Università degli Studi di Genova

Dipartimento di Matematica

Corso di studi in
Matematica

Anno accademico 2023/2024

Tesi di Laurea Magistrale

Learning Theory for Dynamical Systems

Candidato
Leonardo De Scalzi

Relatore
Prof. Lorenzo Rosasco

Relatore
Prof.ssa Silvia Villa

Relatore
Oleksii Kachaiev

Febbraio 2025

Contents

"The poetic aspect of the story is that there are many ways of talking about the natural world. As long as those ways latch on to something real and causally efficacious about the functioning of the world, then we attribute some reality and truth to them."

– Sean Carroll

Introduction

Machine Learning Theory At its core, science aims to construct coherent, testable representations of observed phenomena, models that not only predict accurately but also offer deep insights into underlying processes. In the realm of Artificial Intelligence, Machine Learning develops algorithms that leverage data to build such models, capturing complex patterns for prediction, classification, decision-making, and artificial generation. Recent breakthroughs, such as large language models (e.g., GPT) and robotics capable of real-world interaction, rely on advanced methods that achieve impressive performance, yet they often lack the rigorous theoretical guarantees essential for scientific validity.

Statistical machine learning addresses this gap by providing a mathematical framework to rigorously analyze why learning algorithms succeed or fail in practice. Rather than treating models as "black boxes", this theoretical perspective confronts fundamental challenges: How do we ensure patterns learned from limited data generalize to new scenarios? What conditions prevent complex models from becoming unreliable? By formalizing these questions, statistical machine learning shifts the focus from empirical benchmarks to understanding algorithmic behavior, a fundamental step for deploying models in scientific domains, where interpretability and robustness are as important as accuracy.

The goal of this thesis is to establish such guarantees for the performance of a learning algorithm within a dynamical system setting. To this end, we leverage ideas both from statistical machine learning and dynamical systems theory.

Dynamical Systems: Beyond i.i.d. Data The study of dynamical systems focuses on how states within an environment evolve over time based on specific rules. In many real-world situations, uncertainty and inherent randomness play a significant role, resulting in the development of stochastic dynamical systems. In these systems, the next state depends not only on the current state but also on probabilistic factors.

In discrete time, we formalize stochastic dynamical systems using a state space S and a transition probability function $P(x, A)$, specifying the probability of transitioning from a state $x \in S$ to a measurable set of states $A \subseteq S$ in one time step. We will consider autonomous systems, where the transition probability $P(x, A)$ remains constant over time. Within this framework, sequential data is modeled as a Markov chain with values in S:

$$X = \{X_0, X_1, X_2, \dots\},$$

where each random variable X_t represents the state of the system at time t and is distributed according to $\mu^{[t]}$. Here, μ denotes the initial distribution of the states, and $\mu^{[t]}$ the distribution after t time-steps.

Our objective is to learn, in a supervised setting, the regression function f_ρ, defined pointwise by

$$f_\rho(x) = \mathbb{E}[\, X_{\text{next}} \mid X = x \,],$$

in order to best estimate the next expected value of the system's state. Tackling this forecasting problem requires moving beyond the classical statistical learning assumptions, where data is usually assumed to be independent and identically distributed, an idealization that fails in this setting. Extending classical results to dynamical systems thus presents two main challenges: dependency, which introduces correlations between observations, and non-stationarity, which causes changes in the data distribution over time.

The probabilistic framework of Markov chains and transition probability functions enables analysis of the system's behavior over time, including the study of long-term stability and convergence. To achieve our results, we restrict our analysis to a class of dynamical systems that ensure long-term convergence, effectively weakening the assumption of identical distributions, by employing the following ergodicity assumption: there exists a unique probability measure π on S satisfying

$$\mu^{[t]} \xrightarrow{t\to\infty} \pi$$

for any starting probability measure μ, meaning $\lim_{t\to\infty} \|\mu^{[t]} - \pi\| = 0$. In particular, we utilize the norm $\|\cdot\|_{(C^s(S))^*}$ in the dual of the Hölder space $C^s(S)$, and require that this convergence occurs at a sufficiently fast rate. This assumption ensures that the chain X is asymptotically stationary, and thus it eventually behaves as samples from the limiting measure π, allowing us to derive stable long-term properties of the learning process.

Forecasting Forecasting in stochastic dynamical systems is closely related to the well-studied area of time series analysis, where the goal is to predict future values of a sequence of observations indexed in time. Classic approaches often assume some form of stationarity, where the statistical properties (mean, variance, autocorrelation structure) remain constant over time. Under these assumptions, Autoregressive (AR), Moving Average (MA), and combined ARMA models have become standard tools. These models capture temporal dependencies by expressing the current observation as a function of a finite number of past values and past random errors. Parameter estimation in such models typically involves techniques like Least Squares, Maximum Likelihood Estimation, or Yule-Walker equations, among others.

Beyond these linear models, techniques such as moving averages or differencing can be used to address basic trends, and more specialized time-series methods have been proposed to capture certain behaviors (eg., irregular volatility). While these approaches offer well-established statistical procedures, they typically rely on strong assumptions of stationarity or limited dependence and often target narrow classes of problems. As a result, they do not provide a unifying framework that can handle the rich variety of real-world dynamical systems, particularly those exhibiting highly nonlinear dynamics.

On the side of statistical learning, approaches often assume access to complete datasets upfront (batch or offline learning), with limited emphasis on dynamical scenarios like sequential forecasting. As we mentioned before, these methods are typically framed in static settings rather than through the lens of dynamical systems or temporal dependencies. This is particularly evident in techniques like kernel methods, where explicit learning guarantees for dynamical systems are not established.

Contribution To address these challenges, we propose an Online Learning Algorithm (OLA) in a Reproducing Kernel Hilbert Space (RKHS) for forecasting in stochastic dynamical systems. For tasks such as time series forecasting, optimal control, and system identification, where data arrives sequentially, online learning algorithms are a natural

choice. Unlike traditional batch learning, online learning algorithms process incoming data incrementally, continuously updating the model.

The OLA presented in this work is an SGD-type algorithm given by the update

$$f_{t+1} := f_t - p_t\big[\big(f_t(x_t) - x_{t+1}\big)\, K_{x_t} \;-\; \lambda_t\, f_t\big],$$

where at each step, a new data pair (x_t, x_{t+1}) is observed. Rather than assuming stationarity, we build on the theory of ergodic Markov chains and derive non-asymptotic bounds, thereby bridging a gap between theoretical statistical learning and real-world applications. Specifically, we show convergence rates in expectation for $f_t \to f_\rho$ as $t \to \infty$, where f_ρ is the target regression function. This framework unifies techniques from statistical machine learning and stochastic systems analysis, broadening the applicability of kernel-based methods and offering new insights into learning in dynamical environments.

Outline The thesis is organized into five chapters. *Chapter 1* reviews the foundations of probability theory, recalling basic definitions from measure theory, random variables, and stochastic processes. *Chapter 2* covers Machine Learning with Kernels in the supervised learning framework, including reproducing kernel Hilbert spaces (RKHS), batch and online regression, and learning bounds under i.i.d. assumptions. *Chapter 3* addresses Stochastic Dynamical Systems and Markov Chains, covering fundamental concepts such as ergodicity, irreducibility, and aperiodicity. *Chapter 4* constitutes the core contribution of this work by proposing a kernel-based online learning algorithm and proving the main theorem concerning learning bounds for the algorithm. Finally, *Chapter 5* outlines future directions, proposing potential improvements, addressing learning for more general state-spaces, and relaxing independence assumptions.

Chapter 1

Background: Probability Theory

This chapter provides an introduction to the fundamental concepts of probability theory, essential for both machine learning and stochastic dynamical systems. Key concepts will be briefly introduced, ranging from *random variables* and *probability distributions* to *stochastic processes* and *filtrations*. Probability theory arises a specific instance of the more abstract field of measure theory. The latter provides an appropriate framework to rigorously 'measure', as a unifying language, general mathematical objects such as sets, functions and even random phenomena. We will try to provide sufficient definitions without delving deep into the topic. For more detailed discussions, refer to [18], [23], [17], [1]. For this part, classical notation commonly used in the literature on the subject will be employed.

1.1 Probability Distributions and Random Variables

1.1.1 Measure Theory

Randomness, as an alternative to determinism, in modern mathematics has been formalized systematically with the language of *random variables* and *probability distributions*. These quantitatively describe uncertainty and allow us to make some predictions when incomplete or noisy data is available. This framework is particularly well-suited to our goal of predicting and understanding complex phenomena that involve some degree of randomness.

The key concept of measure theory we introduce is that of a *measure*, which is defined based on a particular set structure, the *σ-algebra*.

Definition 1.1.1 (σ-algebra)**.** Let Ω be a non-empty set and 2^Ω its power set. A collection $\mathcal{A} \subseteq 2^\Omega$ is called a **$\boldsymbol{\sigma}$-algebra** (or σ-field) if it satisfies the following properties:

(i) $\Omega \in \mathcal{A}$;

(ii) if $A \in \mathcal{A}$, then $A^\complement \in \mathcal{A}$;

(iii) if $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.

We call **measurable space** the couple $(\Omega, \mathcal{A})$ and **measurable set** an element $A \in \mathcal{A}$.

We are thus endowing the set Ω with a structure closed under set operations of union and complement. In particular, one has that $\mathcal{A}$ is closed under countable union.

Example 1.1.2 An example of a σ-algebra is the *Borel σ-algebra* $\mathcal{B}(\mathbb{R})$, whose elements are called *Borel sets*. These link the concept of a σ-algebra with another fundamental mathematical structure, that of *topology*[1]. In fact, it is the smallest σ-algebra containing all the open sets of the underlying topology, in this case, the usual euclidean topology on $\mathbb{R}$ generated by all open intervals.

Definition 1.1.3 (Sub-σ-algebra)**.** Given a measurable space $(\Omega, \mathcal{A})$, a **sub-σ-algebra** of $\mathcal{A}$ is any σ-algebra $\mathcal{F}$ such that $\mathcal{F} \subseteq \mathcal{A}$. Moreover, since $\mathcal{F} \subseteq \mathcal{A}$, every set in $\mathcal{F}$ is also in $\mathcal{A}$, and thus every event measurable with respect to $\mathcal{F}$ is also an event measurable with respect to $\mathcal{A}$.

Now let's see how we are actually able to "measure" the measurable sets that we just introduced.

Definition 1.1.4 (Measure)**.** Given a non-empty set Ω and a σ-algebra $\mathcal{A}$ on Ω, a (positive) **measure** is a function $\mu : \mathcal{A} \longrightarrow [0, +\infty]$ such that:

(i) $\mu(\emptyset) = 0$;

(ii) given $(A_n)_{n\geq 1} \in \mathcal{A}$ a countable family of disjoint sets , we have[2]

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

A well-known example of a measure is the *Lebesgue measure*, which generalizes the concept of length, area, and volume in $\mathbb{R}^n$. The Lebesgue measure is pivotal in various fields of mathematics, including integration theory and probability, as it allows for the measurement of more intricate sets that arise in these contexts.

A simpler yet fundamentally important example is the **Dirac measure**, often referred to as the *Dirac delta measure*.

Example 1.1.5 Given a non-empty set Ω, a **Dirac measure** centered at a point $\omega \in \Omega$ is a measure $\delta_\omega : \mathcal{A} \longrightarrow [0, +\infty]$ defined by

$$\delta_\omega(A) := \mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A, \end{cases} \quad \forall A \in \mathcal{A}.$$

where $\mathbf{1}_A$ is the usual indicator function $\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$

The Dirac measure δ_ω assigns full measure to the singleton set containing ω. This measure is particularly useful in probability theory and functional analysis for modeling deterministic outcomes within probabilistic frameworks.

A measure μ on a measurable space $(\Omega, \mathcal{A})$ is called *finite* if $\mu(\Omega) < \infty$. Otherwise, it is called *infinite*. Within the class of infinite measures, there is a subclass with an important property, called *σ-finiteness*. Many fundamental facts of measure and integration theory that we will use in later chapters only hold for measures that are σ-finite.

Definition 1.1.6 (σ-Finite measure)**.** Let μ be a measure on a measurable space $(\Omega, \mathcal{A})$. Then μ is called **σ-finite** if there is a sequence $A_1, A_2, \ldots \in \mathcal{A}$ with $\bigcup_{i=1}^{\infty} A_i = \Omega$ and, for all $i = 1, 2, \ldots$, $\mu(A_i) < \infty$.

[1] A collection of subsets of X containing the whole set X and the empty set, closed under finite intersection and infinite union. Its elements are called *open sets*.

[2] This property is called σ-additivity.

1.1.2 Probability Space

With these notions, we now introduce the framework in which probability theory develops, the *probability space*.

Definition 1.1.7 (Probability Space)**.** A **probability space** is a triplet $(\Omega, \mathcal{A}, \mathbb{P})$, where:

- $\boldsymbol{\Omega}$ is the **sample space**, the set of all possible outcomes of an experiment;
- $\boldsymbol{\mathcal{A}}$ is a $\boldsymbol{\sigma}$**-algebra** on Ω, whose elements are called **events**;
- $\mathbb{P}$ is a **probability measure** on $\mathcal{A}$: a measure whose image is in the interval $[0, 1]$ and such that $\mathbb{P}(\Omega) = 1$.

Remark 1.1.8 We denote with $\mathcal{M}(\Omega, \mathcal{A})$, or simply $\mathcal{M}(\Omega)$ when the σ-algebra is implicit, *the set of all possible measures* on the measurable space; while $\mathcal{M}_+(\Omega)$ and $\mathcal{P}(\Omega)$ denote respectively *the set of positive measures* and *the set of probability measures* on $(\Omega, \mathcal{A})$.

The conditions we have placed on the measure provide us with a tool to consistently quantify the possible outcomes of a random phenomenon. We can model these phenomena with *random variables*.

1.1.3 Random Variables

Definition 1.1.9 (Random Variable)**.** Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, consider a function $X : \Omega \to \Omega'$, where $(\Omega', \mathcal{A}')$ is a measurable space. The function X is said to be a **random variable** if it is $(\mathcal{A}, \mathcal{A}')$**-measurable**, meaning that for every $A \in \mathcal{A}'$,

$$X^{-1}(A) \in \mathcal{A}.$$

We will primarily focus on **real-valued random variables**, where $\Omega' = \mathbb{R}$ and $\mathcal{A}' = \mathcal{B}(\mathbb{R})$, the Borel σ-algebra[3]. In this context, the measurability condition is equivalent to

$$X^{-1}((-\infty, a]) \in \mathcal{A} \quad \forall a \in \mathbb{R}.$$

We will denote random variables with uppercase letters, such as X, and use lowercase letters, such as x, for their corresponding *observations* (or *realizations*), expressed as $X(\omega) = x$. Our focus will be on describing the values that X can assume in relation to the modeled uncertain phenomenon through its *probability distribution*.

Before proceeding, it is useful to introduce the concept of *generated σ-algebra*, which will be important later when we discuss filtrations and conditional expectations.

Definition 1.1.10 (Generated σ-algebra)**.** Let $\mathcal{F}$ be a collection of subsets of a set Ω. The $\boldsymbol{\sigma}$**-algebra generated by** $\boldsymbol{\mathcal{F}}$, denoted by $\sigma(\mathcal{F})$, is the smallest σ-algebra containing $\mathcal{F}$. Formally, it is the intersection of all σ-algebras on Ω that contain $\mathcal{F}$:

$$\sigma(\mathcal{F}) = \bigcap \{\mathcal{A} \subseteq 2^\Omega \mid \mathcal{A} \text{ is a } \sigma\text{-algebra and } \mathcal{F} \subseteq \mathcal{A}\}.$$

In particular, if $X : \Omega \to \Omega'$ is a function (such as a random variable) from Ω to a measurable space $(\Omega', \mathcal{B})$, the $\boldsymbol{\sigma}$**-algebra generated by** $\boldsymbol{X}$, denoted by $\sigma(X)$, is defined as the σ-algebra generated by the collection of pre-images of sets in $\mathcal{B}$ under X:

$$\sigma(X) := \sigma\left(\{X^{-1}(A) \mid A \in \mathcal{B}\}\right).$$

This means that $\sigma(X)$ is the smallest σ-algebra on Ω such that X is measurable with respect to it.

[3]The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ can be generated from the intervals $(-\infty, a]$ with $a \in \mathbb{R}$.

One interesting σ-algebra will be the one *generated* by the Cartesian product of measurable spaces.

Definition 1.1.11 (Product σ-algebra)**.** Let $(\Omega_1, \mathcal{A}_1), \ldots, (\Omega_n, \mathcal{A}_n)$ be measurable spaces and $\Omega := \Omega_1 \times \ldots \times \Omega_n$. Then

$$\mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n := \bigotimes_{i=1}^{n} \mathcal{A}_i := \sigma\left(\left\{\prod_{i=1}^{n} A_i : A_i \in \mathcal{A}_i,\, i = 1, \ldots, n\right\}\right) \tag{1.23}$$

is called the **product σ-algebra** of the σ-algebras $\mathcal{A}_i$, $i = 1, \ldots, n$.

Note that the product σ-algebra $\mathcal{A}_1 \otimes \ldots \otimes \mathcal{A}_n$ is *not* the Cartesian product $\mathcal{A}_1 \times \ldots \times \mathcal{A}_n$. Instead, the product σ-algebra is generated by the set system of all Cartesian products of elements of the σ-algebras $\mathcal{A}_1, \ldots, \mathcal{A}_n$. In Chapter 3, we give an equivalent specification of a product σ-algebra, using projection mappings.

Building on the concept of the product σ-algebra, we now define the associated product measure for σ-finite measures on these spaces.

Proposition 1.1.12 (Product measure)**.** *Let $(\Omega_i, \mathcal{A}_i, \mu_i)$ be measure spaces with σ-finite measures $\{\mu_i\}$, $i = 1, \ldots, n$. Then there exists a uniquely defined measure, denoted by $\mu_1 \otimes \cdots \otimes \mu_n$, on the product space*

$$\left(\prod_{i=1}^{n} \Omega_i, \bigotimes_{i=1}^{n} \mathcal{A}_i\right),$$

satisfying

$$\mu_1 \otimes \cdots \otimes \mu_n(A_1 \times \cdots \times A_n) = \mu_1(A_1) \cdots \mu_n(A_n), \quad \forall (A_1, \ldots, A_n) \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_n.$$

*This measure is σ-finite as well, and it is called the **product measure** of $\mu_1, \ldots, \mu_n$.*

Definition 1.1.13 (Probability distribution)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \Omega'$ be a random variable, where $(\Omega', \mathcal{A}')$ is a measurable space.

(i) The **distribution** (or **law**) of X, denoted by $\mathcal{L}(X)$ or $\mathbb{P}_X$, is the probability measure $\mathbb{P}_X := \mathbb{P} \circ X^{-1}$ on $(\Omega', \mathcal{A}')$, defined by

$$\mathbb{P}_X(A) := \mathbb{P}(X^{-1}(A)), \quad \forall A \in \mathcal{A}'.$$

(ii) For a real random variable $X : \Omega \to \mathbb{R}$, where $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the **distribution function** (or **cumulative distribution function**, CDF) of X is the function $F_X : \mathbb{R} \to [0, 1]$, defined by

$$F_X(x) := \mathbb{P}(X \leq x), \quad x \in \mathbb{R}.$$

(iii) A random variable X is said to have distribution μ if $\mathcal{L}(X) = \mu$.

(iv) A family of random variables $(X_i)_{i \in I}$ is called **identically distributed** if $\mathbb{P}_{X_i} = \mathbb{P}_{X_j}$ for all $i, j \in I$.

This framework allows us to shift our focus from individual outcomes in the sample space Ω to properties of X itself by studying the measure $\mathbb{P}_X$. Particularly for real-valued random variables, this approach facilitates the use of tools from real analysis, such as integration with respect to these measures.

Theorem 1.1.14. *For any distribution function F, there exists a real random variable X with $F_X = F$.*

Definition 1.1.15. A random variable X on $(\Omega, \mathcal{A}, \mathbb{P})$ is called **discrete** if there exists a countable set $\{x_1, x_2, \dots\}$ such that

$$\mathbb{P}(X \in \{x_1, x_2, \dots\}) = 1.$$

In other words, X takes values in a finite, or at most countable, set of points, each with a positive probability.

Definition 1.1.16. A random variable X is called **continuous** if the probability that it takes any *specific* value is zero, that is

$$\mathbb{P}(X = x) = 0 \quad \text{for every } x \in \mathbb{R}.$$

An even stronger condition than continuity is the one characterizing *absolutely continuous random variables*, which possess the valuable property of admitting a *probability density function*.

Definition 1.1.17. A random variable X on $(\Omega, \mathcal{A}, \mathbb{P})$ is said to be **absolutely continuous** if the induced probability measure $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue measure λ on $\mathbb{R}$, that is, if

$$\lambda(A) = 0 \Rightarrow \mathbb{P}_X(A) = 0 \quad \forall A \in \mathcal{B}(\mathbb{R}). \tag{1.1}$$

This property ensures that the variable X does not assign positive probability to "small" sets, where "small" means sets with zero Lebesgue measure. In particular, it prevents X from assigning positive probability to individual points, instead distributing the probability over continuous intervals. As mentioned before, absolute continuity is a stronger condition than mere continuity because it guarantees the existence of *probability density functions*.

Definition 1.1.18 (Density function)**.** The **probability density function** of an absolutely continuous random variable X is a function $p_X : \mathbb{R} \longrightarrow [0, \infty)$ such that for every $A \in \mathcal{B}(\mathbb{R})$

$$\mathbb{P}(X \in A) = \int_A p_X(x)dx.$$

Intuitively, the probability density function p_X describes how probability is concentrated in the outcome space and thus satisfies the important and useful property

$$\int_{\mathbb{R}} p_X(x)dx = 1. \tag{1.2}$$

Example 1.1.19 (Uniform Distribution) A classic example of an absolutely continuous random variable is one that is **uniformly distributed** on the interval $[0, 1]$.

Let X be a random variable with density function

$$p_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

This X is said to be distributed as $U(0, 1)$. Intuitively, X is equally likely to fall anywhere in the interval $[0, 1]$. For any sub-interval $[a, b] \subseteq [0, 1]$, the probability that X lands in $[a, b]$ is simply $b - a$.

Definition 1.1.20 (Expectation)**.** Let X be a real-valued random variable on $(\Omega, \mathcal{A}, \mathbb{P})$. The **expected value** (or **mean**) of X is defined as

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, d\mathbb{P}_X(x),$$

where the distribution of X is $\mathbb{P}_X$ and provided the integral exists (i.e., $\int_{\mathbb{R}} |x| \, d\mathbb{P}_X(x) < \infty$).

Remark 1.1.21 While the expected value of a general random variable is defined via integration with respect to its distribution, in the case of an absolutely continuous random variable X with probability density function p_X, the expected value simplifies to

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \, p_X(x) \, dx.$$

This expression calculates the weighted average of all possible values that X can assume, weighted by their probability density.

In situations involving multiple densities or variables, or when it is necessary to specify the distribution explicitly, we use the notation

$$\mathbb{E}_{x \sim \mathbb{P}_X}[X] = \mathbb{E}_x[X] = \int_{\mathbb{R}} x \, \mathbb{P}_X(x),$$

which we read as *"the expectation of X with respect to observations x distributed according to $\mathbb{P}_X$"*.

Random Vectors

Extending random variables to higher-dimensional spaces is fundamental in applied sciences, allowing the modeling of multidimensional phenomena that describe signals and data in high-dimensional spaces. A (real) **random vector** $\mathbf{X}$ of dimension $k \in \mathbb{N}_{>0}$ is a measurable function from the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ to $\mathbb{R}^k$, that is,

$$\mathbf{X} : (\Omega, \mathcal{A}) \longrightarrow (\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k)),$$

such that for every Borel set $A \in \mathcal{B}(\mathbb{R}^k)$, the preimage $\mathbf{X}^{-1}(A) \in \mathcal{A}$.

Specifically, we can write

$$\mathbf{X} = (X_1, X_2, \ldots, X_k),$$

where each component $X_i : \Omega \to \mathbb{R}$ is a real-valued random variable. The random vector $\mathbf{X}$ induces a probability measure on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$, known as the **joint distribution** of $\mathbf{X}$, defined by

$$\mathbb{P}_{\mathbf{X}}(A) := \mathbb{P}\left(\mathbf{X}^{-1}(A)\right), \quad \forall A \in \mathcal{B}(\mathbb{R}^k).$$

If $\mathbf{X}$ is absolutely continuous, its joint distribution $\mathbb{P}_{\mathbf{X}}$ is absolutely continuous with respect to the Lebesgue measure λ^k on $\mathbb{R}^k$, and there exists a **joint probability density function** $p_{\mathbf{X}} : \mathbb{R}^k \to [0, \infty)$ such that

$$\mathbb{P}_{\mathbf{X}}(A) = \int_A p_{\mathbf{X}}(x_1, \ldots, x_k) \, dx_1 \cdots dx_k, \quad \forall A \in \mathcal{B}(\mathbb{R}^k).$$

The **marginal distribution** of a component X_i is the probability measure $\mathbb{P}_{X_i}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by

$$\mathbb{P}_{X_i}(B) := \mathbb{P}_{\mathbf{X}}\left(\{(x_1, \ldots, x_k) \in \mathbb{R}^k \mid x_i \in B\}\right), \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

If $\mathbf{X}$ has a joint density $p_{\mathbf{X}}$, then the marginal density of X_i is given by

$$p_{X_i}(x_i) = \int_{\mathbb{R}^{k-1}} p_{\mathbf{X}}(x_1, \ldots, x_k)\, dx_1 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_k.$$

The concept of expectation extends naturally to random vectors. The **expected value** (or **mean vector**) of $\mathbf{X}$ is defined as

$$\mathbb{E}[\mathbf{X}] := (\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots, \mathbb{E}[X_k])\,,$$

where each component $\mathbb{E}[X_i]$ is calculated with respect to the marginal distribution $\mathbb{P}_{X_i}$ as

$$\mathbb{E}[X_i] = \int_{\mathbb{R}} x_i\, d\mathbb{P}_{X_i}(x_i).$$

In contexts where the dimension is not crucial, we will refer to multivariate random variables simply as *random variables*. This simplifies the language and emphasizes that many concepts apply regardless of dimensionality.

1.2 Independence

The notion of *independence* is central to probability theory and plays a pivotal role in both theoretical and applied settings. Informally, two events are independent if the occurrence of one does not affect the likelihood of occurrence of the other. This concept naturally extends to collections of events and to random variables, ultimately facilitating factorization properties of joint distributions that greatly simplify analysis in more complex models.

Definition 1.2.1 (Independence of Events)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let $(A_i)_{i\in I}$ be a (finite or countably infinite) collection of events $A_i \in \mathcal{A}$ indexed by some set I. We say that the family $(A_i)_{i\in I}$ is **independent** if for every finite subset $J \subset I$, we have

$$\mathbb{P}\Big(\bigcap_{j\in J} A_j\Big) = \prod_{j\in J} \mathbb{P}(A_j). \tag{1.3}$$

In particular, for two events $A, B \in \mathcal{A}$, this reduces to

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

If no such factorization can be achieved for a given collection of events, we say that the events are **dependent**.

Remark 1.2.2 Independence of an infinite family $(A_i)_{i\in I}$ is thus a strong condition, requiring this multiplicative structure for every finite subfamily. Independence is strictly stronger than mere pairwise independence, as it requires joint factorization across all finite subsets, not just pairs.

The definition of independence for events can be extended to *σ-algebras* and hence to *random variables*.

A family of σ-algebras $(\mathcal{A}_i)_{i\in I}$ is said to be independent if no one of them provides any information about the others. More concretely, for every finite subset $J \subset I$, whenever we pick one event from each $\mathcal{A}_j$, $j \in J$, their joint probability measure factorizes into the product of the individual probabilities. Since each random variable X_i generates a σ-algebra $\sigma(X_i)$ of events in Ω, this notion naturally extends to define **independent random variables**.

Definition 1.2.3 (Independence of Random Variables)**.** Let $(X_i)_{i\in I}$ be a family of random variables on $(\Omega, \mathcal{A}, \mathbb{P})$ taking values in measurable spaces $(\Omega_i, \mathcal{A}_i)$, respectively. The random variables $(X_i)_{i\in I}$ are **independent** if the family of σ-algebras $(\sigma(X_i))_{i\in I}$ is independent. Equivalently, for every finite $J \subset I$ and every collection of measurable sets $(A_j)_{j\in J}$ with $A_j \in \mathcal{A}_j$, we have

$$\mathbb{P}\Big(\bigcap_{j\in J}\{X_j \in A_j\}\Big) = \prod_{j\in J}\mathbb{P}(X_j \in A_j).$$

When each X_i in an independent family $(X_i)_{i\in I}$ shares the same distribution, we say that $(X_i)_{i\in I}$ are **i.i.d.**, i.e., **independent and identically distributed**.

Independence is a fundamental concept for building probabilistic models that assume no underlying dependencies between components. Such assumptions are frequently employed in statistical learning methods, stochastic modeling, and the analysis of algorithms, providing tractability and simplifying both theoretical analysis and numerical computation.

One particularly useful consequence of independence is the following theorem, which relates the expectation of a product of independent random variables to the product of their expectations.

Theorem 1.2.4 (Expectation of the product of random variables)**.** *Let $Y_i : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{B})$, $i = 1, \ldots, n$, be real-valued random variables that are non-negative or with finite expectations, and assume that $Y_1, \ldots, Y_n$ are independent. Then,*

$$\mathbb{E}\left(\prod_{i=1}^{n} Y_i\right) = \prod_{i=1}^{n}\mathbb{E}(Y_i).$$

1.3 Conditioning and Filtrations

1.3.1 Conditional Expectation

In probability theory, the concept of *conditioning* is fundamental when we want to update our understanding of a random phenomenon upon receiving some additional information. We begin by considering conditioning on events and then naturally extend this idea to conditioning on entire collections of events (i.e., on σ-algebras). This leads us to the notion of a *conditional expectation* of a random variable given a σ-algebra.

Definition 1.3.1 (Conditional Probability Given an Event)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and consider an event $A \in \mathcal{A}$ with $\mathbb{P}(A) > 0$. The **conditional probability** of an event $B \in \mathcal{A}$ given A is defined as

$$\mathbb{P}(B \mid A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

This construction induces a new probability measure $\mathbb{P}(\cdot \mid A)$ on A, normalized so that $\mathbb{P}(A \mid A) = 1$.

Similarly, if $X : \Omega \to \mathbb{R}$ is a (real) random variable, we can consider its conditional distribution given A:

$$\mathbb{P}_{X|A}(B) := \frac{\mathbb{P}(X \in B) \cap \mathbb{P}(A)}{\mathbb{P}(A)}, \quad B \in \mathcal{B}(\mathbb{R}).$$

The corresponding **conditional expectation of X given A** is

$$\mathbb{E}[X \mid A] := \int_{\mathbb{R}} x \, d\mathbb{P}_{X|A}(x).$$

Remark 1.3.2 While this notion is straightforward, it is limited because we are only conditioning on a single event A. To capture the notion of conditioning on richer forms of "information", we now move to conditioning on a sub-σ-algebra $\mathcal{F} \subseteq \mathcal{A}$, which can be viewed as encoding a *system* of events that provide partial information.

Definition 1.3.3 (Conditional Expectation with respect to a σ-algebra)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{F} \subseteq \mathcal{A}$ a sub-σ-algebra, and let $X : \Omega \to \mathbb{R}$ be an integrable random variable (i.e. $\mathbb{E}[|X|] < \infty$). A random variable $Y : \Omega \to [0, +\infty]$ is called a **conditional expectation of X given $\mathcal{F}$**, denoted by $\mathbb{E}[X \mid \mathcal{F}]$, if:

1. Y is $\mathcal{F}$-measurable,
2. For all $F \in \mathcal{F}$,
$$\int_F Y(\omega) \, d\mathbb{P}(\omega) = \int_F X(\omega) \, d\mathbb{P}(\omega).$$

Such a Y exists and is unique up to sets of $\mathbb{P}$-measure zero.

Remark 1.3.4 Intuitively, $\mathbb{E}[X \mid \mathcal{F}]$ is the *best $\mathcal{F}$-measurable approximation to X* (in the L^1 sense). If we think of $\mathcal{F}$ as representing the information available to us, then $\mathbb{E}[X \mid \mathcal{F}]$ is the "updated expectation" of X once we incorporate that information. If we choose $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$, the conditional expectation reduces to the constant random variable $\mathbb{E}[X \mid A]$, thus generalizing the event-based case.

From the definition, we can recover **conditional probability given $\mathcal{F}$** by applying the conditional expectation operator to indicator functions as

$$\mathbb{P}(B \mid \mathcal{F}) := \mathbb{E}[\mathbf{1}_B \mid \mathcal{F}], \quad B \in \mathcal{A}.$$

1.3.2 Conditioning on a Random Variable

We have introduced conditional expectation with respect to a sub-σ-algebra $\mathcal{F} \subseteq \mathcal{A}$. One common scenario is when this σ-algebra is generated by a particular random variable $X : \Omega \to (E, \mathcal{E})$. Conditioning on the random variable X can be viewed as conditioning on the information encoded by the values that X takes. Formally, we define the conditional expectation of Y given X as the conditional expectation of Y with respect to $\sigma(X)$.

Definition 1.3.5 (Conditional Expectation Given a Random Variable)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $X : \Omega \to (E, \mathcal{E})$ a random variable, and let $Y \in L^1(\mathbb{P})$, meaning $\mathbb{E}[|Y|] < \infty$.

The **conditional expectation of Y given X** is defined as

$$\mathbb{E}[Y \mid X] := \mathbb{E}[Y \mid \sigma(X)].$$

By definition, $\mathbb{E}[Y \mid X]$ is $\sigma(X)$-measurable and satisfies

$$\int_{X^{-1}(A)} \mathbb{E}[Y \mid X](\omega) \, d\mathbb{P}(\omega) = \int_{X^{-1}(A)} Y(\omega) \, d\mathbb{P}(\omega), \quad \text{for all } A \in \mathcal{E}.$$

In analogy with conditioning on an event, for $x \in E$, we write

$$\mathbb{E}[Y \mid X = x]$$

provided that the integral exists, and call it a conditional expectation of Y with respect to X. If there is no ambiguity, we may omit reference to $\mathbb{P}$ and simply write $\mathbb{E}[Y \mid X]$.

Proposition 1.3.6 (Existence and Uniqueness)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{F} \subseteq \mathcal{A}$ a sub-σ-algebra, and let $Y \in L^1(\mathbb{P})$. Then there exists a random variable $\mathbb{E}[Y \mid \mathcal{F}]$ that is $\mathcal{F}$-measurable and satisfies*

$$\int_F \mathbb{E}[Y \mid \mathcal{F}](\omega)\, d\mathbb{P}(\omega) = \int_F Y(\omega)\, d\mathbb{P}(\omega) \quad \text{for all } F \in \mathcal{F}.$$

Moreover, this conditional expectation is unique up to $\mathbb{P}$-null sets.

Definition 1.3.7. We can define the **conditional probability of an event A given X** as

$$\mathbb{P}(A \mid X) := \mathbb{E}[\mathbf{1}_A \mid X].$$

For each fixed x in the image of X, this gives a probability measure $\mathbb{P}(\cdot \mid X = x)$ on $\mathcal{A}$.

Remark 1.3.8 The expression $\mathbb{E}[Y \mid X = x]$ is defined $\mathbb{P}_X$-almost surely. That is, it is well-defined for almost every x with respect to the distribution of X.

In this way, we are updating our expectations and probabilities based on the specific value X takes. This perspective is especially useful in scenarios involving *Markov processes*, where knowing the current state (the realized value of X) allows us to better understand the behavior of the system moving forward.

Proposition 1.3.9 (Properties of Conditional Expectation)**.** *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $\mathcal{F} \subseteq \mathcal{A}$ a sub-σ-algebra. For integrable random variables $X, Y : \Omega \to \mathbb{R}$, the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}]$ satisfies:*

(i) ***Linearity****: For all constants $\lambda, \mu \in \mathbb{R}$,*

$$\mathbb{E}[\lambda X + \mu Y \mid \mathcal{F}] = \lambda \mathbb{E}[X \mid \mathcal{F}] + \mu \mathbb{E}[Y \mid \mathcal{F}].$$

(ii) ***Stability****: If Z is $\mathcal{F}$-measurable, then*

$$\mathbb{E}[ZX \mid \mathcal{F}] = Z\, \mathbb{E}[X \mid \mathcal{F}] \quad a.s..$$

In particular, $\mathbb{E}[Z \mid \mathcal{F}] = Z$.

(iii) ***Tower property****: If $\mathcal{F}_1 \subseteq \mathcal{F}_2$ are sub-σ-algebras, then*

$$\mathbb{E}\big[\mathbb{E}[X \mid \mathcal{F}_2] \mid \mathcal{F}_1\big] = \mathbb{E}[X \mid \mathcal{F}_1].$$

This is also known as 'Law of total expectation' and, in the case $\mathcal{F}_1 = \{\emptyset, \Omega\}$, it implies

$$\mathbb{E}\big[\mathbb{E}[Z] \mid \mathcal{F}\big] = \mathbb{E}[Z].$$

(iv) ***Independence****: If X is independent of $\mathcal{F}$, then*

$$\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}[X].$$

These fundamental properties mirror those of the usual expectation and ensure that conditioning does not break essential rules like linearity. In particular, all these properties also hold when we consider $\mathbb{E}[Y \mid X]$ by simply replacing $\mathcal{F}$ with $\sigma(X)$.

In machine learning and stochastic dynamical systems, these conditional tools play a fundamental role. They allow us to incorporate known information (such as observed features in supervised learning or past states in a dynamical model) into probability distributions. As a result, conditional expectations and conditional distributions form the backbone of many learning algorithms and methods for handling uncertainty.

1.3.3 Filtrations

As we have seen, conditional expectations and probabilities naturally arise when we focus on a certain subset of events, representing the information currently known. In many applications, especially those involving time-evolution (such as stochastic processes), our information about the system accumulates progressively. To capture this formally, we introduce the concept of a *filtration*.

Definition 1.3.10 (Filtration)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let T be a totally ordered index set (often $T = \{0, 1, 2, \dots\}$ or $T = [0, \infty)$). A **filtration** $(\mathcal{F}_t)_{t \in T}$ is a family of sub-σ-algebras of $\mathcal{A}$ such that for every $s, t \in T$ with $s \leq t$ we have

$$\mathcal{F}_s \subseteq \mathcal{F}_t.$$

Remark 1.3.11 Filtrations are central in the theory of stochastic processes and will play a crucial role in defining *adapted processes*, and *Markov processes*. Intuitively, the idea is that at each time t, certain aspects of the system have been observed, and the collection $\mathcal{F}_t$ encodes exactly what is known at that point. As t grows, we gather more observations and therefore have a richer σ-algebra $\mathcal{F}_t$. This helps to bridge the gap between the intuitive flow of information through time and the formalism of σ-algebras which would otherwise not encapsulate time or causality by itself.

Example 1.3.12 Let us consider a two-step experiment where we roll a fair six-sided die twice. The sample space is

$$\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\},$$

and we equip Ω with the σ-algebra $\mathcal{A} = 2^\Omega$ and the probability measure $\mathbb{P}$ induced by assuming each of the 36 outcomes (ω_1, ω_2) is equally likely with probability $1/36$.

Define the random variables $X_1(\omega) = \omega_1$ and $X_2(\omega) = \omega_2$, corresponding to the results of the first and second rolls, respectively. We consider three distinct times: before any rolls have occurred (time 0), after observing the outcome of the first roll X_1 but before rolling the die the second time (time 1), and finally after both rolls have been observed (time 2).

The filtration $(\mathcal{F}_t)_{t=0,1,2}$ associated with the natural information flow of this experiment is:

$$\mathcal{F}_0 = \{\emptyset, \Omega\}, \quad \mathcal{F}_1 = \sigma(X_1), \quad \mathcal{F}_2 = \sigma(X_1, X_2).$$

At time 0 we know nothing about the outcome. The only events we can determine with certainty, i.e. measure, are the trivial ones: "no outcome" ($\emptyset$) and "some outcome occurs" (Ω). Thus,

$$\mathcal{F}_0 = \{\emptyset, \Omega\}.$$

At time 1 we know $X_1(\omega) = \omega_1$, the result of the first roll. The σ-algebra $\mathcal{F}_1 = \sigma(X_1)$ consists of all events that can be expressed in terms of knowing the first coordinate. Concretely, any event in $\mathcal{F}_1$ looks like

$$A \times \{1, 2, 3, 4, 5, 6\} \quad \text{with } A \subseteq \{1, 2, 3, 4, 5, 6\}.$$

Since there are $2^6 = 64$ subsets of $\{1, \dots, 6\}$, we have

$$|\mathcal{F}_1| = 64.$$

At time 2 we know both X_1 and X_2. The σ-algebra $\mathcal{F}_2 = \sigma(X_1, X_2)$ is just $\mathcal{A} = 2^\Omega$, since with both coordinates known we can distinguish every single outcome in Ω. Thus,

$$\mathcal{F}_2 = 2^\Omega,$$

and since $|\Omega| = 36$, it follows that

$$|\mathcal{F}_2| = 2^{36}.$$

We see the natural chain of information: $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2$. This filtration thus neatly captures the idea of information accumulating over the course of the experiment.

1.4 Stochastic Processes

We conclude this chapter with a brief introduction to stochastic processes, which will serve as a fundamental tool in Chapter 3. Stochastic processes provide a framework for modeling systems that evolve over time with inherent randomness. They are particularly useful for studying phenomena where future states depend on probabilistic rules, making them crucial for understanding dynamical systems subject to random influences.

Definition 1.4.1. A **stochastic process** is a collection of random variables $(X_t)_{t\in T}$, where T is an index set representing time. Each random variable X_t maps the sample space Ω of a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ to a measurable space $(S, \mathcal{B})$ usually called state space, which can be either discrete or continuous. The index set T can be:

- **Discrete**, such as $T = \mathbb{N}$, representing discrete-time processes where the evolution occurs at integer time steps, or
- **Continuous**, such as $T = \mathbb{R}^+$, representing continuous-time processes where the evolution is tracked over a continuous interval.

In both cases the process represents the evolution of the random variable X_t over the parameter $t \in T$, typically representing time.

One important class of stochastic processes is the one of **Markov processes**, which satisfy the **Markov property**: *the future state of the process depends only on the current state, not on the sequence of events that preceded it.* Formally we have the following definition.

Definition 1.4.2. For a **discrete-time Markov process** (also called **Markov chain**) $(X_t)_{t\in\mathbb{N}}$, the property is given by

$$\mathbb{P}(X_{t+1} \in A \mid X_t, X_{t-1}, \ldots, X_0) = \mathbb{P}(X_{t+1} \in A \mid X_t), \tag{1.4}$$

where A is a measurable subset of the state space.

In the case of a discrete-time Markov process with a finite state space, the evolution of the system can be described by a *transition matrix* as follows: let S be a finite state space, and let $(X_n)_{n\in\mathbb{N}}$ be a discrete-time Markov process on S. The Markov property implies that the process can be fully characterized by the probabilities of transitioning from one state to another at each time step. These probabilities are encoded in a matrix called the **transition matrix**.

The transition matrix $P = (p(x, y))_{x,y\in S}$ is defined as

$$p(x, y) := \mathbb{P}(X_{n+1} = y \mid X_n = x), \quad x, y \in S,$$

where $p(x, y)$ represents the probability of transitioning from state x to state y in one time step.

Each row of the matrix P corresponds to a probability distribution over the possible next states, meaning that the entries of each row must sum to 1:

$$\sum_{y \in S} p(x, y) = 1 \quad \text{for all } x \in S.$$

This type of matrix is called a **stochastic matrix**, and it captures the dynamics of the Markov chain over time. By multiplying the current state distribution by the transition matrix, we can obtain the distribution of the process at the next time step. This mechanism is fundamental for modeling the evolution of systems with probabilistic transitions.

We later explore how to transpose this concept in the case of continuous state spaces.

The following definition relates the concept of a filtration to the measurability of a stochastic process, introducing the idea of *adaptedness*.

Definition 1.4.3 (Adapted Process)**.** Let $(X_t)_{t \in T}$ be a stochastic process taking values in a measurable space $(S, \mathcal{B})$, and let $(\mathcal{F}_t)_{t \in T}$ be a filtration on $(\Omega, \mathcal{A}, \mathbb{P})$. The process $(X_t)_{t \in T}$ is said to be **adapted** to $(\mathcal{F}_t)_{t \in T}$ if for every $t \in T$, the random variable X_t is $\mathcal{F}_t$-measurable.

Remark 1.4.4 Naturally, every stochastic process is adapted to its *natural filtration*, defined as $(\mathcal{F}_t^X)_{t \in T}$, where

$$\mathcal{F}_t^X := \sigma(X_s : s \leq t).$$

In this case, the filtration $\mathcal{F}_t^X$ represents all the information that can be derived from the process $(X_s)_{s \leq t}$ up to time t.

In this sense, the Markov property can be stated as follows: the conditional distribution of the next state X_{t+1}, given the entire past $(X_s)_{s \leq t}$, depends only on the current state X_t. Formally, this means that for a stochastic process

$$\mathbb{P}(X_{t+1} \in A \mid \mathcal{F}_t) = \mathbb{P}(X_{t+1} \in A \mid X_t) \quad \text{for all measurable } A \subseteq S.$$

1.4.1 Classification of Stochastic Processes

Stochastic processes can be broadly categorized based on the nature of the state space and the indexing set as follows:

1. **Discrete-Time and Discrete-State Processes**: These processes evolve in discrete time steps, and the state space is finite or countable. A typical example is the one previously mentioned regarding *Markov chains*, where the system moves between states according to probabilities specified by a transition matrix.

2. **Discrete-Time and Continuous-State Processes**: The time is still measured in discrete steps, but the state space is continuous. For instance, Markov processes with transition probability kernels are a specific example that we will examine in the Chapter 3.

3. **Continuous-Time and Discrete-State Processes**: The state changes at random times, which may follow certain probability distributions. For instance, a *Poisson process* counts the occurrences of random events over continuous time.

4. **Continuous-Time and Continuous-State Processes**: These processes involve continuous evolution in both time and state space. *Brownian motion* is a classic example, used to model various physical systems like the motion of atoms.

Chapter 2

Machine Learning with Kernels

Machine learning provides a framework for making predictions and extracting patterns. In Chapter 1, we introduced the probabilistic foundations underpinning much of machine learning theory, focusing on essential concepts such as random variables, distributions, and expectations. Building on these principles, we introduce supervised learning, an approach of learning by examples, where the objective is to infer relationships between inputs and outputs from labeled data.

Specifically, our goal in this chapter is to introduce *kernel methods*, a class of algorithms that combine theoretical rigor with practical flexibility.

Kernel methods offer a powerful alternative for introducing non-linearity into models. While *neural networks* achieve this through hierarchical composition of non-linear functions, kernel methods rely on mapping data into high-dimensional *feature spaces*, where linear algorithms can operate effectively. This approach is grounded in the mathematical framework of *reproducing kernel Hilbert spaces* (RKHS), which we will formally introduce. Kernel methods have achieved widespread success in various tasks, including regression, classification, and clustering, due to their balance between performance, computational efficiency, and interpretability.

The chapter is structured as follows. We begin by formalizing the problem of supervised learning in the classical i.i.d. setting, introducing key notions such as *hypothesis spaces*, *loss functions*, and *risk minimization*. Next, we examine the theory of RKHS and kernel functions, laying the groundwork for kernel-based learning methods. We then discuss specific *online-learning* algorithms as applications of these methods, including *recursive least squares* and *Stochastic Gradient Descent (SGD)*, emphasizing their ability to handle non-linear relationships in data. Finally, we conclude with an overview of learning bounds, which provide theoretical guarantees on the performance of these kernel-based algorithms.

2.1 Supervised Learning: The Classical Setting

The goal of supervised learning is to find an input/output relation $\hat{f}$ from a training set

$$\{(x_1, y_1), \ldots, (x_n, y_n)\}$$

of input/output pairs (called *samples* or *examples*). Given a new input x_{new}, the function $\hat{f}$ should predict y_{new} as the output $\hat{f}(x_{\text{new}})$. When $\hat{f}$ provides good predictions for previously unseen data, we say that $\hat{f}$ *generalizes*.

In this section we formalize this idea by adopting the classical framework of Statistical Learning Theory. We begin with a precise definition of the problem and explore its key components: *probability distribution*, *loss function*, *expected risk*, and *hypothesis space*.

2.1.1 Setting

The fundamental components of supervised learning are:

(i) A probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and a random variable $Z = (X, Y) : \Omega \to \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is called the *input space*, $\mathcal{Y}$ the *output space*, and $\mathcal{X} \times \mathcal{Y}$ the *data space*. We denote the law of Z by $\rho \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$.

(ii) A measurable function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, \infty)$, called *loss function*.

Remark 2.1.1 The data space usually comes with a topology, and the corresponding Borel σ-algebra is considered. However, other choices are also possible and may be used.

It's worth mentioning that in the upcoming chapters, we'll be using the Borel σ-algebra in our interest.

For any measurable function $f : \mathcal{X} \to \mathcal{Y}$ we define the *expected risk* (or expected loss):

$$L(f) \;=\; \mathbb{E}_{(x,y)\sim\rho}\big[\ell\big(y, f(x)\big)\big] \;=\; \int_{\mathcal{X}\times\mathcal{Y}} \ell\big(y, f(x)\big)\, \mathrm{d}\rho(x, y).$$

The learning problem is then the minimization problem

$$\min_{f:\mathcal{X}\to\mathcal{Y}} L(f),$$

assuming that the probability distribution ρ is fixed but unknown, and the only available information about ρ is the finite data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, where we assume that these samples are independent and identically distributed from ρ.

It is evident that minimizing over all measurable functions $\{f : \mathcal{X} \to \mathcal{Y}\}$ is infeasible. Moreover, since the data is finite, we need to define a *data-driven* algorithm that selects f as a good approximation based on the samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and we have to establish a method to evaluate how closely the obtained solution approximates the ideal one.

Remark 2.1.2 In this chapter, the data space $\mathcal{X} \times \mathcal{Y}$ is assumed to be equipped with a fixed probability distribution ρ and each observation (x_i, y_i) is then drawn i.i.d. according to ρ. While the i.i.d. assumption is strong, since it implies that data points do not influence each other and are drawn from the same underlying distribution, it is the classical starting point for theoretical analysis.

Different options for the output space $\mathcal{Y}$ correspond to distinct types of learning problems. The most common choices in the supervised learning setting are:

- **Regression:** This corresponds to $\mathcal{Y} = \mathbb{R}$, while multivariate regression uses $\mathcal{Y} = \mathbb{R}^d$, $d \in \mathbb{N}$.
- **Classification:** Commonly, $\mathcal{Y} = \{-1, 1\}$ for the binary case and $\mathcal{Y} = \{1, 2, \ldots, m\}$ for multiclass classification (in this case with $m \in \mathbb{N}$ distinct categories).

More intricate outputs can also be considered, for instance, by combining the ones above through product spaces.

Let's examine a simple example of regression next.

Example 2.1.3 (Regression) Consider the case where $\mathcal{Y} = \mathbb{R}$. Suppose the joint distribution ρ on $\mathcal{X} \times \mathcal{Y}$ is governed by the relationship

$$y_i \;=\; f_*(x_i) \;+\; \eta\, \epsilon_i,$$

where f_* is an unknown function, $\eta > 0$ is a fixed noise level, and ϵ_i is an i.i.d. random variable with $\mathbb{E}[\epsilon_i] = 0$. A common example is $\epsilon_i \sim \mathcal{N}(0,1)$, implying that conditioned on x_i, each y_i is normally distributed around $f_*(x_i)$ with variance η^2.

In statistical learning, this scenario is often called *random design regression* when (x_i) itself is drawn randomly from a distribution on $\mathcal{X}$. Then each pair (x_i, y_i) is sampled i.i.d. from ρ. Our goal is to estimate the underlying function f_* from these noisy observations.

Remark 2.1.4 (Time-series) In many practical situations, however, the inputs $\{x_1, x_2, x_3, \dots\}$ are not merely independent points but *consecutive* observations in time or space. For instance, consider a time series $\{x_t\}_{t=1,2,\dots}$ describing the state of a system at each discrete time t. Under the classical regression assumption, we still treat $\{(x_t, y_t)\}$ as i.i.d. samples; however, this may be unrealistic in most cases. In Chapter 4, we will revisit regression tasks without the i.i.d. requirement on the data, such as Markov samples, where the underlying distribution is not fixed.

Example 2.1.5 (Classification) For $\mathcal{Y} = \{-1, +1\}$, each $\rho(\cdot \mid x)$ is a distribution on two labels:

$$\rho\big(y \mid x\big) \;=\; \big\{\, \rho\big(y = 1 \mid x\big),\, \rho\big(y = -1 \mid x\big)\big\}.$$

If $f : \mathcal{X} \to \mathbb{R}$ is a real-valued predictor, then the set

$$\{\, x \in \mathcal{X} : f(x) = 0\}$$

is called the *decision boundary*, since classification decisions are usually determined by the sign of $f(x)$. Points where $f(x) = 0$ mark the exact boundary between predicting $+1$ versus -1.

As seen in the example above, the predictor f takes values in $\mathbb{R}$ instead of the binary labeling set $\mathcal{Y} = \{-1, +1\}$. This distinction is often a useful choice for the loss function, which we can then write as $\ell : \mathcal{Y} \times \mathcal{Y}' \to [0, \infty)$, where $\mathcal{Y}'$ is the range of the predictor function f.

2.1.2 Target Function

The set of functions for which the expected error is well-defined is referred to as the *target space* and is denoted by $\mathcal{F}$. When the loss function is measurable with respect to both arguments, this target space corresponds to the collection of all measurable functions. The optimal solution to the learning problem is a function that minimizes the error, specifically:

$$\inf_{f \in \mathcal{F}} L(f). \tag{2.1}$$

Although achieving this infimum may not always be feasible, for many loss functions it is possible to explicitly identify a minimizer f_ρ, known as the *target function*, which satisfies:

$$L(f_\rho) = \min_{f \in \mathcal{F}} L(f).$$

Different choices of the loss function ℓ lead to different expected risks and, consequently, to different solutions (or approximations) for the minimization problem. In the context of regression tasks, which will be our central focus, the loss function is typically expressed as

$$\ell(y, f(x)) = V(y - f(x)),$$

where $V : \mathbb{R} \to [0, \infty)$ serves as a penalty function. This penalty quantifies the cost associated with deviations between the predicted value $f(x)$ and the true label y, assigning higher penalties to larger discrepancies.

Two widely used loss functions in regression are:

- *Square Loss:* $\ell(y,a) = (y-a)^2$,
- *Absolute Loss:* $\ell(y,a) = |y-a|$,

A useful way to analyze the target function and to ensure that the infimum in (2.1) is achieved is through the *inner risk*. Note that, under suitable assumptions ([32] for reference), we can decouple the integral as

$$L(f) \;=\; \int_{\mathcal{X}\times\mathcal{Y}} \ell\big(y, f(x)\big)\,\mathrm{d}\rho(x,y) \;=\; \int_{\mathcal{X}} \Big[\int_{\mathcal{Y}} \ell\big(y, f(x)\big)\,\mathrm{d}\rho\big(y\,|\,x\big)\Big]\mathrm{d}\rho_{\mathcal{X}}(x),$$

where $\rho_{\mathcal{X}}$ is the marginal distribution over $\mathcal{X}$, and $\rho(\cdot \mid x)$ is the conditional distribution given x. For each $x \in \mathcal{X}$, we define the *inner risk*

$$L_x(a) \;:=\; \int_{\mathcal{Y}} \ell\big(y,a\big)\,\mathrm{d}\rho\big(y\,|\,x\big), \quad a \in \mathbb{R}.$$

Then

$$L(f) \;=\; \int_{\mathcal{X}} L_x\big(f(x)\big)\,\mathrm{d}\rho_{\mathcal{X}}(x).$$

If for every x there is a real number a_x minimizing $L_x(a)$, then setting $f_\rho(x) = a_x$ yields

$$L_x\big(f_\rho(x)\big) \;=\; \min_{a\in\mathbb{R}} L_x(a), \quad \text{for almost all } x.$$

By integrating over x with respect to $\rho_{\mathcal{X}}$, it follows that f_ρ is indeed a global minimizer of L. In many standard losses (e.g., squared or absolute), one can solve for a_x explicitly, thereby characterizing the target function f_ρ.

Example 2.1.6 (Square Loss) A fundamental example of the target function can be seen when using the squared or absolute loss function.

Consider

$$\ell\big(y,a\big) \;=\; \big(y-a\big)^2.$$

For a fixed $x \in \mathcal{X}$, the inner risk is

$$L_x(a) \;=\; \int_{\mathcal{Y}} \big(y-a\big)^2\,\mathrm{d}\rho\big(y \mid x\big).$$

To find the minimizing a, one sets the derivative of $L_x(a)$ with respect to a to zero: $\frac{\mathrm{d}}{\mathrm{d}a}L_x(a) \;=\; -2\int \big(y-a\big)\,\mathrm{d}\rho(y \mid x) \;=\; 0$. Solving gives

$$a \;=\; \int y\,\mathrm{d}\rho(y \mid x),$$

the conditional mean of y given x. Consequently, the corresponding target function is

$$f_\rho(x) \;=\; \int y\,\mathrm{d}\rho\big(y \mid x\big).$$

This is called the *regression function* for the squared loss.

If we instead consider the absolute value loss, $\ell(y,a) = |y-a|$, one can similarly show that the optimal choice at each x is the *median* of the conditional distribution $\rho(\cdot \mid x)$.

Remark 2.1.7 As we observed, the selection of loss function directly influences the interpretation of the target function. Both the squared and absolute losses aim to estimate specific characteristics of the conditional distribution $\rho(y \mid x)$. However, the squared loss prioritizes smoothness and sensitivity to variations in the conditional mean, while the absolute loss focuses on robust predictions around the median.

2.1.3 Empirical Risk Minimization (ERM)

As we mentioned earlier, it is clear that directly solving the optimization problem

$$\min_{f:\mathcal{X}\to\mathcal{Y}} L(f)$$

using a finite amount of data is not feasible for two main reasons: the expected risk is defined as an expectation, which may be impossible to evaluate exactly, and searching over the entire space of all measurable functions is practically unmanageable.

A core idea in supervised learning is to replace the intractable expected risk $L(f)$ with a more manageable approximation $\widehat{L}(f)$, called the *empirical risk*, and to constrain the choice of functions f to a chosen subset of the target space of functions $\mathcal{H} \subseteq \mathcal{F} \subseteq \{f : \mathcal{X} \to \mathcal{Y} \text{ measurable}\}$, called the *hypothesis space*.

Specifically, we define the *empirical risk* as:

$$\widehat{L}(f) \;=\; \frac{1}{n}\sum_{i=1}^{n} \ell\big(y_i, f(x_i)\big), \tag{2.2}$$

where $\{(x_i, y_i)\}_{i=1}^n$ are the training samples, and consider the following constrained optimization problem:

$$\min_{f\,\in\,\mathcal{H}} \widehat{L}(f).$$

This approach, called *Empirical Risk Minimization* (ERM), is one of the simplest frameworks for designing learning algorithms.

Linear Least Squares

To illustrate the fundamental principles of ERM, we concentrate on the specific example of *linear least squares* regression, which serves as a foundation for extending these concepts to more advanced methods, such as kernel-based learning, as discussed later in the chapter.

Suppose $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and we adopt the *square loss* $\ell\big(y, f(x)\big) \;=\; \big(y - f(x)\big)^2$. Restricting f to *linear* functions of the form

$$f_w(x) \;=\; x^\top w \quad (\text{for some } w \in \mathbb{R}^d),$$

defines the hypothesis space

$$\mathcal{H} \;=\; \Big\{ f_w : f_w(x) = x^\top w,\; w \in \mathbb{R}^d \Big\}.$$

The empirical risk (2.2) then becomes the well-known *least squares* objective:

$$\widehat{L}(f_w) \;=\; \frac{1}{n}\sum_{i=1}^{n} \big(y_i - x_i^\top w\big)^2 \;=\; \frac{1}{n}\,\|\mathbf{y} - X\,w\|^2,$$

where $X \in \mathbb{R}^{n\times d}$ is the *data matrix* whose i-th row is $x_i^\top$, $\mathbf{y} \in \mathbb{R}^n$ is the vector of outputs $(y_1, \ldots, y_n)$, and $\|\cdot\|$ denotes the euclidean norm. Hence, we can rewrite ERM as,

$$\min_{w\in\mathbb{R}^d} \|\mathbf{y} - X\,w\|^2.$$

Remark 2.1.8 Note that, in this linear least squares setting, each function $f_w \in \mathcal{H}$ corresponds uniquely to the parameter vector $w \in \mathbb{R}^d$. This correspondence is a linear isomorphism between $\mathcal{H}$ and $\mathbb{R}^d$. Concretely,

$$\mathcal{H} \;=\; \big\{\, f_w : w \in \mathbb{R}^d \big\} \quad \longleftrightarrow \quad \mathbb{R}^d,$$

meaning that minimizing over all f in $\mathcal{H}$ is equivalent to minimizing over all $w \in \mathbb{R}^d$. As a result, we can interchangeably think of searching for the "best" linear function f_w or the "best" coefficient vector w. This perspective will remain consistent even when moving to more general (e.g., feature-mapped or kernel-based) settings.

To solve this minimization problem, we divide it into two cases:

Case $n \geq d$ (Under-Parameterized). When the number of samples n is at least as large as the dimension d, we speak of an *under-parameterized* regime. If X has full column rank, the least squares solution is unique, and since the empirical risk function is convex and differentiable, setting its gradient to zero yields the *normal equations*

$$X^\top X\, w \;=\; X^\top \mathbf{y},$$

leading to

$$\widehat{w} \;=\; \big(X^\top X\big)^{-1} X^\top \mathbf{y}.$$

In this scenario, $X^\top X$ is invertible, and $\widehat{w}$ is the unique minimizer of $\|\mathbf{y} - Xw\|^2$.

Case $n < d$ (Over-Parameterized). When the number of parameters d exceeds the sample size n, we are in an *over-parameterized* regime. If X has full row rank, infinitely many solutions can *perfectly* fit the data, i.e. $X\, w = \mathbf{y}$. A classical selection is the *minimal norm* solution

$$\widehat{w} \;:=\; \min_{\substack{w \in \mathbb{R}^d \\ X w = \mathbf{y}}} \|w\|.$$

Solving by Lagrange multipliers shows that

$$\widehat{w} \;=\; X^\top \big(X\, X^\top\big)^{-1} \mathbf{y}.$$

This choice has smallest $\|w\|$ among all exact fits to the training data.

Remark 2.1.9 Regardless of n vs. d, the least squares estimator can be summarized via the pseudoinverse $X^\dagger$:

$$\widehat{w} \;=\; X^\dagger\, \mathbf{y}.$$

When $n \geq d$, $X^\dagger = (X^\top X)^{-1} X^\top$; when $n < d$, $X^\dagger = X^\top (X X^\top)^{-1}$.

Remark 2.1.10 (Stability) If the matrix X has very small singular values (as identified by its singular value decomposition, SVD), small changes in the data vector $\mathbf{y}$ can result in large variations in the estimated solution $\widehat{w}$. This phenomenon, referred to as *instability*, often indicates poor generalization: a model that fits the training data well but is overly sensitive to noise or minor perturbations is unlikely to perform effectively on unseen data. From a linear algebra perspective, small singular values magnify errors, emphasizing the need for *regularization* techniques. For instance, *Tikhonov regularization*, which modifies the objective function to

$$\frac{1}{n}\|\mathbf{y} - Xw\|^2 \;+\; \lambda \|w\|^2,$$

penalizes large norm solutions and alleviates this issue by promoting more stable and robust models.

Moreover, regularization not only improves numerical stability but also helps in proving theoretical learning guarantees. Specifically, it allows for deriving bounds on the error of the algorithm by limiting the complexity of the model (through $\|w\|$ or similar norms). Without regularization (i.e., $\lambda = 0$), these guarantees could break down.

This completes our illustration of ERM in the linear least squares case. While linear models provide a simple and tractable foundation, many real-world tasks demand richer function spaces capable of capturing non-linear dependencies. A classical way to introduce such flexibility is by transforming the data into more complex representations, where linear methods can still operate effectively. We now turn to a brief introduction to the theory of *Reproducing Kernel Hilbert Spaces* (RKHS), which offers a powerful framework for formalizing these ideas, unifying the advantages of linear algorithms with the representational power needed for more complex problems.

2.2 Reproducing Kernel Hilbert Spaces (RKHS)

So far, we focused on *linear* learning models, where $f(x) = w^\top x$ with $w \in \mathbb{R}^d$, which are straightforward to handle but often too limited for complex, real-world data. In fact, we rarely expect intricate dependencies between inputs and outputs to be encapsulated by a linear relationship.

There are two strategies for overcoming linearity:

$$f(x) = \Phi\big(w^\top x\big) \quad \text{or} \quad f(x) = w^\top \Phi(x),$$

where Φ is a non-linear transformation.

The first form underlies *neural network* architectures, which apply non-linear activations to linear combinations of inputs. *Kernel methods*, in contrast, rely on the second approach: they map each x into a (potentially high- or even infinite-dimensional) feature space where linear methods can be applied, while still capturing non-linear effects in the original input space.

Consider a mapping $\Phi : \mathcal{X} \longrightarrow \mathbb{R}^p$, with $\Phi(x) = \big[\varphi_1(x), \ldots, \varphi_p(x)\big]^\top$, so that our predictor is

$$f(x) \;=\; w^\top \Phi(x) \;=\; \sum_{j=1}^{p} w_j\, \varphi_j(x).$$

In effect, the $\{\varphi_j\}$ serve as basis functions that allow us to represent a richer class of predictors while preserving the simplicity of linear parameterization in the transformed space.

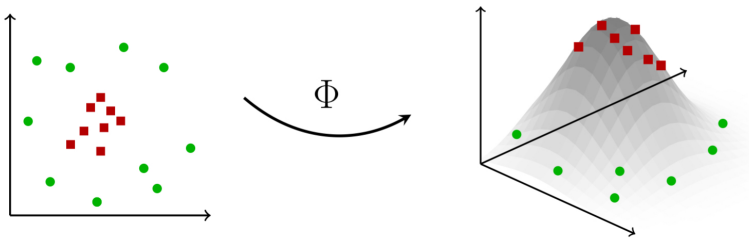

Figure 2.1: Data transformation through feature map

While any high-dimensional feature map can expand representational power, we will leverage a specific structure: the *Reproducing Kernel Hilbert Space* (RKHS). The properties of RKHS enable efficient computations and have profound implications for both theoretical analysis (e.g., generalization bounds, interpretability) and algorithmic design (e.g., the "kernel trick").

To develop a deeper understanding of kernel-based learning methods, we introduce the mathematical foundations of Reproducing Kernel Hilbert Spaces in this section.

2.2.1 Reproducing Kernels

Let (X, μ) be a Hausdorff space equipped with a positive finite Borel measure. We denote L^2_μ as the space of square-integrable functions $f : X \to \mathbb{R}$ with respect to the measure μ, i.e. $\left\{ f : X \to \mathbb{R} \mid f \text{ is measurable and } \int_X |f(x)|^2 \, d\mu(x) < \infty \right\}$, and denote its inner product as $\langle \cdot, \cdot \rangle_\mu = \langle \cdot, \cdot \rangle_{L^2_\mu}$ and the norm by $\| \cdot \|_\mu = \| \cdot \|_{L^2_\mu}$.

Definition 2.2.1 (RKHS)**.** A Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$ is called a **Reproducing Kernel Hilbert Space (RKHS)** if, for each $x \in \mathcal{X}$, the Dirac evaluation functional $\delta_x : f \mapsto f(x)$ is continuous (i.e. a bounded linear functional).

Recall Riesz's representation theorem, which plays a pivotal role in characterizing elements of an RKHS.

Theorem 2.2.2 (Riesz Representation Theorem)**.** *If Φ is a continuous linear functional on a Hilbert space $\mathcal{H}$, then there exists a unique $u \in \mathcal{H}$, called the **representer** of Φ, such that*

$$\Phi(f) = \langle f, u \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

In the specific case of Dirac evaluation functionals, for each δ_x, there exists a unique representer $u_x \in \mathcal{H}$ such that

$$\delta_x f = f(x) = \langle f, u_x \rangle_{\mathcal{H}}.$$

Definition 2.2.3 (Reproducing Kernel)**.** Let $\mathcal{H}$ be a Hilbert space of functions from (a non-empty set) $\mathcal{X}$ to $\mathbb{R}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **Reproducing Kernel** of $\mathcal{H}$ if it satisfies:

(i) $\forall x \in \mathcal{X}$, $k_x := K(\cdot, x) \in \mathcal{H}$,

(ii) $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, K(\cdot, x) \rangle_{\mathcal{H}} = f(x)$, this is referred to as the **reproducing property**.

Remark 2.2.4 Several properties of reproducing kernels follow directly from their definition:

- $k_x \in \mathcal{H}$ is a function from $\mathcal{X}$ to $\mathbb{R}$ such that $k_x(y) = K(x, y)$.
- For any $x, x' \in \mathcal{X}$:

$$K(x, x') = \langle K(\cdot, x), K(\cdot, x') \rangle_{\mathcal{H}} = \langle k_x, k_{x'} \rangle_{\mathcal{H}},$$

- K is symmetric, meaning:

$$K(x, y) = K(y, x) \quad \forall x, y \in \mathcal{X}.$$

From the remark, we can deduce that the Hilbert space $\mathcal{H}$ contains all functions of the form $f = \sum_{j=1}^{N} \alpha_j K(\cdot, x_j)$, where $x_j \in \mathcal{X}$.

We can write the norm of such functions as

$$\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{N}\sum_{i=1}^{N} \alpha_j \alpha_i \langle K(\cdot, x_j), K(\cdot, x_i)\rangle_{\mathcal{H}} = \sum_{j=1}^{N}\sum_{i=1}^{N} \alpha_j \alpha_i \langle k_{x_j}, k_{x_i}\rangle_{\mathcal{H}}.$$

Proposition 2.2.5. *If it exists, the reproducing kernel K for a Hilbert space $\mathcal{H}$ is unique.*

Proof. Assume that $\mathcal{H}$ has two reproducing kernels K_1 and K_2. Then:

$$\langle f, K_1(\cdot, x) - K_2(\cdot, x)\rangle = f(x) - f(x) = 0, \quad \forall f \in \mathcal{H}, \forall x \in \mathcal{X}.$$

In particular, if we take $f = K_1(\cdot, x) - K_2(\cdot, x)$, we obtain:

$$\|K_1(\cdot, x) - K_2(\cdot, x)\|_{\mathcal{H}}^2 = 0 \quad \forall x \in \mathcal{X},$$

implying $K_1 = K_2$. □

Proposition 2.2.6. *Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. Then the evaluation operators δ_x are continuous functionals if and only if $\mathcal{H}$ has a reproducing kernel K.*

Proof. First, assume that $\mathcal{H}$ is a Hilbert space with reproducing kernel K, then:

$$|\delta_x f| = |f(x)| = |\langle f, K(\cdot, x)\rangle| \leq \|K(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}}.$$

Hence, $\delta_x : \mathcal{H} \to \mathbb{R}$ is a bounded linear operator.

Conversely, assume that $\delta_x : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional. By the Riesz representation theorem, there exists a unique representer $u_x \in \mathcal{H}$ such that

$$\delta_x f = \langle f, u_x\rangle_{\mathcal{H}} \quad \forall f \in \mathcal{H}.$$

Define $K(\cdot, x) := u_x(\cdot)$ for all $x \in \mathcal{X}$. Then, it immediately follows:

$$\langle f, K(\cdot, x)\rangle_{\mathcal{H}} = \delta_x f = f(x).$$

Thus, K is a reproducing kernel for $\mathcal{H}$. □

Definition 2.2.7 (Positive Definite Kernel)**.** A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **positive definite kernel** if it is symmetric, i.e. $K(x, y) = K(y, x)$ for all $x, y \in \mathcal{X}$, and if for any $n \in \mathbb{N}$ and any choice of points $x_1, \ldots, x_n \in \mathcal{X}$, the $n \times n$ matrix $[K(x_i, x_j)]_{i,j=1}^n$ is positive semidefinite, i.e.

$$\sum_{i,j=1}^{n} c_i \, c_j \, K\big(x_i, x_j\big) \;\geq\; 0 \quad \text{for all vectors } (c_1, \ldots, c_n) \in \mathbb{R}^n.$$

We refer to $[K(x_i, x_j)]_{i,j=1}^n$ as the **kernel matrix**.

Theorem 2.2.8 (Moore-Aronszajn Theorem)**.** *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. Then there exists a unique (up to isomorphism) RKHS $\mathcal{H}_K \subset \mathbb{R}^{\mathcal{X}}$ with reproducing kernel K.*

Conversely, if K is the reproducing Kernel of an RKHS $\mathcal{H}$, then it is positive definite.

Proof. In the following we give an outline for the construction of $\mathcal{H}_K$. Given a symmetric, positive definite kernel K on $\mathcal{X}$, one can build the corresponding RKHS $\mathcal{H}_K$ in the following steps:

1. *Pre-Hilbert space setup.* Define

$$\mathcal{H}_0 \;=\; \Big\{ \sum_{i=1}^{n} \alpha_i \, K(\cdot, x_i) \;\Big|\; n \in \mathbb{N},\; \alpha_i \in \mathbb{R},\; x_i \in \mathcal{X} \Big\},$$

   the set of all finite linear combinations of kernel sections $K(\cdot, x)$.

2. *Inner product.* For $f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j K(\cdot, y_j)$ in $\mathcal{H}_0$, define

$$\langle f, g \rangle \;:=\; \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \, K(x_i, y_j).$$

3. *Completion.* Endow $\mathcal{H}_0$ with the norm induced by the inner product. Its completion is the Hilbert space $\mathcal{H}_K$. By construction, $\mathcal{H}_K$ satisfies the reproducing property $\langle f, K(\cdot, x) \rangle_{\mathcal{H}_K} = f(x)$. See [26] for details.

This procedure yields a unique RKHS whose kernel is precisely K.

Conversely, assume that K is the reproducing kernel of an RKHS $\mathcal{H}$. The symmetry of K follows directly from the symmetry of the inner product in $\mathcal{H}$:

$$K(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}} = \langle K_y, K_x \rangle_{\mathcal{H}} = K(y, x).$$

For any $n \in \mathbb{N}, x_1, \ldots, x_n \in X$, and $a_1, \ldots, a_n \in \mathbb{R}$, the positive definiteness of K is established by:

$$\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) = \left\| \sum_{i=1}^{n} a_i K_{x_i} \right\|_{\mathcal{H}}^2 \geq 0.$$

□

We will also denote the inner product of $\mathcal{H}_K$ as $\langle f, g \rangle_K := \langle f, g \rangle_{\mathcal{H}_K}$ and, similarly, the norm as $\|f\|_K := \|f\|_{\mathcal{H}_K}$.

Remark 2.2.9 The bilinear form

$$\Big\langle \sum_{i=1}^{n} \alpha_i \, K(\cdot, x_i),\; \sum_{j=1}^{m} \beta_j \, K(\cdot, y_j) \Big\rangle_K \;=\; \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j \, K(x_i, y_j)$$

is a well-defined inner product on the pre-Hilbert space $\mathcal{H}_0 = \operatorname{span}\{K(\cdot, x) \,:\, x \in \mathcal{X}\}$.

Remark 2.2.10 Convergence in $\|\cdot\|_K$-norm implies *pointwise* convergence, thanks to

$$|f_n(x) - f(x)| = \big|\langle f_n - f,\; K(\cdot, x) \rangle_K\big| \;\leq\; \|f_n - f\|_K \, \|K(\cdot, x)\|_K.$$

Hence each Cauchy sequence $\{f_n\} \subset \mathcal{H}_0$ converges pointwise to a limit f. By including all such limits, we obtain the RKHS $\mathcal{H}_K$ where K remains the reproducing kernel.

2.2.2 Mercer's Theorem and the Integral Operator Viewpoint

Thus far, we have seen how a positive definite kernel K defines an RKHS $\mathcal{H}_K$. There is an alternative characterization of K and $\mathcal{H}_K$ via the *integral operator*

$$T_K : L^2_\mu(X) \to L^2_\mu(X), \quad (T_K f)(x) = \int_X K(x,y)\, f(y)\, d\mu(y),$$

also denoted $T_{K,\mu}$ to make the dependency on the measure explicit. This perspective clarifies the connection between kernels and their spectral (eigenfunction) expansions, known as *Mercer's theorem*.

Remark 2.2.11 For T_K to be well-defined as an operator, one usually requires that K is continuous on $\mathcal{X} \times \mathcal{X}$ and satisfies an integrability condition such as $Tr(K) = \iint K^2(x,x')\, \mathrm{d}\mu(x)\, \mathrm{d}\mu(x') < \infty$. This condition is sometimes referred to as *finite trace*, ensuring that $K(\cdot, x) \in L^2_\mu(\mathcal{X})$ for each x. If K is not of finite trace over all of $\mathcal{X}$, one can restrict to a suitable sub-domain where this property holds. In this way the image of T_K in $\mathcal{H}_K$ can be regarded in $L^2_\mu(\mathcal{X})$ by composition with the inclusion $\mathcal{H}_K \hookrightarrow L^2_\mu(\mathcal{X})$.

Now recall the definition of eigenfunction and eigenvalue for a linear operator.

Definition 2.2.12 (Eigenfunction of T_K)**.** A function $\Phi \in L^2_\mu(\mathcal{X})$ is called an **eigenfunction** of T_K if there exists $\lambda \in \mathbb{R}$ such that

$$(T_K \Phi)(x) = \int_X K(x,y)\, \Phi(y)\, d\mu(y) = \lambda\, \Phi(x), \quad \forall\, x \in \mathcal{X}.$$

In this case, λ is the corresponding **eigenvalue**.

Theorem 2.2.13 (Mercer's Theorem)**.** *Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous, positive definite kernel, with $\mathcal{X}$ being compact. Then there exists an orthonormal set of eigenfunctions $\{\Phi_i\}_{i=1}^\infty \subset L^2_\mu(\mathcal{X})$ of T_K, with corresponding non-negative eigenvalues $\{\lambda_i\}_{i=1}^\infty$ that accumulate only at 0, such that*

$$K(x,y) = \sum_{i=1}^\infty \lambda_i\, \Phi_i(x)\, \Phi_i(y),$$

and this series converges uniformly on $\mathcal{X} \times \mathcal{X}$.

Moreover, the RKHS associated with K can be characterized via these eigenfunctions:

$$\mathcal{H}_K = \Big\{ f \in L^2_\mu(\mathcal{X}) \;\Big|\; \sum_{i=1}^\infty \frac{\langle f, \Phi_i\rangle_\mu^2}{\lambda_i} < \infty \Big\}, \quad \text{with inner product } \langle f, g\rangle_K = \sum_{i=1}^\infty \frac{\langle f, \Phi_i\rangle_\mu \langle g, \Phi_i\rangle_\mu}{\lambda_i}.$$

Sketch of proof. Under the stated assumptions, T_K is indeed a compact, self-adjoint, and positive operator on $L^2_\mu(\mathcal{X})$. The spectral theorem for such an operator ensures there is an orthonormal basis of eigenfunctions Φ_i in $L^2_\mu(\mathcal{X})$ with eigenvalues $\lambda_i \geq 0$ converging to 0. One can then expand

$$K(x,y) = \sum_{i=1}^\infty \lambda_i\, \Phi_i(x)\, \Phi_i(y),$$

where the series converges uniformly by standard results on compact operators with continuous kernels. For the RKHS characterization, because each Φ_i is continuous and orthonormal in $L^2_\mu(\mathcal{X})$, we can represent $f \in \mathcal{H}_K$ as

$$f = \sum_{i=1}^\infty c_i\, \Phi_i,$$

subject to the condition $\sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i} < \infty$. The inner product then becomes

$$\langle f, g \rangle_{\mathcal{H}_K} \;=\; \sum_{i=1}^{\infty} \frac{c_i \, d_i}{\lambda_i}$$

for $f = \sum_i c_i \, \Phi_i$ and $g = \sum_i d_i \, \Phi_i$. For details, see [26]. □

Remark 2.2.14 (Compactness Assumption) Classically, Mercer's Theorem assumes that $\mathcal{X}$ is compact and K is continuous on $\mathcal{X} \times \mathcal{X}$. Under these conditions, the operator T_K becomes compact, and one obtains the uniform convergence of the eigenfunction expansion. Moreover, it is well known [6] that $T_K : L^2_\mu(\mathcal{X}) \to L^2_\mu(\mathcal{X})$ is a compact, self-adjoint, positive operator, with $\|T_K\| \leq \kappa^2$, where $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x,x)}$.

In some variants, one can replace or relax the compactness requirement by imposing other conditions. For example, *assuming* that $K(x,x)$ is uniformly bounded (i.e. $\kappa < \infty$) and K satisfies suitable continuity or integrability assumptions relative to the measure μ (eg. sigma-finiteness), then T_K can still be treated as a compact operator on a restricted sub-domain. The key idea is that one needs both boundedness of K (to ensure the images $K(\cdot, x) \in L^2_\mu(\mathcal{X})$) and appropriate continuity properties (to secure the spectral decomposition). In practical settings, these assumptions ensure that K admits the same eigenfunction-based representation without requiring $\mathcal{X}$ to be compact in the strict topological sense.

A useful way to see the integral operator is through the *sampling operator* S_x. For each $x \in \mathcal{X}$, define $S_x = \langle K_x, \, \cdot \, \rangle_K : \mathcal{H}_K \to \mathbb{R}$ by $S_x(f) = \langle K_x, f \rangle_K = f(x)$, which is linear, and let S_x^* be its adjoint, so $S_x^* : \mathbb{R} \to \mathcal{H}_K$ sends $c \mapsto c \, K_x$. Then $S_x^* S_x : \mathcal{H}_K \to \mathcal{H}_K$ is a rank-one positive operator given by $(S_x^* S_x)(f) = f(x) \, K_x$. Taking expectation with respect to μ, we obtain that for $f \in \mathcal{H}_K$,

$$T_{K,\mu} f(x) \;=\; \int_X K(x,y) f(y) \, d\mu(y) \;=\; \int_X (S_y^* S_y f)(x) \, d\mu(y).$$

In this sense, the restriction $T_{K,\mu}|_{\mathcal{H}_K} : \mathcal{H}_K \to \mathcal{H}_K$ can be expressed as

$$T_{K,\mu}|_{\mathcal{H}_K} f = \mathbb{E}_{y \sim \mu}\big[S_y^* S_y f\big],$$

which is usually known as the *covariance operator* of the measure μ in $\mathcal{H}_K$. From now on we are going to denote the operator and its restriction simply as T_K or $T_{K,\mu}$ abusing notation.

Definition 2.2.15 (Features)**.** The Mercer theorem (2.2.13) ensures the existence of a mapping $\Phi : \mathcal{X} \to \mathcal{H}$, such that

$$K(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}},$$

for every positive definite kernel K.

The map Φ is called the **feature map**, and $\mathcal{H}$ is referred to as the **feature space**.

Remark 2.2.16 Since $K(\cdot, x) = \sum_{i=1}^{\infty} \lambda_i \, \Phi_i(\cdot) \, \Phi_i(x)$, we have for any $f \in \mathcal{H}_K$,

$$\langle f, K(\cdot, x) \rangle_K \;=\; \sum_{i=1}^{\infty} \frac{c_i \, \big(\lambda_i \, \Phi_i(x)\big)}{\lambda_i} \;=\; \sum_{i=1}^{\infty} c_i \, \Phi_i(x) \;=\; f(x),$$

thus showing again that K reproduces $f(x)$ via its eigenbasis.

Remark 2.2.17 Mercer's theorem provides a link between the kernel K viewed as an inner product in the RKHS, and the eigenfunction expansion of the compact operator T_K. In analogy to linear algebra, where we diagonalize a matrix to reveal its basis of eigenvectors, here we diagonalize the integral operator T_K to obtain an orthonormal basis $\{\Phi_i\}$ of eigenfunctions in $L^2_\mu(\mathcal{X})$, with corresponding nonnegative eigenvalues $\{\lambda_i\}$.

We will exploit this for defining a regularity condition on our target function, know as *source condition*.

Definition 2.2.18 (Source condition)**.** Given a σ-finite measure μ and a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we say that f_ρ satisfies the **regularity condition of order r**, with $r > \frac{1}{2}$, if

$$f_\rho = T^r_{K,\mu} g, \tag{2.3}$$

for some $g \in L^2_\mu(\mathcal{X})$. Meaning that

$$f_\rho = \sum_{i=1}^{\infty} \lambda^r g_i \Phi_i,$$

where Φ_i, λ_i are the eigenfunctions and eigenvalues of $T_{K,\mu}$ respectively, and $g_i = \langle g, \Phi_i \rangle_\mu$.

Remark 2.2.19 The operator $T_{K,\mu}$ being positive, self-adjoint, and compact on $L^2_\mu(\mathcal{X})$ guarantees we can consider its fractional powers $T^r_{K,\mu}$ via the usual spectral theory. Also note that $T^{1/2}_{K,\mu}(L^2_\mu(\mathcal{X})) = \mathcal{H}_K$, since for $r = 1/2$ the operator is an isometric isomorphism and that $0 < a < b$ implies $T^b_{K,\mu}(L^2_\mu(\mathcal{X})) \subseteq T^a_{K,\mu}(L^2_\mu(\mathcal{X}))$. Therefore $\hat{f} \in \mathcal{H}_K$ for $r \geq 1/2$.

To understand this inclusion, we examine the spectral decomposition of T_K and analyze how the RKHS norm of $\hat{f}$ depends on r. Let $T_{K,\mu}\, \Phi_i \;=\; \lambda_i\, \Phi_i$, with $\lambda_i > 0$, $\lambda_i \to 0$ and $\{\Phi_i\}_{i=1}^{\infty}$ an orthonormal basis in $L^2_\mu(\mathcal{X})$.

Any $g \in L^2_\mu(\mathcal{X})$ can be expanded as

$$g \;=\; \sum_{i=1}^{\infty} g_i\, \Phi_i, \quad g_i = \langle g, \Phi_i \rangle_\mu.$$

Then

$$f_\rho \;=\; T^r_{K,\mu}\, g \;=\; \sum_{i=1}^{\infty} \lambda_i^r\, g_i\, \Phi_i.$$

Since $\langle f_\rho, \Phi_i \rangle_\mu = \lambda_i^r\, g_i$, we obtain

$$\|f_\rho\|^2_K \;=\; \sum_{i=1}^{\infty} \lambda_i^{-1} \left(\lambda_i^r\, g_i\right)^2 \;=\; \sum_{i=1}^{\infty} \lambda_i^{2r-1}\, g_i^2.$$

Whether this series converges depends on $2r - 1$ and the decay of λ_i. If $r > \frac{1}{2}$, then $2r - 1 > 0$ and $\lambda_i^{2r-1} \to 0$ as $i \to \infty$, so the series converges (since $g \in L^2_\mu(\mathcal{X})$ implies $\sum_{i=1}^{\infty} |g_i|^2 < \infty$). Hence, $\|f_\rho\|^2_K < \infty$, thus $f_\rho \in \mathcal{H}_K$. Conversely, if $r < \frac{1}{2}$, one has $2r - 1 < 0$, and for large i the term λ_i^{2r-1} blows up, causing $\|f_\rho\|^2_K$ to diverge. In that case, $f_\rho \notin \mathcal{H}_K$.

2.3 Kernel Methods for Learning

Now that we laid out the mathematical groundwork of RKHS, we can expand our discussion to include learning with non-linear functions through feature maps and kernels. We begin by revisiting some aspects of least squares with linear functions and examine how to incorporate non-linearity using kernels in this framework. This leads us to discuss recursive least squares and stochastic gradient descent algorithms.

Let us start by revisiting the linear least squares problem presented in the first section. The goal is to find the linear function $f(x) = x^\top w$ that minimizes the empirical risk, or equivalently, finding the weight vector w that minimizes the objective, as

$$\hat{w}^\lambda = \arg\min_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(y_i - w^\top x_i)^2 + \lambda\|w\|^2, \quad \lambda \geq 0.$$

The solution to this problem can be expressed as:

$$\hat{w}^\lambda = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + n\lambda I)^{-1}\mathbf{y},$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the input matrix and $\mathbf{y} \in \mathbb{R}^n$ is the output vector. To simplify the interpretation of the solution, we can rewrite it as:

$$\hat{w}^\lambda = \mathbf{X}^\top c = \sum_{i=1}^{n} x_i c_i, \quad \text{where} \quad c = (\mathbf{X}\mathbf{X}^\top + n\lambda I)^{-1}\mathbf{y}.$$

This leads to the equivalent representation:

$$\hat{f}^\lambda(x) = \sum_{i=1}^{n} x^\top x_i c_i,$$

where the coefficients c_i depend on the solution of the linear system.

As we discussed in the previous section, our approach will be similar, but instead of using linear functions, we will consider functions of the form

$$f(x) = w^\top \Phi(x),$$

where Φ is a feature map.

A simple example of a feature map is given by monomials.

Example 2.3.1 Let $\mathcal{X} = \mathbb{R}$, then one can consider the *polynomial feature map* (of degree p)

$$\Phi : \mathbb{R} \to \mathbb{R}^p, \quad \Phi(x) = (x, x^2, x^3, \ldots, x^p)^\top.$$

Another example is the quadratic feature map:

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3, \quad (x_1, x_2) \mapsto (x_1^2, \sqrt{2}x_1x_2, x_2^2).$$

See Figure 2.1 for reference. With an appropriate choice of the mapping Φ, data in higher-dimensional feature spaces can become linearly separable by a hyperplane.

2.3.1 Learning in the Feature Space

Suppose each $x \in \mathcal{X} \subseteq \mathbb{R}^d$ is mapped via

$$\Phi : \mathcal{X} \longrightarrow \mathbb{R}^p, \quad \Phi(x) = [\varphi_1(x), \ldots, \varphi_p(x)]^\top,$$

and we restrict our hypothesis space to

$$\mathcal{H}_\Phi := \left\{ f \;\middle|\; f(x) = w^\top \Phi(x),\ w \in \mathbb{R}^p \right\}.$$

As before, let $\{(x_i, y_i)\}_{i=1}^n$ be the training samples, and consider an empirical risk of the form

$$\widehat{L}_\lambda(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}_\Phi}^2,$$

where ℓ might be the squared loss, and $\|f\|_{\mathcal{H}_\Phi} = \|w\|$ if $f(x) = w^\top \Phi(x)$.

Definition 2.3.2 (Feature matrix)**.** We define the *feature matrix*

$$\widehat{\Phi} \in \mathbb{R}^{n \times p} \quad \text{with} \quad (\widehat{\Phi})_{ij} = \varphi_j(x_i),$$

and let $y \in \mathbb{R}^n$ denote the output vector $(y_1, \ldots, y_n)$.

For the squared-loss case we can write

$$\widehat{L}_\lambda(f_w) = \frac{1}{n} \|\mathbf{y} - \widehat{\Phi}\, w\|^2 + \lambda \|w\|^2.$$

Similarly as before, minimizing over $w \in \mathbb{R}^p$, yields the minimizer

$$\widehat{w}^\lambda = \widehat{\Phi}^\top \left(\widehat{\Phi}\widehat{\Phi}^\top + \lambda I \right)^{-1} \mathbf{y},$$

where we have scaled λ by n inside the matrix inversion, as it is a common notational convention. Setting

$$c = \left(\widehat{\Phi}\widehat{\Phi}^\top + \lambda I \right)^{-1} \mathbf{y} \ \in \mathbb{R}^n, \quad \text{where} \quad (\widehat{\Phi}\widehat{\Phi}^\top)_{i,j} = \widehat{\Phi}(x_i)^\top \widehat{\Phi}(x_j),$$

we can write

$$\widehat{w}^\lambda = \widehat{\Phi}^\top c, \quad = \sum_{i=1}^n \Phi(x_i) c_i.$$

So for each new point x,

$$\widehat{f}^\lambda(x) = (\widehat{w}^\lambda)^\top \Phi(x) = \sum_{i=1}^n c_i \left(\Phi(x_i)^\top \Phi(x) \right).$$

Hence $\widehat{f}^\lambda(x)$ is expressed as a linear combination of the *inner products* $\Phi(x_i)^\top \Phi(x)$. In fact,

$$\widehat{f}^\lambda(x) = \sum_{i=1}^n c_i \langle \Phi(x_i), \Phi(x) \rangle \quad \text{where} \quad c = (\widehat{\Phi}\widehat{\Phi}^\top + \lambda I)^{-1} \mathbf{y}.$$

Remark 2.3.3 (Kernel Trick) Notice that the coefficient vector c and the function values $\widehat{f}^{\lambda}(x)$ never require explicit knowledge of each component φ_j. Instead, all expressions rely on the dot products $\Phi(x_i)^\top \Phi(x_j)$. With this approach, we can replace $\mathcal{H}_\Phi$ with an RKHS and $\Phi(x_i)^\top \Phi(x_j)$ by a kernel

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle,$$

which directly computes the dot product in the (potentially large or infinite-dimensional) feature space. Thus, as it follows from the discussion on RKHS, once we express the solution $\widehat{f}^{\lambda}(x)$ in terms of inner products $\Phi(x_i)^\top \Phi(x)$, we no longer need the explicit feature map Φ.

This insight is commonly referred to as the *kernel trick*, and the procedure by which algorithms can be performed by replacing inner products with a kernel is often called *kernelizing*. Consequently, one can operate in an implicit feature space without explicitly constructing Φ.

Example 2.3.4 (Kernelizing Regularized LLS) Recall the solution in the feature space:

$$\widehat{f}^{\lambda}(x) = \sum_{i=1}^{n} c_i \langle \Phi(x_i), \Phi(x) \rangle, \quad \text{where} \quad c = (\widehat{\Phi}\widehat{\Phi}^\top + \lambda I)^{-1} \mathbf{y}.$$

If we let $K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$, then

$$\widehat{f}^{\lambda}(x) = \sum_{i=1}^{n} c_i K(x_i, x), \quad \text{and} \quad c = (\widehat{K} + \lambda I)^{-1} \mathbf{y},$$

where $\widehat{K} \in \mathbb{R}^{n \times n}$ is the kernel matrix of the training data with $\widehat{K}_{ij} = K(x_i, x_j)$. In particular, $\widehat{f}^{\lambda}$ is an empirical approximation of the (regularized) minimizer of the expected risk, which we can express using the integral operator T_K as

$$f^{\lambda} \;=\; (T_K + \lambda I)^{-1} \, T_K \, f_\rho.$$

A proof of this fact can be found in [26].

This exemplifies a broader principle: solutions to regularized risk minimization in RKHS can always be expressed as linear combinations of kernel evaluations at training points. The *Representer Theorem* formalizes this observation.

Theorem 2.3.5 (Representer Theorem)**.** *Let $\mathcal{H}_K$ be an RKHS with kernel K, and consider the regularized empirical risk minimization problem:*

$$\min_{f \in \mathcal{H}_K} \left\{ \widehat{L}_\lambda(f) \right\}, \quad \widehat{L}_\lambda(f) = \frac{1}{n} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2 + \lambda \|f\|_K^2, \quad \lambda > 0.$$

Any minimizer $\widehat{f}^{\lambda} \in \mathcal{H}_K$ admits a representation:

$$\widehat{f}^{\lambda}(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i), \quad \alpha_i \in \mathbb{R},$$

where the coefficients $\alpha = (\alpha_1, \ldots, \alpha_n)^\top$ solve the finite-dimensional system $(\widehat{K} + \lambda I)\alpha = \mathbf{y}$.

Proof Sketch. Let $\overline{\mathcal{H}} \subset \mathcal{H}_K$ be the subspace spanned by $\{K_{x_i}\}_{i=1}^n$. Decompose $f \in \mathcal{H}_K$ as $f = \overline{f} + f^\perp$, where $\overline{f} \in \overline{\mathcal{H}}$ and $f^\perp \in \overline{\mathcal{H}}^\perp$. By the reproducing property:

$$f^\perp(x_i) = \langle f^\perp, K_{x_i} \rangle_K = 0 \quad \forall i = 1, \ldots, n.$$

Thus, the empirical risk depends only on $\overline{f}$, while $\|f\|_K^2 = \|\overline{f}\|_K^2 + \|f^\perp\|_K^2$. Minimizing $\widehat{L}_\lambda(f)$ forces $\|f^\perp\|_K^2 = 0$, so $f = \overline{f}$. □

Remark 2.3.6 The theorem guarantees that even if $\mathcal{H}_K$ is infinite-dimensional, solutions lie in the n-dimensional subspace spanned by $\{K_{x_i}\}$. This justifies parameterizing f as $f = \sum_{i=1}^n \alpha_i K_{x_i}$ during optimization, avoiding explicit feature maps Φ.

Example 2.3.7 (Polynomial Kernel) Consider the map

$$K(x, z) = (\langle x, z \rangle + 1)^p,$$

with $p \in \mathbb{N}$ and $x, z \in \mathbb{R}^d$. We claim that K can be written in the form

$$K(x, z) = \Phi(x)^\top \Phi(z),$$

for a suitably chosen (finite-dimensional) feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^N$.

Let us first examine the case $x, z \in \mathbb{R}$ for gaining some intuition. Then we can write

$$K(x, z) = (xz + 1)^p = \sum_{k=0}^{p} \binom{p}{k} (xz)^k = \sum_{k=0}^{p} \binom{p}{k} x^k z^k.$$

We can group coefficients expanding the binomial coefficients and define

$$\Phi(x) := \left(\sqrt{\binom{p}{0}}, \sqrt{\binom{p}{1}} x, \sqrt{\binom{p}{2}} x^2, \ldots, \sqrt{\binom{p}{p}} x^p \right)^\top \in \mathbb{R}^{p+1}.$$

A direct check shows $\Phi(x)^\top \Phi(z) = (xz + 1)^p$. Thus

$$K(x, z) = \langle \Phi(x), \Phi(z) \rangle \quad \text{for all } x, z \in \mathbb{R}.$$

For $x, z \in \mathbb{R}^d$, we expand similarly by multinomial coefficients. Each monomial term $\langle x, z \rangle^k$ can be broken down into sums of products of coordinates $x_i z_i$. The resulting finite-dimensional feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^N$ picks out all degree-p monomials in the coordinates of x, up to appropriate constant factors. In particular, each coordinate of $\Phi(x)$ has the form $\sqrt{C_\alpha}(x_1^{\alpha_1} \cdots x_d^{\alpha_d})$ for some multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ with $\|\alpha\|_1 = \alpha_1 + \cdots + \alpha_d \leq p$. Thus, the feature space dimension N is $\binom{d+p}{p}$.

Hence K is a *positive definite kernel*, called the *polynomial kernel of degree p*.

Example 2.3.8 Consider the kernel function:

$$K(x, z) = \frac{1}{1 - \alpha^2 \langle x, z \rangle},$$

where $\alpha^2 \langle x, z \rangle < 1$. This can be expanded as a geometric series:

$$\frac{1}{1 - \alpha^2 \langle x, z \rangle} = \sum_{s=0}^{\infty} (\alpha^2 \langle x, z \rangle)^s.$$

In the scalar case ($d = 1$), this becomes:

$$\frac{1}{1-\alpha^2 xz} = \sum_{s=0}^{\infty} (\alpha^2 xz)^s = \Phi(x)^\top \Phi(z),$$

where:

$$\Phi(x) = (1, \alpha x, \alpha^2 x^2, \alpha^3 x^3, \ldots)^\top.$$

Despite $\Phi(x)$ becoming an infinite-dimensional vector, the kernel can be computed efficiently given α and $\langle x, z\rangle$.

For $d > 1$, an analogous construction exists, where $K(x, z)$ can be interpreted as enumerating all monomials x^s with weights α^s.

Remark 2.3.9 (Taking $p \to \infty$) One can view certain kernels (like the above geometric series) as the limit $p \to \infty$ of polynomial expansions. The previous example illustrates how one can systematically transform "basic" data coordinates into higher-order monomials, enabling linear algorithms to fit more complex relationships. Although the explicit feature map may be large (or even infinite) in general, the kernel function

$$K(x, z) = \langle \Phi(x), \Phi(z)\rangle$$

can be computed directly, rather than writing features explicitly, and efficiently, making such expansions tractable in many learning applications.

Example 2.3.10 Another famous example of an infinite-dimensional feature expansion is given by the *Gaussian kernel*. For $x, z \in \mathbb{R}^d$ and $\gamma > 0$, define

$$K(x, z) = \exp(-\gamma \|x - z\|^2).$$

Below we illustrate how K can be expanded as an infinite sum of monomials, implying that the kernel corresponds to an infinite-dimensional feature map.

Assume $d = 1$ for simplicity and let $z, x \in \mathbb{R}$. Observe that

$$K(x, z) = \exp(-\gamma(x - z)^2) = \exp(-\gamma z^2) \exp(-\gamma x^2) \exp(2\gamma xz).$$

The factor $\exp(2\gamma xz)$ has the power series expansion

$$\exp(2\gamma xz) = \sum_{n=0}^{\infty} \frac{(2\gamma xz)^n}{n!},$$

hence

$$\exp(-\gamma(x - z)^2) = \exp(-\gamma x^2) \exp(-\gamma z^2) \sum_{n=0}^{\infty} \frac{(2\gamma xz)^n}{n!}.$$

Rearranging each term, we see that

$$K(x, z) = \sum_{n=0}^{\infty} \sqrt{\frac{(2\gamma)^n}{n!}} x^n \exp\left(-\frac{\gamma x^2}{2}\right) \sqrt{\frac{(2\gamma)^n}{n!}} z^n \exp\left(-\frac{\gamma z^2}{2}\right).$$

Thus, we may define an infinite-dimensional feature map

$$\Phi(x) = (\Phi_0(x), \Phi_1(x), \Phi_2(x), \ldots)^\top \quad \text{with} \quad \Phi_n(x) = \sqrt{\frac{(2\gamma)^n}{n!}} x^n \exp\left(-\frac{\gamma x^2}{2}\right),$$

so that $K(x, z) = \langle \Phi(x), \Phi(z)\rangle$.

2.3.2 Online Learning

Up to this point, we have discussed *batch (or offline)* learning algorithms, where the entire sample set $\{(x_i, y_i)\}_{i=1}^n$ is given from the start and we may use it all at once (e.g. to minimize an ERM objective). In many real-world scenarios, however, data can arrive sequentially and potentially unbounded in size, making batch learning not advantageous for computations or even impossible. This prompts the study of *online (or incremental) learning*, which processes examples sequentially, one at a time, and updates the hypothesis on the fly.

Regression in Online Learning

We again consider the regression problem from the previous section, but instead of receiving all n data points at once, we observe a sequence of i.i.d. random examples

$$\{z_t\}_{t=1}^{\infty}, \quad z_t = (x_t, y_t) \in \mathcal{X} \times \mathcal{Y},$$

each drawn according to a probability measure ρ. The goal remains to approximate the regression function

$$f_\rho : x \mapsto \int_{\mathcal{Y}} y \, d\rho_{(y|x)},$$

by minimizing the mean squared error,

$$L(f) = \mathbb{E}\big[(f(x) - y)^2\big] = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y)^2 d\rho(x, y),$$

or a regularized version of it, e.g.

$$L_\lambda(f) = L(f) + \lambda \|f\|_K^2 = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y)^2 d\rho(x, y) + \lambda \|f\|_K^2. \tag{2.4}$$

Rather than re-running a batch procedure each time we get a new example, a map T_t updates its current hypothesis f_{t-1} to f_t upon seeing (x_t, y_t). We write

$$f_t = T_t\big(f_{t-1}, x_t, y_t\big),$$

where T_t is an *update map* $T_t : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathcal{H}$. We aim for $f_t \to f_\rho$ in some sense (e.g. in $\mathcal{H}$ or $L^2_{\rho_\mathcal{X}}$-norm). This procedure, in general, is referred to as an *online learning algorithm (OLA)*.

Remark 2.3.11 As each data point is processed in arrival order, the hypothesis can be improved (or at least adapted) at each step, which can be crucial for large-scale or streaming data. Moreover, the computational overhead for each update is typically smaller than a full batch solve, at the expense of possibly more "noisy" updates.

Stochastic Gradient Descent in RKHS

Next, let us explore one specific instance of OLA. Let the hypothesis space $\mathcal{H}_K$ be the RKHS induced by a positive-definite kernel K on $\mathcal{X} \times \mathcal{X}$. Assume there exists constant $\kappa \geq 0$ such that $\kappa := \sup_{x\in\mathcal{X}} \sqrt{K(x, x)} < \infty$, and $M_\rho \geq 0$ such that $\mathrm{supp}(\rho) \subseteq \mathcal{X} \times [-M_\rho, M_\rho]$. We focus on an online algorithm for the squared-loss objective with an RKHS-regularization term (i.e., $L_\lambda(f) = L(f) + \lambda\|f\|_K^2$).

Given the t-th example (x_t, y_t), we update from $f_{t-1} \in \mathcal{H}_K$ to f_t by

$$f_t \;=\; f_{t-1} \;-\; \gamma_t \Big[\big(f_{t-1}(x_t) - y_t\big)\, K_{x_t} + \lambda_t\, f_{t-1}\Big], \tag{2.5}$$

where:

- $(\gamma_t)_{t\in\mathbb{N}}$ is a sequence of positive reals called *step-size (or learning-rate)*,
- $(\lambda_t)_{t\in\mathbb{N}}$ is a sequence of non-negative *regularization-parameters (or gains)*,
- $K_{x_t} := K(\cdot, x_t)$,
- $f_0 \in \mathcal{H}_K$ is an initial guess (often $f_0 = 0$).

Algorithms of this type are referred to as *stochastic gradient descent (SGD)*-type algorithms. In effect, at each iteration, we approximate the gradient of the (regularized) expected risk L_λ, using a single "sample gradient":

$$\nabla L_\lambda(f_{t-1}) \;\approx\; \big(f_{t-1}(x_t) - y_t\big)\, K_{x_t} + \lambda_t\, f_{t-1}.$$

Remark 2.3.12 (Measurability) Let $\mathcal{F} = (\mathcal{F}_t)_{t\in\mathbb{N}_0}$ be the filtration generated by the data, where

$$\mathcal{F}_t = \sigma\{(x_i, y_i) : 1 \le i \le t\}.$$

Here, $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ denotes the conditional expectation with respect to $\mathcal{F}_t$. Note that $(f_t)_{t\in\mathbb{N}}$ is an $\mathcal{F}_t$-adapted stochastic process taking values in the RKHS $\mathcal{H}_K$. Recall that the adaptation to $\mathcal{F}_t$ ensures that f_t depends only on the data observed up to time t.

Remark 2.3.13 (SGD) To understand why the update (2.5) can be viewed as a (stochastic) gradient descent procedure, we first recall how to define a gradient in a Hilbert space.

Let $\mathcal{H}$ be a real Hilbert space, and let $V : \mathcal{H} \to \mathbb{R}$ be a Fréchet-differentiable functional. Then the *gradient* $\nabla V(f) \in \mathcal{H}$ of V at f is the unique element in $\mathcal{H}$ such that, for all $g \in \mathcal{H}$,

$$\langle \nabla V(f),\, g\rangle_{\mathcal{H}} \;=\; DV(f)[\,g\,],$$

where $DV(f)[\,g\,]$ is the directional (or Fréchet) derivative of V at f in the direction g. In the context of (2.5), each data point $z = (x, y)$ defines a local objective

$$V_z(f) \;:=\; \tfrac{1}{2}\Big(\big(f(x) - y\big)^2 \;+\; \lambda\,\|f\|_K^2\Big),$$

where $f \in \mathcal{H}_K$ is a function in the RKHS. We claim that

$$\nabla V_z(f) \;=\; \big(f(x) - y\big)\, K_x + \lambda\, f. \tag{2.6}$$

Indeed, for $g \in \mathcal{H}_K$, we can check the directional derivative at f, which is the linear functional $DV_z(f) : \mathcal{H}_K \to \mathbb{R}$ such that for $g \in \mathcal{H}_K$,

$$\lim_{\|g\|_K \to 0} \frac{|V_z(f+g) - V_z(f) - DV_z(f)(g)|}{\|g\|_K} = 0.$$

By computing the limit and using the reproducing property of the kernel we get

$$DV(f)[\,g\,] = (f(x) - y)g(x) + \lambda\langle f, g\rangle_K = \langle (f(x) - y)K_x + \lambda f, g\rangle_K,$$

which proves (2.6).

Hence, in the case of $z = z_t, f = f_t$, the update (2.5) becomes

$$f_{t+1} \;=\; f_t \;-\; \gamma_t\, \nabla V_{z_t}(f_t),$$

showing that at step t, we descend in the negative gradient of V_{z_t} taken at f_t.

Because $z_t = (x_t, y_t)$ are drawn randomly (i.i.d. from the underlying distribution ρ), we can see the gradient

$$\nabla V_z(f) = (f(x) - y)K_x + \lambda f$$

as a random variable dependent on z. Notably, the expectation satisfies

$$\mathbb{E}[V_z(f)] = \tfrac{1}{2}\left(L(f) + \lambda\|f\|_K^2\right),$$

meaning that the updates in (2.5) can thus be regarded as stochastic approximations of gradient descent methods for solving the regularized expected risk minimization problem (2.4), with time-varying regularization parameters $\lambda = \lambda_t$. In this sense, each iteration is a *stochastic* gradient descent step in $\mathcal{H}_K$. Over many iterations, these individual random steps approximate the global minimizer of the expected risk, while processing one sample at a time.

In particular one can prove the following result.

Proposition 2.3.14. *Fix $\lambda_t = \lambda > 0$ and set $\gamma_t \to 0$ appropriately*[1]*, then*

$$\|f_t - f^\lambda\|_K \to 0, \quad \text{for} \quad t \to \infty;$$

where $f^\lambda = (T_K + \lambda I)^{-1} T_K f_\rho$ is the minimizer of the (regularized) expected risk objective

$$L_\lambda(f) \;=\; \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y)^2 \, d\rho(x,y) \;+\; \lambda\|f\|_K^2.$$

Remark 2.3.15 One can show, under certain conditions on (γ_t, λ_t), that the algorithm's final hypothesis f_t eventually approximates f_ρ at a rate comparable to batch methods.

Before concluding this section, let us state one remark on the role of independence and conditional distribution in this framework.

Remark 2.3.16 (Independence, Conditional Distributions, and Stochastic Updates) Convergence guarantees for the algorithm (2.5) hinge on two aspects of the data and target function. First, we assume that the sequence of examples $\{z_t = (x_t, y_t)\}_{t\in\mathbb{N}}$ is i.i.d. from the underlying distribution ρ. In particular, *independence* ensures that, at each step t, the gradient estimate

$$\nabla V_{z_t}\big(f_{t-1}\big) \;=\; (f_{t-1}(x_t) - y_t)\, K_{x_t} \;+\; \lambda_t\, f_{t-1}$$

is conditionally unbiased given the σ-algebra $\mathcal{F}_{t-1}$ generated by the past data. As we previously observed, the hypothesis f_{t-1} only depends on $z_1, \ldots, z_{t-1}$ and is thus $\mathcal{F}_{t-1}$-measurable. Since z_t is independent of $\mathcal{F}_{t-1}$, the conditional expectation reduces to averaging over z_t while treating f_{t-1} as fixed:

$$\mathbb{E}_{t-1}\big[\nabla V_{z_t}(f_{t-1})\big] \;=\; \nabla L_{\lambda_t}\big(f_{t-1}\big),$$

where $\nabla L_{\lambda_t}(f_{t-1}) = \mathbb{E}_{(x,y)\sim\rho}\big[(f_{t-1}(x) - y)K_x\big] + \lambda_t f_{t-1}$. Each incremental update thereby provides a fresh, unbiased sample of the full gradient. This property results in very useful control of the variance of the updates (e.g., via martingale decomposition). If the data were dependent, additional assumptions (e.g., mixing conditions) would be required to decouple z_t from prior hypotheses $\{f_j\}_{j<t}$.

Second, the *conditional distribution* determines the regression target

$$f_\rho(x) \;=\; \int_{\mathcal{Y}} y \, d\rho(y \mid x).$$

For the iterative sequence $\{f_t\}$ to converge toward f_ρ, the RKHS $\mathcal{H}_K$ must either contain f_ρ exactly, or at least approximate it under some regularity assumption (for instance a *source condition*, $f_\rho = T_K^r g$ with $r > 0$, $g \in L^2_{\rho_\mathcal{X}}$). These regularity assumptions bound the approximation error and govern how effectively finite-sample stochastic updates can learn f_ρ.

[1] $\gamma_t = \frac{1}{(\lambda + k^2)t^\theta}$ with $k = \sup_{x\in X}\sqrt{K(x,x)} < \infty$ and $\theta \in (\frac{1}{2}, 1)$

2.4 Learning Bounds

2.4.1 Excess Risk

The output of a learning algorithm is a function $f_{\mathbf{z}}$ in the hypothesis space $\mathcal{H}$ (eg., RKHS induced by a kernel K), dependent on the training data $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^n$, where each sample $(x_i, y_i) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is drawn from the underlying distribution ρ. A learning algorithm can be thus seen as a map $\mathcal{A} : \mathcal{Z}^n \to \mathcal{H}$ that outputs a hypothesis $f_{\mathbf{z}} = \mathcal{A}(\mathbf{z})$, whose empirical risk is $\widehat{L}(f_{\mathbf{z}}) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_{\mathbf{z}}(x_i))$, given a chosen loss ℓ (eg. squared loss). Since the true distribution ρ is unknown, we generally cannot find f_ρ exactly. Instead, we look for $f_{\mathbf{z}}$ whose expected risk $\widehat{L}(f_{\mathbf{z}}))$ is close to $L(f_\rho)$. The quantity

$$\widehat{L}(f_{\mathbf{z}}) - L(f_\rho),$$

called the *excess risk*, measures how much worse $\widehat{L}(f_{\mathbf{z}})$ is compared to the ideal function.

Consistency means that this excess risk goes to zero as the sample size grows $n \to \infty$. This can be formalized in different ways, for example:

- *Convergence in expectation:* $\mathbb{E}[\,\widehat{L}(f_{\mathbf{z}}) - L(f_\rho)\,] \to 0$,
- *Convergence in probabilty:* $\mathbb{P}\big(\widehat{L}(f_{\mathbf{z}}) - L(f_\rho) \geq \epsilon\big) \to 0$ for all $\epsilon > 0$. Note that the excess risk is a random variable $\Omega \to \mathcal{Z}^n$, $\omega \mapsto \mathbf{z} = (z_1, \ldots, z_n)$.

While consistency is an *asymptotic* notion, *learning bounds* offer finite-sample estimates of how fast a learning algorithm converges. Such results often provide upper bounds of the form

$$\mathbb{E}\Big[\widehat{L}(f_{\mathbf{z}}) - L\big(f_\rho\big)\Big] \leq \epsilon(n, \rho, \mathcal{H}),$$

or, for a confidence parameter $\delta \in (0, 1)$,

$$\mathbb{P}\big(L(f_\rho) \leq \widehat{L}(f_{\mathbf{z}}) + \epsilon(n, \delta, \mathcal{H})\big) \geq 1 - \delta,$$

where ϵ may depend on the sample size n, the probability distribution ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and the properties of $\mathcal{H}$.

For instance, in the case of regularized algorithms in RKHS, the bound might be expressed in terms of the norm $\|\cdot\|_K$, the spectral properties of the kernel integral operator, and the regularization parameter λ. One can then invert this dependence to derive a *sample complexity* (i.e., how large n must be for a desired accuracy) or an *error bound* (i.e., how small the difference in norm between $f_{\mathbf{z}}$ and f_ρ). In the last section of this chapter, for instance, we mention a bound in terms of excess risk and one in terms of probability.

Remark 2.4.1 (No Free Lunch) A natural question is whether one can derive *uniform* guarantees across *all* possible distributions ρ. Ideally, one might seek a single bound of the form

$$\sup_\rho \Big(\mathbb{E}\big[\widehat{L}(f_{\mathbf{z}}) - L(f_\rho)\big]\Big) \;\leq\; \epsilon(n, \mathcal{H}),$$

valid for every distribution on $\mathcal{X} \times \mathcal{Y}$. However, classical "no free lunch" results show that, without additional assumptions, such uniform statements cannot hold in general. Indeed, for any fixed learning algorithm, one can construct distributions ρ that cause arbitrarily poor performance.

This observation does not preclude *universal consistency*, where each distribution is analyzed separately, but it does underline that distribution-free performance bounds require

restricting the set of possible distributions or making further hypotheses (e.g. smoothness conditions). Hence, the assumptions we impose, are crucial for achieving meaningful convergence guarantees in a broader sense.

Having introduced two main algorithms, the batch *regularized linear least squares (LLS)* (Example 2.3.4) and the online *SGD-type* algorithm (2.5), we now discuss their theoretical convergence properties.

2.4.2 Regularized Linear Least Squares: Batch Convergence

Recall the batch *regularized linear least squares* algorithm in feature spaces, described in Example 2.3.4. Recall the hypothesis given by the training data $\{(x_i, y_i)\}_{i=1}^n$,

$$\widehat{f}^{\lambda}(x) \;=\; \sum_{i=1}^{n} c_i\, K(x_i, x),$$

where $c = (\widehat{K} + \lambda I)^{-1}\,\mathbf{y}$ and $\widehat{K}_{ij} = K(x_i, x_j)$, approximates the solution

$$f^{\lambda} \;=\; (T_K + \lambda I)^{-1}\, T_K\, f_\rho,$$

which is the minimizer of $L(f) + \lambda \|f\|_K^2$ in $\mathcal{H}_K$.

Then the following theorem (from [37]) holds.

Theorem 2.4.2 (Regularized LLS Convergence)**.** *Let $\{(x_i, y_i)\}_{i=1}^n$ be i.i.d. samples drawn from ρ, and let $\widehat{f}^{\lambda}$ be the regularized least squares estimator defined above. Suppose there exists constants $\kappa \geq 0$ such that $\kappa := \sup_{x\in\mathcal{X}} \sqrt{K(x,x)} < \infty$, and $M_\rho \geq 0$ such that $\mathrm{supp}(\rho) \subseteq \mathcal{X} \times [-M_\rho, M_\rho]$.*

Then, there exists a constant $C_\delta > 0$ depending on $\delta \in (0,1)$ such that, with probability at least $1-\delta$,

$$L(\widehat{f}^{\lambda}) - L(f^{\lambda}) \leq C_\delta \left(\mathcal{A}(\lambda) + \frac{\kappa^2 \mathcal{B}(\lambda)}{n^2\lambda} + \frac{\kappa \mathcal{A}(\lambda)}{n\lambda} + \frac{\kappa M_\rho}{n^2\lambda} + \frac{M_\rho \mathcal{N}(\lambda)}{n} \right),$$

provided the sample size satisfies

$$n \geq \frac{C_\delta \kappa}{2\lambda} \max\left(\mathcal{N}(\lambda), \sqrt{2/C_\delta}\right),$$

where $\mathcal{A}(\lambda) = \|f^\lambda - f_\rho\|_\rho^2 = \|T_K^{\frac{1}{2}}(f^\lambda - f_\rho)\|_K^2$, $\mathcal{B}(\lambda) = \|f^\lambda - f_\rho\|_K^2$, $\mathcal{N}(\lambda) = \mathrm{Tr}\left[(T_K + \lambda I)^{-1} T_K\right]$, and $C_\eta = 128 \log^2(8/\delta)$.

2.4.3 SGD Algorithm: Convergence Guarantees

We now return to the online case, where the algorithm is given by the update (2.5), with time-varying regularization λ_t and step-size γ_t:

$$f_t \;=\; f_{t-1} \;-\; \gamma_t \Big[\big(f_{t-1}(x_t) - y_t\big)\, K_{x_t} \;+\; \lambda_t\, f_{t-1} \Big],$$

where (x_t, y_t) are i.i.d. samples from ρ, and $K_{x_t}(\cdot) = K(\cdot, x_t)$. We denote the *target function* f_ρ in the RKHS $\mathcal{H}_K$ satisfying

$$f_\rho \;=\; \arg\min_{f\in\mathcal{H}_K} \Big\{ L(f) \;=\; \mathbb{E}[(f(x) - y)^2] \Big\}.$$

Then the following results hold [36].

Theorem 2.4.3 (SGD Algorithm Convergence)**.** *Let* $\{f_t\}$ *be defined by* (2.5)*, then*

$$\limsup_{t\to\infty} \mathbb{E}[\|f_t - f_\rho\|_K^2] = 0,$$

assuming:

1. $f_\rho \in \mathcal{H}_K$ *(or satisfies a suitable source condition,* $f_\rho = T_K^r g$*),*
2. $\gamma_t, \lambda_t \to 0$ *as* $t \to +\infty$*,*
3. $\gamma_t/\lambda_t \to 0$ *and* $\|f_{\lambda_t} - f_{\lambda_{t-1}}\|_K/\gamma_t\lambda_t \to 0$ *,*
4. $\sum_{t=1}^{\infty} \gamma_t\lambda_t = \infty$*.*

Furthermore, if r *quantifies the regularity of* f_ρ*, meaning* $T_K^{-r} f_\rho \in L^2_{\rho_\mathcal{X}}$ *for some* $r \in (1/2, 3/2]$*, and initial regularization parameter* $\lambda_0 \geq 1$*. Then, for all* $t \in \mathbb{N}$*, with probability at least* $1 - \delta$*,*

$$\|f_t - f_\rho\|_K \leq C_0 t_*^{-1} + \left(C_1\lambda_0^{-(r-1/2)} \log\frac{2}{\delta} + C_2\gamma_1\right) t_*^{-\frac{r-\frac{1}{2}}{r+1}},$$

where $t_* := t + t_0$*, with* $t_0 \in \mathbb{N}$ *large enough, and constants*

$$C_0 := 2t_0^{\frac{4r+3}{4r+2}} M_\rho, \quad C_1 := \frac{20r-2}{(2r-1)(2r+3)} \|T_K^{-r} f_\rho\|_\rho, \quad C_2 := \frac{20(\kappa+1)^2 M_\rho}{\kappa}.$$

Corollary 2.4.4. *In particular, in the case* $f_\rho = T_K^r g$ *with* $r \in (1/2, 3/2]$*, we have*

$$\mathbb{E}\big[\|f_t - f_\rho\|_K^2\big] = O\big(t^{-(r-1/2)}\big).$$

Chapter 3

Stochastic Dynamical Systems and Markov Chains

In the earlier chapters, we laid the groundwork for the discussions to come. Chapter 1 introduced the measure-theoretic foundations of probability, providing a rigorous framework for understanding randomness and uncertainty. Building on this, Chapter 2 explored classical machine learning concepts, focusing on kernel methods and learning bounds in the i.i.d. setting.

Yet, many real-world scenarios deviate significantly from the i.i.d. paradigm. Applications such as time series analysis, evolving data streams, and sequences of dependent observations often require a more flexible framework. In this work, we focus on relaxing the 'identical' assumption in i.i.d. data, while avoiding independence through a data collection scheme discussed in Chapter 4. This allows us to address a broader range of processes, including stochastic dynamical systems.

This chapter aims to provide the theoretical tools needed to analyze data arising from *stochastic dynamical systems*. Unlike deterministic systems, stochastic dynamical systems incorporate inherent uncertainty. These systems evolve according to probability kernels, extending deterministic trajectories to stochastic processes as Markov chains. As we progress, we will explore key concepts such as *ergodicity* and the conditions under which system state distributions converge to *stationary distributions*. These results will enable us to work with data that falls outside the usual i.i.d. assumptions.

Our primary goal is to establish key properties, such as measure convergence and dynamics irreducibility, which will form the foundation for the learning framework discussed in the next chapter.

The chapter is organized as follows. We begin by revisiting the concept of a *dynamical system*, starting with the deterministic case before transitioning to the stochastic context. Next, we introduce *Markov processes* and *transition probability kernels*, examining the conditions under which these systems converge to *stationary measures*. Along the way, we study notions such as *irreducibility, aperiodicity*, and various forms of *ergodicity*.

3.1 Dynamical Systems

In this section, we introduce the fundamental concepts of dynamical systems.

Traditionally, dynamical systems are represented by differential or difference equations derived from fundamental physical laws. For instance, Newton's second law $F = ma$ serves as a cornerstone for modeling mechanical systems. Physics-based models like this have

enabled significant advancements by allowing precise simulations and control of various systems.

However, as science ventures into complex and high-dimensional systems such as climate models, neural networks in the brain, financial markets, and language, deriving accurate models from first principles becomes increasingly infeasible. These systems often exhibit nonlinear behavior, high dimensionality, chaos, and stochasticity, making analytical solutions or even numerical simulations challenging, if not impossible.

The explosion of data availability in recent years presents an opportunity to approach these complex systems differently. By leveraging data-driven techniques, we aim to construct models that can accurately predict, control, and provide insights into the underlying dynamics without relying solely on traditional physics-based approaches.

3.1.1 The mathematical modeling

In physics, a *system* is a collection of interacting parts enclosed within a boundary and considered as a single entity. It usually models a specific portion of the universe that is being studied or analyzed. The boundaries of the system might be physical or abstract and they define what is included in the analysis versus what is considered to be the surroundings. A *dynamical system* is a system that evolves in time.

In mathematics, a dynamical system is a collection of elements in a set, possibly together with some structure (e.g., metric, probability measure), equipped with a function that evolves the system over time.

Formally, we define S as the *state space*, whose elements represent the interacting components of the system, and $f : S \to S$ as the *evolution function* that governs the system's dynamics. The specific nature of S and the map f depend on the characteristics of the system being modeled and how it evolves over time. To formalize this concept in a notationally efficient and general manner, we first introduce the notion of a monoid.

Definition 3.1.1 (Monoid)**.** A **monoid** $\mathcal{T}$ is an algebraic structure consisting of a set equipped with a binary operation $*$ satisfying:

(i) **Associativity**: $(g * h) * k = g * (h * k)$ for all $g, h, k \in \mathcal{T}$.

(ii) **Identity Element**: There exists an element $e \in \mathcal{T}$ such that $g * e = e * g = g$ for all $g \in \mathcal{T}$.

utilizing the concept of a monoid, we can now formally define both discrete-time and continuous-time dynamical systems as follows.

Definition 3.1.2 (Dynamical System)**.** A **dynamical system** consists of:

(i) a set S called the **state space**, whose elements represent the states of the system,

(ii) a monoid $\mathcal{T}$, representing the **time domain** or indexing set,

(iii) a map $s : \mathcal{T} \to S$ that assigns to each time index $t \in \mathcal{T}$ the state of the system at that time, denoted as $s_t := s(t) \in S$,

(iv) a function $f : S \to S$ called the **evolution function**, which describes how the system evolves from one state to another.

Depending on the nature of the change, f can be adapted in order to model different dynamics, e.g., $f : S \times \mathcal{T} \to S$ for non-autonomous systems or $f : S \times \Omega \to S$, with Ω being a sample set in a probability space, to introduce stochasticity; similarly the map $s : \mathcal{T} \to S$ will also be adapted. We will explore these cases later.

We classify the system as:

- **Discrete-time** if the monoid is discrete, meaning its elements form a countable set, e.g., $\mathcal{T} = \mathbb{N}$, $\mathcal{T} = \mathbb{Z}/n\mathbb{Z}$.
- **Continuous-time** if the monoid is a continuous monoid, e.g., $\mathcal{T} = \mathbb{R}_{\geq 0}$, $\mathcal{T} = \mathbb{R}$.

In this setting, $\mathcal{T}$ may still represent time in the traditional sense, with the binary operation being ordinary addition, but it can also accommodate more generic indexing structures, such as multidimensional spatial coordinates, the numbering of a word (or token) in a text, or other parameters.

Remark 3.1.3 Although we have defined dynamical systems in a completely abstract setting where S is simply a set, in practice the state space usually has additional structure that is preserved by the map f. For example, $(S, \mathcal{B})$ could be a measurable space and f a measurable map, a topological space and a continuous map, or a metric space and an isometry, or a smooth manifold and a differentiable map.

Example 3.1.4 (Continuous-Time Dynamical System) Consider a continuous-time dynamical system indexed by $\mathcal{T} = \mathbb{R}_{\geq 0}$. Suppose the dynamics are governed by an ordinary differential equation (ODE):

$$\dot{s}(t) = f(s(t)).$$

For example, let the state space $S = \mathbb{R}$ and define $f(s(t)) = cs(t)$, where $c \in \mathbb{R}$. Then the ODE becomes $\dot{s}(t) = cs(t)$. For each initial condition $s(0) = s_0$, the unique solution can be easily found as $s(t) = s_0 e^{ct}$.

Example 3.1.5 (Discrete-Time Dynamical System) Consider a discrete-time dynamical system indexed by $\mathcal{T} = \mathbb{N}$. Suppose the dynamics are governed by a difference equation:

$$s(t+1) = f(s(t)).$$

Let the evolution function be $f(s(t)) = c_1 s(t) + c_2$, with $c_1, c_2 \in \mathbb{R}$. The difference equation becomes $s(t+1) = c_1 s(t) + c_2$. As before, for each initial condition $s(0) = s_0$, the unique solution is $s(t) = s_0 c_1^t + c_2 \frac{1-c_1^t}{1-c_1}$, provided that $c_1 \neq 1$.

In the previous examples, the systems evolve depending only on the initial condition and the choice of constants. Here, the systems' evolution functions are not explicitly dependent on the time parameter $t \in \mathcal{T}$. We call such systems **autonomous**; otherwise, if the dynamics explicitly depend on time, we call them **non-autonomous**, meaning that the dynamics can change over time independently of the state.

3.1.2 From Deterministic to Stochastic Dynamical Systems

So far, we have considered dynamical systems that evolve deterministically, governed by an evolution function f. In practice, however, many systems are better modeled by accounting for stochasticity or are influenced by inherent randomness. There are multiple ways to formalize randomness in this context. In what follows, we will focus on two important perspectives: the one of *stochastic dynamical systems (SDS)* given by a stochastic evolution function, and the one of *Markov chains (MC)* given by a probability kernel. We will show how these two viewpoints are fundamentally intertwined.

First we lay down some fundamental working assumptions.

From now on, we will only consider discrete-time autonomous systems $(S, \mathcal{B})$, where S is a compact state space, and $\mathcal{B}$ is a (countably generated) σ-algebra.

Remark 3.1.6 Note that most of the following discussion applies only when the state space's σ-algebra $\mathcal{B}$ is countably generated. This condition is quite mild. For instance, any subset of $\mathbb{R}^d$ equipped with the standard Borel σ-algebra satisfies this condition, since the Borel σ-algebra is generated by open balls with rational centers and rational radii, which form a countable set.

Definition 3.1.7 (Stochastic Dynamical System (SDS))**.** Let $(S, \mathcal{B})$ be a measurable state space and $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space. Suppose we are given an S-valued random variable $X_0 : \Omega \to S$ and a sequence of i.i.d. random variables $\{\vartheta_t\}_{t\in\mathbb{N}}$ such that $\vartheta_t : \Omega \to [0, 1]$ and each ϑ_t is distributed as $U(0, 1)$.

A **stochastic dynamical system** (SDS) is defined by a measurable function

$$f : S \times [0, 1] \to S,$$

called the *stochastic evolution function*, and the recursion

$$X_{t+1} = f(X_t, \vartheta_t) \quad \text{for all } t \in \mathbb{N}.$$

Each X_t is a measurable function $X_t : \Omega \to S$, representing the system's random state at time t. Each update $X_t \mapsto X_{t+1}$ is driven by the current state and an independent $U(0, 1)$ random input ϑ_t.

Remark 3.1.8 The choice of $\vartheta_t \sim U(0, 1)$ is without loss of generality. By the probability integral transform, any random variable can be generated from a $U(0, 1)$-distributed random variable via an appropriate measurable transformation. It is a convenient choice for representing arbitrary stochastic behavior in the evolution of the system.

As we will see in the next chapter, our ultimate goal is to best approximate certain conditional expectations arising in the stochastic evolution, such as $\mathbb{E}[X_{t+1} \mid X_t = x]$. The aim is to develop a learning algorithm that estimate these conditional expectations from observed data and to establish theoretical guarantees on the sample complexity needed to achieve reliable approximation in this setting.

Given an SDS, we can describe its evolution in probabilistic terms. From any current state $x \in S$, the distribution of the next state X_{t+1} is the law $\mathcal{L}(f(x, \vartheta_t))$. This naturally leads us to define a *transition probability kernel* that encapsulates these probabilities.

Definition 3.1.9 (Transition Probability Function/Markov Kernel)**.** Let $(S, \mathcal{B})$ be a measurable space (with countably generated σ-algebra). A **transition probability function** (or transition probability kernel) is a function

$$P : S \times \mathcal{B} \to [0, 1],$$

satisfying the following properties:

(i) For every $x \in S$, the map $A \mapsto P(x, A)$ is a probability measure on $(S, \mathcal{B})$;

(ii) For every $A \in \mathcal{B}$, the map $x \mapsto P(x, A)$ is $\mathcal{B}$-measurable.

The value $P(x, A)$ represents the probability of transitioning from state $x \in S$ to a measurable set of states $A \in \mathcal{B}$ in one time step. Such a function P is also called a **Markov kernel**.

This definition generalizes the notion of a transition matrix, introduced at the end of the first chapter, from the finite state space setting to more general state spaces. In particular it allows us to move from discrete to continuous state spaces by working with general probability measures rather than discrete probability distributions.

Remark 3.1.10 If we consider a transition kernel $P(x, \cdot)$ that is a Dirac measure at the point $f(x)$, i.e. $P(x, A) = \delta_{f(x)}(A)$, then there is no randomness in the evolution and we recover a deterministic system. Thus, taking $P(x, \cdot) = \delta_x(\cdot)$ for some deterministic update rule is a special case of a stochastic system where the randomness is trivial.

3.1.3 Path Space

In the deterministic setting, a dynamical system is represented by a unique function $s : \mathcal{T} \to S$ mapping each time t to the corresponding state s_t. In the stochastic setting, such a function s is seen as an element of $s \in S^{\mathcal{T}}$, where $S^{\mathcal{T}} = \{s : \mathcal{T} \to S\}$ is the class of functions $s : \mathcal{T} \to S$. The element s, also called a *path* is now a realization of a sequence of random variables $\{X_t\}_{t\in\mathcal{T}}$ defined on the *path space*.

Definition 3.1.11 (Path Space)**.** Let $(S, \mathcal{B})$ be a measurable space and let $\mathcal{T}$ be our discrete time domain ($\mathbb{N}$ in the countable case or $\{0, 1, \dots, T\}$ in the finite case). The measurable space $(S^{\mathcal{T}}, \mathcal{B}^{\mathcal{T}})$ is called a **path space**, where $\mathcal{B}^{\mathcal{T}}$ denotes the σ-algebra on $S^{\mathcal{T}}$ generated by all *evaluation maps* $\pi_t : S^{\mathcal{T}} \to S,\ t \in \mathcal{T}$, given by $\pi_t(s) = s(t)$.

We denote with $(S^T, \mathcal{B}^T)$ the path space for $\mathcal{T} = \{0, 1, \dots, T\}$ and with $(S^\infty, \mathcal{B}^\infty)$ the path space in the case of $\mathcal{T} = \mathbb{N}$.

If $X : \Omega \to U \subset S^{\mathcal{T}}$, then clearly $X_t = \pi_t \circ X$ maps Ω into S. Thus, X may also be regarded as a function $X(t, \omega) = X_t(\omega)$ from $\mathcal{T} \times \Omega$ to S. We will explore this precisely after a few remarks on measurability in the path space.

Remark 3.1.12 Since the finite path space S^T is just the T-times cartesian product $\prod_{n=0}^{T} S$, and $S^\infty = \prod_{n=0}^{\infty} S$, one could ask if the corresponding σ-algebras coincide with the product σ-algebras introduced in the first chapter. It is indeed the case, to see that recall that:

(i) The product σ-algebra $\mathcal{B} \otimes \cdots \otimes \mathcal{B}$ is generated by all cylinder sets of the form

$$C = A_1 \times \cdots \times A_T, \quad \text{with } A_t \in \mathcal{B} \text{ for } t = 1, \dots, T.$$

(ii) The σ-algebra $\mathcal{B}^T$ on the path space is generated by the projections $\pi_t : S^T \to S$.

In particular, every set in $\mathcal{B}^T$ can be expressed as a countable union of intersections of cylinder sets, and every cylinder set belongs to $\mathcal{B}^T$. Hence, $\mathcal{B}^T = \mathcal{B} \otimes \cdots \otimes \mathcal{B}$. Similarly, in the infinite case, we have that $\bigotimes_{n=1}^{\infty} \mathcal{B} = \sigma(\{\pi_t : t \in \mathbb{N}\})$.

A rigorous proof relying on cylinder sets can be found in [34], p.75.

As a consequence we have the following lemma.

Lemma 3.1.13 (Measurability)**.** *Let $(S, \mathcal{B})$ be a measurable space, $\mathcal{T}$ be the index set as above, $U \subset S^{\mathcal{T}}$, and $X : \Omega \to U$ be a function. The following conditions are equivalent:*

(i) X is $\mathcal{B}^{\mathcal{T}} \cap U$-measurable.

(ii) $X_t : \Omega \to S$ is $\mathcal{B}$-measurable for every $t \in \mathcal{T}$.

A mapping X with the properties in Lemma 3.1.13 is called an S-valued (random) process on $\mathcal{T}$ with paths in U. By the lemma it is equivalent to regard X as a collection of random elements X_t in the state space S.

The next theorem shows that from any initial distribution and a transition kernel, we can construct a stochastic process with the corresponding finite-dimensional distributions. This result lays the foundation for relating a Markov chain with a given kernel as a realization of an SDS, and conversely.

Theorem 3.1.14. *For any initial measure μ on S, i.e., $\mu : \mathcal{B} \to [0,1]$, and any transition probability kernel $P, \{P(x,A) : x \in S,\ A \in \mathcal{B}\}$, there exists an S-valued stochastic process $X = \{X_0, X_1, \ldots\}$ on $\mathcal{T}$ with paths in some $U \subset \mathcal{B}^{\mathcal{T}}$, and a probability measure P_μ on $S^{\mathcal{T}}$ such that $P_\mu(A)$ is the probability of the event $\{X \in A\}$ for $A \in U$. Moreover, for measurable sets $A_i \subseteq X_i$, $i = 0, \ldots, n$, and any integer n, we have*

$$P_\mu(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) = \int_{A_0} \cdots \int_{A_{n-1}} \mu(dy_0) P(y_0, dy_1) \cdots P(y_{n-1}, A_n). \tag{3.1}$$

In the case of $\mu = \delta_x$, the dirac measure at a point $x \in S$, we use the notation $P_{\delta_x} = P_x$.

Definition 3.1.15 (time-homogenous Markov chains (MC))**.** The stochastic process X is called a **time-homogeneous Markov chain with transition probability kernel P and initial distribution μ** if the finite-dimensional distributions of X satisfy (3.1) for each n.

With these notions in place, we are now ready to formally establish the connections between SDSs and Markov chains (MC). The following proposition confirms that any S-valued time-homogeneous Markov process with a given transition kernel can be realized as an SDS, and conversely, every SDS induces a time-homogeneous Markov process with a corresponding transition kernel.

Proposition 3.1.16. *Let X be an S-valued process on $\mathcal{T}$. Then the following conditions are equivalent:*

(i) X is a time-homogeneous Markov process with transition kernel P and initial distribution μ,

(ii) There exists a measurable function $f : S \times [0,1] \to S$ and i.i.d. $U(0,1)$ random variables $\vartheta_1, \vartheta_2, \ldots$, independent of $\mathcal{L}(X_0) = \mu$, such that

$$X_n = f(X_{n-1}, \vartheta_n) \quad \text{a.s. for all } n \in \mathbb{N}.$$

In this case, the transition kernel is given by $P(x, \cdot) = \mathcal{L}(f(x, \vartheta))$ almost surely.

This guarantees that discussing SDSs through stochastic evolution functions or as (time-homogenous) Markov chains through transition kernels is essentially equivalent. Hence, the study of the MC and its properties naturally extends to the corresponding SDS.

This framework facilitates the formalization of asymptotic properties and *trajectories* in stochastic systems. Trajectories, in particular, will play a crucial role in the next chapter, as they are fundamental parts of our learning algorithm's data.

Definition 3.1.17 (Trajectory)**.** Let $(S, \mathcal{B})$ be a stochastic dynamical system with initial probability measure μ. A **trajectory** is a realization of the stochastic process given by Theorems 3.1.14 and 3.1.16 with $\mathcal{L}(X_0) = \mu$, that is

$$X(\omega) = \{X_0(\omega), X_1(\omega), X_2(\omega), \dots\}.$$

We call a trajectory of **length** L the finite truncation $\{X_0(\omega), X_1(\omega), \dots, X_{L-1}(\omega)\}$.

3.2 Markov Chains

In the previous sections, we have established that every stochastic dynamical system can be represented as a time-homogeneous Markov chain $\{X_n\}_{n\in\mathbb{N}}$ with a transition probability kernel P and an initial distribution μ. Having identified this correspondence, we now turn our attention to the long-term behavior of such chains. Specifically, we want to understand under what conditions the sequence of measures μP^n converges to a limiting (stationary) distribution and, more importantly, how quickly this convergence occurs.

Our ultimate goal is to characterize conditions that ensure *exponential convergence* of the induced measures. To achieve this, we will begin by examining the n-step transition probability kernels, which describe how distributions evolve over multiple time steps. This approach will guide us through key concepts such as *ergodicity, aperiodicity, and irreducibility*, which together form the backbone of the classical theory ensuring convergence.

We denote by $X_{(P,\mu)}$, or just X when there is no ambiguity, the Markov chain determined by the transition kernel P and the initial measure μ derived from our SDS $(S, \mathcal{B})$.

3.2.1 Evolution of Probability Distributions

Definition 3.2.1 (n-step transition probability kernel)**.** The ***n*-step transition probability kernel** is defined iteratively. We set $P^0(x, A) = \delta_x(A)$ and, for $n \geq 1$, we define inductively

$$P^n(x, A) = \int_S P(x, dy) P^{n-1}(y, A), \quad x \in S,\ A \in \mathcal{B}. \tag{3.2}$$

We write P^n for the n-step transition probability kernel $\{P^n(x, A), x \in S, A \in \mathcal{B}\}$.

Theorem 3.2.2 (Chapman–Kolmogorov)**.** *For any m with $0 \leq m \leq n$, the following Chapman–Kolmogorov equation holds:*

$$P^n(x, A) = \int_S P^m(x, dy) P^{n-m}(y, A), \quad x \in S,\ A \in \mathcal{B}. \tag{3.3}$$

We interpret (3.3) as saying that, as X moves from x into A in n steps, at any intermediate time m, it must take some value $y \in S$; and that, being a Markov chain, it forgets the past at that time m and moves the succeeding $(n - m)$ steps with the law appropriate to starting afresh at y. We can write (3.3) alternatively as

$$P_x(X_n \in A) = \int_S P_x(X_m \in dy) P_y(X_{n-m} \in A). \tag{3.4}$$

Exactly as the one-step transition probability kernel describes a chain X, the m-step kernel satisfies the definition of a transition kernel and thus defines a Markov chain $X^m = \{X_n^m\}$ with transition probabilities

$$P_x(X_n^m \in A) = P^{mn}(x, A). \tag{3.5}$$

Definition 3.2.3 (m-skeleton of a Markov chain)**.** The chain X^m with transition law (3.5) is called the ***m*-skeleton** of the chain X.

We now explore how the evolution of measures and observables in terms of the transition kernel P.

Given an initial (σ-finite) measure μ on S, it evolves by recursively applying the Markov transition function. The n-step evolution of μ acts pointwise on $A \in \mathcal{B}$ as

$$\mu^{[n]}(A) := \int_S \mu^{[n-1]}(dx)\, P(x, A), \tag{3.6}$$

which, in terms of the n-step transition kernel P^n, is equivalent to

$$\mu^{[n]}(A) = \mu P^n(A) = \int_S \mu(dx)\, P^n(x, A).$$

As an operator, P^n acts on continuous measurable functions f on S as

$$P^n f(x) = \int_S P^n(x, dy) f(y).$$

Remark 3.2.4 Since S is compact, every continuous function $f : S \to \mathbb{R}$ is bounded and thus integrable with respect to any probability measure on S. In particular, for any initial measure μ and any Markov kernel P, the measures μP^n are probability measures, and the integrals defining $P^n f$ and μP^n are well-defined.

From now on, our main focus will be on the sequence of ditributions $\{\mu^{[t]}\}_{t\in\mathbb{N}}$, with $\mu^{[t]} := \mu P^t$. We will see what it means for this distributions to converge and study its rate of convergence.

3.2.2 Ergodicity

The main results regarding the learning algorithm presented in Chapter 4 largely depend on the convergence of distributions, for which we will need a fast rate of convergence. Our aim for the rest of this chapter is to formalize this convergence precisely and study sufficient conditions on our Markov Chains for this convergence to happen. To begin, let's recall the definition of the total variation norm in the classical setting.

Definition 3.2.5 (Total Variation Norm)**.** If $\mu \in \mathcal{M}(S)$, then the **total variation norm** $\|\mu\|_{TV}$ is defined as

$$\|\mu\|_{TV} := \sup_{f:|f|\leq 1} |\mu(f)| = \sup_{A\in\mathcal{B}} \mu(A) - \inf_{A\in\mathcal{B}} \mu(A).$$

Definition 3.2.6 (Ergodicity)**.** We call the system **ergodic** if there exists a unique probability measure $\pi \in \mathcal{P}(S)$ such that $\mu^{[t]} = \mu P^t \xrightarrow{t\to\infty} \pi$ for any starting probability measure μ, meaning $\|\mu^{[t]} - \pi\|_{TV} \xrightarrow{t\to\infty} 0 \quad \forall \mu \in \mathcal{P}(S)$.

Although ergodicity is typically expressed in terms of the TV-norm, we will make use of a different version, in particular with respect to the Hölder norm. To achieve this, let's first revisit the definition of Hölder space.

Definition 3.2.7 (Hölder Space)**.** For $s \in (0, 1]$, the **Hölder space** $C^s(S)$ is defined as the set of all functions $f : S \to \mathbb{R}$ such that the Hölder norm $\|f\|_{C^s}$ is finite, where the Hölder norm is defined by

$$\|f\|_{C^s(S)} = \|f\|_{C(S)} + |f|_{C^s(S)}, \quad \text{with} \quad |f|_{C^s(S)} = \sup_{x\neq y} \frac{|f(x)-f(y)|}{d(x,y)^s}.$$

Note that $C^s(S)$ is a Banach space for each $s \in (0,1]$, as it is complete with respect to the Hölder norm $\|\cdot\|_{C^s(S)}$. Moreover, a probability measure μ on S can be regarded as an element of the dual space $(C^s(S))^*$. This is because the inclusion $C^s(S) \hookrightarrow C(S)$ is a continuous embedding due to the simple inequality $\|f\|_{C(S)} \leq \|f\|_{C^s(S)}$. Recall that S is compact, and so every continuous function on S has a finite supremum norm $\|f\|_{C(S)} = \|f\|_\infty < +\infty$. Therefore, we have the following inclusion $\mathcal{P}(S) \subset \mathcal{M}(S) = C(S)^* \hookrightarrow (C^s(S))^*$, where $\mathcal{P}(S)$ is the space of probability measures and $\mathcal{M}(S)$ is the space of signed bounded measures on S.

Definition 3.2.8 (Ergodicity with Hölder norm)**.** We call the system **ergodic**, or simply ergodic, if there exists a unique probability measure $\pi \in \mathcal{P}(S)$ such that $\mu^{[t]} = \mu P^t \xrightarrow{t\to\infty} \pi$ for any starting probability measure μ, meaning $\|\mu^{[t]} - \pi\|_{(C^s(S))^*} \xrightarrow{t\to\infty} 0 \quad \forall \mu \in \mathcal{P}(S)$. Equivalently, by definition of the dual, for each $\mu \in \mathcal{P}(S)$

$$\frac{\left|\int_S f(x)d\mu^{[t]} - \int_S f(x)d\pi\right|}{\|f\|_{C^s(S)}} \xrightarrow{t\to\infty} 0 \quad \forall f \in C^s(S),\ \forall t. \tag{3.7}$$

From now on we will always assume this definition of ergodicity.

Remark 3.2.9 In general, a Markov chain can exhibit multiple "ergodic behaviors" if there are several invariant measures, each attracting different initial distributions. Concretely, the state space may decompose into distinct "ergodic regions," and starting the chain in one region leads to convergence to one particular invariant measure, while starting in another region may lead to a different limit. However, in the setting where a *unique* invariant measure π exists and attracts *all* initial distributions, we say the chain is ergodic. This is precisely the scenario we are interested in: the limiting distribution is unique and does not depend on the initial distribution.

To gain a clearer understanding of the limit measure of an ergodic system, we introduce the concept of *invariant measure* and examine its connection to the long-term behavior.

Definition 3.2.10 (Invariant measures)**.** A measure $\pi \in \mathcal{M}(S)$ is called **invariant** if it satisfies $\pi P = \pi$, i.e.

$$\pi(A) = \int_S \pi(dx)P(x,A) \quad \forall A \in \mathcal{B}.$$

Let us first recall what a stationary chain is and how it relates to the invariant measure we just introduced.

A process is called stationary if, for any k, the marginal distribution of $\{X_n, \ldots, X_{n+k}\}$ stays the same regardless of the value of n. While most Markov chains aren't stationary by default, we can sometimes create a stationary process $\{X_n, n \in \mathbb{N}\}$ by choosing the initial distribution μ appropriately.

To generate an entire stationary process, it's enough to ensure stationarity at the first step. Starting with an initial invariant probability measure π, we can iterate as follows:

$$\pi(A) = \int_S \pi(dx)P^2(x,A)$$

$$\vdots$$

$$= \int_S \pi(dx)P^n(x,A) = P_\pi(X_n \in A),$$

for any n and any $A \in \mathcal{B}$.

By the Markov property, X is stationary if and only if the distribution of X_n does not change over time.

Invariant probability measures are important not just because they define stationary processes, but also because they determine the ergodic, behavior of the chain. To see why, consider $P_\mu(X_n \in \cdot)$ for any initial distribution μ. If a limiting measure γ_μ exists on the space of probability measures, such that

$$P_\mu(X_n \in A) \to \gamma_\mu(A) \quad \forall A \in \mathcal{B} \text{ as } n \to \infty,$$

then

$$\begin{aligned}\gamma_\mu(A) &= \lim_{n\to\infty} \int_S \mu(dx) P^n(x, A) \\ &= \lim_{n\to\infty} \int_S \mu(dx) \int P^{n-1}(x, dw) P(w, A) \\ &= \int_S \gamma_\mu(dw) P(w, A),\end{aligned}$$

This shows that γ_μ is invariant under P. In particular, if there is a unique invariant probability measure π, then $\gamma_\mu = \pi$ for any initial μ (provided the limit exists), showing that the long-term behavior of the chain does not depend on the initial distribution μ.

Example 3.2.11 Consider a Markov chain $\{X_n\}$ on the real line, where $P(x, \cdot) = N\left(\frac{x}{2}, \frac{3}{4}\right)$ for each $x \in \mathbb{R}$. Equivalently,

$$X_{n+1} = \frac{1}{2} X_n + U_{n+1},$$

where $\{U_n\}$ are i.i.d. with $\mathcal{L}(U_n) = N\left(0, \frac{3}{4}\right)$.

Note that the standard normal distribution $\pi = N(0, 1)$ is invariant. In fact, for any $A \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned}\int_{\mathbb{R}} \pi(dx) P(x, A) &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \left(\int_{y\in A} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}\left(y-\frac{x}{2}\right)^2} dy \right) dx, \\ &= \int_{x\in\mathbb{R}} \int_{y\in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}\left(y-\frac{x}{2}\right)^2} dy dx, \\ &= \int_{y\in A} \int_{x\in\mathbb{R}} \frac{2}{2\pi\sqrt{3}} e^{-\frac{1}{2}x^2 - \frac{2}{3}\left(y-\frac{x}{2}\right)^2} dx dy, \\ &= \int_{y\in A} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \pi(A).\end{aligned}$$

which means the Markov chain $\{X_n\}$ is stationary with respect to $N(0, 1)$.

Remark 3.2.12 For this reason, we can express the limit measure of an ergodic system as $\mu P^n \to \pi$, independently of the measure μ. Formally, in the sense of $(C^s(S))^*$, we have:

$$\|\psi P^t - \phi P^t\|_{(C^s(S))^*} \xrightarrow{t\to\infty} 0 \quad \forall \psi, \phi \in \mathcal{P}(S).$$

In particular, by fixing $x \in S$, we obtain $P^n(x, \cdot) \to \pi$.

As we have observed, the relationship between the limit measure and the invariant measure is valid as long as we can formalize the limit and, hence, the distance between two measures. To achieve this, it is essential to define a norm for our measure space. The Hölder norm we previously introduced is merely one specific option. In reality, a wide variety of norms is used.

Remark 3.2.13 Although the total variation (TV) norm is the most common choice for measuring the distance between probability measures, in this chapter (and especially in the next), we employ the Hölder norm because it allows us to derive explicit convergence rates under additional smoothness assumptions on the Markov chain. Working in $(C^s(S))^*$ gives us finer control over certain error terms related to the chain's transition dynamics, which is crucial for obtaining quantitative bounds in Chapter 4. Nevertheless, convergence in the TV-norm guarantees convergence for the Hölder norm as well, since the TV-norm imposes a stronger condition in the space of measures. Keeping this in mind, we can be reassured that the usual theoretical guarantees one has in the classical framework still apply to our case of interest.

To derive explicit bounds for our use case rather than just asymptotic results, it is essential to introduce a stronger form of ergodicity that requires a convergence rate: *geometric ergodicity*. Geometric ergodicity corresponds to the property, which not always holds, that this convergence occurs exponentially quickly.

Definition 3.2.14 (Geometric Ergodicity)**.** We say that the system above is **geometrically ergodic** if:

(i) it is ergodic – in the sense of $(C^s(S))^*$;

(ii) there exists $V : S \to [0, +\infty)$ measurable, $C > 0$ and $\alpha \in [0, 1)$ such that

$$\int_S V(x)d\pi(x) < \infty \text{ and } \|P^t(x,\cdot) - \pi(\cdot)\|_{(C^s(S))^*} \leq C\alpha^t V(x) \quad \forall x \in S,\ \forall t \in \mathbb{N}.$$

This definition formalizes what it means for an ergodic system to have a geometric (or exponential) rate of convergence. There are several other definitions of ergodicity, the one of our interest is *uniformly geometric ergodicity*. It is a stronger form of geometric ergodicity where the function $V(x)$ in the definition is constant. This means that the rate of convergence to the limit distribution does not depend on the initial state x. Instead, the convergence happens at a uniform exponential rate for all starting points in the state space.

Remark 3.2.15 (Uniform ergodicity) Uniform ergodicity per se does not necessarily imply an exponential rate of convergence, but together with the requirement of a geometric rate, it insists that the same constants C and α in (4.2) work uniformly across the space.

Definition 3.2.16 (Uniformly geometric ergodicity)**.** Let $0 \leq s \leq 1$. We call the system **uniformly geometrically ergodic** if there exist constants $C > 0$ and $0 < \alpha < 1$ such that,

$$\|P^t(x,\cdot) - \pi(\cdot)\|_{(C^s(S))^*} \leq C\alpha^t, \quad \forall t. \tag{3.8}$$

Or, equivalently, for any initial measure μ,

$$\left| \int_S f(x)d(\mu P^t) - \int_S f(x)d\pi \right| \leq C\alpha^t \|f\|_{C^s(X)}, \quad \forall f \in C^s(X),\ \forall t. \tag{3.9}$$

Remark 3.2.17 Trivially we have:

$$\textit{uniformly geometric ergodicity} \implies \textit{geometric ergodicity} \implies \textit{simple ergodicity}.$$

This ergodicity assumption is equivalent to, or implied by, a number of different statements, which in some cases might be easier to check. We explore this in the following by introducing concepts such as *irreducibility*, *aperiodicity* and *Harris recurrence*.

3.2.3 Irreducibility

Definition 3.2.18 (φ-irreducible)**.** The process X with Markov kernel P, is said to be **φ-irreducible** if there exists a non-zero σ-finite measure φ on S such that,

$$\varphi(A) > 0 \implies \exists n > 0 \ : \ P^n(x, A) > 0 \quad \forall x \in S.$$

This basically means that every 'relevant' measurable set, in the sense of φ, is always accessible from any point $x \in S$ in a finite amount of steps.

Example 3.2.19 Back to the example from the previous section (3.2.11), where $\{X_n\}$ is a Markov chain on the real line, with $P(x, \cdot) = N\left(\frac{x}{2}, \frac{3}{4}\right)$ for each $x \in \mathbb{R}$.

Then, for any $A \subset \mathcal{B}(\mathbb{R})$ such that the Lebesgue measure $\lambda(A) > 0$, for all $x \in \mathbb{R}$,

$$P(x, A) = \int_{y \in A} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}\left(y - \frac{x}{2}\right)^2} dy > 0.$$

It follows that $\{X_n\}$ is λ-irreducible.

While the concept of φ-irreducibility ensures that the Markov chain can, with some positive probability, reach all sets of positive φ-measure from any starting point, it does not guarantee uniqueness of the measure φ. In practice, there could be many different φ that make the chain irreducible. To obtain a canonical irreducibility measure and a more intrinsic notion of irreducibility, it is useful to introduce the idea of a *maximal irreducibility measure*.

Proposition 3.2.20 (Maximal Irreducibility Measure)**.** *Suppose the Markov chain X is φ-irreducible for some (nonzero) measure φ on S. Then there exists a measure ψ, called a **maximal irreducibility measure**, such that:*

(i) the chain is ψ-irreducible;

(ii) for any other measure φ', the chain is φ'-irreducible if and only if $\psi \succ \varphi'$ (i.e., any set $A \in \mathcal{B}$ for which $\varphi'(A) > 0$ also satisfies $\psi(A) > 0$);

(iii) $\psi(A) = 0 \implies \psi\{x \mid \exists n > 0 \ \ s.t. \ \ P^n(x, A) > 0\} = 0.$

We will consistently use ψ to denote an arbitrary maximal irreducibility measure for the chain X.

Definition 3.2.21 (ψ-irreducible)**.** The Markov chain is called **ψ-irreducible** if it is φ-irreducible for some φ, and the measure ψ is a maximal irreducibility measure satisfying the conditions of Proposition 3.2.20.

We write

$$\mathcal{B}^+ := \{A \in \mathcal{B} : \psi(A) > 0\}$$

for the sets of positive ψ-measure; the equivalence of maximal irreducibility measures implies that $\mathcal{B}^+$ is uniquely defined.

With the definition of ψ-irreducibility as our baseline, we can further explore conditions ensuring convergence.

3.2.4 Aperiodicity

While ψ-irreducibility is a fundamental notion, it does not preclude the existence of periodic structures that can hinder convergence. To address this, we introduce the concept of *small sets*, which play an important role in establishing certain uniformity conditions and in proving aperiodicity.

Definition 3.2.22 (Small set)**.** A set $C \in \mathcal{B}$ is called a **small set** if there exists an integer $m > 0$ and a nontrivial measure ν_m on S such that for all $x \in C$ and all $B \in \mathcal{B}$,

$$P^m(x, B) \geq \nu_m(B).$$

When this condition holds, we say that C is ν_m*-small*.

Intuitively, when a set C is small, and ν_m being non-trivial meaning that there exists B measurable such that $\nu_m(B) > 0$, then there is a positive chance for the chain to move from the small set C to B in m steps, independently of the state $x \in C$.

The existence of small sets is central to establishing various ergodic properties of Markov chains. In particular, for a ψ-irreducible chain, it can be shown that every set in $\mathcal{B}^+$ contains a small set. Another important fact is the following

Proposition 3.2.23. *Suppose the chain X is ψ-irreducible. If $C \in \mathcal{B}^+(X)$ is ν_n-small, then there exists $M \in \mathbb{N}$, and a measure ν_M such that C is ν_M-small, and $\nu_M(C) > 0$.*

Hence we have $P^M(x, \cdot) \geq \nu_M(\cdot)$, $x \in C$, and $\nu_M(C) > 0$, so that, when the chain starts in C, there is a positive probability that the chain will return to C at time M. We will use the set C and the corresponding measure ν_M to define a cycle for irreducible Markov chains.

Consider the set of time points for which C is ν_M-small with a *minorizing measure* ν_n (meaning that for every $x \in C$ and $B \in \mathcal{B}$, we have $P^n(x, B) \geq \nu_n(B)$) proportional to ν_M.

Formally the set is

$$E_C = \{n \geq 1 : C \text{ is } \nu_n\text{-small, with } \nu_n = \varepsilon_n \nu_M \text{ for some } \varepsilon_n > 0\}.$$

For any $B \subseteq C$, if $n, m \in E_C$, we know C is both ν_n-small and ν_m-small. Thus,

$$P^m(x, B) \geq \nu_m(B) = \varepsilon_m \nu_M(B), \quad x \in C,$$

and

$$P^n(y, B) \geq \nu_n(B) = \varepsilon_n \nu_M(B), \quad y \in C.$$

Using the Chapman-Kolmogorov equations, and restricting the integral to C, for $x \in C$ we obtain

$$P^{n+m}(x, B) = \int_S P^m(x, dy) P^n(y, B) \geq \int_C P^m(x, dy) P^n(y, B).$$

Substituting the minorization bounds:

$$P^{n+m}(x, B) \geq \int_C \varepsilon_m \nu_M(dy)\, \varepsilon_n \nu_M(B) = [\varepsilon_m \varepsilon_n \nu_M(C)]\, \nu_M(B),$$

which shows that E_C is closed under addition. Thus, there is a natural "period" for the set C, given by the greatest common divisor (gcd) of E_C. It can be shown that C is ν_{nd}-small for all sufficiently large n, where $d = \gcd(E_C)$.

We show that this value is in fact a property of the whole chain X, and is independent of the particular small set chosen, in the following

Theorem 3.2.24 (Periodic cycle)**.** *Suppose that X is a ψ-irreducible Markov chain on S. Let $C \in \mathcal{B}^+(X)$ be a ν_M-small set and let d be the greatest common divisor of the set E_C. Then there exist disjoint sets $D_1, \ldots, D_d \in \mathcal{B}$ (a "d-cycle") such that:*

(i) For $x \in D_i$, $P(x, D_{i+1}) = 1$, $i = 0, \ldots, d-1$ (mod d);

(ii) The complementary set $N = [\bigcup_{i=1}^{d} D_i]^c$ is ψ-null, i.e., $\psi(N) = 0$.

The d-cycle $\{D_i\}$ is maximal in the sense that for any other collection $\{d', D'_k, k = 1, \ldots, d'\}$ satisfying (i)-(ii), we have d' dividing d; while for $d = d'$, then, by reordering the indices (if necessary), $D'_i = D_i$ ψ-almost everywhere.

Proof. For $i = 0, 1, \ldots, d-1$ set

$$D_i^* = \left\{ y : \sum_{n=1}^{\infty} P^{nd-i}(y, C) > 0 \right\},$$

by irreducibility, $S = \bigcup D_i^*$.

The D_i^* are in general not disjoint, but we can show that their intersection is ψ-null. Suppose there exists i, k such that $\psi(D_i^* \cap D_k^*) > 0$. Then for some fixed $m, n > 0$, there is a subset $A \subseteq D_i^* \cap D_k^*$ with $\psi(A) > 0$ such that

$$P^{md-i}(w, C) \geq \varepsilon_m > 0, \quad w \in A$$

$$P^{nd-k}(w, C) \geq \varepsilon_n > 0, \quad w \in A$$

and since ψ is the maximal irreducibility measure, we can also find r such that

$$\int_C \nu_M(dy) P^r(y, A) = \varepsilon_c > 0.$$

Now we use the fact that C is a ν_M-small set. For $x \in C$, $B \subseteq C$, one can derive

$$\begin{aligned} P^{2M+md-i+r}(x, B) &\geq \int_C P^M(x, dy) \int_A P^r(y, dw) \int_C P^{md-i}(w, dz) P^M(z, B) \\ &\geq [\varepsilon_c \varepsilon_m] \nu_M(B), \end{aligned}$$

so that $[2M+md+r]-i \in E_C$. By identical reasoning, we also have $[2M+nd+r]-k \in E_C$. This contradicts the definition of d, and we have shown that $\psi(D_i^* \cap D_k^*) = 0$, $i \neq k$.

Let $N = \bigcup_{i,j} (D_i^* \cap D_k^*)$, so that $\psi(N) = 0$. The sets $\{D_i^* \setminus N\}$ form a disjoint class of sets, for which the complementary of its union satisfies $\psi([\bigcup_i (D_i^* \setminus N)]^c) = 0$. We can find a set D such that $P(x, D) = 1$ for any $x \in D$ and $D_i = D \cap (D_i^* \setminus N)$ are disjoint and $D = \bigcup D_i$. By the Chapman-Kolmogorov equations, if $x \in D$ is such that $P(x, D_j) > 0$, then we have $x \in D_{j-1}$, by definition, for $j = 0, \ldots, d-1$ (mod d). Thus $\{D_i\}$ is a d-cycle.

To prove the maximality and uniqueness result, suppose $\{D'_i\}$ is another cycle with period d', with $N = [\bigcup D'_i]^c$ such that $\psi(N) = 0$. Let k be any index with $\nu_M(D'_k \cap C) > 0$, since $\psi(N) = 0$ and $\psi \geq \nu_M$, such a k exists. We then have, since C is a ν_M-small set,

$$P^M(x, D'_k \cap C) \geq \nu_M(D'_k \cap C) > 0 \quad \text{for every } x \in C.$$

Since $(D'_k \cap C)$ is non-empty, this implies that M is a multiple of d'; since this happens for any $n \in E_C$, by definition of d we have d' divides d as required. Also, we must have $C \cap D'_j$ empty for any $j \neq k$; if not we would have some $x \in C$ with $P^M(x, C \cap D'_k) = 0$, which contradicts the properties of C.

Hence we have $C \subseteq (D'_k \cup N)$, for some particular k. It follows by the definition of the original cycle that each D'_j is a union up to ψ-null sets of d/d_i elements of D_i. $\square$

From the proof, it is clear that the cycle does not depend, except perhaps for ψ-null sets, on the small set initially chosen, and that any small set must be essentially contained inside one specific member of the cyclic class $\{D_i\}$.

Definition 3.2.25 (Aperiodic, strongly aperiodic)**.** Suppose that X is a φ-irreducible Markov chain.

The largest d for which a d-cycle occurs for X is called the **period** of X.

When $d = 1$, the chain is called **aperiodic**.

When there exists a ν_1-small set A with $\nu_1(A) > 0$, then the chain is called **strongly aperiodic**.

Example 3.2.26 Using the same example as in the previous sections (3.2.11), we can show that $\{X_n\}$ is aperiodic. Suppose, in the contrary, that $\{X_n\}$ is periodic with periodic cycle $D_1, \ldots, D_d$.

Let $x \in D_1$, then

$$P(x, D_2) = \int_{y \in D_2} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}\left(y-\frac{x}{2}\right)^2} dy = 1.$$

It follows that

$$\int_{y \in D_2^c} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}\left(y-\frac{x}{2}\right)^2} dy = 0.$$

Since $0 < \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}\left(y-\frac{x}{2}\right)^2} < \infty$, and that the chain is $\lambda-$irreducible, we have $\lambda(D_2^c) = 0$, (where λ is the Lebesgue measure on $\mathbb{R}$). Since $D_1 \subset D_2^c$, $\lambda(D_1) = 0$. Hence, for $x \in D_d$,

$$P(x, D_1) = \int_{y \in D_1} \frac{2}{\sqrt{3}\sqrt{2\pi}} e^{-\frac{2}{3}\left(y-\frac{x}{2}\right)^2} dy = 0,$$

which contradicts periodicity.

3.2.5 Recurrence

In developing concepts of recurrence for sets $A \in \mathcal{B}$, we will consider the event that $X \in A$ infinitely often (i.o.) defined by

$$\{X \in A \text{ i.o.}\} := \bigcap_{N=1}^{\infty} \bigcup_{k=N}^{\infty} \{X_k \in A\}$$

which is well defined as an $\mathcal{B}^\infty$-measurable event on the path space S^∞. For $x \in S$, $A \in \mathcal{B}$ we write

$$Q(x, A) := P_x\{X \in A \text{ i.o.}\}$$

Definition 3.2.27 (Harris recurrence)**.** The set A is called *Harris recurrent* if

$$Q(x, A) = 1, \quad x \in A.$$

A chain X is called **Harris recurrent** (or just Harris) if it is ψ-irreducible and every set in $\mathcal{B}^+$ is Harris recurrent.

The following proposition provides a sufficient condition for a set to be Harris recurrent.

Proposition 3.2.28. *Given a set $A \in \mathcal{B}$, if there exists $n \in \mathbb{N}_{>0}$ such that $P^n(x, A) \equiv 1$, $x \in A$, then $Q(x, A) = P^n(x, A)$ for every $x \in S$, in particular A is Harris recurrent.*

Thanks to irreducibility and aperiodicity, we can now strengthen the connection between the chain X and its skeletons using the following theorem.

Theorem 3.2.29. *If the chain X is ψ-irreducible and aperiodic, then*

$$X \text{ is Harris} \iff \text{the skeletons } X^m \text{ are Harris for all } m.$$

To study the long-term behavior of the chain (and hence of the system), we will divide recurrent chains into two classes: the one of *positive* recurrent chains, which provides a strong kind of stability, and the one of *null* recurrent chains, which will not be of our interest.

The strongest form of stability is when the distribution of X_n remains unchanged for different n, which is exactly the case of stationary processes induced by the invariant measure π.

Definition 3.2.30 (Positive chains)**.** A ψ-irreducible chain X is called **positive** if it admits an invariant probability measure π. It is called **null** otherwise.

3.2.6 Convergence

In this concluding section of the chapter, we examine the convergence of distributions and their ergodic behavior by integrating all the concepts we've explored so far. Let us state the most important results of this section, which provide guarantees on the ergodic behavior of the chain.

Theorem 3.2.31 (Aperiodic Ergodic Theorem)**.** *Consider the chain X to be aperiodic Harris recurrent, with invariant measure π'. Then the following are equivalent:*

(i) X is positive Harris

(ii) π' is a finite measure

(iii) there exists a unique probability measure $\pi \in \mathcal{P}(S)$ such that, for every $x \in S$,

$$\sup_{A \in \mathcal{B}} |P^n(x, A) - \pi(A)| \longrightarrow 0 \quad \text{as } n \to \infty.$$

In particular, π is a constant multiple of π', so if any of the above conditions are met, we can always consider the invariant measure to be a unique probability measure.

Proof. [20, p. 314] □

Theorem 3.2.32. *If X is positive Harris recurrent and aperiodic, then for any initial measure μ*

$$\lim_{n \to \infty} \|\mu P^n - \pi\|_{TV} = 0.$$

In particular,

$$\lim_{n \to \infty} \mu^{[n]} = \lim_{n \to \infty} \mu P^n(A) = \pi(A) \quad \text{for all measurable } A \in \mathcal{B}.$$

Proof. [20, p. 328] □

Corollary 3.2.33. *If X is positive Harris recurrent and aperiodic, then X is ergodic in the sense of $(C^s(S))^*$ for $s \in [0,1]$.*

Proof. From the previous theorem, for any initial probability measure μ, we have convergence in total variation (TV) norm:

$$\lim_{n\to\infty} \|\mu P^n - \pi\|_{\mathrm{TV}} = 0.$$

Since $\|f\|_\infty \leq \|f\|_{C^s(S)}$ for all $f \in C^s(S)$ and S is compact, it follows that any function f with $\|f\|_{C^s(S)} \leq 1$ also satisfies $\|f\|_\infty \leq 1$.

Hence, we have

$$\|\mu\|_{(C^s(S))^*} = \sup_{\|f\|_{C^s(S)}\leq 1} |\mu(f)| \leq \sup_{\|f\|_\infty \leq 1} |\mu(f)| = \|\mu\|_{\mathrm{TV}}.$$

This implies

$$\|\mu P^n - \pi\|_{(C^s(S))^*} \leq \|\mu P^n - \pi\|_{\mathrm{TV}}.$$

Since we know $\|\mu P^n - \pi\|_{\mathrm{TV}} \to 0$ as $n \to \infty$, it follows that

$$\|\mu P^n - \pi\|_{(C^s(S))^*} \to 0.$$

Thus, convergence in TV norm implies convergence in $(C^s(S))^*$ norm. Since the chain is positive Harris recurrent and aperiodic, it converges to the unique stationary distribution π, and therefore it is ergodic in the sense of $(C^s(S))^*$. □

Remark 3.2.34 Once again, relating the chain with its skeletons, one can derive that if X is ψ-irreducible and aperiodic, then for each m, a measure π is invariant for X if and only if it is invariant for X^m.

Let us conclude with a result that not only guarantees ergodicity but also the desired uniformly geometric ergodicity.

Theorem 3.2.35 (Uniformly Geometric Ergodic Theorem)**.** *Suppose X ψ-irreducible and aperiodic. Then the following conditions are equivalent:*

(i) there exist some $\alpha < 1, R < \infty$ such that

$$\|P^n - \pi\|_{TV} \leq C\alpha^n;$$

(ii) for some $n \in \mathbb{N}_{>0}$,

$$\sup_{x\in S} \|P^n(x,\cdot) - \pi(\cdot)\| < 1;$$

(iii) X satisfies Doeblin's condition, i.e., $\exists \phi \in \mathcal{P}(S)$ such that for some integer m, $\varepsilon < 1, \delta > 0$,

$$\phi(A) > \varepsilon \implies P^m(x,A) \geq \delta,$$

for every $x \in S$.

Furthermore, any of these conditions implies that X is uniformly geometrically ergodic.

Proof. [20, p. 401] □

Corollary 3.2.36. *If X is positive Harris recurrent and aperiodic, then it is also* uniformly geometrically ergodic.

In the specific case of absolutely continuous Markov kernels, the conditions become significantly easier to verify, ensuring that the same result as in Theorem 3.2.35 holds.

Proposition 3.2.37. *If the transition probability function $P(x, \cdot)$ is absolutely continuous with respect to a strictly positive measure λ on S, that is, there exists a measurable function $\psi(x, y)$ such that $P(x, A) = \int_A \psi(x, y)\, d\lambda(y)$, $\forall x \in S$, $A \in \mathcal{B}(S)$, where $\psi(x, y) > 0$ for all $x, y \in S$, then the system is uniformly geometrically ergodic.*

Proof. [10, p. 249] and [30, p. 13] □

Chapter 4

Statistical Learning Bounds for SDS & MC

In the previous chapter we examined the theory of stochastic dynamical systems from a Markov chain perspective. We established some conditions that guarantee stability and convergence of marginal distributions starting from arbitrary initial measures, ensuring that under suitable assumptions the chain converges to a unique stationary distribution at a controlled rate.

Our goal in this chapter is to introduce a machine learning algorithm that, given a state $x \in S$, provides an approximation of the average system's next state, formally $\mathbb{E}[X_{t+1}|X_t = x]$, where X_{t+1} is the random variable with distribution $P(x, \cdot)$.

This chapter is based on the foundational work by Smale and Zhou on online learning with Markov sampling where the i.i.d. assumption is weakened [30], meaning that we will still assume our samples to be independent but not identically distributed. The techniques developed by Smale and Zhou provide a framework for learning from non-identically sampled data, using a reproducing kernel Hilbert space (RKHS) and a Mercer kernel to formulate an online learning algorithm. It's evident that to maintain some statistical guarantees when dealing with changing distributions, we require rather strong assumptions about how these distributions evolve and eventually converge. This motivates the study conducted in the previous chapter, as it provides us with convergence rates rather than merely asymptotic results.

In the first part of this chapter, we introduce the setting and the data generation process. Then we introduce the online learning algorithm (OLA). After discussing some assumptions on the learning problem we examine the main learning bounds for the algorithm, along with a detailed error analysis.

4.1 Setting

Let $(S, \mathcal{B})$ be a discrete-time autonomous systems, where $S = [a, b] \subset \mathbb{R}$ is the state space, and $\mathcal{B}$ is a σ-algebra on S. As in the previous chapter, we will focus on the time-homogenous Markov chain X characterizing the system, where $X = \{X_0, X_1, X_2, \dots\}$ is the process evolving by the Markov transition Kernel P.

Our goal is to learn, in a supervised setting, the function, defined pointwise as

$$f_\rho(x) = \mathbb{E}[X_{t+1}|X_t = x] = \int_S P(x, dy)y.$$

Remark 4.1.1 Recall Subsection 2.2.2, where we rely on Mercer's Theorem under the assumption that S is compact, ensuring T_K is a compact, self-adjoint, positive operator. However, compactness can often be replaced by other assumptions, such as *bounded support* of μ (and suitable continuity conditions), or *finite-trace*: $\iint K^2(x,x')\,d\mu(x)\,d\mu(x') < \infty$. Restricting ourselves to a sub-domain where K satisfies the above, then still guarantees a spectral decomposition analogous to Mercer's Theorem. Hence, requiring a compact domain is a classic but not exclusive choice to ensure T_K remains well-defined and admits an eigenfunction expansion.

4.1.1 Data Collection

When observing a dynamical system, it seems natural to look at how an initial state evolves over time and take these consecutive observations as the data for our learning algorithm. Recall that, a *trajectory* of the system is a realization of the stochastic process

$$X(\omega) = \{X_0(\omega), X_1(\omega), \dots\},$$

where ω is an event in Ω. We will always denote the *initial distribution* as $\mu = \mathcal{L}(X_0)$.

Suppose we have a number T of trajectories $X(\omega_0), X(\omega_1), \dots, X(\omega_{T-1})$, where $\omega_0, \dots, \omega_{T-1}$ are *independent events* in Ω.

Let us consider the sequence of *examples* $\{z_t\}$, defined by the trajectories above, as

$$z_t = (X_t(\omega_t), X_{t+1}(\omega_t)).$$

We will consider $\{z_t\}_{t=0,1,\dots,T-1}$ as our *data* and we will write $z_t = (x_t, x_{t+1})$ for ease of notation. Note that, the examples $\{z_t\}$ take values in $S \times S$ and are *independent* by construction:

$$\begin{array}{lll}
\underbrace{X_0(\omega_0),\, X_1(\omega_0)}_{z_0}, \dots & & \\
X_0(\omega_1),\, \underbrace{X_1(\omega_1),\, X_2(\omega_1)}_{z_1}, \dots & & \\
X_0(\omega_2),\, X_1(\omega_2),\, \underbrace{X_2(\omega_2),\, X_3(\omega_2)}_{z_2}, \dots & & \\
\vdots & & \ddots \\
X_0(\omega_t), & \dots & \underbrace{X_t(\omega_t),\, X_{t+1}(\omega_t)}_{z_t}, \dots
\end{array}$$

This approach addresses the challenge of handling dependent data, a common issue when working with dynamical systems (without relying on any additional conditions, e.g. mixing).

Moreover, we focus on trajectories rather than just a single evolution (e.g., $X_0 \to X_1$) to leverage the system's ergodicity assumption. This ensures that the algorithm's convergence does not depend on the initial distribution.

Remark 4.1.2 (Length of trajectories) In practice, the problem of sampling trajectories is highly task-dependent, and obtaining an arbitrary number of trajectories with arbitrary length is often impractical or even impossible.

Our data collection scheme requires each trajectory $X(\omega_i)$ to have a length $L_i \geq i$, meaning

$$X(\omega_i) = \{X_0(\omega_i), X_1(\omega_i), \dots, X_i(\omega_i), \dots, X_{L_i-1}(\omega_i)\}.$$

These truncated trajectories are sufficient to construct the dataset required for our online algorithm.

In the final chapter, we briefly explore how to decrease the number of required trajectories by introducing mixing conditions for the system. This method is useful in practical scenarios where obtaining multiple initializations is difficult, while acquiring longer trajectories, though in smaller amounts, is possible.

Remark 4.1.3 We can see the sample z_t as a realization of the random variable Z_t (with values in $S \times S$).

We have that $Z = \{Z_t\}_{t=0,\dots,T-1}$ are independent random variables such that, for $A, B \subseteq S$ measurable sets, we have

$$P_\mu\big(Z_t \in A \times B\big) = \int_A \mu^{[t-1]}(dx) P(x, B).$$

4.2 Online Learning Algorithm (OLA)

Consider a Mercer kernel $K : S \times S \to \mathbb{R}$ and $\mathcal{H}_K$ the corresponding RKHS. Learning is performed with the following regularized algorithm:

$$\begin{cases} f_0 & := 0, \\ f_{t+1} & := f_t - p_t\big((f_t(x_t) - x_{t+1})K_{x_t} - \lambda_t f_t\big), \quad t \geq 0 \end{cases} \tag{4.1}$$

where p_t is the step size parameter and λ_t is the regularization parameter. Note that the i.i.d. case corresponds to $z_t = (x_t, x_{t+1})$ being drawn from the same identical distribution, as the usual supervised setting discussed in Chapter 2.

We will show that under regularity assumptions, rapidly converging distributions and an appropriate choice for the parameters the algorithm effectively learns the regression function, i.e. $f_t \to f_\rho$.

4.2.1 Convergence of distributions

The convergence of the algorithm largely depends on the convergence of the marginal distributions $\mu^{[t]} = \mu P^t$, for which we will need an exponential rate of convergence in the dual of the Hölder space $C^s(S)$, with $s \in [0, 1]$:

$$\|\mu^{[t]} - \pi\|_{(C^s(X))^*} \leq C\alpha^t, \quad \forall t. \tag{4.2}$$

for some $C > 0$ and $0 < \alpha < 1$.

Recall that, this convergence is obtained by requiring the system to be *uniformly geometrically ergodic*, and that the convergence does not depend on the initial probability measure μ.

4.2.2 Regularity conditions

Now consider a fixed Hölder exponent $s \in [0, 1]$ and let us take a look at the assumption on the Kernel.

Definition 4.2.1 (Kernel Condition)**.** We say that the Mercer kernel K satisfies the **kernel condition (of order s)** if:

(i) $K \in C^s(S \times S)$,

(ii) there exists $\kappa_{2s} > 0$ such that, for all $u_1, u_2, v_1, v_2 \in S$,

$$|K(u_1, v_1) + K(u_2, v_2) - K(u_1, v_2) - K(u_2, v_1)| \leq \kappa_{2s} \left(d(u_1, v_1)\right)^s \left(d(u_2, v_2)\right)^s . \tag{4.3}$$

Remark 4.2.2 If the Kernel is C^2, then it satisfies the *Kernel condition*. Alternatively, one might assume other conditions, as mentioned in previous chapters (2.2.14), such as $\kappa = \sup_{x \in \mathcal{X}} \sqrt{K(x,x)} < \infty$; note that in our case this condition is implied by the other assumptions. In more generality, when $S \subset \mathbb{R}^n$ with smooth boundary and K is C^2, then the kernel condition holds. A proof can be found in [44].

Now let us introduce a regularity assumption on the target function, involving the integral operator $T_{K,\mu}$.

Remark 4.2.3 (Source condition) Recall from 2.2.2 that the integral operator $T_{K,\mu}$ on $\mathcal{H}_K$ is the covariance operator

$$T_{K,\mu} = \mathbb{E}_{x \sim \mu}\left[S_x^* S_x\right] = \int_S S_x^* S_x \, d\mu(x) = \int_S K_x \langle K_x, \cdot \rangle_K \, d\mu(x).,$$

where $S_x : \mathcal{H}_K \to \mathbb{R}$, $S_x(f) = f(x)$ is the sampling operator and S^* its adjoint.

We say that f_ρ satisfies the *regularity condition of order* r, with $1/2 < r \leq 3/2$, if there exists $g \in L^2_\mu(S)$, such that

$$f_\rho = T^r_{K,\mu} g. \tag{4.4}$$

See definition 2.2.18 for details.

We are now ready to state the main result.

4.2.3 Learning bounds

Let us first list the assumptions needed:

(A) **Exponential convergence:** The sequence $\{\mu^{[t]}\}$ converges exponentially in the sense of (4.2).

(B) **Kernel condition:** The kernel K satisfies the kernel condition (4.3).

(C) **Regularity condition:** The element f_ρ satisfies the regularity condition (4.4) with $1/2 < r \leq 3/2$.

Theorem 4.2.4 (Learning bounds)**.** *Under Assumptions (A–C) listed above, consider the* online learning algorithm (4.1) *with parameters*

$$\lambda_t := \lambda_0 (t+1)^{-\beta}, \quad p_t := p_0 (t+1)^{-\theta}, \tag{4.5}$$

where $\lambda_0, p_0 > 0$, $\theta \in (0,1)$, *and* $\beta \in (0, 1-\theta]$.

Then the following bounds hold for $t \in \mathbb{N}$*:*

$$\begin{aligned}
&\mathbb{E}\left[\|f_t - f_\rho\|_K\right] \leq \|g\|_\pi \lambda_0^{r-\frac{1}{2}} t^{-\beta(r-\frac{1}{2})} \\
&+ \begin{cases}
\left(p_0 C_1^* + \left(C\lambda_0^{r-\frac{3}{2}} + \lambda_0^{r-\frac{1}{2}}\right) C_2^*\right) t^{-\min\{\beta(r-\frac{1}{2}), \frac{\theta-\beta}{2}\}}, & \text{if } 0 < \beta < 1-\theta, \\
\left(p_0 C_1^* + \left(C\lambda_0^{r-\frac{3}{2}} + \lambda_0^{r-\frac{1}{2}}\right) C_2^*\right) t^{-\min\{\beta(r-\frac{1}{2}), \theta-\frac{1}{2}, p_0\lambda_0\}} \log(t+1), & \text{if } \beta = 1-\theta.
\end{cases}
\end{aligned} \tag{4.6}$$

Remark 4.2.5 The constants C_1^* and C_2^* depend on the kernel constant $k := \sup_{x\in S}\sqrt{K(x,x)}$, on κ_{2s}, α, β, θ, r, C_K, and on the product $p_0\lambda_0$.

Corollary 4.2.6. *If the Markov Chain X of the system is positive Harris recurrent and aperiodic, then condition (A) is implied by the Uniformly Geometric Ergodic Theorem 3.2.35.*

Remark 4.2.7 Moreover if $\beta = 0$, meaning the regularization parameter is fixed ($\lambda_t \equiv \lambda_0$), then

$$\mathbb{E}\left[\|f_t - f_\rho\|_K\right] \leq \|g\|_\pi \lambda_0^{r-\frac{1}{2}} + \begin{cases} \left(p_0 C_1^* + C\lambda_0^{r-\frac{3}{2}} C_2^*\right) t^{-\frac{\theta}{2}}, & \text{if } \alpha < 1, \\ p_0 C_1^* t^{-\frac{\theta}{2}} + C\lambda_0^{r-\frac{3}{2}} C_2^* t^{\theta}, & \text{if } \alpha = 1. \end{cases}$$

Note that in both cases, the algorithm is not guaranteed to converge. In the first case, where $\alpha < 1$, the bound asymptotically reaches the bias term $\|g\|_\pi \lambda_0^{r-1/2}$. In contrast, for $\alpha = 1$, which corresponds to the situation where there is no exponential convergence for the distributions, there is no assurance that the algorithm will converge. In this scenario, the bound states

$$\mathbb{E}\left[\|f_t - f_\rho\|_K\right] = O\big(t^{\theta}\big).$$

4.3 Error Analysis

4.3.1 Error Decomposition

The *offline (batch)* version in the i.i.d. case of our OLA would be:

$$f_{\lambda,\mu} := \arg\min_{f\in\mathcal{H}_K} \left\{ \int_S (f(x) - f_\rho(x))^2 \, d\mu(x) + \lambda \|f\|_K^2 \right\},$$

where μ replaces our $\mu^{[t]} = \mu P^t$. So in the OLA, there is an error caused by the changing measures $\{\mu^{[t]}\}$.

Remark 4.3.1 The function above can also be written as:

$$f_{\lambda,\mu} = (T_{K,\mu} + \lambda I)^{-1} T_{K,\mu} f_\rho. \tag{4.7}$$

The error can be decomposed into three parts:

$$f_{t+1} - f_\rho = \left(f_{t+1} - f_{\lambda_t,\mu^{[t]}}\right) + \left(f_{\lambda_t,\mu^{[t]}} - f_{\lambda_t,\pi}\right) + \left(f_{\lambda_t,\pi} - f_\rho\right). \tag{4.8}$$

- $f_{\lambda_t,\pi} - f_\rho$ is the *approximation error*;
- $f_{\lambda_t,\mu^{[t]}} - f_{\lambda_t,\pi}$ is the *drift error*. It depends on the measure difference, while the regularization parameter is the same λ_t;
- $f_{t+1} - f_{\lambda_t,\mu^{[t]}}$ is the *sample error*.

4.3.2 Approximation Error

Proposition 4.3.2. *If f_ρ satisfies the regularity condition, then $\forall\lambda > 0$:*

$$\|f_{\lambda,\pi} - f_\rho\|_K \leq \|g\|_\pi \lambda^{r-\frac{1}{2}}.$$

Proof. To ease up the notation: $T = T_{K,\pi}$.

$$(T + \lambda I) f_{\lambda,\pi} = T f_\rho.$$

Subtracting from both sides $(T + \lambda I) f_\rho$ leads to:

$$(T + \lambda I) f_{\lambda,\pi} - (T + \lambda I) f_\rho = -(\lambda I) f_\rho.$$

Hence,

$$f_{\lambda,\pi} - f_\rho = -\lambda (T + \lambda I)^{-1} f_\rho.$$

Splitting the power $-1 = (r - \frac{3}{2}) + (\frac{1}{2} - r)$ of the term $(T + \lambda I)$ and writing $f_\rho = T^r g = T^{r-\frac{1}{2}} T^{\frac{1}{2}} g$, we get:

$$f_{\lambda,\pi} - f_\rho = -\lambda (T + \lambda I)^{r-\frac{3}{2}} [(T + \lambda I)^{\frac{1}{2}-r} T^{r-\frac{1}{2}}] T^{\frac{1}{2}} g.$$

Now recall that $T^{\frac{1}{2}} : \mathcal{H}_K \hookrightarrow L^2_\pi(S)$ is an isometric isomorphism, since

$$\|T^{\frac{1}{2}} g\|_K^2 = \langle T^{\frac{1}{2}} g, T^{\frac{1}{2}} g \rangle_K,$$

and T being self-adjoint, yields

$$\langle Tg, g \rangle_K = \left\langle \int_S g(x) K_x d\pi,\ g \right\rangle_K.$$

Using the reproducing property we get

$$\int_S g^2(x) d\pi = \|g\|_\pi^2.$$

Thus,

$$\|T^{\frac{1}{2}} g\|_K = \|g\|_\pi.$$

Also observe that the eigenvalues of $(T + \lambda I)^{\frac{1}{2}-r} T^{r-\frac{1}{2}}$ are

$$\left(\frac{\sigma_i}{\sigma_i + \lambda} \right)^{r-\frac{1}{2}},$$

with $\{\sigma_i\}_{i=1}^\infty$ the eigenvalues of T (which is positive self-adjoint).

Since T is positive ($\sigma_i \geq 0$) and since $\frac{1}{2} < r \leq \frac{3}{2}$, we have:

$$\|(T + \lambda I)^{\frac{1}{2}-r} T^{r-\frac{1}{2}}\|_K \leq \sup_i \left| \frac{\sigma_i}{\sigma_i + \lambda} \right|^{r-\frac{1}{2}} \leq 1,$$

and

$$\|(T + \lambda I)^{r-\frac{3}{2}}\|_K \leq \lambda^{r-\frac{3}{2}}.$$

Combining the other bounds concludes the proof. □

4.3.3 Drift Error

It follows from the reproducing property $\langle f, K_x\rangle_K = f(x)$, that

$$|f(x)| \leq \|f\|_K \|K_x\|_K \leq k\|f\|_K \quad \text{with } k = \sup_{x\in S}\sqrt{K(x,x)}.$$

If K satisfies the kernel condition, then one could prove that:

$$\mathcal{H}_K \hookrightarrow C^s(S) \quad \text{with the immersion bounded by } \|f\|_{C^s(S)} \leq (k+\kappa_{2s})\|f\|_K,$$

with κ_{2s} coming from the *kernel condition*.

Proposition 4.3.3. *If $f_\rho \in C^s(S)$, and K satisfies the kernel condition, then:*

$$\|f_{\lambda,\mu} - f_{\lambda,\mu'}\|_K \leq \frac{C_K}{\lambda}\|\mu-\mu'\|_{(C^s(S))^*}\|f_{\lambda,\mu'} - f_\rho\|_{C^s(S)},$$

with

$$C_K = \sqrt{k^2 + 2|K|_{C^s(S\times S)} + \kappa_{2s}},$$

which is a constant depending on the Kernel K.

Corollary 4.3.4. *In particular, for $\mu^{[t]} \to \pi$ converging exponentially, and f_ρ satisfying the regularity condition, the following holds:*

$$\|f_{\lambda_t,\mu^{[t]}} - f_{\lambda_t,\pi}\|_K \leq \tilde{C}_K C\|g\|_\pi \alpha^t (t+1)^{-\beta(r-\frac{3}{2})},$$

with the parameters defined in Theorem 4.2.4 and

$$\tilde{C}_K := \frac{C_K(k+\kappa_{2s})}{\lambda_0^{\frac{3}{2}-r}}.$$

Proof of Proposition. Notation: Let $T_{K,\mu} = T_\mu$.

Having $(T_{\mu'} + \lambda I)f_{\lambda,\mu'} = T_{\mu'} f_\rho$, we can write:

$$(T_\mu + \lambda I)f_{\lambda,\mu} - (T_{\mu'} + \lambda I)f_{\lambda,\mu'} = (T_\mu - T_{\mu'})f_\rho.$$

Then, subtracting $(T_\mu + \lambda I)f_{\lambda,\mu'}$ from both sides and rearranging the terms, we get:

$$f_{\lambda,\mu} - f_{\lambda,\mu'} = (T_\mu + \lambda I)^{-1}[(T_\mu - T_{\mu'})f_\rho + T_{\mu'}f_{\lambda,\mu'} - T_\mu f_{\lambda,\mu'}].$$

Simplifying further:

$$f_{\lambda,\mu} - f_{\lambda,\mu'} = (T_\mu + \lambda I)^{-1}(T_\mu - T_{\mu'})(f_\rho - f_{\lambda,\mu'}).$$

Taking norms:

$$\|f_{\lambda,\mu} - f_{\lambda,\mu'}\|_K \leq \frac{1}{\lambda}\|(T_\mu - T_{\mu'})(f_\rho - f_{\lambda,\mu'})\|_K. \tag{4.9}$$

Now call $f = f_\rho - f_{\lambda,\mu'} \in C^s(S)$, since $f_{\lambda,\mu'} \in \mathcal{H}_K \subseteq C^s(S)$.

We estimate $\|(T_\mu - T_{\mu'})f\|_K^2$ in the following by first writing:

$$\|(T_\mu - T_{\mu'})f\|_K^2 = \langle \int_S f(u)K_u d(\mu-\mu')(u), \int_S f(v)K_v d(\mu-\mu')(v)\rangle_K,$$

so, expanding:

$$\|(T_\mu - T_{\mu'})f\|_K^2 = \int_S f(u)\left[\int_S f(v)K(v,u)d(\mu-\mu')(v)\right]d(\mu-\mu')(u).$$

Define the auxiliary function

$$h(u) := f(u)\left[\int_S f(v)K(v,u)d(\mu-\mu')(v)\right].$$

Since the space of measures $\mathcal{M}(S) \subseteq (C^s(S))^*$, we get:

$$\|(T_\mu - T_{\mu'})f\|_K^2 \leq \|\mu-\mu'\|_{(C^s(S))^*}\|h\|_{C^s(S)}.$$

Estimating $\|h\|_{C^s(S)}$

Recall that

$$\|h\|_{C^s(S)} = \|h\|_{C(S)} + |h|_{C^s(S)}, \quad \text{where} \quad |h|_{C^s(S)} := \sup_{u\neq v}\frac{|h(u)-h(v)|}{d(u,v)^S}.$$

1. Estimating $\|h\|_{C(S)}$

$$\|h\|_{C(S)} \leq \|f\|_{C(S)} \cdot \sup_{u\in S}\left|\int_S f(v)K(v,u)d(\mu-\mu')(v)\right|.$$

2. Estimating $|h|_{C^s(S)}$

$$|h|_{C^s(S)} = \left|f\cdot\int_S f(v)K(v,u)d(\mu-\mu')(v)\right|_{C^s(S)}$$

$$\leq |f|_{C^s(S)}\sup_{u\in S}\left|\int_S f(v)K(v,u)d(\mu-\mu')(v)\right| + \|f\|_{C(S)}\sup_{u\in S}\left|\int_S f(v)K(v,u)d(\mu-\mu')(v)\right|_{C^s(S)},$$

where the inequality is given by the following remark.

Remark 4.3.5 Note that

$$|h_1h_2|_{C^s(S)} = \sup\frac{|h_1(x)h_2(x) - h_1(y)h_2(y)|}{d(x,y)^s}.$$

Expanding the difference,

$$|h_1h_2|_{C^s(S)} \leq \sup\frac{|h_1(x)-h_1(y)|}{d(x,y)^s}\cdot\sup|h_2| + \sup\frac{|h_2(x)-h_2(y)|}{d(x,y)^s}\cdot\sup|h_1|,$$

thus,

$$|h_1h_2|_{C^s(S)} \leq |h_1|_{C^s(S)}\sup|h_2| + |h_2|_{C^s(S)}\sup|h_1|.$$

For the first estimation we have:

$$\sup_{u\in S}\left|\int_S f(v)K(v,u)d(\mu-\mu')(v)\right| \leq \|\mu-\mu'\|_{(C^s(S))^*}\|f(\cdot)K(u,\cdot)\|_{C^s(S)},$$

with

$$\|f(\cdot)K(u,\cdot)\|_{C^s(S)} = \|f\cdot K(u,\cdot)\|_{C(S)} + |f\cdot K(u,\cdot)|_{C^s(S)} \leq \|f\|_{C(S)}k^2 + |f|_{C^s(S)}k^2 + \|f\|_{C(S)}\kappa_{2s},$$

where we applied the remark once again for the inequality.

For the second term, we focus on the $C^s(X)$-seminorm:

$$\left|\int_S f(v)K(v,u)\,d(\mu-\mu')(v)\right|_{C^s(X)} = \sup_{u\neq u'} \frac{\left|\int_S f(v)[K(u,v)-K(u',v)]\,d(\mu-\mu')(v)\right|}{d(u,u')^s}.$$

$$\leq \sup_{u\neq u'} \|\mu-\mu'\|_{(C^s(X))^*} \left\| f(\cdot)\frac{K(\cdot,u)-K(\cdot,u')}{d(u,u')^s}\right\|_{C^s(X)}.$$

As before, the last $C^s(X)$ norm is bounded:

$$\|f(\cdot)(K(\cdot,u)-K(\cdot,u'))/d(u,u')^s\|_{C^s(X)} \leq \|f\|_{C^s(X)}(|K|_{C^s(S\times S)}+\kappa_{2s}).$$

This gives us the upper bound for the second term.

Combining (1) and (2), we get:

$$\|h\|_{C^s(S)} \leq \|\mu-\mu'\|_{(C^s(S))^*}\|f\|^2_{C^s(S)}\left(k^2+\kappa_{2s}+2|K|_{C^s(S\times S)}\right).$$

Finally, using what we derived at the beginning, we have:

$$\|(T_\mu - T_{\mu'})f\|^2_K \leq \|\mu-\mu'\|^2_{(C^s(S))^*}\|f\|^2_{C^s(S)}\left(k^2+\kappa_{2s}+2|K|_{C^s(S\times S)}\right).$$

Together with (4.9), this ends the proof. □

Proof of the Corollary. We start by recalling the results we have established so far:

From Proposition 1, we have:

$$\|f_{\lambda,\mu}-f_{\lambda,\mu'}\|_K \leq \frac{C_K}{\lambda}\|\mu-\mu'\|_{(C^s(X))^*}\|f_{\lambda,\mu'}-f_\rho\|_{C^s(X)}.$$

From the exponential convergence:

$$\|\mu^{[t]}-\pi\|_{(C^s(X))^*} \leq C\alpha^t.$$

By the remark on the embedding $\mathcal{H}_K \hookrightarrow C^s(X)$:

$$\|f\|_{C^s(X)} \leq (k+\kappa_{2s})\|f\|_K.$$

For the approximation error derived previously:

$$\|f_{\lambda,\pi}-f_\rho\|_K \leq \|g\|_\pi \lambda^{r-\frac{1}{2}}.$$

Recall the parameters: $\lambda_t := \lambda_0(t+1)^{-\beta}$ with $\lambda_0 > 0$ and $\beta \in [0, 1-\theta]$.

Combining the above results:

$$\|f_{\lambda_t,\mu^{[t]}}-f_{\lambda_t,\pi}\|_K \leq \frac{C_K}{\lambda_t}\|\mu^{[t]}-\pi\|_{(C^s(X))^*}\|f_{\lambda_t,\pi}-f_\rho\|_{C^s(X)},$$

substituting the exponential convergence of $\mu^{[t]} \to \pi$ and using the embedding inequality $\|f_{\lambda_t,\pi}-f_\rho\|_{C^s(X)} \leq (k+\kappa_{2s})\|f_{\lambda_t,\pi}-f_\rho\|_K$,

$$\leq \frac{C_K}{\lambda_t}(C\alpha^t)(k+\kappa_{2s})\|f_{\lambda_t,\pi}-f_\rho\|_K.$$

Now use the approximation error bound we get

$$\leq C_K C\alpha^t(k+\kappa_{2s})\|g\|_\pi \lambda_t^{r-\frac{3}{2}}.$$

Setting $\tilde{C}_K := C_K(k+\kappa_{2s})\lambda_0^{r-\frac{3}{2}}$ and substituting the parameters, we recover the bound stated in the corollary. □

4.3.4 Sample error

We study by iteration what changes when going from f_t to f_{t+1}. This involves examining the transition from $f_{\lambda_{t-1},\mu^{[t-1]}}$ to $f_{\lambda_t,\mu^{[t]}}$.

We have:

$$f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_t,\mu^{[t]}} = \big(f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_{t-1},\pi} + f_{\lambda_t,\pi} - f_{\lambda_t,\mu^{[t]}}\big) + \big(f_{\lambda_{t-1},\pi} - f_{\lambda_t,\pi}\big). \quad (4.10)$$

Since the regularization parameter changes with each step, we need some additional results.

Proposition 4.3.6 (Tarres and Yao)**.** *If f_ρ satisfies the regularity condition and $\lambda, \lambda' > 0$, then*

$$\|f_{\lambda,\pi} - f_{\lambda',\pi}\|_K \leq |\lambda^{r-\frac{1}{2}} - (\lambda')^{r-\frac{1}{2}}| \, \|g\|_\pi (r-\tfrac{1}{2})^{-1}.$$

See [36] for more details.

Corollary 4.3.7. *Assuming all the convergence, regularity, and kernel conditions, as well as the parameters given in Theorem 4.2.4, we have that for each t:*

$$\|f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_t,\mu^{[t]}}\|_K \leq \begin{cases} 4\|g\|_\pi (\tilde{C}_K C\, \alpha^{t-1}(t+1)^{\beta(\frac{3}{2}-r)} + \lambda_0^{r-\frac{1}{2}}(t+1)^{-\beta(r-\frac{1}{2})-1}), & \text{if } \beta > 0, \\ 2\|g\|_\pi \tilde{C}_K C\, \alpha^{t-1}, & \text{if } \beta = 0. \end{cases}$$

Proof. We start from the identity:

$$f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_t,\mu^{[t]}} = \big(f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_{t-1},\pi} + f_{\lambda_t,\pi} - f_{\lambda_t,\mu^{[t]}}\big) + \big(f_{\lambda_{t-1},\pi} - f_{\lambda_t,\pi}\big).$$

Taking the $\|\cdot\|_K$ norm, we have:

$$\|f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_t,\mu^{[t]}}\|_K \leq \|f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_{t-1},\pi}\|_K + \|f_{\lambda_t,\pi} - f_{\lambda_t,\mu^{[t]}}\|_K + \|f_{\lambda_{t-1},\pi} - f_{\lambda_t,\pi}\|_K.$$

The first two terms on the right-hand side can be bounded using Corollary 4.3.4, while the last term is controlled by Proposition 4.3.6.

If $\beta = 0$, then $\lambda_t \equiv \lambda_0$ is constant. Hence, the last term is zero. Thus, combining the first two terms, which also become simpler, we deduce:

$$\|f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_t,\mu^{[t]}}\|_K \leq \|g\|_\pi \tilde{C}_K C\alpha^{t-1} + \|g\|_\pi \tilde{C}_K C\alpha^t \leq 2\|g\|_\pi \tilde{C}_K C\alpha^{t-1}.$$

When $\beta > 0$, $\lambda_t = \lambda_0(t+1)^{-\beta}$ introduces a more complex time dependence:

$$\|f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_{t-1},\pi}\|_K \leq \tilde{C}_K C\|g\|_\pi \alpha^{t-1} t^{\beta(\frac{3}{2}-r)}$$

$$\|f_{\lambda_t,\pi} - f_{\lambda_t,\mu^{[t]}}\|_K \leq \tilde{C}_K C\|g\|_\pi \alpha^t (t+1)^{\beta(\frac{3}{2}-r)}$$

$$\|f_{\lambda_{t-1},\pi} - f_{\lambda_t,\pi}\|_K \leq |\lambda_{t-1}^{r-\frac{1}{2}} - \lambda_t^{r-\frac{1}{2}}| \|g\|_\pi (r-\tfrac{1}{2})^{-1}$$

Using again $\alpha^t \leq \alpha^{t-1}$ and $t^{\beta(\frac{3}{2}-r)} \leq (t+1)^{\beta(\frac{3}{2}-r)}$ (since $r \leq \frac{3}{2}$), we get:

$$\|f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_{t-1},\pi}\|_K + \|f_{\lambda_t,\pi} - f_{\lambda_t,\mu^{[t]}}\|_K \leq 2\tilde{C}_K C\|g\|_\pi \alpha^{t-1}(t+1)^{\beta(\frac{3}{2}-r)}.$$

Let us now focus on the last term. By the Mean Value Theorem applied to the auxiliary function $h(t) := \lambda_t^{r-\frac{1}{2}} = (\lambda_0(t+1)^{-\beta})^{r-\frac{1}{2}}$, for some $t_* \in (t-1, t)$ we have

$$\left|\lambda_{t-1}^{r-\frac{1}{2}} - \lambda_t^{r-\frac{1}{2}}\right| = \left|h'(t_*)\right| = \lambda_0^{r-\frac{1}{2}} (t_*+1)^{-\beta(r-\frac{1}{2})-1}(r-\tfrac{1}{2})\beta.$$

Since $t < t_* + 1$ and $0 < \beta < 1$, we obtain

$$\left|\lambda_{t-1}^{r-\frac{1}{2}} - \lambda_t^{r-\frac{1}{2}}\right|(r-\tfrac{1}{2})^{-1} \leq \lambda_0^{r-\frac{1}{2}}\, t^{-\beta(r-\frac{1}{2})-1}.$$

Hence, the last term is bounded as

$$\|f_{\lambda_{t-1},\pi} - f_{\lambda_t,\pi}\|_K \leq \lambda_0^{r-\frac{1}{2}}\, \|g\|_\pi t^{-\beta(r-\frac{1}{2})-1}$$

Since $t > \frac{t+1}{2}$ for $t \geq 2$, it follows that

$$t^{-\beta(r-\frac{1}{2})-1} \;<\; \left(\tfrac{t+1}{2}\right)^{-\beta(r-\frac{1}{2})-1} = 2^{\beta(r-\frac{1}{2})+1}\,(t+1)^{-\beta(r-\frac{1}{2})-1} \leq 4\,(t+1)^{-\beta(r-\frac{1}{2})-1},$$

where we used $\beta(r-\frac{1}{2}) + 1 \leq 2$.

Putting together the bounds above and adjusting the constants to group the terms, we can conclude the proof. □

Let us state the following technical results that will be used in the subsequent proof.

Lemma 4.3.8 (Technical Lemma)**.** *(a) For $c, a > 0$, there holds:*

$$\exp(-cx) \leq \left(\frac{a}{ec}\right)^a x^{-a}, \quad \forall x > 0. \tag{4.11}$$

(b) Let $c > 0$ and $q_2 \geq 0$. If $0 < q_1 < 1$, then for any $t \in \mathbb{N}$ we have:

$$\sum_{i=0}^{t-2}(i+1)^{-q_2}\exp\Big(-c\sum_{j=i+2}^{t+1} j^{-q_1}\Big) \leq \left(\frac{2^{q_1+q_2}}{c} + \left(\frac{1+q_2}{ec(1-2^{q_1-1})}\right)^{\frac{1+q_2}{1-q_1}}\right)(t+1)^{q_1-q_2}. \tag{4.12}$$

In particular, for $q_1 = 1$, we have:

$$\sum_{i=0}^{t-2}(i+1)^{-q_2}\exp\Big(-c\sum_{j=i+2}^{t+1} j^{-1}\Big) \leq \begin{cases} \frac{2^{q_2}}{|c-q_2+1|}t^{-\min\{c,q_2-1\}}, & \text{if } c \neq q_2-1, \\ 2^{q_2}(t+1)^{-c}\log(t+2), & \text{if } c = q_2-1. \end{cases} \tag{4.13}$$

Proof. See [30]. □

Before stating the full proof of Theorem 4.2.4, let us look at an outline first.

Outline of the Proof. Recall from (4.8) that the total error $f_{t+1} - f_\rho$ splits into three parts:

$$\underbrace{\left(f_{t+1} - f_{\lambda_t,\mu^{[t]}}\right)}_{\textit{sample error}} + \underbrace{\left(f_{\lambda_t,\mu^{[t]}} - f_{\lambda_t,\pi}\right)}_{\textit{drift error}} + \underbrace{\left(f_{\lambda_t,\pi} - f_\rho\right)}_{\textit{approximation error}}.$$

In the preceding sections, we derived bounds for the approximation error, which is bounded by a regularization argument (Proposition 4.3.2) and the source condition is utilized to demonstrate that f_ρ is approximable in the Reproducing Kernel Hilbert Space (RKHS). The drift error is estimated using the measure-difference result (Corollary 4.3.4), which relies on the exponential convergence of $\{\mu^{[t]}\}$ toward π in the dual of Hölder space $C^s(S)$. The remaining task is to address the sample error. To accomplish this, the proof is divided as follows.

1. **Sample Error Bound** First we set $W_{t+1} = f_{t+1} - f_{\lambda_t,\mu^{[t]}}$ and rewrite it via a one-step recursion of the online algorithm. By iterating this recursion and separating out certain operators (A_t) and correction terms (χ_t), we decompose W_{t+1} in equation (4.15) into two main summations:
   - One summation captures how changes in the measure $\mu^{[t]}$ affect $f_{\lambda_t,\mu^{[t]}}$; tools like Proposition 4.3.6 and Lemma 4.12 provide estimates under exponential decay conditions for this summation.
   - The other summation consists of stochastic increments χ_t; we exploit the independence of samples $\{z_t\}$ to show that cross-terms vanish in expectation, reducing the analysis to a diagonal sum of $\|\chi_i\|_K^2$. Additional bounds follow from uniform control of $\|f_{\lambda_i,\mu^{[i]}}\|_K$ and a final application of the technical lemmas. (The analysis for this summation is also known as *Reverse Martingale Decomposition.*)
2. **Combining All Bounds.** After bounding these two summations, we add the bounds for the approximation error and the drift error. A final application of the triangle inequality yields the stated convergence rates, with explicit constants C_1^*, C_2^* given by combining the stepwise estimates.

Proof of Theorem 4.2.4. Denote the *sample error* term of the error decomposition as:

$$W_{t+1} = f_{t+1} - f_{\lambda_t,\mu^{[t]}}.$$

The *first step of the proof* is to establish a simple expression for W_{t+1}, by iterating a one-step recursion.

From the definition of the sampling operator, we notice that $x_{t+1}K_{x_t} = S_{x_t}^* x_{t+1}$ and $f_t(x_t)K_{x_t} = S_{x_t}(f_t)K_{x_t} = S_{x_t}^* S_{x_t}(f_t)$. Then, by definition of f_t, we know that:

$$W_{t+1} = f_t - f_{\lambda_t,\mu^{[t]}} - p_t\Big(S_{x_t}^* S_{x_t}(f_t) - S_{x_t}^* x_{t+1} + \lambda_t f_t\Big).$$

Expanding the terms we get:

$$W_{t+1} = f_t - f_{\lambda_t,\mu^{[t]}} - p_t\Big(S_{x_t}^* S_{x_t}(f_t - f_{\lambda_t,\mu^{[t]}}) + S_{x_t}^* S_{x_t} f_{\lambda_t,\mu^{[t]}} - S_{x_t}^* x_{t+1} + \lambda_t f_t\Big).$$

Grouping in terms of $f_t - f_{\lambda_t,\mu^{[t]}}$, we write $\lambda_t f_t$ as $\lambda_t(f_t - f_{\lambda_t,\mu^{[t]}}) + \lambda_t f_{\lambda_t,\mu^{[t]}}$. Definition (4.7) with the measure $\mu^{[t]}$ yields $\lambda_t f_{\lambda_t,\mu^{[t]}} = T_{K,\mu^{[t]}}(f_\rho - f_{\lambda_t,\mu^{[t]}})$. Therefore, we have:

$$W_{t+1} = \Big((1-p_t\lambda_t)I - p_t S_{x_t}^* S_{x_t}\Big)(f_t - f_{\lambda_t,\mu^{[t]}}) - p_t\Big(S_{x_t}^* S_{x_t} f_{\lambda_t,\mu^{[t]}} - S_{x_t}^* x_{t+1} + T_{K,\mu^{[t]}}(f_\rho - f_{\lambda_t,\mu^{[t]}})\Big).$$

Denote:

$$A_t = (1 - p_t\lambda_t)I - p_t S_{x_t}^* S_{x_t}, \quad \chi_t = p_t\Big(S_{x_t}^* S_{x_t} f_{\lambda_t,\mu^{[t]}} - S_{x_t}^* x_{t+1} + T_{K,\mu^{[t]}}(f_\rho - f_{\lambda_t,\mu^{[t]}})\Big).$$

Observe that $\mathbb{E}\big[S_{x_t}^* S_{x_t} f_{\lambda_t,\mu^{[t]}}\big] = T_{K,\mu^{[t]}} f_{\lambda_t,\mu^{[t]}}$, and $\mathbb{E}\big[S_{x_t}^* x_{t+1}\big] = \mathbb{E}\big[K_{x_t} x_{t+1}\big] = T_{K,\mu^{[t]}} f_\rho$, hence

$$\begin{aligned}\mathbb{E}[\chi_t] &= p_t\,\mathbb{E}\big[S_{x_t}^* S_{x_t} f_{\lambda_t,\mu^{[t]}} - S_{x_t}^* x_{t+1} + T_{K,\mu^{[t]}}(f_\rho - f_{\lambda_t,\mu^{[t]}})\big] \\ &= p_t\left(\mathbb{E}\big[S_{x_t}^* S_{x_t} f_{\lambda_t,\mu^{[t]}}\big] - T_{K,\mu^{[t]}} f_{\lambda_t,\mu^{[t]}} + T_{K,\mu^{[t]}} f_\rho - \mathbb{E}\big[S_{x_t}^* x_{t+1}\big]\right) \\ &= 0 \end{aligned} \tag{4.14}$$

By definition of W_t, we have

$$W_t = f_t - f_{\lambda_{t-1},\mu^{[t-1]}} \quad \text{and} \quad f_t - f_{\lambda_t,\mu^{[t]}} = W_t + \Big(f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_t,\mu^{[t]}}\Big).$$

If we denote $f_{\lambda_{-1},\mu^{[-1]}} = 0$, then there holds:

$$W_{t+1} = A_t W_t + A_t\Big(f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_t,\mu^{[t]}}\Big) - \chi_t, \quad \forall t \in \mathbb{N}.$$

Denote $\Pi_i = A_t A_{t-1} \cdots A_i$ and $\Pi_{t+1} = I$. Since $f_0 = 0$ gives $W_0 = 0$, by iteration we obtain:

$$W_{t+1} = \sum_{i=0}^{t} \Pi_i\Big(f_{\lambda_{i-1},\mu^{[i-1]}} - f_{\lambda_i,\mu^{[i]}}\Big) - \sum_{i=0}^{t} \Pi_{i+1}\chi_i, \quad \forall t \in \mathbb{N}. \tag{4.15}$$

The operator $p_i\lambda_i I + p_i S_{x_i}^* S_{x_i}$ is positive and bounded by $(p_i\lambda_i + p_i k^2)I$. So for $i \geq t_0$, the smallest integer greater than $(p_0\lambda_0 + p_0 k^2)^{1/\theta}$, the operator $A_i : \mathcal{H}_K \to \mathcal{H}_K$ is positive and bounded by $(1 - p_i\lambda_i)I$. Hence $\|A_i\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq 1 - p_i\lambda_i \leq \exp(-p_i\lambda_i)$. For $i < t_0$, $\|A_i\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq 1 + p_i\lambda_i + p_i k^2$. It follows that the operator norm of Π_i satisfies:

$$\|\Pi_i\|_{\mathcal{H}_K \to \mathcal{H}_K} \leq D_0 \exp\Big(-p_0\lambda_0 \sum_{j=i}^{t} j^{-\beta-\theta}\Big), \quad \forall\, 0 \leq i \leq t, \tag{4.16}$$

where D_0 is the constant given by:

$$D_0 = (1 + p_0\lambda_0 + p_0 k^2)^{(p_0\lambda_0 + p_0 k^2)^{1/\theta}} \exp\Big(p_0\lambda_0 (p_0\lambda_0 + p_0 k^2)^{1/\theta}\Big).$$

The *second step of the proof* is to bound the first term in (4.15). Apply Corollary 4.3.7 and (4.16). We find that:

$$\Big\|\sum_{i=0}^{t} \Pi_i\Big(f_{\lambda_{i-1},\mu^{[i-1]}} - f_{\lambda_i,\mu^{[i]}}\Big)\Big\|_K \tag{4.17}$$

$$\leq \begin{cases} 4\|g\|_\pi D_0 \sum_{i=0}^{t} \exp\big(-p_0\lambda_0 \sum_{j=i}^{t} j^{-\beta-\theta}\big)\Big(\tilde{C}_K C \alpha^{i-1}(i+1)^{\beta(\frac{3}{2}-r)} + \lambda_0^{r-\frac{1}{2}}(i+1)^{-\beta(r-\frac{1}{2})-1}\Big), \\ 2\|g\|_\pi D_0 \sum_{i=0}^{t} \exp\Big(-p_0\lambda_0 \sum_{j=i}^{t} j^{-\theta}\Big)\tilde{C}_K C \alpha^{i-1}, \end{cases} \tag{4.18}$$

for $\beta > 0$ and $\beta = 0$ respectively.

Consider the case $\beta > 0$ and $\alpha < 1$. Because the exponential decay is faster than any polynomial decay, we know that the term with $\alpha^i i^{\beta(\frac{3}{2}-r)}$ is dominated by the polynomial term $i^{-\beta(r-\frac{1}{2})-1}$. In fact, by Lemma 4.11 with $c = \log(1/\alpha)$ and $a = 2$, we have:

$$\alpha^i = \exp\big(-i\log(1/\alpha)\big) \leq \left(\frac{2}{e\log(1/\alpha)}\right)^2 i^{-2}. \tag{4.19}$$

For each $i \in \mathbb{N}$,

$$\alpha^{i-1}(i+1)^{\beta(\frac{3}{2}-r)} \leq \left(\frac{4}{e\log(1/\alpha)}\right)^2 (i+1)^{-\beta(r-\frac{1}{2})-1}.$$

It follows that:

$$\left\|\sum_{i=0}^{t}\Pi_i\Big(f_{\lambda_{i-1},\mu^{[i-1]}}-f_{\lambda_i,\mu^{[i]}}\Big)\right\|_K \le D_2\sum_{i=0}^{t}(i+1)^{-\beta(r-\frac{1}{2})-1}\exp\Big(-p_0\lambda_0\sum_{j=i}^{t}j^{-\beta-\theta}\Big),$$

where D_2 is the constant:

$$D_2 = 4\|g\|_\pi D_0\Big(\tilde{C}_K C\left(\frac{4}{e\log(1/\alpha)}\right)^2+\lambda_0^{r-\frac{1}{2}}\Big).$$

Applying Lemma 4.12 with $c=p_0\lambda_0$, $q_2=\beta(r-\frac{1}{2})+1$, and $q_1=\beta+\theta$, we obtain a bound for the first term of (4.15) as:

$$\left\|\sum_{i=0}^{t}\Pi_i\Big(f_{\lambda_{i-1},\mu^{[i-1]}}-f_{\lambda_i,\mu^{[i]}}\Big)\right\|_K \le \begin{cases} D_4(t+1)^{-\beta(r-\frac{1}{2})-1+\beta+\theta}, & \text{if } \beta+\theta<1,\\ D_4(t+1)^{-\min\{\beta(r-\frac{1}{2}),p_0\lambda_0\}}, & \text{if } \beta+\theta=1,\ p_0\lambda_0\neq\beta(r-\frac{1}{2}),\\ D_4(t+1)^{-\beta(r-\frac{1}{2})}\log(t+2), & \text{if } \beta+\theta=1,\ p_0\lambda_0=\beta(r-\frac{1}{2}),\end{cases} \tag{4.20}$$

where D_4 is the constant given by $D_4=D_2D_3$ with:

$$D_3 := \begin{cases} \frac{8}{p_0\lambda_0}+1+\left(\frac{2+\beta(r-\frac{1}{2})}{ep_0\lambda_0(1-2^{\beta+\theta-1})}\right)^{\frac{2+\beta(r-\frac{1}{2})}{1-\beta-\theta}}, & \text{if } \beta+\theta<1,\\ \frac{4}{|p_0\lambda_0-\beta(r-\frac{1}{2})|}+1, & \text{if } \beta+\theta=1 \text{ and } p_0\lambda_0\neq\beta(r-\frac{1}{2}),\\ 5, & \text{if } \beta+\theta=1 \text{ and } p_0\lambda_0=\beta(r-\frac{1}{2}).\end{cases}$$

The case $\beta=0$ is easier. We apply (4.19) when $\alpha<1$. Lemma 4.12 with $c=p_0\lambda_0$ and $q_1=\theta$ yields:

$$\left\|\sum_{i=0}^{t}\Pi_i\Big(f_{\lambda_{i-1},\mu^{[i-1]}}-f_{\lambda_i,\mu^{[i]}}\Big)\right\|_K \le \begin{cases} D_4(t+1)^{-1}, & \text{if } \beta=0,\alpha<1,\\ D_4(t+1)^{\theta}, & \text{if } \beta=0,\alpha=1,\end{cases}$$

where the constant D_4 is given by $D_4=2\|g\|_\pi D_0\tilde{C}_K C D_3$ with:

$$D_3 := \begin{cases} \left(\frac{4}{e\log(1/\alpha)}\right)^2\left(\frac{8}{p_0\lambda_0}+1+\left(\frac{3}{ep_0\lambda_0(1-2^{\theta-1})}\right)^{\frac{3}{1-\theta}}\right), & \text{if } \beta=0,\alpha<1,\\ \frac{2}{p_0\lambda_0}+1+\left(\frac{1}{ep_0\lambda_0(1-2^{\theta-1})}\right)^{\frac{1}{1-\theta}}, & \text{if } \beta=0,\alpha=1.\end{cases}$$

The *third step of the proof* is to estimate the second term of (4.15), which is

$$\left\|\sum_{i=0}^{t}\Pi_{i+1}\chi_i\right\|_K.$$

First, expand

$$\left\|\sum_{i=0}^{t}\Pi_{i+1}\chi_i\right\|_K^2 = \sum_{i=0}^{t}\sum_{\ell=0}^{t}\big\langle\Pi_{i+1}\chi_i,\ \Pi_{\ell+1}\chi_\ell\big\rangle_K. \tag{4.21}$$

Define $\mathcal{F}_i^t := \sigma(z_i,\dots,z_t)$. Using the law of total expectation (see Proposition 1.3.9), we have

$$\mathbb{E}\big[\langle\Pi_{i+1}\chi_i,\ \Pi_{\ell+1}\chi_\ell\rangle_K\big] = \mathbb{E}\Big[\mathbb{E}\big[\langle\Pi_{i+1}\chi_i,\ \Pi_{\ell+1}\chi_\ell\rangle_K \mid \mathcal{F}_{i+1}^t\big]\Big].$$

By construction, χ_i is a function solely of the variable z_i, and Π_{i+1} only of the random variables $\{z_{i+1}, z_{i+2}, \dots, z_t\}$. Therefore, for $\ell > i$ we have

$$\mathbb{E}\Big[\mathbb{E}\big[\langle \Pi_{i+1}\,\chi_i,\ \Pi_{\ell+1}\,\chi_\ell\rangle_K \ \big|\ \mathcal{F}_{i+1}^t\big]\Big] \ = \ \mathbb{E}\Big[\langle \mathbb{E}[\Pi_{i+1}\chi_i \ \big|\ \mathcal{F}_{i+1}^t], \Pi_{\ell+1}\,\chi_\ell\rangle_K\Big],$$

where we use linearity of the inner product, linearity of expectation, and measurability of $\Pi_{\ell+1}\,\chi_\ell$ with respect to $\mathcal{F}_{i+1}^t$ for moving the term $\Pi_{\ell+1}\,\chi_\ell$ out of the conditional expectation. Finally, since Π_{i+1} is measurable with respect to $\mathcal{F}_{i+1}^t$, we can pull it out of the conditional expectation by stability

$$\mathbb{E}\big[\Pi_{i+1}\,\chi_i\big|\ \mathcal{F}_{i+1}^t\big] \ = \ \Pi_{i+1}\,\mathbb{E}\big[\chi_i\big|\ \mathcal{F}_{i+1}^t\big].$$

Now we use the fact that $z_1, \dots, z_t$ are *independent*, which implies that χ_i is independent of $\mathcal{F}_{i+1}^t$, hence

$$\Pi_{i+1}\,\mathbb{E}\big[\chi_i\big|\ \mathcal{F}_{i+1}^t\big] \ = \ \Pi_{i+1}\,\mathbb{E}\big[\chi_i\big] \ = \ 0,$$

based on (4.14).

Therefore we have

$$\mathbb{E}\Big[\langle \Pi_{i+1}\,\chi_i,\ \Pi_{\ell+1}\,\chi_\ell\rangle_K\Big] \ = \ 0 \quad \text{for } \ell > i.$$

Thus, only the terms with $i = \ell$ remain in the expectation of (4.21), yielding

$$\mathbb{E}\left[\Big\|\sum_{i=0}^{t}\Pi_{i+1}\,\chi_i\Big\|_K^2\right] \ = \ \sum_{i=0}^{t}\mathbb{E}\big[\|\Pi_{i+1}\,\chi_i\|_K^2\big].$$

It follows from (4.16) that:

$$\mathbb{E}\left[\Big\|\sum_{i=0}^{t}\Pi_{i+1}\chi_i\Big\|_K^2\right] \le \sum_{i=0}^{t} D_0^2 \exp\Big(-2p_0\lambda_0\sum_{j=i+1}^{t} j^{-\beta-\theta}\Big)\mathbb{E}\left(\|\chi_i\|_K^2\right).$$

Since

$$\chi_i = p_i\left\{\big(f_{\lambda_i,\mu^{[i]}}(x_i) - x_{i+1}\big)K_{x_i} + T_{K,\mu^{[i]}}(f_\rho - f_{\lambda_i,\mu^{[i]}})\right\},$$

we see that

$$\|\chi_i\|_K^2 \le 2p_i^2k^2\left\{\big(f_{\lambda_i,\mu^{[i]}}(x_i) - x_{i+1}\big)^2 + \|f_\rho - f_{\lambda_i,\mu^{[i]}}\|_{\mu^{[i]}}^2\right\}.$$

Then

$$\mathbb{E}\big[\|\chi_i\|_K^2\big] \le 4p_i^2k^2\left(\|f_\rho - f_{\lambda_i,\mu^{[i]}}\|_{\mu^{[i]}}^2 + M^2\right).$$

To bound the norm, we take 4.7 with $\lambda = \lambda_i$ and $\mu = \mu^{[i]}$, and bound

$$\|f_\rho - f_{\lambda_i,\mu^{[i]}}\|_{\mu^{[i]}}^2 + \lambda_i\|f_{\lambda_i,\mu^{[i]}}\|_K^2 \le \|f_\rho\|_{\mu^{[i]}}^2 \le M^2.$$

Hence

$$\|f_\rho - f_{\lambda_i,\mu^{[i]}}\|_{\mu^{[i]}}^2 \le M^2 \quad \text{and} \quad \mathbb{E}\big[\|\chi_i\|_K^2\big] \le 8p_i^2k^2M^2.$$

Therefore:

$$\mathbb{E}\left[\Big\|\sum_{i=0}^{t}\Pi_{i+1}\chi_i\Big\|_K^2\right] \le 8p_0^2k^2M^2D_0^2\sum_{i=0}^{t}(i+1)^{-2\theta}\exp\Big(-2p_0\lambda_0\sum_{j=i+2}^{t+1} j^{-\beta-\theta}\Big).$$

Applying Lemma 4.12 with $c = 2p_0\lambda_0$, $q_2 = 2\theta$, and $q_1 = \beta + \theta$, and the Schwarz inequality, we know that

$$\mathbb{E}\left[\left\|\sum_{i=0}^{t}\Pi_{i+1}\chi_i\right\|_K^2\right] \leq \begin{cases} 3p_0kMD_0D_1(t+1)^{\frac{\beta-\theta}{2}} & \text{if } \beta+\theta<1, \\ 3p_0kMD_0D_1(t+1)^{-\min\{\theta-\frac{1}{2},p_0\lambda_0\}} & \text{if } \beta+\theta=1, p_0\lambda_0\neq\theta-\frac{1}{2}, \\ 3p_0kMD_0D_1(t+1)^{\frac{1}{2}-\theta}\sqrt{\log(t+2)} & \text{if } \beta+\theta=1, p_0\lambda_0=\theta-\frac{1}{2}. \end{cases}$$

where D_1 is the constant given by

$$D_1 = \begin{cases} \frac{2}{\sqrt{p_0\lambda_0}} + 1 + \left(\frac{2}{ep_0\lambda_0(1-2^{\beta+\theta-1})}\right)^{\frac{2}{1-\beta-\theta}}, & \text{if } \beta+\theta<1, \\ \frac{2}{\sqrt{|2p_0\lambda_0-2\theta+1|}}+1, & \text{if } \beta+\theta=1, p_0\lambda_0\neq\theta-\frac{1}{2}, \\ 3, & \text{if } \beta+\theta=1, p_0\lambda_0=\theta-\frac{1}{2}. \end{cases}$$

This, in conjunction with (4.20), provides a bound for the error decomposition's *sample error* term.

The *last step of the proof* is to estimate the total error $\|f_{t+1} - f_\rho\|_K$ by applying the triangle inequality to the error decomposition. The approximation error is estimated in Proposition 4.3.2 as

$$\|f_{\lambda_t,\mu} - f_\rho\|_K \leq \|g\|_\pi \lambda_0^{r-\frac{1}{2}}(t+1)^{-\beta(r-\frac{1}{2})},$$

while the drift error is bounded in Corollary 4.3.4 as

$$\|f_{\lambda_t,\pi} - f_{\lambda_t,\mu^t}\|_K \leq \tilde{C}_K C\|g\|_\pi \alpha^t (t+1)^{-\beta(r-\frac{3}{2})}.$$

Note that when $\alpha < 1$, we have $\alpha^t = \exp\{-\log\frac{1}{\alpha}t\} \leq \frac{1}{e\log\frac{1}{\alpha}}t^{-1}$ by Lemma 4.11 with $a = 1$ and $c = \log\frac{1}{\alpha}$. Adding bounds for the three terms verifies the error estimate in Theorem 4.2.4 with the constants C_1^*, C_2^* given explicitly by

$$C_1^* = 3kMD_0D_1,$$

$$C_2^* = \|g\|_\pi \begin{cases} \frac{C_K(k+\kappa_{2s})}{e\log\frac{1}{\alpha}} + 4D_3D_0\left(C_K(k+\kappa_{2s})\left(\frac{4}{e\log\frac{1}{\alpha}}\right)^2+1\right), & \text{if } \beta>0, \alpha<1, \\ \frac{C_K(k+\kappa_{2s})}{e\log\frac{1}{\alpha}} + 2D_3D_0C_K(k+\kappa_{2s}), & \text{if } \beta=0, \alpha<1, \\ C_K(k+\kappa_{2s})\Big(1+2D_3D_0\Big), & \text{if } \beta=0, \alpha=1. \end{cases}$$

Finally, by changing $t+1 \to t$ and re-indexing f_{t+1} as f_t, we obtain the stated bound in terms of t. This completes the proof of Theorem 4.2.4. □

4.4 Discussion

4.4.1 Interpretation of the Main Results

In this chapter, we extended the framework of learning a regression function in an RKHS to the setting of Markovian data, where a stochastic dynamical system (SDS) or Markov chain (MC) generates sequential samples. Our core problem was to learn the map

$$f_\rho(x) \;=\; \mathbb{E}\big[X_{t+1} \,\big|\, X_t = x\big],$$

which plays the role of the next-step predictor. As we mentioned before, the challenge when dealing with dynamical systems is that data streams are not always stationary nor independent. Hence, moving beyond i.i.d. samples requires a focus on these two issues.

Below, we discuss how we tackled these challenges and interpret the main technical results by making a more intuitive sense of the online learning algorithm and its convergence properties. Finally we briefly contextualize this work in the literature.

Data Collection and Online Learning

- *Independence:* By collecting "diagonal" slices from multiple trajectories, we formed a sequence of state transitions $z_t = (x_t, x_{t+1})$ that are *independent*, despite coming from dynamical trajectories, yet each z_t follows a distinct distribution. This allows us to merge Markov chain evolution with an online (i.e., incremental) update scheme, while still retaining the key benefit of sample independence.

- *Drifting distributions:* To build intuition on how we handle non-identical distributions, recall the error decomposition

$$\underbrace{\big(f_{t+1} - f_{\lambda_t,\mu^{[t]}}\big)}_{\textit{sample error}} + \underbrace{\big(f_{\lambda_t,\mu^{[t]}} - f_{\lambda_t,\pi}\big)}_{\textit{drift error}} + \underbrace{\big(f_{\lambda_t,\pi} - f_\rho\big)}_{\textit{approximation error}},$$

  and the update rule in (4.1):

$$f_{t+1} \;=\; f_t \;-\; p_t\Big[\big(f_t(x_t) - x_{t+1}\big)\,K_{x_t} \;+\; \lambda_t\, f_t\Big].$$

  The difference between the i.i.d. case and our case shows in the drift error, which is controlled by Corollary 4.3.4 through the geometric rate $C\alpha^t$, and in the iterations of the sample error (4.10). When analyzing the latter, we encounter the transition from $f_{\lambda_{t-1},\mu^{[t-1]}} \to f_{\lambda_t,\mu^{[t]}}$, which involves both the change in the distributions and in the regularization parameter, and decompose it as

$$f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_t,\mu^{[t]}} = \big(f_{\lambda_{t-1},\mu^{[t-1]}} - f_{\lambda_{t-1},\pi} + f_{\lambda_t,\pi} - f_{\lambda_t,\mu^{[t]}}\big) + \big(f_{\lambda_{t-1},\pi} - f_{\lambda_t,\pi}\big).$$

  The terms stemming from the decaying regularization $\lambda_t \to 0$ associated with a fixed measure are reminiscent of classic regularized SGD algorithms [36]. In the limit, these terms are controlled as discussed in Chapter 2 (see 2.3.2).

  In our case we have an additional complexity given by the changing distributions, hence we don't have a clear expected risk to minimize, nor a fixed gradient to approximate. The decay of the regularization happens simultaneously with the change of distributions, thus we accumulate an iterative error.

Learning Bounds

Theorem 4.2.4. Our main result shows that, under the assumptions above, the function produced by the online algorithm (4.1) converges to f_ρ at different rates depending on the choice of parameters. In particular, for a decaying regularization $\lambda_t \to 0$, i.e. $\beta > 0$, we distinguish two cases:

Case $\beta < 1 - \theta$

$$\mathbb{E}\big[\|f_t - f_\rho\|_K\big] = O\big(\, t^{-\min\{\beta(r-\frac{1}{2}),\frac{\theta-\beta}{2}\}}\,\big),$$

Case $\beta = 1 - \theta$

$$\mathbb{E}\big[\|f_t - f_\rho\|_K\big] = O\big(\, t^{-\min\{\beta(r-\frac{1}{2}),\theta-\frac{1}{2},p_0\lambda_0\}} \log(t)\,\big).$$

Note that the term $t^{-\beta(r-\frac{1}{2})}$ is reminiscent of classic SGD in the i.i.d. setting (see 2.4.4), it decays faster with bigger r, meaning a stronger source condition, or if we pick β suitably. Here, we should note that we cannot choose $r > \frac{3}{2}$, as several bounds in the error analysis depend on $r \leq \frac{3}{2}$. Moreover, finding an optimal choice of parameters is not trivial, since the asymptotic behavior depends both on the initialization and decay of gain and learning rate, and on the regularity of f_ρ; it should thus be studied case by case depending on the known conditions.

It is worth mentioning that this abrupt change in the bound for $\beta = 1 - \theta$ stems directly from Lemma 4.12, which is a technical bound relying on the anti-derivative of a term of the form $t^{-\beta-\theta}$. Hence for this specific choice of β we get two qualitatively different guarantees on the algorithm's convergence. Future work might include improvements for this particular case.

Overall, the interpretation of these results is that as soon as the underlying Markov chain's distribution stabilizes (i.e. $\mu^{[t]} \to \pi$ are close enough), one can treat incremental regression in an RKHS with only a mild penalty in convergence speeds compared to the standard i.i.d. setting. This confirms the robustness of kernel-based online algorithms in scenarios where data are generated by a stable stochastic dynamical system.

4.4.2 The context in the broader literature

The analysis of online learning in the context of non-stationary and dependent data has long been a challenge in statistical learning. Our results contribute to the ongoing effort to extend classical (i.i.d.) convergence guarantees to broader scenarios, particularly in forecasting for Markov chains and stochastic dynamical systems. Here, we highlight the areas where our framework integrates with existing literature.

Moving beyond i.i.d.

We can identify two lines of research on this topic. From a statistical and mathematical perspective, initial analysis of RKHS-based (particularly ridge) regression focused on *i.i.d.* data and established consistency along with optimal rates [6]. These assumptions were systematically expanded: allowing non-identical yet independent samples [30, 16], incorporating mixing conditions for dependent but stationary processes [41, 47, 22, 42, 25, 45], and covering more general dynamical systems. In contrast, applied domains such as system identification, optimal control, or time series forecasting typically adopt more general dynamical models (where dependence and non-stationarity are inherent) and add simplifying assumptions to approximate classical learning frameworks, since strictly *i.i.d.* conditions are often impractical in real-world scenarios.

A similar distinction arises between *offline* and *online* learning paradigms. In statistical machine learning, iterative techniques are driven mainly by memory constraints and the growth of datasets. However, for real-time data streams in practical settings, the motivation for incremental (online) methods becomes natural. Within the *i.i.d.* online learning literature in RKHS, Smale and Yao [29], Ying and Pontil [40], Tarres and Yao [36], and others [39, 13, 19, 11, 12, 7] have shown that suitably tuned step sizes and regularization can yield consistency rates matching those of batch algorithms (see Corollary 2.4.3).

Moving beyond *identical* distributions, Smale and Zhou [30] and Hu [16] examined independent but non-identical data (e.g., drifting marginals) under exponential or polynomial

conditions, confirming convergence of regularized solutions. Our approach in Chapter 4 is closely related to [30] but derives $\mu^{[t]}$ from a stable Markov chain whose marginal converges geometrically (uniformly) to π. Theorem 4.2.4 shows that this additional distributional complexity is contained, leading to non-asymptotic error bounds. Several works [33, 38] treat Markovian (or time-correlated) data in offline or semi-offline frameworks, but explicit *online* convergence analyses in an RKHS remain relatively scarce.

In this work, we leverage ergodicity to address Markovian drift and use an online algorithm suitable for real-world streaming data. This strategy bridges a gap between the statistical perspective of relaxing *i.i.d.* assumptions with the applied perspective of examining systems that exhibit an inherent more complex dynamics.

Outlook

Overall, Chapter 4 reinforces the viewpoint, initially proposed by Smale and Zhou [30], that one can achieve nearly i.i.d. rates for online regression in an RKHS, provided the time-varying measures converge at a sufficient pace and the target function f_ρ is regular enough. We connect this distributional convergence directly to the ergodicity of a Markov chain, providing a clear method to implement kernel-based online learning for stochastic dynamical systems. This bridges a gap between purely i.i.d. or mixing-based offline analysis in kernel methods and Markov chain/SDS stability analysis, enabling new applications where data come from stable but nonstationary processes. Improvements might include relaxing geometric ergodicity to sub-exponential (eg. polynomial), even non-uniform rates, or replacing the ergodicity assumption by other conditions (eg. persistence of excitation). Future directions could extend this setting to forecasting in higher-dimensional dynamical systems, or including mixing conditions for weakening independence assumptions.

Chapter 5

Perspectives and Improvements

In this final chapter, we briefly discuss several directions for extending and enhancing our framework. Now that we have established statistical learning bounds for one-dimensional state spaces $S \subseteq \mathbb{R}$, we propose improvements aimed at broadening both the theoretical and practical applicability of our approach. In particular, we outline potential extensions to multi-dimensional state spaces, the estimation of higher-order moments, and strategies to weaken the independence assumption via mixing techniques.

5.1 Multidimensional state space

For a more general state space $S \subset \mathbb{R}^n$, the stochastic process $X = \{X_i\}_i$ becomes vector valued and so the problem of finding the regression function $f_\rho(x) = \mathbb{E}[x_{\text{next}} \mid x] \in S$. To extend our learning algorithm we first need to adapt the RKHS and its elements to accomodate multidimensional outputs.

To handle multidimensional outputs we can consider a *vector-valued* RKHS [21, 5]. In this setting, a reproducing kernel is a symmetric function

$$K : S \times S \;\to\; \mathbb{R}^{d\times d},$$

such that for any $x, x' \in S$, the matrix $x'K(x, x')$ is positive semidefinite. A *vector-valued* RKHS $\mathcal{H}_K$ is the Hilbert space of functions $f : S \to \mathbb{R}^d$ with inner product $\langle \cdot, \cdot \rangle_K$ satisfying the reproducing property:

$$\langle f,\; K(\cdot, x)\, c\rangle_K \;=\; f(x)^\top c \quad \text{for all} \quad c \in \mathbb{R}^d,\, x \in S.$$

The choice of K corresponds to how one parameterizes the function of interest. In fact, any function in $\mathcal{H}_K$ lies in the closure of finite linear combinations of the form

$$f(x) \;=\; \sum_{i=1}^{p} K(x_i, x)\, c_i, \quad c_i \in \mathbb{R}^d,$$

where each $K(x_i, x)$ is a $d \times d$ matrix acting on the vector c_i. The norm $\|f\|_K$ typically measures the complexity of f, reflecting the scalar-valued scenario.

Once the vector-valued kernel framework is in place, we can formulate an online update rule analogous to the one-dimensional case, replacing scalar operations with matrix-vector operations. For instance, a gradient-based approach might yield an update of the form:

$$f_{t+1} \;=\; f_t \;-\; p_t \Big[\big(f_t(x_t) - x_{t+1}\big)\, K(x_t, \cdot) \;+\; \lambda_t\, f_t\Big],$$

where $f_t : S \to \mathbb{R}^d$, p_t is the gain, λ_t is the regularization vector, and $K(x_t, \cdot)$ is the operator applied to each $x \in S$. The same high-level analysis on approximation, sample, and drift errors, carries over, although technically more involved due to the matrix-valued nature of K. Notably, the increased computational overhead due to matrix operations can significantly impact the algorithm's practical implementation, highlighting an important aspect of cost analysis for future research.

5.2 Learning More Than One Moment

In this work, our primary focus has been on learning the conditional expectation

$$f_\rho(x) = \mathbb{E}[X_{t+1} \mid X_t = x],$$

which represents the first moment of the transition distribution.

Remark 5.2.1 In our setting, the Markov chain $X = \{X_0, X_1, \dots\}$ is time-homogeneous, meaning that the transition probabilities, and thus the conditional distribution, are invariant with respect to time. This invariance ensures that the regression function f_ρ is consistent across all time steps.

However, a more complete statistical description of the underlying dynamical system can be obtained by estimating additional moments. In particular:

- **Variance Estimation (Second Moment):** Beyond the mean, learning the conditional variance
$$\sigma^2(x) = \mathbb{E}\Big[\Big(X_{t+1} - f_\rho(x)\Big)^2 \mid X_t = x\Big]$$
would provide valuable information about the uncertainty and variability in the system's evolution. One approach is to extend the current online learning framework to simultaneously estimate both the first and second moments by considering a vector-valued RKHS for multi-output regression with the approach mentioned in the section above. For instance, we may consider estimating the second moment function
$$f_\rho^{(2)}(x) = \mathbb{E}\Big[X_{t+1}^2 \mid X_t = x\Big],$$
and then recover the variance via $\sigma^2(x) = f_\rho^{(2)}(x) - \big(f_\rho(x)\big)^2$. For further details on estimation of conditional variance functions, see [9] and [8].

- **Kalman Filtering:** In settings where the dynamics are linear or can be locally approximated as linear, Kalman filtering offers estimates of both the state and its uncertainty. Consider a linear state-space model
$$x_{t+1} = Ax_t + w_t, \quad y_{t+1} = Cx_{t+1} + v_{t+1},$$
where x_t is the true state fo the system at time t and y_t is its measurement. The variables w_t and v_t represent process and measurement noise, respectively. The Kalman filter recursively updates the state estimate via
$$\hat{x}_{t+1} = A\hat{x}_t + K_t\big(y_t - C\hat{x}_t\big),$$
where the matrix K_t is known as the Kalman gain.

  Using kernel-based online learning with Kalman filtering techniques may lead to hybrid algorithms that capture both the mean and covariance structures of the transition distribution. A kernel-based extension of the Kalman filter is discussed in [35] and related works.

- **Moment Generating Functions:** The moment generating function (MGF) of a random variable X is defined as

$$M_X(t) = \mathbb{E}\left[e^{tX}\right],$$

for t in an open interval around zero where the expectation exists. Since the MGF uniquely characterizes the distribution of X (when it is finite in a neighborhood of zero), its Taylor expansion yields

$$M_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}[X^k].$$

The cumulant generating function (CGF) is then given by

$$K_X(t) = \log M_X(t),$$

and, when differentiable, the kth cumulant κ_k is recovered via

$$\kappa_k = \left.\frac{d^k}{dt^k} K_X(t)\right|_{t=0}.$$

In particular, $\kappa_1 = \mathbb{E}[X]$ and $\kappa_2 = \mathrm{Var}(X)$.

Recent work [2] has proposed a kernel-based approach for learning the cumulants by embedding the space of such functions into a reproducing kernel Hilbert space (RKHS). In this framework, one constructs an estimator $\hat{K}_X(t)$ in an RKHS $\mathcal{H}_K$ induced by a kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.

A possible direction is to extend this framework to ergodic systems by leveraging the techniques developed in this work.

5.3 Weakening independence: Mixing

In our data collection we assumed that the trajectories $\{X(\omega_i)\}$ are independent. However, in many practical applications obtaining a large number of independent trajectories can be challenging. A possible direction is to relax the independence assumption by assuming mixing properties on the dynamical system.

Definition 5.3.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space, and consider sub σ-algebras $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$. Define

$$\alpha(\mathcal{A}, \mathcal{B}) := \sup_{A \in \mathcal{A}, B \in \mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|,$$

$$\beta(\mathcal{A}, \mathcal{B}) := \sup_{A \in \mathcal{A}, B \in \mathcal{B}} \tfrac{1}{2} \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \cap B_j) - P(A_i)P(B_j)|,$$

where the supremum is over all (finite) partitions $\{A_1, \ldots, A_I\}$ and $\{B_1, \ldots, B_J\}$ of Ω, with $A_i \in \mathcal{A}$, $B_j \in \mathcal{B}$ for all i, j respectively.

Definition 5.3.2. Let $X = \{X_t\}_{t=0}^{\infty}$ be a stochastic process. Define the σ-algebra $\mathcal{F}_i^j = \sigma(X_i, \ldots, X_j)$, and for each $n \in \mathbb{N}$, let

$$\alpha(n) = \sup_j \alpha(\mathcal{F}_0^j, \mathcal{F}_{j+n}^{\infty}),$$

$$\beta(n) = \sup_j \beta(\mathcal{F}_0^j, \mathcal{F}_{j+n}^{\infty}).$$

These are known as the **α-mixing** and **β-mixing** coefficients, respectively.

The process X is called **α-mixing** if $\alpha(n) \to 0$ as $n \to \infty$. Similarly, X is **β-mixing** if $\beta(n) \to 0$ as $n \to \infty$.

Remark 5.3.3 The α-mixing and β-mixing conditions quantify how the dependence between past and future observations in a stochastic process diminishes as the time gap increases. These conditions ensure that events occurring before time t and after time $t+n$ become asymptotically independent as $n \to \infty$, uniformly over t. Although consecutive observations may exhibit strong dependence, observations that are sufficiently far apart can be treated as approximately independent.

A key difference is that α-mixing measures the maximal deviation of joint probabilities from the product of marginals, whereas β-mixing quantifies their total variation distance. In fact, one can show that

$$\alpha(\mathcal{A}, \mathcal{B}) \leq \beta(\mathcal{A}, \mathcal{B})$$

for any pair of σ-algebras, meaning that β-mixing is a stronger condition than α-mixing.

Remark 5.3.4 These definitions represent only a few of the many ways to quantify dependence in stochastic processes. In addition to the α- and β-mixing coefficients defined above, other strong mixing conditions (such as ϕ-mixing and ψ-mixing) offer alternative formulations. In our framework, β-mixing is particularly well-suited, as stationary, aperiodic, Harris recurrent chains are known to be β-mixing (and therefore α-mixing) [4]. Recall that we assumed our chain to be positive Harris recurrent and aperiodic to ensure ergodicity, thus we can think of the chain as asymptotically β-mixing.

Mixing Time

One practical approach is to partition a long trajectory into blocks that are separated by a parameter, which defines how long it takes for the dependency of the chain to be sufficiently small.

Definition 5.3.5 (Mixing Time)**.** Given the definitions above and $\varepsilon > 0$, we define

$$t^{\alpha}_{\text{mix}}(\varepsilon) := \min_n\{\alpha(n) \leq \varepsilon\}, \qquad t^{\beta}_{\text{mix}}(\varepsilon) := \min_n\{\beta(n) \leq \varepsilon\}$$

Over an interval of length t_{mix}, observations from distinct blocks can be regarded as approximately independent. Note that choosing different mixing coefficients leads to different mixing times, in particular $t^{\alpha}_{\text{mix}} \leq t^{\beta}_{\text{mix}}$.

One practical advantage of assuming mixing conditions for our data collection scheme in 4.1.1, is that it allows us to extract multiple effective samples from a single trajectory, thereby reducing the need for many independent trajectories:

$$\begin{array}{lll}
\underbrace{X_0(\omega_0),\, X_1(\omega_0)}_{z_0},\ldots & & \underbrace{X_{t_{\text{mix}}}(\omega_0),\, X_{t_{\text{mix}}+1}(\omega_0)}_{z_{t_{\text{mix}}}}, \\
X_0(\omega_1),\, \underbrace{X_1(\omega_1),\, X_2(\omega_1)}_{z_1},\ldots & & \underbrace{X_{t_{\text{mix}}+1}(\omega_1),\, X_{t_{\text{mix}}+2}(\omega_1)}_{z_{t_{\text{mix}}+1}}, \\
\vdots & \ddots & \ddots \\
X_0(\omega_{t_{\text{mix}}-1}),\ldots & \underbrace{X_{t_{\text{mix}}-1}(\omega_{t_{\text{mix}}-1}),\, X_{t_{\text{mix}}}(\omega_{t_{\text{mix}}-1})}_{z_{t_{\text{mix}}-1}},\ldots &
\end{array}$$

Remark 5.3.6 For our algorithm (4.1), the number of independent trajectories required to compute the update f_t is $t+1$, which grows without bound as $t \to \infty$. However, assuming mixing, when $t \geq t_{\text{mix}}$ no more than t_{mix} trajectories are required. Thus, by

sampling long trajectories, we can cap the maximum number of independent trajectories necessary for implementing the algorithm. This approach is particularly useful in practical scenarios where initializing new trajectories is costly and obtaining long trajectories is more feasible.

Another advantage of mixing conditions is that by selecting blocks from the chain X that are at least t_{mix} apart, we obtain segments that are approximately independent. This allows us to work directly with these blocks rather than relying on a data collection scheme based on fully independent trajectories. Two common strategies in the literature exist: one is to partition a long trajectory into disjoint blocks separated by intervals of length t_{mix}, ensuring negligible dependence between blocks; the other is to construct a family of skeletons, $X^{t_{\text{mix}}+c}$ for $0 \leq c \leq t_{\text{mix}}$, by sampling the chain at different offsets from the initial state. This approach effectively increases the number of independent samples available from a single long trajectory. Future research could focus on quantifying the trade-offs between block length and estimation accuracy, as well as on developing optimized algorithms that automatically determine the optimal block-sampling strategy based on the observed mixing rate [4, 3, 14].

Bibliography

[1] Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[2] Patric Bonnier, Harald Oberhauser, and Zoltán Szabó. Kernelized cumulants: Beyond kernel mean embeddings. *arXiv preprint arXiv:2301.12466*, 2023.

[3] D. Bosq. Bernstein-type large deviations inequalities for partial sums of strong mixing processes. *Statistics*, 24(1):59–70, 1993.

[4] Richard C. Bradley. Basic properties of strong mixing conditions: A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[5] C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Anal. Appl. (Singap.)*, 4(4):377–408, 2006.

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

[7] A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

[8] Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. 1996.

[9] Jianqing Fan and Qiwei Yao. Efficient estimation of conditional variance functions in nonparametric regression. *Journal of the American Statistical Association*, 93(442):1152–1160, 1998.

[10] William Feller et al. An introduction to probability theory and its applications. 1971.

[11] X. Guo, Z. C. Guo, and L. Shi. Capacity dependent analysis for functional online learning algorithms. *Applied and Computational Harmonic Analysis*, 67:101567, 2023.

[12] Z. C. Guo, A. Christmann, and L. Shi. Optimality of robust online learning. *arXiv preprint*, 2304.10060, 2023.

[13] Z. C. Guo and L. Shi. Fast and strong convergence of online learning algorithms. *Advances in Computational Mathematics*, 45:2745–2770, 2019.

[14] H. Hang and I. Steinwart. A bernstein-type inequality for some mixing processes and dynamical systems with an application to learning, 2015.

[15] M.W. Hirsch, S. Smale, and R.L. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Pure and Applied Mathematics - Academic Press. Elsevier Science, 2004.

[16] T. Hu and D.-X. Zhou. Online learning with samples drawn from non-identical distributions. *Journal of Machine Learning Research*, 10(12):2873–2898, 2009.

[17] Olav Kallenberg and Olav Kallenberg. *Foundations of modern probability*, volume 2. Springer, 1997.

[18] Achim Klenke. *Probability Theory: A Comprehensive Course*. Universitext. Springer, Cham, 3rd edition, 2020.

[19] J. Lin and V. Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21(147):1–63, 2020.

[20] Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[21] C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17:177–204, 2005.

[22] M. Mohri and A. Rostamizadeh. Stability bounds for stationary phi-mixing and beta-mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.

[23] David Pollard. *A user's guide to measure theoretic probability*. Number 8. Cambridge University Press, 2002.

[24] Ashish Sabharwal and Bart Selman. S. russell, p. norvig, artificial intelligence: A modern approach, 2011.

[25] A. Sancetta. Estimation in reproducing kernel hilbert spaces with dependent data. *IEEE Transactions on Information Theory*, 67(3):1782–1795, 2020.

[26] Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.

[27] S. Smale and F. Cucker. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2001.

[28] S. Smale and F. Cucker. Best choice for regularization parameters in learning theory: on the bias-variance problem. *Foundations of Computational Mathematics*, 2:413–418, 2002.

[29] Steve Smale and Yuan Yao. Online learning algorithms. *Foundations of Computational Mathematics*, 6:145–170, 06 2006.

[30] Steve Smale and Ding-Xuan Zhou. Online learning with markov sampling. *Analysis and Applications*, 7(01):87–113, 2009.

[31] John Stachurski. A hilbert space central limit theorem for geometrically ergodic markov chains. Technical report, mimeo, Kyoto University, 2009.

[32] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008.

[33] I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

[34] R. Steyer and W. Nagel. *Probability and Conditional Expectation: Fundamentals for the Empirical Sciences*. John Wiley & Sons, Incorporated, 2017.

[35] Mengwei Sun, Mike E. Davies, Ian K. Proudler, and James R. Hopgood. Adaptive kernel kalman filter. *IEEE Transactions on Signal Processing*, 2022.

[36] Pierre Tarrès and Yuan Yao. Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9):5716–5735, September 2014.

[37] Ernesto De Vito and Andrea Caponnetto. Risk bounds for regularized least-squares algorithm with operator-valued kernels. Technical Report MIT-CSAIL-TR-2005-031, AIM-2005-015, CBCL-249, MIT Computer Science and Artificial Intelligence Laboratory, May 2005.

[38] J. Xu, Y. Tang, B. Zou, Z. Xu, L. Li, and Y. Lu. The generalization ability of online svm classification based on markov sampling. *IEEE Transactions on Neural Networks and Learning Systems*, 26(3):628–639, 2014.

[39] Y. Ying and D.-X. Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224–244, 2017.

[40] Yiming Ying and Massimiliano Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

[41] B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.

[42] M. J. Zhang and H. W. Sun. Regression learning with non-identically and non-independently sampling. *International Journal of Wavelets, Multiresolution and Information Processing*, 15(1), 2017.

[43] Xiwei Zhang and Tao Li. Convergence conditions of online regularized statistical learning in reproducing kernel hilbert space with non-stationary data. *arXiv preprint arXiv:2404.03211*, 2024.

[44] Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.

[45] T. Ziemann and S. Tu. Learning with little mixing. *Advances in Neural Information Processing Systems*, 35:4626–4637, 2022.

[46] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, Washington DC, USA, 2003.

[47] B. Zou, L. Q. Li, and Z. B. Xu. The generalization performance of erm algorithm with strongly mixing observations. *Journal of Machine Learning Research*, 75(3):275–295, 2009.