter 9, and their application in the context of statistical learning theory in a mean field limit setup in Chapter 10.

Discrete-time mean field limits In Chapter 11, we consider discrete-time multiagent systems and their mean field limit. As explained in detail in the introduction there, trying to adapt existing tools from mean field theory to this task leads to technical problems. Surprisingly, kernels offer an elegant solution, leading to an ex-

istence result for the mean field limit of discrete-time multiagent systems, the first such results to the best of our knowledge.

8.5. Technical background: Kernel mean embeddings

In Chapter 9 and 11 we will need kernel mean embeddings, so we collect some background on this concept here. Let $k : X \times X \to \mathbb{R}$ be a Borel-measurable kernel. Furthermore, assume that it is bounded, i.e., $\sup_{x,x' \in X} |k(x, x')| < \infty$. In this case, $X \ni x \mapsto k(\cdot, x) \in H_k$ is Bochner-integrable w.r.t. every $\mu \in \mathcal{P}(X)$, and we define

$$\Pi_k : \mathcal{P}(X) \to H_k, \quad \mu \mapsto \int_X k(\cdot, x) \mathrm{d}\mu. \tag{8.11}$$

For $\mu \in \mathcal{P}(X)$, we call $\Pi_k(\mu)$ the kernel mean embedding (KME) of μ in H_k. This terminology is explained by the fact that in the present setting, for all $\mu \in \mathcal{P}(X)$, any $f \in H_K$ is μ-integrable, and $\int_X f(x) \mathrm{d}\mu = \langle f, \Pi_k(\mu) \rangle_k$. If the map Π_k is injective, then we call k *characteristic*, cf. [187] for many examples and conditions for this property. To simplify notation, define additionally $\hat{\Pi}_k(\vec{x}) = \Pi_k(\hat{\mu}[\vec{x}])$ for $\vec{x} \in X^M$. For convenience, we record the following simple fact.

Lemma 8.5.1. The set $\Pi_k\left(\mathcal{P}(X)\right)$ is convex.

Proof. Let $g, h \in \Pi_k\left(\mathcal{P}(X)\right)$ and $\lambda \in (0, 1)$ be arbitrary. By definition there exist $\mu, \nu \in \mathcal{P}(X)$ with $g = \Pi_k(\mu)$ and $h = \Pi_k(\nu)$. We then have

$$\begin{aligned}
\lambda g + (1-\lambda)h &= \lambda \Pi_k(\mu) + (1-\lambda)\Pi_k(\nu) \\
&= \lambda \int_X k(\cdot, x) \mathrm{d}\mu(x) + (1-\lambda) \int_X k(\cdot, x) \mathrm{d}\nu(x) \\
&= \int_X k(\cdot, x) \mathrm{d}(\lambda\mu + (1-\lambda)\nu) \\
&= \Pi_k(\lambda\mu + (1-\lambda)\nu) \in \Pi_k\left(\mathcal{P}(X)\right),
\end{aligned}$$

where we used the fact that $\lambda\mu + (1-\lambda)\nu \in \mathcal{P}(X)$. □

Finally, define the maximum mean discrepancy (MMD) w.r.t. k by

$$\gamma_k : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}_{\geq 0}, \quad \gamma_k(\mu, \nu) = \|\Pi_k(\mu) - \Pi_k(\nu)\|_k.$$

γ_k is a semimetric on $\mathcal{P}(X)$, and if k is characteristic, γ_k is a metric.

Lemma 8.5.2. If (X, d_k) is a compact metric space, then $\Pi_k(\mathcal{P}(X))$ is compact.

Proof. Recall that $(\mathcal{P}(X), d_{\mathrm{KR}})$ is compact, where d_{KR} is defined with $d_X = d_k$. Next, since (X, d_k) is compact, it is separable, and hence [188, Theorem 21] implies that for all $\mu, \nu \in \mathcal{P}(X)$

$$\gamma_k(\mu, \nu) = \|\Pi_k(\mu) - \Pi_k(\nu)\|_k \leq d_{\mathrm{KR}}(\mu, \nu),$$

which shows that $\Pi_k : (\mathcal{P}, d_{\mathrm{KR}}) \to (H_k, \|\cdot\|_k)$ is 1-Lipschitz continuous. Altogether, $\Pi_k(\mathcal{P}(X))$ is the image of a compact set under a continuous map, so it is compact. □ □

For additional background on KMEs and MMD, see [188].

8.6. Comments

Section 8.1 is based on and partially taken verbatim from [CF6]. Section 8.2 has been written from scratch for this thesis, and the author is not aware of a similar exposition carefully describing the intuition behind the definition of the mean field limit of functions, though the ideas are of course implicit in existing presentations like [50]. Section 8.3 is based on a talk given by the author of this thesis at RWTH Aachen University in 2023. Section 8.4 is based on and partially has been taken verbatim from [CF6], and Section 8.5 has been taken from [CF5].

9. Kernels and their RKHSs in the mean field limit

In the preceding chapter, we presented the notion of a mean field limit of a sequence of functions, and motivated that this concept can appear also in the context of kernels and kernel methods. In this chapter, we now start the investigation of kernels and their RKHSs in the mean field limit. First, we define the appropriate notion of convergence and prove an existence result of the mean field limit of kernels in Section 9.1. Large classes of examples of appropriate kernels are provided in Section 9.2, where we also discuss an interesting connection to kernel mean embeddings. Finally, in Section 9.3 we investigate the RKHSs of kernels in the mean field limit. Unless noted otherwise, in this chapter we use the setting that was introduced in the preceding chapter.

This chapter is based on, and in parts taken verbatim, from the articles [CF3, CF6, CF4]. Detailed comments on the author's contribution and the relation of this chapter to existing work are provided in Section 9.5.

9.1. The mean field limit of kernels

The definition of mean field convergence of a sequence of functions can be easily generalized to bivariate functions, or more precisely, functions with two blocks of arguments that are growing in size.

Definition 9.1.1. Let $\mathcal{X}$ be a measurable space, and let $\mathcal{P}$ be a set of probability distributions on $\mathcal{X}$ that contains all empirical probability measures with finitely many atoms. Consider bivariate functions $\kappa_M : \mathcal{X}^M \times \mathcal{X}^M \to \mathbb{R}$, $M \in \mathbb{N}_+$, and $\kappa : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$. We say that κ is the mean field limit of $(\kappa_M)_M$, or that $(\kappa_M)_M$

converges in mean field to κ, if

$$\lim_{M\to\infty} \sup_{\vec{x},\vec{x}'\in\mathcal{X}^M} |\kappa_M(\vec{x},\vec{x}') - \kappa(\hat{\mu}[\vec{x}],\hat{\mu}[\vec{x}'])| = 0.$$

Since a kernel on X^M is just a bivariate function, the preceding definition immediately applies to sequence of kernels with an increasing number of inputs. This immediately raises several questions. Does a mean field limit exist in this case? If κ_M are kernels and they converge in mean field to κ, is the limit also a kernel? And if they are all kernels, what happens to their RKHSs? In this section, we start with the first two questions.

Recall that X is a compact metric space and d_{KR} is the Kantorowich-Rubinstein metric on $\mathcal{P}(X)$, the set of Borel probability measures on X. Define $d^2_{\mathrm{KR}} : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}_{\geq 0}$ by

$$d^2_{\mathrm{KR}}((\mu_1,\mu_1'),(\mu_2,\mu_2')) = d_{\mathrm{KR}}(\mu_1,\mu_2) + d_{\mathrm{KR}}(\mu_1',\mu_2'),$$

and note that $(\mathcal{P}(X)\times\mathcal{P}(X), d^2_{\mathrm{KR}})$ is a compact metric space. Consider now a sequence

$$k^{[M]} : X^M \times X^M \to \mathbb{R},\ M \in \mathbb{N}_+,$$

of kernels on input space X^M.

Assumption 9.1.2. 1. *(Symmetry in $\vec{x}$)* For all $M \in \mathbb{N}_+$, $\vec{x},\vec{x}' \in X^M$ and permutations $\sigma \in \mathcal{S}_M$ we have

$$k^{[M]}(\sigma\vec{x},\vec{x}') := k^{[M]}((x_{\sigma(1)},\ldots,x_{\sigma(M)}),\vec{x}') = k^{[M]}(\vec{x},\vec{x}')$$

2. *(Uniform boundedness)* There exists $C_k \in \mathbb{R}_{\geq 0}$ such that

$$\forall M \in \mathbb{N}_+, \vec{x},\vec{x}' \in X^M : |k^{[M]}(\vec{x},\vec{x}')| \leq C_k$$

3. *(Uniform continuity)* There exists a modulus of continuity $\omega_k : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that for all $M \in \mathbb{N}_+$, $\vec{x}_1,\vec{x}_1',\vec{x}_2,\vec{x}_2' \in X^M$

$$|k^{[M]}(\vec{x}_1,\vec{x}_1') - k^{[M]}(\vec{x}_2,\vec{x}_2')| \leq \omega_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_1],\hat{\mu}[\vec{x}_1']),(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}_2'])\right]\right)$$

For later reference, we record two immediate consequences of the preceding assumption.

Proposition 9.1.3. Consider the situation of Assumption 9.1.2. For $M \in \mathbb{N}_+$, let $(\mathcal{H}_M, \Phi_M)$ be *any* feature space-feature map pair for $k^{[M]}$.

1. For all $M \in \mathbb{N}_+$, Φ_M is invariant under permutations, i.e., for all $\vec{x} \in X^M$ and $\sigma \in \mathcal{S}_M$ we have $\Phi_M(\sigma\vec{x}) = \Phi_M(\vec{x})$.

2. For all $M \in \mathbb{N}_+$ and $\vec{x} \in X^M$ we have $\|\Phi_M(\vec{x})\|_{\mathcal{H}_M} \leq \sqrt{C_k}$.

3. $\sqrt{2\omega_k}$ is a modulus of continuity for Φ_M for all $M \in \mathbb{N}_+$, i.e., for all $\vec{x}_1, \vec{x}_2 \in X^M$ we have

$$\|\Phi_M(\vec{x}_1) - \Phi_M(\vec{x}_2)\|_{\mathcal{H}_M} \leq \sqrt{2\omega_k\left(d_{\mathrm{KR}}\left[\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2]\right]\right)}$$

Proof. 1. Let $M \in \mathbb{N}_+$, $\vec{x} \in X^M$ and $\sigma \in \mathcal{S}_M$ be arbitrary. From

$$\begin{aligned}\|\Phi_M(\sigma\vec{x}) - \Phi_M(\vec{x})\|^2_{\mathcal{H}_M} &= \langle \Phi_M(\sigma\vec{x}, \Phi_M(\sigma\vec{x})\rangle_{\mathcal{H}_M} - 2\langle \Phi_M(\sigma\vec{x}), \Phi_M(\vec{x})\rangle_{\mathcal{H}_M} \\ &\quad + \langle \Phi_M(\vec{x}), \Phi_M(\vec{x})\rangle_{\mathcal{H}_M} \\ &= k^{[M]}(\sigma\vec{x}, \sigma\vec{x}) - 2k^{[M]}(\sigma\vec{x}, \vec{x}) + k^{[M]}(\vec{x}, \vec{x}) \\ &= k^{[M]}(\vec{x}, \vec{x}) - 2k^{[M]}(\vec{x}, \vec{x}) + k^{[M]}(\vec{x}, \vec{x}) \\ &= 0\end{aligned}$$

(where we used the symmetry and permutation invariance of $k^{[M]}$) we find that $\Phi_M(\sigma\vec{x}) = \Phi_M(\vec{x})$, hence the permutation invariance of all Φ_M.

2. Let $M \in \mathbb{N}_+$ and $\vec{x} \in X^M$ be arbitrary, then

$$\|\Phi_M(\vec{x})\|_{\mathcal{H}_M} = \sqrt{\langle \Phi_M(\vec{x}), \Phi_M(\vec{x})\rangle_{\mathcal{H}_M}} = \sqrt{k^{[M]}(\vec{x}, \vec{x})} \leq \sqrt{C_k}.$$

3. Let $M \in \mathbb{N}_+$ and $\vec{x}_1, \vec{x}_2 \in X^M$ be arbitrary, then

$$\begin{aligned}\|\Phi_M(\vec{x}_1) - \Phi_M(\vec{x}_2)\|^2_{\mathcal{H}_M} &= \langle \Phi_M(\vec{x}_1), \Phi_M(\vec{x}_1)\rangle_{\mathcal{H}_M} - \langle \Phi_M(\vec{x}_2), \Phi_M(\vec{x}_1)\rangle_{\mathcal{H}_M} \\ &\quad - \langle \Phi_M(\vec{x}_1), \Phi_M(\vec{x}_2)\rangle_{\mathcal{H}_M} + \langle \Phi_M(\vec{x}_2), \Phi_M(\vec{x}_2)\rangle_{\mathcal{H}_M} \\ &= k^{[M]}(\vec{x}_1, \vec{x}_1) - k^{[M]}(\vec{x}_2, \vec{x}_1) - k^{[M]}(\vec{x}_1, \vec{x}_2) + k^{[M]}(\vec{x}_2, \vec{x}_2) \\ &\leq |k^{[M]}(\vec{x}_1, \vec{x}_1) - k^{[M]}(\vec{x}_2, \vec{x}_1)| + |k^{[M]}(\vec{x}_1, \vec{x}_2) - k^{[M]}(\vec{x}_2, \vec{x}_2)| \\ &\leq 2\omega_k(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2])),\end{aligned}$$

hence $\|\Phi_M(\vec{x}_1) - \Phi_M(\vec{x}_2)\|_{\mathcal{H}_M} \leq \sqrt{2\omega_k(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2]))}$.

□

Next, we turn to properties of functions f from the RKHSs of $k^{[M]}$.

Proposition 9.1.4. Consider the situation of Assumption 9.1.2. Let $M \in \mathbb{N}_+$, denote for simplicity $H_M = H_{k^{[M]}}$, and let $f \in H_M$ be arbitrary.

1. For all $\vec{x} \in X^M$ and $\sigma \in \mathcal{S}_M$ we have

$$f(\sigma\vec{x}) = f(\vec{x}).$$

2. For all $\vec{x} \in X^M$ we get

$$|f(\vec{x})| \leq \|f\|_{H_M}\sqrt{C_k}.$$

3. Let $\vec{x}_1, \vec{x}_2 \in X^M$ be arbitrary, then

$$|f(\vec{x}_1) - f(\vec{x}_2)| \leq \sqrt{2\omega_k(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2]))}.$$

The arguments used in the proof are standard, but for completeness we provide all details.

Proof. Using the reproducing property and symmetry of $k^{[M]}$, we find for $\vec{x} \in X^M$ and $\sigma \in \mathcal{S}_M$

$$f(\sigma\vec{x}) = \langle f, k^{[M]}(\sigma\vec{x}, \cdot)\rangle_{H_M} = \langle f, k^{[M]}(\vec{x}, \cdot)\rangle_{H_M} = f(\vec{x}),$$

establishing the first claim. Next, using again the reproducing property of $k^{[M]}$, Cauchy-Schwarz and the boundedness of $k^{[M]}$ we get

$$\begin{aligned}|f(\vec{x})| &= |\langle f, k^{[M]}(\vec{x}, \cdot)\rangle_{H_M}| \\ &\leq \|f\|_{H_M} \|k^{[M]}(\vec{x}, \cdot)\|_{H_M} \\ &= \|f\|_{H_M} \sqrt{k^{[M]}(\vec{x}, \vec{x})} \\ &\leq \|f\|_{H_M} \sqrt{C_k},\end{aligned}$$

showing the second statement. Similarly, for $\vec{x}_1, \vec{x}_2 \in X^M$ we get

$$\begin{aligned}|f(\vec{x}_1) - f(\vec{x}_2)| &= |\langle f, k^{[M]}(\vec{x}_1, \cdot) - k^{[M]}(\vec{x}_2, \cdot)\rangle_{H_M}| \leq \|f\|_{H_M} \|k^{[M]}(\vec{x}_1, \cdot) - k^{[M]}(\vec{x}_2, \cdot)\|_{H_M} \\ &= \|f\|_{H_M} \sqrt{k^{[M]}(\vec{x}_1, \vec{x}_1) - k^{[M]}(\vec{x}_1, \vec{x}_2) + k^{[M]}(\vec{x}_2, \vec{x}_2) - k^{[M]}(\vec{x}_2, \vec{x}_1)} \\ &\leq \|f\|_{H_M} \sqrt{|k^{[M]}(\vec{x}_1, \vec{x}_1) - k^{[M]}(\vec{x}_1, \vec{x}_2)| + |k^{[M]}(\vec{x}_2, \vec{x}_2) - k^{[M]}(\vec{x}_2, \vec{x}_1)|} \\ &\leq \|f\|_{H_M} \sqrt{2\omega_k(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2]))}.\end{aligned}$$

□

We now turn to the central existence result of the mean field limit of kernels. The next theorem extends the proof in [49, Theorem 2.1] and shows that, if a sequence of kernels fulfills Assumption 9.1.2, then a mean field limit exists, which is again a kernel.

Theorem 9.1.5. Under Assumption 9.1.2, there exists a subsequence $(k^{[M_\ell]})_\ell$ and a continuous, bounded kernel $k : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$ such that

$$\lim_{\ell \to \infty} \sup_{\vec{x}, \vec{x}' \in X^{M_\ell}} |k^{[M_\ell]}(\vec{x}, \vec{x}') - k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])| = 0. \tag{9.1}$$

The first part of the proof is based on the same arguments as in [49, Theorem 2.1] and repeated only for convenience.

Proof. We construct a sequence of uniformly bounded and equi–continuous kernels $k^{[M]}_{\mathrm{McK}}$ for $M \in \mathbb{N}_+$. Its limit will be the desired kernel k.

Step 1. In the first step we define $k^{[M]}_{\mathrm{McK}}$ and show that it is bounded on $\mathcal{P}(X) \times \mathcal{P}(X)$ and coincides with the kernel $k^{[M]}$ on $X^M \times X^M$. Since $\mathcal{P}(X)$ is compact, it has a finite diameter $D_{\mathcal{P}(X)} \in \mathbb{R}_{\geq 0}$. Let $\tilde{\omega}_k : [0, 2D_{\mathcal{P}(X)}] \to \mathbb{R}_{\geq 0}$ be a

modulus of continuity, that is a pointwise upper bound to ω_k. For all $M \in \mathbb{N}_+$, define now the McKean extension $k^{[M]}_{\mathrm{McK}} : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$ by

$$k^{[M]}_{\mathrm{McK}}(\mu, \mu') := \inf_{\vec{x}, \vec{x}' \in X^M} k^{[M]}(\vec{x}, \vec{x}') + \tilde{\omega}_k \left(d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']), (\mu, \mu') \right] \right).$$

Note that for all $M \in \mathbb{N}_+$, $k^{[M]}_{\mathrm{McK}}$ is well-defined. For this, we show that $d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']), (\mu, \mu') \right]$ belongs to the domain of $\tilde{\omega}_k$. This holds true, since

$$d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']), (\mu, \mu') \right] \leq d_{\mathrm{KR}}(\hat{\mu}[\vec{x}], \mu) + d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'], \mu') \leq 2D_{\mathcal{P}(X)}.$$

Second, we show that $k^{[M]}_{\mathrm{McK}}(\mu, \mu')$ is bounded. Since X and hence $\mathcal{P}(X)$ are non-empty, we have $k^{[M]}_{\mathrm{McK}}(\mu, \mu') < \infty$. The uniform continuity assumption on $k^{[M]}$ implies that all kernels are continuous as functions on X^{2M} and therefore (recall that $\tilde{\omega}_k \geq 0$)

$$k^{[M]}_{\mathrm{McK}}(\mu, \mu') \geq \inf_{\vec{x}, \vec{x}' \in X^M} k^{[M]}(\vec{x}, \vec{x}') > -\infty$$

by compactness of $X^M \times X^M$.

Furthermore, observe that for all $M \in \mathbb{N}_+$ and $\vec{x}, \vec{x}' \in X^M$, we have

$$k^{[M]}_{\mathrm{McK}}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) = k^{[M]}(\vec{x}, \vec{x}'). \tag{9.2}$$

For arbitrary $\vec{x}, \vec{x}' \in X^M$ it holds by construction

$$\begin{aligned} k^{[M]}_{\mathrm{McK}}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) &\leq k^{[M]}(\vec{x}, \vec{x}') + \tilde{\omega}_k \left(d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']), (\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) \right] \right) \\ &= k^{[M]}(\vec{x}, \vec{x}'). \end{aligned}$$

Let additionally $\vec{x}_1, \vec{x}'_1 \in X^M$ be arbitrary, then we obtain

$$\begin{aligned} &k^{[M]}(\vec{x}_1, \vec{x}'_1) + \tilde{\omega}_k \left(d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}'_1]), (\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) \right] \right) \\ &\quad \geq k^{[M]}(\vec{x}, \vec{x}') - |k^{[M]}(\vec{x}_1, \vec{x}'_1) - k^{[M]}(\vec{x}, \vec{x}')| + \tilde{\omega}_k \left(d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}'_1]), (\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) \right] \right) \\ &\quad \geq k^{[M]}(\vec{x}, \vec{x}') - \omega_k \left(d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}'_1]), (\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) \right] \right) + \tilde{\omega}_k \left(d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}'_1]), (\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) \right] \right) \\ &\quad \geq k^{[M]}(\vec{x}, \vec{x}') - \tilde{\omega}_k \left(d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}'_1]), (\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) \right] \right) + \tilde{\omega}_k \left(d^2_{\mathrm{KR}} \left[(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}'_1]), (\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) \right] \right) \\ &\quad = k^{[M]}(\vec{x}, \vec{x}'), \end{aligned}$$

where we used the uniform continuity of $k^{[M]}$ in the second inequality and the

definition of $\tilde{\omega}_k$ (together with $d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_1],\hat{\mu}[\vec{x}'_1]),(\hat{\mu}[\vec{x}],\hat{\mu}[\vec{x}'])\right] \leq 2D_{\mathcal{P}X}$) in the third inequality. This implies that $k^{[M]}_{\mathrm{McK}}(\hat{\mu}[\vec{x}],\hat{\mu}[\vec{x}']) \geq k^{[M]}(\vec{x},\vec{x}')$.

Step 2 We now show equi-boundedness of $(k^{[M]}_{\mathrm{McK}})_M$. Let $M \in \mathbb{N}_+$ and $\mu,\mu' \in \mathcal{P}(X)$ be arbitrary, then

$$\begin{aligned}
|k^{[M]}_{\mathrm{McK}}(\mu,\mu')| &= \left| \inf_{\vec{x},\vec{x}' \in X^M} k^{[M]}(\vec{x},\vec{x}') + \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}],\hat{\mu}[\vec{x}']),(\mu,\mu')\right]\right)\right| \\
&\leq \inf_{\vec{x},\vec{x}' \in X^M} |k^{[M]}(\vec{x},\vec{x}')| + \left|\tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}],\hat{\mu}[\vec{x}']),(\mu,\mu')\right]\right)\right| \\
&\leq C_k + \tilde{\omega}_k(2D_{\mathcal{P}(X)}) =: \tilde{C}_k,
\end{aligned}$$

where we used the uniform boundedness of $k^{[M]}$ and the compactness of $\mathcal{P}(X)$.

Step 3 Next, we show that $\tilde{\omega}_k$ is a modulus of continuity, i.e., for all $M \in \mathbb{N}_+$, $\mu_1,\mu'_1,\mu_2,\mu'_2 \in \mathcal{P}(X)$ we have

$$|k^{[M]}_{\mathrm{McK}}(\mu_1,\mu'_1) - k^{[M]}_{\mathrm{McK}}(\mu_2,\mu'_2)| \leq \tilde{\omega}_k(d^2_{\mathrm{KR}}[(\mu_1,\mu'_1),(\mu_2,\mu'_2)]).$$

To establish this, let $M \in \mathbb{N}_+$, $\mu_1,\mu'_1,\mu_2,\mu'_2 \in \mathcal{P}(X)$ and $\epsilon > 0$ be arbitrary. Now, let $(\vec{x}_2,\vec{x}'_2) \in X^{2M}$ be ϵ-close, i.e.,

$$k^{[M]}(\vec{x}_2,\vec{x}'_2) + \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_2,\mu'_2)\right]\right) \leq k^{[M]}_{\mathrm{McK}}(\mu_2,\mu'_2) + \epsilon.$$

Then, it holds

$$\begin{aligned}
k^{[M]}_{\mathrm{McK}}(\mu_1,\mu'_1) &\leq k^{[M]}(\vec{x}_2,\vec{x}'_2) + \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_1,\mu'_1)\right]\right) \\
&= k^{[M]}(\vec{x}_2,\vec{x}'_2) + \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_2,\mu'_2)\right]\right) \\
&\quad - \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_2,\mu'_2)\right]\right) + \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_1,\mu'_1)\right]\right) \\
&\leq k^{[M]}_{\mathrm{McK}}(\mu_2,\mu'_2) + \epsilon - \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_2,\mu'_2)\right]\right) \\
&\quad + \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_1,\mu'_1)\right]\right) \\
&\leq k^{[M]}_{\mathrm{McK}}(\mu_2,\mu'_2) + \epsilon - \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_2,\mu'_2)\right]\right) \\
&\quad + \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_2],\hat{\mu}[\vec{x}'_2]),(\mu_2,\mu'_2)\right] + d^2_{\mathrm{KR}}\left[(\mu_2,\mu'_2),(\mu_1,\mu'_1)\right]\right) \\
&\leq k^{[M]}_{\mathrm{McK}}(\mu_2,\mu'_2) + \epsilon + \tilde{\omega}_k\left(d^2_{\mathrm{KR}}\left[(\mu_2,\mu'_2),(\mu_1,\mu'_1)\right]\right),
\end{aligned}$$

where we used the definition of $k^{[M]}_{\mathrm{McK}}(\mu_1, \mu_1')$ in the first inequality, the choice of $(\vec{x}_2, \vec{x}_2')$ in the second inequality, the triangle inequality for d_{KR} together with the monotonicity of $\tilde{\omega}_k$ in the third inequality and finally the subadditivity. Repeating these steps with the roles interchanged shows that

$$|k^{[M]}_{\mathrm{McK}}(\mu_1, \mu_1') + k^{[M]}_{\mathrm{McK}}(\mu_2, \mu_2')| \leq \tilde{\omega}_k(d^2_{\mathrm{KR}}\left[(\mu_1, \mu_1'), (\mu_2, \mu_2')\right]) + \epsilon$$

and since $\epsilon > 0$ was arbitrary and $\tilde{\omega}_k$ does not depend on M, the claim follows.

Step 4 Summarizing, $(k^{[M]}_{\mathrm{McK}})_{M\in\mathbb{N}_+} \subseteq C^0(\mathcal{P}(X)\times\mathcal{P}(X), \mathbb{R})$ is a uniformly bounded, equi-continuous sequence. The Arzela-Ascoli theorem guarantees existence of $k \in C^0(\mathcal{P}(X) \times \mathcal{P}(X), \mathbb{R})$ and an unbounded sequence $(M_\ell)_{\ell\in\mathbb{N}_+}$ such that

$$\lim_{\ell\to\infty} \sup_{\mu,\mu'\in\mathcal{P}(X)} |k^{[M_\ell]}_{\mathrm{McK}}(\mu, \mu') - k(\mu, \mu')| = 0.$$

This implies also (9.1). To prove this, note that for all $\ell \in \mathbb{N}_+$ and $\vec{x} \in X^{M_\ell}$ we have $\hat{\mu}[\vec{x}] \in \mathcal{P}(X)$, and hence

$$\begin{aligned}\lim_{\ell\to\infty} \sup_{\vec{x},\vec{x}'\in X^{M_\ell}} |k^{[M_\ell]}(\vec{x}, \vec{x}') - k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])| &= \lim_{\ell\to\infty} \sup_{\vec{x},\vec{x}'\in X^{M_\ell}} |k^{[M_\ell]}_{\mathrm{McK}}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) - k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])| \\ &\leq \lim_{\ell\to\infty} \sup_{\mu,\mu'\in\mathcal{P}(X)} |k^{[M_\ell]}(\mu, \mu') - k(\mu, \mu')| \\ &= 0,\end{aligned}$$

where we used (9.2) in the first equality.

Step 5 Next, we show that for all $\mu_1, \mu_2 \in \mathcal{P}(X)$, $|k(\mu_1, \mu_2)| \leq C_k$, i.e., the function k is bounded. For this, let $\mu_1, \mu_2 \in \mathcal{P}(X)$ and $\epsilon > 0$ be arbitary. Choose $n \in \mathbb{N}_+$ such that

$$\|k^{[M_n]}_{\mathrm{McK}} - k\|_\infty = \sup_{\mu,\mu'\in\mathcal{P}(X)} |k^{[M_n]}_{\mathrm{McK}}(\mu, \mu') - k(\mu, \mu')| \leq \epsilon.$$

We then have

$$|k(\mu_1, \mu_2)| \leq |k(\mu_1, \mu_2) - k^{[M_n]}_{\mathrm{McK}}(\mu_1, \mu_2)| + |k^{[M_n]}_{\mathrm{McK}}(\mu_1, \mu_2)| \leq \epsilon + C_k,$$

due to the uniform boundedness of $k^{[M_n]}$. Since $\epsilon > 0$ was arbitrary, the claim follows.

Step 6 Finally, we show that k is a kernel, i.e., k is a symmetric and positive definite function on $\mathcal{P}(X)$.

Symmetry Let $\mu, \mu' \in \mathcal{P}(X)$ and $\vec{x}_M, \vec{x}'_M \in X^M$ such that $d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_M], \mu)$ and $d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'_M], \mu')$ converge to zero. For convenience, define $\hat{\mu}_\ell = \hat{\mu}[\vec{x}_{M_\ell}]$ and $\hat{\mu}'_\ell = \hat{\mu}[\vec{x}'_{M_\ell}]$. We then have

$$\begin{aligned}|k(\mu,\mu') - k(\mu',\mu)| &\leq |k(\mu,\mu') - k(\hat{\mu}_\ell,\hat{\mu}'_\ell)| + |k(\hat{\mu}_\ell,\hat{\mu}'_\ell) - k^{[M_\ell]}(\vec{x}_{M_\ell},\vec{x}'_{M_\ell})| \\ &\quad + |k^{[M_\ell]}(\vec{x}'_{M_\ell},\vec{x}_{M_\ell}) - k(\hat{\mu}'_k,\hat{\mu}_k)| + |k(\hat{\mu}'_\ell,\hat{\mu}_\ell) - k(\mu',\mu)| \\ &\to 0,\end{aligned}$$

where we used the symmetry of $k^{[M_\ell]}$ in the inequality and then the continuity of k (w.r.t. d^2_{KR}) as well as (9.1).

Positive definiteness Let $N \in \mathbb{N}_+$, $\alpha \in \mathbb{R}^N$ and $\mu_1, \ldots, \mu_N \in \mathcal{P}(X)$ as well as $\vec{x}_n^{[M]} \in X^M$ such that for all $n = 1, \ldots, N$, $d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_n^{[M]}], \mu_n) \to 0$. For convenience, define $\hat{\mu}_n^{[M]} = \hat{\mu}[\vec{x}_n^{[M]}]$. Let $\epsilon > 0$ be arbitrary. For all $i, j = 1, \ldots, N$ and M we have

$$\begin{aligned}k(\mu_i,\mu_j) &\geq k(\hat{\mu}_i^{[M]},\hat{\mu}_j^{[M]}) - |k(\mu_i,\mu_j) - k(\hat{\mu}_i^{[M]},\hat{\mu}_j^{[M]})| \\ &\geq k^{[M]}(\vec{x}_i^{[M]},\vec{x}_j^{[M]}) - |k(\hat{\mu}_i^{[M]},\hat{\mu}_j^{[M]}) - k^{[M]}(\vec{x}_i^{[M]},\vec{x}_j^{[M]})| - |k(\mu_i,\mu_j) - k(\hat{\mu}_i^{[M]},\hat{\mu}_j^{[M]})|\end{aligned}$$

Choosing ℓ large enough and setting $M = M_\ell$ ensures

$$k(\mu_i,\mu_j) \geq k^{[M_k]}(\vec{x}_i^{[M_k]},\vec{x}_j^{[M_\ell]}) - 2\epsilon$$

due to the continuity of the k and (9.1). Repeating this for all pairs (i, j) and taking the maximum over all resulting k then leads to

$$\sum_{i,j=1}^N \alpha_i\alpha_j k(\mu_i,\mu_j) \geq \sum_{i,j=1}^N \alpha_i\alpha_j k^{[M_\ell]}(\vec{x}_i^{[M_\ell]},\vec{x}_j^{[M_\ell]}) - 2N^2\epsilon \geq -2N^2\epsilon,$$

where we used that $k^{[M_\ell]}$ is a kernel. Since $\epsilon > 0$ was arbitrary, we find that

$$\sum_{i,j=1}^N \alpha_i\alpha_j k(\mu_i,\mu_j) \geq 0.$$

□

Remark 9.1.6. The function $\tilde{\omega}_k$ from the proof of Theorem 9.1.5 is also a modulus of continuity for k, i.e., for all $\mu_i \in \mathcal{P}(X)$, $i = 1, \dots, 4$,

$$|k(\mu_1, \mu_2) - k(\mu_3, \mu_4)| \leq \tilde{\omega}_k(d^2_{\mathrm{KR}}[(\mu_1, \mu_2), (\mu_3, \mu_4)]).$$

Proof. Let $\mu_i \in \mathcal{P}(X)$, $i = 1, \dots, 4$, and $\epsilon > 0$ be arbitrary. Choose $n \in \mathbb{N}_+$ such that

$$\|k^{[M_n]}_{\mathrm{McK}} - k\|_\infty = \sup_{\mu,\mu' \in \mathcal{P}(X)} |k^{[M_n]}_{\mathrm{McK}}(\mu, \mu') - k(\mu, \mu')| \leq \frac{\epsilon}{2}$$

(exists due to the Arzela-Ascoli Theorem). We then have

$$\begin{aligned} |k(\mu_1, \mu_2) - k(\mu_3, \mu_4)| &\leq |k(\mu_1, \mu_2) - k^{[M_n]}_{\mathrm{McK}}(\mu_1, \mu_2)| \\ &\quad + |k^{[M_n]}_{\mathrm{McK}}(\mu_1, \mu_2) - k^{[M_n]}_{\mathrm{McK}}(\mu_3, \mu_4)| + |k^{[M_n]}_{\mathrm{McK}}(\mu_3, \mu_4) - k(\mu_3, \mu_4)| \\ &\leq \frac{\epsilon}{2} + \tilde{\omega}_k(d^2_{\mathrm{KR}}[(\mu_1, \mu_2), (\mu_3, \mu_4)]) + \frac{\epsilon}{2} \end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we find that

$$|k(\mu_1, \mu_2) - k(\mu_3, \mu_4)| \leq \tilde{\omega}_k(d^2_{\mathrm{KR}}[(\mu_1, \mu_2), (\mu_3, \mu_4)]).$$

This finishes the proof. □

Remark 9.1.7. It is also possible to generalize Assumption 9.1.2 and Theorem 9.1.5 to kernel sequences of the form $k^{[M]} : (Y \times X^M) \times (Y \times X^M) \to \mathbb{R}$ for some compact metric space Y, leading to a mean field kernel $k : (Y \times \mathcal{P}(X)) \times (Y \times \mathcal{P}(X)) \to \mathbb{R}$ using techniques presented for example in [33].

9.2. Examples of kernel mean field limits

We now turn to examples of kernels and their mean field. We will consider two large classes of suitable kernels, pull-back kernels and double-sum kernels. To close, we will discuss and intriguing connection between double-sum kernels in the mean field and kernel mean embeddings.

9.2.1. Pullback kernels

Our first example are sequences of kernels that arise as the pull-backs [152, Section 5.4] of a sufficiently regular kernel along mean field compatible functions.

Proposition 9.2.1. Let Y be a Banach space, $k_0 : Y \times Y \to \mathbb{R}$ be a kernel on Y and $\phi^{[M]} : X^M \to Y$ a sequence of functions. Furthermore, assume that

1. *(Boundedness of k_0)* There exists a $C_{k_0} \in \mathbb{R}_{\geq 0}$ with $|k_0(y, y')| \leq C_{k_0}$ for all $y, y' \in Y$.

2. *(Continuity of k_0)* The kernel k_0 has a modulus of continuity ω_{k_0}, i.e.,

$$|k_0(y_1, y_1') - k_0(y_2, y_2')| \leq \omega_{k_0}(\|y_1 - y_2\|_Y + \|y_1' - y_2'\|_Y)$$

   for all $y_1, y_1', y_2, y_2' \in Y$.

3. *(Symmetry of $\phi^{[M]}$)* For all $M \in \mathbb{N}$, the function $\phi^{[M]}$ is permutation invariant, i.e., for all $\vec{x} \in X^M$ and $\sigma \in \mathcal{S}_M$ we have $\phi^{[M]}(\sigma\vec{x}) = \phi^{[M]}(\vec{x})$.

4. *(Uniform continuity of $\phi^{[M]}$)* There exists a modulus of continuity $\omega_\phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that for all $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}' \in X^M$

$$\|\phi^{[M]}(\vec{x}) - \phi^{[M]}(\vec{x}')\|_Y \leq \omega_\phi \left(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])\right).$$

Then $k^{[M]} : X^M \times X^M \to \mathbb{R}$ with $k^{[M]}(\vec{x}, \vec{x}') = k_0(\phi^{[M]}(\vec{x}), \phi^{[M]}(\vec{x}'))$ is a sequence of kernels on X^M fulfilling Assumption 9.1.2.

Proof. Since $k^{[M]}$ is the pull-back of k_0 along $\phi^{[M]}$, it is a kernel on X^M. Symmetry is clear,

$$k^{[M]}(\sigma\vec{x}, \vec{x}') = k_0(\phi^{[M]}(\sigma\vec{x}), \phi^{[M]}(\vec{x}')) = k_0(\phi^{[M]}(\vec{x}), \phi^{[M]}(\vec{x}')) = k^{[M]}(\vec{x}, \vec{x}').$$

Uniform boundedness follows from boundedness of k_0, hence $C_k = C_{k_0}$. For the

uniform continuity, let $M \in \mathbb{N}_+$, $\vec{x}_1, \vec{x}_1', \vec{x}_2, \vec{x}_2' \in X^M$, then

$$\begin{aligned}
|k^{[M]}(\vec{x}_1, \vec{x}_1') - k^{[M]}(\vec{x}_2, \vec{x}_2')| &= |k_0(\phi^{[M]}(\vec{x}_1), \phi^{[M]}(\vec{x}_1')) - k_0(\phi^{[M]}(\vec{x}_2), \phi^{[M]}(\vec{x}_2'))| \\
&\leq \omega_{k_0}\left(\|\phi^{[M]}(\vec{x}_1) - \phi^{[M]}(\vec{x}_2)\|_Y + \|\phi^{[M]}(\vec{x}_1') - \phi^{[M]}(\vec{x}_2')\|_Y\right) \\
&\leq \omega_{k_0}\left(\omega_\phi(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2])) + \omega_\phi(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2']))\right) \\
&\leq \omega_k\left(d_{\mathrm{KR}}^2\left[(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_1']), (\hat{\mu}[\vec{x}_2], \hat{\mu}[\vec{x}_2'])\right]\right)
\end{aligned}$$

for an appropriate modulus of continuity ω_k. □

9.2.2. Double-sum kernels: Abstract perspective

The next class of examples has been introduced by [108] and extended by [39], though similar constructions have been used earlier [78]. However, the connection to mean field limits and kernel mean embeddings has not yet been investigated.

Proposition 9.2.2. Let $k_0 : X \times X \to \mathbb{R}$ be a kernel bounded by $|k_0(x, x')| \leq C_{k_0}$ for some $C_{k_0} \in \mathbb{R}_{\geq 0}$. Define for $M \in \mathbb{N}_+$ the map $k^{[M]} : X^M \times X^M \to \mathbb{R}$ by

$$k^{[M]}(\vec{x}, \vec{x}') = \frac{1}{M^2} \sum_{m,m'=1}^{M} k_0(x_m, x_{m'}'). \tag{9.3}$$

Then $k^{[M]}$ are kernels that are permutation invariant in their first argument, and that are uniformly bounded.

Proof. Let $M \in \mathbb{N}_+$ be arbitrary. First, we establish that $k^{[M]}$ is indeed a kernel by showing that it is a symmetric, positive definite function. Note that this fact has been established earlier, cf. e.g. [39], but for convenience we provide a full proof. For all $\vec{x}, \vec{x}' \in X^M$ we have (using the symmetry of k)

$$k^{[M]}(\vec{x}, \vec{x}') = \frac{1}{M^2} \sum_{m,m'=1}^{M} k_0(x_m, x_{m'}') = \frac{1}{M^2} \sum_{m,m'=1}^{M} k_0(x_{m'}', x_m) = k^{[M]}(\vec{x}', \vec{x}),$$

i.e., $k^{[M]}$ is symmetric. Let $N \in \mathbb{N}_+$ and $\vec{x}^1, \ldots, \vec{x}^N \in X^M$, $\alpha \in \mathbb{R}^N$ be arbitrary,

then

$$\begin{aligned}\sum_{i,j=1}^{N} \alpha_i \alpha_j k^{[M]}(\vec{x}^i, \vec{x}^j) &= \sum_{i,j=1}^{N} \alpha_i \alpha_j \frac{1}{M^2} \sum_{m,m'=1}^{M} k_0(x_m^i, x_{m'}^j) \\ &= \sum_{i,j=1}^{N} \sum_{m,m'=1}^{M} \alpha_i \alpha_j \frac{1}{M^2} k_0(x_m^i, x_{m'}^j) \\ &= \sum_{(i,m),(j,m') \in \mathcal{I}} \frac{\alpha_i}{M} \frac{\alpha_j}{M} k_0(x_m^i, x_{m'}^j) \geq 0,\end{aligned}$$

where we defined $\mathcal{I} = \{1, \ldots, N\} \times \{1, \ldots, M\}$ and used that k_0 is positive definite.

For the uniform boundedness, let $\vec{x}, \vec{x}' \in X^M$, then

$$|k^{[M]}(\vec{x}, \vec{x}')| \leq \frac{1}{M^2} \sum_{m,m'=1}^{M} |k_0(x_m, x'_{m'})| \leq \frac{1}{M^2} M^2 C_{k_0} = C_{k_0}.$$

□

In addition to permutation-invariance and boundedness, we also have a form of uniform continuity of double sum kernels.

Proposition 9.2.3. Let $k_0 : X \times X \to \mathbb{R}$ be a kernel bounded by $|k_0(x, x')| \leq C_{k_0}$ for some $C_{k_0} \in \mathbb{R}_{\geq 0}$, and assume that (X, d_{k_0}) is a separable metric space, where

$$d_{k_0} : X \times X \to \mathbb{R}_{\geq 0},\ d_{k_0}(x, x') = \|\Phi_{k_0}(x) - \Phi_{k_0}(x')\|_{k_0}$$

is the usual kernel metric. Then the double sum kernels $k^{[M]}$ defined in (9.3) are uniformly continuous with respect to the Kantorowich-Rubinstein distance induced by d_{k_0}.

Proof. Observe that for $\vec{x}, \vec{x}' \in X^M$ we have

$$\begin{aligned}k^{[M]}(\vec{x}, \vec{x}') &= \frac{1}{M^2} \sum_{m,m'=1}^{M} k_0(x_m, x'_{m'}) = \frac{1}{M^2} \sum_{m,m'=1}^{M} \langle k_0(\cdot, x'_{m'}), k_0(\cdot, x_m) \rangle_{k_0} \\ &= \left\langle \frac{1}{M} \sum_{m'=1}^{M} k_0(\cdot, x'_{m'}), \frac{1}{M} \sum_{m=1}^{M} k_0(\cdot, x_m) \right\rangle_{k_0} = \langle f_{\hat{\mu}[\vec{x}']}, f_{\hat{\mu}[\vec{x}]} \rangle_{k_0}.\end{aligned}$$

Furthermore, we also have for any $\vec{x} \in X^M$

$$\begin{aligned}
\|f^{k_0}_{\hat{\mu}[\vec{x}]}\|_{k_0} &= \sqrt{\langle \int k_0(\cdot, x) \mathrm{d}\hat{\mu}[\vec{x}](x), \int k_0(\cdot, x') \mathrm{d}\hat{\mu}[\vec{x}](x') \rangle_{k_0}} \\
&= \sqrt{\int \int \langle k_0(\cdot, x), k_0(\cdot, x') \rangle_{k_0} \mathrm{d}\hat{\mu}[\vec{x}](x) \mathrm{d}\hat{\mu}[\vec{x}](x')} \\
&\leq \sqrt{\int \int |k_0(x', x)| \mathrm{d}\hat{\mu}[\vec{x}](x) \mathrm{d}\hat{\mu}[\vec{x}](x')} \\
&\leq \sqrt{C_{k_0}}.
\end{aligned}$$

Let $\vec{x}_1, \vec{x}_2, \vec{x}'_1, \vec{x}'_2 \in X^M$, then

$$\begin{aligned}
|k^{[M]}(\vec{x}_1, \vec{x}'_1) - k^{[M]}(\vec{x}_2, \vec{x}'_2)| &= |\langle f^{k_0}_{\hat{\mu}[\vec{x}'_1]}, f^{k_0}_{\hat{\mu}[\vec{x}_1]} \rangle_{k_0} - \langle f^{k_0}_{\hat{\mu}[\vec{x}'_2]}, f^{k_0}_{\hat{\mu}[\vec{x}_2]} \rangle_{k_0}| \\
&= |\langle f^{k_0}_{\hat{\mu}[\vec{x}'_1]} - f^{k_0}_{\hat{\mu}[\vec{x}'_2]}, f^{k_0}_{\hat{\mu}[\vec{x}_1]} \rangle_{k_0} + \langle f^{k_0}_{\hat{\mu}[\vec{x}'_2]}, f^{k_0}_{\hat{\mu}[\vec{x}_1]} - f^{k_0}_{\hat{\mu}[\vec{x}_2]} \rangle_{k_0}| \\
&\leq \|f^{k_0}_{\hat{\mu}[\vec{x}'_1]} - f^{k_0}_{\hat{\mu}[\vec{x}'_2]}\|_{k_0} \|f^{k_0}_{\hat{\mu}[\vec{x}_1]}\|_{k_0} + \|f^{k_0}_{\hat{\mu}[\vec{x}'_2]}\|_{k_0} \|f^{k_0}_{\hat{\mu}[\vec{x}_1]} - f^{k_0}_{\hat{\mu}[\vec{x}_2]}\|_{k_0} \\
&\leq \sqrt{C_k}(\|f^{k_0}_{\hat{\mu}[\vec{x}'_1]} - f^{k_0}_{\hat{\mu}[\vec{x}'_2]}\|_{k_0} + \|f^{k_0}_{\hat{\mu}[\vec{x}_1]} - f^{k_0}_{\hat{\mu}[\vec{x}_2]}\|_{k_0})
\end{aligned}$$

Next, since (X, d_{k_0}) is separable, [188, Theorem 21] shows that $\|f^{k_0}_{\hat{\mu}[\vec{x}_1]} - f^{k_0}_{\hat{\mu}[\vec{x}_2]}\|_{k_0} \leq \widetilde{d_{\mathrm{KR}}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2])$ and $\|f^{k_0}_{\hat{\mu}[\vec{x}'_1]} - f^{k_0}_{\hat{\mu}[\vec{x}'_2]}\|_{k_0} \leq \widetilde{d_{\mathrm{KR}}}(\hat{\mu}[\vec{x}'_1], \hat{\mu}[\vec{x}'_2])$, where

$$\widetilde{d_{\mathrm{KR}}}(\mu_1, \mu_2) = \sup \left\{ \int_X \phi(x) \mathrm{d}(\mu_1 - \mu_2)(x) \mid \phi : X \to \mathbb{R} \text{ is 1-Lipschitz w.r.t. } d_{k_0} \right\},$$

the Kantorowich-Rubinstein distance induced by d_{k_0}. Altogether, we find that

$$|k^{[M]}(\vec{x}_1, \vec{x}'_1) - k^{[M]}(\vec{x}_2, \vec{x}'_2)| \leq \sqrt{C_{k_0}}(\widetilde{d_{\mathrm{KR}}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2]) + \widetilde{d_{\mathrm{KR}}}(\hat{\mu}[\vec{x}'_1], \hat{\mu}[\vec{x}'_2])),$$

but since $\sqrt{C_{k_0}}$ does not depend on M, this establishes uniform continuity of $k^{[M]}$ w.r.t. $\widetilde{d_{\mathrm{KR}}}$. □

In Proposition 9.2.3, we have not established uniform continuity of the double sum kernels (9.3) with respect to the Kantorowich-Rubinstein distance induced by the metric d_X. In particular, combining Propositions 9.2.2 and 9.2.3 is not enough to ensure that the double sum kernels fulfill Assumption 9.1.2. However, if (X, d_{k_0}) is a compact, separable metric space, then Proposition 9.2.3 implies that Assumption

9.1.2, now with d_{k_0} instead of d_X, applies to the kernel sequence (9.3). In this case, Theorem 9.1.5 shows the existence of the mean field limit kernel and its associated RKHS, again with d_{k_0} instead of d_X. These observations motivate the developments in the next section.

9.2.3. Double sum kernels: Elementary approach

In the preceding section, we dealt with double sums kernels within the general theory for mean field limits of kernels. We will now revisit this class of kernels from a more elementary perspective. Let k_0 be a bounded kernel on X, so there exists $B_0 \in \mathbb{R}_{\geq 0}$ such that $|k_0(x, x')| \leq B_0$ for all $x, x' \in X$. Define now $k : \mathcal{P}(X) \times \mathcal{P}(X) \to \mathbb{R}$ by

$$k(\mu, \nu) = \int_X \int_X k_0(x, x') \mathrm{d}\mu(x) \mathrm{d}\nu(x'). \tag{9.4}$$

Since $\mu, \nu \in \mathcal{P}(X)$ are finite measures and k_0 is bounded, the double integral above is well-defined. Furthermore, we have

$$\begin{aligned} k(\mu, \nu) &= \int_X \int_X k_0(x, x') \mathrm{d}\mu(x) \mathrm{d}\nu(x') \\ &= \int_X \int_X \langle k_0(\cdot, x'), k_0(\cdot, x) \rangle_{k_0} \mathrm{d}\mu(x) \mathrm{d}\nu(x') \\ &= \Big\langle \int_X k_0(\cdot, x') \mathrm{d}\nu(x), \int_X k_0(\cdot, x) \mathrm{d}\mu(x) \Big\rangle_{k_0}, \end{aligned}$$

where the integrals in the last line are in the sense of Bochner, cf. [188, Theorem 1], and we used in the last step that the scalar product as a continuous linear functional commutes with the Bochner integral. The above equality shows that k is indeed a kernel on $\mathcal{P}(X)$.

For $M \in \mathbb{N}_+$, define $k^{[M]} : X^M \times X^M \to \mathbb{R}$ by

$$k^{[M]}(\vec{x}, \vec{x}') = \frac{1}{M^2} \sum_{i,j=1}^{M} k_0(x_i, x'_j). \tag{9.5}$$

These bivariate maps are called double sum kernels, and it is well-known that they are indeed kernels, and permutation-invariant. Furthermore, since for $M \in \mathbb{N}_+$ and

$\vec{x}, \vec{x}' \in X^M$ we have

$$|k^{[M]}(\vec{x}, \vec{x}')| = \left| \frac{1}{M^2} \sum_{i,j=1}^{M} k_0(x_i, x'_j) \right| \leq \frac{1}{M^2} \sum_{i,j=1}^{M} |k_0(x_i, x'_j)| \leq B_0,$$

the kernels $k^{[M]}$ are uniformly bounded.

Observe now that for all $M \in \mathbb{N}_+$ and $\vec{x}, \vec{x}' \in X^M$ we have

$$\begin{aligned} k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) &= \int_X \int_X k_0(x, x') \mathrm{d}\hat{\mu}[\vec{x}](x) \mathrm{d}\hat{\mu}[\vec{x}'](x') \\ &= \frac{1}{M} \sum_{i=1}^{M} \frac{1}{M} \sum_{j=1}^{M} k(x_i, x'_j) \\ &= k^{[M]}(\vec{x}, \vec{x}'), \end{aligned}$$

which implies that

$$\lim_{M \to \infty} \sup_{\vec{x}, \vec{x}' \in X^M} |k^{[M]}(\vec{x}, \vec{x}') - k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])| = 0,$$

so the kernels $k^{[M]}$ converge to k in mean field in the sense of Definition 9.1.1.

Remark 9.2.4. The preceding developments work for any measurable space $(X, \mathcal{A}_X)$, where $\mathcal{P}(X)$ is now the set of probability measures defined on this measurable space, and any $\mathcal{A} \otimes \mathcal{A}$-$\mathcal{B}(\mathbb{R})$-measurable (here $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$) and bounded kernel $k_0 : X \times X \to \mathbb{R}$.

For some applications, we might need additional properties of the involved kernels. For a particular and broad class of base kernels, the following result provides a sufficient condition for uniform Lipschitz continuity. To the best of our knowledge, this result appeared for the first time in [CF4].

Proposition 9.2.5. Let $\mathcal{X}$ be a normed vectorspace, $X \subseteq \mathcal{X}$ a non-empty Borel-measurable subset, $\phi : X \to \mathbb{R}$ a L-Lipschitz continuous function, define $\kappa_0 : X \times X \to \mathbb{R}$ by $\kappa_0(x, x') = \phi(\|x - x'\|)$, and for $M \in \mathbb{N}_+$ define $\kappa_M(\vec{x}, \vec{x}') = \frac{1}{M^2} \sum_{i,j=1}^{M} \kappa_0(\vec{x}_i, \vec{x}'_j)$. We then have for all $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}', \vec{y}, \vec{y}' \in X^M$ that

$$|\kappa_M(\vec{x}, \vec{x}') - \kappa_M(\vec{y}, \vec{y}')| \leq L \, d_{\mathrm{KR}}^2 \left((\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']), (\hat{\mu}[\vec{y}], \hat{\mu}[\vec{y}']) \right).$$

Proof. Without loss of generality we can assume that $L \in \mathbb{R}_{>0}$. Define for $x \in X$ the function $\varphi_x : X \to \mathbb{R}$ by $\varphi_x(x') = L^{-1}\phi(\|x' - x\|)$, and observe that since that for all $x', y' \in X$

$$\begin{aligned}
|\varphi_x(x') - \varphi_x(y')| &= L^{-1}|\phi(\|x' - x\|) - \phi(\|y' - x\|)| \\
&\leq L^{-1}L|\|x' - x\| - \|y' - x\|| \\
&\leq \|(x' - x) - (y' - x)\| \\
&= \|x' - y'\|
\end{aligned}$$

the function φ_x is 1-Lipschitz continuous.

Let now $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}', \vec{y}, \vec{y}' \in X^M$, then we get

$$\begin{aligned}
|\kappa_M(\vec{x}, \vec{x}') - \kappa_M(\vec{y}, \vec{y}')| &= \left| \frac{1}{M^2} \sum_{i,j=1}^{M} \kappa_0(x_i, x'_j) - \frac{1}{M^2} \sum_{i,j=1}^{M} \kappa_0(y_i, y'_j) \right| \\
&= \left| \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{M} \sum_{j=1}^{M} \phi(\|x_i - x'_j\|) - \frac{1}{M} \sum_{j=1}^{M} \phi(\|y_i - y'_j\|) \right) \right| \\
&\leq \left| \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{M} \sum_{j=1}^{M} \phi(\|x_i - x'_j\|) - \frac{1}{M} \sum_{j=1}^{M} \phi(\|x_i - y'_j\|) \right) \right| \\
&\quad + \left| \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{M} \sum_{j=1}^{M} \phi(\|x_i - y'_j\|) - \frac{1}{M} \sum_{j=1}^{M} \phi(\|y_i - y'_j\|) \right) \right| \\
&= L \left| \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1}{M} \sum_{j=1}^{M} \varphi_{x_i}(x'_j) - \frac{1}{M} \sum_{j=1}^{M} \varphi_{x_i}(y'_j) \right) \right| \\
&\quad + L \left| \frac{1}{M} \sum_{j=1}^{M} \left(\frac{1}{M} \sum_{i=1}^{M} \varphi_{y'_j}(x_i) - \frac{1}{M} \sum_{i=1}^{M} \varphi_{y'_j}(y_i) \right) \right| \\
&\leq L \frac{1}{M} \sum_{i=1}^{M} \left| \frac{1}{M} \sum_{j=1}^{M} \varphi_{x_i}(x'_j) - \frac{1}{M} \sum_{j=1}^{M} \varphi_{x_i}(y'_j) \right| \\
&\quad + L \frac{1}{M} \sum_{j=1}^{M} \left| \frac{1}{M} \sum_{i=1}^{M} \varphi_{y'_j}(x_i) - \frac{1}{M} \sum_{i=1}^{M} \varphi_{y'_j}(y_i) \right|.
\end{aligned}$$

Observe now that for all Borel-measurable $f : X \to \mathbb{R}$ we have

$$\frac{1}{M}\sum_{i=1}^{M} f(x_i) = \int_X f(x)\mathrm{d}\hat{\mu}[\vec{x}](x),$$

so we can continue with

$$\begin{aligned}
|\kappa_M(\vec{x},\vec{x}') - \kappa_M(\vec{y},\vec{y}')| &\leq L\frac{1}{M}\sum_{i=1}^{M}\left|\int_X \varphi_{x_i}(x')\mathrm{d}\hat{\mu}[\vec{x}'](x') - \int_X \varphi_{x_i}(y')\mathrm{d}\hat{\mu}[\vec{y}'](y')\right| \\
&\quad + L\frac{1}{M}\sum_{j=1}^{M}\left|\int_X \varphi_{y'_j}(x)\mathrm{d}\hat{\mu}[\vec{x}](x) - \int_X \varphi_{y'_j}(y)\mathrm{d}\hat{\mu}[\vec{y}](y)\right| \\
&\leq L\frac{1}{M}\sum_{i=1}^{M}\sup_{\substack{f:X\to\mathbb{R}\\ f \text{ 1-Lipschitz}}}\left|\int_X f(x')\mathrm{d}\hat{\mu}[\vec{x}'](x') - \int_X f(y')\mathrm{d}\hat{\mu}[\vec{y}'](y')\right| \\
&\quad + L\frac{1}{M}\sum_{j=1}^{M}\sup_{\substack{f:X\to\mathbb{R}\\ f \text{ 1-Lipschitz}}}\left|\int_X f(x)\mathrm{d}\hat{\mu}[\vec{x}](x) - \int_X f(y)\mathrm{d}\hat{\mu}[\vec{y}](y)\right| \\
&= L\frac{1}{M}\sum_{i=1}^{M} d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'],\hat{\mu}[\vec{y}']) + L\frac{1}{M}\sum_{j=1}^{M} d_{\mathrm{KR}}(\hat{\mu}[\vec{x}],\hat{\mu}[\vec{y}]) \\
&= L(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'],\hat{\mu}[\vec{y}']) + d_{\mathrm{KR}}(\hat{\mu}[\vec{x}],\hat{\mu}[\vec{y}])) \\
&= L d^2_{\mathrm{KR}}\left((\hat{\mu}[\vec{x}],\hat{\mu}[\vec{x}']),(\hat{\mu}[\vec{y}],\hat{\mu}[\vec{y}'])\right),
\end{aligned}$$

establishing the claim. □

We illustrate the results with a concrete example.

Example 9.2.6. Let consider now $X \subseteq \mathcal{H}$ a nonempty subset of a Hilbert space. A kernel k_0 on X of the form $k_0(x,x') = \phi(\|x - x'\|)$ is called a *radial kernel*, or a *radial basis function (kernel)*. In the following, we consider $\mathcal{H} = \mathbb{R}^d$, $X \subseteq \mathbb{R}^d$ a nonempty compact subset, and choose k_0 as the Gaussian kernel, so in this case $\phi(s) = \exp(-\frac{s^2}{2\gamma^2})$. Observe that this ϕ is bounded, and (globally) Lipschitz-continuous with Lipschitz bound given by $\max_{s\in\mathbb{R}}|\phi'(s)|$, so the resulting sequence of double sum kernel fulfills all conditions from Theorem 9.1.5.

Finally, we would like to point out the following delicate aspect of the preceding developments. By direct calculation, we have established the mean field convergence

of the double sum kernels k_M, as defined in (9.5), to k given by (9.4). Furthermore, the sequence of double sum kernels based on the Gaussian kernel fulfills all the conditions of Theorem 9.1.5, so there exists a mean field limit kernel that is bounded and Lipschitz continuous, and a subsequence of the double sum kernel sequence, that converges in mean field to this latter kernel. However, we *did not prove that this kernel is* (9.4). If we had uniqueness of the mean field limit kernel in Theorem 9.1.5, then this would trivially follow. Investigation of this uniqueness question is beyond the scope of the present work. However, it is clear that (9.4) is bounded, and by using mutatis mutandis the arguments from the proof of Proposition 9.2.5, one can verify that (9.4) is Lipschitz continuous. This means that (9.4) fulfills the properties from the limit kernel in Theorem 9.1.5.

9.2.4. Double sum kernels, mean field limits, and kernel mean embeddings

Recall from the proof of Proposition 9.2.3 that for all $M \in \mathbb{N}_+$ and $\vec{x}, \vec{x}' \in X^M$ we have

$$\begin{aligned} k^{[M]}(\vec{x}, \vec{x}') &= \frac{1}{M^2} \sum_{m,m'=1}^{M} k_0(x_m, x'_{m'}) = \frac{1}{M^2} \sum_{m,m'=1}^{M} \langle k_0(\cdot, x'_{m'}), k_0(\cdot, x_m) \rangle_{k_0} \\ &= \left\langle \frac{1}{M} \sum_{m'=1}^{M} k_0(\cdot, x'_{m'}), \frac{1}{M} \sum_{m=1}^{M} k_0(\cdot, x_m) \right\rangle_{k_0} = \langle f^{k_0}_{\hat{\mu}[\vec{x}']}, f^{k_0}_{\hat{\mu}[\vec{x}]} \rangle_{k_0}. \end{aligned}$$

This equality implies that for all $M \in \mathbb{N}_+$ the RKHS H_0 is a feature space and $\Phi_M : X^M \to H_0$, $\Phi_M(\vec{x}) = f^{k_0}_{\hat{\mu}[\vec{x}]}$ is a feature map for $k^{[M]}$. Furthermore, defining $e^{[M]}(x) = (x \cdots x) \in X^M$ for $x \in X$ and $M \in \mathbb{N}_+$, we obtain that for all $M \in \mathbb{N}_+$, $\vec{x} \in X^M$ and $\bar{x} \in X$

$$\Phi^{[M]}(\vec{x})(e^{[M]}(\bar{x})) = k^{[M]}(e^{[M]}(\bar{x}), \vec{x}) = f^{k_0}_{\hat{\mu}[\vec{x}]}(\bar{x})$$

Note that $e^{[M]}(\bar{x})$ can be interpreted as a representation of $\delta_{\bar{x}}$ in X^M since $\hat{\mu}[e^{[M]}(\bar{x})] = \delta_{\bar{x}}$. Altogether, we have now two different kernel-based embeddings of empirical probability distributions: We can embed $\hat{\mu}[\vec{x}]$ into H_0 via the kernel mean embedding $f^{k_0}_{\hat{\mu}[\vec{x}]}$ or we can identify $\hat{\mu}[\vec{x}]$ with $\vec{x}$ and embed into H_M with the canonical feature map $\Phi^{[M]}(\vec{x}) = k^{[M]}(\cdot, \vec{x})$. Those two embeddings are connected by evalua-

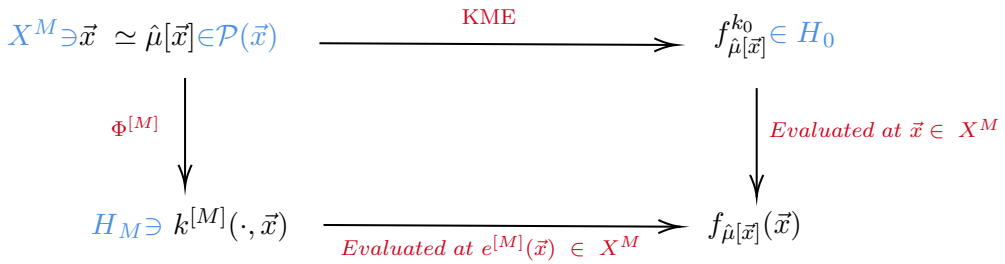

Figure 9.1.: Commutative diagram on the relation of canonical feature map of $k^{[M]}$ and KMEs.

tions on a Dirac distribution, represented by $\bar{x} \in X$ and $e^{[M]}(\bar{x}) \in X^M$, respectively. This leads to the commutative diagram in Figure 9.1.

An interesting situation arises if we consider the weak$*$ convergence of empirical probability measures, metrized by d_{KR}, and the convergence of their embeddings. Consider the setting of Propositions 9.2.2 and 9.2.3 and assume additionally that the double sum kernels (9.3) are uniformly continuous, so that Theorem 9.1.5 applies and we have the mean field limit kernel k and its associated RKHS H_k, as well as convergence (of a subsequence) of $k^{[M]}$ to k. Let $\vec{x}_M \in X^M$ with $\hat{\mu}[\vec{x}_M] \xrightarrow{d_{\mathrm{KR}}} \mu$ for some $\mu \in \mathbb{P}(X)$. Each empirical measure $\hat{\mu}[\vec{x}_M]$ can be embedded into H_0 via the kernel mean embeddings $f^{k_0}_{\hat{\mu}[\vec{x}_M]}$ and into H_M by first identifying it with $\vec{x}_M$ and then using the canonical feature map $\Phi^{[M]}$. Assume now that k_0 is characteristic, i.e., the map $\mathcal{P}(X) \to H_k$, $\mu \mapsto f^k_\mu$ is injective. Under this assumption, convergence of the kernel mean embeddings metrizes the weak$*$ topology [183, Theorem 12], so we get that $f^{k_0}_{\hat{\mu}[\vec{x}_M]} \xrightarrow{H_0} f^{k_0}_\mu$. Since k is the MFL of $k^{[M]}$ and the former is continuous w.r.t. d_{KR}, we also get up to a subsequence $k^{[M]}(\cdot, \vec{x}_M) \to k(\cdot, \mu)$ as a mean field limit. Note that the kernel mean embeddings that appear here are all well-defined, cf. [188, Theorem 1].

The preceding discussion is summarized as a diagram in Figure 9.2.

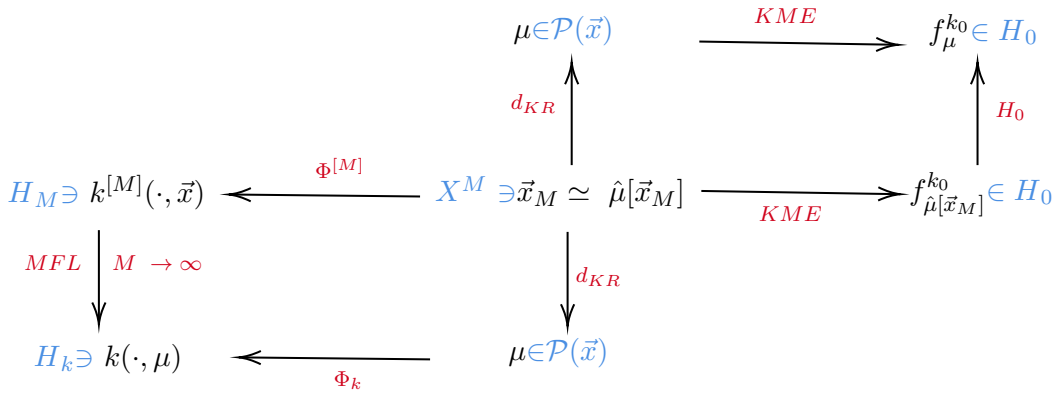

Figure 9.2.: Diagram illustration of the relations of double sum kernel, KME and MFL.

9.3. The reproducing kernel Hilbert space of the mean field limit kernel

The mean field limit k established above is a kernel and therefore it is associated with a unique RKHS. We will now investigate the relation of this RKHS with the RKHSs of the finite-input kernels. We slightly simplify the setting and notation to ease the exposition of the following results. Consider a sequence of kernels $k_M : X^M \times X^M \to \mathbb{R}$, $M \in \mathbb{N}_+$, with the following properties[1].

1. For all $M \in \mathbb{N}_+$, $\vec{x}, \vec{x}' \in X^M$ and permutations $\sigma \in \mathcal{S}_M$ we have $k_M(\sigma\vec{x}, \vec{x}') = k_M(\vec{x}, \vec{x}')$.

2. There exists $C_k \in \mathbb{R}_{\geq 0}$ such that $\forall M \in \mathbb{N}_+, \vec{x}, \vec{x}' \in X^M : |k_M(\vec{x}, \vec{x}')| \leq C_k$.

3. There exists some $L_k \in \mathbb{R}_{>0}$ such that for all $M \in \mathbb{N}_+$, $\vec{x}_1, \vec{x}_1', \vec{x}_2, \vec{x}_2' \in X^M$ we have $|k_M(\vec{x}_1, \vec{x}_1') - k_M(\vec{x}_2, \vec{x}_2')| \leq L_k d^2_{\mathrm{KR}}\left[(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_1']), (\hat{\mu}[\vec{x}_2], \hat{\mu}[\vec{x}_2'])\right]$.

Theorem 9.1.5 then ensures the existence of a mean field limit kernel $k : \mathbb{P}(X) \times \mathbb{P}(X) \to \mathbb{R}$ which is also bounded C_k and is L_k-Lipschitz continuous. By potentially

[1] The only different to Theorem 9.1.5 is the restriction of uniform Lipschitz continuity instead of an arbitrary modulus of continuity. This simplification is common in the related literature, e.g. [50, Lemma 1.2], and the following developments could be generalized to an arbitrary modulus of continuity.

re-indexing, we have $k_M \xrightarrow{\mathcal{P}_1} k$. Finally, denote by $H_M := H_{k_M}$ the (unique) RKHS corresponding to kernel k_M and denote by H_k the unique RKHS of k.

We clarify the relation between $H_M$and H_k in the next result.

Theorem 9.3.1. 1. For every $f \in H_k$, there exists a sequence $f_M \in H_M$, $M \in \mathbb{N}_+$, such that $f_M \xrightarrow{\mathcal{P}_1} f$.

2. Let $f_M \in H_M$ be sequence such that there exists $B \in \mathbb{R}_{\geq 0}$ with $\|f_M\|_M \leq B$ for all $M \in \mathbb{N}_+$. Then there exists a subsequence $(f_{M_\ell})_\ell$ and $f \in H_k$ with $f_{M_\ell} \xrightarrow{\mathcal{P}_1} f$ and $\|f\|_k \leq B$.

In other words, on the one hand, every RKHS function from $H_k$arises as a mean field limit of RKHS functions from H_M.On the other hand, every uniformly norm-bounded sequence of RKHS functions $(f_M)_M$ has a mean field limit in H_k.

The relation between the kernels k_M and their RKHSs H_M, and the mean field limit kernel k and its RKHS H_k is illustrated as a commutative diagram in Figure 9.3. In order to arrive at the mean field RKHS H_k, on the one hand, we consider the mean field limit k of the k_M, and then form the corresponding RKHS H_k. This is essentially the content of Theorem 9.1.5. On the other hand, we can first go from the kernel k_M to the associated unique RKHS H_M (for each $M \in \mathbb{N}_+$). Theorem 9.3.1 then says that H_k can be interpreted as a mean field limit of the RKHSs H_M, since every function in H_k arises as a mean field limit of a sequence of functions from the H_M, and every uniformly norm-bounded sequence of such functions has a mean field limit that is in H_k. Next, we state two technical results that will play an important role in the following developments, and which might be of independent interest. They describe lim inf and lim sup inequalities required for Γ-convergence arguments used later on.

Lemma 9.3.2. Let $f_M \in H_M$, $M \in \mathbb{N}_+$, and $f \in H_k$ such that $f_M \xrightarrow{\mathcal{P}_1} f$, then

$$\|f\|_k \leq \liminf_{M\to\infty} \|f_M\|_M. \tag{9.6}$$

Proof. Assume the statement is not true, i.e., $\|f\|_k > \liminf_{M\to\infty} \|f_M\|_M$. This means that there exists a subsequence M_ℓ and $C \in \mathbb{R}_{\geq 0}$ such that $\|f\|_k > \lim_\ell \|f_{M_\ell}\|_{M_\ell} = C$. Note that this implies that $\|f\|_k > 0$.

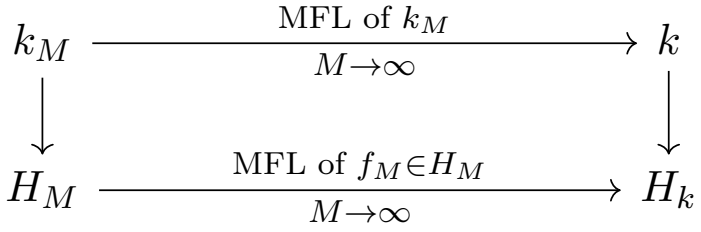

Figure 9.3.: The kernel k arises as the mean field limit (MFL) of the kernels k_M. Every uniformly norm-bounded sequence $f_M \in H_M$, $M \in \mathbb{N}_+$, has an MFL in H_k, and every function $f \in H_k$ arises as such an MFL (Theorem 9.3.1).

Let $\epsilon_1, \epsilon_2 > 0$ and $\alpha > 1$, $\beta \in (0,1)$ be arbitrary. From Theorem 9.4.1, there exists $(\vec{\mu}, \vec{\alpha}) \in \mathbb{P}(X)^N \times \mathbb{R}^N$ such that

$$\mathcal{D}(\vec{\mu}, \vec{\alpha}, f, k) + \epsilon_1 \geq \|f\|_k,$$

and w.l.o.g. we can assume that $\epsilon_1 > 0$ is small enough so that $\mathcal{D}(\vec{\mu}, \vec{\alpha}, f, k) > 0$. The latter implies that $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)$, $\mathcal{W}(\vec{\mu}, \vec{\alpha}, k) > 0$, so defining

$$\begin{aligned}\epsilon_\alpha &= \frac{\alpha - 1}{\alpha}\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)\\ \epsilon_\beta &= (1/\beta - 1)\mathcal{W}(\vec{\mu}, \vec{\alpha}, k)\end{aligned}$$

we get $\epsilon_\alpha, \epsilon_\beta > 0$. For each $n = 1, \ldots, N$, choose $\vec{x}_n^{[M]} \in X^M$ such that $\vec{x}_n^{[M]} \xrightarrow{d_{\mathrm{KR}}} \mu_n$ for $M \to \infty$. Choose now $L_1 \in \mathbb{N}$ such that for all $\ell \geq L_1$ we get

$$\begin{aligned}|\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) - \mathcal{E}(\vec{\mu}, \vec{\alpha}, f)| &\leq \epsilon_\alpha\\ |\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) - \mathcal{W}(\vec{\mu}, \vec{\alpha}, k)| &\leq \epsilon_\beta.\end{aligned}$$

(cf. also the proof of Theorem 9.3.1) and $\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k^{[M_\ell]}) > 0$. We then get

$$\begin{aligned}\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) &\leq \alpha\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell})\\ \mathcal{W}(\vec{\mu}, \vec{\alpha}, k) &\geq \beta\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k^{[M_\ell]}),\end{aligned}$$

so altogether

$$\frac{\mathcal{E}(\vec{\mu},\vec{\alpha},f)}{\mathcal{W}(\vec{\mu},\vec{\alpha},k)} \leq \frac{\alpha\mathcal{E}(\vec{X}^{[M_\ell]},\vec{\alpha},f_{M_\ell})}{\beta\mathcal{W}(\vec{X}^{[M_\ell]},\vec{\alpha},k_{M_\ell})}.$$

Using Theorem 9.4.1 again leads to

$$\frac{\alpha\mathcal{E}(\vec{X}^{[M_\ell]},\vec{\alpha},f_M)}{\beta\mathcal{W}(\vec{X}^{[M_\ell]},\vec{\alpha},k_{M_\ell})} = \mathcal{D}(\vec{X}^{[M_\ell]},\vec{\alpha},f_{M_\ell},k_{M_\ell}) \leq \|f_{M_\ell}\|_{M_\ell}.$$

Finally, let L_2 such that for all $\ell \geq L_2$ we have $\|f_{M_\ell}\|_{M_\ell} \leq C + \epsilon_2$. For $\ell \geq L_1, L_2$ we then get

$$\begin{aligned}
C < \|f\|_k &\leq \mathcal{D}(\vec{\mu},\vec{\alpha},f,k) + \epsilon_1 \\
&= \frac{\mathcal{E}(\vec{\mu},\vec{\alpha},f)}{\mathcal{W}(\vec{\mu},\vec{\alpha},k)} + \epsilon_1 \\
&\leq \frac{\alpha\mathcal{E}(\vec{X}^{[M_\ell]},\vec{\alpha},f_{M_\ell})}{\beta\mathcal{W}(\vec{X}^{[M_\ell]},\vec{\alpha},k_{M_\ell})} + \epsilon_1 \\
&\leq \frac{\alpha}{\beta}\|f_{M_\ell}\|_{M_\ell} + \epsilon_1 \\
&\leq \frac{\alpha}{\beta}C + \frac{\alpha}{\beta}\epsilon_2 + \epsilon_1.
\end{aligned}$$

Since $\epsilon_1, \epsilon_2 > 0$ and $\alpha > 1$, $\beta \in (0,1)$ were arbitrary, this implies that

$$C < \|f\|_k \leq C,$$

a contradiction. □

Lemma 9.3.3. Let $f \in H_k$. Then there exist $f_M \in H_M$, $M \in \mathbb{N}_+$, such that $\lim_{M\to\infty} \sup_{\vec{x}\in X^M} |f_M(\vec{x}) - f(\hat{\mu}[\vec{x}])| = 0$, and

$$\limsup_{M\to\infty} \|f_M\|_M \leq \|f\|_k. \tag{9.7}$$

Proof. Let $f \in H_k$ be arbitrary and choose $(\epsilon_n)_n \subseteq \mathbb{R}_{>0}$ with $\epsilon_n \searrow 0$.

Step 1 For each $n \in \mathbb{N}$ choose

$$f_n^{\text{pre}} = \sum_{\ell=1}^{L_n} \alpha_\ell^{(n)} k(\cdot, \mu_\ell^{(n)}) \in H_k^{\text{pre}},$$

where $\alpha_1^{(n)}, \dots, \alpha_{L_n}^{(n)} \in \mathbb{R}$ and $\mu_1^{(n)}, \dots, \mu_{L_n}^{(n)} \in \mathbb{P}(X)$, with

$$\|f - f_n^{\mathrm{pre}}\|_k \leq \frac{\epsilon_n}{3\sqrt{C_k}}$$

and $\|f_n^{\mathrm{pre}}\|_k \leq \|f\|_k$. To see that such a sequence of functions exists, choose some sequence $(\bar{f}_n)_n \in H_k^{\mathrm{pre}}$ with $\bar{f}_n = \sum_{\ell=1}^{\bar{L}_n} \bar{\alpha}_\ell^{(n)} k(\cdot, \bar{\mu}_\ell^{(n)})$, where $\bar{\alpha}_\ell^{(n)} \in \mathbb{R}$, $\bar{\mu}_\ell^{(n)} \in \mathbb{P}(X)$, with $\bar{f}_n \xrightarrow{\|\cdot\|_k} f$ (exists since H_k^{pre} is dense in H_k). Define now for $n \in \mathbb{N}$

$$\bar{H}_n = \mathrm{span}\{k(\cdot, \bar{\mu}_\ell^{(m)}) \mid m = 1, \dots, n,\ \ell = 1, \dots, \bar{L}_m\}$$

and $\hat{f}_n = P_{\bar{H}_n} f$, where $P_{\bar{H}_n}$ is the orthogonal projection onto $\bar{H}_n$. Then $\bar{H}_n \subseteq H_k^{\mathrm{pre}}$, $\|\hat{f}_n\|_k = \|P_{\bar{H}_n} f\|_k \leq \|f\|_k$ and $\|f - \hat{f}_n\|_k \leq \|f - \bar{f}_n\|_k \to 0$ (since $\hat{f}_n = P_{\bar{H}_n} f$ is the orthogonal projection of f onto $\bar{H}_n$ and $\bar{f}_n \in \bar{H}_n$), hence $\hat{f}_n \xrightarrow{\|\cdot\|_k} f$. We can now choose $(f_n^{\mathrm{pre}})_n$ as a subsequence of $(\hat{f}_n)_n$.

Next, for all $n \in \mathbb{N}$ and $\ell = 1, \dots, L_n$ choose $\vec{x}_M^{(n,\ell)} \in X^M$ with $\hat{\mu}[\vec{x}_M^{(n,\ell)}] \xrightarrow{d_{\mathrm{KR}}} \mu_\ell^{(n)}$ for $M \to \infty$. Furthermore, for all $n \in \mathbb{N}$ choose $M_n \in \mathbb{N}$ such that for all $M \geq M_n$ and $\ell = 1, \dots, L_n$ we have

$$d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_M^{(n,\ell)}], \mu_\ell^{(n)}) \leq \min\left\{ \frac{\epsilon_n}{3\left(1 + L_k \sum_{\ell'=1}^{L_n} |\alpha_{\ell'}^{(n)}|\right)}, \frac{\epsilon_n^2}{2\left(1 + 2L_k \sum_{i,j=1}^{L_n} |\alpha_i^{(n)}||\alpha_j^{(n)}|\right)} \right\}$$

and

$$\sup_{\vec{x},\vec{x}' \in X^M} |k_M(\vec{x}, \vec{x}') - k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])| \leq \min\left\{ \frac{\epsilon_n}{3\left(1 + \sum_{\ell'=1}^{L_n} |\alpha_{\ell'}^{(n)}|\right)}, \frac{\epsilon_n^2}{2\left(1 + \sum_{i,j=1}^{L_n} |\alpha_i^{(n)}||\alpha_j^{(n)}|\right)} \right\}.$$

W.l.o.g. we can assume that $(M_n)_n$ is strictly increasing. For $M \in \mathbb{N}$, let $n(M)$ be the largest integer such that $M_{n(M)} \leq M$ and define

$$\hat{f}_M^{\mathrm{pre}} = \sum_{\ell=1}^{L_{n(M)}} \alpha_\ell^{(n(M))} k(\cdot, \hat{\mu}[\vec{x}_M^{(n(M),\ell)}]) \in H_k^{\mathrm{pre}}$$
$$f_M = \sum_{\ell=1}^{L_{n(M)}} \alpha_\ell^{(n(M))} k_M(\cdot, \vec{x}_M^{(n(M),\ell)}) \in H_M^{\mathrm{pre}}.$$

Step 2 We now show that $f_M \xrightarrow{\mathcal{P}_1} f$. For this, let $\epsilon > 0$ be arbitrary and $n_\epsilon \in \mathbb{N}$

such that $\epsilon_n \leq \epsilon$. Let now $M \geq M_{n_\epsilon}$ (note that this implies that $n(M) \geq n_\epsilon$ and hence $\epsilon_{n(M)} \leq \epsilon_n$) and $\vec{x} \in X^M$, then we have

$$|f(\hat{\mu}[\vec{x}]) - f_M(\vec{x})| \leq \underbrace{|f(\hat{\mu}[\vec{x}]) - f_{n(M)}(\hat{\mu}[\vec{x}])|}_{=I} + \underbrace{|f_{n(M)}(\hat{\mu}[\vec{x}]) - \hat{f}_M^{\mathrm{pre}}(\hat{\mu}[\vec{x}])|}_{=II} + \underbrace{|\hat{f}_M^{\mathrm{pre}}(\hat{\mu}[\vec{x}]) - f_M(\vec{x})|}_{=III}$$

We continue with

$$\begin{aligned} I &= |f(\hat{\mu}[\vec{x}]) - f_{n(M)}(\hat{\mu}[\vec{x}])| \\ &= |\langle f - f_{n(M)}, k(\cdot, \hat{\mu}[\vec{x}])\rangle_k| \\ &\leq \|f - f_{n(M)}\|_k \|k(\cdot, \hat{\mu}[\vec{x}])\|_k \\ &= \|f - f_{n(M)}\|_k \sqrt{k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}])} \\ &\leq \frac{\epsilon_{n(M)}}{3\sqrt{C_k}} \sqrt{C_k} \end{aligned}$$

where we first used the reproducing property of k, then Cauchy-Schwarz, again the reproducing property of k, and finally the choice $f_{n(M)}$ and the boundedness of k.

Next,

$$\begin{aligned} II &= |f_{n(M)}(\hat{\mu}[\vec{x}]) - \hat{f}_M^{\mathrm{pre}}(\hat{\mu}[\vec{x}])| \\ &= \left| \sum_{\ell=1}^{L_{n(M)}} \alpha_\ell^{(n(M))} k(\hat{\mu}[\vec{x}], \mu_\ell^{(n(M))}) - \sum_{\ell=1}^{L_{n(M)}} \alpha_\ell^{(n(M))} k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}_M^{(n(M),\ell)}]) \right| \\ &\leq \sum_{\ell=1}^{L_{n(M)}} \left| \alpha_\ell^{(n(M))} \right| |k(\hat{\mu}[\vec{x}], \mu_\ell^{(n(M))}) - k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}_M^{(n(M),\ell)}])| \\ &\leq L_k \sum_{\ell=1}^{L_{n(M)}} \left| \alpha_\ell^{(n(M))} \right| d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_M^{(n(M),\ell)}], \mu_\ell^{(n(M))}) \\ &\leq \frac{\epsilon_{n(M)}}{3}, \end{aligned}$$

where we used the triangle inequality, the Lipschitz continuity of k, and then the choice of the sequence $(M_n)_n$.

Finally,

$$\begin{aligned} III &= |\hat{f}_M^{\mathrm{pre}}(\hat{\mu}[\vec{x}]) - f_M(\vec{x})| \\ &= \left| \sum_{\ell=1}^{L_{n(M)}} \alpha_\ell^{(n(M))} k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}_M^{(n(M),\ell)}]) - \sum_{\ell=1}^{L_{n(M)}} \alpha_\ell^{(n(M))} k_M(\vec{x}, \vec{x}_M^{(n(M),\ell)}) \right| \\ &\leq \sum_{\ell=1}^{L_{n(M)}} \left| \alpha_\ell^{(n(M))} \right| |k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}_M^{(n(M),\ell)}]) - k_M(\vec{x}, \vec{x}_M^{(n(M),\ell)})| \\ &\leq \frac{\epsilon_{n(M)}}{3}, \end{aligned}$$

where the triangle inequality has been used in the first step and then again the choice of the sequence $(M_n)_n$.

Altogether,

$$\begin{aligned} |f(\hat{\mu}[\vec{x}]) - f_M(\vec{x})| &\leq I + II + III \\ &\leq \frac{\epsilon_{n(M)}}{3} + \frac{\epsilon_{n(M)}}{3} + \frac{\epsilon_{n(M)}}{3} \\ &\leq \epsilon, \end{aligned}$$

establishing $f_M \xrightarrow{\mathcal{P}_1} f$.

Step 3 We now show $\limsup_{M\to\infty} \|f_M\|_M \leq \|f\|_k$. Let $\epsilon > 0$ be arbitrary and $n_\epsilon \in \mathbb{N}$ such that $\epsilon_n \leq \epsilon$ and let $M \geq M_{n_\epsilon}$. We have

$$\begin{aligned} \|f_M\|_M^2 &= \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))} \alpha_{\ell'}^{(n(M))} k_M(\vec{x}_M^{(n(M),\ell')}, \vec{x}_M^{(n(M),\ell')}) \\ &\leq \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))} \alpha_{\ell'}^{(n(M))} k(\mu_{\ell'}^{(n(M))}, \mu_\ell^{(n(M))}) \ + |R_1| + |R_2| \\ &= \|f_{n(M)}^{\mathrm{pre}}\|_k^2 + R_1 + R_2 \\ &\leq \|f\|_k^2 + R_1 + R_2. \end{aligned}$$

with remainder terms

$$R_1 = \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))}\alpha_{\ell'}^{(n(M))} k_M(\vec{x}_M^{(n(M),\ell')}, \vec{x}_M^{(n(M),\ell')}) - \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))}\alpha_{\ell'}^{(n(M))} k(\hat{\mu}[\vec{x}_M^{(n(M),\ell')}], \hat{\mu}[\vec{x}_M^{(n(M),\ell')}])$$

$$R_2 = \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))}\alpha_{\ell'}^{(n(M))} k(\hat{\mu}[\vec{x}_M^{(n(M),\ell')}], \hat{\mu}[\vec{x}_M^{(n(M),\ell')}]) - \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))}\alpha_{\ell'}^{(n(M))} k(\mu_{\ell'}^{(n(M))}, \mu_\ell^{(n(M))})$$

We now bound these terms, so that

$$\begin{aligned} R_1 &= \left| \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))}\alpha_{\ell'}^{(n(M))} k_M(\vec{x}_M^{(n(M),\ell')}, \vec{x}_M^{(n(M),\ell')}) - \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))}\alpha_{\ell'}^{(n(M))} k(\hat{\mu}[\vec{x}_M^{(n(M),\ell')}], \hat{\mu}[\vec{x}_M^{(n(M),\ell')}]) \right| \\ &\leq \sum_{\ell,\ell'=1}^{L_{n(M)}} |\alpha_\ell^{(n(M))}||\alpha_{\ell'}^{(n(M))}| |k_M(\vec{x}_M^{(n(M),\ell')}, \vec{x}_M^{(n(M),\ell')}) - k(\hat{\mu}[\vec{x}_M^{(n(M),\ell')}], \hat{\mu}[\vec{x}_M^{(n(M),\ell')}])| \\ &\leq \frac{\epsilon_{n(M)}^2}{2}, \end{aligned}$$

and

$$\begin{aligned} R_2 &= \left| \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))}\alpha_{\ell'}^{(n(M))} k(\hat{\mu}[\vec{x}_M^{(n(M),\ell')}], \hat{\mu}[\vec{x}_M^{(n(M),\ell')}]) - \sum_{\ell,\ell'=1}^{L_{n(M)}} \alpha_\ell^{(n(M))}\alpha_{\ell'}^{(n(M))} k(\mu_{\ell'}^{(n(M))}, \mu_\ell^{(n(M))}) \right| \\ &\leq \sum_{\ell,\ell'=1}^{L_{n(M)}} |\alpha_\ell^{(n(M))}||\alpha_{\ell'}^{(n(M))}| |k(\hat{\mu}[\vec{x}_M^{(n(M),\ell')}], \hat{\mu}[\vec{x}_M^{(n(M),\ell')}]) - k(\mu_{\ell'}^{(n(M))}, \mu_\ell^{(n(M))})| \\ &\leq L_k \sum_{\ell,\ell'=1}^{L_{n(M)}} |\alpha_\ell^{(n(M))}||\alpha_{\ell'}^{(n(M))}| \left(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_M^{(n(M),\ell)}], \mu_\ell^{(n(M))}) + d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_M^{(n(M),\ell')}], \mu_{\ell'}^{(n(M))}) \right) \\ &\leq \frac{\epsilon_{n(M)}^2}{2}. \end{aligned}$$

Altogether,

$$\begin{aligned} \|f_M\|_M^2 &\leq \|f\|_k^2 + |R_1| + |R_2| \\ &\leq \|f\|_k^2 + \frac{\epsilon_{n(M)}^2}{2} + \frac{\epsilon_{n(M)}^2}{2} \\ &\leq \|f\|_k^2 + \epsilon^2, \end{aligned}$$

so $\|f_M\|_M \leq \|f\|_k + \epsilon$ for all $M \geq M_{n_\epsilon}$, and since $\epsilon > 0$ was arbitrary, we finally get

$\limsup_{M\to\infty} \|f_M\|_M \leq \|f\|_k$. □

Finally, we can now provide the proof for the central Theorem 9.3.1.

Proof of Theorem 9.3.1. The first statement is part of Lemma 9.3.3. Let us turn to the second statement. Proposition 9.1.4 ensures that $(f_M)_M$ is a bounded, equicontinuous sequence, so Proposition 8.2.2 ensures the existence of a subsequence $(f_{M_\ell})_\ell$ and a continuous function $f : \mathbb{P}(X) \to \mathbb{R}$ with $f_{M_\ell} \xrightarrow{\mathcal{P}_1} f$, so we only have to ensure that $f \in H_k$ with $\|f\|_k \leq B$. For this, we use the characterization of RKHS functions from Theorem 9.4.1. In particular, we will utilize the notation introduced there.

Step 1 Let $(\vec{\mu}, \vec{\alpha}) \in \mathbb{P}(X)^N \times \mathbb{R}^N$. We show that if $\mathcal{W}(\vec{\mu}, \vec{\alpha}, k) = 0$, then $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) = 0$.

Assume that $\mathcal{W}(\vec{\mu}, \vec{\alpha}, k) = 0$. If $B = 0$, then $f_M \equiv 0$ and $f_{M_\ell} \xrightarrow{\mathcal{P}_1} f$ implies that $f \equiv 0$, so the claim is clear in this case. Assume now $B > 0$, let $\epsilon > 0$ be arbitary and for $n = 1, \dots, N$, choose sequences $\vec{x}_n^{[M]} \in X^M$ such that $\vec{x}_n^{[M]} \xrightarrow{d_{\mathrm{KR}}} \mu_n$ for $M \to \infty$. For convenience, define $\vec{X}^{[M]} = \begin{pmatrix} \vec{x}_1^{[M]} & \cdots & \vec{x}_N^{[M]} \end{pmatrix}$. Choose now $\ell_\epsilon \in \mathbb{N}$ such that for all $M \geq M_{\ell_\epsilon}$ we get $\mathcal{W}(\vec{X}^{[M]}, \vec{\alpha}, k_M) \leq \epsilon/B$. This is possible since $k_M \xrightarrow{\mathcal{P}_1} k$ together with the continuity of k_M and k as well as $\vec{x}_n^{[M]} \xrightarrow{d_{\mathrm{KR}}} \mu_n$ for $M \to \infty$ and all $n = 1, \dots, N$ implies that $\mathcal{W}(\vec{X}^{[M]}, \vec{\alpha}, k_M) \to \mathcal{W}(\vec{\mu}, \vec{\alpha}, k) = 0$. Let now $\ell \geq \ell_\epsilon$ be arbitrary and observe that $f_M \in H_M$ implies $\mathcal{N}(f_M, k_M) < \infty$ according to Theorem 9.4.1, so in particular $\mathcal{D}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}, k_{M_\ell}) < \infty$.

If $\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) = 0$, then we get that $\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) = 0 \leq \epsilon$ since $\mathcal{D}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}, k_{M_\ell}) < \infty$, which implies by definition that $\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) = 0$.

If $\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) > 0$, then we have

$$\frac{\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell})}{\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell})} = \mathcal{D}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}, k_{M_\ell}) \leq \mathcal{N}(f_{M_\ell}, k_{M_\ell}) = \|f_{M_\ell}\|_{M_\ell} \leq B,$$

which implies

$$\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) \leq B\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) \leq \epsilon.$$

Since $f_{M_\ell} \xrightarrow{\mathcal{P}_1} f$ together with the continuity of f_M and f as well as $\vec{x}_n^{[M]} \xrightarrow{d_{\mathrm{KR}}} \mu_n$ implies that $\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) \to \mathcal{E}(\vec{\mu}, \vec{\alpha}, f)$, we get that $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) \leq \epsilon$, and since $\epsilon > 0$ was arbitrary we arrive at $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) \leq 0$.

Assume now that $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) < 0$. This implies that there exist $\delta > 0$ and $\ell_\delta \in \mathbb{N}$ such that for all $\ell \geq \ell_\delta$ we have $\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) \leq -\delta < 0$, since $\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) \to \mathcal{E}(\vec{\mu}, \vec{\alpha}, f)$. Let $\ell \geq \ell_\delta$, then we get that $\mathcal{E}(\vec{X}^{[M_\ell]}, -\vec{\alpha}, f_{M_\ell}) \geq \delta > 0$ and we have $\mathcal{W}(\vec{X}^{[M_\ell]}, -\vec{\alpha}, k_{M_\ell}) = \mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) > 0$. We can then continue with

$$\begin{aligned}\frac{\delta}{\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell})} &\leq \frac{\mathcal{E}(\vec{X}^{[M_\ell]}, -\vec{\alpha}, f_{M_\ell})}{\mathcal{W}(\vec{X}^{[M_\ell]}, -\vec{\alpha}, k_{M_\ell})} \\ &\leq \mathcal{D}(\vec{X}^{[M_\ell]}, -\vec{\alpha}, f_{M_\ell}, k_{M_\ell}) \\ &\leq \mathcal{N}(f_{M_\ell}, k_{M_\ell}) \\ &= \|f_{M_\ell}\|_{M_\ell} \leq B,\end{aligned}$$

which implies that $\mathcal{W}(\vec{X}^{[M_\ell]}, -\vec{\alpha}, k_{M_\ell}) = \mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) \geq \delta/B$. But since $\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) \to \mathcal{W}(\vec{\mu}, \vec{\alpha}, k)$, this implies that $\mathcal{W}(\vec{\mu}, \vec{\alpha}, k) \geq \delta/B > 0$, a contradiction. Altogether, $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) = 0$.

Step 2 Let $(\vec{\mu}, \vec{\alpha}) \in \mathbb{P}(X)^N \times \mathbb{R}^N$. If $\mathcal{W}(\vec{\mu}, \vec{\alpha}, k) > 0$ and $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) > 0$, then

$$\frac{\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)}{\mathcal{W}(\vec{\mu}, \vec{\alpha}, k)} \leq B.$$

To show this, let $\alpha > 1$ and $\beta \in (0, 1)$ be arbitrary. Define

$$\begin{aligned}\epsilon_\alpha &= \frac{\alpha - 1}{\alpha} \mathcal{E}(\vec{\mu}, \vec{\alpha}, f) \\ \epsilon_\beta &= (1/\beta - 1) \mathcal{W}(\vec{\mu}, \vec{\alpha}, k)\end{aligned}$$

and observe that $\epsilon_\alpha, \epsilon_\beta > 0$. Furthermore, for all $n = 1, \ldots, N$ choose a sequence $\vec{x}_n^{[M]} \in X^M$ such that $\vec{x}_n^{[M]} \xrightarrow{d_{\mathrm{KR}}} \mu_n$ for $M \to \infty$, and define $\vec{X}^{[M]} = \begin{pmatrix} \vec{x}_1^{[M]} & \cdots & \vec{x}_N^{[M]} \end{pmatrix}$. Choose $\ell_\epsilon \in \mathbb{N}_+$ such that for all $\ell \geq \ell_\epsilon$ we have

$$\begin{aligned}|\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) - \mathcal{E}(\vec{\mu}, \vec{\alpha}, f)| &\leq \epsilon_\alpha \\ |\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) - \mathcal{W}(\vec{\mu}, \vec{\alpha}, k)| &\leq \epsilon_\beta\end{aligned}$$

and $\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) > 0$. Such an ℓ_ϵ exists because $k_M \xrightarrow{\mathcal{P}_1} k$ together with the continuity of k_M and k as well as the convergence of $\vec{x}_n^{[M]}$ to μ_n imply that $\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) \to \mathcal{W}(\vec{\mu}, \vec{\alpha}, k)$, and $f_{M_\ell} \xrightarrow{\mathcal{P}_1} f$ together with the continuity of f_M

and f imply that $\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) \to \mathcal{E}(\vec{\mu}, \vec{\alpha}, f)$.

Let now $\ell \geq \ell_\epsilon$ be arbitrary. By definition of ϵ_α we get $\alpha\epsilon_\alpha \leq (\alpha-1)\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)$, which in turn leads to

$$\begin{aligned}\epsilon_\alpha &\leq \epsilon_\alpha - \alpha\epsilon_\alpha + (\alpha-1)\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)\\ &= -(\alpha-1)\epsilon_\alpha + (\alpha-1)\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)\\ &= (\alpha-1)(\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) - \epsilon_\alpha)\\ &\leq (\alpha-1)\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}),\end{aligned}$$

where we used in the last inequality that $\alpha - 1 > 0$ and by choice of ℓ_ϵ we have $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) \leq \mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) + \epsilon_\alpha$. We can then continue with

$$\begin{aligned}\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) &\leq \mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) + \epsilon_\alpha\\ &\leq \mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}) + (\alpha-1)\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell})\\ &= \alpha\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell}).\end{aligned}$$

Next, by definition of ϵ_β and choice of ℓ_ϵ we find that

$$\begin{aligned}\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell}) &\leq \mathcal{W}(\vec{\mu}, \vec{\alpha}, k) + \epsilon_\beta\\ &= \mathcal{W}(\vec{\mu}, \vec{\alpha}, k) + (1/\beta - 1)\mathcal{W}(\vec{\mu}, \vec{\alpha}, k)\\ &= (1/\beta)\mathcal{W}(\vec{\mu}, \vec{\alpha}, k),\end{aligned}$$

hence

$$\frac{1}{\mathcal{W}(\vec{\mu}, \vec{\alpha}, k)} \leq \frac{1}{\beta\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell})}.$$

Combining these results, we get that for all $\ell \geq \ell_\epsilon$

$$\frac{\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)}{\mathcal{W}(\vec{\mu}, \vec{\alpha}, k)} \leq \frac{\alpha}{\beta}\frac{\mathcal{E}(\vec{X}^{[M_\ell]}, \vec{\alpha}, f_{M_\ell})}{\mathcal{W}(\vec{X}^{[M_\ell]}, \vec{\alpha}, k_{M_\ell})} \leq \frac{\alpha}{\beta}\mathcal{N}(f_{M_\ell}, k_{M_\ell}) = \frac{\alpha}{\beta}\|f_{M_\ell}\|_{M_\ell} \leq \frac{\alpha}{\beta}B.$$

Since $\alpha > 1$ and $\beta \in (0,1)$ were arbitrary, this shows that

$$\frac{\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)}{\mathcal{W}(\vec{\mu}, \vec{\alpha}, k)} \leq B.$$

Step 3 Let $(\vec{\mu}, \vec{\alpha}) \in \mathbb{P}(X)^N \times \mathbb{R}^N$ be arbitrary. If $\mathcal{W}(\vec{\mu}, \vec{\alpha}, k) = 0$, then we get from Step 1 that $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) = 0 \leq B$. Assume now $\mathcal{W}(\vec{\mu}, \vec{\alpha}, k) > 0$. If $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) = 0$, then again $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) = 0 \leq B$. If $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) > 0$, then Step 2 ensures that

$$\frac{\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)}{\mathcal{W}(\vec{\mu}, \vec{\alpha}, k)} = \mathcal{D}(\vec{\mu}, \vec{\alpha}, f, k) \leq B.$$

Finally, if $\mathcal{E}(\vec{\mu}, \vec{\alpha}, f) < 0$, then again

$$\frac{\mathcal{E}(\vec{\mu}, \vec{\alpha}, f)}{\mathcal{W}(\vec{\mu}, \vec{\alpha}, k)} = \mathcal{D}(\vec{\mu}, \vec{\alpha}, f, k) < 0 \leq B.$$

Altogether, we get that $\mathcal{D}(\vec{\mu}, \vec{\alpha}, f, k) \leq B$. Since $(\vec{\mu}, \vec{\alpha})$ was arbitrary, maximization leads to $\mathcal{N}(f, k) \leq B < \infty$, hence $f \in H_k$ and $\|f\|_k = \mathcal{N}(f, k) \leq B$. □

9.4. Technical background: A characterization of RKHS functions

Here we recall the following characterization of RKHS functions from [17, Section I.4]. Let $\mathcal{X} \neq \emptyset$ be arbitrary. For $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric and positive semidefinite and some $f \in \mathbb{R}^{\mathcal{X}}$ as well as $N \in \mathbb{N}_+$, $\vec{x} \in \mathcal{X}^N$, $\vec{\alpha} \in \mathbb{R}^N$ define

$$\begin{aligned}\mathcal{E}(\vec{x}, \vec{\alpha}, f) &= \sum_{n=1}^{N} \alpha_n f(x_n)\\ \mathcal{W}(\vec{x}, \vec{\alpha}, k) &= \sqrt{\sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_j, x_i)},\end{aligned}$$

where we might omit some arguments if they are clear. Furthermore, define

$$\mathcal{D}(\vec{x}, \vec{\alpha}, f, k) = \begin{cases} \frac{\mathcal{E}(\vec{x}, \vec{\alpha}, f)}{\mathcal{W}(\vec{x}, \vec{\alpha}, k)} & \text{if } \mathcal{E}(\vec{x}, \vec{\alpha}, f) \neq 0, \mathcal{W}(\vec{x}, \vec{\alpha}, k) \neq 0 \\ 0 & \text{if } \mathcal{E}(\vec{x}, \vec{\alpha}, f) = \mathcal{W}(\vec{x}, \vec{\alpha}, k) = 0 \\ \infty & \text{if } \mathcal{E}(\vec{x}, \vec{\alpha}, f) \neq 0, \mathcal{W}(\vec{x}, \vec{\alpha}, k) = 0 \end{cases}$$

and

$$\mathcal{N}(f, k) = \sup_{\substack{(\vec{x}, \vec{\alpha}) \in \mathcal{X}^N \times \mathbb{R}^N \\ N \in \mathbb{N}_+}} \mathcal{D}(\vec{x}, \vec{\alpha}, f, k).$$

We collect now some simple facts that will be used repeatedly.

Let $\vec{x} \in \mathcal{X}^N$, $\vec{\alpha} \in \mathbb{R}^N$, $N \in \mathbb{N}_+$, be arbitrary, and define

$$f = \sum_{n=1}^{N} \alpha_n k(\cdot, x_n) \in H_k^{\mathrm{pre}}.$$

1. By construction, $\mathcal{W}(\vec{x}, \vec{\alpha}, k) \in \mathbb{R}_{\geq 0}$ (recall that k is positive semidefinite).

2. Since $f \in H_k^{\mathrm{pre}}$, its RKHS norm has an explicit form and we find

$$\|f\|_k = \sqrt{\sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_j, x_i)} = \mathcal{W}(\vec{x}, \vec{\alpha}, k).$$

   This also implies that $f \equiv 0$ if and only if $\mathcal{W}(\vec{x}, \vec{\alpha}, k) = 0$.

3. If $\mathcal{W}(\vec{x}, \vec{\alpha}, k) > 0$, then

$$\begin{aligned}
\mathcal{D}(\vec{x}, \vec{\alpha}, f, k) &= \frac{\mathcal{E}(\vec{x}, \vec{\alpha}, f)}{\mathcal{W}(\vec{x}, \vec{\alpha}, k)} \\
&= \frac{\sum_{i=1}^{N} \alpha_i f(x_i)}{\sqrt{\sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_j, x_i)}} \\
&= \frac{\sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_j, x_i)}{\sqrt{\sum_{i,j=1}^{N} \alpha_i \alpha_j k(x_j, x_i)}} \\
&= \frac{\mathcal{W}(\vec{x}, \vec{\alpha}, k)^2}{\mathcal{W}(\vec{x}, \vec{\alpha}, k)} = \mathcal{W}(\vec{x}, \vec{\alpha}, k).
\end{aligned}$$

We can now state the characterization result.

Theorem 9.4.1. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel and $f \in \mathbb{R}^{\mathcal{X}}$. Then $f \in H_k$ if and only if $\mathcal{N}(f, k) < \infty$. If $f \in H_k$, then $\|f\|_k = \mathcal{N}(f, k)$.

For convenience, we provide a full self-contained proof of this result.

Proof. **Step 1** First, we show that for $f \in H_k$, we have $\|f\|_k = \mathcal{N}(f, k)$.

$\mathcal{N}(f,k) \leq \|f\|_k$*:* Let $N \in \mathbb{N}_+$ and $(\vec{x},\vec{\alpha}) \in \mathcal{X}^N \times \mathbb{R}^N$ be arbitrary. Observe that

$$\begin{aligned}\mathcal{E}(\vec{x},\vec{\alpha},f) &= \sum_{n=1}^N \alpha_n f(x_n)\\ &= \sum_{n=1}^N \alpha_n \langle f, k(\cdot,x_n)\rangle_k\\ &= \langle f, \sum_{n=1}^N \alpha_n k(\cdot,x_n)\rangle_k\\ &\leq \|f\|_k \left\|\sum_{n=1}^N \alpha_n k(\cdot,x_n)\right\|_k\\ &= \|f\|_k \mathcal{W}(\vec{x},\vec{\alpha},k).\end{aligned}$$

If $\mathcal{W}(\vec{x},\vec{\alpha},k) = \|\sum_{n=1}^N \alpha_n k(\cdot,x_n)\|_k = 0$, then $\sum_{n=1}^N \alpha_n k(\cdot,x_n) = 0_{H_k}$, hence $\mathcal{E}(\vec{x},\vec{\alpha},f) = \langle f, 0_{H_k}\rangle_k = 0$ and by definition $\mathcal{D}(\vec{x},\vec{\alpha},f,k) = 0 \leq \|f\|_k$.

If $\mathcal{W}(\vec{x},\vec{\alpha},k) > 0$, we can rearrange to get

$$\frac{\mathcal{E}(\vec{x},\vec{\alpha},f)}{\mathcal{W}(\vec{x},\vec{\alpha},k)} = \mathcal{D}(\vec{x},\vec{\alpha},f,k) \leq \|f\|_k.$$

Since $(\vec{x},\vec{\alpha})$ was arbitrary, we find that $\mathcal{N}(\vec{x},\vec{\alpha},f,k) \leq \|f\|_k$.

$\mathcal{N}(f,k) \geq \|f\|_k$*:* Let $\epsilon > 0$ and choose $f_\epsilon = \sum_{n=1}^N \alpha_n k(\cdot,x_n) \in H_k^{\text{pre}}$ such that $\|f - f_\epsilon\|_k < \epsilon$. If $\mathcal{W}(\vec{x},\vec{\alpha},k) = \|f_\epsilon\|_k = 0$, then $f_\epsilon = 0_{H_k}$ and hence $\mathcal{E}(\vec{x},\vec{\alpha},f) = \langle f, f_\epsilon\rangle_k = \langle f, 0_{H_k}\rangle_k = 0$. By definition, this then shows

$$\mathcal{D}(\vec{x},\vec{\alpha},f) = 0 = \|f_\epsilon\|_k \geq \|f\|_k - \epsilon.$$

Before we continue, note that for all $f_1, f_2 \in H_k$ we have

$$\begin{aligned}|\mathcal{E}(\vec{x},\vec{\alpha},f_1) - \mathcal{E}(\vec{x},\vec{\alpha},f_2)| &= \left|\sum_{n=1}^N \alpha_n (f_1(x_n) - f_2(x_n))\right| = \left|\sum_{n=1}^N \alpha_n \langle f_1 - f_2, k(\cdot,x_n)\rangle_k\right|\\ &= \left|\langle f_1 - f_2, \sum_{n=1}^N \alpha_n k(\cdot,x_n)\rangle_k\right|\\ &\leq \|f_1 - f_2\|_k \|f_\epsilon\|_k.\end{aligned}$$

Assume now that $\mathcal{W}(\vec{x}, \vec{\alpha}, k) > 0$, then we get

$$\begin{aligned}
\mathcal{D}(\vec{x}, \vec{\alpha}, f, k) &= \frac{\mathcal{E}(\vec{x}, \vec{\alpha}, f)}{\mathcal{W}(\vec{x}, \vec{\alpha}, k)} \\
&\geq \frac{\mathcal{E}(\vec{x}, \vec{\alpha}, f_\epsilon)}{\mathcal{W}(\vec{x}, \vec{\alpha}, k)} - \frac{\|f - f_\epsilon\|_k \|f_\epsilon\|_k}{\mathcal{W}(\vec{x}, \vec{\alpha}, k)} \\
&\geq \frac{\mathcal{E}(\vec{x}, \vec{\alpha}, f_\epsilon)}{\mathcal{W}(\vec{x}, \vec{\alpha}, k)} - \frac{\epsilon \|f_\epsilon\|_k}{\mathcal{W}(\vec{x}, \vec{\alpha}, k)} \\
&= \mathcal{W}(\vec{x}, \vec{\alpha}, k) - \epsilon \\
&= \|f_\epsilon\|_k - \epsilon \\
&\geq \|f\|_k - 2\epsilon
\end{aligned}$$

Altogether, by definition of $\mathcal{N}(f, k)$, we get that

$$\mathcal{N}(f, k) \geq \mathcal{D}(\vec{x}, \vec{\alpha}, f, k) \geq \|f\|_k - 2\epsilon.$$

Since $\epsilon > 0$ was arbitrary, we find that $\mathcal{N}(f, k) \geq \|f\|_k$.

Step 2 Let $f \in \mathbb{R}^{\mathcal{X}}$ be arbitrary. We show that if $\mathcal{N}(f, k) < \infty$, then

$$\begin{aligned}
\ell_f : H_k^{\text{pre}} &\to \mathbb{R} \\
\sum_{n=1}^{N} \alpha_n k(\cdot, x_n) &\mapsto \sum_{n=1}^{N} \alpha_n f(x_n)
\end{aligned}$$

is a well-defined, linear and continuous (w.r.t. $\|\cdot\|_k$) map.

To establish the *well-posedness*, let $(\vec{x}, \vec{\alpha}) \in \mathcal{X}^N \times \mathbb{R}^N$ and $(\vec{y}, \vec{\beta}) \in \mathcal{X}^M \times \mathbb{R}^M$ such that

$$\sum_{n=1}^{N} \alpha_n k(\cdot, x_n) = \sum_{m=1}^{M} \beta_m k(\cdot, y_m) \in H_k^{\text{pre}}.$$

This implies that

$$\sum_{n=1}^{N} \alpha_n k(\cdot, x_n) + \sum_{m=1}^{M} (-\beta_m) k(\cdot, y_m) = 0_{H_k}$$

and hence $\mathcal{W}((\vec{x}, \vec{y}), (\vec{\alpha}, -\vec{\beta}), k) = \| \sum_{n=1}^{N} \alpha_n k(\cdot, x_n) + \sum_{m=1}^{M} (-\beta_m) k(\cdot, y_m) \|_k = 0$.

Assume now that

$$\sum_{n=1}^{N} \alpha_n f(x_n) \neq \sum_{m=1}^{m} \beta_m f(x_m),$$

then we get that

$$\sum_{n=1}^{N} \alpha_n f(x_n) + \sum_{m=1}^{m} (-\beta_m) f(x_m) = \mathcal{E}((\vec{x}, \vec{y}), (\vec{\alpha}, -\vec{\beta}), f) \neq 0$$

which by definition implies that $\mathcal{D}((\vec{x}, \vec{y}), (\vec{\alpha}, -\vec{\beta}), f, k) = \infty$ and therefore $\mathcal{N}(f, k) = \infty$, a contradiction.

The *linearity* is then clear. Finally, to show the *continuity*, let $H_k^{\text{pre}} \ni f_0 = \sum_{n=1}^{N} \alpha_n k(\cdot, x_n)$ be arbitrary and set $\vec{x} = \begin{pmatrix} x_1 & \cdots & x_N \end{pmatrix}$, $\vec{\alpha} = \begin{pmatrix} \alpha_1 & \cdots & \alpha_N \end{pmatrix}$, then

$$\begin{aligned} |\ell_f(f_0)| &= \left| \sum_{n=1}^{N} \alpha_n f(x_n) \right| \\ &= |\mathcal{E}(\vec{x}, \vec{\alpha}, f)| \\ &\leq \mathcal{N}(f, k) \mathcal{W}(\vec{x}, \vec{\alpha}, k) \\ &= \mathcal{N}(f, k) \|f_0\|_k. \end{aligned}$$

Since $\mathcal{N}(f, k)$ is finite and independent of f_0, and ℓ_f is a linear map, this shows the continuity of ℓ_f.

Step 3 Let $f \in \mathbb{R}^{\mathcal{X}}$ such that $\mathcal{N}(f, k) < \infty$. Since according to Step 2 ℓ_f is a linear and continuous map on H_k^{pre} and the latter is dense in H_k, there exists a unique linear and continuous extension $\bar{\ell}_f : H_k \to \mathbb{R}$ of ℓ_f. Furthermore, from the Riesz Representation Theorem there exists a unique $\hat{f} \in H_k$ with $\bar{\ell}_f = \langle \cdot, \hat{f} \rangle_k$. For all $x \in \mathcal{X}$ we then get

$$\hat{f}(x) = \langle \hat{f}, k(\cdot, x) \rangle_k = \langle k(\cdot, x), \hat{f} \rangle_k = \bar{\ell}_f(k(\cdot, x)) = \ell_f(k(\cdot, x)) = f(x),$$

hence $f = \hat{f} \in H_k$. □

9.5. Comments

Section 9.1 and 9.2 are based on, and to a large extent taken verbatim, from [CF3], with the exception of Section 9.2.3, which is from [CF4]. Section 9.3 is based on, and to a large extent taken verbatim, from [CF3, CF6]. The idea to investigate kernels and their RKHSs in the mean field limit is due to the author of the present thesis, who also established all of the theoretical results above. The elementary approach to the double sum kernel in Section 9.2.3 has been started by M. Herty and C. Segala on a formal level, the fully rigorous treatment here is due to the author of the thesis. The idea to describe the relation between the RKHSs as a commutative diagram is due to M. Herty. The article [CF3] was written mostly by the present author and M. Herty, with editorial input from all remaining authors. The parts of the article [CF4] that appear in this thesis have been written by the present author. The article [CF6] was written mostly by the author of this thesis, with editorial input by M. Herty and S. Trimpe.

10. Kernel-based statistical learning in the mean field limit

One motivation for the study of kernels in the mean field limit is given by statistical learning problems arising in the context of kinetic theory, cf. our discussion in Chapter 8. In the preceding Chapter 9, we introduced the mean field limit of kernels, and we achieved a fairly complete description of the resulting RKHSs, so we are ready to turn to the to the investigation of kernel-based statistical learning in the mean field limit, the goal of this chapter. In Section 10.1, we start with results on the approximation capabilities of RKHSs in a mean field context. We briefly review the standard setup of statistical learning theory in 10.2, before investigating kernel-based learning in the mean field limit in Section 10.3. For the reader's convience, we provide some background on Γ-convergence in Section 10.4.

This chapter is based on, and in large parts taken verbatim from, the article [CF6]. Detailed comments on the author's contribution and the relation of this chapter to existing work are provided in Section 10.5.

10.1. Approximation with kernels in the mean field limit

Kernel-based machine learning methods use in general an RKHS as the hypothesis space, and learning often reduces to a search or optimization problem over this function space. For this reason, it is important to investigate the approximation properties of a given kernel and its associated RKHS as well as to ensure that the learning problem over an RKHS (which is in general an infinite-dimensional object) can be tackled with finite computations.

The next result asserts that, under a uniformity condition, the approximation power of the finite-input kernels k_M is inherited by the mean field limit kernel.

Proposition 10.1.1. For $M \in \mathbb{N}_+$, let $\mathcal{F}_M$ be the set of symmetric functions that are continuous w.r.t. $(\vec{x}, \vec{x}') \mapsto d_{\mathrm{KR}}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}'])$. Let $\mathcal{F} \subseteq C^0(\mathcal{P}(X), \mathbb{R})$ such that for all $f \in \mathcal{F}$ and $\epsilon > 0$ there exist $B \in \mathbb{R}_{\geq 0}$ and sequences $f_M \in \mathcal{F}_M$, $\hat{f}_M \in H_M$, $M \in \mathbb{N}_+$, such that

1. $f_M \xrightarrow{\mathcal{P}_1} f$
2. $\|f_M - \hat{f}_M\|_\infty \leq \epsilon$ for all $M \in \mathbb{N}_+$
3. $\|\hat{f}_M\|_M \leq B$ for all $M \in \mathbb{N}_+$

Then for all $f \in \mathcal{F}$ and $\epsilon > 0$, there exists $\hat{f} \in H_k$ with $\|f - \hat{f}\|_\infty \leq \epsilon$.

Proof. Let $f \in \mathcal{F}$ and $\epsilon > 0$ be arbitrary. Let $B \in \mathbb{R}_{\geq 0}$ and $f_M \in \mathcal{F}_M$, $\hat{f}_M \in H_M$, $M \in \mathbb{N}_+$, such that $f_M \xrightarrow{\mathcal{P}_1} f$, $\|f_M - \hat{f}_M\| \leq \frac{\epsilon}{5}$ and $\|\hat{f}_M\|_M \leq B$ for all $M \in \mathbb{N}_+$ (exist by definition of $\mathcal{F}$). Theorem 9.3.1 ensures that there exists a subsequence $(f_{M_\ell})_\ell$ and $\hat{f} \in H_k$ with $\|\hat{f}\|_k \leq B$ such that $\hat{f}_{M_\ell} \xrightarrow{\mathcal{P}_1} \hat{f}$ for $\ell \to \infty$. Choose now $L_1 \in \mathbb{N}_+$ such that for all $\ell \geq L_1$ we have

$$\sup_{\vec{x} \in X^{M_\ell}} |\hat{f}_{M_\ell}(\vec{x}) - \hat{f}(\hat{\mu}[\vec{x}])| \leq \frac{\epsilon}{5}$$
$$\sup_{\vec{x} \in X^{M_\ell}} |f_{M_\ell}(\vec{x}) - f(\hat{\mu}[\vec{x}])| \leq \frac{\epsilon}{5}.$$

Let now $\mu \in \mathcal{P}(X)$ be arbitrary and choose a sequence $\vec{x}_M \in X^M$ with $\hat{\mu}[\vec{x}_M] \xrightarrow{d_{\mathrm{KR}}} \mu$. Finally, let $L_2 \in \mathbb{N}_+$ such that for all $\ell \geq L_2$ we have

$$|f(\mu) - f(\hat{\mu}[\vec{x}_{M_\ell}])| \leq \frac{\epsilon}{5}$$
$$|\hat{f}(\mu) - \hat{f}(\hat{\mu}[\vec{x}_{M_\ell}])| \leq \frac{\epsilon}{5}$$

(such an L_2 exists due to the continuity of f and $\hat{f}$).

We now have for $\ell \geq \max\{L_1, L_2\}$ that

$$\begin{aligned} |f(\mu) - \hat{f}(\mu)| &\leq |f(\mu) - f(\hat{\mu}[\vec{x}_{M_\ell}])| + |f(\hat{\mu}[\vec{x}_{M_\ell}]) - f_{M_\ell}(\vec{x}_{M_\ell})| + |f_{M_\ell}(\vec{x}_{M_\ell}) - \hat{f}_{M_\ell}(\vec{x}_{M_\ell})| \\ &\quad + |\hat{f}_{M_\ell}(\vec{x}_{M_\ell}) - \hat{f}(\hat{\mu}[\vec{x}_{M_\ell}])| + |\hat{f}(\hat{\mu}[\vec{x}_{M_\ell}]) - \hat{f}(\mu)| \\ &\leq \frac{\epsilon}{5} + \frac{\epsilon}{5} + \frac{\epsilon}{5} + \frac{\epsilon}{5} + \frac{\epsilon}{5} = \epsilon. \end{aligned}$$

Since μ was arbitrary, the result follows. □

Intuitively, the set $\mathcal{F}$ consists of all continuous functions on $\mathcal{P}(X)$ that arise as a mean field limit of functions which can be uniformly approximated by uniformly norm-bounded RKHS functions. The result then states (to use a somewhat imprecise terminology) that the RKHS H_k is dense in $\mathcal{F}$. We can interpret this as an appropriate mean field variant of the universality property of kernels: a kernel on a compact metric space is called universal if its associated RKHS is dense w.r.t. the supremum norm in the space of continuous functions, and many common kernels are universal, cf. e.g. [189, Section 4.6]. In our setting, ideally universality of the finite-input kernels k_M is inherited by the mean field limit kernel k. However, since the mean field limit can be interpreted as a form of smoothing limit, some uniformity requirements should be expected. Proposition 10.1.1 provides exactly such a condition.

Remark 10.1.2. In Proposition 10.1.1, the set $\mathcal{F}$ is a subvectorspace of $C^0(\mathcal{P}(X), \mathbb{R})$. Furthermore, if the $\mathcal{P}_1$-convergence in the definition of $\mathcal{F}$ is uniform, then $\mathcal{F}$ is closed.

Proof. We first show that $\mathcal{F}$ is a subvectorspace. Let $f, g \in \mathcal{F}$ and $\lambda \in \mathbb{R}$, $\epsilon > 0$ be arbitrary. W.l.o.g. we can assume $\lambda \neq 0$. Choose sequences $f_M, g_M \in \mathcal{F}_M$, $\hat{f}_M, \hat{g}_M \in H_M$, $M \in \mathbb{N}_+$, and constants $B_f, B_g \in \mathbb{R}_{\geq 0}$ from the definition of $\mathcal{F}$ for f, $\frac{\epsilon}{2|\lambda|}$, and g, $\frac{\epsilon}{2}$, respectively. Let $M \in \mathbb{N}_+$, $\vec{x} \in X^M$ be arbitrary, then

$$|\lambda f_M(\vec{x}) + g(\vec{x}) - (\lambda f(\hat{\mu}[\vec{x}]) - g(\hat{\mu}[\vec{x}]))| \leq |\lambda||f_M(\vec{x}) - f(\hat{\mu}[\vec{x}])| + |g_M(\vec{x}) - g(\hat{\mu}[\vec{x}])|$$

together with $f_M \xrightarrow{\mathcal{P}_1} f$, $g_M \xrightarrow{\mathcal{P}_1} g$ shows that $\lambda f_M + g_M \xrightarrow{\mathcal{P}_1} \lambda f + g$.

Next, we have for all $M \in \mathbb{N}_+$ that

$$\|(\lambda f_M + g_M) - (\lambda \hat{f}_M + \hat{g}_M)\|_\infty \leq |\lambda|\|f_M - \hat{f}_M\|_\infty + \|g_M - \hat{g}_M\|_\infty \leq |\lambda|\frac{\epsilon}{2|\lambda|} + \frac{\epsilon}{2} = \epsilon.$$

Finally,

$$\|\lambda \hat{f}_M + \hat{g}_M\|_M \leq |\lambda|\|\hat{f}_M\|_M + \|\hat{g}_M\|_M \leq |\lambda|B_f + B_g,$$

establishing that $(\lambda \hat{f}_M + \hat{g}_M)_M$ is uniformly norm-bounded. Altogether, we have that $\lambda f + g \in \mathcal{F}$.

We now turn to the second claim. Let $(f^{(n)})_n \subseteq \mathcal{F}$ such that $f^{(n)} \to f$ for some $f \in C^0(\mathcal{P}(X), \mathbb{R})$ and for all $\bar{\epsilon} > 0$ there exist $f_M^{(n)} \in \mathcal{F}_M$, $\hat{f}_M^{(n)} \in H_M$, $(\rho_M)_M \subseteq \mathbb{R}_{\geq 0}$

and $B^{(n)} \in \mathbb{R}_{\geq 0}$ with $\rho_M \searrow 0$, $\|f_M^{(n)} - \hat{f}_M^{(n)}\|_\infty \leq \bar{\epsilon}$ and $\|\hat{f}_M^{(n)}\|_M \leq B^{(n)}$ for all $n, M \in \mathbb{N}_+$, and

$$\sup_{\vec{x} \in X^M} |f_M^{(n)}(\vec{x}) - f^{(n)}(\hat{\mu}[\vec{x}])| \leq \rho_M$$

for all $n, M \in \mathbb{N}_+$. We now show that $f \in \mathcal{F}$. For this, let $\epsilon > 0$ be arbitrary and choose $f_M^{(n)} \in \mathcal{F}_M$, $\hat{f}_M^{(n)} \in H_M$, $(\rho_M)_M \subseteq \mathbb{R}_{\geq 0}$ and $B^{(n)} \in \mathbb{R}_{\geq 0}$ as above with $\bar{\epsilon} = \frac{\epsilon}{4}$. Let $N \in \mathbb{N}_+$ be such that $\|f^{(m)} - f^{(n)}\|_\infty \leq \frac{\epsilon}{4}$ for all $m, n \geq N$ (such an N exists since $(f^{(n)})_n$ converges in $C^0(\mathcal{P}(X), \mathbb{R})$ and hence is a Cauchy sequence). Furthermore, let $M_\rho \in \mathbb{N}_+$ be such that for all $M \geq M_\rho$ we have $\rho_M \leq \frac{\epsilon}{4}$. Define now $f_M = f_M^{(M)}$ and $\hat{f}_M = \hat{f}_M^{(M)}$ for $M = 1, \ldots, M_\rho - 1$, and $f_M = f_M^{(M+N)}$, $\hat{f}_M = \hat{f}_M^{(N)}$ for $M \geq M_\rho$.

Step 1 Let $M \geq M_\rho$ and $\vec{x} \in X^M$ be arbitrary. We have

$$\begin{aligned} |f_M(\vec{x}) - f(\hat{\mu}[\vec{x}])| &= |f_M^{(N+M)}(\vec{x}) - f(\hat{\mu}[\vec{x}])| \\ &\leq |f_M^{(N+M)}(\vec{x}) - f^{(N+M)}(\hat{\mu}[\vec{x}])| + |f^{(N+M)}(\hat{\mu}[\vec{x}]) - f(\hat{\mu}[\vec{x}])| \\ &\leq \rho_M + \|f^{(N+M)} - f\|_\infty, \end{aligned}$$

and since the right hand side (which is independent of $\vec{x}$) converges to 0 for $M \to \infty$, we get $f_M \xrightarrow{\mathcal{P}_1} f$.

Step 2 For $M = 1, \ldots, M_\rho$ we get

$$\|f_M - \hat{f}_M\|_\infty = \|f_M^{(M)} - \hat{f}_M^{(M)}\|_\infty \leq \bar{\epsilon} \leq \epsilon.$$

Let now $M \geq M_\rho$ and $\vec{x} \in X^M$ be arbitrary. We have

$$\begin{aligned}
|f_M(\vec{x}) - \hat{f}_M(\vec{x})| &= |f_M^{(M+N)}(\vec{x}) - \hat{f}_M^{(N)}(\vec{x})| \\
&\leq |f_M^{(M+N)}(\vec{x}) - f^{(N+M)}(\hat{\mu}[\vec{x}])| + |f^{(N+M)}(\hat{\mu}[\vec{x}]) - f^{(N)}(\hat{\mu}[\vec{x}])| \\
&\qquad + |f^{(N)}(\hat{\mu}[\vec{x}]) - f_M^{(N)}(\vec{x})| + |f_M^{(N)}(\vec{x}) - \hat{f}_M^{(N)}(\vec{x})| \\
&\leq \sup_{\vec{x}' \in X^M} |f_M^{(M+N)}(\vec{x}') - f^{(M+N)}(\hat{\mu}[\vec{x}'])| + \|f^{(M+N)} - f^{(N)}\|_\infty \\
&\qquad + \sup_{\vec{x}' \in X^M} |f^{(N)}(\hat{\mu}[\vec{x}']) - f_M^{(N)}(\vec{x}')| + \|f_M^{(N)} - \hat{f}_M^{(N)}\|_\infty \\
&\leq \rho_M + \frac{\epsilon}{4} + \rho_M + \bar{\epsilon} \\
&\leq 4\frac{\epsilon}{4} = \epsilon,
\end{aligned}$$

and since $\vec{x} \in X^M$ was arbitrary, we get $\|f_M - \hat{f}_M\|_\infty \leq \epsilon$.

Step 3 For $M = 1, \ldots, M_\rho - 1$ we get by construction that $\|\hat{f}_M\|_M = \|\hat{f}_M^{(M)}\|_M \leq B^{(M)}$, and for $M \geq M_\rho$ we find $\|\hat{f}_M\|_M = \|\hat{f}_M^{(N)}\|_M \leq B^{(N)}$. Altogether, we get for $M \in \mathbb{N}_+$ that

$$\|\hat{f}_M\|_M \leq \max\{B^{(1)}, \ldots, B^{(M_\rho - 1)}, B^{(N)}\}.$$

Combining the three steps establishes that $f \in \mathcal{F}$. □

Since k_M and k are kernels, we have the usual representer theorem for their corresponding RKHSs, cf. e.g. [180]. A natural question is then whether we have mean field convergence of the minimizers and their representation. This is clarified by the next result.

Theorem 10.1.3. Let $N \in \mathbb{N}_+$, $\mu_1, \ldots, \mu_N \in \mathcal{P}(X)$ and for $n = 1, \ldots, N$ let $\vec{x}_n^{[M]} \in X^M$, $M \in \mathbb{N}_+$, such that $\hat{\mu}[\vec{x}_n^{[M]}] \xrightarrow{d_{\mathrm{KR}}} \mu_n$ for $M \to \infty$. Let $L : \mathbb{R}^N \to \mathbb{R}_{\geq 0}$ be continuous and strictly convex and $\lambda > 0$. For each $M \in \mathbb{N}_+$ consider the problem

$$\min_{f \in H_M} L(f(\vec{x}_1^{[M]}), \ldots, f(\vec{x}_N^{[M]})) + \lambda \|f\|_M, \tag{10.1}$$

as well as the problem

$$\min_{f \in H_k} L(f(\mu_1), \ldots, f(\mu_N)) + \lambda \|f\|_k. \tag{10.2}$$

Then for each $M \in \mathbb{N}_+$ problem (10.1) has a unique solution f_M^*, which is of the form $f_M^* = \sum_{n=1}^N \alpha_n^{[M]} k_M(\cdot, \vec{x}_n^{[M]}) \in H_M$, with $\alpha_1^{[M]}, \dots, \alpha_N^{[M]} \in \mathbb{R}$, and problem (10.2) has a unique solution f^*, which is of the form $f^* = \sum_{n=1}^N \alpha_n k(\cdot, \mu_n) \in H_k$, with $\alpha_1, \dots, \alpha_N \in \mathbb{R}$. Furthermore, there exists a subsequence $(f_{M_\ell}^*)_\ell$ such that $f_{M_\ell}^* \xrightarrow{\mathcal{P}_1} f^*$ and

$$L(f_{M_\ell}^*(\vec{x}_1^{[M_\ell]}), \dots, f_{M_\ell}^*(\vec{x}_N^{[M_\ell]})) + \lambda \|f_{M_\ell}^*\|_{M_\ell} \to L(f^*(\mu_1), \dots, f^*(\mu_N)) + \lambda \|f^*\|_k. \tag{10.3}$$

for $\ell \to \infty$.

The main point of this result is the convergence of the minimizers, which we will establish using a Γ-convergence argument. This approach seems to have been introduced by [75, 35, 74] originally in the context of multi-agent systems. For the reader's convenience, we briefly recall in Section 10.4.

Proof. The existence and uniqueness of f_M and f follows from the well-known representer theorem (applied to all k_M and k).

We now turn to the convergence of the minimizers. For all $M \in \mathbb{N}_+$ we have

$$\lambda \|f_M^*\|_M \le L(f_M^*(\vec{x}_1^{[M]}), \dots, f_M^*(\vec{x}_N^{[M]})) + \lambda \|f\|_M \le L(0, \dots, 0),$$

i.e., $\|f_M^*\|_M \le L(0, \dots, 0)/\lambda$. Define

$$\mathcal{L}_M : H_M \to \mathbb{R}_{\ge 0},\ f \mapsto L(f(\vec{x}_1^{[M]}), \dots, f(\vec{x}_N^{[M]})) + \lambda \|f\|_M$$
$$\mathcal{L} : H_k \to \mathbb{R}_{\ge 0},\ f \mapsto L(f(\mu_1), \dots, f(\mu_N)) + \lambda \|f\|_k,$$

and let $f_M \in H_M$ with $f_M \xrightarrow{\mathcal{P}_1} f$ for some $f \in H_k$. The continuity of f_M, f and L as well as $\vec{x}_n^{[M]} \xrightarrow{d_{\mathrm{KR}}} \mu_n$ for $M \to \infty$ and all $n = 1, \dots, N$, imply then that $\lim_{M\to\infty} L(f_M(\vec{x}_1^{[M]}), \dots, f_M(\vec{x}_N^{[M]})) = L(f(\mu_1), \dots, f(\mu_N))$. Combining this with Lemma 9.3.2 leads to

$$\mathcal{L}(f) \le \liminf_{M\to\infty} \mathcal{L}_M(f).$$

Let now $f \in H_k$ be arbitrary and let $f_M \in H_M$ be the sequence from Lemma 9.3.3. Using the same arguments as above we find that

$$\limsup_{M\to\infty} \mathcal{L}_M(f_M) \le \|f\|_k.$$

We have shown that $\mathcal{L}_M \xrightarrow{\Gamma} \mathcal{L}$ and hence Proposition 10.4.1 ensures that there exists a subsequence $(f^*_{M_\ell})_\ell$ such that $f^*_{M_\ell} \xrightarrow{\mathcal{P}_1} f^*$ and $\mathcal{L}_{M_\ell}(f^*_{M_\ell}) \to \mathcal{L}(f^*)$. □

Remark 10.1.4. An inspection of the proof reveals that in Theorem 10.1.3 we can replace the term $\lambda\|\cdot\|_M$ and $\lambda\|\cdot\|_k$ by $\Omega(\|\cdot\|_M)$ and $\Omega(\|\cdot\|_k)$, where $\Omega : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a nonnegative, strictly increasing and continuous function.

10.2. The setup of statistical learning theory

We now introduce the standard setup of statistical learning theory, following mostly [189, Chapters 2 and 5]. Let $\mathcal{X} \neq \emptyset$ (associated with some σ-algebra) and $\emptyset \neq Y \subseteq \mathbb{R}$ closed (associated with the corresponding Borel σ-algebra). A *loss function* is in this setting a measurable function $\ell : \mathcal{X} \times Y \times \mathbb{R} \to \mathbb{R}_{\geq 0}$. Let P be a probability distribution on $\mathcal{X} \times Y$ and $f : \mathcal{X} \to \mathbb{R}$ a measurable function, then the *risk of f w.r.t. P and loss function ℓ* is defined by

$$\mathcal{R}_{\ell,P}(f) = \int_{\mathcal{X}\times Y} \ell(x, y, f(x))\mathrm{d}P.$$

Note that this is always well-defined since $(x, y) \mapsto \ell(x, y, f(x))$ is a measurable and nonnegative function. For a set $H \subseteq \mathbb{R}^{\mathcal{X}}$ of measurable functions we also define the *minimal risk over H* by

$$\mathcal{R}^{H*}_{\ell,P} = \inf_{f\in H} \mathcal{R}_{\ell,P}(f).$$

If H is a normed vector space, we additionally define the *regularized risk of $f \in H$* and the *minimal regularized risk over H* by

$$\mathcal{R}_{\ell,P,\lambda}(f) = \mathcal{R}_{\ell,P}(f) + \lambda\|f\|^2_H, \qquad \mathcal{R}^{H*}_{\ell,P,\lambda} = \inf_{f\in H} \mathcal{R}_{\ell,P,\lambda}(f),$$

where $\lambda \in \mathbb{R}_{>0}$ is the *regularization parameter*. A *data set of size $N \in \mathbb{N}_+$* is a tuple $D_N = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X} \times Y)^N$ and for a function $f : \mathcal{X} \to \mathbb{R}$ we define its *empirical risk* by

$$\mathcal{R}_{\ell,D_N}(f) = \frac{1}{N}\sum_{n=1}^{N} \ell(x_n, y_n, f(x_n)).$$

If H is a normed vector space and $f \in H$, we define additionally the *regularized empirical risk* and the *minimal regularized empirical risk over H* by

$$\mathcal{R}_{\ell,D_N,\lambda}(f) = \mathcal{R}_{\ell,D_N}(f) + \lambda\|f\|_H^2, \qquad \mathcal{R}^{H*}_{\ell,D_N,\lambda} = \inf_{f\in H} \mathcal{R}_{\ell,D_N,\lambda}(f),$$

where $\lambda \in \mathbb{R}_{>0}$ is again the regularization parameter. Note that the notation for the empirical risks is consistent with the risk w.r.t. a probability distribution P, if we identify a data set D_N by the corresponding empirical distribution $\frac{1}{N}\sum_{n=1}^N \delta_{(x_n,y_n)}$.

In the following, H will be a RKHS and a minimizer (assuming existence and uniqueness) of $\mathcal{R}^{H*}_{\ell,P,\lambda}$ will be called an *infinite-sample support vector machine (SVM)*. Similarly, $\mathcal{R}^{H*}_{\ell,D_N,\lambda}$ will be called the *empirical solution of the SVM w.r.t. the data set D_N*. Note that this is the common terminology in statistical learning theory, cf. [189], and corresponds to (empirical) risk minization with Tikhonov regularization.

10.3. Statistical learning theory in the mean field limit

We now turn to kernel-based statistical learning in the mean field limit.

10.3.1. Setup

We start by translating the setup of statistical learning theory to a mean field setting. This will require the mean field limit of loss functions. However, to the best of our knowledge, existing mean field limit existence result are not applicable to general loss functions. We therefore first state and prove the following existence result for mean field limits of functions, which might be of independent interest.

Proposition 10.3.1. Let (X, d_X) be a compact metric space and (Z, d_Z) a metric space that has a countable basis $(U_n)_n$ such that $\bar{U}_n$ is compact for all $n \in \mathbb{N}$. Let $f_M : X^M \times Z \to \mathbb{R}$, $M \in \mathbb{N}_+$, be a sequence of functions fulfilling the following conditions.

1. *(Symmetry in $\vec{x}$)*[1] For all $M \in \mathbb{N}_+$, $\vec{x} \in X^M$, $z \in Z$ and permutations $\sigma \in \mathcal{S}_M$, we have $f_M(\sigma\vec{x}, z) = f_M(\vec{x}, z)$.

[1] As mentioned before, this condition is actually implied by the next condition. However, as usual in the kinetic theory literature, we kept this condition for emphasis.

2. *(Uniform boundedness)* There exists $B_f \in \mathbb{R}_{\geq 0}$ and a function $b : Z \to \mathbb{R}_{\geq 0}$ such that $\forall M \in \mathbb{N}_+, \vec{x} \in X^M, z \in z : |f_M(\vec{x}, z)| \leq B_f + b(z)$.

3. *(Uniform Lipschitz continuity)* There exists some $L_f \in \mathbb{R}_{>0}$ such that for all $M \in \mathbb{N}_+$, $\vec{x}_1, \vec{x}_2 \in X^M$, $z_1, z_2 \in Z$ we have $|f_M(\vec{x}_1, z_1) - f_M(\vec{x}_2, z_2)| \leq L_f \left(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2]) + d_Z(z_1, z_2)\right)$.

Then there exists a subsequence $(f_{M_\ell})_\ell$ and a continuous function $f : \mathcal{P}(X) \times Z \to \mathbb{R}$ such that $f_{M_\ell} \xrightarrow{\mathcal{P}_1} f$ for $\ell \to \infty$. Furthermore, f is L_f-Lipschitz continuous and there exists $B_F \in \mathbb{R}_{\geq 0}$ such that for all $\mu \in \mathcal{P}(X)$, $z \in Z$ we have $|f(\mu, z)| \leq B_F + b(z)$.

Proof. For $M \in \mathbb{N}_+$ define the McShane extension $F_M : \mathcal{P}(X) \times Z \to \mathbb{R}$ by

$$F_M(\mu, z) = \inf_{\vec{x} \in X^M} f_M(\vec{x}, z) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}], \mu).$$

Observe that F_M is well-defined (i.e., $\mathbb{R}$-valued) since $f_M(\cdot, z)$ and $L_f d_{\mathrm{KR}}(\hat{\mu}[\cdot], \mu)$ are bounded for every $z \in Z$ (since f_M and $d_{\mathrm{KR}}(\hat{\mu}[\cdot], \mu)$ are continuous and $\mathcal{P}(X)$ is compact, hence bounded).

Step 1 F_M extends f_M, i.e., for all $M \in \mathbb{N}_+$, $\vec{x} \in X^M$ and $z \in Z$ we have $F_M(\hat{\mu}[\vec{x}], z) = f_M(\vec{x}, z)$. To show this, let $\vec{x} \in X^M$ and $z \in Z$ be arbitrary and observe that by definition

$$\begin{aligned} F_M(\hat{\mu}[\vec{x}], z) &= \inf_{\vec{x}' \in X^M} f_M(\vec{x}', z) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'], \hat{\mu}[\vec{x}]) \\ &\leq f_M(\vec{x}, z) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}]) = f_M(\vec{x}, z). \end{aligned}$$

If $F_M(\hat{\mu}[\vec{x}], z) < f_M(\vec{x}, z)$, then there exists some $\vec{x}' \in X^M$ such that

$$f_M(\vec{x}', z) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'], \hat{\mu}[\vec{x}]) < f_M(\vec{x}, z),$$

but this means that

$$L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}'], \hat{\mu}[\vec{x}]) < f_M(\vec{x}, z) - f_M(\vec{x}', z) \leq |f_M(\vec{x}, z) - f_M(\vec{x}', z)|,$$

contradicting the L_f-Lipschitz continuity of f_M.

Step 2 All F_M are L_f-continuous: Let $M \in \mathbb{N}_+$, $\mu_i \in \mathcal{P}(X)$ and $z_i \in Z$, $i = 1, 2$, be arbitrary. Since X^M is compact and $f_M(\cdot, z)$ and $L_f d_{\mathrm{KR}}(\hat{\mu}[\cdot], \mu_i)$, $i = 1, 2$, are

continuous, the infimum in the definition of F_M is actually attained. Let $\vec{x}_2 \in X^M$ such that $F_M(\mu_2, z_2) = f_M(\vec{x}_2, z_2) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_2)$, then we have

$$\begin{aligned}
F_M(\mu_1, z_1) &\leq f_M(\vec{x}_2, z_1) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_1) \\
&= f_M(\vec{x}_2, z_1) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_2) - L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_2) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_1) \\
&\leq f_M(\vec{x}_2, z_2) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_2) + L_f d_Z(z_1, z_2) - L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_2) \\
&\qquad + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_1) \\
&\leq F_M(\mu_2, z_2) + L_f d_Z(z_1, z_2) - L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_2) + L_f d_{\mathrm{KR}}(\mu_1, \mu_2) \\
&\qquad + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_2], \mu_2) \\
&= F_M(\mu_2, z_2) + L_f(d_{\mathrm{KR}}(\mu_1, \mu_2) + d_Z(z_1, z_2)),
\end{aligned}$$

where we used the definition of F_M in the first inequality, the Lipschitz continuity of f_M (w.r.t. the second argument) for the second inequality, and then the fact that $\vec{x}_2$ attains the infimum in the definition of $F_M(\mu_2, z_2)$ and the triangle inequality for d_{KR}. Interchanging the roles of μ_1, z_1 and μ_2, z_2 then establishes the claim.

Step 3 There exists $B_F \in \mathbb{R}_{\geq 0}$ such that for all $M \in \mathbb{N}_+$, $\mu \in \mathcal{P}(X)$ and $z \in Z$ we have $|F_M(\mu, z)| \leq B_F + h(z)$: Let $D_{\mathcal{P}(X)}$ be the diameter of $\mathcal{P}(X)$ (which is finite since $\mathcal{P}(X)$ is compact), then for all $M \in \mathbb{N}_+$ and $\vec{x} \in X^M$, $z \in Z$, $\mu \in \mathcal{P}(X)$ we have

$$-(B_f + L_f D_{\mathcal{P}(X)} + b(z)) \leq f_M(\vec{x}, z) + L_f d_{\mathrm{KR}}(\hat{\mu}[\vec{x}], \mu) \leq B_f + L_f D_{\mathcal{P}(X)} + b(z),$$

therefore $|F_M(\mu, z)| \leq B_f + L_f D_{\mathcal{P}(X)} + b(z)$, showing the claim with $B_F = B_f + L_f D_{\mathcal{P}(X)}$.

Step 4 Summarizing, $(F_M)_M$ is a sequence of L_f-Lipschitz continuous and hence equicontinuous functions such that for all $\mu \in \mathcal{P}(X)$ and $z \in Z$, the set $\{F_M(\mu, z) \mid M \in \mathbb{N}_+\}$ is relatively compact (since it is a bounded subset of $\mathbb{R}$). We can now use a variant of the Arzela-Ascoli theorem, cf. [114, Corollary III.3.3]. From the assumption on Z, we can find a sequence $(V_n)_n$ of open subsets of Z such that all $\bar{V}_n$ are compact, $\bar{V}_n \subseteq V_{n+1}$ and we have $\bigcup_n V_n = Z$. Then $(F_M|_{\bar{V}_n})_M$ is a sequence of functions that fulfills the conditions of the Arzela-Ascoli theorem (since $\mathcal{P}(X) \times K_n$ is compact), so there exists a subsequence $(F_{M_\ell^{(n)}}|_{\bar{V}_n})_\ell$ that converges uniformly to a continuous function on $\mathcal{P}(X) \times \bar{V}_n$. Denote the diagonal subsequence of all these

subsequences by $(F_{M_\ell})_\ell$, then there exists a continuous $f : \mathcal{P}(X) \times Z \to \mathbb{R}$ such that $(F_{M_\ell})_\ell$ converges uniformly on compact subsets to f. Since $\mathcal{P}(X)$ is compact, this means that for all compact $K \subseteq Z$

$$\lim_\ell \sup_{\substack{\mu \in \mathcal{P}(X) \\ z \in K}} |F_{M_\ell}(\mu, z) - f(\mu, z)| = 0.$$

This also implies that for all $\mu \in \mathcal{P}(X)$ and $z \in Z$ we have $|f(\mu, z)| \leq B_F + b(z)$.

Furthermore, f is also L_f-Lipschitz continuous: Let $\mu_i \in \mathcal{P}(X)$, $z_i \in Z$, $i = 1, 2$, and $\epsilon > 0$ be arbitrary. Let $K \subseteq Z$ be compact with $z_1, z_2 \in K$ and choose $\ell \in \mathbb{N}_+$ such that

$$\sup_{\substack{\mu \in \mathcal{P}(X) \\ z \in K}} |F_{M_\ell}(\mu, z) - f(\mu, z)| \leq \frac{\epsilon}{2}.$$

We then have

$$\begin{aligned} |f(\mu_1, z_1) - f(\mu_2, z_2)| &\leq |f(\mu_1, z_1) - F_{M_\ell}(\mu_1, z_1)| + |F_{M_\ell}(\mu_1, z_1) - F_{M_\ell}(\mu_2, z_2)| \\ &\quad + |F_{M_\ell}(\mu_2, z_2) - f(\mu_2, z_2)| \\ &\leq L_f \left(d_{\mathrm{KR}}(\mu_1, \mu_2) + d_Z(z_1, z_2) \right) + \epsilon, \end{aligned}$$

and since $\epsilon > 0$ was arbitrary, the claim follows.

Step 5 For $\ell \in \mathbb{N}_+$ and $\vec{x} \in X^{M_\ell}$, $z \in Z$ we have

$$|f_{M_\ell}(\vec{x}, z) - f(\hat{\mu}[\vec{x}], z)| = |F_{M_\ell}(\hat{\mu}[\vec{x}], z) - f(\hat{\mu}[\vec{x}], z)|$$

since F_{M_ℓ} extends f_{M_ℓ}, and hence

$$\sup_{\substack{\vec{x} \in X^{M_\ell} \\ z \in K}} |f_{M_\ell}(\vec{x}, z) - f(\hat{\mu}[\vec{x}], z)| \to 0.$$

□

Let now $\emptyset \neq Y \subseteq \mathbb{R}$ be compact and $\ell_M : X^M \times Y \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, $M \in \mathbb{N}$, such that the following holds.

1. $\ell_M(\sigma \vec{x}, y, t) = \ell_M(\vec{x}, y, t)$ for all $\vec{x} \in X^M$, $\sigma \in \mathcal{S}_M$, $y \in Y$, $t \in \mathbb{R}$.
2. There exists $C_\ell \in \mathbb{R}_{\geq 0}$ and a nondecreasing function $b : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with

$|\ell_M(\vec{x}, y, t)| \leq C_\ell + b(|t|)$ for all $M \in \mathbb{N}$ and $\vec{x} \in X^M, y \in Y, t \in \mathbb{R}$.

3. There exists $L_\ell \in \mathbb{R}_{\geq 0}$ with

$$|\ell_M(\vec{x}_1, y_1, t_1) - \ell_M(\vec{x}_2, y_2, t_2)| \leq L_\ell(d_{\mathrm{KR}}(\hat{\mu}[\vec{x}_1], \hat{\mu}[\vec{x}_2]) + |y_1 - y_2| + |t_1 - t_2|)$$

for all $\vec{x}_1, x_2 \in X^M$, $y_1, y_1' \in Y, t_1, t_2 \in \mathbb{R}$.

In particular, all ℓ_M are measurable (assuming the Borel σ-algebra on X^M) and hence are loss functions on $X^M \times Y$. Proposition 10.3.1 ensures the existence of a subsequence $(\ell_{M_m})_m$ and an L_ℓ-Lipschitz continuous function $\ell : \mathcal{P}(X) \times Y \times \mathbb{R} \to \mathbb{R}$ with

$$\lim_{M \to \infty} \sup_{\substack{\vec{x} \in X^{M_m} \\ y \in Y, t \in K}} |\ell_{M_m}(\vec{x}, y, t) - \ell(\hat{\mu}[\vec{x}], y, t)| = 0 \tag{10.4}$$

for all compact $K \subseteq \mathbb{R}$, and we write again $\ell_{M_m} \xrightarrow{\mathcal{P}_1} \ell$. *For readability, from now on we switch to this subsequence.* Furthermore, we also get from Proposition 10.3.1 that there exists some $C_L \in \mathbb{R}_{\geq 0}$ such that $|\ell(\mu, y, t)| \leq C_L + b(|t|)$ for all $\mu \in \mathcal{P}(X), y \in Y, t \in \mathbb{R}$.

Remark 10.3.2. Note that, for Proposition 10.3.1 to apply, it is enough to assume in item 2) above the existence of a function $b : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with $|\ell_M(\vec{x}, y, t)| \leq C_\ell + b(|t|)$. However, we chose the slightly stronger condition that b is nondecreasing, since then ℓ_M is a *Nemitskii loss* according to [189, Definition 2.16]. Since the function with constant value C_ℓ is actually P_M-integrable, this means that ℓ_M is even a *P_M-integrable Nemitskii loss* according to [189]. A similar remark then applies to ℓ.

Lemma 10.3.3. The function ℓ is nonnegative. Furthermore, if all ℓ_M are convex loss functions [189, Definition 2.12], i.e., if for all $M \in \mathbb{N}_+$, $\vec{x} \in X^M, y \in Y, t_1, t_2 \in \mathbb{R}$ and $\lambda \in (0, 1)$ we have

$$\ell_M(\vec{x}, y, \lambda t_1 + (1 - \lambda) t_2) \leq \lambda \ell_M(\vec{x}, y, t_1) + (1 - \lambda) \ell_M(\vec{x}, y, t_2), \tag{10.5}$$

then so is ℓ.

Proof. That ℓ is nonnegative is clear from the proof of Proposition 10.3.1. Let now all ℓ_M be convex and let $\mu \in \mathcal{P}(X)$, $y \in Y, t_1, t_2 \in \mathbb{R}$ and $\lambda \in (0, 1)$ be arbitrary, and

define $I = [\min\{t_1, t_2\}, \max\{t_1, t_2\}]$. Furthermore, let $\vec{x}_M \in X^M$ with $\vec{x}_M \xrightarrow{d_{\mathrm{KR}}} \mu$ for $M \to \infty$ and $\epsilon > 0$ be arbitrary. Choose now M so large that

$$\begin{aligned}
&|\ell(\mu, y, \lambda t_1 + (1-\lambda)t_2) - \ell(\hat{\mu}[\vec{x}_M], y, \lambda t_1 + (1-\lambda)t_2)| \\
&\quad \leq \frac{\epsilon}{6} \sup_{\substack{\vec{x} \in X^M \\ y' \in Y, t \in I}} |\ell_M(\vec{x}, y', t') - \ell(\hat{\mu}[\vec{x}], y', t')| \\
&\quad \leq \frac{\epsilon}{6}.
\end{aligned}$$

This is possible due to the continuity of ℓ, as well as $\ell_M \xrightarrow{\mathcal{P}_1} \ell$. We then have

$$\begin{aligned}
\ell(\mu, y, \lambda t_1 + (1-\lambda)t_2) &\leq \ell(\hat{\mu}[\vec{x}], y, \lambda t_1 + (1-\lambda)t_2) + \frac{\epsilon}{6} \\
&\leq \ell_M(\vec{x}_M, y, \lambda t_1 + (1-\lambda)t_2) + \frac{\epsilon}{3} \\
&\leq \lambda \ell_M(\vec{x}_M, y, t_1) + (1-\lambda)\ell_M(\vec{x}_M, y, t_2) + \frac{\epsilon}{3} \\
&\leq \lambda \ell(\hat{\mu}[\vec{x}_M], y, t_1) + (1-\lambda)\ell(\hat{\mu}[\vec{x}_M], y, t_2) + \frac{\epsilon}{3} + (\lambda + 1 - \lambda)\frac{\epsilon}{6} \\
&\leq \lambda \ell(\mu, y, t_1) + (1-\lambda)\ell(\mu, y, t_2) + \epsilon,
\end{aligned}$$

and since $\epsilon > 0$ was arbitrary, this establishes

$$\ell(\mu, y, \lambda t_1 + (1-\lambda)t_2) \leq \lambda \ell(\mu, y, t_1) + (1-\lambda)\ell(\mu, y, t_2),$$

i.e., convexity of ℓ. □

10.3.2. Empirical SVM solutions

Given data sets $D_N^{[M]} = \left((\vec{x}_1^{[M]}, y_1^{[M]}), \ldots, (\vec{x}_N^{[M]}, y_N^{[M]})\right)$ for all $M \in \mathbb{N}_+$ with $\vec{x}_n^{[M]} \in X^M$, $y_n^{[M]} \in Y$, and $D_N = ((\mu_1, y_1), \ldots, (\mu_N, y_N))$ with $\mu_n \in \mathcal{P}(X)$ and $y_n \in Y$, we write $D_N^{[M]} \xrightarrow{\mathcal{P}_1} D_N$ if $\hat{\mu}[\vec{x}_n^{[M]}] \xrightarrow{d_{\mathrm{KR}}} \mu_n$ and $y_n^{[M]} \to y_n$ (where $M \to \infty$) for all $n = 1, \ldots, N$. We can interpret this as mean field convergence of the data sets.

Furthermore, consider the empirical risk of hypothesis $f_M \in H_M$ (and $f \in H_k$)

on data set $D_N^{[M]}$ (and D_N)

$$\mathcal{R}_{\ell_M, D_N^{[M]}}(f_M) = \frac{1}{N}\sum_{n=1}^{N} \ell_M(\vec{x}_n^{[M]}, y_n^{[M]}, f_M(\vec{x}_n^{[M]}))$$
$$\mathcal{R}_{\ell, D_N}(f) = \frac{1}{N}\sum_{n=1}^{N} \ell(\mu_n, y_n, f(\mu_n)),$$

and the corresponding regularized risk

$$\mathcal{R}_{\ell_M, D_N^{[M]}, \lambda}(f_M) = \frac{1}{N}\sum_{n=1}^{N} \ell_M(\vec{x}_n^{[M]}, y_n^{[M]}, f_M(\vec{x}_n^{[M]})) + \lambda\|f_M\|_M^2$$
$$\mathcal{R}_{\ell, D_N, \lambda}(f) = \frac{1}{N}\sum_{n=1}^{N} \ell(\mu_n, y_n, f(\mu_n)) + \lambda\|f\|_k^2,$$

where $\lambda \in \mathbb{R}_{>0}$ is the regularization parameter.

Proposition 10.3.4. Let $\lambda > 0$, assume that all ℓ_M are convex and let $D_N^{[M]}$, D_N be finite data sets with $D_N^{[M]} \xrightarrow{\mathcal{P}_1} D_N$. Then for all $M \in \mathbb{N}_+$, $H_M \ni f_M \mapsto \mathcal{R}_{\ell_M, D_N^{[M]}, \lambda}(f_M)$ has a unique minimizer $f_{M,\lambda}^* \in H_M$ and $H_k \ni f \mapsto \mathcal{R}_{\ell, D_N, \lambda}(f)$ has a unique minimizer $f_\lambda^* \in H_k$. Furthermore, for all $M \in \mathbb{N}_+$ there exist $\alpha_n^{[M]} \in \mathbb{R}$, $n = 1, \ldots, N$, such that $f_{M,\lambda}^* = \sum_{n=1}^N \alpha_n^{[M]} k_M(\cdot, \vec{x}_n^{[M]})$, and there exist $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$ such that $f_\lambda^* = \sum_{n=1}^N \alpha_n k(\cdot, \mu_n)$. Finally, there exists a subsequence $(f_{M_m,\lambda}^*)_m$ such that $f_{M_m,\lambda}^* \xrightarrow{\mathcal{P}_1} f_\lambda^*$ and $\mathcal{R}_{\ell_{M_m}, D_N^{[M_m]}, \lambda}(f_{M_m,\lambda}^*) \to \mathcal{R}_{\ell, D_N, \lambda}(f_\lambda^*)$ for $m \to \infty$.

Proof. From Lemma 10.3.3 we get that ℓ is nonnegative and convex. The existence, uniqueness and the representation formulas follow then from the standard representer theorem, cf. e.g., [189, Theorem 5.5].

Furthermore, for all $M \in \mathbb{N}_+$ we have

$$\begin{aligned}\lambda\|f_{M,\lambda}^*\|_M^2 &\leq \frac{1}{N}\sum_{n=1}^{N} \ell_M(\vec{x}_n^{[M]}, y_n^{[M]}, f_{M,\lambda}^*(\vec{x}_n^{[M]})) + \lambda\|f_{M,\lambda}^*\|_M^2 \\ &\leq \mathcal{R}_{\ell_M, D_N^{[M]}, \lambda}(0) \\ &\leq NC_\ell,\end{aligned}$$

hence $\|f_{M,\lambda}^*\|_M \leq \sqrt{\frac{NC_\ell}{\lambda}}$.

Let $f \in H_k$ and $(f_M)_M$, $f_M \in H_M$, such that $f_M \xrightarrow{\mathcal{P}_1} f$. From $D_N^{[M]} \xrightarrow{\mathcal{P}_1} D_N$ and the continuity of ℓ_M, ℓ, together with $\ell_M \xrightarrow{\mathcal{P}_1} \ell$ and the boundedness of $\{y_n^{[M]} \mid M \in \mathbb{N}_+, n = 1, \ldots, N\} \subseteq Y$ and $\{f_M(\vec{x}_n^{[M]}) \mid M \in \mathbb{N}_+, N = 1, \ldots, N\}$ we find that

$$\lim_M \frac{1}{N} \sum_{n=1}^{N} \ell_M(\vec{x}_n^{[M]}, y_n^{[M]}, f_M(\vec{x}_n^{[M]})) = \frac{1}{N} \sum_{n=1}^{N} \ell(\mu_n, y_n, f(\mu_n)).$$

Combining this with Lemma 9.3.2 and Lemma 9.3.3 then establishes that $\mathcal{R}_{\ell_M, D_N^{[M]}, \lambda} \xrightarrow{\Gamma} \mathcal{R}_{\ell, D_N, \lambda}$ and the remaining claims follow from Proposition 10.4.1 and the uniqueness of the minimizers. □

10.3.3. Convergence of distributions and infinite-sample SVMs in the mean field limit

We now turn to the question of mean field limits of distributions and the associated learning problems and SVM solutions. Let $(P^{[M]})_M$ be a sequence of distributions, where $P^{[M]}$ is a probability distribution on $X^M \times Y$, and let P be a probability distribution on $\mathcal{P}(X) \times Y$. We say that $P^{[M]}$ *converges in mean field to* P and write $P^{[M]} \xrightarrow{\mathcal{P}_1} P$, if for all continuous (w.r.t. the product topology on $\mathcal{P}(X) \times Y$) and bounded [2] f we have

$$\int_{X^M \times Y} f(\hat{\mu}[\vec{x}], y) \mathrm{d}P^{[M]}(\vec{x}, y) \to \int_{\mathcal{P}(X) \times Y} f(\mu, y) \mathrm{d}P(\mu, y). \tag{10.6}$$

This convergence notion of probability distributions (on different input spaces) appears to be not standard, but it is a natural concept in the present context. Essentially, it is weak (also called narrow) convergence of probability distributions adapated to our setting.

Consider now data sets $D_N^{[M]}$, D_N, with $D_N^{[M]} \xrightarrow{\mathcal{P}_1} D_N$, then we also have convergence in mean field of the datasets, interpreted as empirical distributions: let

[2] Of course, since Y is compact, all continuous f are bounded in our present setting.

$f \in C^0(\mathcal{P}(X) \times Y, \mathbb{R})$ be bounded, then

$$\int_{X^M \times Y} f(\hat{\mu}[\vec{x}], y) \mathrm{d}D_N^{[M]}(\vec{x}, y) = \frac{1}{N} \sum_{n=1}^{N} f(\hat{\mu}[\vec{x}_n^{[M]}], y_n^{[M]}) \xrightarrow{M \to \infty} \frac{1}{N} \sum_{n=1}^{N} f(\mu_n, y_n) = \int_{\mathcal{P}(X) \times Y} f(\mu, y) \mathrm{d}D_N(\mu, y).$$

This shows that the mean field convergence of probability distributions as defined here is a direct generalization of the natural notion of mean field convergence of data sets.

Finally, consider the risk of hypothesis $f_M \in H_M$ and $f \in H_k$ w.r.t. the distribution $P^{[M]}$ and P, respectively,

$$\mathcal{R}_{\ell_M, P^{[M]}}(f_M) = \int_{X^M \times Y} \ell_M(\vec{x}, y, f_M(\vec{x})) \mathrm{d}P^{[M]}(\vec{x}, y)$$
$$\mathcal{R}_{\ell, P}(f) = \int_{\mathcal{P}(X) \times Y} \ell(\mu, y, f(\mu)) \mathrm{d}P(\mu, y),$$

as well as the minimal risks

$$\mathcal{R}^{H_M *}_{\ell_M, P^{[M]}} = \inf_{f_M \in H_M} \mathcal{R}_{\ell_M, P^{[M]}}(f_M) \qquad \mathcal{R}^{H_k *}_{\ell, P} = \inf_{f \in H_k} \mathcal{R}_{\ell, P}(f).$$

Our first result ensures that mean field convergence of distributions $P^{[M]}$, loss functions ℓ_M and data sets $D_N^{[M]}$ ensures the convergence of the corresponding risks of the empirical SVM solutions.

Lemma 10.3.5. Consider the situation and notation of Proposition 10.3.4 and assume that $P^{[M]} \xrightarrow{\mathcal{P}_1} P$. We then have $\mathcal{R}_{\ell_{M_m}, P^{[M_m]}}(f^*_{M_m, \lambda}) \to \mathcal{R}_{\ell, P}(f^*_\lambda)$ for $m \to \infty$.

Proof. Let $\epsilon > 0$ be arbitrary. Recall from the proof of Proposition 10.3.4 that for all $M \in \mathbb{N}_+$ we have $\|f^*_{M,\lambda}\|_M \leq \sqrt{\frac{NC_\ell}{\lambda}}$, and hence for all $\vec{x} \in X^M$ we have

$$|f^*_{M,\lambda}(\vec{x})| \leq \|f^*_{M,\lambda}\|_k \|k_M(\cdot, \vec{x})\|_k \leq \sqrt{\frac{NC_\ell}{\lambda}} \sqrt{C_k}.$$

A similar argument applies to $f^*_\lambda \in H_k$, so we can find a compact set $K \subseteq \mathbb{R}$ with

$$\{f^*_{M,\lambda}(\vec{x}_n^{[M]}) \mid M \in \mathbb{N}_+, n = 1, \dots, N\} \cup \{f^*_\lambda(\mu_n) \mid n = 1, \dots, N\} \subseteq K.$$

Choose now $m_\epsilon \in \mathbb{N}_+$ such that for all $m \geq m_\epsilon$ we have

$$\sup_{\substack{\vec{x} \in X^{M_m} \\ y \in Y}} |\ell_{M_m}(\vec{x}, y, f^*_{M_m,\lambda}(\vec{x})) - \ell_{M_m}(\vec{x}, y, f^*_\lambda(\hat{\mu}[\vec{x}]))| \leq \frac{\epsilon}{3}$$

$$\sup_{\substack{\vec{x} \in X^{M_m} \\ y \in Y, t \in K}} |\ell_{M_m}(\vec{x}, y, t) - \ell(\hat{\mu}[\vec{x}], y, t)| \leq \frac{\epsilon}{3}$$

$$\left| \int_{X^{M_m} \times Y} \ell(\hat{\mu}[\vec{x}], y, f^*_\lambda(\hat{\mu}[\vec{x}])) \mathrm{d}P^{[M_m]}(\vec{x}, y) - \int_{\mathcal{P}(X) \times Y} \ell(\mu, y, f^*_\lambda(\mu)) \mathrm{d}(\mu, y) \right| \leq \frac{\epsilon}{3}.$$

Such a m_ϵ exists since $f^*_{M_m,\lambda} \xrightarrow{\mathcal{P}_1} f^*_\lambda$ and all ℓ_{M_m} are uniformly Lipschitz continuous (first inequality), $\ell_{M_m} \xrightarrow{\mathcal{P}_1} \ell$ and Y and K are compact (second inequality), and $P^{[M]} \xrightarrow{\mathcal{P}_1} P$ as well as that $(\mu, y) \mapsto \ell(\mu, y, f^*_\lambda(\mu))$ is continuous and bounded (third inequality). We now have

$$\begin{aligned}
&\left| \mathcal{R}_{\ell_{M_m}, P^{[M_m]}}(f^*_{M_m,\lambda}) - \mathcal{R}_{\ell,P}(f^*_\lambda) \right| \\
&\quad \leq \left| \int_{X^{M_m} \times Y} \ell_{M_m}(\vec{x}, y, f^*_{M_m,\lambda}(\vec{x})) - \ell_{M_m}(\vec{x}, y, f^*_\lambda(\hat{\mu}[\vec{x}])) \mathrm{d}P^{[M_m]}(\vec{x}, y) \right| \\
&\qquad + \left| \int_{X^{M_m} \times Y} \ell_{M_m}(\vec{x}, y, f^*_\lambda(\hat{\mu}[\vec{x}])) - \ell(\hat{\mu}[\vec{x}], y, f^*_\lambda(\hat{\mu}[\vec{x}])) \mathrm{d}P^{[M_m]}(\vec{x}, y) \right| \\
&\qquad + \left| \int_{X^{M_m} \times Y} \ell(\hat{\mu}[\vec{x}], y, f^*_\lambda(\hat{\mu}[\vec{x}])) \mathrm{d}P^{[M_m]}(\vec{x}, y) - \int_{\mathcal{P}(X) \times Y} \ell(\mu, y, f^*_\lambda(\mu)) \mathrm{d}(\mu, y) \right| \\
&\quad \leq \int_{X^{M_m} \times Y} |\ell_{M_m}(\vec{x}, y, f^*_{M_m,\lambda}(\vec{x})) - \ell_{M_m}(\vec{x}, y, f^*_\lambda(\hat{\mu}[\vec{x}]))| \mathrm{d}P^{[M_m]}(\vec{x}, y) \\
&\qquad + \int_{X^{M_m} \times Y} |\ell_{M_m}(\vec{x}, y, f^*_\lambda(\hat{\mu}[\vec{x}])) - \ell(\hat{\mu}[\vec{x}], y, f^*_\lambda(\hat{\mu}[\vec{x}]))| \mathrm{d}P^{[M_m]}(\vec{x}, y) \\
&\qquad + \frac{\epsilon}{3} \\
&\quad \leq \epsilon,
\end{aligned}$$

and since $\epsilon > 0$ was arbitrary, the claim follows. □

Next, we investigate the mean field convergence of infinite-sample SVM solutions

and their associated risks. Define for $\lambda \in \mathbb{R}_{\geq 0}$ (and all $M \in \mathbb{N}_+$) the regularized risk of $f_M \in H_M$ and $f \in H_k$, respectively, by

$$\mathcal{R}_{\ell_M,P^{[M]},\lambda}(f_M) = \mathcal{R}_{\ell_M,P^{[M]}}(f_M) + \lambda\|f_M\|_M^2, \qquad \mathcal{R}_{\ell,P,\lambda}(f) = \mathcal{R}_{\ell,P}(f) + \lambda\|f\|_k^2,$$

and the corresponding minimal risks by

$$\mathcal{R}^{H_M*}_{\ell_M,P^{[M]},\lambda} = \inf_{f_M \in H_M} \mathcal{R}_{\ell_M,P^{[M]},\lambda}(f_M), \qquad \mathcal{R}^{H_k*}_{\ell,P,\lambda} = \inf_{f \in H_k} \mathcal{R}_{\ell,P,\lambda}(f).$$

Proposition 10.3.6.[3] Let $\lambda > 0$, assume that all ℓ_M are convex loss functions and let $P^{[M]}$ and P be probability distributions on $X^M \times Y$ and $\mathcal{P}(X) \times Y$, respectively, with $P^{[M]} \xrightarrow{\mathcal{P}_1} P$. Then for all $M \in \mathbb{N}_+$, $H_M \ni f_M \mapsto \mathcal{R}_{\ell_M,P^{[M]},\lambda}(f_M)$ has a unique minimizer $f^*_{M,\lambda} \in H_M$ and $H_k \ni f \mapsto \mathcal{R}_{\ell,P,\lambda}(f)$ has a unique minimizer $f^*_\lambda \in H_k$. Furthermore, there exists a subsequence $(f^*_{M_m,\lambda})_m$ such that $f^*_{M_m,\lambda} \xrightarrow{\mathcal{P}_1} f^*_\lambda$ and $\mathcal{R}_{\ell_{M_m},P^{[M_m]},\lambda}(f^*_{M_m,\lambda}) \to \mathcal{R}_{\ell,P,\lambda}(f^*_\lambda)$ for $m \to \infty$. In particular, $\mathcal{R}^{H_{M_m}*}_{\ell_{M_m},P^{[M_m]},\lambda} \to \mathcal{R}^{H_k*}_{\ell,P,\lambda}$.

Proof. Observe that all k_M are bounded measurable kernels, $\mathcal{R}_{\ell_M,P^{[M]}}(f_M) < \infty$ for all $f \in H_M$, ℓ_M is a convex, $P^{[M]}$-integrable Nemitskii loss (cf. Remark 10.3.2) and hence [189, Lemma 5.1, Theorem 5.2] guarantee the existence and uniqueness of $f^*_{M,\lambda}$. A completely analogous argument shows the existence and uniqueness of f^*_λ.

We now show that $\mathcal{R}_{\ell_M,P^{[M]},\lambda} \xrightarrow{\Gamma} \mathcal{R}_{\ell,P,\lambda}$. For the Γ-lim inf-inequality, let $f_M \in H_M$, $f \in H_k$ be arbitrary with $f_M \xrightarrow{\mathcal{P}_1} f$, and let $\epsilon > 0$. Choose $M_\epsilon \in \mathbb{N}_+$ so large that for all $M \geq M_\epsilon$

$$\left| \int \ell(\hat{\mu}[\vec{x}], y, f(\hat{\mu}[\vec{x}]) \mathrm{d}P^{[M]}(\vec{x}, y)) - \int \ell(\mu, y, f(\mu)) \mathrm{d}P(\mu, y) \right| \leq \frac{\epsilon}{2}$$

(this is possible since $(\mu, y) \mapsto \ell(\mu, y, f(\mu))$ is bounded and continuous and $P^{[M]} \xrightarrow{\mathcal{P}_1} P$) and

$$|\ell_M(\vec{x}, y, f_M(\vec{x})) - \ell(\hat{\mu}[\vec{x}], y, f(\hat{\mu}[\vec{x}]))| \leq \frac{\epsilon}{2}$$

for all $\vec{x} \in X^M$, $y \in Y$ (this is possible due to the same argument used in the proof

[3] Note that Proposition 10.3.4 is actually a corollary of this result. However, since the former result is independent of the notion of mean field convergence of probability distributions, we stated and proved it separately.

of Lemma 10.3.5). For $M \geq M_\epsilon$ we then find

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f) &= \int \ell(\mu, y, f(\mu))\mathrm{d}P(\mu, y) + \lambda\|f\|_k^2 \\
&\leq \int \ell_M(\vec{x}, y, f_M(\vec{x}))\mathrm{d}P^{[M]}(\vec{x}, y) \\
&\qquad + \left| \int \ell(\hat{\mu}[\vec{x}], y, f(\hat{\mu}[\vec{x}])\mathrm{d}P^{[M]}(\vec{x}, y)) - \int \ell(\mu, y, f(\mu))\mathrm{d}P(\mu, y) \right| \\
&\qquad + \left| \int \ell_M(\vec{x}, y, f_M(\vec{x})) - \ell(\hat{\mu}[\vec{x}], y, f(\hat{\mu}[\vec{x}]))\mathrm{d}P^{[M]}(\vec{x}, y) \right| + \lambda\|f\|_k^2 \\
&\leq \int \ell_M(\vec{x}, y, f_M(\vec{x}))\mathrm{d}P^{[M]}(\vec{x}, y) + \lambda \liminf_M \|f_M\|_M^2 + \epsilon,
\end{aligned}$$

where we used Lemma 9.3.2 in the last inequality.

For the Γ-lim sup-inequality, let $f \in H_k$ be arbitrary and let $(f_M)_M$ be the recovery sequence from Lemma 9.3.3. The desired inequality then follows by repeating the arguments from above.

Finally, using exactly the same argument as in the proof of Proposition 10.3.4 shows that $\|f^*_{M,\lambda}\|_M \leq \sqrt{\frac{NC_\ell}{\lambda}}$, so we can apply Proposition 10.4.1 and the result follows. □

Finally, we would like to show that $\mathcal{R}^{H_M*}_{\ell_M,P^{[M]}} \to \mathcal{R}^{H_k*}_{\ell,P}$ for $P^{[M]} \xrightarrow{\mathcal{P}_1} P$. Up to a subsequence, this is established under Assumption 10.3.7. Define the *approximation error functions*, cf. [189, Definition 5.14], by

$$A_2^{[M]}(\lambda) = \inf_{f \in H_M} \mathcal{R}_{\ell_M,P^{[M]},\lambda}(f) - \mathcal{R}^{H_M*}_{\ell_M,P^{[M]}} \qquad A_2(\lambda) = \inf_{f \in H_k} \mathcal{R}_{\ell,P,\lambda}(f) - \mathcal{R}^{H_k*}_{\ell,P},$$

where $M \in \mathbb{N}_+$ and $\lambda \in \mathbb{R}_{\geq 0}$. Note that (for all $M \in \mathbb{N}_+$) $A_2^{[M]}, A_2 : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ are increasing, concave and continuous, and $A_2^{[M]}, A_2(0) = 0$, cf. [189, Lemma 5.15]. We need essentially equicontinuity of $(A_2^{[M]})_M$ in 0, which is formalized in the following assumption.

Assumption 10.3.7. For all $\epsilon > 0$ there exists $\lambda_\epsilon > 0$ such that for all $0 < \lambda \leq \lambda_\epsilon$ and $M \in \mathbb{N}_+$ we have $A_2^{[M]}(\lambda) \leq \epsilon$.

Proposition 10.3.8. Assume that all ℓ_M are convex loss functions, let $P^{[M]}$ and P be probability distributions on $X^M \times Y$ and $\mathcal{P}(X) \times Y$, respectively, with $P^{[M]} \xrightarrow{\mathcal{P}_1} P$.

If Assumption 10.3.7 holds, there exists a strictly increasing sequence $(M_m)_m$ with $\mathcal{R}^{H_{M_m}*}_{\ell_{M_m},P^{[M_m]}} \to \mathcal{R}^{H_k*}_{\ell,P}$ for $m \to \infty$.

Proof. Let $(\epsilon_n)_n \subseteq \mathbb{R}_{>0}$ with $\epsilon_m \searrow 0$. We construct a strictly increasing sequence $(M_n)_n$ such that

$$\left|\mathcal{R}^{H_{M_n}*}_{\ell_{M_n},P^{[M_n]}} - \mathcal{R}^{H_k*}_{\ell,P}\right| \leq \epsilon_n$$

for all $n \in \mathbb{N}_+$.

We start with $n = 1$: Since $A_2(0) = 0$ and A_2 is continuous in 0, cf. [189, Lemma 5.15], there exists $\lambda_1' \in \mathbb{R}_{>0}$ such that $A_2(\lambda) \leq \frac{\epsilon_1}{3}$ for all $0 < \lambda \leq \lambda_1'$. From Assumption 10.3.7 we get $\lambda_1'' \in \mathbb{R}_{>0}$ such that for all $M \in \mathbb{N}_+$ we have $A_2^{[M]}(\lambda) \leq \frac{\epsilon_1}{3}$ for all $0 < \lambda \leq \lambda_1''$. Define now $\lambda_1 = \min\{\lambda_1', \lambda_1''\}$, and observe that $\lambda_1 > 0$. Proposition 10.3.6 ensures the existence of a strictly increasing sequence $(M_m^{(1)})_m \subseteq \mathbb{N}_+$ with

$$\mathcal{R}^{H_{M_m^{(1)}}*}_{\ell_{M_m^{(1)}},P^{[M_m^{(1)}]},\lambda_1} \to \mathcal{R}^{H_k*}_{\ell,P,\lambda_1}$$

for $m \to \infty$. Choose $m_1 \in \mathbb{N}_+$ such that for all $m \geq m_1$ we have

$$\left|\mathcal{R}^{H_{M_m^{(1)}}*}_{\ell_{M_m^{(1)}},P^{[M_m^{(1)}]},\lambda_1} - \mathcal{R}^{H_k*}_{\ell,P,\lambda_1}\right| \leq \frac{\epsilon_1}{3}.$$

We now set $M_1 = M_{m_1}^{(1)}$ and get that

$$\begin{aligned}
\left|\mathcal{R}^{H_{M_1}*}_{\ell_{M_1},P^{[M_1]}} - \mathcal{R}^{H_k*}_{\ell,P}\right| &\leq \left|\mathcal{R}^{H_{M_{m_1}^{(1)}}*}_{\ell_{M_{m_1}^{(1)}},P^{[M_{m_1}^{(1)}]}} - \mathcal{R}^{H_{M_{m_1}^{(1)}}*}_{\ell_{M_{m_1}^{(1)}},P^{[M_{m_1}^{(1)}]},\lambda_1}\right| + \left|\mathcal{R}^{H_{M_{m_1}^{(1)}}*}_{\ell_{M_{m_1}^{(1)}},P^{[M_{m_1}^{(1)}]},\lambda_1} - \mathcal{R}^{H_k*}_{\ell,P,\lambda_1}\right| \\
&\quad + \left|\mathcal{R}^{H_k*}_{\ell,P,\lambda_1} - \mathcal{R}^{H_k*}_{\ell,P}\right| \\
&\leq A_2^{[M_m^{(1)}]}(\lambda_1) + \frac{\epsilon_1}{3} + A_2(\lambda_1) \\
&\leq \epsilon_1.
\end{aligned}$$

We can now repeat the argument from above inductively: Suppose we have constructed our subsequence up to $n \in \mathbb{N}_+$, i.e., $M_1, \ldots, M_n$. Choose $\lambda' \in \mathbb{R}_{>0}$ such that $A_2(\lambda) \leq \frac{\epsilon_{n+1}}{3}$ for all $0 < \lambda \leq \lambda'$ (exists due to continuity), and $\lambda'' \in \mathbb{R}_{>0}$ such that for all $M \in \mathbb{N}_+$ we have $A_2^{[M]}(\lambda) \leq \frac{\epsilon_{n+1}}{3}$ for all $0 < \lambda \leq \lambda''$ (using Assumption

10.3.7). Define now $\lambda_{n+1} = \min\{\lambda', \lambda''\}$, and observe that $\lambda_{n+1} > 0$. Proposition 10.3.6 ensures the existence of a strictly increasing sequence $\left(M_m^{(n+1)}\right)_m$ such that

$$\mathcal{R}^{H_{M_m^{(n+1)}}*}_{\ell_{M_m^{(n+1)}}, P^{[M_m^{(n+1)}]}, \lambda_{n+1}} \to \mathcal{R}^{H_k*}_{\ell,P,\lambda_{n+1}}$$

for $m \to \infty$. Choose m_{n+1} such that for all $m \geq m_{n+1}$ we have

$$\left| \mathcal{R}^{H_{M_m^{(n+1)}}*}_{\ell_{M_m^{(n+1)}}, P^{[M_m^{(n+1)}]}, \lambda_{n+1}} - \mathcal{R}^{H_k*}_{\ell,P,\lambda_{n+1}} \right| \leq \frac{\epsilon_{n+1}}{3}.$$

Define now $M_{n+1} = \max\{M_n + 1, M_{m_{n+1}}^{(n+1)}\}$, then we get

$$\begin{aligned}
\left| \mathcal{R}^{H_{M_{n+1}}*}_{\ell_{M_{n+1}}, P^{[M_{n+1}]}} - \mathcal{R}^{H_k*}_{\ell,P} \right| \leq & \left| \mathcal{R}^{H_{M_{m_{n+1}}^{(n+1)}}*}_{\ell_{M_{m_{n+1}}^{(n+1)}}, P^{[M_{m_{n+1}}^{(n+1)}]}} - \mathcal{R}^{H_{M_{m_{n+1}}^{(n+1)}}*}_{\ell_{M_{m_{n+1}}^{(n+1)}}, P^{[M_{m_{n+1}}^{(n+1)}]}, \lambda_{n+1}} \right| \\
& + \left| \mathcal{R}^{H_{M_{m_{n+1}}^{(n+1)}}*}_{\ell_{M_{m_{n+1}}^{(n+1)}}, P^{[M_{m_{n+1}}^{(n+1)}]}, \lambda_{n+1}} - \mathcal{R}^{H_k*}_{\ell,P,\lambda_{n+1}} \right| \\
& + \left| \mathcal{R}^{H_k*}_{\ell,P,\lambda_{n+1}} - \mathcal{R}^{H_k*}_{\ell,P} \right| \\
\leq & A_2^{M_{m_{n+1}}^{(n+1)}}(\lambda_{n+1}) + \frac{\epsilon_{n+1}}{3} + A_2(\lambda_{n+1}) \\
\leq & \epsilon_{n+1}.
\end{aligned}$$

The resulting sequence $(M_n)_n$ fulfills then

$$\mathcal{R}^{H_{M_n}*}_{\ell_{M_n}, P^{[M_n]}} \to \mathcal{R}^{H_k*}_{\ell,P}$$

for $n \to \infty$. □

10.4. Technical background: A Γ-convergence argument

We use repeatedly the concept of Γ-convergence, see for example [60]. For convenience, in this section we summarize the well-known and standard main argument, roughly following [34, Chapter 5]. Let $F_M : H_M \to \mathbb{R} \cup \{\infty\}$ and $F : H_k \to \mathbb{R} \cup \{\infty\}$.

We say that F_M Γ-converges to F and write $F_M \xrightarrow{\Gamma} F$, if

1. For all sequences $(f_M)_M$, $f_M \in H_M$, with $f_M \xrightarrow{\mathcal{P}_1} f$ for some $f \in H_k$, we have

$$F(f) \leq \liminf_M F_M(f_M).$$

2. For all $f \in H_k$ there exists a sequence $(f_M)_M$ with $f_M \in H_M$ such that $f_M \xrightarrow{\mathcal{P}_1} f$ and

$$F(f) \geq \limsup_M F_M(f_M).$$

The sequence in the second item is commonly called a *recovery sequence* (for f).

Proposition 10.4.1. Let $F_M \xrightarrow{\Gamma} F$ and $f_M^* \in \operatorname{argmin}_{f \in H_M} F_M(f)$ for all $M \in \mathbb{N}$ (in particular, all the minima are attained). If there exists $B \in \mathbb{R}_{\geq 0}$ such that $\|f_M^*\|_M \leq B$ for all $M \in \mathbb{N}$, then there exists a subsequence $(f_{M_\ell}^*)_\ell$ and $f^* \in H_k$ such that $f_{M_\ell}^* \xrightarrow{\mathcal{P}_1} f^*$. Furthermore, $F_{M_\ell}(f_{M_\ell}^*) \to F(f^*)$.

Proof. From Theorem 9.3.1 we get the existence of $(f_{M_\ell}^*)_\ell$ and $f^* \in H_k$, and that $f_{M_\ell}^* \xrightarrow{\mathcal{P}_1} f^*$. Let $f \in H_k$ be arbitrary and let $(f_M)_M$ be a recovery sequence for f. We then have

$$\begin{aligned} F(f) &\geq \limsup_M F_M(f_M) \\ &\geq \limsup_{M_\ell} F_{M_\ell}(f_{M_\ell}) \\ &\geq \liminf_{M_\ell} F_{M_\ell}(f_{M_\ell}) \\ &\geq \liminf_{M_\ell} F_{M_\ell}(f_{M_\ell}^*) \\ &\geq F(f^*), \end{aligned}$$

where we used the lim sup-inequality of Γ-convergence in the first step, standard properties of lim sup and lim inf in the second and third step, the fact that $f_{M_\ell}^*$ is a minimizer of F_{M_ℓ} in the fourth step, and finally the lim inf-inequality of Γ-convergence. Since $f \in H_k$ was arbitrary, this shows that f^* is a minimizer of F.

Furthermore, let $(f_M)_M$ be a recovery sequence for f^*, then

$$\begin{aligned} F(f^*) &\geq \limsup_M F_M(f_M) \\ &\geq \limsup_\ell F_{M_\ell}(f_{M_\ell}) \\ &\geq \limsup_\ell F_{M_\ell}(f^*_{M_\ell}), \end{aligned}$$

where we used the lim sup-inequality in the first step, an elementary property of lim sup in the second step, and finally that $f^*_{M_\ell}$ is a minimizer of F_{M_ℓ}. Since $f^*_{M_\ell} \xrightarrow{\mathcal{P}_1} f^*$, the lim inf-inequality of Γ-convergence implies that

$$F(f^*) \leq \liminf_\ell F_{M_\ell}(f^*_{M_\ell}),$$

so we find that

$$\limsup_\ell F_{M_\ell}(f^*_{M_\ell}) \leq F(f^*) \leq \liminf_\ell F_{M_\ell}(f^*_{M_\ell}),$$

establishing that $F_{M_\ell}(f^*_{M_\ell}) \to F(f^*)$. □

10.5. Comments

This chapter is based on, and to a large extent taken verbatim from [CF6]. All theoretical results presented above have been established by the author of the present thesis.

11. Mean field limits of discrete-time multiagent systems via kernel mean embeddings

In the present part of this thesis, we are concerned with kernels in the context of mean field limits as arising in kinetic theory. Motivated by learning problems on large scale interacting particle systems, cf. Chapter 8, we investigated in Chapters 9 and 10 kernels, RKHSs, and kernel-based learning methods in the mean field limit, which led to a rich and fairly complete theory. This means that problems from kinetic theory motivated interesting questions and theoretical results for kernels and associated learning methods. In this chapter, we will find that conversely kernels and their theory are also helpful in kinetic theory itself. The starting point is the investigation of the mean field limit of discrete-time multiagent systems (MAS). In Section 11.1, we provide some background and explain why new results are needed in this context, and why kernels and their theory will be helpful. In Section 11.2, we state and prove two existence results for sequence of functions that are immediately applicable to discrete-time MAS. Finally, in Section 11.3 we apply the results to the standard optimal control setup, and establish a relaxed dynamic programming principle in the mean field.

This chapter is based on and taken to a large extent verbatim from the article [CF5].[1] Detailed comments on the author's contribution are provided in Section 11.5.

[1] © IEEE 2023. Reprinted, with permission, from Christian Fiedler, Michael Herty, and Sebastian Trimpe. *Mean field limits for discrete-time dynamical systems via kernel mean embeddings.* IEEE Control Systems Letters, 2023.

11.1. Introduction

In continuous time, MAS can be described on the microscopic level by ordinary differential equations (ODEs). As an example, let us consider first-order dynamics, as appearing for example when considering alignment. Let $x_i(t) \in \mathbb{R}^d$ be the state of agent $i = 1, \dots, M$ at time $t \geq 0$, and let $u(t) \in \mathbb{R}^m$ be some control input. A very general class of such dynamics is given by the ODE system

$$\begin{aligned} \dot{x}_i &= \frac{1}{M} \sum_{j=1}^{M} \Psi(x_i, x_j)(x_j - x_i) + Bu \\ x_i(0) &= x_i^0 \qquad i = 1, \dots, M \end{aligned} \tag{11.1}$$

with *interaction function* $\Psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, control matrix $B \in \mathbb{R}^{d \times m}$ and initial states $x_1^0, \dots, x_M^0 \in \mathbb{R}^d$. If the number of agents or particle M is very large, then it can be benefitial to go to the mesoscopic level, and model only the evolution of the distribution of the agents over time. One way to do this transition is the mean field limit, cf. Section 8, leading to a partial differential equation (PDE) describing the evolution of the particle distribution.

But what about discrete-time MAS? Typical examples include [58, 221], and they are particularly relevant for engineering applications [41]. On the microscopic level, the direct analogon of the ODE system (11.1) is given by

$$\begin{aligned} x_i^+ &= x_i + \frac{1}{M} \sum_{j=1}^{M} \Psi(x_i, x_j)(x_j - x_i) + Bu \\ x_i(0) &= x_i^0 \qquad i = 1, \dots, M. \end{aligned} \tag{11.2}$$

Formally, we have discrete-time control system of the form

$$\vec{x}^+ = f_M(\vec{x}, u), \tag{11.3}$$

where the transition function is of the form $f_M : X^M \times U \to X^M$, where the state space of the individual agents X and the space of control inputs U does not have to have any special structure a priori. Note that not any function $f : X^M \times U \to X^M$ can be interpreted as the transition function of a discrete-time MAS. Inspecting the structure of systems like (11.2) shows that they should be permutation equivariant

in the state space, i.e., for all $\vec{x} \in X^M$, $u \in U$, and $\sigma \in \mathcal{S}_M$, we need

$$f(\sigma\vec{x}, u) = \sigma f(\vec{x}, u). \tag{11.4}$$

Suppose now that we have a large M, and just as in the continuous time case, we would like to go to the mesoscopic level via the mean field limit. Since discrete-time dynamical systems are completely described by their transition functions, the mean field limit should be formalized on the level of transition functions. Adapting the reasoning from Section 8.2, this can be done as follows. Let X be a measurable space, U some set, and $\mathcal{P}$ a space of probability distributions over X that contains all empirical measures with finitely many atoms, and let $d_{\mathcal{P}}$ be a metric on $\mathcal{P}$. Consider a sequence of functions $f_M : X^M \times U \to X^M$, $M \in \mathbb{N}_+$, that is permutation equivariant in the first argument, and a function $F : \mathcal{P} \times U \to \mathcal{P}$. We say that F is the mean field limit of $(f_M)_M$, or that the latter converges to F in mean field, if

$$\lim_{M\to\infty} \sup_{\substack{\vec{x}\in X^M \\ u\in U}} d_{\mathcal{P}}(\hat{\mu}[f_M(\vec{x}, u)], F(\hat{\mu}[\vec{x}], u)) = 0. \tag{11.5}$$

It is now tempting to adapt the proof strategy from Proposition 8.2.2, or the simpler variant [50, Lemma 1.2], to the present situation. However, one quickly finds a severe difficulty. The functions f_M have an increasing number of inputs *and* outputs. An important step in the proof of [50, Lemma 1.2] is the extension of these functions from $\mathcal{E}_M(X)$ to all Borel probability measures while preserving the uniform Lipschitz bound, which is done with the McShane extension [134]. But this explicit construction works only for functions with a scalar output space, whereas in our case, we have an increasing number of outputs, even growing unboundedly.

We now present a way to circumvent this difficulty. The overall proof strategy remains roughly the same, however, we need to adapt the extension step. Instead of the explicit McShane extension, we use a non-constructive extension result. Since we need to preserve the uniform Lipschitz bound, we choose *Kirszbraun's Theorem*. As is well-known, in general this only works for functions between Hilbert spaces, and apart from a slight weaking, this cannot be overcome [56]. So we need to translate our setting, which involves empirical measures and Borel probability measures, into a Hilbert space setting. But we have already a very convenient tool at hand for this – kernel mean embeddings. Here is now our approach. Instead of more tra-

ditional probability metrics like Kantorowich-Rubinstein, we work with the MMD, which is compatible with the RKHS framework. We embed the empirical measures into an RKHS via kernel mean embeddings, perform the Lipschitz extension via Kirszbraun's theorem, and then project onto the subspace of all kernel mean embeddings in the RKHS, since the Lipschitz extension will in general result in RKHS elements that are not necessarily embeddings of probability distributions. From now on, the rest of the proof works as before, leading to a mean field limit on KMEs. Essentially, we have lifted the microscopic dynamics into an RKHS, where a mean field limit exists. Finally, if we use a characteristic kernel, we can translate this back into the original state space.

11.2. New Mean Field Limit Existence Results

We now present and prove new mean field limit existence results that are tailored to discrete-time control systems.

Systems with control input Our first main result concerns functions of the form of transition functions of discrete-time dynamics with input. We will apply this result to such systems in Section 11.3.

Theorem 11.2.1. Let (X, d_X) be a metric space, H a Hilbert space and $U \subseteq H$ compact, k a measurable, bounded and characteristic kernel such that $\Pi_k(\mathcal{P}(X))$ is compact. Consider a sequence of functions $f_M : X^M \times U \to X^M$, $M \in \mathbb{N}_+$, such that

1. $\forall M \in \mathbb{N}_+ \, \forall \vec{x} \in X^M, \sigma \in \mathcal{S}_M, u \in U\colon \, \hat{\mu}[f_M(\sigma\vec{x}, u)] = \hat{\mu}[f_M(\vec{x}, u)]$
2. $\exists L \in \mathbb{R}_{\geq 0} \, \forall M \in \mathbb{N}_+ \forall \vec{x}, x' \in X^M, \, u, u' \in U$:

$$\gamma_k(\hat{\mu}[f_M(\vec{x}, u)], \hat{\mu}[f_M(\vec{x}', u')]) \leq L \left\| (\Pi_k(\hat{\mu}[\vec{x}]), u) - (\Pi_k(\hat{\mu}[\vec{x}']), u') \right\|_{H_k \times H}, \tag{11.6}$$

where $\| \cdot \|_{H_k \times H}$ is the norm for the product of the Hilbert spaces H_k and H.

Then there exist a subsequence $(f_{M_\ell})_\ell$ and an L-Lipschitz continuous function F :

$\Pi_k\left(\mathcal{P}(X)\right) \times U \to \Pi_k\left(\mathcal{P}(X)\right)$ such that

$$\lim_{\ell\to\infty} \sup_{\vec{x}\in X^{M_\ell}, u\in U} \|\hat{\Pi}_k\left(f_{M_\ell}(\vec{x},u)\right) - F(\hat{\Pi}_k(\vec{x}),u)\|_k = 0 \tag{11.7}$$

and

$$\lim_{\ell\to\infty} \sup_{\vec{x}\in X^{M_\ell}, u\in U} \gamma_k(\hat{\mu}[f_{M_\ell}(\vec{x},u)], f(\hat{\mu}[\vec{x}],u)) = 0, \tag{11.8}$$

where we defined $f(\mu,u) = \Pi_k^{-1}(F(\Pi_k(\mu),u))$.

Proof. Due to property 1) of the f_M, the mappings $\tilde{f}_M : \mathcal{E}_M(X) \times U \to \mathcal{E}_M(X)$ given by

$$\tilde{f}_M\left(\frac{1}{M}\sum_{m=1}^{M}\delta_{x_m}, u\right) = \hat{\mu}[f_M(x_1,\ldots,x_M,u)]$$

are well-defined for all $M \in \mathbb{N}_+$. Furthermore, since k is characteristic, the mappings

$$\tilde{F}_M = \Pi_k \circ \tilde{f}_M \circ \Pi_k|^{-1}_{\Pi_k(\mathcal{E}_M(X))}$$

are well-defined. Let now $M \in \mathbb{N}_+$, $g, g' \in \Pi_k(\mathcal{E}_M(X))$, $u, u' \in U$ be arbitrary, and choose $\vec{x}, \vec{x}' \in X^M$ such that

$$g = \Pi_k\left(\frac{1}{M}\sum_{m=1}^{M}\delta_{x_m}\right), \quad g' = \Pi_k\left(\frac{1}{M}\sum_{m=1}^{M}\delta_{x'_m}\right).$$

We then have

$$\begin{aligned}
\|\tilde{F}_M(g,u) - \tilde{F}_M(g',u')\|_k &= \left\|\hat{\Pi}_k\left(f_M(\vec{x},u)\right) - \hat{\Pi}_k\left(f_M(\vec{x}',u')\right)\right\|_k \\
&= \gamma_k(\hat{\mu}[f_M(\vec{x},u)], \hat{\mu}[f_M(\vec{x}',u')]) \\
&\leq L\left\|(\Pi_k(\hat{\mu}[\vec{x}]),u) - (\Pi_k(\hat{\mu}[\vec{x}']),u')\right\|_{H_k\times H} \\
&= L\left\|(g,u) - (g',u')\right\|_{H_k\times H},
\end{aligned}$$

which shows that all $\tilde{F}_M$ are L-Lipschitz continuous. Since $H_k \times H$ is a Hilbert space, Kirszbraun's theorem ensures that there exist L-Lipschitz continuous mappings $\bar{F}_M : H_k \times H \to H_k$ with $\bar{F}_M|_{\Pi_k(\mathcal{P}(X))\times U} = \tilde{F}_M$, for all $M \in \mathbb{N}_+$. Recall that $\Pi_k(\mathcal{P}(X))$ is convex and by assumption also compact, so there exists the orthogonal projection $P_{\Pi_k(\mathcal{P}(X))}$ from H_k onto $\Pi_k(\mathcal{P}(X))$. Define now for all $M \in \mathbb{N}_+$ the mappings

$F_M : \Pi_k(\mathcal{P}(X)) \to \Pi_k(\mathcal{P}(X))$ by

$$F_M = P_{\Pi_k(\mathcal{P}(X))} \circ \bar{F}_M|_{\Pi_k(\mathcal{P}(X))\times U}.$$

We have for all $M \in \mathbb{N}_+$, $g, g' \in \Pi_k(\mathcal{E}_M(X))$, $u, u' \in U$ that

$$\begin{aligned}\|F_M(g,u) - F_M(g',u')\|_k &= \|P_{\Pi_k(\mathcal{P}(X))}\left(\bar{F}_M(g,u) - \bar{F}_M(g',u')\right)\|_k \\ &\leq \|P_{\Pi_k(\mathcal{P}(X))}\|_{L(H_k)}\|\bar{F}_M(g,u) - \bar{F}_M(g',u')\|_k \\ &\leq L\left\|(g,u) - (g',u')\right\|_{H_k\times H},\end{aligned}$$

where we used that $\|P_{\Pi_k(\mathcal{P}(X))}\|_{L(H_k)} \leq 1$ (since $P_{\Pi_k(\mathcal{P}(X))}$ is a projection). This shows that all F_M are L-Lipschitz continuous. Since $\Pi_k(\mathcal{P}(X)) \times U$ and $\Pi_k(\mathcal{P}(X))$ are compact, we now have a sequence $(F_M)_{M\in\mathbb{N}_+}$ of equicontinuous functions defined on a compact input set, with their range contained in a compact set, so the Arzela-Ascoli theorem asserts that there exist a subsequence $(F_{M_\ell})_\ell$ and an L-Lipschitz continuous function $F : \Pi_k(\mathcal{P}(X)) \to \Pi_k(\mathcal{P}(X))$ with

$$\lim_{\ell\to\infty} \sup_{g\in\Pi_k(\mathcal{P}(X)),u\in U} \|F_{M_\ell}(g,u) - F(g,u)\|_k = 0,$$

which implies that

$$\lim_{\ell\to\infty} \sup_{\vec{x}\in X^{M_\ell},u\in U} \|F_{M_\ell}(\hat{\Pi}_k(\vec{x}),u) - F(\hat{\Pi}_k(\vec{x}),u)\|_k = 0.$$

Observe that for all $M \in \mathbb{N}_+$, $\vec{x} \in X^M$ and $u \in U$ we have

$$\begin{aligned}F_M(\Pi_k(\hat{\mu}[\vec{x}]),u) &= P_{\Pi_k(\mathcal{P}(X))}\left(\bar{F}_M(\Pi_k(\hat{\mu}[\vec{x}]),u)\right) \\ &= P_{\Pi_k(\mathcal{P}(X))}\left(\tilde{F}_M(\Pi_k(\hat{\mu}[\vec{x}]),u)\right) \\ &= P_{\Pi_k(\mathcal{P}(X))}\left(\Pi_k(\hat{\mu}[f_M(\vec{x},u)])\right) \\ &= \Pi_k(\hat{\mu}[f_M(\vec{x},u)]),\end{aligned}$$

which implies

$$\lim_{\ell\to\infty} \sup_{\vec{x}\in X^{M_\ell},u\in U} \|\hat{\Pi}_k(f_{M_\ell}(\vec{x},u)) - F(\hat{\Pi}_k(\vec{x}),u)\|_k = 0,$$

and since Π_k is injective, we can set $f(\mu, u) = \Pi_k^{-1}(F(\Pi_k(\mu), u))$ and get (11.8). $\square$

Feedback maps Our second main result is concerned with functions that can be used as feedback maps for discrete-time control systems, cf. Theorem 11.3.6 for an example of such a situation.

Theorem 11.2.2. Let (X, d_X) be a metric space, $k : X \times X \to \mathbb{R}$ a bounded, Borel-measurable and characteristic kernel on X such that $\Pi_k(\mathcal{P}(X))$ is compact, H a Hilbert space and $C \subseteq H$ a compact and convex subset. Consider maps $g_M : X^M \to C$, $M \in \mathbb{N}_+$, such that

1. $\forall M \in \mathbb{N}_+$, $\vec{x} \in X^M$, $\sigma \in \mathcal{S}_M$, $u \in U$: $g_M(\sigma\vec{x}) = g_M(\vec{x})$
2. $\exists L \in \mathbb{R}_{\geq 0} \forall M \in \mathbb{N}_+$, $\vec{x}, x' \in X^M$:

$$\|g_M(\vec{x}) - g_M(\vec{x}')\|_H \leq L\gamma_k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']). \tag{11.9}$$

Then there exist a subsequence $(g_{M_\ell})_\ell$ and an L-Lipschitz continuous map $G : \Pi_k(\mathcal{P}(X)) \to C$ such that

$$\lim_{\ell\to\infty} \sup_{\vec{x}\in X^{M_\ell}} \|g_{M_\ell}(\vec{x}) - G(\Pi_k(\hat{\mu}[\vec{x}]))\|_H = 0 \tag{11.10}$$

and

$$\lim_{\ell\to\infty} \sup_{\vec{x}\in X^{M_\ell}} \|g_{M_\ell}(\vec{x}) - g(\hat{\mu}[\vec{x}])\|_H = 0, \tag{11.11}$$

where we defined $g = G \circ \Pi_k$.

Remark 11.2.3. Observe that g is also L-Lipschitz continuous as a map on $(\mathcal{P}(X), \gamma_k)$, since for all $\mu, \mu' \in \mathcal{P}(X)$ we have

$$\begin{aligned}\|g(\mu) - g(\mu')\|_H &= \|G(\Pi_k(\mu)) - G(\Pi_k(\mu'))\|_H \\ &\leq L\|\Pi_k(\mu) - \Pi(\mu')\|_k = L\gamma_k(\mu, \mu').\end{aligned}$$

Proof of Theore 11.2.2. The proof follows a similar strategy as used in the proof of Theorem 11.2.1. Since all g_M are permutation invariant, we can define the maps

$\tilde{g}_M : \mathcal{E}_M(X) \to C$ by

$$\tilde{g}_M\left(\frac{1}{M}\sum_{m=1}^{M}\delta_{x_m}\right) = g_M(x_1,\dots,x_M),$$

and since k is characteristic, we can further define $\tilde{G}_M = \tilde{g}_M \circ \Pi_k^{-1}|_{\Pi_k(\mathcal{E}_M(X))}$. Observe that for all $M \in \mathbb{N}_+$ and $\vec{x} \in X^M$ we have by construction $\tilde{G}_M(\Pi_k(\hat{\mu}[\vec{x}])) = g_M(\vec{x})$. Furthermore, for all $M \in \mathbb{N}_+$ and $\vec{x}, \vec{x}' \in X^M$ we have

$$\|\tilde{G}_M(\hat{\Pi}_k(\vec{x})) - \tilde{G}_M(\hat{\Pi}_k(\vec{x}'))\|_H = \|g_M(\vec{x}) - g_M(\vec{x}')\|_H \leq L\gamma_k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']),$$

so Kirszbraun's theorem ensures the existence of L-Lipschitz continuous maps $\bar{G}_M : H_k \to H$ with $\bar{G}_M|_{\Pi_k(\mathcal{E}_M(X))} = g_M$ for all $M \in \mathbb{N}_+$. Since C is compact and convex by assumption, the orthogonal projection P_C exists, and we can define $G_M = P_C \circ \bar{G}_M|_{\Pi_k(\mathcal{P}(X))}$. Since $\|P_C\|_{L(H)} \leq 1$, we have for all $M \in \mathbb{N}_+$, $f_1, f_2 \in \Pi_k(\mathcal{P}(X))$ that

$$\begin{aligned}\|G_M(f_1) - G_M(f_2)\|_H &= \|P_C(\bar{G}_M(f_1) - \bar{G}_M(f_2))\|_H \\ &\leq \|P_C\|_{L(H)}\|\bar{G}_M(f_1) - \bar{G}_M(f_2)\|_H \\ &\leq L\|f_1 - f_2\|_k.\end{aligned}$$

The sequence $(G_M)_M$ is therefore equicontinuous, defined on a compact set (since by assumption $\Pi_k(\mathcal{P}(X))$ is compact) with a compact codomain, so by the Arzela-Ascoli theorem there exist a subsequence $(G_{M_\ell})_\ell$ and an L-Lipschitz continuous map $G : \Pi_k(\mathcal{P}(X)) \to C$ such that

$$\lim_{\ell\to\infty}\sup_{f\in\Pi_k(\mathcal{P}(X))}\|G_{M_\ell}(f) - G(f)\|_H = 0,$$

which implies

$$\lim_{\ell\to\infty}\sup_{\vec{x}\in X^{M_\ell}}\|g_{M_\ell}(\vec{x}) - G(\Pi_k(\hat{\mu}[\vec{x}]))\|_H = 0.$$

Finally, since k is characteristic, we can set $g = G \circ \Pi_k$, and by definition we then have (11.11). □

11.3. Application to Discrete-Time Systems

We now apply our existence results to discrete-time control systems, in particular, large-scale MAS. First, we specify an appropriate setup that ensures the existence of the mean field dynamics and mean field stage cost. Next, we show that in our setting also the corresponding total cost functional has a mean field limit. Finally, we prove a result on relaxed dynamic programming in the mean field limit.

11.3.1. Setup

Let $X \neq \emptyset$ be some set, $k : X \times X \to \mathbb{R}$ a kernel on X and denote by d_k the corresponding kernel metric. From now on, we make the following assumption.

Assumption 11.3.1. The (semi)metric space (X, d_k) is compact, and the kernel k is bounded, Borel-measurable (w.r.t. the Borel σ-algebra on (X, d_k)) and characteristic.

Recall that under these assumptions $\Pi_k(\mathcal{P}(X))$ is compact. Consider now a sequence of discrete-time dynamical systems

$$\vec{x}_+^{[M]} = f^{[M]}(\vec{x}^{[M]}, u),\ M \in \mathbb{N}_+ \tag{11.12}$$

with transition functions $f^{[M]} : X^M \times U \to X^M$, where X^M is the state space and U the input space. Given an initial state $\vec{x}_0^{[M]} \in X^M$ and a control input sequence $\underline{u} \in U^N$, a state-trajectory $\vec{x}^{[M]}(\cdot; \vec{x}_0^{[M]}, \underline{u})$ is induced by

$$\begin{aligned} \vec{x}^{[M]}(0; \vec{x}_0^{[M]}, \underline{u}) &= \vec{x}_0^{[M]} \\ \vec{x}^{[M]}(n+1; \vec{x}_0^{[M]}, \underline{u}) &= f^{[M]}(\vec{x}^{[M]}(n; \vec{x}_0^{[M]}, \underline{u}), \underline{u}(n)), \end{aligned}$$

where $n = 0, \dots, N-1$. We make the following assumption on these systems.

Assumption 11.3.2.

1. $U \subseteq H$ is compact, where H is a Hilbert space.
2. All f_M are permutation equivariant in the state, i.e., $\forall M \in \mathbb{N}_+, \vec{x} \in X^M, \sigma \in \mathcal{S}_M, u \in U :\ f^{[M]}(\sigma\vec{x}, u) = \sigma f^{[M]}(\vec{x}, u)$.
3. The $f^{[M]}$ are uniformly Lipschitz-continuous, i.e., $\exists L_f \in \mathbb{R}_{\geq 0} \forall M \in \mathbb{N}_+, \vec{x}, x' \in$

X^M, $u, u' \in U$:

$$\gamma_k(\hat{\mu}[f^{[M]}(\vec{x}, u)], \hat{\mu}[f^{[M]}(\vec{x}', u')]) \leq L_f \left\| (\hat{\Pi}_k(\vec{x}), u) - (\hat{\Pi}_k(\vec{x}'), u') \right\|_{H_k \times H}.$$

Assumptions 11.3.1 and 11.3.2 together allow to apply Theorem 11.2.1, so there exist a subsequence $(f^{[M_m]})_m$ and a map $F : \Pi_k(\mathcal{P}(X)) \to \Pi_k(\mathcal{P}(X))$ such that

$$\lim_{\ell \to \infty} \sup_{\substack{\vec{x} \in X^{M_m} \\ u \in U}} \|\hat{\Pi}_k \left(f^{[M_m]}(\vec{x}, u) \right) - F(\hat{\Pi}_k(\vec{x}), u)\|_k = 0.$$

This map is also L_f-Lipschitz continuous. Since k is characteristic, we can define the function $f = \Pi_k^{-1} \circ F \circ \Pi_k : \mathcal{P}(X) \times U \to \mathcal{P}(X)$, so

$$\lim_{m \to \infty} \sup_{\vec{x} \in X^{M_m}, u \in U} \gamma_k(\hat{\mu}[f^{[M_\ell]}(\vec{x}, u)], f(\hat{\mu}[\vec{x}], u)) = 0.$$

The map f induces another discrete-time dynamical system

$$\mu_+ = f(\mu, u), \tag{11.13}$$

where $\mathcal{P}(X)$ is the state space and U the input space. A given initial state $\mu_0 \in \mathcal{P}(X)$ and control sequence $\underline{u} \in U^N$ induce a state trajectory $\mu(\cdot; \mu_0, \underline{u})$ by

$$\begin{aligned} \mu(0; \mu_0, \underline{u}) &= \mu_0 \\ \mu(n+1; \mu_0, \underline{u}) &= f(\mu(n; \mu_0, \underline{u}), \underline{u}(n)) \, \forall n = 0, \dots, N-1. \end{aligned}$$

Motivated by optimal control applications, consider a sequence of stage cost functions $\ell^{[M]} : X^M \times U \to \mathbb{R}$, and the associated finite-horizon total cost functionals $J_N^{[M]} : X^M \times U^N \to \mathbb{R}$, $N \in \mathbb{N}_+$, defined by

$$J_N^{[M]}(\vec{x}_0, \underline{u}) = \sum_{n=0}^{N-1} \ell^{[M]}(\vec{x}^{[M]}(n; \vec{x}_0, \underline{u}), \underline{u}(n)). \tag{11.14}$$

We make the following assumption on the stage cost functions $\ell^{[M]}$.

Assumption 11.3.3. 1. All $\ell^{[M]}$ are permutation-invariant in the state variable, i.e., $\forall M \in \mathbb{N}_+, \vec{x} \in X^M, \sigma \in \mathcal{S}_M, u \in U$: $\ell^{[M]}(\sigma\vec{x}, u) = \ell^{[M]}(\vec{x}, u)$.

2. The $\ell^{[M]}$ are uniformly bounded, i.e., $\exists B_\ell \in \mathbb{R}_{\geq 0}\, \forall M \in \mathbb{N}_+, \vec{x} \in X^M, u \in U$: $|\ell^{[M]}(\vec{x}, u)| \leq B_\ell$.

3. The stage cost functions are uniformly Lipschitz-continuous, $\exists L_\ell \in \mathbb{R}_{\geq 0} \forall M \in \mathbb{N}_+, \vec{x}, x' \in X^M, u, u' \in U$

$$|\ell^{[M]}(\vec{x}, u) - \ell^{[M]}(\vec{x}', u')| \leq L_\ell(\gamma_k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']) + \|u - u'\|_H).$$

An inspection of the proof of Theorem 8.2.2 shows that under Assumptions 11.3.1 and 11.3.3 there exist a subsequence $(\ell^{[M_{m_p}]})_p$ and a function $\ell : \mathcal{P}(X) \times U \to \mathbb{R}$ such that

$$\lim_{p\to\infty} \sup_{\vec{x}\in X^{M_{m_p}}, u\in U} |\ell^{[M_{m_p}]}(\vec{x}, u) - \ell(\hat{\mu}[\vec{x}], u)| = 0. \tag{11.15}$$

This can be used to define a corresponding total cost functional $J_N : \mathcal{P}(X) \times U \to \mathbb{R}$, $N \in \mathbb{N}_+$, by

$$J_N(\mu_0, \underline{u}) = \sum_{n=0}^{N-1} \ell(\mu(n; \mu_0, \underline{u}), \underline{u}(n)). \tag{11.16}$$

We now switch to the subsequence $(M_{m_p})_p$ and reindex by M for readability, so that we can write

$$\lim_{M\to\infty} \sup_{\vec{x}\in X^M, u\in U} \gamma_k(\hat{\mu}[f^{[M]}(\vec{x}, u)], f(\hat{\mu}[\vec{x}], u)) = 0 \tag{11.17}$$

$$\lim_{M\to\infty} \sup_{\vec{x}\in X^M, u\in U} |\ell^{[M]}(\vec{x}, u) - \ell(\hat{\mu}[\vec{x}], u)| = 0. \tag{11.18}$$

11.3.2. Mean field limit of $J_N^{[M]}$

Next, we show that J_N is indeed the MFL of $J_N^{[M]}$.

Proposition 11.3.4. For all $N \in \mathbb{N}_+$ we have

$$\lim_{M\to\infty} \sup_{\substack{\vec{x}_0\in X^M \\ \underline{u}\in U^N}} |J_N^{[M]}(\vec{x}_0, \underline{u}) - J_N(\hat{\mu}[\vec{x}_0], \underline{u})| = 0. \tag{11.19}$$

For the proof we need a technical lemma.

Lemma 11.3.5. For all $M \in \mathbb{N}_+$, $\vec{x}_0 \in X^M$, $N \in \mathbb{N}_+$ and $\underline{u} \in U^N$ we have

$$\gamma_k(\hat{\mu}(N), \mu(N) \leq \sum_{n=1}^{N} L_f^{n-1} \|\Pi_k(\hat{\mu}(N-n+1)) - F(\hat{\mu}(N-n), \underline{u}(N-n))\|_k,$$

where we defined for brevity $\hat{\mu}(n) = \hat{\mu}[\vec{x}^{[M]}(n; \vec{x}_0, \underline{u})]$ and $\mu(n) = \mu(n; \hat{\mu}[\vec{x}], \underline{u})$.

This result can be shown using a standard induction argument, and hence the proof is omitted.

Proof (of Proposition 11.3.4). Let $N \in \mathbb{N}_+$, $M \in \mathbb{N}_+$, $\vec{x}_0 \in X^M$ and $\underline{u} \in U^N$ be arbitrary, and define $\vec{x}(n) = \vec{x}^{[M]}(n; \vec{x}_0, \underline{u})$ and $\hat{\mu}(n) = \hat{\mu}[\vec{x}(n)]$, then we have

$$\begin{aligned} |J_N^{[M]}(\vec{x}_0, \underline{u}) - J_N(\hat{\mu}[\vec{x}_0], \underline{u})| &\leq \sum_{n=0}^{N-1} |\ell^{[M]}(\vec{x}(n), \underline{u}(n)) - \ell(\mu(n; \hat{\mu}[\vec{x}_0], \underline{u}), \underline{u}(n))| \\ &\leq \sum_{n=0}^{N-1} |\ell^{[M]}(\vec{x}(n), \underline{u}(n)) - \ell(\hat{\mu}(n), \underline{u}(n))| \\ &\quad + |\ell(\hat{\mu}(n), \underline{u}(n)) - \ell(\mu(n; \hat{\mu}[\vec{x}_0], \underline{u}), \underline{u}(n))|. \end{aligned}$$

Since ℓ is L_ℓ-Lipschitz continuous, we have for all $n = 0, \ldots, N-1$ that

$$\begin{aligned} |\ell(\hat{\mu}(n), \underline{u}(n)) - \ell(\mu(n; \hat{\mu}[\vec{x}_0], \underline{u}), \underline{u}(n))| &\leq L_\ell \gamma_k(\hat{\mu}[\vec{x}^{[M]}(n; \vec{x}_0, \underline{u})], \mu(n; \hat{\mu}[\vec{x}_0], \underline{u})) \\ &\leq L_\ell \sum_{i=1}^{n} L_f^{i-1} \|\hat{\Pi}_k(\vec{x}(n-i+1)) - F(\hat{\mu}(n-i), \underline{u}(n-i))\|_k, \end{aligned}$$

where we used Lemma 11.3.5 in the second inequality.

Combining these bounds results in

$$
\begin{aligned}
&\sup_{\substack{\vec{x}_0 \in X^M \\ \underline{u} \in U^N}} |J_N^{[M]}(\vec{x}_0, \underline{u}) - J_N(\hat{\mu}[\vec{x}_0], \underline{u})| \\
&\quad \leq \sup_{\substack{\vec{x}_0 \in X^M \\ \underline{u} \in U^N}} \sum_{n=0}^{N-1} |\ell^{[M]}(\vec{x}(n), \underline{u}(n)) - \ell(\hat{\mu}(n), \underline{u}(n))| \\
&\qquad + \sup_{\substack{\vec{x}_0 \in X^M \\ \underline{u} \in U^N}} \sum_{n=0}^{N-1} L_\ell \sum_{i=1}^{n} L_f^{i-1} \|\Pi_k(\hat{\mu}(n-i+1)) - F(\hat{\mu}(n-i), \underline{u}(n-i))\|_k, \\
&\quad \leq \sum_{n=0}^{N-1} \sup_{\substack{\vec{x}_0 \in X^M \\ \underline{u} \in U^N}} |\ell^{[M]}(\vec{x}(n), \underline{u}(n)) - \ell(\hat{\mu}(n), \underline{u}(n))| \\
&\qquad + \sum_{n=0}^{N-1} \sum_{i=1}^{n-1} L_\ell L_f^{i-1} \sup_{\substack{\vec{x}_0 \in X^M \\ \underline{u} \in U^N}} \|\Pi_k(\hat{\mu}(n-i+1)) - F(\hat{\mu}(n-i), \underline{u}(n-i))\|_k \\
&\quad \longrightarrow 0 \quad \text{for } M \to \infty.
\end{aligned}
$$

This concludes the proof. □

11.3.3. Relaxed dynamic programming

Relaxed dynamic programming has been frequently used in the analysis of NMPC. We now present a mean field limit variant thereof, which can be used to derive performance bounds for mean field NMPC. This result generalizes [91, Proposition 1] to a wide class of systems and feedback maps.

Theorem 11.3.6. Assume that U is convex. Consider $\tilde{V}_M : X^M \to \mathbb{R}_{\geq 0}$, $M \in \mathbb{N}_+$, such that

1. $\forall M \in \mathbb{N}_+, \vec{x} \in X^M, \sigma \in \mathcal{S}_M : \tilde{V}_M(\sigma\vec{x}) = \tilde{V}_M(\vec{x})$
2. $\exists L_{\tilde{V}} \in \mathbb{R}_{\geq 0}\, \forall M \in \mathbb{N}_+, \vec{x}, x' \in X^M$:

$$
|\tilde{V}_M(\vec{x}) - \tilde{V}_M(\vec{x}')| \leq L_{\tilde{V}} \gamma_k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']). \tag{11.20}
$$

Let $\kappa_M : X^M \to U$, $M \in \mathbb{N}_+$, such that

1. $\forall M \in \mathbb{N}_+, \forall \vec{x} \in X^M, \sigma \in \mathcal{S}_M, u \in U:\ \kappa_M(\sigma\vec{x}) = \kappa_M(\vec{x})$

2. $\exists L_\kappa \in \mathbb{R}_{\geq 0}\, \forall M \in \mathbb{N}_+, \vec{x}, x' \in X^M$:

$$\|\kappa_M(\vec{x}) - \kappa_M(\vec{x}')\|_H \leq L_\kappa \gamma_k(\hat{\mu}[\vec{x}], \hat{\mu}[\vec{x}']). \tag{11.21}$$

Assume that there exists $\alpha \in (0, 1]$ such that for all $M \in \mathbb{N}_+$ and $\vec{x} \in X^M$ we have

$$\tilde{V}_M(\vec{x}) \geq \tilde{V}_M(f^{[M]}(\vec{x}, \kappa_M(\vec{x}))) + \alpha \ell^{[M]}(\vec{x}, \kappa_M(\vec{x})). \tag{11.22}$$

Then there exists a strictly increasing sequence $(M_m)_m$, an $L_{\tilde{V}}$-Lipschitz continuous function $\tilde{V} : (\mathcal{P}(X), \gamma_k) \to \mathbb{R}_{\geq 0}$ and a map $\kappa : \mathcal{P}(X) \to H$ such that for all $\mu \in \mathcal{P}(X)$ we have

$$\tilde{V}(\mu) \geq \tilde{V}(f(\mu, \kappa(\mu))) + \alpha \ell(\mu, \kappa(\mu)). \tag{11.23}$$

Proof. Under the given assumptions, Theorem 11.2.2 is applicable to $(\kappa_M)_M$, so there exist a subsequence $(\kappa_{M_p})_p$ and an L_κ-Lipschitz continuous map $\kappa : \mathcal{P}(X) \to U$ such that $\sup_{\vec{x} \in X^{M_p}} \|\kappa_{M_p}(\vec{x}) - \kappa(\hat{\mu}[\vec{x}])\|_H \to 0$ for $p \to \infty$. Since k is characteristic, we can define $\kappa = K \circ \Pi_k^{-1}|_{\Pi_k(\mathcal{P}(X))}$. An inspection of the proof of Theorem 8.2.2 reveals that it applies to $(\tilde{V}_{M_p})_p$, so there exists a subsequence $(M_{p_m})_m$ and a function $\tilde{V} : \mathcal{P}(X) \to \mathbb{R}_{\geq 0}$ such that $\tilde{V}_{p_m} \xrightarrow{\mathcal{P}_1} \tilde{V}$ for $m \to \infty$, and for all $\mu, \mu' \in \mathcal{P}(X)$ we also have $|\tilde{V}(\mu) - \tilde{V}(\mu')| \leq L_{\tilde{V}} \gamma_k(\mu, \mu')$. To simplify the notation, we denote $(M_{p_m})_m$ by $(M_m)_m$ from now on. Let now $\mu \in \mathcal{P}(X)$ and $\epsilon > 0$ be arbitrary. There exists $\vec{x}_M \in X^M$ such that $\gamma_k(\hat{\mu}[\vec{x}_M], \mu) \to 0$. Define $\hat{\epsilon} = \epsilon/5(L_{\tilde{V}}(1 + L_f(1 + L_\kappa)) + \alpha L_\ell(1 + L_\kappa))$ and choose $m \in \mathbb{N}_+$ such that

$$\begin{aligned}
\gamma_k(\hat{\mu}[\vec{x}], \mu) &\leq \hat{\epsilon} \\
\sup_{\vec{x} \in X^{M_m}} |\tilde{V}_{M_m}(\vec{x}) - \tilde{V}(\hat{\mu}[\vec{x}])| &\leq \frac{\epsilon}{10} =: \epsilon_{\tilde{V}} \\
\sup_{\vec{x} \in X^{M_m}} \|\kappa_{M_m}(\vec{x}) - \kappa(\hat{\mu}[\vec{x}])\|_H &\leq \epsilon/5(L_{\tilde{V}} L_f + \alpha L_\ell) =: \epsilon_\kappa \\
\sup_{\substack{\vec{x} \in X^{M_m} \\ u \in U}} |\ell^{[M_m]}(\vec{x}, u) - \ell(\hat{\mu}[\vec{x}], u)| &\leq \epsilon/5\alpha =: \epsilon_\ell
\end{aligned}$$

and

$$\sup_{\substack{\vec{x}\in X^{M_m}\\ u\in U}} \|\hat{\Pi}_k(f^{[M_m]}(\vec{x},u)) - F(\hat{\Pi}_k(\vec{x}),u)\|_k \leq \epsilon/5L_{\tilde{V}} =: \epsilon_f.$$

We now have

$$\begin{aligned}\|\kappa(\mu) - \kappa_{M_m}(\vec{x}_{M_m})\|_H &= \|\kappa(\mu) - \kappa(\hat{\mu}[\vec{x}_{M_m}])\|_H + \|\kappa(\hat{\mu}[\vec{x}_{M_m}]) - \kappa_{M_m}(\vec{x}_{M_m})\|_H\\ &\leq L_\kappa \gamma_k(\hat{\mu}[\vec{x}_{M_m}],\mu) + \epsilon_\kappa\\ &\leq L_\kappa\hat{\epsilon} + \epsilon_\kappa\end{aligned}$$

and

$$\begin{aligned}&\gamma_k\left(\hat{\mu}\left[f^{[M_m]}(\vec{x}_{M_m},\kappa_{M_m}(\vec{x}_{M_m}))\right], f(\mu,\kappa(\mu))\right)\\ &\quad\leq \|F(\mu,\kappa(\mu)) - F(\hat{\mu}[\vec{x}_{M_m}],\kappa_{M_m}(\vec{x}_{M_m}))\|_k + \|F(\hat{\mu}[\vec{x}_{M_m}],\kappa_{M_m}(\vec{x}_{M_m}))\\ &\qquad - \Pi_k\left(f^{[M_m]}(\vec{x}_{M_m},\kappa_{M_m}(\vec{x}_{M_m}))\right)\|_k\\ &\quad\leq L_f\,\|(\Pi_k(\mu),\kappa(\mu)) - (\Pi_k(\hat{\mu}[\vec{x}_{M_m}],\kappa_{M_m}(\vec{x}_{M_m}))\|_{H_k\times H}\\ &\qquad + \sup_{\substack{\vec{x}\in X^{M_m}\\ u\in U}} \|\Pi_k(f^{[M_m]}(\vec{x},u)) - F(\Pi_k(\hat{\mu}[\vec{x}]),u)\|_k\\ &\quad\leq L_f\left(\gamma_k(\hat{\mu}[\vec{x}_{M_m}],\mu) + \|\kappa_{M_m}(\vec{x}_{M_m}) - \kappa(\mu)\|_H\right) + \epsilon_f\\ &\quad\leq L_f\hat{\epsilon} + L_f(L_\kappa\hat{\epsilon} + \epsilon_\kappa) + \epsilon_f\end{aligned}$$

as well as $\tilde{V}(\mu) \geq \tilde{V}_{M_m}(\vec{x}_{M_m}) - \epsilon_{\tilde{V}} - L_{\tilde{V}}\hat{\epsilon}$. Finally,

$$\begin{aligned}\tilde{V}(\mu) &\geq \tilde{V}_{M_m}(\vec{x}_{M_m}) - \epsilon_{\tilde{V}} - L_{\tilde{V}}\hat{\epsilon}\\ &\geq \tilde{V}_{M_m}(f^{[M_m]}(\vec{x}_{M_m},\kappa_{M_m}(\vec{x}_{M_m}))) + \alpha\ell^{[M_m]}(\vec{x}_{M_m},\kappa_{M_m}(\vec{x}_{M_m})) - \epsilon_{\tilde{V}} - L_{\tilde{V}}\hat{\epsilon}\\ &\geq \tilde{V}(\hat{\mu}[f^{[M_m]}(\vec{x}_{M_m},\kappa_{M_m}(\vec{x}_{M_m}))]) + \alpha\ell(\hat{\mu}[\vec{x}_{M_m}],\kappa_{M_m}(\vec{x}_{M_m})) - 2\epsilon_{\tilde{V}} - \alpha\epsilon_\ell - L_{\tilde{V}}\hat{\epsilon}\\ &\geq V(f(\mu,\kappa(\mu))) + \alpha\ell(\mu,\kappa(\mu)) - L_{\tilde{V}}\gamma_k\left(\hat{\mu}\left[f^{[M_m]}(\vec{x}_{M_m},\kappa_{M_m}(\vec{x}_{M_m}))\right], f(\mu,\kappa(\mu))\right)\\ &\qquad - \alpha L_\ell\left(\gamma_k(\hat{\mu}[\vec{x}],\mu) + \|\kappa(\mu) - \kappa_{M_m}(\vec{x}_{M_m})\|_k\right) - 2\epsilon_{\tilde{V}} - \alpha\epsilon_\ell - L_{\tilde{V}}\hat{\epsilon}\\ &\geq \tilde{V}(f(\mu,\kappa(\mu))) + \alpha\ell(\mu,\kappa(\mu)) - L_{\tilde{V}}\left(L_f\hat{\epsilon} + L_f(L_\kappa\hat{\epsilon} + \epsilon_\kappa) + \epsilon_f\right)\\ &\qquad - \alpha L_\ell\left(\hat{\epsilon} + L_\kappa\hat{\epsilon} + \epsilon_\kappa\right) - 2\epsilon_{\tilde{V}} - \alpha\epsilon_\ell - L_{\tilde{V}}\hat{\epsilon}\\ &= \tilde{V}(f(\mu,\kappa(\mu))) + \alpha\ell(\mu,\kappa(\mu)) - \epsilon.\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we get (11.23). □

11.4. Conclusion

We would like to close with some concluding remarks. First, to the best of our knowledge, the mean field limit existence results in Section 11.2 are the first result that generalize the well-known [50, Lemma 1.2] to functions with essentially non-scalar output spaces. Second, in this way we have also achieved, to the best of our knowledge, the first existence results for the mean field limit of discrete-time multiagent systems. In contrast to the continuous-time situation, we do not arrive at an explicit description of the mean field dynamics, so at present, we have a pure existence result.

The developments in this chapter can be the starting point of many promising additional investigations. As already mentioned in the introduction, to establish the existence of the mean field limit, we have lifted the dynamics into an RKHS, where also the mean field limit dynamics act. It seems therefore very promising to use kernel methods to solve numerical problems associated with the mean field dynamics, like simulation or control. Furthermore, the proof strategy from Section 11.2 works actually with any suitable Hilbertian embedding of probability distributions. It would be interesting to investigate other embeddings which might offer advantages over KMEs. Since so far we worked with a very general class of systems, it would also be interesting to specialize our results to more concrete system classes. Finally, since we now have a rigorous mean field limit existence result for discrete-time MAS available, the investigation of further control and system-theoretic questions in this context are interesting.

11.5. Comments

This chapter is based on, and to a large extent taken verbatim from, the article [CF5], which was published in the IEEE Control Systems Letters, and presented at the American Control Conference 2024. All results are due to the author of this thesis, who would like to thank P.-F. Massiani for pointing out a problem with the original formulation of Theorem 11.2.1.

12. Conclusions

This thesis presented a broad range of contributions to kernel methods and their theory in the context of systems and control. In this concluding chapter, we will briefly summarize and discuss this work, and provide some pointers to future directions.

Summary To provide a solid foundation, in Part I of this thesis we present some background on kernels and RKHSs. Our introduction to these topics in Chapter 2 does not only serve as background material for this thesis, but also aims at closing a gap in the literature by providing a perspective-agnostic exposition. Furthermore, Chapter 3 contains a comprehensive treatment of Lipschitz and Hölder continuity in RKHSs, which collects and improves many known results, and presents some new ones.

In Part II, the focus is on uncertainty bounds for kernel methods and their use in learning-based control, which is the first major topic of this thesis. The general discussion in Chapter 4 suggests that in the context of learning-based control, especially when using establish robust control methods, frequentist uncertainty bounds or alternatively worst-case uncertainty bounds should be used. Furthermore, as is well-known but often not clearly articulated in the literature, for their use in learning-based control these bounds need to be numerically computable, not too conservative, and based on reasonable prior knowledge. This leads us to frequentist uncertainty bounds for GP regression and kernel ridge regression, as outlined in Chapter 5. We collect and refine several such bounds, and provide an elementary derivation of the well-known uncertainty bound based on self-normalization. Furthermore, we state and prove new bounds that are robust to certain classes of model misspecifications. In turn, in Chapter 6 the uncertainty bounds are then empirically evaluated and actually used in learning-based control applications. We conduct thorough numerical experiments to evaluate the uncertainty bounds which indicate that they are tight enough to be useful in practice, which is also confirmed by a

first, simple learning-based control application based on MPC. In addition, we also use them together with modern robust controller synthesis. The resulting modular approach allows the seamless integration of prior knowledge and machine learning components in an established framework of modern control, while retaining rigorous statistical and control-theoretic guarantees. However, in order to be useful in practice, uncertainty bounds should only rely on reasonable prior knowledge, ideally in the form of assumptions that are meaningful and clearly interpretable by users. This prompts us to revisit in Chapter 7 the foundations of uncertainty bounds for kernels methods. Essentially all frequentist uncertainty bounds rely on the knowledge of a concrete upper bound on the RKHS norm of the target function, and we argue that this forms a severe obstacle to the practical applicability of these approaches, since at the moment it is appears very difficult, if not impossible to derive such a bound from established prior knowledge in non-trivial situations. While this problem is known to some extent in the learning-based control literature, the severity appears to be underappreciated, and we provide the most thourough discussion of this issue so far. To overcome this issue, we propose to rely on geometric assumptions for the uncertainty sets, since many relevant forms of prior knowledge can be encoded in this way, and some special cases like Lipschitz bounds are already very established in systems and control. As a first step in this direction, we show how the resulting uncertainty sets can be combined with kernel methods by using hard shape constrained kernel machines. Our results and insights from Part II demonstrate the power of kernel methods in the context of learning-based control with rigorous guarantees, highlight practically relevant limitations, and point to novel approaches to overcome the latter.

Uncertainty bounds for kernel methods and their use in learning-based control is an established and active topic, and our contributions in Part II of this thesis are about revisiting and consolidating the foundations of this area. In contrast, in Part III of this thesis we turn to a novel and underexplored topic: kernels and kernel methods in the mean field limit. We start in Chapter 8 with a gentle introduction to mean field limits and the relevance of this concept, and motivate the connection to kernels and kernel methods. In Chapter 9, we introduce the notion of a mean field limit of kernels, prove an existence result, and present and investigate a large class of suitable kernels. Furthermore, we provide an essentially complete theoretical description of the interaction between the mean field limit of kernels and their

RKHSs. These results allow us to investigate in Chapter 10 kernel-based statistical learning methods in the mean field limit. We establish a representer theorem in the mean field limit, and show that the risks and learning outcomes of regularized empirical risk minimization schemes converge in the mean field limit. From a practical perspective, this justifies switching between the microscopic and mesoscopic levels in kernel-based learning methods on large-scale multiagent systems, analogously to the situation of numerical methods in kinetic theory. From a theoretical perspective, we introduce and investigate a new large-scale limit in theoretical machine learning, complementing many existing approaches for such limits. The results in Chapters 9 and 10 show that kinetic theory motivates interesting problems in kernel-based statistical learning with a rich theory. Chapter 11 demonstrates that conversely kernels and their theory can be useful in kinetic theory. Using kernel mean embeddings, we are able to prove an existence result for the mean field limit of very general discrete-time multiagent systems, the first such result to the best of our knowledge. As a first application, we formulate in a rigorous manner the standard optimal control setup in the mean field limit, and prove a relaxed dynamic programming principle in this setting. In summary, in Part III we uncover very fruitful connections between kernels and kinetic theory, with many new theoretical results and lots of opportunities for further investigations and new applications of kernel methods.

Future work To close, we would like to point out some open questions and interesting directions for future work, starting with uncertainty bounds for kernel methods as outlined in Chapters 4 and 5. In Section 5.5 we stated and proved uncertainty bounds that are robust to some model misspecifications, focusing on lengthscale and unstructured, but bounded kernel misspecifications. Using the techniques from [47], these results should be easy to extend to more general classes of hyperparameter misspecficiations. Similarly, analogous robust uncertainty bounds should be possible in a worst-case setting with bounded additive noise, by combining the results from [127, 176] with our techniques and the ones from [47]. Another avenue for future work are uncertainty bounds for more general measurement models, like derivatives or linear integral functionals. The latter should lead in a straightforward manner to frequentist (and possibly worst-case) uncertainty bounds for kernel-based linear system identification. Furthermore, it should be possible to establish all results in Chapter 5 and the generalizations outlined here also for vector RKHSs, cf. [152,

Chapter 6]. For example, using the results in [142], the simple uncertainty bounds from Sections 5.1 and 5.2 can be immediately generalized to this setting with essentially the same arguments. Finally, the question of tightness of the bounds should be further investigated. For example, when not the target function itself is of interest, but rather a finite dimensional projection thereof, using techniques from optimal design of experiments can result in tighter uncertainty sets [145]. Furthermore, it would be interesting to investigate how much tighter the uncertainty bounds can be made. In the case of bounded noise, it is possible to derive such bounds at the expense of relying on an optimization problem instead of a closed-form solution [176], and it is a very interesting problem whether this could be achieved for stochastic noise. A similar question appears for time-uniform bounds, and in some situations a tight bound can be provided [95].

Regarding the empirical evaluation of uncertainty bounds as started in Chapter 6, further experiments with synthetic target functions are another interesting direction. This includes more kernels, more RKHS function sampling methods (e.g., using other orthonormal bases if available), and additional noise distributions. Furthermore, our refined versions of the simple uncertainty bounds from Sections 5.1 and 5.2 should be now competitive with the standard time-uniform bounds, and should be included in such an empirical evaluation. Similarly, the additional theoretical results outlined above should be also systematically evaluated with numerical experiments. All of these generalizations and improvements can be immediately transferred to applications in learning-based control.

In addition, the methodology for learning-enhanced robust controller synthesis in Section 6.2 points to many directions for future work. For example, it can be applied to a more complex control scenario, where its advantages – seamless integration of prior knowledge and learning components, state-of-the-art robust controller synthesis methods, providing rigorous control-theoretical and statistical guarantees – become even more apparent. Furthermore, the integration of the learning component for static nonlinearities can be improved, by tailoring the uncertainty sets to the form of the IQC multipliers, which is a challenging but very promising direction. Moreover, frequentist uncertainty bounds for kernel-based linear system identification, as described above, would allow the generalization of the methodology to dynamic uncertainties.

Finally, Chapter 7 provides also many interesting open questions and directions

for future work. First, as discussed there, at present it appears that deriving quantitative RKHS norm bounds from common engineering prior knowledge is not possible with existing methods in non-trivial situations. Nonetheless, given the importance of this problem and the fact that a large amount of work in learning-based control is affected by this issue, it should be further investigated. We suspect that at least in the context kernel-based linear system identification it might be possible to relate the RKHS norm of suitable kernels to frequency domain properties, which are well-established in control theory [65]. Second, more complex examples of geometric constraints and their relation to practical prior knowledge should be investigated and used in kernel machines. Third, and perhaps the most important aspect, the initial uncertainty bounds derived from geometric assumptions should be improved. This includes the development of algorithms that can automatically compute such uncertainty bounds, replacing manual derivation on a case-by-case basis, as well as the the generalization to stochastic and potentially unbounded noise.

Turning to Part III of this thesis, despite our extensive theoretical investigations, many interesting open questions remain for kernels and kernel methods in the context mean field limits. In the following, we point out just some of them. In Chapter 9 we present a rather complete theory for kernels and their RKHSs in the mean field limit. Ongoing work is concerned with the generalization to vRKHSs, and initial results indicate that indeed most of the theory generalizes immediately to this setting. In terms of concrete kernels in the mean field limit, only the case of double sum kernels has been investigated in detail, and it would be interesting to extend this to other classes of suitable kernels. In the context of statistical learning with kernels in the mean field limit, cf. Chapter 10, two questions are of particular interest. First, one motivation for the mean field limit in kinetic theory is given by the availability of suitable numerical methods for kinetic PDEs [151]. This poses the question whether such methods can be adapted to the case of kernels in the mean field limit and associated learning problems. Second, the convergence results for regularized empirical risk minimization in Chapter 10 are all in the setting of finite data sets, but in statistical learning theory, a major aspect is the investigation of the limit of infinitely many i.i.d. data points, leading to questions of consistency and learning rates [189, Chapter 6]. It is therefore very interesting to investigate the case of a double limit, with an increasing number of particles and an increasing number of data points in the data set, and ideally provide consistency and learning

rate results in the mean field limit.

Finally, also Chapter 11 leads to many interesting directions for further investigations. Ongoing work is concerned with generalizing the results to other types of Hilbertian embeddings, cf. also Chapter A, and various classes of multiagent systems, as well as using these to establish control-theoretic results, for example related to controllability and dissipativity. Furthermore, while we have used kernel mean embeddings in Chapter 11 for theoretical reasons, it turns out that our proof strategy actually amounts to lifting the microscopic dynamic to mesoscopic dynamics in RKHSs. This immediately suggests to use establish kernel methods to solve numerical tasks like simulation or control. Investigating this approach appears to be a very promising direction for future work.

Appendix

A. Towards statistical learning theory with distributional inputs

This thesis is focused on kernel methods in the context of systems and control. However, during the work leading to this thesis, we could also achieve new theoretical results in the area of kernel-based statistical learning with distributional inputs. While not directly related to systems and control, the techniques we are utilizing are closely related to the work reported in this thesis, so we include this additional work here.

Apart from minor modifications, this chapter is taken verbatim from the work [CF7].

A.1. Introduction

Supervised statistical learning with distributional inputs has received significant attention lately, cf. [196, 70, 137], both from practical and theoretical perspectives. The goal is to learn a relation between inputs and outputs from data, where the inputs are probability distributions on some measurable space. Furthermore, the inputs (probability distributions) are not directly accessible, but the data contains only samples thereof. A classic example is the prediction of some health indicator of a patient from several clinical measurements [195], which we recall now. Let $\mathcal{S}$ be the set of outcomes of some diagnosis tools (e.g., electrocardiogram characteristics, or the blood serum concentration of some substance). Since these measurements will vary even when coming from the same patient, it is reasonable to assume that an individual patient with a specific health status has a certain distribution Q on $\mathcal{S}$ that generates the measurements and that can be a predictor for some health indicator $y \in \mathcal{Y}$ (e.g., healthy or not). However, during training, Q is not directly accessible, but rather through independent and identically distributed (i.i.d.) samples

$S_1, \dots, S_M \overset{\text{i.i.d.}}{\sim} Q$. For example, this could correspond to daily blood measurements of a patient during a week-long hospital stay, assuming the patient's distribution Q has not changed during the week (e.g., when the health status has not changed). The training data consists of such data from N different patients, so the data set is not of the form $\bar{\mathcal{D}} = ((Q_n, y_n))_{n=1,\dots,N}$, but rather $\mathcal{D} = ((S^{(n)}, y_n))_{n=1,\dots,N}$, where $S_1^{(n)}, \dots, S_{M_n}^{(n)} \overset{\text{i.i.d.}}{\sim} Q_n$. The goal is to learn a map $f_{\mathcal{D}}$ from distributions Q over $\mathcal{S}$ to outcomes $\mathcal{Y}$ (e.g., from distributions over diagnostic measurements to health status).

Among such learning problems, the focus of previous theoretical investigations has been on *distributional regression*. In this setting, one is interested in predicting a real-valued quantity from a distributional input, so $\mathcal{Y} = \mathbb{R}$. While the early work [162] relied on density estimation, starting with [195], kernel mean embeddings (KMEs) together with kernel ridge regression (KRR) have been used. For concreteness, let us review this latter approach. Consider a data set $\mathcal{D}$ as introduced above. A single input item $S^{(n)}$ is first interpreted as an empirical measure $\hat{\mu}[S^{(n)}] = \frac{1}{M_n} \sum_{m=1}^{M_n} \delta_{S_m^{(n)}}$, where δ_s is the Dirac measure with atom on s, which is then embedded into a reproducing kernel Hilbert space (RKHS) H_κ using the KME, $\hat{\mu}[S^{(n)}] \mapsto \Pi_k \hat{\mu}[S^{(n)}]$. Assuming access to a (second) kernel k on the RKHS H_κ, one then performs KRR on the transformed data set $\mathcal{D}_{\hat{\Pi}_k} = ((\Pi_k \hat{\mu}[S^{(n)}], y_n))_{n=1,\dots,N}$. The resulting learned function $f_{\mathcal{D}_{\hat{\Pi}_k}}$ can then be used for prediction by composing it with the KME map. A distribution Q on $\mathcal{S}$ would therefore lead to prediction $f_{\mathcal{D}_{\hat{\Pi}_k}}(\Pi_k Q)$. This approach has been thoroughly analyzed [195, 196, 70]. All of these investigations rely on the seminal analysis [48] of the regularized least-squares algorithm for regression.

Recently, [137] developed a much more general perspective on this problem. Instead of KMEs, they consider suitable embeddings Π of probability distributions into some Hilbert space $\mathcal{H}$, and then utilize distance substitution kernels [88] with the induced Hilbertian (semi)metric $(P, Q) \mapsto \|\Pi P - \Pi Q\|_{\mathcal{H}}$ on probability distributions. In particular, they apply this construction to sliced Wasserstein (SW) distances [36] and construct corresponding SW kernels. The resulting method has been theoretically analyzed, building again on [48].

Despite this multitude of activity, many interesting and practically relevant problems in this area are still open. In this work, we focus on two theoretical aspects.

First, most theoretically grounded works in the context of distributional learning methods have focused almost exclusively on the distributional regression problem. However, other learning scenarios are also highly relevant, in particular, distributional classification. For example, in the medical example outlined above, a natural task is to predict a binary health status of a patient (e.g., having a certain disease or not), corresponding to (distributional) binary classification. As another example, in [123], distributional classification is applied to the problem of predicting causal directions and causal graphs from data. KMEs are also used there, though empirical risk minization (ERM) is then applied on the transformed data set. To the best of our knowledge, this reference is also the only one investigating distributional classification with KMEs from a theoretical perspective. While they establish some generalization bounds based on margin theory, no consistency results or oracle inequalities in the sense of [189] are provided. Recall that an oracle inequality in this context is a high probability bound on the excess risk of the learned hypothesis over what an oracle, that has access to the true data-generating distribution, can achieve. In turn, these inequalities allow derivation of consistency results, and under suitable distributional assumptions also of learning rates, and it is hence highly desirable that such inequalities are also available in the distributional learning setting. *Second*, the theoretical analyses of distributional learning have been restricted to rather specific settings. Even in the context of distributional regression, the learning setups considered have been fairly specific. In particular, in the context of KME-based distributional regression, to the best of our knowledge only KRR has been considered so far, and analyzed exclusively using [48]. This technique is inherently limited to KRR, and hence cannot be used to analyze inter alia support vector regression (SVR) with the ϵ-insensitive loss. It is furthermore also not suitable to analyze classification using support vector machines (SVMs), or more general regularized empirical risk approaches.

We therefore tackle these open issues. First, for the distributional learning setting outlined above, we provide two oracle inequalities for the risk of SVMs (in the sense of regularized risk minimization over RKHSs) that cover a multitude of learning scenarios. To the best of our knowledge, both of these results are completely new in the context of learning on distributional inputs. Second, we establish a generalization bound for distributional learning based on algorithmic stability, and apply it to SVMs. Third, inspired by [137], we formulate all of this for kernel-based methods

that rely on a general Hilbertian embedding of probability distributions. In this manner, our results apply to the case of KMEs and SW kernels, and *any future method that provides Hilbertian embeddings.* Fourth, we specialize our results to KMEs and SW distances for the Hilbertian embedding.

A.2. Distributional Learning Setup

In this section, we introduce necessary technical background, and formalize the learning setup that we consider in the following.

Preliminaries For a measurable space $(\mathcal{Z}, \mathcal{A}_\mathcal{Z})$, we denote the set of all probability distributions on it by $\mathcal{M}_1(\mathcal{Z})$, suppressing the σ-algebra if no confusion can arise, and the set of measurable real-valued functions is denoted by $\mathcal{L}^0(\mathcal{Z})$. If $(\mathcal{X}, \mathcal{A}_\mathcal{X})$, $(\mathcal{Y}, \mathcal{A}_\mathcal{Y})$ are measurable spaces, $f : \mathcal{X} \to \mathcal{Y}$ is a measurable map, and $\mu \in \mathcal{M}_1(\mathcal{X})$ is a probability measure, then the *pushforward of μ along f* is defined as $f\sharp\mu(A') = \mu(f^{-1}(A'))$ for all $A' \in \mathcal{A}_\mathcal{Y}$. For a topological space $(\mathcal{X}, \tau_\mathcal{X})$, we denote the associated Borel σ-algebra by $\mathcal{B}(\tau_\mathcal{X})$. Given $\mu_n, \mu \in \mathcal{M}_1(\mathcal{X})$, $n \in \mathbb{N}_+$, we say that $(\mu_n)_n$ converges weakly[1] to μ, if for all bounded and continuous $f : \mathcal{X} \to \mathbb{R}$, we have $\int_\mathcal{X} f(x)\mathrm{d}\mu_n(x) \to \int_\mathcal{X} f(x)\mathrm{d}\mu(x)$. This induces a topology τ_w on $\mathcal{M}_1(\mathcal{X})$, called the topology of weak convergence. If $(\mathcal{X}, \|\cdot\|_\mathcal{X})$ is a normed space, we define $\mathcal{B}(\mathcal{X})$ as the Borel σ-algebra generated by the open sets w.r.t. to the topology induced by the norm.

We use the notation and terminology on kernels as used in the rest of this thesis. In addition, we also define $\|k\|_\infty = \sup_{x\in\mathcal{X}} \sqrt{k(x,x)}$, and call k bounded if it is bounded as a map on $\mathcal{X} \times \mathcal{X}$, which is the case if and only if $\|k\|_\infty < \infty$.

Furthermore, in order to balance generality and simplicity of notation, we use *comparison functions*, a very successful formalism in control theory [105]. Define $\mathcal{K}$ as the set of continuous functions $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that $\alpha(0) = 0$ and α is strictly increasing. Operations and relations are declared pointwise on $\mathcal{K}$, so for $\alpha_1, \alpha_2 \in \mathcal{K}$, $\alpha_1 \leq \alpha_2$ means that $\alpha_1(s) \leq \alpha_2(s)$ for all $s \in \mathbb{R}_{\geq 0}$. Note that $\mathcal{K}$ is closed under addition and scalar multiplication with positive real numbers. Finally, we call $(\alpha_B)_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ a *nondecreasing family*, if $\alpha_a \leq \alpha_b$ for all $0 < a \leq b < \infty$.

Furthermore, we use the basic setup of statistical learning theory, as presented in

[1] In the sense of probability theory, not functional analysis.

Section 10.2. For the reader's convience, we recall this here, and add some additional material that is necessary for the following developments. Let $\mathcal{X}$ be a measurable space, and let $\emptyset \neq \mathcal{Y} \subseteq \mathbb{R}$ be closed. A *loss function* $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a measurable function. We call ℓ convex, differentiable, etc., if for all $x \in \mathcal{X}, y \in \mathcal{Y}$ the function $\ell(x, y, \cdot)$ has the corresponding property. If ℓ is locally Lipschitz continuous, we define for all $B \in \mathbb{R}_{>0}$

$$|\ell|_{1,B} = \sup_{\substack{t_1,t_2 \in [-B,B] \\ t_1 \neq t_2 \\ x \in \mathcal{X}, y \in \mathcal{Y}}} \frac{|\ell(x,y,t_1) - \ell(x,y,t_2)|}{|t_1 - t_2|}. \tag{A.1}$$

Given $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ and $f : \mathcal{X} \to \mathbb{R}$ measurable, we define the *risk* $\mathcal{R}_{\ell,P}(f) = \int \ell(x, y, f(x)) \mathrm{d}P(x, y)$ and the *Bayes risk* $\mathcal{R}^*_{\ell,P} = \inf_{f \in \mathcal{L}^0(\mathcal{X})} \mathcal{R}_{\ell,P}(f)$. Let k be a kernel on $\mathcal{X}$ such that all functions in H_k are measurable. For $f \in H_k$ and a *regularization parameter* $\lambda \in \mathbb{R}_{>0}$, we define the *regularized risk* $\mathcal{R}_{\ell,P,\lambda}(f) = \mathcal{R}_{\ell,P}(f) + \lambda \|f\|_k^2$, as well as $\mathcal{R}^{H_k*}_{\ell,P,\lambda} = \inf_{f \in H_k} \mathcal{R}_{\ell,P,\lambda}(f)$ and $\mathcal{R}^{H_k*}_{\ell,P} = \inf_{f \in H_k} \mathcal{R}_{\ell,P}(f)$. Additionally, if $\mathcal{R}^{H_k*}_{\ell,P} < \infty$, define the *approximation error function* $A^{(2)}_{\ell,P} : \mathbb{R}_{\geq 0} \to \mathbb{R}$ by

$$A^{(2)}_{\ell,P}(\lambda) = \mathcal{R}^{H_k*}_{\ell,P,\lambda} - \mathcal{R}^{H_k*}_{\ell,P}. \tag{A.2}$$

Furthermore, define the *empirical risk* of a function $f \in H_k$ w.r.t. data $D = ((x_n, y_n))_{n=1,\ldots,N} \in (\mathcal{X} \times \mathcal{Y})^N$ by $\mathcal{R}_{\ell,D}(f) = \frac{1}{N} \sum_{n=1}^N \ell(x_n, y_n, f(x_n))$, and the *regularized empirical risk* $\mathcal{R}_{\ell,D,\lambda}(f) = \mathcal{R}_{\ell,D}(f) + \lambda \|f\|_k^2$. If it exists, a solution to the optimization problem

$$\min_{f \in H_k} \mathcal{R}_{\ell,D,\lambda}(f) \tag{A.3}$$

is called an *(empirical) SVM solution* and we denote it by $f_{D,\lambda}$. Similarly, if a solution to the optimization problem

$$\min_{f \in H_k} \mathcal{R}_{\ell,P,\lambda}(f) \tag{A.4}$$

exists, we called it an *infinite-sample SVM solution* or just *SVM solution*, and denote it by $f_{P,\lambda}$.

Two-stage sampling setup We now introduce the concrete distributional learning setup that we consider, roughly following [195] and [137]. *Unless noted otherwise,*

this will be the setup that we use in the remainder of this work. Let $(\mathcal{S}, \tau_{\mathcal{S}})$ be a topological space and consider the set of Borel probability measures $\mathcal{M}_1(\mathcal{S})$ on $\mathcal{S}$. Let τ_w be the topology induced by weak convergence in $\mathcal{M}_1(\mathcal{S})$, and consider the measurable space $(\mathcal{M}_1(\mathcal{S}), \mathcal{B}(\tau_w))$.

Let $\mathcal{H}$ be a Hilbert space, which we endow with the Borel σ-algebra $\mathcal{B}(\mathcal{H})$, let $\Pi : \mathcal{M}_1(\mathcal{S}) \to \mathcal{H}$ be some map, and define the Hilbertian semimetric $\mathrm{d}_{\mathcal{H}}(P, Q) = \|\Pi(P) - \Pi(Q)\|_{\mathcal{H}}$, and the set $\mathcal{X} = \Pi(\mathcal{M}_1(\mathcal{S}))$. Additionally, we assume access to a family of maps $(\hat{\Pi}_M)_{M\in\mathbb{N}_+}$ with $\hat{\Pi}_M : \mathcal{S}^M \to \mathcal{X}$, and we define $\mathcal{S}^* = \bigcup_{M\in\mathbb{N}_+} \mathcal{S}^M$ and $\hat{\Pi} : \mathcal{S}^* \to \mathcal{X}$ by $\hat{\Pi}(S) = \hat{\Pi}_M(S)$ for all $S \in \mathcal{S}^M$ and $M \in \mathbb{N}_+$. The usual example is $\hat{\Pi}(S) = \Pi\left(\frac{1}{M}\sum_{m=1}^{M} \delta_{S_m}\right)$ for $S \in \mathcal{S}^M$ and all $M \in \mathbb{N}_+$. However, our setup allows also different choices. For the analysis of this setting, measurability of various components needs to be ensured, for which the following assumption can be invoked.

Assumption A.2.1. $\mathcal{H}$ is separable, Π is $\mathcal{B}(\tau_w)$-$\mathcal{B}(\mathcal{H})$-measurable, and $\mathcal{X} \in \mathcal{B}(\mathcal{H})$. Furthermore, for all $M \in \mathbb{N}_+$, $\hat{\Pi}_M$ is $\mathcal{B}(\tau_{\mathcal{S}})^{\otimes M}$-$\mathcal{B}(\mathcal{X})$-measurable.

The following technical result now ensures that we can apply the usual statistical learning theory setup.

Lemma A.2.2. Under Assumption A.2.1, the map Π is $\mathcal{B}(\tau_w)$-$\mathcal{B}(\tau_{\mathcal{H}}|_{\mathcal{X}})$-measurable, where $\tau_{\mathcal{H}}|_{\mathcal{X}}$ is the subspace topology on $\mathcal{X}$ induced by the topology on $\mathcal{H}$. Furthermore, every $P \in \mathcal{M}_1(\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y})$ induces a distribution P_Π on $\mathcal{X} \times \mathcal{Y}$ as the pushforward of P along $(Q, y) \mapsto (\Pi Q, y)$.

A proof of this result is provided in Section A.1.1 in [195] and the supplementary to [123]. For the remainder of this subsection, we work under Assumption A.2.1. Let P be a probability distribution on $\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y}$ (often called *meta-distribution*), *and to ease the notational load, in the following we will use P also for the push-forward*[2] *P_Π, if no confusion can arise.* Furthermore, we assume the following data-generating model. We sample $(Q_1, y_1), \ldots, (Q_N, y_N)$ i.i.d. from P, and for each $n = 1, \ldots, N$, we assume that $S^{(n)} \sim Q_n^{\otimes M_n}$ for some $M_n \in \mathbb{N}_+$, and that $S^{(1)}, \ldots, S^{(N)}$ are also independent. The data sets used for training are then of the

[2] Formally, $P_\Pi = g\sharp P$, where the measurable map $g : \mathcal{M}_1(\mathcal{S}) \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y}$ is defined by $g(Q, y) = (\Pi Q, y)$.

form $\mathcal{D} = ((S^{(n)}, y_n))_{n=1,\dots,N} \in (\mathcal{S}^* \times \mathcal{Y})^N$. Furthermore, we define

$$\mathcal{D}_{\hat{\Pi}} = ((\hat{\Pi}(S^{(n)}), y_n))_{n=1,\dots,N}$$

and for $\bar{\mathcal{D}} = ((Q_n, y_n))_{n=1,\dots,N} \in (\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y})^N$, define

$$\bar{\mathcal{D}} = ((Q_n, y_n))_{n=1,\dots,N} \in (\mathcal{M}_1(\mathcal{S}) \times \mathcal{Y})^N$$
$$\bar{\mathcal{D}}_{\Pi} = ((\Pi Q_n, y_n))_{n=1,\dots,N} \in (\mathcal{X} \times \mathcal{Y})^N.$$

To summarize, we have to deal with two sampling stages. First, sampling input-output pairs $(Q, y) \sim P$, and then sampling from the distributions Q. Let now $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function and k a kernel on $\mathcal{X}$. Given a data set $\mathcal{D}$ as above, consider the regularized empirical risk minimization problem

$$\min_{f \in H_k} \mathcal{R}_{\ell, \mathcal{D}_{\hat{\Pi}}, \lambda}(f) \tag{A.5}$$

where $\lambda \in \mathbb{R}_{>0}$ is the regularization parameter. If a solution $f_{\mathcal{D}_{\hat{\Pi}}, \lambda}$ to (A.5) exists, it can be used for a prediction task with distributional inputs by composing it with the map Π, so given input $Q \in \mathcal{M}_1(\mathcal{S})$, it leads to prediction $f_{\mathcal{D}_{\hat{\Pi}}, \lambda}(\Pi Q)$. Using Assumption A.2.1 and Lemma A.2.2, we can now consider various risks[3] like $\mathcal{R}_{\ell, P, \lambda}(f_{\mathcal{D}_{\hat{\Pi}}, \lambda})$.

Remark A.2.3. Note that $\mathcal{X}$ is a subset of a Hilbert space $\mathcal{H}$, so in order to implement the approach just described, we need kernels k on (subsets of) Hilbert spaces. On the one hand, *any* such kernel can in principle be used for this task, cf. [55] for some examples. On the other hand, constructing kernels on (potentially infinite-dimensional) Hilbert spaces can be challenging. To tackle this, [137] suggested a general framework based on *distance substitution kernels* [88]. The Hilbertian embedding $(\mathcal{H}, \Pi)$ is used to construct a kernel on probability distributions by defining $k(P, Q) = \phi(\|\Pi P - \Pi Q\|_{\mathcal{H}})$, where ϕ is a function that induces a radial kernel. Note that all of our general results immediately apply to this framework, covering for example sliced 1- and 2-Wasserstein distances and the induced distance substitution kernels. For details and concrete examples, we refer to [137].

[3] Note that we tacitly assume that the learning methods induced by the regularized (empirical) risk minimization problems are *measurable* learning methods. In the setting we consider, this does not pose a problem, cf. the thorough discussion in Chapter 6 in [189].

Special case: Kernel mean embeddings The first works on distributional learning using Hilbertian embeddings relied on kernel mean embeddings (KMEs), cf. Section 8.5. For convenience, we summarize the necessary background in the following result.

Proposition A.2.4. Let $(\mathcal{S}, \mathcal{A}_\mathcal{S})$ be a measurable space, and κ a measurable and bounded kernel on $\mathcal{S}$.

1. The map
$$\Pi_k : \mathcal{M}_1(\mathcal{S}) \to H_\kappa,\ \Pi_k Q = \int \kappa(\cdot, s)\mathrm{d}Q(s) \tag{A.6}$$
is well-defined, and we call $\Pi_k Q$ the *kernel mean embedding (KME)* of $Q \in \mathcal{M}_1(\mathcal{S})$ w.r.t. κ.

2. Define $\hat{\Pi}_k : \mathcal{S}^* \to H_\kappa$ by
$$\hat{\Pi}_k((s_1, \ldots, s_M)) = \frac{1}{M}\sum_{m=1}^{M} \kappa(\cdot, s_m). \tag{A.7}$$
For all $Q \in \mathcal{M}_1(\mathcal{S})$ and $S \sim Q^{\otimes M}$, $M \in \mathbb{N}_+$, and $\delta \in (0,1)$, we have that
$$\|\hat{\Pi}_k S - \Pi_k Q\|_\kappa \leq 2\sqrt{\frac{\|\kappa\|_\infty^2}{M}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(1/\delta)}{M}} \tag{A.8}$$
holds with probability at least $1 - \delta$.

3. Let $(\mathcal{S}, \tau_\mathcal{S})$ be a separable topological space, choose $\mathcal{A}_\mathcal{S} = \mathcal{B}(\tau_\mathcal{S})$, and assume that κ is continuous. Then Π_k is $(\mathcal{M}_1(\mathcal{S}), \mathcal{B}(\tau_w))$-$(H_\kappa, \mathcal{B}(H_\kappa))$-measurable, where τ_w is the topology induced by weak convergence in $\mathcal{M}_1(\mathcal{S})$.

A proof can be found in Appendix A.5. This result allows to use KMEs as the Hilbertian embedding, i.e., setting $\mathcal{H} = H_\kappa$, $\Pi = \Pi_k$ and $\hat{\Pi} = \hat{\Pi}_k$.

A.3. Oracle Inequalities

Oracle inequalities are important tools in modern statistical learning theory [189]. Roughly speaking, they are concentration inequalities for the excess risk of the learning outcome over the risk that is achieved by an oracle which has access to the

true underlying distribution. In particular, oracle inequalities provide finite-sample guarantees, and can be used to derive consistency of a learning method, as well as (under additional assumptions on the data-generating distribution) learning rates. We now present our two oracle inequalities for the risk of SVMs in the distributional learning setting. The first one is based on a form of Lipschitz-continuity of SVMs, and can be interpreted as a distributional analogon of Theorem 6.24 in [189].

Theorem A.3.1. Let Assumption A.2.1 hold. Assume that ℓ is convex, differentiable, and that there exists $B_\ell \in \mathbb{R}_{\geq 0}$ such that $\ell(x, y, 0) \leq B_\ell$. Furthermore, assume that there exists $B'_\ell \in \mathbb{R}_{\geq 0}$ such that $|\ell'(x, y, 0)| \leq B'_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, and that there exist $\gamma_1 \in \mathcal{K}$ and a nondecreasing family $(\gamma_{3,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ such that $|\ell'(x_1, y, t_1) - \ell'(x_2, y, t_2)| \leq \gamma_1(\|x_1 - x_2\|) + \gamma_{3,B}(|t_1 - t_2|)$ for all $B \in \mathbb{R}_{>0}$, $x_1, x_2 \in \mathcal{X}, y \in \mathcal{Y}$ and $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$. Let k be a kernel on $\mathcal{H}$ that is measurable, bounded, has a separable RKHS H_k, and assume that there exists[4] $\alpha_k \in \mathcal{K}$ such that $\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \alpha_k(\|x_1 - x_2\|_{\mathcal{H}})$ for all $x_1, x_2 \in \mathcal{X}$. Finally, assume that for all $n = 1, \ldots, N$, there exists $B_n : (0, 1) \to \mathbb{R}_{\geq 0}$ such that $\mathbb{P}[\|\hat{\Pi}(S^{(n)}) - \Pi(Q_n)\|_{\mathcal{H}} > B_n(\delta)] < \delta$ for all $\delta \in (0, 1)$. We then have for all $\lambda \in \mathbb{R}_{>0}$ and $\delta \in (0, 1)$ that with probability at least $1 - \delta$

$$\begin{aligned}\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P} \leq A^{(2)}_{\ell,P}(\lambda) + \frac{2\sqrt{B_\ell/\lambda} + |\ell|_{1,B_f}\|k\|_\infty/\lambda}{N} \sum_{n=1}^{N} \alpha_\lambda(B_n(\delta/(2N))) \\ + 2\frac{|\ell|_{1,B_f}\|k\|^2_\infty}{\lambda}\left(B'_\ell + \gamma_{3,B_f}\left(\|k\|_\infty\sqrt{B_\ell/\lambda}\right)\right) \\ \times \left(\sqrt{\frac{2\ln(2/\delta)+1}{N}} + \frac{4\ln(2/\delta)}{3N}\right),\end{aligned}$$

where we defined $B_f = \|k\|_\infty\sqrt{B_\ell/\lambda}$ and $\alpha_\lambda = \|k\|_\infty(\gamma_1 + \gamma_{3,B_f} \circ \left(\sqrt{B_\ell/\lambda}\alpha_k\right)) + \left(B'_\ell + \gamma_{3,B_f}(B_f)\right)\alpha_k$.

The functions B_n in the statement of the result are used to provide estimation bounds of the Hilbertian embeddings of the input distributions, i.e., how close $\hat{\Pi}S^{(n)}$ (which can be computed from data) is to ΠQ_n (which cannot be computed from

[4] The latter condition implies that Φ_k is continuous, which implies that k is continuous, which in turn implies that k is measurable and has a separable RKHS. However, we kept these two conditions for emphasis.

data). In particular, the functions B_n depend on M_n, but we suppressed this dependency to ease the notation. Similarly, $\alpha_\lambda \in \mathcal{K}$ describes (up to a multiplicative factor) how the estimation error of the Hilbertian embedding that arises from a single data set item, influences the risk. Using this abstraction allows us to formulate our results for *any* Hilbertian embedding approach. Specializing to a concrete embedding approach then boils down to checking the well-posedness assumptions (cf. Assumption A.2.1), and replacing the B_n by concrete estimation bounds. While this approach makes Theorem A.3.1 (and similarly Theorem A.3.4 presented below) broadly applicable to various Hilbertian embedding methods, as a result the bounds do not directly help in choosing an appropriate embedding.

Proof sketch for Theorem A.3.1. The basic idea is to apply the proof strategy of Theorem 6.24 in [189] to the ideal, but inaccessible data set $\bar{\mathcal{D}}_\Pi$, and then use estimation error bounds for the Hilbertian embeddings (encoded by the functions B_n) to translate this to the accessible data set $\mathcal{D}_{\hat{\Pi}}$. To do so, we use a known generalized Representer Theorem (recalled as Proposition A.6.4 later on) together with the continuity property of the canonical feature map and the regularity and boundedness properties of the loss function, which allows us to propagate the estimation error through the bounds. A detailed proof is provided in Section A.6.2. □

Example A.3.2. Let us provide some concrete examples for the ingredients of the preceding result. For instance, consider loss functions of the form $\ell(x, y, t) = \psi(y-t)$ (which are called *distance-based supervised losses* in [189]), and assume that ψ is continuously differentiable and that $\mathcal{Y} \subseteq [-M, M]$ for some $M \in \mathbb{R}_{>0}$. In this case, suitable constants B_ℓ and B'_ℓ exist, and we can choose an arbitrary $\gamma_1 \in \mathcal{K}$ (since ℓ does not depend on the first argument) and $\gamma_{3,B}(s) = C_B s$ for suitable constants $C_B \in \mathbb{R}_{>0}$. An example of the condition on Φ_k is given by Hölder-continuity of the canonical feature map Φ_k, which has been used in previous works like [195]. This means that there exist $C_k \in \mathbb{R}_{>0}$, $\alpha \in (0, 1]$, such that $\|\Phi_k(x_1) - \Phi_k(x_2)\|_\mathcal{H} \leq C_k \|x_1 - x_2\|^\alpha$ for all $x_1, x_2 \in \mathcal{X}$, and we can set $\alpha_k(s) = C_k s^\alpha$.

When using KMEs for the Hilbertian embedding, we get the following oracle inequality as a special case.

Corollary A.3.3. Let $\mathcal{S}$ be a compact metric space, κ be a measurable, bounded, continuous and universal kernel on $\mathcal{S}$, and set $\mathcal{H} = H_\kappa$, $\Pi = \Pi_k$, and $\hat{\Pi} = \hat{\Pi}_k$.

Assume that ℓ is convex, differentiable, ℓ' is locally Lipschitz continuous, and that there exists $B_\ell, B'_\ell \in \mathbb{R}_{\geq 0}$ such that $\ell(x,y,0) \leq B_\ell$ and $|\ell'(x,y,0)| \leq B'_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let k be a universal kernel on $\mathcal{H}$ that is measurable and bounded, and that there exists $\alpha_k \in \mathcal{K}$ such that $\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \alpha_k(\|x_1 - x_2\|_{\mathcal{H}})$ for all $x_1, x_2 \in \mathcal{X}$. We then have for all $\lambda \in \mathbb{R}_{>0}$ and $\delta \in (0,1)$ that with probability at least $1-\delta$

$$\begin{aligned}
&\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^*_{\ell,P} \\
&\quad \leq A^{(2)}_{\ell,P}(\lambda) + \frac{2\sqrt{\lambda B_\ell} + |\ell|_{1,B_f}\|k\|_\infty}{N} \times \sum_{n=1}^{N} \alpha_\lambda \left(2\sqrt{\frac{\|\kappa\|^2_\infty}{M_n}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(2N/\delta)}{M_n}} \right) \\
&\qquad + 2|\ell|_{1,B_f}\|k\|_\infty \left(B'_\ell + \gamma_{3,B_f}(B_f) \right) \left(\sqrt{\frac{2\ln(2N/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2N/\delta)}{3N} \right),
\end{aligned}$$

with B_f and α_λ as in Theorem A.3.1.

Theorem A.3.1 puts strong regularity requirements on the loss function, but needs only mild assumptions for the kernel used in the empirical risk minimization. The following oracle inequality, a distributional analogon of Theorem 6.25 from [189], is complementary, putting only mild requirements on the loss function, but strong structural results on the RKHS are used.

Theorem A.3.4. Assume that ℓ is convex, that there exist $\gamma_1 \in \mathcal{K}$ and a nondecreasing family $(\gamma_{3,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ such that for all $x_1, x_2 \in \mathcal{X}, y \in \mathcal{Y}$, and all $B \in \mathbb{R}_{>0}$ and $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$, it holds that $|\ell(x_1,y,t_1) - \ell(x_2,y,t_2)| \leq \gamma_1(\|x_1 - x_2\|_{\mathcal{H}}) + \gamma_{3,B}(|t_1 - t_2|)$, and that there exists $B_\ell \in \mathbb{R}_{\geq 0}$ such that $\ell(x,y,0) \leq B_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let k be a kernel on $\mathcal{X}$ that is measurable, bounded, and has a separable RKHS H_k. Assume that there exists an nondecreasing family $(\alpha_{f,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ such that for $B \in \mathbb{R}_{>0}$ and $f \in H_k$ with $\|f\|_k \leq B$, we have $|f(x_1) - f(x_2)| \leq \alpha_{f,B}(\|x_1 - x_2\|_{\mathcal{H}})$ for all $x_1, x_2 \in \mathcal{X}$. Furthermore, let $\epsilon, \lambda \in \mathbb{R}_{>0}$, and let $\mathcal{F} \subseteq H_k$ be a finite set such that for all $f \in H_k$ with $\|f\|_k \leq \sqrt{B_\ell/\lambda}$ there exists $\tilde{f} \in \mathcal{F}$ with $\|f - \tilde{f}\|_\infty \leq \epsilon$. Finally, assume that for all $n = 1, \ldots, N$, there exists $B_n : (0,1) \to \mathbb{R}_{\geq 0}$ such that $\mathbb{P}[\|\hat{\Pi}(S^{(n)}) - \Pi(Q_n)\|_{\mathcal{H}} > B_n(\delta)] < \delta$ for all $\delta \in (0,1)$. We then have for all $\delta \in (0,1)$ that with probability at least $1-\delta$ it holds

that

$$\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D},\lambda}) - \mathcal{R}_{\ell,P}^{H_k*} \leq A_{\ell,P}^{(2)}(\lambda) + 4\gamma_{3,\tilde{B}_f}(\epsilon) + \frac{2}{N}\sum_{n=1}^{N} \alpha_\lambda \left(B_n \left(\frac{\delta}{N+|\mathcal{F}|} \right) \right) + 2\left(B_\ell + \gamma_{3,\tilde{B}_f}(\tilde{B}_f) \right) \sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}},$$

where we defined $\tilde{B}_f = \|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}} + \epsilon$ and $\alpha_\lambda = \gamma_1 + \gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}$.

The central assumption of Theorem A.3.4 is the existence of a suitable discretization $\mathcal{F}$ of $\bar{B}^{H_k}_{\sqrt{B_\ell/\lambda}}$, the closed centered ball with radius $\sqrt{B_\ell/\lambda}$ in the RKHS H_k. Under suitable assumptions, a *finite* $\mathcal{F}$ exists, and one can set $|\mathcal{F}| = \mathcal{N}(\bar{B}^{H_k}_{\sqrt{B_\ell/\lambda}}, \|\cdot\|_\infty, \epsilon)$, where $\mathcal{N}(T,d,\epsilon)$ is the *ϵ-covering number* of a metric space (T,d). For more details and pointers to the literature, we refer to Chapters 6, 7 in [189].

Proof sketch for Theorem A.3.4. Similarly to the proof of Theorem A.3.1, we apply the proof strategy of Theorem 6.25 in [189] to the ideal, but inaccessible data set $\bar{\mathcal{D}}_\Pi$, and translate the result to the accessible data set $\mathcal{D}_{\hat{\Pi}}$ by the estimation bounds described by the functions B_n, using the continuity and boundedness properties of the loss function (which can be milder now, since we do not use Proposition A.6.4 anymore) and the canonical feature map. A detailed proof is provided in Section A.6.2 later on. □

Example A.3.5. A sufficient condition for the existence of the nondecreasing family $(\gamma_{3,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ is Hölder-continuity. If $\mathrm{d}_{\mathcal{H}}(\mu,\nu) = \|(\Phi_k \circ \Pi)(\mu) - (\Phi_k \circ \Pi)(\nu)\|_k$, then it is well-known that one can choose $\alpha_{f,B}(s) = Bs$. If there exists $C_k, \alpha_k \in \mathbb{R}_{>0}$ such that $|k(x_1,x) - k(x_2,x)| \leq C_k \|x_1 - x_2\|_{\mathcal{H}}^{\alpha_k}$ for all $x_1, x_2 \in \mathcal{X}$, then one can choose $\alpha_{f,B}(s) = \sqrt{2C_k}s^{\alpha_k/2}$. For proofs of these facts and more general conditions, we refer to [CF1].

We can immediately specialize Theorem A.3.4 to the case of KMEs for the Hilbertian embedding.

Corollary A.3.6. Consider the situation of Theorem A.3.4. Additionally, let $\mathcal{S}$ be a compact metric space, κ be a measurable, bounded, continuous and universal

kernel on $\mathcal{S}$, and set $\mathcal{H} = H_\kappa$, $\Pi = \Pi_k$, and $\hat{\Pi} = \hat{\Pi}_k$. We then have for all $\delta \in (0,1)$, that

$$\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D},\lambda}) - \mathcal{R}_{\ell,P}^{H_k *} \leq A_{\ell,P}^{(2)}(\lambda) + 4\gamma_{3,\tilde{B}_f}(\epsilon) + 2\left(B_\ell + \gamma_{3,\tilde{B}_f}(\tilde{B}_f)\right)\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}} + \frac{2}{N}\sum_{n=1}^{N}\alpha_\lambda\left(2\sqrt{\frac{\|k\|_\infty^2}{M}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(\frac{N+|\mathcal{F}|}{\delta})}{M}}\right),$$

holds with probability at least $1-\delta$, with $\tilde{B}_f$ and α_λ as in Theorem A.3.4.

The proof is completely analogous to the one of Corollary A.3.3.

A.4. Stability-based Generalization Bound

The oracle inequalities from the previous section allow us to compare the risk of the learned hypothesis (i.e., of the empirical SVM solution) to the minimum risk that could be achieved by an oracle (having access to the true underlying meta-distribution). We now consider a slightly different question: How accurate is the empirical risk of the learned hypothesis (which can be computed from data) as an estimate of the true risk of the learned hypothesis (which cannot be computed, since we do not know the true underlying data-generating distribution)? In other words, how well does the learned hypothesis generalize from the training data to the population, as measured by its risk?

We investigate this using a variation of our basic setup. Let $(Q,y) \sim P$ as before, but now assume that the number of samples from Q (collected in S) is also random. Denote the joint distribution of (Q,S,y) by $\bar{P}$, the marginal distribution of (S,y) by $\tilde{P}$, and the number of samples in S by M, an $\mathbb{N}_+$-valued random variables. A special case covered by this setup is a constant M, a setting which is often considered in related works like [195] or [137]. The data set $\mathcal{D}$ is therefore now generated by sampling $(Q_1, S^{(1)}, y), \ldots, (Q_N, S^{(N)}, y_N)$ i.i.d. from $\bar{P}$, and then setting $\mathcal{D} = ((S^{(n)}, y_n))_{n=1,\ldots,N}$, hence $\mathcal{D} \sim \tilde{P}^{\otimes N}$.

The generalization bounds that follow are based on the concept of algorithmic stability [38], which applies to very general learning methods. A learning method is a map[5] $\bigcup_{N\in\mathbb{N}_+}(\mathcal{X}\times\mathcal{Y})^N \ni D \mapsto \mathcal{L}_D$, where $\mathcal{L}_D : \mathcal{X} \to \mathbb{R}$ is measurable. We call $\mathcal{L}$

[5] In the present setting, it is safe to ignore measurability issues, cf. the discussion in Chapter 6 in

β-stable (w.r.t. the loss function ℓ) if there exists $(\beta_N)_{N\in\mathbb{N}_+}$, $\beta_N \in \mathbb{R}_{\geq 0}$, such that for all $N \in \mathbb{N}_+$, $x \in \mathcal{X}, y \in \mathcal{Y}$ we have

$$|\ell(x, y, \mathcal{L}_D(x)) - \ell(x, y, \mathcal{L}_{\tilde{D}}(x))| \leq \beta_N, \tag{A.9}$$

for all $D, \tilde{D} \in (\mathcal{X} \times \mathcal{Y})^N$ such that there exists $1 \leq i \leq N$ with $D_n = \tilde{D}_n$, $n \in \{1, \ldots, N\} \setminus \{i\}$. In other words, a learning method is β-stable, if changing just one sample in a data set of size $N \in \mathbb{N}_+$, changes the loss of the resulting hypothesis by at most β_N. We are now ready to present the announced generalization result. It is a distributional-input analogon of Theorem 14.2 in [141].

Theorem A.4.1. Consider a β-stable learning method $\mathcal{L}$. Assume that there exists a concave $\alpha \in \mathcal{K}$ such that for all $x_1, x_2 \in \mathcal{X}, y \in \mathcal{Y}$ and all $D \in (\mathcal{S}^* \times \mathcal{Y})^N$ we have $|\ell(x_1, y, \mathcal{L}_D(x_1)) - \ell(x_2, y, \mathcal{L}_D(x_2))| \leq \alpha(\|x_1 - x_2\|_{\mathcal{H}})$. For all $\delta \in (0, 1)$, it holdst with probability at least $1 - \delta$ that

$$\mathcal{R}_{\ell,P}(\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}) \leq \mathcal{R}_{\ell,\mathcal{D}_\Pi}(\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}) + (2N\beta_N + B)\sqrt{\frac{\ln(1/\delta)}{2N}} + \alpha\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi} S\|_{\mathcal{H}}\right]\right) + \beta_N.$$

The proof of this result can be found in Appendix A.7.2. We now present and prove a generalization bound for SVMs in the two-stage sampling setup, which is based on Theorem A.4.1.

Theorem A.4.2. Let ℓ be convex, locally Lipschitz continuous, and assume that there exists $\gamma_1 \in \mathcal{K}$ such that $|\ell(x_1, y, t) - \ell(x_2, y, t)| \leq \gamma_1(\|x_1 - x_2\|_{\mathcal{H}})$ for all $x_1, x_2 \in \mathcal{X}, y \in \mathcal{Y}$ and $t \in \mathbb{R}$. Furthermore, assume that there exists $B_\ell \in \mathbb{R}_{>0}$ such that $\ell(x, y, 0) \leq B_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let k be measurable and bounded, and assume that there exists a nondecreasing family $(\alpha_{f,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ such that for all $x_1, x_2 \in \mathcal{X}$, $B \in \mathbb{R}_{>0}$, and all $f \in H_k$ with $\|f\|_k \leq B$ we have $|f(x_1) - f(x_2)| \leq \alpha_{f,B}(\|x_1 - x_2\|_{\mathcal{H}})$. Assume that for $\lambda \in \mathbb{R}_{>0}$, there exists a concave $\alpha_\lambda \in \mathcal{K}$ with $\gamma_1 + |\ell|_{1,B_f}\alpha_{f,\sqrt{B_\ell/\lambda}} \leq \alpha_\lambda$, where we defined $B_f = \|k\|_\infty\sqrt{B_\ell/\lambda}$. We then have for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, that

$$\begin{aligned}\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) \leq \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \frac{|\ell|^2_{1,B_f}\|k\|^2_\infty}{\lambda N} + \alpha\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi} S\|_{\mathcal{H}}\right]\right) \\ + \left(\frac{2|\ell|^2_{1,B_f}\|k\|^2_\infty}{\lambda} + B_\ell + |\ell|_{1,B_f}B_f\right)\sqrt{\frac{\ln(1/\delta)}{2N}}.\end{aligned}$$

[189].

Before turning to the proof of Theorem A.4.2, we describe two example classes of suitable ℓ and α_λ.

Example A.4.3. Assume that there exist $(C_{f,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathbb{R}_{>0}$, $(\alpha_{f,B})_{B\in\mathbb{R}_{>0}} \subseteq (0,1]$ such that $|f(x_1) - f(x_2)| \leq C_{f,B}\|x_1 - x_2\|_\Pi^{\alpha_{f,B}}$ for all $B \in \mathbb{R}_{>0}$, $x_1, x_2 \in \mathcal{X}$, and $f \in H_k$ with $\|f\|_k \leq B$. Let us call a function $\phi : \mathbb{R} \to \mathbb{R}$ locally Hölder-continuous, if there exist $(C_{\phi,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathbb{R}_{>0}$, $(\alpha_{\phi,B})_{B\in\mathbb{R}_{>0}} \subseteq (0,1]$, such that for all $B \in \mathbb{R}_{>0}$, $|\phi(s_1) - \phi(s_2)| \leq C_{\phi,B}|s_1 - s_2|^{\alpha_{\phi,B}}$ for all $s_1, s_2 \in [-B, B]$. We refer to [CF1] for a discussion of these properties, including characterizations of suitable k. (i) Assume that $\ell(x, y, t) = \psi(y - t)$, where ψ is a nonnegative, locally Hölder-continuous function. Given $\lambda \in \mathbb{R}_{>0}$, we can then choose $\alpha_\lambda(s) = C_{\psi,\|k\|_\infty\sqrt{B_\ell/\lambda}} C_{f,\sqrt{B_\ell/\lambda}}^{\alpha_\psi} s^{\alpha_\psi\alpha_f}$ with $\alpha_\psi = \alpha_{\psi,\|k\|_\infty\sqrt{B_\ell/\lambda}}$ and $\alpha_f = \alpha_{f,\sqrt{B_\ell/\lambda}}$. (ii) Assume that $\ell(x, y, t) = \varphi(yt)$ (called a *margin-based loss function* in [189]) for a nonnegative, locally Hölder-continuous function, and that $\mathcal{Y} \subseteq [-M, M]$ for some $M \in \mathbb{R}_{>0}$. Given $\lambda \in \mathbb{R}_{>0}$, we can then choose $\alpha_\lambda(s) = C_\varphi M^{\alpha_\varphi} C_{f,\sqrt{\varphi(0)/\lambda}}^{\alpha_\varphi} s^{\alpha_\varphi\alpha_f}$ with $C_\varphi = C_{\varphi,M\|k\|_\infty\sqrt{\varphi(0)/\lambda}}$, $\alpha_\varphi = \alpha_{\varphi,M\|k\|_\infty\sqrt{\varphi(0)/\lambda}}$, and $\alpha_f = \alpha_{f,\sqrt{\varphi(0)/\lambda}}$.

Proof of Theorem A.4.2. Let Q be a distribution on $\mathcal{M}_1(\mathcal{S})\times\mathcal{Y}$. From Lemma A.5.4 we have $|f_{Q,\lambda}(x)| \leq \|k\|_\infty\|f_{Q,\lambda}\|_k \leq \|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}} = B_f$, so

$$\begin{aligned}
|\ell(x_1, y, f_{Q,\lambda}(x_1)) - \ell(x_2, y, f_{Q,\lambda}(x_2))| &\leq \gamma_1(\|x_1 - x_2\|_\mathcal{H}) + |\ell|_{1,B_f}|f_{Q,\lambda}(x_1) - f_{Q,\lambda}(x_2)| \\
&\leq \left(\gamma_1 + |\ell|_{1,B_f}\alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right)(\|x_1 - x_2\|_\mathcal{H}) \\
&\leq \alpha_\lambda(\|x_1 - x_2\|_\mathcal{H}),
\end{aligned}$$

hence α_λ fulfills the requirements of Theorem A.4.1. Furthermore, as before we have $\ell(x, y, f_{Q,\lambda}(x)) \leq B_\ell + |\ell|_{1,B_f}B_f = B$.

Next[6], for all $f, g \in H_k$ with $\|f\|_k, \|g\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$ and all $x \in \mathcal{X}, y \in \mathcal{Y}$, we have $|\ell(x, y, f(x)) - \ell(x, y, g(x))| \leq |\ell|_{1,B_f}|f(x) - g(x)|$. An inspection of the proof of Proposition 14.4 in [141] then shows that the learning method $(\mathcal{X} \times \mathcal{Y})^N \ni D \mapsto f_{\ell,D}\lambda \in H_k$ is $\beta_N = \frac{|\ell|_{1,B_f}^2\|k\|_\infty^2}{\lambda N}$ stable.

The result now follows from Theorem A.4.1. □

[6]The following is a generalization of the property from Definition 14.3 in [141].

Remark A.4.4. An inspection of the proof of Theorem A.4.2 and how Proposition 14.4 in [141] is used there, reveals that instead of local Lipschitz continuity of ℓ the following continuity property is sufficient: There exists a family $(C_B, p_B)_{B\in\mathbb{R}_{>0}}$ with $C_B \in \mathbb{R}_{>0}$, $0 < p_B < 2$ for all $B \in \mathbb{R}_{>0}$, such that for all $x \in \mathcal{X}, y \in \mathcal{Y}$, $B \in \mathbb{R}_{>0}$ and all $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$ we have that $|\ell(x,y,t_1) - \ell(x,y,t_2)| \leq C_B|t_1 - t_2|^{p_B}$. Furthermore, we now need a concave $\alpha_\lambda \in \mathcal{K}$ such that $\gamma_1 + C_B \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}(\cdot)^{p_B} \leq \alpha_\lambda$ with $B = \|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}}$. In this case, we have $\beta_N = C_B^{1+\frac{1}{2-p_B}} \|k\|_\infty^{p_B + \frac{p_B}{2-p_B}} \left(\frac{1}{N\lambda}\right)^{\frac{1}{2-p_B}}$.

Once again, we can immediately specialize to the case of using KMEs for the Hilbertian embedding.

Corollary A.4.5. Consider the situation of Theorem A.4.2. Additionally, let $\mathcal{S}$ be a compact metric space, κ be a measurable, bounded, continuous and universal kernel on $\mathcal{S}$, and set $\mathcal{H} = H_\kappa$, $\Pi = \Pi_k$, and $\hat{\Pi} = \hat{\Pi}_k$. We then have for all $\delta \in (0,1)$, with probability at least $1 - \delta$, that

$$\begin{aligned}\mathcal{R}_{\ell,P}(f_{\ell,\mathcal{D}_{\hat{\Pi}}}\lambda) \leq \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\ell,\mathcal{D}_{\hat{\Pi}}}\lambda) + \alpha_\lambda\left(\frac{\sqrt{2\|\kappa\|_\infty}}{\mathbb{E}[\sqrt{M}]}\right) \\ + \left(\frac{2|\ell|_{1,B_f}^2\|k\|_\infty^2}{\lambda} + B_\ell + |\ell|_{1,B_f}B_f\right)\sqrt{\frac{\ln(1/\delta)}{2N}} + \frac{|\ell|_{1,B_f}^2\|k\|_\infty^2}{\lambda N},\end{aligned}$$

where we defined $B_f = \|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}$.

Proof. First, as in the proof of Corollary A.3.3, the KME setup fulfills Assumption A.2.1. Let $Q \in \mathcal{M}_1(\mathcal{S})$, $M \in \mathbb{N}_+$, and $S \sim Q^{\otimes M}$. According to Lemma 4 in [85], $\|\Pi_k Q - \hat{\Pi}_k S\|_\kappa$ is the maximum mean discrepancy between Q and the empirical measure $\frac{1}{M}\sum_{m=1}^M \delta_{S_m}$, so we get from Equation (19) in the same reference that

$$\mathbb{E}_{S\sim Q^{\otimes M}}\left[\|\Pi_k Q - \hat{\Pi}_k S\|_\kappa\right] \leq \sqrt{\frac{2\|\kappa\|_\infty}{M}},$$

which implies that

$$\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi} S\|_{\mathcal{H}}\right] \leq \mathbb{E}\left[\sqrt{\frac{2\|\kappa\|_\infty}{M}}\right] = \frac{\sqrt{2\|\kappa\|_\infty}}{\mathbb{E}[\sqrt{M}]}.$$

Combining this with Theorem A.4.2 and using that α_λ is increasing, establishes the result. □

A.5. Additional Technical Background

Comparison functions In addition to $\mathcal{K}$, we define

$$\mathcal{L} = \{\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0} \mid \rho \text{ continuous, strictly decreasing, } \lim_{s\to\infty} \rho(s) = 0\}.$$

Observe that if $\rho \in \mathcal{L}$, then $\rho(s) > 0$ for all $s \in \mathbb{R}_{\geq 0}$, and its inverse is defined on its range, i.e., $\rho^{-1} : (0, \rho(0)] \to \mathbb{R}_{\geq 0}$. We define addition and scalar multiplication in $\mathcal{K}$ and $\mathcal{L}$ pointwise, i.e., if $\alpha_1, \alpha_2 \in \mathcal{K}$ (respectively, $\mathcal{L}$), then $\alpha_1 + \alpha_2$ is defined by $(\alpha_1+\alpha_2)(s) = \alpha_1(s)+\alpha_2(s)$ for all $s \in \mathbb{R}_{\geq 0}$, and if $c_1 \in \mathbb{R}_{>0}$, then $c_1\alpha_1$ is defined by $(c_1\alpha_1)(s) = c_1\alpha_1(s)$. Note that $c_1\alpha_1 + \alpha_2 \in \mathcal{K}$ (respectively, in $\mathcal{L}$), so $\mathcal{K}$ and $\mathcal{L}$ form a cone. Furthermore, $\alpha_1 \circ \alpha_2 \in \mathcal{K}$. We also define comparison relations pointwise, e.g., if $\alpha_1, \alpha_1 \in \mathcal{K}$, then $\alpha_1 \leq \alpha_2$ means that $\alpha_1(s) \leq \alpha_2(s)$ for all $s \in \mathbb{R}_{\geq 0}$. For more background on comparison functions, including historical remarks and application examples, we refer to [105].

More on loss functions For technical reasons, we need some additional concepts from [189]. A loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is called a *Nemitskii loss*, if there exists a measurable function $b : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and an increasing function $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and all $t \in \mathbb{R}$ we have

$$\ell(x, y, t) \leq b(x, y) + h(|t|).$$

Let P be a probability distribution on $\mathcal{X} \times \mathcal{Y}$. We call ℓ a *P-integrable Nemitskii loss*, if it is a Nemitskii loss, and the function b from the definition of this concept is P-integrable.

Boundedness in RKHSs For convenience, we summarize some well-known results on boundedness of kernels and RKHS functions.

Lemma A.5.1 (Boundedness in RKHSs)**.** Let $\mathcal{X}$ be an arbitrary nonempty set and k a kernel on it, and define

$$\|k\|_\infty = \sup_{x\in\mathcal{X}} \sqrt{k(x, x)}. \tag{A.10}$$

1. k is bounded if and only if $\|k\|_\infty < \infty$.
2. All $f \in H_k$ are bounded if and only if k is bounded.
3. For all $f \in H_k$ and $x \in \mathcal{X}$, $|f(x)| \leq \|f\|_k \|k\|_\infty$.

Proof. For the first item, assume that k is bounded, then obviously $\|k\|_\infty < \infty$. Conversely, if $\|k\|_\infty < \infty$, then we have for all $x, x' \in \mathcal{X}$

$$|k(x,x')| = |\langle k(\cdot,x'), k(\cdot,x)\rangle_k| \leq \|k(\cdot,x')\|_k \|k(\cdot,x)\|_k = \sqrt{k(x',x')}\sqrt{k(x,x)} \leq \|k\|_\infty^2 < \infty$$

so k is indeed bounded.

The second statement is given by Lemma 4.23 in [189].

For the last statement, let $f \in H_k$ and $x \in \mathcal{X}$ be arbitrary, then

$$|f(x)| = |\langle f, k(\cdot,x)\rangle_k| \leq \|f\| \|k(\cdot,x)\|_k = \|f\|_k \sqrt{k(x,x} \leq \|f\|_k \|k\|_\infty.$$

□

Properties of loss functions and their risks Next, we present two technical results on loss functions and their associated risks. These results are essentially known (cf. Chapter 2 in [189]), however, we formulate them in greater generality using comparison functions.

Lemma A.5.2 (Condition for P-integrable Nemitskii loss)**.** Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function such that there exists $B_\ell \in \mathbb{R}_{\geq 0}$ with $\ell(x,y,0) \leq B_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$, and a nondecreasing family $(\alpha_{\ell,B})_{B \in \mathbb{R}_{>0}} \subseteq \mathcal{K}$ with $|\ell(x,y,t_1) - \ell(x,y,t_2)| \leq \alpha_{\ell,B}(|t_1 - t_2|)$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$, then ℓ is a P-integrable Nemitskii loss for all distributions P on $\mathcal{X} \times \mathcal{Y}$.

In particular, this result applies to locally Lipschitz continuous functions, where $\alpha_{\ell,B}(t) = |\ell|_{1,|t|}|t|$.

Proof. Let $x \in \mathcal{X}, y \in \mathcal{Y}, t \in \mathbb{R}$ be arbitrary, then we have

$$\begin{aligned}\ell(x,y,t) &\leq \ell(x,y,0) + |\ell(x,y,t) - \ell(x,y,0)| \\ &\leq B_\ell + \alpha_{\ell,|t|}(|t|)\end{aligned}$$

Since $\int B_\ell \mathrm{d}P = B_\ell < \infty$ and $t \mapsto \alpha_{\ell,|t|}(|t|)$ is nondecreasing, the statement follows. □

Lemma A.5.3 (Continuity of risk from continuity of loss function)**.** Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a loss function such that there exists a nondecreasing family $(\alpha_{\ell,B})_{B\in\mathbb{R}_{>0}} \subseteq \mathcal{K}$ with $|\ell(x,y,t_1) - \ell(x,y,t_2)| \leq \alpha_{\ell,B}(|t_1 - t_2|)$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$ and $t_1, t_2 \in \mathbb{R}$ with $|t_1|, |t_2| \leq B$. Let P be a distribution such that ℓ is a P-integrable Nemitskii loss.

1. For all $B \in \mathbb{R}_{>0}$ and all measurable and bounded[7] f, g with $\|f\|_\infty, \|g\|_\infty \leq B$, we have
$$|\mathcal{R}_{\ell,P}(f) - \mathcal{R}_{\ell,P}(g)| \leq \alpha_{\ell,B}(\|f - g\|_\infty). \tag{A.11}$$

2. Let k be a measurable and bounded kernel on $\mathcal{X}$. For all $B \in \mathbb{R}_{>0}$ and $f, g \in H_k$ with $\|f\|_k, \|g\|_k \leq B$, we have
$$|\mathcal{R}_{\ell,P}(f) - \mathcal{R}_{\ell,P}(g)| \leq \alpha_{\ell,\|k\|_\infty \cdot B}(\|f - g\|_k \|k\|_\infty). \tag{A.12}$$

Proof. For the first claim, let $B \in \mathbb{R}_{>0}$ and f, g be measurable functions with $\|f\|_\infty, \|g\|_\infty \leq B$. We then have

$$\begin{aligned}
|\mathcal{R}_{\ell,P}(f) - \mathcal{R}_{\ell,P}(g)| &\leq \int |\ell(x,y,f(x)) - \ell(x,y,g(x))| \mathrm{d}P(x,y) \\
&\leq \int \alpha_{\ell,B}(|f(x) - g(x)|) \mathrm{d}P(x,y) \\
&\leq \int \alpha_{\ell,B}(\|f - g\|_\infty) \mathrm{d}P(x,y) \\
&= \alpha_{\ell,B}(\|f - g\|_\infty),
\end{aligned}$$

where we used the triangle inequality in the first step, the existence of $(\alpha_{\ell,B})_B$ in the second step, the fact that $\alpha_{\ell,B}$ is increasing in the third step, and finally that P is a probability distribution.

For the second claim, let $B \in \mathbb{R}_{>0}$ and $f, g \in H_k$ with $\|f\|_k, \|g\|_k \leq B$. Since k is measurable and bounded, also f, g are measurable and bounded. From Lemma A.5.1 we get $\|f\|_\infty \leq \|f\|_k \|k\|_\infty \leq B\|k\|_\infty$, and similarly for g, as well as $\|f - g\|_\infty \leq \|f - g\|_k \|k\|_\infty$. The result now follows from the first claim. □

[7] Measurably essentially bounded would be enough.

Bound on norm of regularized risks minimizers Finally, we recall a well-known result providing a bound on the norm of minimizers of regularized risks minimization problems, cf. the beginning of Section 5.1 in [189].

Lemma A.5.4 (Regularized risk minimization over RKHSs)**.** Let $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a convex, locally Lipschitz continuous loss function, such that there exists $B_\ell \in \mathbb{R}_{\geq 0}$ with $\ell(x, y, 0) \leq B_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let k be a kernel on $\mathcal{X}$ that is measurable, bounded, and with separable H_k. For all distributions P on $\mathcal{X} \times \mathcal{Y}$ and all $\lambda \in \mathbb{R}_{>0}$, there exists a unique solution $f_{P,\lambda}$ of

$$\min_{f \in H_k} \mathcal{R}_{\ell,P}(f) + \lambda \|f\|_k^2,$$

and $\|f_{P,\lambda}\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$.

Proof. Lemma A.5.2 ensures that ℓ is a P-integrable Nemitskii loss, so Lemma 5.1 and Theorem 5.2 from [189] are applicable and ensure that a unique solution $f_{P,\lambda}$ exists.

Additionally, we have

$$\begin{aligned}
\lambda \|f_{P,\lambda}\|_k^2 &\leq \mathcal{R}_{\ell,P}(f_{P,\lambda}) + \lambda \|f_{P,\lambda}\|_k^2 = \mathcal{R}_{\ell,P,\lambda}(f_{P,\lambda}) \\
&\leq \mathcal{R}_{\ell,P,\lambda}(0) = \mathcal{R}_{\ell,P}(0) + \lambda \|0\|_k^2 \\
&= \int \ell(x, y, 0) \mathrm{d}P(x, y) \\
&\leq B_\ell,
\end{aligned}$$

where we first used the nonnegativity of ℓ (and monotonicity of the integral) in the first step, followed by the definition of $f_{P,\lambda}$, and finally the boundedness assumption of ℓ in zero. Rearranging shows that indeed $\|f_{P,\lambda}\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$. □

Kernel mean embeddings

Proof of Proposition A.2.4. The first statement is contained in Theorem 2 and Proposition 2 in [188], and the discussion following it.

The second statement follows from Theorem 1 in [123], by two minor modifications. First, applying Lemma A.5.1 to $f \in H_\kappa$ with $\|f\|_\kappa \leq 1$ leads to $|f(s)| \leq \|f\|_\kappa \|\kappa\|_\infty \leq \|\kappa\|_\infty$ for all $s \in \mathcal{S}$, which shows that the constant of the bounded

difference property in the proof of Theorem 1 in [123] needs to be set to $2\|\kappa\|_\infty/M$. Second, we use $\int \kappa(s,s)\mathrm{d}Q(s) \leq \int \|\kappa\|_\infty^2 \mathrm{d}Q(s) \leq \|\kappa\|_\infty^2$ for all $Q \in \mathcal{M}_1(\mathcal{S})$.

The third statement is shown in Section A.1.1 in [195]. □

Sliced Wasserstein distances Let $\mathcal{S} = \mathbb{R}^d$ and denote by $\underline{\mathcal{W}}_2(\mu,\nu)$ the sliced 2-Wasserstein distance, cf. equation (13) in [137]. It has been shown in Proposition 5 in the same reference that there exists a Hilbert space $\mathcal{H}_2$ and a map $\Pi_2 : \mathcal{M}_1(\mathcal{S}) \to \mathcal{H}_2$ such that $\underline{\mathcal{W}}_2(\mu,\nu) = \|\Pi_2\mu - \Pi_2\nu\|_{\mathcal{H}_2}$. Setting $\hat{\Pi}_2 S = \Pi_2\hat{\mu}[S]$ for all $S \in \mathcal{S}^M$ and $M \in \mathbb{N}_+$, where $\hat{\mu}[S] = \frac{1}{M}\sum_{m=1}^M \delta_{S_m}$ is the empirical measure having the components of S as atoms, and assuming that Assumption A.2.1 is fulfilled, our main results Theorems A.3.1, A.3.4 and A.4.2 apply to the case of sliced 2-Wasserstein-based Hilbertian embeddings. For more details, as well as the case of sliced 1-Wasserstein-based Hilbertian embeddings, and concrete constructions of suitable kernels k on $\mathcal{H}_2$, we refer to [137].

A.6. Additional Material on the Oracle Inequalities

In this section, we present the proofs of our oracle inequalities Theorem A.3.1 and Theorem A.3.4. Furthermore, we state and prove specializations to the case of sliced 2-Wasserstein embeddings, analogous to the results for KMEs, cf. Corollary A.3.3 and Corollary A.3.6.

A.6.1. Sliced Wasserstein Distances

Our specialization of the oracle inequalities to sliced 2-Wasserstein embeddings are based on the following error bound, which might be of independent interest.

Proposition A.6.1. Let P be a distribution on $\mathcal{M}_1(\mathbb{R}^d) \times \mathcal{Y}$ and $(Q,y) \sim P$. Assume that P-a.s. Q is a log-concave distribution, and denote its (P-a.s. defined) covariance matrix by Σ_Q. Furthermore, assume that there exists $\rho_\Sigma \in \mathcal{L}$ such that for all $t \in \mathbb{R}_{\geq 0}$, $\mathbb{P}[\|\Sigma_Q\| \geq t] \leq \rho_\Sigma(t)$ P-a.s. Let $M \in \mathbb{N}_+$ and $S \sim Q^{\otimes M}$, then for all $0 < \delta < \min\{1/4, 2\rho_\Sigma(1/\tilde{C}_d)\}$, we have

$$\mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq \frac{\rho_\Sigma^{-1}(\delta/2)}{\sqrt{M}}\left(C_d\sqrt{\ln(M)} + \tilde{C}_d \ln(4/\delta)\right)\right] \leq \delta,$$

where C_d and $\tilde{C}_d$ are universal constants that only depend on d.

To simplify the notation in the following proof, we define $a \wedge b = \min\{a, b\}$ for $a, b \in \mathbb{R}$.

Proof. As shown in the proof of Proposition 7 in [147], there exists a universal constant $c_d \in \mathbb{R}_{>0}$, depending only on $d \in \mathbb{N}_+$, such that $\frac{1}{P_\mu} \geq \frac{1}{c_d \|\Sigma_\mu\|}$ for all log-concave distributions μ on $\mathbb{R}^d$, where P_μ is the Poincare constant of μ. Furthermore, according to Theorem 1 (choosing $p = 2$ there) in the same reference, there exists a universal constant $C_d \in \mathbb{R}_{>0}$, depending only on d, such that

$$\mathbb{E}[\underline{\mathcal{W}}_2(\mu, \hat{\mu}[X])] \leq \sqrt{\frac{\|\Sigma_\mu\| \ln(M)}{M}} \tag{A.13}$$

for all log-concave distributions μ on $\mathbb{R}^d$ and $X \sim \mu^{\otimes M}$.

Let $0 < \delta < \min\{1/4, 2\rho_\Sigma(1/\tilde{C}_d)\}$ be arbitrary, and $x, t \in \mathbb{R}_{>0}$ two constants to be chosen later. We start with

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t\right] \\
&= \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t,\ \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad + \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t,\ \|\Sigma_Q\| > x \wedge \sqrt{x}\right] \\
&\leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t \mid \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \mathbb{P}[\|\Sigma_Q\| \leq x \wedge \sqrt{x}] \\
&\qquad + \mathbb{P}[\|\Sigma_Q\| > x \wedge \sqrt{x}] \\
&\leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t \mid \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] + \rho_\Sigma(x \wedge \sqrt{x}),
\end{aligned}$$

where we used in the last step that probabilities are always from $[0, 1]$, and the assumption on $\|\Sigma_Q\|$.

We continue with the first term,

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq (x\wedge\sqrt{x})C_d\sqrt{\frac{\ln(M)}{M}}+t \;\middle|\; \|\Sigma_Q\| \leq x\wedge\sqrt{x}\right] \\
&\quad\leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq C_d\sqrt{\frac{x\ln(M)}{M}}+t \;\middle|\; \|\Sigma_Q\| \leq x\wedge\sqrt{x}\right] \\
&\quad\leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq C_d\sqrt{\frac{\|\Sigma_Q\|\ln(M)}{M}}+t \;\middle|\; \|\Sigma_Q\| \leq x\wedge\sqrt{x}\right] \\
&\quad\leq \mathbb{P}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) \geq \mathbb{E}[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S])]+t \mid \|\Sigma_Q\| \leq x\wedge\sqrt{x}\right] \\
&\quad\leq \mathbb{P}\left[|\underline{\mathcal{W}}_2(Q,\hat{\mu}[S]) - \mathbb{E}[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S])]| \geq t \mid \|\Sigma_Q\| \leq x\wedge\sqrt{x}\right] \\
&\quad\leq \mathbb{E}\left[2\exp\left(-\frac{\sqrt{M}t\wedge Mt^2}{\min\{2\sqrt{P_Q},6e^5P_Q\}}\right) \;\middle|\; \|\Sigma_Q\| \leq x\wedge\sqrt{x}\right],
\end{aligned}$$

where we used Theorem 3.8 from [119], as used in the proof of Proposition 7 in [147], and the last equality holds almost surely.

Conditional on $\|\Sigma_Q\| \leq x\wedge\sqrt{x}$, we get that

$$\begin{aligned}
\frac{1}{\min\{2\sqrt{P_Q},6e^5P_Q\}} &= \max\left\{\frac{1}{2\sqrt{P_Q}},\frac{1}{6e^5P_Q}\right\} \\
&\geq \frac{1}{6e^5}\max\left\{\frac{1}{\sqrt{P_Q}},\frac{1}{P_Q}\right\} \\
&\geq \frac{1}{6e^5}\max\left\{\frac{1}{\sqrt{c_d\|\Sigma_Q\|}},\frac{1}{c_d\|\Sigma_Q\|}\right\} \\
&\geq \frac{1}{6e^5\max\{\sqrt{c_d},c_d\}}\max\left\{\frac{1}{\sqrt{\|\Sigma_Q\|}},\frac{1}{\|\Sigma_Q\|}\right\} \\
&= \frac{1}{6e^5\max\{\sqrt{c_d},c_d\}}\frac{1}{\min\{\sqrt{\|\Sigma_Q\|},\|\Sigma_Q\|\}} \\
&\leq \frac{1}{6e^5\max\{\sqrt{c_d},c_d\}}\frac{1}{x\wedge\sqrt{x}} \\
&= \frac{1}{\tilde{C}_d(x\wedge\sqrt{x})}.
\end{aligned}$$

In the last inequality we used that

$$\min\{\sqrt{\|\Sigma_Q\|}, \|\Sigma_Q\|\} \leq \min\{\sqrt{x \wedge \sqrt{x}}, x \wedge \sqrt{x}\} \leq x \wedge \sqrt{x},$$

and in the last step we defined $\tilde{C}_d = 6e^5 \max\{\sqrt{c_d}, c_d\}$.

We therefore get (again almost surely) that

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad \leq \mathbb{E}\left[2\exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\min\{2\sqrt{P_Q}, 6e^5 P_Q\}}\right) \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad \leq \mathbb{E}\left[2\exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d(x \wedge \sqrt{x})}\right) \;\middle|\; \|\Sigma_Q\| \leq x \wedge \sqrt{x}\right] \\
&\quad = 2\exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d(x \wedge \sqrt{x})}\right).
\end{aligned}$$

Observe now that

$$2\exp\left(-\frac{\sqrt{M}t^2 \wedge Mt^2}{\tilde{C}_d(x \wedge \sqrt{x})}\right) = \frac{\delta}{2} \quad \Leftrightarrow \quad x \wedge \sqrt{x} = \frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)}$$

and since $\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)} > 0$ (recall that we restricted δ to $(0, 1/4)$), we can choose $x \in \mathbb{R}_{>0}$ such that the last display holds.

With this choice of x, we are now at

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq \frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)} C_d \sqrt{\frac{\ln(M)}{M}} + t\right] \\
&\quad = \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq (x \wedge \sqrt{x}) C_d \sqrt{\frac{\ln(M)}{M}} + t\right] \\
&\quad \leq 2\exp\left(-\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d(x \wedge \sqrt{x})}\right) + \rho_\Sigma\left(x \wedge \sqrt{x}\right) \\
&\quad \leq \frac{\delta}{2} + \rho_\Sigma\left(\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)}\right).
\end{aligned}$$

Note that this holds since the above computation works for any version of the con-

ditional expectation.

Next, let $s > 1$ and set $t = \frac{\ln(4/\delta)}{\sqrt{M}} s$, then

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq s\frac{C_d}{\tilde{C}_d}\sqrt{\frac{\ln(M)}{M}} + \frac{\ln(4/\delta)}{\sqrt{M}}s\right] \\
&= \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq \frac{\ln(4/\delta)s \wedge \ln(4/\delta)^2 s^2}{\tilde{C}_d \ln(4/\delta)} C_d\sqrt{\frac{\ln(M)}{M}} + \frac{\ln(4/\delta)}{\sqrt{M}}s\right] \\
&= \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq \frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)} C_d\sqrt{\frac{\ln(M)}{M}} + t\right] \\
&\leq \frac{\delta}{2} + \rho_\Sigma\left(\frac{\sqrt{M}t \wedge Mt^2}{\tilde{C}_d \ln(4/\delta)}\right) \\
&= \frac{\delta}{2} + \rho_\Sigma\left(\frac{s}{\tilde{C}_d}\right),
\end{aligned}$$

where we used that $\ln(4/\delta)s \wedge \ln(4/\delta)^2 s^2 = \ln(4/\delta)s$ since $\ln(4/\delta), s > 1$.

The condition $\mathbb{P}[\|\Sigma_Q\| \geq x] \leq \rho_\Sigma(x)$ for all $x \in \mathbb{R}_{\geq 0}$ implies that $\rho_\Sigma([0,\infty)) = (0,1]$, so we have

$$\rho_\Sigma\left(\frac{s}{\tilde{C}_d}\right) = \frac{\delta}{2} \quad \Leftrightarrow \quad s = \tilde{C}_d \rho_\Sigma^{-1}(\delta/2)$$

and since

$$s > 1 \quad \Leftrightarrow \quad \tilde{C}_d \rho_\Sigma^{-1}(\delta/2) > 1 \quad \Leftrightarrow \quad \delta < 2\rho_\Sigma(1/\tilde{C}_d),$$

our requirements on δ ensures that we can set $s = \tilde{C}_d \rho_\Sigma^{-1}(\delta/2)$. Altogether, we arrived at

$$\begin{aligned}
&\mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq \tilde{C}_d \rho_\Sigma^{-1}(\delta/2)\frac{C_d}{\tilde{C}_d}\sqrt{\frac{\ln(M)}{M}} + \frac{\ln(4/\delta)}{\sqrt{M}}\tilde{C}_d \rho_\Sigma^{-1}(\delta/2)\right] \\
&= \mathbb{P}\left[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S]) \geq s\frac{C_d}{\tilde{C}_d}\sqrt{\frac{\ln(M)}{M}} + \frac{\ln(4/\delta)}{\sqrt{M}}s\right] \leq \frac{\delta}{2} + \rho_\Sigma\left(\frac{s}{\tilde{C}_d}\right) = \delta
\end{aligned}$$

□

We can now formulate and prove the announced specializations of the oracle inequalities.

Corollary A.6.2. Consider the situation of Theorem A.3.1. Let $\mathcal{S} = \mathbb{R}^d$, set $\mathcal{H} = \mathcal{H}_2$, $\Pi = \Pi_2$, and $\hat{\Pi} = \hat{\Pi}_2$, and assume that Assumption A.2.1 holds in this case. Furthermore, for $(Q, y) \sim P$, assume that P-a.s. Q is a log-concave distribution, and denote its (P-a.s. defined) covariance matrix by Σ_Q. Assume that ℓ is convex, differentiable, ℓ' is locally Lipschitz continuous, and that there exists $B_\ell, B'_\ell \in \mathbb{R}_{\geq 0}$ such that $\ell(x, y, 0) \leq B_\ell$ and $|\ell'(x, y, 0)| \leq B'_\ell$ for all $x \in \mathcal{X}, y \in \mathcal{Y}$. Let k be a kernel on $\mathcal{H}$ that is measurable and bounded, and that there exists $\alpha_k \in \mathcal{K}$ such that $\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \alpha_k(\|x_1 - x_2\|)$. We then have for all $\lambda \in \mathbb{R}_{>0}$ and $\delta \in (0, 1)$ that with probability at least $1 - \delta$

$$\begin{aligned}
&\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^*_{\ell,P} \leq A^{(2)}_{\ell,P}(\lambda) \\
&\quad + \frac{2\sqrt{\lambda B_\ell} + |\ell|_{1,B_f}\|k\|_\infty}{N} \sum_{n=1}^{N} \alpha_\lambda \left(\frac{\rho_\Sigma^{-1}\left(\frac{\delta}{2(N+|\mathcal{F}|)}\right)}{\sqrt{M}} \left(C_d \sqrt{\ln(M)} + \tilde{C}_d \ln\left(\frac{4(N+|\mathcal{F}|)}{\delta}\right)\right)\right) \\
&\quad + 2|\ell|_{1,B_f}\|k\|_\infty \left(B'_\ell + \gamma_{3,B_f}(B_f)\right) \left(\sqrt{\frac{2\ln(2N/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2N/\delta)}{3N} \right),
\end{aligned}$$

with B_f and α_λ as in Theorem A.3.1, and C_d and $\tilde{C}_d$ are universal constants that only depend on d.

Proof. The result follows immediately by combining Theorem A.3.1 with Proposition A.6.1. □

Corollary A.6.3. Consider the situation of Theorem A.3.1. Let $\mathcal{S} = \mathbb{R}^d$, set $\mathcal{H} = \mathcal{H}_2$, $\Pi = \Pi_2$, and $\hat{\Pi} = \hat{\Pi}_2$, and assume that Assumption A.2.1 holds in this case. Furthermore, for $(Q, y) \sim P$, assume that P-a.s. Q is a log-concave distribution, and denote its (P-a.s. defined) covariance matrix by Σ_Q. Finally, assume that there exists $\rho_\Sigma \in \mathcal{L}$ such that for all $t \in \mathbb{R}_{\geq 0}$, $\mathbb{P}[\|\Sigma_Q\| \geq t] \leq \rho_\Sigma(t)$ P-a.s. We then have for all $0 < \delta < \min\{1/4, 2\rho_\Sigma(1/\tilde{C}_d)\}$ that with probability at least $1 - \delta$ it holds that

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D},\lambda}) - \mathcal{R}^{H_k *}_{\ell,P} &\leq A^{(2)}_{\ell,P}(\lambda) + 2\left(B_\ell + \gamma_{3,\tilde{B}_f}(\tilde{B}_f)\right)\sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}} + 4\gamma_{3,\tilde{B}_f}(\epsilon) \\
&\quad + \frac{2}{N}\sum_{n=1}^{N} \alpha_\lambda \left(\frac{\rho_\Sigma^{-1}\left(\frac{\delta}{2(N+|\mathcal{F}|)}\right)}{\sqrt{M}} \left(C_d \sqrt{\ln(M)} + \tilde{C}_d \ln\left(\frac{4(N+|\mathcal{F}|)}{\delta}\right)\right)\right),
\end{aligned}$$

where we defined $\tilde{B}_f = \|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}} + \epsilon$, $\alpha_\lambda = \gamma_1 + \gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}$, and C_d and

$\tilde{C}_d$ are universal constants that only depend on d.

Proof. The result follows immediately by combining Theorem A.3.4 with Proposition A.6.1. □

A.6.2. Proof of the Oracle Inequalities

We will need the following result, which is derived at the beginning of Section 5.2 in [189], but not stated as a theorem there. For convenience, we repeat it here.

Proposition A.6.4. Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ a loss function that is convex, differentiable, and define $\ell' = \frac{\mathrm{d}}{\mathrm{dt}}\ell$. Let k be a kernel on $\mathcal{X}$ that is measurable, bounded, and has a separable RKHS H_k. For all $P \in \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ such that ℓ and $|\ell'|$ are P-integrable Nemitskii losses, and for all $\lambda \in \mathbb{R}_{>0}$, there exists a unique solution $f_{P,\lambda}$ of

$$\min_{f \in H_k} \mathcal{R}_{\ell,P}(f) + \lambda \|f\|_k^2, \tag{A.14}$$

and this solution fulfills the equation

$$f_{P,\lambda} = -\frac{1}{2\lambda} \int_{\mathcal{X}\times\mathcal{Y}} \ell'(x, y, f_{P,\lambda}(x)) \Phi_k(x) \mathrm{d}P(x, y). \tag{A.15}$$

Note that in (A.15) a Bochner integral appears.

Proof of Theorem A.3.1. Let $\lambda \in \mathbb{R}_{>0}$ be arbitrary and define $\bar{\mathcal{D}} = ((Q_n, y_n))_{n\in[N]}$. We then have

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P,\lambda} &= \mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P,\lambda}(f_{P,\lambda}) \\
&= \mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 + \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) \\
&\quad + \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) + \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{P,\lambda}) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{P,\lambda}) \\
&\quad - \mathcal{R}_{\ell,P}(f_{P,\lambda}) - \lambda\|f_{P,\lambda}\|_k^2 + \lambda\|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 - \lambda\|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 \\
&= \underbrace{\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda})}_{I} + \underbrace{\mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,P}(f_{P,\lambda})}_{II} \\
&\quad + \underbrace{\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) + \lambda\|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2 - (\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{P,\lambda}) + \lambda\|f_{P,\lambda}\|_k^2)}_{=III} \\
&\quad + \underbrace{\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{P,\lambda}) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f_{\bar{\mathcal{D}}_\Pi,\lambda})}_{=IV} + \underbrace{\lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \lambda\|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k^2}_{=V}
\end{aligned}$$

We now upper bound terms I to V. First, by definition of $f_{\bar{\mathcal{D}}_\Pi,\lambda}$, term III is nonpositive, and hence can be discarded.

In order to bound the remaining terms, we need some preparations. Lemma A.5.2 ensures that for all distributions Q on $\mathcal{X}\times\mathcal{Y}$, ℓ is a Q-integrable Nemitskii loss. Furthermore, repeating the proof of Lemma A.5.2 on ℓ' shows that also $|\ell'|$ is a Q-integrable Nemitskii loss. Altogether, we can apply Proposition A.6.4 to ℓ for any distribution Q on $\mathcal{X}\times\mathcal{Y}$. An inspection of the proof of Theorem 5.9 in [189] reveals that (5.14) in this reference applies to the present situation, so for all distributions $Q,\tilde{Q}$ on $\mathcal{X}\times\mathcal{Y}$, unique SVM solutions $f_{Q,\lambda}$ and $f_{\tilde{Q},\lambda}$ exist, and we have

$$\|f_{Q,\lambda}-f_{\tilde{Q},\lambda}\|_k \leq \frac{1}{\lambda}\left\|\int h_Q(x,y)\Phi_k(x)\mathrm{d}Q(x,y)-\int h_Q(x,y)\Phi_k(x)\mathrm{d}\tilde{Q}(x,y)\right\|_k, \tag{A.16}$$

where we defined $h_Q(x,y)=\ell'(x,y,f_{Q,\lambda}(x))$.

Bounding I Using Lemma A.5.4, we have $\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k, \|f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$, hence we get from Lemma A.5.1 that $|f_{\mathcal{D}_{\hat{\Pi}},\lambda}(x)|, |f_{\bar{\mathcal{D}}_\Pi,\lambda}(x)| \leq \|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}} =: B_f$. Define now for brevity $L_\ell := |\ell|_{1,B_f}$, then we get

$$\begin{aligned}
&\left|\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda})-\mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda})\right| \leq L_\ell\|k\|_\infty\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}-f_{\bar{\mathcal{D}}_\Pi,\lambda}\|_k\\
&\quad\leq \frac{L_\ell\|k\|_\infty}{\lambda}\left\|\frac{1}{N}\sum_{n=1}^N h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)},y_n)\Phi_k(\hat{\Pi}S^{(n)})-\frac{1}{N}\sum_{n=1}^N h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n,y_n)\Phi_k(\Pi Q_n)\right\|_k\\
&\quad\leq \frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^N\left\|h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)},y_n)\Phi_k(\hat{\Pi}S^{(n)})-h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n,y_n)\Phi_k(\Pi Q_n)\right\|_k\\
&\quad\leq \frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^N|h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)},y_n)-h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n,y_n)|\|\Phi_k(\hat{\Pi}S^{(n)})\|_k\\
&\qquad+|h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n,y_n)|\|\Phi_k(\hat{\Pi}S^{(n)})-\Phi_k(\Pi Q_n)\|_k
\end{aligned}$$

where we used Lemma A.5.3 in the first inequality, in the second step the bound (A.16), followed by using the triangle inequality twice. For each $n=1,\ldots,N$, we have

$$\begin{aligned}
|h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)},y_n)-h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n,y_n)| &= |\ell'(\hat{\Pi}S^{(n)},y_n,f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\hat{\Pi}S^{(n)}))-\ell'(\Pi Q_n,y_n,f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n))|\\
&\leq \gamma_1(\|\hat{\Pi}S^{(n)}-\Pi Q_n\|_{\mathcal{H}})+\gamma_{3,B_f}(|f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\hat{\Pi}S^{(n)})-f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n)|)\\
&\leq \left(\gamma_1+\gamma_{3,B_f}\circ\left(\sqrt{B_\ell/\lambda}\cdot\alpha_k\right)\right)(\|\hat{\Pi}S^{(n)}-\Pi Q_n\|_{\mathcal{H}}),
\end{aligned}$$

where we used the definition of $h_{\mathcal{D}_{\hat{\Pi}}}$ in the first step, and in the following inequality we used the assumed continuity property of ℓ' (together with the previously derived bound B_f on the values of $f_{\mathcal{D}_{\hat{\Pi}},\lambda}$ and $f_{\mathcal{D}_{\hat{\Pi}},\lambda}$). In the last inequality we used that for all $f \in H_k$ and $x_1, x_2 \in \mathcal{X}$,

$$|f(x_1) - f(x_2)| = |\langle f, \Phi_k(x_1) - \Phi_k(x_2)\rangle_k| \leq \|f\|_k \|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \|f\|_k \alpha_k(\|x_1 - x_2\|_{\mathcal{H}}).$$

Furthermore, we also have $\|\Phi_k(\hat{\Pi}S^{(n)}) - \Phi_k(\Pi Q_n)\|_k \leq \alpha_k(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}})$ and $\|\Phi_k(\hat{\Pi}S^{(n)})\|_k \leq \|k\|_\infty$. Finally,

$$\begin{aligned}
|h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)| &= |\ell'(\Pi Q_n, y_n, f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n))| \\
&\leq |\ell'(\Pi Q_n, y_n, 0)| + |\ell'(\Pi Q_n, y_n, f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n)) - \ell'(\Pi Q_n, y_n, 0)| \\
&\leq B'_\ell + \gamma_{3,B_f}(|f_{\mathcal{D}_{\hat{\Pi}},\lambda}(\Pi Q_n)|) \\
&\leq B'_\ell + \gamma_{3,B_f}(B_f).
\end{aligned}$$

Altogether, we can continue with

$$\begin{aligned}
&\left|\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda})\right| \\
&\quad\leq \frac{L_\ell \|k\|_\infty}{\lambda} \frac{1}{N} \sum_{n=1}^{N} |h_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}, y_n) - h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)| \|\Phi_k(\hat{\Pi}S^{(n)})\|_k \\
&\qquad + |h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)| \|\Phi_k(\hat{\Pi}S^{(n)}) - \Phi_k(\Pi Q_n)\|_k \\
&\quad\leq \frac{L_\ell \|k\|_\infty}{\lambda} \frac{1}{N} \sum_{n=1}^{N} \|k\|_\infty \left(\gamma_1 + \gamma_{3,B_f} \circ \left(\sqrt{B_\ell/\lambda} \cdot \alpha_k\right)\right) (\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}}) \\
&\qquad + \left(B'_\ell + \gamma_{3,B_f}(B_f)\right) \alpha_k(\|\hat{\Pi}S^{(n)} - \Pi Q_n)\|_{\mathcal{H}}) \\
&\quad\leq \frac{L_\ell \|k\|_\infty}{\lambda} \frac{1}{N} \sum_{n=1}^{N} \left(\|k\|_\infty \left(\gamma_1 + \gamma_{3,B_f} \circ \left(\sqrt{B_\ell/\lambda} \cdot \alpha_k\right)\right)\right. \\
&\qquad\qquad \left. + \left(B'_\ell + \gamma_{3,B_f}(B_f)\right) \alpha_k\right) (\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}})
\end{aligned}$$

Defining $\alpha_\lambda = \|k\|_\infty(\gamma_1 + \gamma_{3,B_f} \circ \left(\sqrt{B_\ell/\lambda}\alpha_k\right)) + \left(B'_\ell + \gamma_{3,B_f}(B_f)\right) \alpha_k$ and using a union bound, we finally get with probability at least $1 - \delta/2$ that

$$\left|\mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P}(f_{\bar{\mathcal{D}}_\Pi,\lambda})\right| \leq \frac{L_\ell \|k\|_\infty}{\lambda} \frac{1}{N} \sum_{n=1}^{N} \alpha_\lambda(B_n(\delta/(2N))).$$

Bounding II and IV Let $Q = P$ or $\bar{\mathcal{D}}_\Pi$. We have

$$\begin{aligned}
&\mathcal{R}_{\ell,Q}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,Q}(f_{P,\lambda}) \leq L_\ell \|k\|_\infty \|f_{\bar{\mathcal{D}}_\Pi,\lambda} - f_{P,\lambda}\|_k \\
&\quad \leq \frac{L_\ell \|k\|_\infty}{\lambda} \left\| \frac{1}{N} \sum_{n=1}^N h_{\bar{\mathcal{D}}_\Pi}(\Pi Q_n, y_n) \Phi_k(\Pi Q_n) - \int h_{\bar{\mathcal{D}}_\Pi}(x,y) \Phi_k(x) \mathrm{d}P(x,y) \right\|_k \\
&\quad = \frac{L_\ell \|k\|_\infty}{\lambda} \left\| \frac{1}{N} \sum_{n=1}^N \xi_n - \mathbb{E}[\xi_n] \right\|_k
\end{aligned}$$

where the first two steps are similar as in bounding I, and in the last step we defined $\xi_n = h_{\bar{\mathcal{D}}_\Pi}(\Pi Q_n, y_n)\Phi_k(\Pi Q_n)$. Since $(Q_1, y_1) \ldots, (Q_N, y_N) \overset{\text{i.i.d.}}{\sim} P$, also $\xi_1, \ldots, \xi_N$ are i.i.d. Furthermore,

$$\begin{aligned}
\|\xi_n\|_k &= \|h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)\Phi_k(\Pi Q_n)\|_k = |h_{\mathcal{D}_{\hat{\Pi}}}(\Pi Q_n, y_n)| \|\Phi_k(\Pi Q_n)\|_k \\
&\leq |\ell'(\Pi Q_n, y_n, f_{\bar{\mathcal{D}}_\Pi,\lambda}(\Pi Q_n))| \|k\|_\infty \\
&\leq \left(B'_\ell + \gamma_{3,B_f}(B_f) \right) \|k\|_\infty,
\end{aligned}$$

so $\xi_1, \ldots, \xi_N$ are H_k-valued i.i.d. random variables bounded by $B_\xi := \left(B'_\ell + \gamma_{3,B_f}(B_f) \right) \|k\|_\infty$. Hoeffding's inequality for random variables in a separable Hilbert space, cf. Corollary 6.15 in [189], now ensures that with probability at least $1 - \delta/2$

$$\left\| \frac{1}{N} \sum_{n=1}^N \xi_n - \mathbb{E}[\xi_n] \right\|_k \leq B_\xi \left(\sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N} \right).$$

This implies that with probability at least $1 - \delta/2$

$$\mathcal{R}_{\ell,Q}(f_{\bar{\mathcal{D}}_\Pi,\lambda}) - \mathcal{R}_{\ell,Q}(f_{P,\lambda}) \leq \frac{L_\ell \|k\|_\infty}{\lambda} B_\xi \left(\sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N} \right),$$

so with same probability the bound

$$II + IV \leq 2\frac{L_\ell \|k\|_\infty}{\lambda} B_\xi \left(\sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N} \right)$$

holds.

Bounding V Using elementary computations, we get

$$\begin{aligned}
\lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \lambda\|f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k^2 &= \lambda\left(\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \|f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k^2\right) \\
&= \lambda\left(\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k + \|f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k\right)\left(\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k - \|f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k\right) \\
&\leq \lambda\left(\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k + \|f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k\right)\|f_{\mathcal{D}_{\hat{\Pi}},\lambda} - f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k \\
&\leq 2\lambda\sqrt{\frac{B_\ell}{\lambda}}\|f_{\mathcal{D}_{\hat{\Pi}},\lambda} - f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k \\
&\leq 2\sqrt{\frac{B_\ell}{\lambda}}\frac{1}{N}\sum_{n=1}^{N}\alpha_\lambda(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}}),
\end{aligned}$$

where we used Lemma A.5.4 in the second to last step, and the bound on $\|f_{\mathcal{D}_{\hat{\Pi}},\lambda} - f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k$ from bounding I. In particular, with probability at least $1-\delta/2$ we get that

$$\lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 - \lambda\|f_{\bar{\mathcal{D}}_{\Pi},\lambda}\|_k^2 \leq 2\sqrt{\frac{B_\ell}{\lambda}}\frac{1}{N}\sum_{n=1}^{N}\alpha_\lambda(B_n(\delta/(2N))).$$

Finishing Using again a union bound, we finally get that with probability at least $1-\delta$ we have

$$\begin{aligned}
\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P,\lambda}^{H_k*} &\leq \underbrace{\frac{L_\ell\|k\|_\infty}{\lambda}\frac{1}{N}\sum_{n=1}^{N}\alpha_\lambda(B_n(\delta/(2N)))}_{\text{from I}} \\
&\quad + \underbrace{2\frac{L_\ell\|k\|_\infty}{\lambda}B_\xi\left(\sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N}\right)}_{\text{from II and IV}} \\
&\quad + \underbrace{2\sqrt{\frac{B_\ell}{\lambda}}\frac{1}{N}\sum_{n=1}^{N}\alpha_\lambda(B_n(\delta/(2N)))}_{\text{from V}} \\
&= \left(2\sqrt{\frac{B_\ell}{\lambda}} + \frac{L_\ell\|k\|_\infty}{\lambda}\right)\frac{1}{N}\sum_{n=1}^{N}\alpha_\lambda(B_n(\delta/(2N))) \\
&\quad + 2\frac{L_\ell\|k\|_\infty}{\lambda}B_\xi\left(\sqrt{\frac{2\ln(2/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2/\delta)}{3N}\right)
\end{aligned}$$

The result now follows from the definition of $A^{(2)}_{\ell,P}(\lambda)$. □

Proof of Corollary A.3.3. Since $\mathcal{S}$ is compact, it is in particular separable, so Proposition A.2.4 ensures that Π_k is $(\mathcal{M}_1(\mathcal{S}), \mathcal{B}(\tau_w))$-$(H_\kappa, \mathcal{B}(H_\kappa))$-measurable. Furthermore, since $\mathcal{S}$ is a compact metric space, $\mathcal{M}_1(\mathcal{S})$ with the topology of weak convergence is compact. Since κ is universal, Π_k is continuous, and hence $\mathcal{X} = \Pi_k(\mathcal{M}_1(\mathcal{S}))$ is a compact metric space. In particular, it is closed, and hence $\mathcal{X} \in \mathcal{B}(H_\kappa)$, and it is also separable. By definition, for all $M \in \mathbb{N}_+$ and $S \in \mathcal{S}^M$, $\hat{\Pi}_M(S) = \hat{\Pi}_k(S) = \frac{1}{M}\sum_{m=1}^M k(\cdot, S_m)$, and hence measurable. Altogether, Assumption A.2.1 is fulfilled.

Next, for all $x_1, x_2 \in \mathcal{X}$ we have $\|\Phi_k(x_1) - \Phi_k(x_2)\|_k \leq \alpha_k(\|x_1 - x_2\|_{\mathcal{H}})$, which shows that Φ_k is continuous, so according to Lemma 4.29 in [189] also k is continuous. Since $\mathcal{X}$ is separable, this shows that also H_k is separable.

Using the KME estimation bound from Proposition A.2.4 to find appropriate B_n, all assumptions of Theorem A.3.1 are fulfilled, and we get

$$\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}^{H_k*}_{\ell,P,\lambda} \leq (2\sqrt{\lambda B_\ell} + L_\ell\|k\|_\infty)\frac{1}{N}\sum_{n=1}^N \alpha_\lambda \left(2\sqrt{\frac{\|k\|^2_\infty}{M_n}} + \sqrt{\frac{2\|\kappa\|_\infty \ln(2N/\delta)}{M_n}}\right)$$
$$+ 2L_\ell\|k\|_\infty \left(B'_\ell + L'_\ell\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}\right)\left(\sqrt{\frac{2\ln(2N/\delta)}{N}} + \sqrt{1/N} + \frac{4\ln(2N/\delta)}{3N}\right),$$

where we defined

$$\alpha_\lambda = \|k\|_\infty L'_\ell \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}} + \left(B'_\ell + L'_\ell\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}\right)\alpha_k.$$

Finally, since ℓ is locally Lipschitz continuous, it is in particular continuous, and as shown by Lemma A.5.2, it is also a P-integrable Nemitskii loss. Together with the fact that $\mathcal{X}$ is a compact metric space and k is universal, Corollary 5.29 in [189] shows that $\mathcal{R}^{H_k*}_{\ell,P,\lambda} = \mathcal{R}^*_{\ell,P}$, and the result follows. □

The strategy of the following proof follows the one for Theorem 6.25 in [189], however, several adaptions are necessary to deal with the two-stage sampling.

Proof of Theorem A.3.4. Let $\lambda \in \mathbb{R}_{>0}$ be arbitrary. We start with

$$\begin{aligned}\mathcal{R}_{\ell,P,\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,P,\lambda}(f_{P,\lambda}) &= \mathcal{R}_{\ell,P}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) - \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) + \lambda\|f_{\mathcal{D}_{\hat{\Pi}},\lambda}\|_k^2 \\ &\quad - \left(\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{P,\lambda}) + \lambda\|f_{P,\lambda}\|_k^2\right) + \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{P,\lambda}) - \mathcal{R}_{\ell,P}(f_{P,\lambda}) \\ &\leq 2 \sup_{\substack{f\in H_k \\ \|f\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}}} \left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,P}(f)\right|,\end{aligned}$$

where we used in the last step that $\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}},\lambda}(f_{\mathcal{D}_{\hat{\Pi}},\lambda}) \leq \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}},\lambda}(f_{P,\lambda})$ by definition of $f_{\mathcal{D}_{\hat{\Pi}},\lambda}$, and we applied Lemma A.5.4 to $f_{\mathcal{D}_{\hat{\Pi}},\lambda}$ and $f_{P,\lambda}$.

Let $f \in H_k$ with $\|f\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$, and choose $\tilde{f} \in \mathcal{F}$ such that $\|f - \tilde{f}\|_k \leq \epsilon$. Observe that $|\tilde{f}(x)| \leq |f(x)| + |\tilde{f}(x) - f(x)| \leq \|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}} + \epsilon = \tilde{B}_f$, where we used the choice of $\tilde{f}$ together with (the proof of) Lemma A.5.2. We then have

$$\begin{aligned}\left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,P}(f)\right| &\leq \left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f)\right| + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(\tilde{f})\right| \\ &\quad + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(\tilde{f}) - \mathcal{R}_{\ell,P}(\tilde{f})\right| + \left|\mathcal{R}_{\ell,P}(\tilde{f}) - \mathcal{R}_{\ell,P}(f)\right| \\ &\leq \left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f)\right| + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(\tilde{f}) - \mathcal{R}_{\ell,P}(\tilde{f})\right| + 2\gamma_{3,\tilde{B}_f}(\epsilon),\end{aligned}$$

where we used (a modified variant of) Lemma A.5.3 in the last step together with $|f(x)|, |\tilde{f}(x)| \leq \tilde{B}_f$.

We now bound the first two terms. First,

$$\begin{aligned}\left|\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}_\Pi}(f)\right| &\leq \frac{1}{N}\sum_{n=1}^N \left|\ell(\hat{\Pi}S^{(n)}, y_n, f(\hat{\Pi}S^{(n)})) - \ell(\Pi Q_n, y_n, f(\Pi Q_n))\right| \\ &\leq \frac{1}{N}\sum_{n=1}^N \gamma_1(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}}) + \gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}}\left(\left|f(\hat{\Pi}S^{(n)}) - f(\Pi Q_n)\right|\right) \\ &\leq \frac{1}{N}\sum_{n=1}^N \left(\gamma_1 + \gamma_{3,\|k\|_\infty\sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right)(\|\hat{\Pi}S^{(n)} - \Pi Q_n\|_{\mathcal{H}}),\end{aligned}$$

where we used the triangle inequality, then the continuity property of ℓ, and then the continuity property of f.

Second,

$$\left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{f}) - \mathcal{R}_{\ell,P}(\tilde{f})\right| = \left|\frac{1}{N}\sum_{n=1}^N \ell(\Pi Q_n, y_n, \tilde{f}(\Pi Q_n)) - \int \ell(\Pi Q, y, \tilde{f}(\Pi Q))\mathrm{d}P(Q,y)\right|,$$

$\ell(\Pi Q_1, y_1, \tilde{f}(\Pi Q_1)), \ldots, \ell(\Pi Q_N, y_N, \tilde{f}(\Pi Q_N))$ are i.i.d. random variables (since the (Q_n, y_n) are i.i.d.), and for all $n = 1, \ldots, N$ we have $|\ell(\Pi Q_n, y_n, \tilde{f}(\Pi Q_n))| \leq B_\ell + \gamma_{3,\tilde{B}_f}(\tilde{B}_f) = B_\xi$ according to (the proof of) Lemma A.5.2. All of this means that we can use Hoeffding's inequality to bound this term.

Third, we can combine the previous two bounds. Using the union bound we have

$$\begin{aligned}
&\mathbb{P}\left[\max_{n=1,\ldots,N} \|\hat{\Pi} S^{(n)} - \Pi Q_n\|_{\mathcal{H}} > B_n(\delta/(N+|\mathcal{F}|)) \text{ or } \max_{\tilde{g}\in\mathcal{F}} \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{g}) - \mathcal{R}_{\ell,P}(\tilde{g})\right| > B_\xi \sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}}\right] \\
&\leq \sum_{n=1}^{N} \mathbb{P}\left[\|\hat{\Pi} S^{(n)} - \Pi Q_n\|_{\mathcal{H}} > B_n(\delta/(N+|\mathcal{F}|)\right] + \sum_{\tilde{g}\in\mathcal{F}} \mathbb{P}\left[\left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{g}) - \mathcal{R}_{\ell,P}(\tilde{g})\right| > B_\xi \sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}}\right] \\
&\leq N\frac{\delta}{N+|\mathcal{F}|} + |\mathcal{F}|\frac{\delta}{N+|\mathcal{F}|} = \delta
\end{aligned}$$

Together with our previous two bounds this implies that with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\mathcal{R}_{\ell,\mathcal{D}}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}}(f)\right| + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{f}) - \mathcal{R}_{\ell,P}(\tilde{f})\right| \\
&\leq \frac{1}{N}\sum_{n=1}^{N} \left(\gamma_1 + \gamma_{3,\|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right) (\|\hat{\Pi} S^{(n)} - \Pi Q_n\|_{\mathcal{H}}) + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{f}) - \mathcal{R}_{\ell,P}(\tilde{f})\right| \\
&\leq \frac{1}{N}\sum_{n=1}^{N} \left(\gamma_1 + \gamma_{3,\|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right) B_n(\delta/(N+|\mathcal{F}|)) + B_\xi \sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}}.
\end{aligned}$$

This also implies that with probability at least $1-\delta$,

$$\begin{aligned}
|\mathcal{R}_{\ell,\mathcal{D}}(f) - \mathcal{R}_{\ell,P}(f)| &\leq \left|\mathcal{R}_{\ell,\mathcal{D}}(f) - \mathcal{R}_{\ell,\bar{\mathcal{D}}}(f)\right| + \left|\mathcal{R}_{\ell,\bar{\mathcal{D}}}(\tilde{f}) - \mathcal{R}_{\ell,P}(\tilde{f})\right| + 2|\ell|_{1,\tilde{B}_f}\epsilon \\
&\leq \frac{1}{N}\sum_{n=1}^{N} \left(\gamma_1 + \gamma_{3,\|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}}\right) (B_n(\delta/(N+|\mathcal{F}|)) \\
&\quad + B_\xi \sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}} + 2|\ell|_{1,\tilde{B}_f}\epsilon,
\end{aligned}$$

and since $f \in H_k$ with $\|f\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}$ was arbitrary, this in turn implies that

$$\sup_{\substack{f \in H_k \\ \|f\|_k \leq \sqrt{\frac{B_\ell}{\lambda}}}} |\mathcal{R}_{\ell,\mathcal{D}}(f) - \mathcal{R}_{\ell,P}(f)| \leq \frac{1}{N} \sum_{n=1}^{N} \left(\gamma_1 + \gamma_{3,\|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}}} \circ \alpha_{f,\sqrt{\frac{B_\ell}{\lambda}}} \right) (B_n(\delta/(N+|\mathcal{F}|)) + B_\xi \sqrt{\frac{2\ln((N+|\mathcal{F}|)/\delta)}{N}} + 2\gamma_{3,\tilde{B}_f}(\epsilon),$$

with probability at least $1-\delta$, and the result follows. □

A.7. Additional Material on Generalization via Algorithmic Stability

A.7.1. Sliced Wasserstein

Corollary A.7.1. Consider the situation of Theorem A.4.2. Additionally, assume $\mathcal{S} = \mathbb{R}^d$, let $(\mathcal{H}, \Pi)$ be the sliced 2-Wasserstein embedding, and assume that the support of $\pi_{\mathcal{S}} \sharp P$[8] is contained in the set of log-concave distributions, and for $(Q, y) \sim P$, denote by Σ_Q the (a.s.) defined covariance matrix of Q. We then have for all $\delta \in (0,1)$, with probability at least $1-\delta$, that

$$\mathcal{R}_{\ell,P}(f_{\ell,\mathcal{D}_{\hat{\Pi}}}\lambda) \leq \mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(f_{\ell,\mathcal{D}_{\hat{\Pi}}}\lambda) + \alpha_\lambda \left(C_d \mathbb{E}\left[\sqrt{\frac{\|\Sigma_Q\| \ln(M)}{M}} \right] \right) + \left(\frac{2|\ell|_{1,B_f}^2 \|k\|_\infty^2}{\lambda} + B_\ell + |\ell|_{1,B_f} B_f \right) \sqrt{\frac{\ln(1/\delta)}{2N}} + \frac{|\ell|_{1,B_f}^2 \|k\|_\infty^2}{\lambda N},$$

where we defined $B_f = \|k\|_\infty \sqrt{\frac{B_\ell}{\lambda}}$, and $C_d \in \mathbb{R}_{>0}$ is a universal constant that depends only on d.

Proof. Let $Q \in \mathcal{M}_1(\mathcal{S})$ and $M \in \mathbb{N}_+$. According to Theorem 1 in [147], we have

$$\mathbb{E}_{S \sim Q^{\otimes M}}[\underline{\mathcal{W}}_2(Q, \hat{\mu}[S])] \leq C_d \sqrt{\frac{\|\Sigma_Q\| \ln(M)}{M}},$$

[8] $\pi_{\mathcal{S}}$ is the usual coordinate projection onto $\mathcal{S}$.

where $C_d \in \mathbb{R}_{>0}$ is a universal constant that depends only on d. This implies that

$$\begin{aligned}\alpha_\lambda\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi}S\|_{\mathcal{H}}\right]\right) &= \alpha_\lambda\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\underline{\mathcal{W}}_2(Q,\hat{\mu}[S])\right]\right)\\ &\leq \alpha_\lambda\left(\mathbb{E}\left[C_d\sqrt{\frac{\|\Sigma_Q\|\ln(M)}{M}}\right]\right),\end{aligned}$$

with α_λ defined in Theorem A.4.2. This result now establishes the claim. □

A.7.2. Proof of the general result

Our proof follows the one of Theorem 14.2 in [141], adapted to the present distributional setting.

Proof of Theorem A.4.1. Define $F : (\mathcal{X}\times\mathcal{Y})^N \to \mathbb{R}$ by $F(D) = \mathcal{R}_{\ell,P}(\mathcal{L}_D) - \mathcal{R}_{\ell,D}(\mathcal{L}_D)$. Let $N \in \mathbb{N}_+$, $D \in (\mathcal{X}\times\mathcal{Y})^N$, $1 \leq i \leq N$ and $(\tilde{x},\tilde{y}) \in \mathcal{X}\times\mathcal{Y}$ be arbitrary. Define $\tilde{D} \in (\mathcal{X}\times\mathcal{Y})^N$ by

$$\tilde{D}_n = \begin{cases} D_n & \text{if } n \neq i \\ (\tilde{x},\tilde{y}) & \text{if } n = i \end{cases}$$

and for all $1 \leq n \leq N$, define also $(\tilde{x}_n,\tilde{y}_n) = D_n$. We then have

$$\begin{aligned}|F(D) - F(\tilde{D})| &= \left|\mathcal{R}_{\ell,P}(\mathcal{L}_D) - \mathcal{R}_{\ell,D}(\mathcal{L}_D) - \left(\mathcal{R}_{\ell,P}(\mathcal{L}_{\tilde{D}}) - \mathcal{R}_{\ell,\tilde{D}}(\mathcal{L}_{\tilde{D}})\right)\right|\\ &\leq |\mathcal{R}_{\ell,P}(\mathcal{L}_D) - \mathcal{R}_{\ell,P}(\mathcal{L}_{\tilde{D}})| + \left|\frac{1}{N}\sum_{n=1}^N \ell(x_n,y_n,\mathcal{L}_D(x_n)) - \frac{1}{N}\sum_{n=1}^N \ell(\tilde{x}_n,\tilde{y}_n,\mathcal{L}_{\tilde{D}}(\tilde{x}_n))\right|\\ &\leq \int |\ell(\Pi Q,y,\mathcal{L}_D(\Pi Q)) - \ell(\Pi Q,y,\mathcal{L}_{\tilde{D}}(\Pi Q))|\mathrm{d}P(Q,y)\\ &\quad + \frac{1}{N}\left|\ell(x_i,y_i,\mathcal{L}_D(x_i)) - \ell(\tilde{x}_i,\tilde{y}_i,\mathcal{L}_{\tilde{D}}(\tilde{x}_i)) + \sum_{\substack{n=1\\n\neq i}}^N \ell(x_n,y_n,\mathcal{L}_D(x_n)) - \ell(\tilde{x}_n,\tilde{y}_n,\mathcal{L}_{\tilde{D}}(\tilde{x}_n))\right|\\ &\leq \beta_N + \frac{1}{N}|\ell(x_i,y_i,\mathcal{L}_D(x_i)) - \ell(\tilde{x},\tilde{y},\mathcal{L}_{\tilde{D}}(\tilde{x}))| + \frac{1}{N}\sum_{\substack{n=1\\n\neq i}}^N |\ell(x_n,y_n,\mathcal{L}_D(x_n)) - \ell(x_n,y_n,\mathcal{L}_{\tilde{D}}(x_n))|\\ &\leq \beta_N + \frac{B}{N} + \frac{N-1}{N}\beta_N = \left(1+\frac{N-1}{N}\right)\beta_N + \frac{B}{N} = C.\end{aligned}$$

McDiarmid's bounded difference inequality then shows that for all $\delta \in (0,1)$, we

have with probability at least $1-\delta$ that

$$\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) \leq \mathbb{E}\left[\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})\right] + C\sqrt{\frac{N\ln(1/\delta)}{2}}$$

We now bound upper bound the expectation in the preceding display. We have

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})\right] = &\underbrace{\mathbb{E}\left[\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) - \mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right]\right]}_{=I} \\ &+ \underbrace{\mathbb{E}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right] - \mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})\right]}_{=II}\end{aligned}$$

and bound the two terms separately. Observe that

$$\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}) = \mathbb{E}_{(Q,y)\sim P}\left[\ell(\Pi Q, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\Pi Q))\right] = \mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\Pi Q, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\Pi Q))\right],$$

so we have

$$\begin{aligned}I &= \mathbb{E}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\Pi Q, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\Pi Q)) - \ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right]\right] \\ &\leq \mathbb{E}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\left|\ell(\Pi Q, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\Pi Q)) - \ell(\hat{\Pi}S, y, \mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right|\right]\right] \\ &\leq \mathbb{E}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\alpha(\|\Pi Q - \hat{\Pi}S\|_{\mathcal{H}})\right]\right] \\ &\leq \alpha\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q - \hat{\Pi}S\|_{\mathcal{H}}\right]\right),\end{aligned}$$

where we used Jensen's inequality together with the concavity of α in the last step.

We turn to term II. Let $(S^{(N+1)}, y_{N+1}) \sim \tilde{P}$ such that $(S^{(1)}, y_1), \ldots, (S^{(N+1)}, y_{N+1})$ are i.i.d., and define $\tilde{\mathcal{D}} = ((S^{(2)}, y_2), \ldots, (S^{(N+1)}, y_{N+1})$. Note that $\mathcal{D}$ and $\tilde{\mathcal{D}}$ have the same distribution. We then have

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}})\right] = \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\frac{1}{N}\sum_{n=1}^{N}\ell(\hat{\Pi}S^{(n)}, y_n, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}))\right] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\ell(\hat{\Pi}S^{(n)}, y_n, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}))\right] \\ &= \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\ell(\hat{\Pi}S^{(1)}, y_1, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))\right] = \mathop{\mathbb{E}}_{\substack{(S^{(n)},y_n)\\ n=1,\ldots,N}}\left[\ell(\hat{\Pi}S^{(1)}, y_1, \mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))\right],\end{aligned}$$

which we can upper bound by

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{(S^{(n)},y_n)\\ n=1,\dots,N}}\left[\ell(\hat{\Pi}S^{(1)},y_1,\mathcal{L}_{\bar{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))\right]+\mathop{\mathbb{E}}_{\substack{(S^{(n)},y_n)\\ n=1,\dots,N}}\left[|\ell(\hat{\Pi}S^{(1)},y_1,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))-\ell(\hat{\Pi}S^{(1)},y_1,\mathcal{L}_{\bar{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))|\right]\\
&\quad\le \mathop{\mathbb{E}}_{\substack{(S^{(n)},y_n)\\ n=1,\dots,N}}\left[\ell(\hat{\Pi}S^{(1)},y_1,\mathcal{L}_{\bar{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S^{(1)}))\right]+\beta_N\\
&\quad= \mathop{\mathbb{E}}_{\mathcal{D},(S,y)}\left[\ell(\hat{\Pi}S,y,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right]+\beta_N.
\end{aligned}$$

Furthermore, observe that

$$\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\hat{\Pi}S,y,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right]=\mathbb{E}_{(S,y)\sim\tilde{P}}\left[\ell(\hat{\Pi}S,y,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right].$$

We now get

$$\begin{aligned}
II &= \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\ell(\hat{\Pi}S,y,\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S))\right]-\frac{1}{N}\sum_{n=1}^{N}\ell(\hat{\Pi}S^{(n)},y_n,\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}}(\hat{\Pi}S^{(n)}))\right]\\
&= \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathbb{E}_{(S,y)\sim\tilde{P}}\left[\ell(\hat{\Pi}S,y,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right]\right]-\mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathcal{R}_{\ell,\mathcal{D}_{\hat{\Pi}}}(\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}})\right]\\
&\le \mathbb{E}_{\mathcal{D}\sim\tilde{P}^{\otimes N}}\left[\mathbb{E}_{(S,y)\sim\tilde{P}}\left[\ell(\hat{\Pi}S,y,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right]\right]-\mathop{\mathbb{E}}_{\mathcal{D},(S,y)}\left[\ell(\hat{\Pi}S,y,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right]+\beta_N\\
&= \mathbb{E}_{\mathcal{D},(S,y)}\left[\ell(\hat{\Pi}S,y,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))-\ell(\hat{\Pi}S,y,\mathcal{L}_{\mathcal{D}_{\hat{\Pi}}}(\hat{\Pi}S))\right]+\beta_N\\
&= \beta_N.
\end{aligned}$$

Altogether we have

$$\mathbb{E}\left[\mathcal{R}_{\ell,P}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})-\mathcal{R}_{\ell,\hat{\mathcal{D}}_{\hat{\Pi}}}(\mathcal{L}_{\hat{\mathcal{D}}_{\hat{\Pi}}})\right]\le\alpha\left(\mathbb{E}_{(Q,S,y)\sim\bar{P}}\left[\|\Pi Q-\hat{\Pi}S\|_{\mathcal{H}}\right]\right)+\beta_N,$$

and the result follows. □

A.8. Conclusion

We continued the investigation of kernel-based statistical learning with distributional inputs from the perspective of modern statistical learning theory. To the best of our knowledge, we provided the first general oracle inequalities in this setting, complementing the existing excess risk bounds for distributional regression using kernel ridge regression. In particular, our analysis covers rather general loss functions en-

coding a multitude of learning scenarios. Additionally, we provided generalization bounds based on algorithmic stability, a result and setting which has not been analyzed at all in the distributional learning literature. We formulated all of this in a very general setup based on Hilbertian embeddings of probability distributions. On the one hand, in this manner the kernel construction approach from [137] is applicable, and on the other hand, our main results apply directly to any existing and future embedding approach. For example, if appropriate estimation tools become available, our results will be directly applicable to the recently introduced kernel cumulant embeddings [37]. Finally, we provided specializations of our results to the important cases of KMEs as well as the recent sliced 2-Wasserstein distances.

Many relevant questions are still open, and our results form the starting point for a multitude of future investigations. First, while oracle inequalities can be used to derive consistency results, in order to guarantee learning rates, one needs suitable assumptions to derive bounds on the approximation error function. Finding such conditions in the present setting is an important open problem. Second, while the setting of our main results is rather general, we need various boundedness assumptions on the loss functions. Removing these assumptions, or replacing them by clippability (cf. Chapters 2 and 9 in [189]), is another interesting problem. Third, both of our oracle inequalities are based on classic arguments, and it is known, cf. Chapter 7 in [189], that using more advanced tools from empirical process theory, one can derive sharper oracle inequalities, which eventually can lead to better learning rates. We expect that this applies also in the distributional setting, and that the resulting analysis approach for kernel ridge regression from [190] then provides an alternative to the integral operator technique from [48], which so far was the main focus in the distributional regression literature.

A.9. Comments

This chapter is taken mostly verbatim from [CF7]. This work arose from discussions of the author with P.-F. Massiani in the context of a biomedical applications. Most of the theoretical results have been developed by the author, with help from P.-F. Massiani.

Acronyms

BO	Bayesian Optimization
GP	Gaussian Process
HSKM	Hard Shape Constrained Kernel Machine
i.i.d.	Independent and identically distributed
IPS	Interacting Particle System
IQC	Integral Quadratic Constraint
KRR	Kernel Ridge Regression
LFR	Linear Fractional Representation
LMI	Linear Matrix Inequality
LS	Least-Squares
MAS	Multiagent System
MPC	Model Predictive Control
RKHS	Reproducing Kernel Hilbert Space
RL	Reinforcement learning
SVM	Support Vector Machine

List of Mathematical Symbols

$X \sim P$	X is distributed according to P.
$\mathbb{E}[X]$	Expectation of random variable X
$\mathcal{GP}_{\mathcal{X}}(m, k)$	Gaussian process on $\mathcal{X}$ with mean function m, covariance function k
I	Identity matrix. If necessary, I_N makes dimension N explicit.
id	Identity map. If necessary, $\mathrm{id}_{\mathcal{X}}$ makes the underlying set $\mathcal{X}$ explicit.
$\mathcal{Y}^{\mathcal{X}}$	Set of all maps $f : \mathcal{X} \to \mathcal{Y}$
$\mathcal{N}(\mu, \Sigma)$	Normal distribution with mean μ, covariance matrix Σ
$A \preceq B$	Positive semidefinite ordering for self-adjoint matrices or linear operators, i.e., $B - A$ is positive semidefinite
$\mathbb{R}$	Real numbers
$\mathbb{R}_{\geq 0}$	Nonnegative real numbers
$\mathbb{R}_{>0}$	Positive real numbers
$(H_k, \langle \cdot, \cdot \rangle_k)$	RKHS with reproducing kernel k

$\mathrm{Var}[X]$	Variance of random variable X

Bibliography

[CF1] Christian Fiedler. “Lipschitz and Hölder Continuity in Reproducing Kernel Hilbert Spaces”. In: *arXiv preprint arXiv:2310.18078* (2023).

[CF2] Christian Fiedler, Massimo Fornasier, Timo Klock, and Michael Rauchensteiner. “Stable recovery of entangled weights: Towards robust identification of deep neural networks from minimal samples”. In: *Applied and Computational Harmonic Analysis* 62 (2023), pp. 123–172.

[CF3] Christian Fiedler, Michael Herty, Michael Rom, Chiara Segala, and Sebastian Trimpe. “Reproducing kernel Hilbert spaces in the mean field limit”. In: *Kinetic and Related Models* 16.6 (2023), pp. 850–870.

[CF4] Christian Fiedler, Michael Herty, Chiara Segala, and Sebastian Trimpe. “Recent kernel methods for interacting particle systems: first numerical results”. In: *European Journal of Applied Mathematics* 36 (2025), pp. 464 –489.

[CF5] Christian Fiedler, Michael Herty, and Sebastian Trimpe. “Mean field limits for discrete-time dynamical systems via kernel mean embeddings”. In: *IEEE Control Systems Letters* (2023).

[CF6] Christian Fiedler, Michael Herty, and Sebastian Trimpe. “On kernel-based statistical learning theory in the mean field limit”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[CF7] Christian Fiedler, Pierre-François Massiani, Friedrich Solowjow, and Sebastian Trimpe. “On Statistical Learning Theory for Distributional Inputs”. In: *Forty-first International Conference on Machine Learning (ICML)*. 2024.

[CF8] Christian Fiedler, Johanna Menn, Lukas Kreisköther, and Sebastian Trimpe. “On Safety in Safe Bayesian Optimization”. In: *Transactions on Machine Learning Research (TMLR)* (2024).

[CF9] Christian Fiedler, Johanna Menn, and Sebastian Trimpe. “Safety in safe Bayesian optimization and its ramifications for control”. In: *Extended abstract, presented as poster at Symposium on Systems Theory in Data and Optimization (SysDO)* (2024).

[CF10] Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. “Learning-enhanced robust controller synthesis with rigorous statistical and control-theoretic guarantees”. In: *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE. 2021, pp. 5122–5129.

[CF11] Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. “Learning Functions and Uncertainty Sets Using Geometrically Constrained Kernel Regression”. In: *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE. 2022, pp. 2141–2146.

[CF12] Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. “Practical and rigorous uncertainty bounds for Gaussian process regression”. In: *Proceedings of the AAAI Conference on artificial intelligence*. Vol. 35. 8. 2021, pp. 7439–7447.

[CF13] Christian Fiedler, Carsten W. Scherer, and Sebastian Trimpe. “Practical and Rigorous Uncertainty Bounds for Gaussian Process Regression”. In: *arXiv preprint arXiv:2105.02796v2* (2023).

[CF14] Christian Fiedler and Sebastian Trimpe. “A panoramic introduction to reproducing kernel Hilbert spaces”. In: *Manuscript in preparation* (2024).

[CF15] Christian Fiedler and Sebastian Trimpe. “Analysis of EMPC schemes without terminal constraints via local incremental stabilizability”. In: *2024 European Control Conference (ECC)*. IEEE. 2024, pp. 1830–1836.

[CF16] Christian Fiedler and Sebastian Trimpe. “Revisiting the derivation of stage costs in infinite horizon discrete-time optimal control”. In: *2022 30th Mediterranean Conference on Control and Automation (MED)*. IEEE. 2022, pp. 211–216.

[CF17] Christian Fiedler, Sebastian Trimpe, and Michael Herty. “Reproducing Kernels in and for the Mean Field Limit”. In: *Extended abstract and oral presentation, DEEPK2024, Leuven, Belgium* (2024).

[CF18] Friedrich Solowjow, Dominik Baumann, Christian Fiedler, Andreas Jocham, Thomas Seel, and Sebastian Trimpe. "A kernel two-sample test for dynamical systems". In: *arXiv preprint arXiv:2004.11098* (2020).

[CF19] Abdullah Tokmak, Christian Fiedler, Melanie N Zeilinger, Sebastian Trimpe, and Johannes Köhler. "Automatic nonlinear MPC approximation with closed-loop guarantees". In: *arXiv preprint arXiv:2312.10199. Accepted for publication in IEEE Transactions on Automatic Control* (2023).

[1] Yasin Abbasi-Yadkori. "Online learning for linearly parametrized control problems". PhD thesis. University of Alberta, 2012.

[2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits". In: *Advances in neural information processing systems* 24 (2011).

[3] Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer Science & Business Media, 2009.

[4] Christian Agrell. "Gaussian Processes with Linear Operator Inequality Constraints". In: *Journal of Machine Learning Research* 20 (2019), pp. 1–36.

[5] Amir Ali Ahmadi and Bachir El Khadir. "Learning dynamical systems with side information". In: *Learning for Dynamics and Control*. PMLR. 2020, pp. 718–727.

[6] Giacomo Albi, Mattia Bongini, Emiliano Cristiani, and Dante Kalise. "Invisible control of self-organizing agents leaving unknown environments". In: *SIAM Journal on Applied Mathematics* 76.4 (2016), pp. 1683–1710.

[7] Syed Twareque Ali, Jean-Pierre Antoine, and Jean-Pierre Gazeau. *Coherent states, wavelets and their generalizations*. Vol. 3. Springer, 2000.

[8] Carmen Amo Alonso, Jerome Sieber, and Melanie N Zeilinger. "State space models as foundation models: A control theoretic overview". In: *arXiv preprint arXiv:2403.16899* (2024).

[9] Daniel Alpay and Palle ET Jorgensen. "New characterizations of reproducing kernel Hilbert spaces and applications to metric geometry". In: *Opuscula Mathematica* 41.3 (2021).

[10] Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. "Kernels for vector-valued functions: A review". In: *Foundations and Trends® in Machine Learning* 4.3 (2012), pp. 195–266.

[11] Herbert Amann. *Ordinary differential equations: an introduction to nonlinear analysis*. Vol. 13. Walter de gruyter, 2011.

[12] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[13] MOSEK ApS. *MOSEK Fusion API for Python 9.3.18*. 202. URL: `https://docs.mosek.com/latest/pythonfusion/index.html`.

[14] Nachman Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[15] Karl Johan Åström and Richard Murray. *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2021.

[16] Karl Johan Åström and Richard M Murray. *Feedback systems: an introduction for scientists and engineers*. Princeton university press, 2021.

[17] Marc Atteia. *Hilbertian kernels and spline functions*. Elsevier, 1992.

[18] Pierre-Cyril Aubin-Frankowski and Zoltan Szabo. "Handling Hard Affine SDP Shape Constraints in RKHSs". In: *arXiv preprint arXiv:2101.01519* (2021).

[19] Pierre-Cyril Aubin-Frankowski and Zoltán Szabó. "Hard Shape-Constrained Kernel Machines". In: *Advances in Neural Information Processing Systems (NeurIPS-2020)*. 2020.

[20] Pierre-Cyril Aubin-Frankowski and Zoltán Szabó. "Hard shape-constrained kernel machines". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 384–395.

[21] Michele Ballerini, Nicola Cabibbo, Raphael Candelier, Andrea Cavagna, Evaristo Cisbani, Irene Giardina, Vivien Lecomte, Alberto Orlandi, Giorgio Parisi, Andrea Procaccini, et al. "Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from

a field study". In: *Proceedings of the national academy of sciences* 105.4 (2008), pp. 1232–1237.

[22] Dominik Baumann, Alonso Marco, Matteo Turchetta, and Sebastian Trimpe. "Gosafe: Globally optimal safe robot learning". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 4452–4458.

[23] Nicola Bellomo, Pierre Degond, and Eitan Tadmor. *Active Particles, Volume 1: Advances in Theory, Models, and Applications*. Birkhäuser, 2017.

[24] Nicola Bellomo, Pierre Degond, and Eitan Tadmor. "Active Particles, Volume 2". In: (2019).

[25] Julian Berberich, Anne Koch, Carsten W Scherer, and Frank Allgöwer. "Robust data-driven state-feedback design". In: *2020 American Control Conference (ACC)*. IEEE. 2020, pp. 1532–1538.

[26] Julian Berberich, Anne Koch, Carsten W Scherer, and Frank Allgöwer. "Robust data-driven state-feedback design". In: *2020 American Control Conference (ACC)*. IEEE. 2020, pp. 1532–1538.

[27] Julian Berberich, Carsten W Scherer, and Frank Allgöwer. "Combining prior knowledge and data for robust controller design". In: *arXiv preprint arXiv:2009.05253* (2020).

[28] Felix Berkenkamp and Angela P Schoellig. "Learning-based robust control: Guaranteeing stability while improving performance". In: *IEEE/RSJ Proceedings of International Conference on Intelligent Robots and Systems (IROS)*. 2014.

[29] Felix Berkenkamp and Angela P Schoellig. "Safe and robust learning control with Gaussian processes". In: *2015 European Control Conference (ECC)*. IEEE. 2015, pp. 2496–2501.

[30] Felix Berkenkamp and Angela P Schoellig. "Safe and robust learning control with Gaussian processes". In: *2015 European Control Conference (ECC)*. IEEE. 2015, pp. 2496–2501.

[31] Felix Berkenkamp, Angela P Schoellig, and Andreas Krause. "Safe controller optimization for quadrotors with Gaussian processes". In: *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2016, pp. 491–496.

[32] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2004.

[33] Adrien Blanchet and Guillaume Carlier. "From Nash to Cournot–Nash equilibria via the Monge–Kantorovich problem". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372.2028 (2014), p. 20130398.

[34] Mattia Bongini. "Sparse optimal control of multiagent systems". PhD thesis. Technische Universität München, 2016.

[35] Mattia Bongini, Massimo Fornasier, Markus Hansen, and Mauro Maggioni. "Inferring interaction rules from observations of evolutive systems I: The variational approach". In: *Mathematical Models and Methods in Applied Sciences* 27.05 (2017), pp. 909–951.

[36] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. "Sliced and Radon Wasserstein barycenters of measures". In: *Journal of Mathematical Imaging and Vision* 51 (2015), pp. 22–45.

[37] Patric Bonnier, Harald Oberhauser, and Zoltán Szabó. "Kernelized Cumulants: Beyond Kernel Mean Embeddings". In: *Advances in Neural Information Processing Systems* (2023).

[38] Olivier Bousquet and André Elisseeff. "Stability and generalization". In: *The Journal of Machine Learning Research* 2 (2002), pp. 499–526.

[39] Poompol Buathong, David Ginsbourger, and Tipaluck Krityakierne. "Kernels over sets of finite sets using RKHS embeddings, with application to Bayesian (combinatorial) optimization". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2731–2741.

[40] Mona Buisson-Fenet, Friedrich Solowjow, and Sebastian Trimpe. "Actively learning gaussian process dynamics". In: *Learning for dynamics and control*. PMLR. 2020, pp. 5–15.

[41] Francesco Bullo. *Lectures on network systems*. Kindle Direct Publishing, 2022.

[42] Giuseppe C Calafiore and Marco C Campi. "The scenario approach to robust control design". In: *IEEE Transactions on automatic control* 51.5 (2006), pp. 742–753.

[43] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. "Gaussian process optimization with adaptive sketching: Scalable and no regret". In: *Conference on Learning Theory*. PMLR. 2019, pp. 533–557.

[44] J Calliess. "Conservative decision-making and inference in uncertain dynamical systems". PhD thesis. Oxford University, UK, 2014.

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

[46] Marco C Campi and Simone Garatti. *Introduction to the scenario approach*. SIAM, 2018.

[47] Alexandre Capone, Armin Lederer, and Sandra Hirche. "Gaussian process uniform error bounds with unknown hyperparameters for safety-critical applications". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 2609–2624.

[48] Andrea Caponnetto and Ernesto De Vito. "Optimal rates for the regularized least-squares algorithm". In: *Foundations of Computational Mathematics* 7 (2007), pp. 331–368.

[49] Pierre Cardaliaguet. *Notes on mean field games*. Tech. rep. 2010.

[50] René Carmona and François Delarue. *Probabilistic theory of mean field games with applications I-II*. Springer, 2018.

[51] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. "Statistical physics of social dynamics". In: *Reviews of modern physics* 81.2 (2009), p. 591.

[52] Carlo Cercignani. *Rarefied gas dynamics: from basic concepts to actual calculations*. Vol. 21. Cambridge Texts in Applied Mathematics. Cambridge university press, 2000.

[53] Young-Pil Choi, Dante Kalise, Jan Peszek, and Andrés A Peters. "A collisionless singular Cucker–Smale model with decentralized formation control". In: *SIAM Journal on Applied Dynamical Systems* 18.4 (2019), pp. 1954–1981.

[54] Sayak Ray Chowdhury and Aditya Gopalan. "On kernelized multi-armed bandits". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 844–853.

[55] Andreas Christmann and Ingo Steinwart. "Universal kernels on non-standard input spaces". In: *Advances in neural information processing systems* 23 (2010).

[56] Ştefan Cobzaş, Radu Miculescu, Adriana Nicolae, et al. *Lipschitz functions*. Springer, 2019.

[57] Emiliano Cristiani, Benedetto Piccoli, and Andrea Tosin. *Multiscale modeling of pedestrian dynamics*. Vol. 12. Modeling, Simulation and Applications. Springer, 2014.

[58] Felipe Cucker and Steve Smale. "Emergent behavior in flocks". In: *IEEE Transactions on automatic control* 52.5 (2007), pp. 852–862.

[59] Giuseppe Da Prato. *An introduction to infinite-dimensional analysis*. 2nd ed. Springer Science & Business Media, 2006.

[60] Gianni Dal Maso. *An introduction to* Γ*-convergence*. Vol. 8. Springer Science & Business Media, 2012.

[61] Alex Devonport, He Yin, and Murat Arcak. "Bayesian safe learning and control with sum-of-squares analysis and polynomial kernels". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE. 2020, pp. 3159–3165.

[62] Steven Diamond and Stephen Boyd. "CVXPY: A Python-embedded modeling language for convex optimization". In: *Journal of Machine Learning Research* 17.83 (2016), pp. 1–5.

[63] Devdatt P Dubhashi and Alessandro Panconesi. *Concentration of measure for the analysis of randomized algorithms*. Cambridge University Press, 2009.

[64] Richard M Dudley. *Real analysis and probability*. Cambridge University Press, 2002.

[65] Geir E Dullerud and Fernando Paganini. *A course in robust control theory: a convex approach*. Vol. 36. Springer Science & Business Media, 2000.

[66] Geir E Dullerud and Fernando Paganini. *A course in robust control theory: a convex approach*. Vol. 36. Springer Science & Business Media, 2013.

[67] Audrey Durand, Odalric-Ambrym Maillard, and Joelle Pineau. "Streaming kernel regression with provably adaptive mean, variance, and regularization". In: *Journal of Machine Learning Research* 19.17 (2018), pp. 1–34.

[68] John RG Dyer, Anders Johansson, Dirk Helbing, Iain D Couzin, and Jens Krause. "Leadership, consensus decision making and collective behaviour in humans". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 364.1518 (2009), pp. 781–789.

[69] Lawrence C Evans. *Partial differential equations*. Vol. 19. American Mathematical Society, 2022.

[70] Zhiying Fang, Zheng-Chu Guo, and Ding-Xuan Zhou. "Optimal learning rates for distribution regression". In: *Journal of complexity* 56 (2020), p. 101426.

[71] Gregory E Fasshauer and Michael J McCourt. *Kernel-based approximation methods using Matlab*. Vol. 19. World Scientific Publishing Company, 2015.

[72] JC Ferreira and Valdir Antônio Menegatto. "Positive definiteness, reproducing kernel Hilbert spaces and beyond". In: *Annals of Functional Analysis* 4.1 (2013).

[73] Simon Fischer and Ingo Steinwart. "Sobolev norm learning rates for regularized least-squares algorithms". In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 8464–8501.

[74] Massimo Fornasier, Stefano Lisini, Carlo Orrieri, and Giuseppe Savaré. "Mean-field optimal control as Gamma-limit of finite agent controls". In: *European Journal of Applied Mathematics* 30.6 (2019), pp. 1153–1186.

[75] Massimo Fornasier and Francesco Solombrino. “Mean-field optimal control”. In: *ESAIM: Control, Optimisation and Calculus of Variations* 20.4 (2014), pp. 1123–1152.

[76] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, 2013.

[77] Antonio G Garcia. “Orthogonal sampling formulas: a unified approach”. In: *SIAM review* 42.3 (2000), pp. 499–512.

[78] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alexander J Smola. “Multi-instance kernels”. In: *ICML*. Vol. 2. 3. 2002, p. 7.

[79] Andreas Geist and Sebastian Trimpe. “Learning constrained dynamics with Gauss’ principle adhering Gaussian processes”. In: *Learning for Dynamics and Control*. PMLR. 2020, pp. 225–234.

[80] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*. Vol. 44. Cambridge University Press, 2017.

[81] François Golse. “On the dynamics of large particle systems in the mean field limit”. In: *Macroscopic and large scale phenomena: coarse graining, mean field limits and ergodicity* (2016), pp. 1–144.

[82] Xiaoqian Gong, Michael Herty, Benedetto Piccoli, and Giuseppe Visconti. “Crowd Dynamics: Modeling and Control of Multiagent Systems”. In: *Annual Review of Control, Robotics, and Autonomous Systems* 6 (2022).

[83] Robert B Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.

[84] Dennis Gramlich, Christian Ebenbauer, and Carsten W Scherer. “Synthesis of accelerated gradient algorithms for optimization and saddle point problems using Lyapunov functions and LMIs”. In: *Systems & Control Letters* 165 (2022), p. 105271.

[85] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.

[86] Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*. Vol. 38. Cambridge University Press, 2014.

[87] Lars Grüne and Jürgen Pannek. *Nonlinear model predictive control*. 2nd ed. Springer, 2017.

[88] Bernard Haasdonk and Claus Bahlmann. "Learning with distance substitution kernels". In: *Joint pattern recognition symposium*. Springer. 2004, pp. 220–227.

[89] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. 2009.

[90] Mohamed K Helwa, Adam Heins, and Angela P Schoellig. "Provably robust learning-based approach for high-accuracy tracking control of Lagrangian systems". In: *IEEE Robotics and Automation Letters* 4.2 (2019), pp. 1587–1594.

[91] Michael Herty and Mattia Zanella. "Performance bounds for the mean-field limit of constrained dynamics". In: *Discrete and Continuous Dynamical Systems* 37.4 (2017), p. 2023.

[92] Lukas Hewing, Kim P Wabersich, Marcel Menner, and Melanie N Zeilinger. "Learning-based model predictive control: Toward safe learning in control". In: *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020), pp. 269–296.

[93] Zhong-Sheng Hou and Zhuo Wang. "From model-based control to data-driven control: Survey, classification and perspective". In: *Information Sciences* 235 (2013), pp. 3–35.

[94] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. "Time-uniform Chernoff bounds via nonnegative supermartingales". In: (2020).

[95] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. "Time-uniform, nonparametric, nonasymptotic confidence sequences". In: (2021).

[96] Daniel Hsu, Sham Kakade, and Tong Zhang. "A tail inequality for quadratic forms of subgaussian random vectors". In: (2012).

[97] Petros Ioannou and Bariş Fidan. *Adaptive control tutorial*. SIAM, 2006.

[98] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". In: *Advances in neural information processing systems* 31 (2018).

[99] Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

[100] Carl Jidling, Niklas Wahlstrom, Adrian Wills, and Thomas B Schön. "Linearly constrained Gaussian processes". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 1215–1224.

[101] Carl Jidling, Niklas Wahlström, Adrian Wills, and Thomas B Schön. "Linearly constrained Gaussian processes". In: *Advances in neural information processing systems* 30 (2017).

[102] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. "Gaussian processes and kernel methods: A review on connections and equivalences". In: *arXiv preprint arXiv:1807.02582* (2018).

[103] Toni Karvonen, George Wynne, Filip Tronarp, Chris Oates, and Simo Sarkka. "Maximum likelihood estimation and uncertainty quantification for Gaussian process approximation of deterministic functions". In: *SIAM/ASA Journal on Uncertainty Quantification* 8.3 (2020), pp. 926–958.

[104] Yael Katz, Kolbjørn Tunstrøm, Christos C Ioannou, Cristián Huepe, and Iain D Couzin. "Inferring the structure and dynamics of interactions in schooling fish". In: *Proceedings of the National Academy of Sciences* 108.46 (2011), pp. 18720–18725.

[105] Christopher M Kellett. "A compendium of comparison function results". In: *Mathematics of Control, Signals, and Systems* 26 (2014), pp. 339–374.

[106] Hassan K Khalil. *Nonlinear systems*. Vol. 3. Prentice hall Upper Saddle River, NJ, 2002.

[107] Mohammad Khosravi and Roy S Smith. "Nonlinear system identification with prior knowledge on the region of attraction". In: *IEEE Control Systems Letters* 5.3 (2020), pp. 1091–1096.

[108] Jungtaek Kim, Michael McCourt, Tackgeun You, Saehoon Kim, and Seungjin Choi. "Bayesian optimization with approximate set kernels". In: *Machine Learning* 110.5 (2021), pp. 857–879.

[109] Achim Klenke. *Probability theory: a comprehensive course*. 2nd ed. Springer Science & Business Media, 2014.

[110] Anne Koch, Julian Berberich, and Frank Allgöwer. "Verifying dissipativity properties from noise-corrupted input-state data". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE. 2020, pp. 616–621.

[111] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. "Learning-based model predictive control for safe exploration". In: *2018 IEEE conference on decision and control (CDC)*. IEEE. 2018, pp. 6059–6066.

[112] Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. "Learning-based model predictive control for safe exploration". In: *2018 IEEE Conference on Decision and Control (CDC)*. IEEE. 2018, pp. 6059–6066.

[113] Hayri Korezlioglu. "Reproducing kernels in separable Hilbert spaces". In: *Pacific Journal of Mathematics* 25.2 (1968), pp. 305–314.

[114] Serge Lang. *Real and functional analysis*. Vol. 142. Springer Science & Business Media, 2012.

[115] Markus Lange-Hegermann. "Algorithmic linearly constrained Gaussian processes". In: *Advances in Neural Information Processing Systems* 31 (2018).

[116] Jean-François Le Gall. *Brownian motion, martingales, and stochastic calculus*. Springer, 2016.

[117] Armin Lederer, Jonas Umlauft, and Sandra Hirche. "Uniform error bounds for Gaussian process regression with application to safe control". In: *Advances in Neural Information Processing Systems* 32 (2019).

[118] Olli Lehto. *Some remarks on the kernel function in Hilbert function space*. Suomalainen tiedeakatemia, 1952.

[119] Tianyi Lin, Zeyu Zheng, Elynn Chen, Marco Cuturi, and Michael I Jordan. "On projection robust optimal transport: Sample complexity and model misspecification". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 262–270.

[120] Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. "Random features for kernel approximation: A survey on algorithms, theory, and beyond". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2021), pp. 7128–7148.

[121] Lennart Ljung. *System identification: Theory for the user*. 2nd ed. Pearson, 1998.

[122] Lennart Ljung. *System Identification: Theory for the User*. Pearson Education, 1998.

[123] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. "Towards a learning theory of cause-effect inference". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1452–1461.

[124] Fei Lu, Ming Zhong, Sui Tang, and Mauro Maggioni. "Nonparametric inference of interaction laws in systems of agents from trajectory data". In: *Proceedings of the National Academy of Sciences* 116.29 (2019), pp. 14424–14433.

[125] Milan Lukić and Jay Beder. "Stochastic processes with sample paths in reproducing kernel Hilbert spaces". In: *Transactions of the American Mathematical Society* 353.10 (2001), pp. 3945–3969.

[126] Emilio T Maddalena and Colin N Jones. "Learning non-parametric models with guarantees: A smooth Lipschitz regression approach". In: *IFAC-PapersOnLine* 53.2 (2020), pp. 965–970.

[127] Emilio T Maddalena, Paul Scharnhorst, and Colin N Jones. "Deterministic error bounds for kernel-based learning techniques under bounded noise". In: *arXiv preprint arXiv:2008.04005* (2020).

[128] Odalric-Ambrym Maillard. "Self-normalization techniques for streaming confident regression". In: *Preprint* (2016). URL: `https://hal.science/hal-01349727/document`.

[129] Horia Mania, Michael I Jordan, and Benjamin Recht. "Active learning for nonlinear system identification with guarantees". In: *Journal of Machine Learning Research* 23.32 (2022), pp. 1–30.

[130] Jose Maria Manzano, Daniel Limon, David Muñoz de la Peña, and Jan-Peter Calliess. "Output feedback MPC based on smoothed projected kinky inference". In: *IET Control Theory & Applications* 13.6 (2019), pp. 795–805.

[131] Ivan Markovsky. *Low rank approximation: algorithms, implementation, applications*. 2nd ed. Springer, 2019.

[132] Ivan Markovsky, Linbin Huang, and Florian Dörfler. "Data-driven control based on the behavioral approach: From theory to applications in power systems". In: *IEEE Control Systems Magazine* 43.5 (2023), pp. 28–68.

[133] Alexandre Mauroy, Y Susuki, and Igor Mezic. *Koopman operator in systems and control*. Springer, 2020.

[134] Edward James McShane. "Extension of range of functions". In: *Bulletin of the American Mathematical Association* (1934).

[135] Giacomo Meanti, Luigi Carratino, Lorenzo Rosasco, and Alessandro Rudi. "Kernel methods through the roof: handling billions of points efficiently". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 14410–14422.

[136] Alexandre Megretski and Anders Rantzer. "System analysis via integral quadratic constraints". In: *IEEE Transactions on Automatic Control* 42.6 (1997), pp. 819–830.

[137] Dimitri Meunier, Massimiliano Pontil, and Carlo Ciliberto. "Distribution regression with sliced Wasserstein kernels". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 15501–15523.

[138] Charles A Micchelli and Massimiliano Pontil. "On learning vector-valued functions". In: *Neural computation* 17.1 (2005), pp. 177–204.

[139] Mario Milanese and Carlo Novara. "Set membership identification of nonlinear systems". In: *Automatica* 40.6 (2004), pp. 957–975.

[140] Mario Milanese and Carlo Novara. “Set membership identification of nonlinear systems”. In: *Automatica* 40.6 (2004), pp. 957–975.

[141] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. 2nd ed. MIT press, 2018.

[142] Mattes Mollenhauer and Claudia Schillings. “On the concentration of subgaussian vectors and positive quadratic forms in Hilbert spaces”. In: *arXiv preprint arXiv:2306.11404* (2023).

[143] Michael Muehlebach and Michael I Jordan. “Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives”. In: *Journal of Machine Learning Research* 22.73 (2021), pp. 1–50.

[144] Kevin P Murphy. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

[145] Mojmir Mutny and Andreas Krause. “Experimental design for linear functionals in reproducing kernel hilbert spaces”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 20175–20188.

[146] Giovanni Naldi, Lorenzo Pareschi, and Giuseppe Toscani. *Mathematical modeling of collective behavior in socio-economic and life sciences*. Springer Science & Business Media, 2010.

[147] Sloan Nietert, Ziv Goldfeld, Ritwik Sadhu, and Kengo Kato. “Statistical, robustness, and computational guarantees for sliced Wasserstein distances”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 28179–28193.

[148] Kwang-Kyo Oh, Myoung-Chul Park, and Hyo-Sung Ahn. “A survey of multi-agent formation control”. In: *Automatica* 53 (2015), pp. 424–440.

[149] Baver Okutmuştur. “Reproducing kernel Hilbert spaces”. PhD thesis. Bilkent Universitesi (Turkey), 2005.

[150] Houman Owhadi and Clint Scovel. “Separability of reproducing kernel spaces”. In: *Proceedings of the American Mathematical Society* 145.5 (2017), pp. 2131–2138.

[151] Lorenzo Pareschi and Giuseppe Toscani. *Interacting multiagent systems: kinetic equations and Monte Carlo methods*. OUP Oxford, 2013.

[152] Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*. Vol. 152. Cambridge university press, 2016.

[153] George Pedrick. "Theory of reproducing kernels for Hilbert spaces of vector valued functions". PhD thesis. University of Kansas, 1957.

[154] Victor H de la Pena. "From decoupling and self-normalization to machine learning". In: *Notices of the American Mathematical Society* 66.10 (2019).

[155] Victor H de la Pena, Michael J Klass, and Tze Leung Lai. "Theory and applications of multivariate self-normalized processes". In: *Stochastic Processes and their Applications* 119.12 (2009), pp. 4210–4227.

[156] Victor H de la Pena, Michael J Klass, and Tze Leung Lai. "Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws". In: (2004).

[157] Victor H de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer, 2009.

[158] Andrés A Peters, Richard H Middleton, and Oliver Mason. "Leader tracking in homogeneous vehicle platoons with broadcast delays". In: *Automatica* 50.1 (2014), pp. 64–74.

[159] Gianluigi Pillonetto, Tianshi Chen, Alessandro Chiuso, Giuseppe De Nicolao, and Lennart Ljung. *Regularized system identification: Learning dynamic models from data*. Springer Nature, 2022.

[160] Gianluigi Pillonetto, Francesco Dinuzzo, Tianshi Chen, Giuseppe De Nicolao, and Lennart Ljung. "Kernel methods in system identification, machine learning and function estimation: A survey". In: *Automatica* 50.3 (2014), pp. 657–682.

[161] János D Pintér. *Global optimization in action: continuous and Lipschitz optimization: algorithms, implementations and applications*. Vol. 6. Springer Science & Business Media, 1995.

[162] Barnabás Póczos, Aarti Singh, Alessandro Rinaldo, and Larry Wasserman. "Distribution-free distribution regression". In: *artificial intelligence and statistics*. PMLR. 2013, pp. 507–515.

[163] Joaquin Quinonero-Candela and Carl Edward Rasmussen. “A unifying view of sparse approximate Gaussian process regression”. In: *The Journal of Machine Learning Research* 6 (2005), pp. 1939–1959.

[164] Ali Rahimi and Benjamin Recht. “Random features for large-scale kernel machines”. In: *Advances in neural information processing systems* 20 (2007).

[165] Anders Rantzer. “On the kalman-yakubovich-popov lemma”. In: *Systems & control letters* 28.1 (1996), pp. 7–10.

[166] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[167] James Blake Rawlings, David Q Mayne, and Moritz Diehl. *Model predictive control: theory, computation, and design*. Vol. 2. Nob Hill Publishing Madison, WI, 2017.

[168] Alexander von Rohr, Dmitrii Likhachev, and Sebastian Trimpe. “Robust direct data-driven control for probabilistic systems”. In: *Systems & Control Letters* 196 (2025), p. 106011.

[169] Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. “Less is more: Nyström computational regularization”. In: *Advances in neural information processing systems* 28 (2015).

[170] Alessandro Rudi, Luigi Carratino, and Lorenzo Rosasco. “Falkon: An optimal large scale kernel method”. In: *Advances in neural information processing systems* 30 (2017).

[171] Stuart Russell and Peter Norvig. *Artificial Intelligence. A Modern Approach*. 4th ed. Pearson, 2021.

[172] Michael G Safonov and Tung-Ching Tsao. “The unfalsified control concept: A direct path from experiment to controller”. In: *Feedback Control, Nonlinear Systems, and Complexity*. Springer. 1995, pp. 196–214.

[173] Saburou Saitoh. “Hilbert spaces induced by Hilbert space valued functions”. In: *Proceedings of the American Mathematical Society* 89.1 (1983), pp. 74–78.

[174] Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. 2nd ed. Springer, 2018.

[175] Robert Schaback. *Kernel-based meshless methods*. 2011.

[176] Paul Scharnhorst, Emilio T Maddalena, Yuning Jiang, and Colin N Jones. "Robust uncertainty bounds in reproducing kernel hilbert spaces: A convex optimization approach". In: *IEEE Transactions on Automatic Control* 68.5 (2022), pp. 2848–2861.

[177] Carsten Scherer. "Theory of robust control". In: *Delft University of Technology* (2001), pp. 1–160.

[178] Carsten Scherer and Christian Ebenbauer. "Convex synthesis of accelerated gradient algorithms". In: *SIAM Journal on Control and Optimization* 59.6 (2021), pp. 4615–4645.

[179] Carsten Scherer and Siep Weiland. "Linear matrix inequalities in control". In: *Lecture Notes, Dutch Institute for Systems and Control, Delft, The Netherlands* 3.2 (2000).

[180] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. "A generalized representer theorem". In: *International conference on computational learning theory*. Springer. 2001, pp. 416–426.

[181] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2018.

[182] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[183] Carl-Johann Simon-Gabriel and Bernhard Schölkopf. "Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 1708–1736.

[184] Edward Snelson and Zoubin Ghahramani. "Sparse Gaussian processes using pseudo-inputs". In: *Advances in neural information processing systems* 18 (2005).

[185] Raffaele Soloperto, Matthias A Müller, Sebastian Trimpe, and Frank Allgöwer. "Learning-based robust model predictive control with state-dependent uncertainty". In: *IFAC-PapersOnLine* 51.20 (2018), pp. 442–447.

[186] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. "Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design". In: *Proceedings of the 27th International Conference on Machine Learning*. Omnipress. 2010, pp. 1015–1022.

[187] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures." In: *Journal of Machine Learning Research* 12.7 (2011).

[188] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. "Hilbert space embeddings and metrics on probability measures". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1517–1561.

[189] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[190] Ingo Steinwart, Don R Hush, Clint Scovel, et al. "Optimal Rates for Regularized Least Squares Regression." In: *COLT*. 2009, pp. 79–93.

[191] Ingo Steinwart and Clint Scovel. "Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs". In: *Constructive Approximation* 35 (2012), pp. 363–417.

[192] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. "Safe exploration for optimization with Gaussian processes". In: *International conference on machine learning*. PMLR. 2015, pp. 997–1005.

[193] Richard S. Sutton and Andrew Barto. *Reinforcement learning: An introduction*. 2nd ed. MIT press, 2018.

[194] Botond Szabó, Aad W Van Der Vaart, and JH Van Zanten. "Frequentist coverage of adaptive nonparametric Bayesian credible sets". In: (2015).

[195] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath Sriperumbudur. "Two-stage sampled learning theory on distributions". In: *Artificial Intelligence and Statistics*. PMLR. 2015, pp. 948–957.

[196] Zoltán Szabó, Bharath K Sriperumbudur, Barnabás Póczos, and Arthur Gretton. "Learning theory for distribution regression". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 5272–5311.

[197] Aretha L Teckentrup. "Convergence of Gaussian process regression with estimated hyper-parameters and applications in Bayesian inverse problems". In: *SIAM/ASA Journal on Uncertainty Quantification* 8.4 (2020), pp. 1310–1337.

[198] Roberto Tempo, Giuseppe Calafiore, and Fabrizio Dabbene. *Randomized algorithms for analysis and control of uncertain systems: with applications*. Springer Science & Business Media, 2012.

[199] Roberto Tempo, Giuseppe Calafiore, Fabrizio Dabbene, et al. *Randomized algorithms for analysis and control of uncertain systems: with applications*. Vol. 7. Springer, 2013.

[200] Giuseppe Toscani. "Kinetic models of opinion formation". In: *Communications in mathematical sciences* 4.3 (2006), pp. 481–496.

[201] Andrea Tosin and Mattia Zanella. "Kinetic-controlled hydrodynamics for traffic models with driver-assist vehicles". In: *Multiscale Modeling & Simulation* 17.2 (2019), pp. 716–749.

[202] Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. "Statistical learning theory for control: A finite-sample perspective". In: *IEEE Control Systems Magazine* 43.6 (2023), pp. 67–97.

[203] Alexandre B Tsybakov and Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[204] Jonas Umlauft, Lukas Pöhler, and Sandra Hirche. "An uncertainty-based control Lyapunov approach for control-affine systems modeled by Gaussian process". In: *IEEE Control Systems Letters* 2.3 (2018), pp. 483–488.

[205] Sattar Vakili, Nacime Bouziani, Sepehr Jalali, Alberto Bernacchia, and Dashan Shiu. "Optimal order simple regret for Gaussian process bandits". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 21202–21215.

[206] Joost Veenman and Carsten W Scherer. “A synthesis framework for robust gain-scheduling controllers”. In: *Automatica* 50.11 (2014), pp. 2799–2812.

[207] Joost Veenman and Carsten W Scherer. “IQC-synthesis with general dynamic multipliers”. In: *International Journal of Robust and Nonlinear Control* 24.17 (2014), pp. 3027–3056.

[208] Joost Veenman, Carsten W Scherer, Carlos Ardura, Samir Bennani, Valentin Preda, and Bénédicte Girouart. “IQClab: A new IQC based toolbox for robustness analysis and control design”. In: *IFAC-PapersOnLine* 54.8 (2021), pp. 69–74.

[209] Joost Veenman, Carsten W Scherer, and Hakan Köroğlu. “Robust stability and performance analysis based on integral quadratic constraints”. In: *European Journal of Control* 31 (2016), pp. 1–32.

[210] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press, 2018.

[211] Martin Vetterli, Jelena Kovačević, and Vivek K Goyal. *Foundations of signal processing*. Cambridge University Press, 2014.

[212] Tamás Vicsek and Anna Zafeiris. “Collective motion”. In: *Physics reports* 517.3-4 (2012), pp. 71–140.

[213] Curtis R Vogel. *Computational methods for inverse problems*. SIAM, 2002.

[214] Alexander von Rohr, Matthias Neumann-Brosig, and Sebastian Trimpe. “Probabilistic robust linear quadratic regulators with Gaussian processes”. In: *3rd Annual Learning for Dynamics and Control Conference*. PMLR. 2021 (forthcoming).

[215] Wenjia Wang, Rui Tuo, and CF Jeff Wu. “On prediction properties of kriging: Uniform error bounds and robustness”. In: *Journal of the American Statistical Association* 115.530 (2020), pp. 920–930.

[216] Joachim Weidmann. *Lineare Operatoren in Hilberträumen: Teil 1 Grundlagen*. Springer-Verlag, 2000.

[217] Holger Wendland. *Scattered data approximation*. Vol. 17. Cambridge university press, 2004.

[218] Holger Wendland. *Scattered data approximation*. Vol. 17. Cambridge university press, 2004.

[219] Justin Whitehouse, Aaditya Ramdas, and Steven Z Wu. "On the sublinear regret of GP-UCB". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 35266–35276.

[220] Lingfei Wu, Ian En-Hsu Yen, Fangli Xu, Pradeep Ravikumar, and Michael Witbrock. "D2ke: From distance to kernel and embedding". In: *arXiv preprint arXiv:1802.04956* (2018).

[221] Xiuxia Yin, Zhiwei Gao, Dong Yue, and Yichuan Fu. "Convergence of velocities for the short range communicated discrete-time Cucker–Smale model". In: *Automatica* 129 (2021), p. 109659.

[222] Ding-Xuan Zhou. "Derivative reproducing properties for kernel methods in learning theory". In: *Journal of computational and Applied Mathematics* 220.1-2 (2008), pp. 456–463.

[223] Kemin Zhou, John Comstock Doyle, Keith Glover, et al. *Robust and optimal control*. Vol. 40. Prentice hall New Jersey, 1996.