DISS. ETH NO. 26163

IRREGULAR SAMPLING OF BANDLIMITED FUNCTIONS: RECONSTRUCTION AND FILTERING

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

ADRIAN EMMANUEL GEORG HUBER
M.Sc. RWTH Aachen University

born on 06 March 1990
citizen of Germany

accepted on the recommendation of

Prof. Dr. S.-C. Liu, examiner
Prof. Dr. H.-A. Loeliger, co-examiner

2019

Meinen Eltern und meiner Frau gewidmet

ABSTRACT

Sampling in both the signal processing as well as statistical sense is studied in this thesis. We first start by investigating the mathematical properties of a common event-based sampling method, send-on-delta sampling, in detail. The analysis focuses on the extent to which bandlimited functions can be either reconstructed or approximated from the samples that are obtained if the function is sampled by the send-on-delta scheme. It turns out that for square-integrable, continuous and bandlimited functions the main deficit of send-on-delta sampling is that it generates only finitely many sampling points. Standard stability results therefore preclude an exact reconstruction of the underlying function from the acquired samples. We next investigate whether under additional assumptions on the sampled functions besides bandlimitedness, stable reconstruction is feasible from finitely many sampling points only. We find that this is indeed the case, as every square-integrable and bandlimited function is compressible in the prolate spheroidal wave function basis. We use this basis, together with results from infinite-dimensional compressed sensing, to show that stable recovery is possible from finitely many sampling points that are uniformly distributed (in the probabilistic sense) in a time interval in which most of the function energy is concentrated. We then show how continuous-time filtering can be carried out on a sequence of samples that are obtained by sampling a bandlimited function irregularly (the particular sampling mechanism is not important for these findings). Discrete-time filtering therefore does not need to be carried out on samples obtained from a regular sampling pattern. Finally, we investigate sampling in the statistical sense. Using ideas from information theory, specifically universal coding, we show explicitly by example how sampling priors on parametric families influence the amount of information transmitted from the parametric family to the data obtained. We show that the choice of sampling prior has implications for machine learning in that the solution obtained by a machine learning algorithm depends on the sampling prior.

ZUSAMMENFASSUNG

Im Rahmen dieser Arbeit wird Abtastung im Sinne der Signalverarbeitung und Statistik untersucht. Zuerst analysieren wir detailliert die mathematischen Eigenschaften des ereignisbasierten Abtastverfahrens Send-on-Delta Abtastung. Schwerpunkt der Analyse ist dabei die Frage, ob bandbegrenzte Funktionen aus den mittels Send-on-Delta Abtastung gewonnenen Abtastwerten entweder rekonstruiert oder aber approximiert werden können. Für quadratintegrable, stetige und bandbegrenzte Funktionen liegt das Hauptproblem der Send-on-Delta Abtastung darin, dass nur endlich viele Abtastwerte generiert werden. Aus der Literatur bekannte Stabilitätsresultate schliessen daher eine exakte Rekonstruktion der abgetasteten Funktion aus ebendiesen Abtastwerten aus. Als nächstes analysieren wir daher, inwiefern durch Hinzunahme weiterer Annahmen (d.h. zusätzlich zur Annahme der Bandbegrenztheit) eine stabile Rekonstruktion der abgetasteten Funktion aus endlich vielen Abtastwerten möglich ist. Dass dies in der Tat der Fall ist, zeigen wir mit Hilfe der Prolate-Spheroidal-Wave-Funktionen Basis, in der jede quadratintegrable und bandbegrenzte Funktion kompressibel ist. Mit Hilfe dieser Basis und Compressed Sensing in unendlichdimensionalen Räumen zeigen wir dann, dass eine stabile Rekonstruktion aus endlich vielen in der Zeit gleichverteilten Abtastwerten möglich ist. Die Abtastzeitpunkte befinden sich dabei in dem Intervall, in dem der Grossteil der Funktionsenergie konzentriert ist. Wir zeigen dann, wie zeitkontinuierliche Filterung auf Funktionswerten ausgeführt werden kann, die einer irregulären Abtastsequenz entsprechen, wobei der Abtastmechanismus für diese Untersuchung belanglos ist. Zeitdiskrete Filterung bedarf daher nicht einer regulären Abtastfolge. Schliesslich untersuchen wir die Abtastung im statistischen Sinne. Unter Zuhilfenahme von Theorien aus der Informationstheorie, vor allem aus der Universal Coding Theorie, zeigen wir den Einfluss des Sampling Priors, der auf einer parametrischen Familie definiert ist, auf den Informationsfluss von der parametrischen Familie zu der Datenstichprobe auf. Wir zeigen dann, dass die Wahl des Sampling Priors Auswirkungen auf das Maschinelle Lernen aufweist, da die vom Algorithmus gefundenen Lösungen vom Sampling Prior abhängig sind.

ACKNOWLEDGEMENTS

This work would not have been possible without the support of many people. First I would like to thank my supervisor Shih-Chii Liu for creating a great environment for free and independent research. To provide a PhD student with so much freedom in terms of choosing research projects cannot be taken for granted. I would also like to thank the members of my institute with whom I have spent many interesting and joyful discussions. Particular mention has to go to my two colleagues and friends Gregor Schuhknecht and Stefan Braun, as well as to Jithendar Anumula, with whom I had many inspiring discussions on science.

My time in England set me on the path of science and engineering. My teachers there kindled within me a love for abstract reasoning and thought. For this profound influence on my life I am very grateful. Without their input, my life would have taken a very different turn.

This work would not have been possible without my family. I want to thank my sister who always supported me in my private life. My wife has been a constant support and help during the years of my PhD studies. Without her providing a solid emotional fundament during the last years of my studies, it is doubtful whether I would have finished them. Thank you for everything you have given me. Finally I would like to thank my parents who have always encouraged me to follow on an academic path, and without their support – both financially and spiritually – I would not be writing these words. I am deeply grateful for the many things they have given me.

CONTENTS

1 INTRODUCTION

The word *sampling* as used in the mathematical sciences possesses different meanings. One interpretation of the word, derived from statistics, is based on the process of obtaining realizations of a random variable/vector and using these realizations to infer something about the underlying distribution(s). The other meaning, traditionally used in the field of signal processing, is based on the process of pointwise evaluation of a function such that the collected information is sufficient for reconstruction/filtering. The latter sampling process is mostly understood to be *deterministic*, i.e. the pointwise samples are taken at a predetermined set of places (in time or space), tailored only to the space in which the sampled function lives. The deterministic nature of pointwise sampling can be relaxed, however, and replaced by *random* sampling. Random sampling can arise through different mechanisms. One mechanism that leads to random sampling is *event-based* sampling of stochastic processes. In this paradigm, samples are generated whenever some deterministic criterion is fulfilled by the underlying function. Another mechanism is a random sampling pattern that is uncoupled from the sampled function/stochastic process. For randomness to be mathematically tractable, it requires structure. Structure is in particular required if a question such as "Is there enough information recorded in the sampled values for the function to be reconstructible?" is to be answered. Another question that requires assumptions on the structure of the random sampling pattern is whether a sampling pattern will result in aliasing. The more precise the questions asked, the more structure is imposed on the sampling pattern. Many properties such as reconstructability of a sampled function from the sample values alone directly translate into *yes/no* statements on the sufficiency of a sampling pattern structure for the preservation of the desired property. As such, with respect to a specific property that a sampling pattern should preserve, different sampling methods can only be compared in terms of whether or not the desired property is retained.

To what extent is randomness beneficial when compared to deterministic sampling schemes? An answer to such a question depends on the specific operation that should be carried out on the sampled function values. To shed light on the possible benefits of randomness, two possible operations on the sampled values are discussed in Sections 1.1 and 1.2: reconstruction and spectrum estimation.

Sampling in the statistical sense leads to a gradual reduction in uncertainty. A similar statement can only be made partially for signal processing sampling. As stated above, sampling patterns need to be tailored to the function space in which the sampled function resides. Consider the space of square-integrable (in the Lebesgue sense) and bandlimited functions. Given a function from such a space, *finitely* many samples cannot characterize a function uniquely. In fact, a function of arbitrarily smaller bandwidth than the sampled function can perfectly interpolate the finitely many samples, provided the energy of the interpolating function is not constrained. As such, moving from n to $n+1$ samples does not in itself lead to a better identifiability of the original function from its samples in the discussed function space. Additional assumptions need to be placed on the function space, such as sparsity constraints in some frame/basis or stronger concentration in time properties (effectively moving to a different function space, for example to some weighted L^1 space), such that an increase from n to $n+1$ samples leads to an effective gain in information. Alternatively, some additional constraints on the sampled function besides the sampling values themselves can be provided by the sampling mechanism itself. Event-based sampling can in particular provide such implicit constraints. The differential gain in information provided by each new sample therefore strongly depends on the *a priori* assumptions which structure the underlying function space.

Reduction in uncertainty is in some sense equivalent to better predictability. As predictability can be assessed in statistical terms (in case that either the sampled function is stochastic or the sampling pattern is random), it is a property that should be fulfilled by both sampling in the statistical as well as signal processing sense. In time-series analysis, for example, predicting the future from finitely many past observations (samples) is a common task. If the time-series structure is partly unknown, then prediction consists of two parts, that is model identification and forecasting. The operations of reconstruction and filtering from signal processing can be understood as predictions as well, as they consist of an (optimal) inference of the desired function from observations. Both prediction in the statistical and signal processing sense are partly ambiguous, if either the model/function space is too large or too few observations are given. The study of sampling schemes that encode some property of the sampled function optimally can be effectively carried out statistically.

In Sections 1.1 and 1.2, reconstruction and spectrum estimation from both deterministic as well as randomly distributed sampling patterns are discussed.

1.1 RECONSTRUCTABILITY

Under which conditions is it possible to fully reconstruct a function from a set of samples? The number and type of samples required clearly depends on the function space in which the sampled function lives. Polynomials of fixed and finite order, for example, are in general fully specified by a finite number of samples, while piecewise linear functions are determined by specifying the function values at the ends of the linear segments. Both examples allow for the use of irregularly spaced samples in general. In both examples the number of required sampling points is inherently connected to the degrees of freedom that a function from the respective space possesses.

These general ideas apply to function spaces of bandlimited functions as well. Consider the Paley-Wiener space of square-integrable and continuous functions whose Fourier transform is restricted to a compact interval on the real line. in general, absent any further assumptions, functions from the Paley-Wiener space have infinitely many degrees of freedom, and so infinitely many (countable infinity) sampling values will have to be taken for the function to be in principle reconstructible from its samples. Functions living in the Paley-Wiener space are entire functions, i.e. they are holomorphic on the entire complex plane. Since holomorphic functions are analytic, they can be expressed as a power series. A local power series expansion such as a Taylor series therefore fully describes the function everywhere. In the Taylor series case, the samples would correspond to (higher-order) derivatives of the function at the expansion point, yielding a full description of the sampled function. The description of a function from the Paley-Wiener space via a Taylor series at an expansion point is, however, not stable with respect to noise. Small perturbations of the expansion coefficients will not lead to small changes in the reconstructed function in general. As such, limited numerical precision implies that a Taylor series representation of Paley-Wiener functions is not sufficient for practical implementations. Stability with respect to noise necessitates a certain distribution of sampling times along the real line. The set of conditions imposed on sampling patterns such that they are stable is known for Paley-Wiener spaces. Both necessary and sufficient conditions are fully specified. A particularly important part in these conditions is played by the Nyquist-Landau rate which is the lowest average sampling rate below which stable sampling is infeasible [1]. For lowpass functions, the Nyquist-Landau rate is equal to the more well-known Nyquist rate. Two different sampling regimes can now be differentiated. The differentiation is with respect to whether the average sampling rate corresponds exactly to the Nyquist-Landau rate or whether it is strictly larger. In case the average sampling rate is equal to the Nyquist-Landau rate, then reconstruction is possible if the set $\left\{e^{it_n x}\right\}$, with $(t_n)_{n\in\mathbb{Z}}$ the sampling pattern and x the frequency variable, constitutes a Riesz basis for the space of square-integrable functions with the same support as the frequency support of the Paley-Wiener function. Rather complicated conditions on $(t_n)_{n\in\mathbb{Z}}$ that ensure the Riesz basis property have been described [2]. Simplified sufficient conditions are known that guarantee the Riesz basis property as well [3]. These conditions stipulate that the sampling pattern should be a perturbed uniform sampling pattern, with the maximum amount of perturbation of each sample from its nominal uniform position bounded from above by known constants. In case the average sampling rate is larger than the Nyquist-Landau rate, then the set $\left\{e^{it_n x}\right\}$ should form a frame for the space of square-integrable functions with the same support as the frequency support of the Paley-Wiener function. Necessary as well as sufficient conditions for this frame property have been fully described [4]. Once the frame property is ensured, then powerful iterative reconstruction algorithm can be leveraged [5].

1.2 SPECTRUM ESTIMATION AND ALIASING

A common problem in applications is the estimation of the power spectrum (or the power spectral density, in case it exists) from the sampled values of a weakly-stationary and mean-square continuous stochastic process. If uniform samples are taken, it is well known that a bandlimited function has to be sampled at the Nyquist rate or higher for the estimation of the power spectrum to be possible [6]. Sufficieny of sampling at the Nyquist rate is not given for bandlimited functions in general, but depends in addition on other properties of the function, such as in which function space it is contained. To be precise, if the power spectral density belongs to $L^p\left(-\pi w, \pi w\right)$, $p > 1$ and $w > 0$, then Nyquist rate uniform sampling is sufficient, otherwise oversampling is required. If non-bandlimited functions or functions with unknown bandlimit are sampled at a fixed uniform rate, then a consistent estimation of the power spectrum of the sampled stochastic process cannot be carried out from the samples alone. The reason for this behavior is the presence of aliasing, that is the overlap of shifted copies of the power spectral density of the sampled stochastic process. For uniform sampling, absense of aliasing is therefore equivalent to the fact that the power spectrum can be estimated from the samples alone.

Random sampling has therefore been investigated as a possible mechanism to circumvent this problem which is inherently associated with uniform sampling. Given the crucial observation that a consistent estimation of the power spectrum is possible in the absence of aliasing for uniform sampling, conditions were sought that guarantee this alias-free property for larger classes of spectra under random sampling, in particular for non-bandlimited spectra. A first definition specifying when a sampling pattern is alias-free with respect to a set of spectra reads as follows:

> The sampling sequence $\{t_n\}$ is alias free relative to $\mathcal{S}$ (a family of spectra) if no two random processes with different spectra belonging to $\mathcal{S}$ yield the same correlation sequence $\{r(n)\}$. [7]

The correlation sequence $\{r(n)\}$ is specified by $r(n) = \mathbb{E}[x(t_{m+n})\, x^*(t_m)]$, where x is the stochastic process, $(\cdot)^*$ denotes complex conjugation and the expectation $\mathbb{E}[\cdot]$ is taken over the process x and the sampling pattern $\{t_n\}$. One of the first random sampling schemes investigated was additive random sampling which can be described by the following sampling mechanism: $t_n = t_{n-1} + \gamma_n$ with $\gamma_n \sim p(\tau)$ and $p(\tau)$ a probability density function in $L^2(0, \infty)$. Additive random sampling is alias-free with respect to any power spectral density contained in $L^1 \cap L^2$ if the characteristic function $\phi(\omega) = \int_0^\infty p(\tau)\, e^{i\omega\tau} d\tau$ takes no value more than once on the real axis [8]. This condition is sufficient, but not necessary. It holds in particular also for non-bandlimited spectra. A specific instance of additive random sampling which is alias-free (in the sense of the above definition) is Poisson sampling, for which $p(\tau)$ is taken as $p(\tau) = \beta e^{-\beta\tau}$ for $\tau \geq 0$ and $\beta > 0$. Poisson sampling is alias-free for arbitrary $\beta > 0$, implying that a sampling scheme can be alias-free even if the average sampling rate is far below the Nyquist-Landau rate. Another sampling scheme which was found to be alias-free (in the sense of the above definition) with respect to the set of spectra with power spectral densities in $L^1 \cap L^2$ is the scheme $t_n = t_{n-1} + \gamma$, where γ is a fixed random variable with distribution on $[0, \infty)$ and finite mean [9]. This sampling scheme corresponds to a uniform sampling pattern with a random but fixed spacing. Since no uniform sampling scheme can lead to a consistent spectrum estimator for arbitrary power spectra, the alias-free property described by the above definition does not guarantee that the spectrum can be consistently estimated from the randomly sampled values. It is therefore a necessary, but not a sufficient condition. The main problem of the above definition is that one is averaging over the sampling pattern $\{t_n\}$, which implies that the specific sampling times are not taken into account when estimating the spectrum. In fact, only the set of values $\{x(t_n)\}$ can be used for the spectrum estimation. A new definition of alias-free sampling that takes into account specific sampling times was therefore required. It is given in [10]. Sampling schemes that are alias-free relative to this latter definition but not to the first definition given can be found, implying that the two definitions are not equivalent. Poisson sampling is alias-free with respect to the latter definition as well [10]. Consistent estimators have been described for Poisson sampling [10, 11].

If a minimum separation between consecutive samples is prescribed, then it was shown that no sampling scheme is alias-free relative to both definitions for the set of all spectra [12]. For Poisson sampling, for example, no such minimum separation can be found if the number of samples grows to infinity. Consistent estimators have been described that take such a minimum separation into account [13]. A minimum separation is a practical constraint, as sampling devices cannot sample at an arbitrary fast rate.

The careful use of randomness in the construction of a sampling scheme has led to benefits not available to uniform sampling schemes. It is interesting to note that Poisson sampling, being closely related to stationary Poisson point processes, is the point process with the largest entropy among all point processes fulfilling some conditions (cf. [14]). It is hence a process with maximal uncertainty. Uniform sampling, on the other hand, is by definition as regular as possible. Relaxing the structure in one domain (the sampling scheme) has led to an increased set of spectra that can be resolved from the samples alone.

The first definition describing when a sampling scheme is alias-free relative to a family of spectra $\mathcal{S}$ is centered around the correlation sequence $\{r(n)\}$. It demands in particular that a one-to-one mapping exists between spectra and correlation sequences. Spectrum estimation can then proceed based solely on the knowledge of $\{r(n)\}$ (a consistent estimator does not exist in the general case). Recently, event-based sampling schemes have gained popularity. An example of such a scheme is level-crossing (cf. [15]). Here we show that level crossing sampling cannot be used to resolve any spectrum from the samples alone. The output of a level-crossing scheme is always the same value, implying that an empirical determination of the correlation sequence from the samples alone always yields the same sequence for all input processes. There is hence no one-to-one mapping from spectra to correlation sequences from which it follows that level-crossing is not alias-free. Another common example of an event-based sampling scheme is send-on-delta sampling (cf. [16]). A new sample is generated whenever the function has changed by some preset threshold since the last recorded sample. Multiplying the stochastic process with a fixed scalar changes the power spectral density only by a scalar as well, while the output of a send-on-delta sampler depends in a nonlinear way on the scalar. Spectrum estimators will in general produce different estimates depending on whether an unscaled/scaled stochastic process is sampled by a send-on-delta sampler, the relationship between the two estimates being nonlinear.

1.3 OUTLOOK AND STRUCTURE OF THESIS

The discussion in Sections 1.1 and 1.2 has shown that under certain conditions random sampling can be beneficial. In the case of reconstruction, stable reconstructability can be guaranteed for some types of irregular sampling. Uniform sampling is therefore not special, being just a specific case that fulfills some reconstructability conditions. For the estimation of the spectrum of a stochastic process, however, it has been seen that random sampling (which is different from deterministic irregular sampling) can provide benefits not available to uniform sampling, as it can lead to the existence of consistent power spectrum estimators even for non-bandlimited processes.

Chapter 2 investigates send-on-delta sampling in detail. It will be shown that functions living in Paley-Wiener spaces (which contain most real-world bandlimited functions) cannot be stably reconstructed from the samples. Modifications of the sampling scheme are then proposed that ensure reconstructability. The reason for the lack of stability of this sampling scheme in Paley-Wiener spaces is that only finitely many samples are generated irrespective of the precise threshold chosen. The global sampling density is hence equal to zero which is below the Nyquist-Landau rate. To ensure stability of reconstruction from finitely many samples, other restrictions have to be placed on the function besides the Paley-Wiener space restriction. Sparsity in some basis/frame is such an additional assumption. Chapter 3 uses this additional assumption of sparsity to show that stable reconstruction of a bandlimited function is indeed possible from finitely many samples. For this analysis, a prolate spheroidal wave functions basis is used, which for many reasons constitutes a preferred basis for bandlimited functions. This basis is in particular suited for describing effectively timelimited bandlimited functions, i.e. bandlimited functions for which most of the energy is concentrated in a finite time interval. For such effectively timelimited bandlimited functions, only finitely many expansion coefficients in a prolate spheroidal wave function basis expansion will have values departing significantly from zero, while the remaining coefficients will be (numerically speaking) very close to zero. The precise number of nonzero coefficients depends on the time-bandwidth product of the sampled functions, where time refers to the effective time interval to which the function is constrained. Using such sparsity results, infinite-dimensional compressed sensing theory can be leveraged. Infinite-dimensional compressed sensing is an extension of classical compressed sensing theory (which operates on finite dimensional spaces) to infinite-dimensional spaces such as Hilbert spaces of functions. Combining the reproducing kernel Hilbert space nature of Paley-Wiener spaces (pointwise sampling corresponds to inner products of the function with some measurement kernel) with the properties of the prolate spheroidal wave function basis, we can then show that finitely many samples are sufficient to guarantee stable reconstruction with a certain probability in case the samples are distributed with uniform probability in the interval of interest. The resulting reconstruction algorithm does not correspond to some series expansion any more (as in the usual sampling series for uniform sampling), but an l_1-optimization algorithm needs to be used in line with compressed sensing theory to reconstruct the function.

In Chapter 4 we discuss the filtering of irregularly sampled bandlimited functions. Conventionally, discrete-time filtering has been restricted to uniformly sampled bandlimited functions, as such sampling ensures (provided the sampling rate is higher than the Nyquist rate) that discrete-time processing is equivalent to an underlying continuous-time filtering. Under uniform sampling, analog filtering can therefore equivalently be directly carried out on a digital computer. In this chapter, we discuss a method that enables an equivalent operation on irregular samples, i.e. a mapping from continuous-time convolution to operations carried out only on irregularly sampled values. Conditions on the required sampling patterns are discussed in this chapter that enable such an equivalent representation. The results described in Chapter 4 therefore directly reflect those described in Section 1.1: under certain conditions, both reconstruction and filtering can be done based on irregular samples only. The mathematical description of reconstruction/filtering from irregular samples, however, is more involved than for the uniformly sampled counterpart.

In Chapter 5 sampling from a statistical perspective is discussed. The task studied is sequential prediction of stochastic processes. Having observed n samples, sample $n+1$ is to be predicted. As a specific instantiation of a strictly stationary stochastic process an Ornstein-Uhlenbeck process is taken. The observed realizations of the stochastic process, however, do not correspond to one stochastic process, but are instead drawn from a parametric family of stochastic processes. The parametric family indexes the parameters of the stochastic process. No fixed model can be optimal for all realizations from the parametric family at all times, for an optimal model is tailored to the specific set of parameters describing an element of the parametric family. A model that is optimal for the entire parametric family can therefore only be asymptotically optimal for a member of the parametric family. Notions of optimality are discussed in the literature on universal coding. An optimal model is dependent on the way realizations are drawn from the parametric family which in turn depends on the sampling prior on the parameters of the parametric family. For the case of the Ornstein-Uhlenbeck parametric family, a specific sampling prior exists which maximizes the information transmitted from the obtained realizations to the parameters indexing the parametric family, which is Jeffreys' prior. Jeffreys' prior is explicitly derived in Chapter 5 for the Ornstein-Uhlenbeck parametric family. We connect these results to notions of indistinguishability: having observed n samples, is it possible to statistically distinguish two different stochastic processes from the same parametric family from the n samples alone? Jeffreys' prior underlying the data generating process is then an optimal way to transmit as much information as possible from the parametric family to the stochastic process realizations. We then show by a concrete example that a machine learning algorithm attempting to learn a model for the entire parametric family will only generalize if the data are drawn according to Jeffreys' prior. By generalization we mean that prediction quality does not degrade if the test prior is different from the training prior. Jeffreys' prior turns out to be a worst-case prior: if a model is learned for this prior, it will equally work for any other (less information transmitting) prior. We then discuss the implications of these findings for machine learning.

The thesis concludes with a summary and an outlook to potential future research that could be based on the results presented herein.

2

ON SEND-ON-DELTA SAMPLING OF BANDLIMITED FUNCTIONS

Event-based sampling schemes have been used to exploit the sparse nature of many real-world signals. These sampling schemes follow a general abstract principle: a sample is generated whenever a suitably defined event has occurred in the measured signal. In the absence of activity, no samples are generated by event-based sampling schemes. Signals with intermittent activity are ubiquitous in the real world, be it speech or movement captured by a camera, for example. Another potential advantage of event-based sampling schemes is the fact that the sampling rate is dynamically adjusted by the activity that is being measured. Quickly changing scenes – quick has to be understood with respect to the events that are detected – produce more samples than slowly changing signals. The sampling pattern that results from the application of an event-based sampler to a signal is in general irregular.

Sampling schemes are classically formulated for bandlimited functions. In this chapter, we study an event-based sampling scheme known as Send-on-Delta sampling (cf. [18]) which is applied to bandlimited functions. This sampling scheme has found use in the design of biologically inspired hardware [19, 20]. Particular emphasis is placed on the question whether the resulting sampling pattern is sufficient for a full reconstruction of the underlying function, provided the function is bandlimited. As such, this chapter forms a bridge between the world of signal processing on the one hand and modern sampling schemes with origins in different domains.

2.1 INTRODUCTION

Classical sampling schemes such as the Whittaker-Kotelnikov-Shannon sampling method [21] sample a signal at a predetermined set of time instants $\{t_n\}_{n\in\mathbb{Z}}$, recording the amplitude values $\{f(t_n)\}_{n\in\mathbb{Z}}$ at these times. The sampling pattern itself is matched to the nature of the sampled function. The classical function space used for sampling expansions is the Paley-Wiener space, defined as $PW_{\pi w} := \{f : f \in L^2(\mathbb{R}) \bigcap C(\mathbb{R}), \mathrm{supp}\hat{f} \subseteq$

[1] Published in [17] © 2017 IEEE

$[-\pi w, \pi w], w > 0\}$, where $L^2(\mathbb{R})$ is the space of all square-integrable (Lebesgue) functions over $\mathbb{R}$, $C(\mathbb{R})$ is the space of all continuous functions on the real line, supp is the support of a function, and $\hat{f}$ is the Fourier transform of the function f: $\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt$. For $PW_{\pi w}$, the question of which sets of sampling times allow stable reconstruction was solved both for the case of sampling at the Nyquist-Landau rate (leading to a Riesz basis expansion) and for oversampling (leading to a frame expansion) [4, 21]. There are few studies such as the one in [22], where functions belonging to larger function spaces than $PW_{\pi w}$ are considered. [22] studies in particular functions belonging to the Bernstein class B_σ^∞. B_σ^∞ is defined as the space consisting of entire functions of exponential type at most σ that belong to $L^\infty(\mathbb{R})$ on the real line. Sufficient conditions on an oversampled sampling set that ensure reconstructability are given. In [23, 24], sampling of PW_π^∞ and B_π^∞ is studied. The sampling instants are the zeros of sine-type functions.

Event-based sampling schemes generate sampling points whenever the sampled function fulfills a condition for the generation of an event. Classical examples are Level-Crossing [15], Send-on-Delta [16] and Sine-Wave-Crossing sampling [25]. Reconstruction is possible from Sine-Wave-Crossing samples as long as certain requirements on the frequency and amplitude of the sine-wave sampler are met. Reconstruction from Level-Crossing samples has been studied with an approximate reconstruction solution for the special case of zero-crossings given in [26]. Another approach for event-driven sampling has been proposed in [27]. Time-encoding ensures reconstructability of bandlimited functions. In [28], the time-encoding approach is extended to general shift-invariant subspaces.

In this chapter, we study the behavior of Send-on-Delta sampling which has recently been used in event-based sensors [19, 20] that attempt to mimic biological sensors such as the cochlea or the retina. Send-on-Delta sampling is studied in $PW_{\pi w}$ and in B_σ^∞. In the latter case, we restrict our study to those functions that are not constant on time intervals above a certain length. Only real-valued functions are considered in this chapter. We show that Send-on-Delta sampling is a sampling scheme that generates a finite number of samples for functions in $PW_{\pi w}$ making stable reconstruction infeasible. For functions in our restricted B_σ^∞ function space, however, a sufficient amount of samples can be generated provided that the threshold chosen is a function of time. To the best of our knowledge we are the first to analyze the global behavior of Send-on-Delta sampling, contrasting with local analysis as carried out in [16].

2.2 SEND-ON-DELTA SAMPLING IN THE PALEY-WIENER SPACE

We show that Send-on-Delta sampling generates finitely many samples when sampling signals from $PW_{\pi w}$ irrespective of the chosen threshold. Our proof follows [29]. We start by noting the following:

$$f \in PW_{\pi w} \Rightarrow f \to 0 \quad \text{as} \quad t \to \pm\infty \tag{2.1}$$

The time at which sampling begins is denoted by $t_0 \in \mathbb{R}$. $\theta > 0$ is the Send-on-Delta threshold. A new sample is generated at time t_{j+1} if the following condition is met: $|f(t_{j+1}) - f(t_j)| = \theta$. We now prove the following proposition:

Proposition 1. *Given θ and t_0 as above, we have for $f \in PW_{\pi w}$:*

1. *A finite number of samples are produced by Send-on-Delta sampling.*
2. *The total number of samples is bounded from above by $\frac{\|f'\|_1}{\theta} + 1$, where f' is the derivative of f.*

Proof.

$$0 < \theta = \left|f(t_{j+1}) - f(t_j)\right| \leq \left|f(t_{j+1})\right| + \left|f(t_j)\right| \tag{2.2}$$

If the t_n tend to infinity, then by observation (3.23) the values of f at the sampling points tend to zero. This, however, contradicts inequality (2.2). All t_n must therefore be finite. Samples will therefore only be generated in a compact interval $[a, b]$. Consider such an interval $[a, b]$ and assume that m consecutive sampling times $\{t_j, ..., t_{j+m-1}\}$ are contained in $[a, b]$. If m is equal to 1, it is clear that the number of samples in this interval is finite. We can therefore restrict our attention to the case $m \geq 2$. Let $0 \leq k \leq m-2$. We have

$$\begin{aligned}\theta = |f(t_{j+k+1}) - f(t_{j+k})| &= \Big|\int_{t_{j+k}}^{t_{j+k+1}} f'(t)dt\Big| \\ &\leq \int_{t_{j+k}}^{t_{j+k+1}} |f'(t)|dt\end{aligned}$$

We sum up over all intervals set up by the sampling times $\{t_j, ..., t_{j+m-1}\}$ in $[a, b]$. We obtain

$$(m-1)\theta \leq \int_a^b |f'(t)|dt \tag{2.3}$$

With $a \to -\infty$ and $b \to \infty$ it follows from (2.3):

$$m \leq \frac{\|f'\|_1}{\theta} + 1 \tag{2.4}$$

Using Hölder's inequality on (2.3), we obtain the following bound for the number of sampling times in $[a, b]$:

$$m \leq \frac{\|f'\|_2}{\theta}(b-a)^{\frac{1}{2}} + 1$$

Since $\|f\|_2$ is finite by assumption, $\|f'\|_2$ is finite by Bernstein's inequality. The number of samples generated must therefore be finite. □

Remark 1. Level-Crossing sampling behaves fundamentally different from Send-on-Delta sampling as it can produce both a finite and an infinite amount of samples. Consider in particular the function $f(t) = \frac{1-\operatorname{sinc}(t)}{t}$ with $\operatorname{sinc}(t) = \frac{\sin(t)}{t}$. This function is bandlimited, its Fourier transform being $\hat{f}(\omega) = \frac{i}{2}\sqrt{\frac{\pi}{2}}(\operatorname{sgn}(1-\omega) + \omega\operatorname{sgn}(\omega-1) + 2\operatorname{sgn}(\omega) - \omega\operatorname{sgn}(\omega+1) - \operatorname{sgn}(\omega+1))$. f is furthermore in L^2: $\int_{-\infty}^{\infty} |f(t)|^2 dt = \frac{\pi}{3}$. f must therefore decay to zero for $t \to \pm\infty$. Consider now the intersection of f with a function $\forall t\ g(t) = c$, $c \in \mathbb{R}$. g corresponds to sampling levels, while c describes the position of the level. One obtains $\operatorname{sinc}(t) = 1 - ct$. As sinc is amplitude bounded for all t, the line $1 - ct$ will (for $c \neq 0$) eventually leave (for sufficiently large or small t) the amplitude region swept out by sinc. The time spent in this amplitude region is therefore limited. No line, however, can cross a Sinus Cardinalis function infinitely often in a finite time interval. For $c \neq 0$ the number of intersection points is therefore finite. For $c = 0$ the only intersection is at $t = 0$. In order to obtain infinitely many samples, one would therefore require an infinite set of sampling levels. Consider now $f(t) = \operatorname{sinc}(t)$. This function has infinitely many zero crossings. We conclude: contrary to Send-on-Delta sampling which always produces finitely many samples for signals from $PW_{\pi w}$, Level-Crossing sampling can result in both a finite and an infinite number of samples.

It is now clear that functions in $PW_{\pi w}$ cannot be reconstructed from samples generated with Send-on-Delta sampling. The global sampling density is in particular equal to zero and therefore below the Nyquist-Landau rate. Send-on-Delta is hence not a stable sampling scheme. Locally, the number of samples generated depends both on the threshold chosen and the local derivative of the function considered. The distribution of local samples is therefore dependent on both the function and the threshold chosen. In section 2.4 it will be seen, however, that the approximation error is bounded due to the very nature of Send-on-Delta sampling.

2.3 SEND-ON-DELTA SAMPLING IN A RESTRICTED BERNSTEIN CLASS

To produce infinitely many samples, it seems intuitively clear that what is needed as input are oscillatory functions with infinite energies, i.e. functions in $L^{\infty}(\mathbb{R})$ with minimum deviations from their local means. From now on, θ is a function of time: $\theta = \theta(t)$. By varying θ, we aim at obtaining sampling patterns that meet the requirements of [22] for reconstructability. The method described in [22] could then be used for reconstruction.

We consider functions f that live in B^{∞}_{σ}, are continuous and that adhere to the following restriction:

1. Oscillation = $\int_I |f(t) - m_I(f)|dt \geq C_I > 0$, where $m_I(f) = \frac{1}{|I|}\int_I f(t)dt$

Given a function f in B^{∞}_{σ}, I is an interval in which at least one sample is demanded. The size of I depends on σ. Our aim is therefore, considering interval I, to find a θ so that at least one sample is generated, assuming C_I is known. The required minimum local oscillation of the function rules out describing such a function in $PW_{\pi w}$.

Proposition 2. *Given $C_I > 0$, the number of samples in an interval I obtained from sampling a function living in the restricted B^{∞}_{σ} function space with Send-on-Delta will be greater than one if θ fulfills the following condition:*

$$\theta \leq \frac{C_I}{|I|}$$

Proof. As the mean of a function has no influence on the number of generated samples in an interval (we assume that a first sampling point is given), we henceforth consider only zero-mean functions in interval I. The functions which we consider therefore must fulfill the following conditions:

$$\int_I |f(t)|dt \geq C_I > 0 \tag{2.5}$$

$$\int_I f(t)dt = 0 \tag{2.6}$$

Assume first that only condition (2.5) is valid. A function that generates the least number of samples (no samples) is the constant function in I

with constant value $h = C_I/|I|$. Condition (2.6) is now assumed to hold, too. We split up I into two disjoint sets I_1 and I_2 ($|I_1| + |I_2| = |I|$). On I_1, $f(t)$ is assumed to have positive values, on I_2, $f(t)$ is assumed to have negative values. I_1 and I_2 are in general unions of countably many disjoint sets. Adding up all values of integrals evaluated on the disjoint sets that constitute I_1, we obtain a value of $C_I/2$. Equivalently we obtain a value of $-C_I/2$ for I_2. Let us now consider I_1. Irrespective of the ratio $|I_1|/|I|$, the mean of $f(t)$ on I_1 is equal to $\frac{C_I}{2|I_1|}$ and on I_2 equal to $-\frac{C_I}{2|I_2|}$. We study three cases:

$$|I_1|/|I| < 1/2 \text{ and } |I_2|/|I| > 1/2 \tag{2.7}$$

$$|I_1|/|I| > 1/2 \text{ and } |I_2|/|I| < 1/2 \tag{2.8}$$

$$|I_1|/|I| = 1/2 \text{ and } |I_2|/|I| = 1/2 \tag{2.9}$$

We introduce the following notation (see Fig. 2.1): $h_1^< = \frac{C_I}{2|I_1|} > \frac{C_I}{|I|} = h^{(0)}$, where $h_1^<$ corresponds to the mean of $f(t)$ in I_1 in case condition (2.7) applies, and $h^{(0)}$ corresponds to the mean of $f(t)$ in I in case only condition (2.5) applies (we only consider the lower limit). $h_1^>$ is the mean in I_1 in case condition (2.8) is valid. $h_2^<$ and $h_2^>$ are defined similarly. $h_1^= = -h_2^=$ correspond to the mean in I_1 and I_2 in case condition (2.9) applies. We start with condition (2.7) and deduce:

$$h_1^< > h^{(0)} \tag{2.10}$$

$$h_2^> > -h^{(0)} \tag{2.11}$$

Let us fix a coordinate system: $I = [0, x]$, $I_1 = [0, x_1]$ and $I_2 = (x_1, x]$, $x > x_1 > 0$. It is assumed that I_1 and I_2 are connected sets, as this minimizes the number of samples generated. We obtain

$$\frac{d}{dx_1} h_1^<(x_1) = -\frac{C_I}{2x_1^2} \tag{2.12}$$

$$\frac{d}{dx_1} h_2^>(x_1) = -\frac{C_I}{2(x - x_1)^2} \tag{2.13}$$

Since we assume at the moment that condition (2.7) holds, it follows that $-\frac{C_I}{2(x-x_1)^2} > -\frac{C_I}{2x_1^2}$. It then follows that $h_1^<$ grows faster than $h_2^>$ for decreasing x_1. This implies that with decreasing x_1, the difference between $h_1^<$ and $h_2^>$ grows. Equivalent results for $h_1^>$ and $h_2^<$ can be obtained if (2.8) is valid. The minimum difference between the mean values in I_1 and I_2 will therefore be found under condition (2.9), namely $h_1^= - h_2^= = 2h^{(0)}$. Since $f(t)$ is

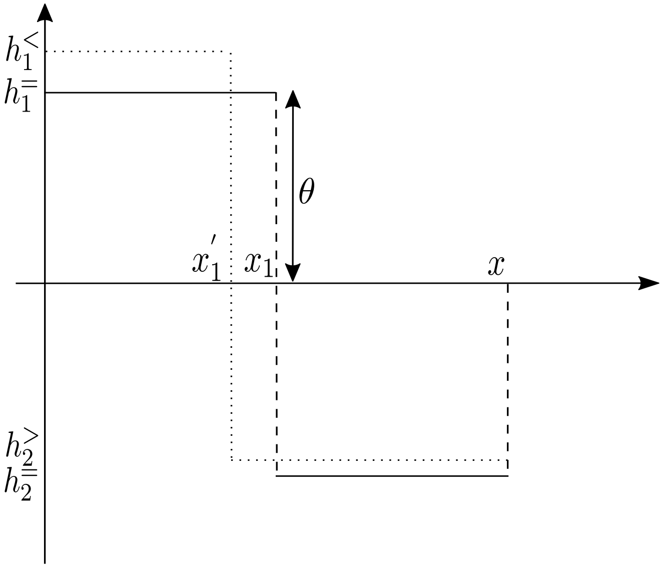

FIGURE 2.1: Illustration of the basic principle behind the proof of proposition 2 © 2017 IEEE

continuous, every point between either $h_1^<$ and $h_2^>$ or between $h_1^>$ and $h_2^<$ will have to be passed at least once. The minimum number of samples can therefore be found under condition (2.9). The number of samples generated in I depends on the last sample before I as well. By choosing $\theta = h^{(0)}$, we ensure that at least one sample will be generated in I, irrespective of the previous sample. □

Remark 2. In case the function f is amplitude bounded, we can determine a miminum separation between consecutive samples generated from sampling a function from the Paley-Wiener space with Send-on-Delta. Such a minimum separation result easily follows from Bernstein's inequality [30]: If $f \in PW_{\pi w}$, then $\|f'\|_\infty \leq \pi w \|f\|_\infty$. If $\|f\|_\infty = c, c > 0$, then $\Delta_t = \frac{\theta}{\pi w c}$, where Δ_t is the minimum separation between two consecutive samples. For functions for which no amplitude bound is known, such a miminum separation result cannot be found, at least not for functions that are sampled during a finite time interval [31].

Remark 3. The method described in proposition 2 for choosing θ to ensure at least one sample in an interval can also be used for functions from $PW_{\pi w}$. The local distribution of sampling times can be partially influenced this way. Since $\theta \to 0$ is not feasible in practice (this would be required due to notion (3.23)), the method does not generate infinitely many samples.

Remark 4. Let us give an example: for $f \in B^{\infty}_{\pi-\delta}$, $0 < \delta < \pi$, $|I|$ is equal to $1/2$. I can be shifted once a new sample has been obtained in such a way that the first time instant of I coincides with the last sampling time recorded.

Remark 5. If θ is varied with time, both the sample times and θ need to be stored for reconstruction.

Remark 6. If too many samples are generated in a specific I, samples should be deleted to meet the requirements of [22] for reconstruction. For $f \in B^{\infty}_{\pi-\delta}$, $0 < \delta < \pi$, only those samples closest to the integers should be kept, i.e. only one sample should remain in $|t - n| < 1/4$, $n \in \mathbb{N}$; all other samples should be deleted. If a local approximation is performed, however, samples should be kept to ensure good local approximation to the underlying function.

Remark 7. Estimation of the precise value of C_I seems to be difficult in general. If a minimal C_I was known for the entire function, θ could be adjusted once and for all, making sure that a sufficient amount of samples would be generated. This, however, would incur a surplus in samples in regions of the function where there is a larger local oscillation than such a minimum oscillation. Setting θ once by adjusting it to the minimum oscillation, however, alleviates one from having to store the temporal development of θ.

Remark 8. If no upper bound for the L^2 norm of the function is known, then no upper bound on the number of samples generated in a finite interval by means of Send-on-Delta sampling can be given. This follows directly from the theory of superoscillations [31]. Functions in $PW_{\pi w}$ can have arbitrary finite derivatives locally, irrespective of the particular bandwidth πw.

2.4 APPROXIMATION IN THE PALEY-WIENER SPACE

As shown above, stable reconstruction is not feasible for functions in $PW_{\pi w}$ sampled with Send-on-Delta. Given that an ideal reconstruction cannot be achieved, one can attempt to obtain a faithful approximation.

Besides providing timing information and (indirectly) amplitude information at the sampling times, Send-on-Delta implicitly restricts possible function values between two sampling points. For $t \in (t_j, t_{j+1})$, where t_j and t_{j+1} are consecutive sampling times,

$$|f(t) - f(t_j)| < \theta \tag{2.14}$$

is valid. By the above observation (equation (3.23)), $f(t)$ decays to zero for $t \to \pm\infty$. For $t_0 \ll 0$, we assume that

$$|f(t_0) - f(t)| < \theta, t < t_0, \tag{2.15}$$

i.e. no sample would have been generated prior to the first sample (cf. [29]). By choosing a Schwartz class function ψ with $\hat{\psi} = 1$ on $[-\pi w, \pi w]$, $w > 0$ and $\hat{\psi}$ compactly supported, $f(t)$ can be reconstructed from an oversampled sampling set:

$$f(t) = \sum_{k \in \mathbb{Z}} f(\beta k)\psi(t - \beta k), 0 < \beta < 1/w$$

β depends on the particular ψ chosen. Approximations $\tilde{f}(\beta k)$ to the $f(\beta k)$ values on the oversampled grid can be obtained by any interpolation method that fulfills conditions (2.14) and (2.15). Feasible options are in particular linear interpolation and nonlinear constrained parametric spline interpolation [32]. Both methods lead to error estimates on the estimated samples:

$$|f(\beta k) - \tilde{f}(\beta k)| \leq 2\theta, k \in \mathbb{Z}$$

The infinite sum

$$\tilde{f}(t) = \sum_{k \in \mathbb{Z}} \tilde{f}(\beta k)\psi(t - \beta k)$$

converges uniformly [29].

If the approximated/reconstructed function is itself not required to be in $PW_{\pi w}$, other solutions can be found [33, 34]. Using such variable bandwidth theorems, however, does not ensure that condition (2.14) is satisfied in the approximated function. A very precise reconstruction can be obtained for those intervals of the sampled function, however, that contain large oscillations away from their local mean.

2.5 CONCLUSION

We have shown that Send-on-Delta is a sampling scheme that allows for only approximate reconstructions of functions in $PW_{\pi w}$ as it generates finitely many samples. For functions in a restricted B^{∞}_{σ} space, a method for changing θ is described that ensures that a sufficient number of samples are recorded. For $PW_{\pi w}$, only an approximation to the true function is feasible. Due to the particular nature of Send-on-Delta sampling, however, the error in the approximate reconstruction can be bounded from above. Send-on-Delta sampling can therefore be used only if mathematically precise reconstruction is not necessary since all real-world signals will belong to $PW_{\pi w}$ than to infinite-energy signals contained in B_σ^∞. Since all practically implementable sampling schemes will only yield approximations to a signal (sampling can only occur during a finite time interval and with finite precision), however, it is justified to use Send-on-Delta sampling, particularly if sampling hardware is required to be energy-efficient.

3

ON APPROXIMATION OF BANDLIMITED FUNCTIONS WITH COMPRESSED SENSING

Compressed sensing as a field was started with two seminal publications by Donoho [36] and Candès, Romberg and Tao [37]. Most compressed sensing theory was developed for finite dimensional spaces to deal with the following problem: given a vector in $\mathbb{C}^N$ or $\mathbb{R}^N$, is it possible to reconstruct it from $m < N$ measurements provided the vector is sparse in some transform domain? Here we briefly outline major developments made in answering this question. The description is based on a standard exposition of the field by Foucart and Rauhut [38].

Given a sparse vector $x \in \mathbb{C}^N$ and a measurement matrix $A \in \mathbb{C}^{m \times N}$, is it possible to reconstruct x from $y = Ax$, knowing that x is sparse? Let us assume that x is s-sparse, i.e. it has at most s non-zero entries. Two different cases can now be distinguished. The first case aims at determining conditions that the measurement matrix A needs to fulfill such that all s-sparse vectors x can be recovered from the measurements y (the uniform setting), while the second case looks at conditions A needs to fulfill for a fixed x to be recoverable from y (the nonuniform setting). The minimal number of measurements m is equal to $2s$ in the first case and to $s+1$ in the second. Explicit measurement matrices A can be constructed that reach this lower bound in the number of measurements. $2s$ resp. $s+1$ measurements, however, do not provide stability with respect to a lack of precise sparsity and measurement errors. For the reconstruction from $2s$ resp. $s+1$ measurements (the minimal number of measurements) it is necessary to solve the following optimization problem

$$\min_{z \in \mathbb{C}^N} \|z\|_0 \quad \text{subject to } Az = y, \tag{3.1}$$

where $\|\cdot\|_0$ counts the number of non-zero entries in the argument. No algorithms are known that solve the optimization problem (3.1) efficiently for realistic m and N. The optimization problem is therefore relaxed to a convex optimization problem

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \quad \text{subject to } Az = y, \tag{3.2}$$

[2] Published in [35] © 2018 IEEE

which is also known as basis pursuit. A more general optimization problem known as quadratically constrained basis pursuit

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \quad \text{subject to } \|Az - y\|_2 \leq \eta \tag{3.3}$$

takes into account measurement noise by introducing η, the noise level, into the optimization procedure. Conditions on A can now be derived that ensure that (3.2) and (3.3) yield the correct result x. In the uniform setting, conditions known under the name null space property need to be fulfilled. In the noiseless case, the null space property demands that vectors living in the null space of matrix A are not too sparse. If the null space property is fulfilled, then (3.2) yields the same solution as (3.1). For the nonuniform recovery setting, finer conditions on A and x can be derived (the conditions obviously depend now on x as well) which ensure reconstructability of x via (3.2) or (3.3). If A guarantees reconstructability in the uniform sense for all s-sparse x, then it guarantees it for the nonuniform setting as well.

As a verification of the null space property is not trivial for a given matrix, other conditions have been derived that ensure that (3.2) or (3.3) find the true solution. One such condition is the restricted isometry property. A measurement matrix A fulfills the restricted isometry property if the smallest $\delta \geq 0$ such that

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \quad \text{for all } s\text{-sparse } x \tag{3.4}$$

is small for large enough s. Explicitly constructed matrices are known that fulfill the restricted isometry property if the number of measurements m is bounded from below by $m \geq Cs^2$ with C a constant. It is important to iterate that the discussion so far has not assumed that any quantity, that is neither A nor x, is random. If the measurement matrix has a sufficiently small restricted isometry constant, then the optimization procedures (3.2) and (3.3) yield an exact or approximate reconstruction of x from the measurements y. The introduction of randomness (subgaussian random matrices can be used) in the construction of A can be used to lower the required number of measurements from Cs^2 to $m \geq C_\delta s \log(eN/s)$, where C_δ is a constant depending on δ (an upper bound that the restricted isometry constant should fulfill). If the measurement matrix is random, however, then the restricted isometry constant is only smaller than δ with a probability at least $1-\epsilon$ ($\epsilon > 0$ enters in a precise lower bound for the minimal number of measurement m – the smaller ϵ, the larger m has to be). Reconstruction results obtained from (3.2) or (3.3) with A a subgaussian random matrix, however, are then also only exact or approximate (with error bounds depending on the noise η) with a probability at least equal to $1-\epsilon$.

Such lower bounds on the required number of measurements m are essentially optimal. To see this, the notion of Gelfand m-width is introduced. Given a normed space X and a subset K thereof, define the Gelfand m-width as

$$d^m(K,X) := \inf_{L^m}\left\{\sup_{x\in K\cap L^m}\|x\|, L^m \text{ subspace of } X \text{ with } \operatorname{codim}(L^m)\le m\right\}. \tag{3.5}$$

$\|\cdot\|$ denotes the norm with which the normed space is equipped. The Gelfand m-width therefore attempts to find a subspace in which x is small. A related notion is the Kolmogorov m-width

$$d_m(K,X) := \inf_{X_m}\left\{\sup_{x\in K}\inf_{z\in X_m}\|x-z\|, X_m \text{ subspace of } X \text{ with } \dim(X_m)\le m\right\}, \tag{3.6}$$

which measures the extent to which K can be approximated by a linear subspace. Define the notion of compressive m-width

$$E^m(K,X) := \inf_{A,\Delta}\left\{\sup_{x\in K}\|x-\Delta(Ax)\|, A : X\to\mathbb{R}^m \text{ linear}, \Delta : \mathbb{R}^m \to X\right\}, \tag{3.7}$$

where A is a nonadaptive linear map, and Δ an arbitrary reconstruction map. If the subset K satisfies $-K=K$ and $K+K\subset aK$ for some constant $a>0$, then

$$d^m(K,X)\le E^m(K,X)\le ad^m(K,X). \tag{3.8}$$

The unit ball $B_1^N := \left\{z\in\mathbb{C}^N : \|z\|_1 \le 1\right\}$ in l_1^N models compressible vectors well. If K is chosen as B_1^N and X as l_p^N for $1<p\le 2$, then the Gelfand m-width of K in X can be upper and lower bounded as follows:

$$c_1\min\left\{1,\frac{\log(eN/m)}{m}\right\}^{1-1/p}\le d^m\left(B_1^N,l_p^N\right)\le c_2\min\left\{1,\frac{\log(eN/m)}{m}\right\}^{1-1/p} \tag{3.9}$$

for $m<N$ and constants $c_1,c_2>0$. Eq.(3.9) leads directly to the following statement (Proposition 10.7 in [38]) for the recovery of vectors that are not exactly sparse, but only compressible (stability property): If some matrix $A\in\mathbb{R}^{m\times N}$ and a reconstruction map $\Delta:\mathbb{R}^m\to\mathbb{R}^N$ exist such that

$$\|x-\Delta(Ax)\|_p \le \frac{C}{s^{1-1/p}}\inf_{z\in\mathbb{C}^N}\left\{\|x-z\|_1, z \text{ is } s\text{-sparse}\right\} \tag{3.10}$$

for $1 < p \leq 2$ and all $x \in \mathbb{R}^N$, then there exist constant $c_1, c_2 > 0$ (which depend only on C) such that

$$m \geq c_1 s \log\left(\frac{eN}{s}\right) \tag{3.11}$$

if $s > c_2$. Eq. (3.10) can be compared to the sharp estimate

$$\inf_{z \in \mathbb{C}^N} \left\{ \|x - z\|_p \,,\; z \text{ is } s\text{-sparse} \right\} \leq \frac{c_{1,p}}{s^{1-1/p}} \|x\|_1 \tag{3.12}$$

with $c_{1,p} = \left(\frac{1}{p}\right)^{1/p} \left(1 - \frac{1}{p}\right)^{1-1/p}$. Eq. (3.12) and (3.10) are bounded by similar terms from the right. The optimal sharp estimate in Eq. (3.12) bounds the best s-term approximation to x from above (this best estimate requires knowing the s most important entries of x). The minimal number of measurements needed as described by Eq. (3.11) enable a similar reconstruction error, but now the knowledge of the s most important terms of x is not required. Only a slightly larger number of nonadaptive measurements are necessary to obtain the same information as knowing the s most important entries *a priori*.

The introduction of randomness in the construction of the measurement matrix allows for Eq. (3.10) and (3.11) to be reached. Moreover, the reconstruction map Δ is explicitly given as well. The careful use of randomness has therefore enabled more efficient sampling than what can be achieved with the currently known best deterministic sampling schemes.

In the rest of this chapter, we describe the use of compressed sensing (specifically infinite-dimensional versions thereof) to enable the stable sampling and reconstruction of bandlimited functions from only finitely many pointwise samples.

3.1 INTRODUCTION

The approximate reconstruction of a bandlimited function in an interval from a finite number of samples is a well-studied problem. The classical Whittaker-Shannon-Kotelnikov sampling theorem requires infinitely many samples for the reconstruction of the entire bandlimited function [21]; reconstruction within a compact interval from samples taken at the Nyquist rate can lead to major errors as the sinc-function decays only with a rate of $\frac{1}{t}$ for $t \to \infty$, i.e. samples from far outside the interval to be reconstructed can influence the values of the function in the interval to a major extent. Such truncation errors have been thoroughly studied in the literature. By introducing oversampling, local reconstruction can be achieved with far greater accuracy, as the reconstruction function can be chosen to decay much faster in time than the sinc-function associated with Nyquist-rate sampling. Errors arising from truncated sinc-expansions in the oversampling regime have been studied by Helms and Thomas [39]; the same authors also developed bounds for the truncation errors in case of self-truncating sampling expansions. Truncation error bounds for finite-energy signals were derived by Brown [40]. A related approach was followed by Knab who analyzed error bounds arising from estimating a bandlimited function in an interval from a finite number of equidistant samples using Lagrange interpolation [41]. Error bounds for Lagrange interpolation from equidistant samples of a bandlimited function depending on the sampling rate and the Nyquist rate were developed by Radzyner and Bason [42]. Klamer and Masry studied error bounds arising from Lagrange interpolation of bandlimited functions with finitely many sampling points distributed according to a point process [43]. They derived error bounds for sampling points distributed according to a Poisson point process in particular. Strohmer and Tanner considered nonuniform periodic sampling, deriving a reconstruction algorithm using a finite number of samples [44]. Returning to Lagrange interpolation, Selva considered a weighted Lagrange interpolation scheme for the local approximation of a bandlimited function from nonuniform samples [45]. Explicit error bounds were given for nonuniform sampling schemes with a maximum deviation of individual samples from a uniform grid.

As described in the beginning of this chapter, compressed sensing (CS) studies the solution of underdetermined linear systems, exploiting random measurements and the sparsity of the signal to be reconstructed. Classical CS theory was developed for finite-dimensional spaces. Recently, Adcock and Hansen extended CS theory to infinite-dimensional spaces, thereby enabling the application of CS to functions living in Hilbert spaces [46].

Through the combination of infinite-dimensional CS with the theory of Prolate Spheroidal Wave Functions (PSWF), we derive approximation methods for bandlimited functions. A similar approach which can be found in [47] uses CS to recover functions sampled pointwise, assuming that the functions are sparse in a PSWF basis. The main difference of our work from [47] is that we lower the lower bounds on the number of measurements sufficient for faithful approximation; additionally, our method does not require sampling points distributed according to a Chebyshev distribution for reconstructing expansion coefficients above a certain index. Instead, uniform sampling can be used throughout in our formulation. In Section 3.2 we discuss the main aspects of infinite-dimensional CS and recapitulate the basics of PSWF and Reproducing Kernel Hilbert Spaces (RKHS). Thereafter we derive our main results.

3.1.1 *Notation*

The version of the Fourier transform used in this chapter is

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt.$$

All considered functions live in the Paley-Wiener space $PW_{\pi w} := \{f : f \in L^2(\mathbb{R}) \bigcap C(\mathbb{R}), \text{supp}\hat{f} \subseteq [-\pi w, \pi w], w > 0\}$, where $L^2(\mathbb{R})$ is the space of all square-integrable (Lebesgue) functions over $\mathbb{R}$, $C(\mathbb{R})$ is the space of all continuous functions on the real line, supp is the support of a function, and $\hat{f}$ is the Fourier transform of the function f. The reproducing kernel of $PW_{\pi w}$ is $k_{\pi w}(t,s) = w \cdot \text{sinc}(w(t-s))$, where $\text{sinc}(t) = \frac{\sin(\pi t)}{\pi t}$. The coherence $v(U)$ of an infinite matrix U is defined as $v(U) = \sup_{i,j \in \mathbb{N}} |u_{ij}|$ with u_{ij} the entries of matrix U. $\langle \cdot, \cdot \rangle$ denotes the inner product in a generic Hilbert space H over $\mathbb{C}$. The effective interval in which most of the function energy is concentrated has length $T > 0$.

3.2 INFINITE-DIMENSIONAL COMPRESSED SENSING

Let H be a separable Hilbert space with an orthonormal basis $\{\phi_j\}_{j \in \mathbb{N}}$. Then every function in H can be expanded as $f = \sum_{j=1}^{\infty} \alpha_j \phi_j$ with $\alpha_j = \langle f, \phi_j \rangle$. Let $\Delta = \text{supp}(f) \subset \{1, \ldots, M\}$ with $M \in \mathbb{N}$ and $\text{supp}(f) = \{j \in \mathbb{N} : \alpha_j \neq 0\}$. If $|\Delta| = r$, f is (r, M)-sparse in the basis $\{\phi_j\}_{j \in \mathbb{N}}$. The best approximation error for compressible signals (see [38]) can then be defined as

$$\sigma_{r,M}(\alpha) = \min\{\|\alpha - \eta\|_1 : \eta \text{ is } (r, M)\text{-sparse}\}. \tag{3.13}$$

Let $\zeta_1(f)$, $\zeta_2(f)$, ..., be a countable collection of samples with $\zeta_j(f) = \langle f, \psi_j \rangle$ and $\{\psi_j\}_{j \in \mathbb{N}}$ an orthonormal basis for H. The infinite matrix

$$U = \begin{pmatrix} \langle \phi_1, \psi_1 \rangle & \langle \phi_2, \psi_1 \rangle & \langle \phi_3, \psi_1 \rangle & \cdots \\ \langle \phi_1, \psi_2 \rangle & \langle \phi_2, \psi_2 \rangle & \langle \phi_3, \psi_2 \rangle & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{3.14}$$

is an isometry. Suppose that m of the first N measurements $\zeta_j(f)$ are chosen uniformly at random with the position of the chosen measurements being indicated by $\Omega \subset \{1, \ldots, N\}$ and $|\Omega| = m$. Define P_Ω to be the orthogonal projection from $l^2(\mathbb{N})$ to $\text{span}\{e_j : j \in \Omega\}$ with $\{e_j : j \in \mathbb{N}\}$ the canonical basis of $l^2(\mathbb{N})$ and P_M the orthogonal projection to $\text{span}\{e_j : j = 1, \ldots, M\}$. One result from [46] reads then as follows: Provided certain technical requirements on N and m are met, then by solving the finite-dimensional problem

$$\inf_{\eta \in l^1(\mathbb{N})} \|\eta\|_1 \text{ subject to } P_\Omega U P_M \eta = P_\Omega \zeta, \tag{3.15}$$

a solution ξ can be found with probability $1 - \epsilon$, $\epsilon > 0$, which is close in norm to the true solution α:

$$\|\xi - \alpha\| \leq 8 \left(1 + \frac{2N}{m}\right) \sigma_{|\Delta|,M}(\alpha). \tag{3.16}$$

One requirement on m is

$$m \geq C \cdot N \cdot v^2(U) \cdot |\Delta| \cdot \left(\log\left(\epsilon^{-1}\right) + 1\right) \cdot \log\left(\frac{MN\sqrt{|\Delta|}}{m}\right), \tag{3.17}$$

i.e. the number of necessary measurements m in order to obtain a faithful reconstruction with sufficiently high probability is bounded from below. C is a fixed constant in Eq. (3.17).

3.2.1 *Prolate Spheroidal Wave Functions*

We recapitulate the basics of Prolate Spheroidal Wave Functions (PSWF) briefly. The PSWF were introduced into signal analysis in a series of papers [48–50]. They are solutions to both a differential equation and to an integral equation, forming an orthonormal basis for the Paley-Wiener space on the real line. The PSWF satisfy the integral equation

$$\int_{-1}^{1} \frac{\sin(c(x-y))}{\pi(x-y)} \phi(y) dy = \lambda \phi(x), \; |x| \leq 1, \tag{3.18}$$

where $c = \frac{\pi w T}{2}$. The differential equation which the PSWF satisfy is

$$\frac{d}{dx}(1-x^2)\frac{d\phi}{dx} + (\chi - c^2 x^2)\phi = 0. \tag{3.19}$$

Equation (3.18) has solutions for discrete values of λ which can be sorted in a descending order:

$$\lambda_0 > \lambda_1 > \lambda_2 > \cdots > 0\,. \tag{3.20}$$

Possible eigenvalues λ are functions of c, i.e. $\lambda_i = \lambda_i(c), i \in \mathbb{N}_0$. The PSWF basis numbering starts with 0, contrary to the generic basis $\{\phi\}_{i\in\mathbb{N}}$. The PSWF constitute an optimal basis for the space of bandlimited functions [51]. We consider a subclass of the Paley-Wiener space of functions with maximum energy E. Then the Kolmogorov n-width d_n in $L_2(-\frac{T}{2}, \frac{T}{2})$ of the energy-bounded subclass of the Paley-Wiener space equipped with the L_2-norm in $\left(-\frac{T}{2}, \frac{T}{2}\right)$ is equal to $d_n = \sqrt{E\lambda_n}$ and the subspace which leads to this infinum is $S_n = \text{span}(\phi_0, \phi_1, \ldots, \phi_{n-1})$, ϕ_j being the PSWF. The best approximation to a function f in an interval living in the energy-bounded Paley-Wiener space in any n-dimensional subspace is then given by $\sum_{j=0}^{n-1}\langle f, \phi_j\rangle\phi_j$. The worst case error arising from this approximation is equal to $\sqrt{E\lambda_n}$.

3.2.2 *Reproducing Kernel Hilbert Space*

$PW_{\pi w}$ is a Reproducing Kernel Hilbert Space (RKHS). The reproducing kernel is given by $k_{\pi w}(t,s) = w \cdot \text{sinc}(w(t-s))$. If s is chosen to correspond to $\{\frac{n}{w}\}_{n\in\mathbb{Z}}$, then the set $\frac{1}{\sqrt{w}}\{k_{\pi w}(t, \frac{n}{w})\}_{n\in\mathbb{Z}}$ is an orthonormal basis for $PW_{\pi w}$. Sampling at the Nyquist rate corresponds therefore implicitly (after normalization of sample values by $\frac{1}{\sqrt{w}}$) to inner products with an orthonormal basis. In the case of oversampling (samples are taken at the rate $\{\frac{n}{w'}\}_{n\in\mathbb{Z}}$, $w' > w$) the induced set of functions

$$\frac{1}{\sqrt{w'}}\{k_{\pi w'}(t, \frac{n}{w'})\}_{n\in\mathbb{Z}}, \tag{3.21}$$

forms a tight frame with unit frame bound in $PW_{\pi w}$. Consider now the infinite matrix U from Eq. (3.14) with $\{\psi_i\}_{i\in\mathbb{N}}$ a tight frame with unit frame bound and $\{\phi_i\}_{i\in\mathbb{N}}$ an orthonormal basis for $PW_{\pi w}$. U is an isometry. Indeed,

$$U\alpha = U\begin{pmatrix}\langle f, \phi_1\rangle \\ \langle f, \phi_2\rangle \\ \langle f, \phi_3\rangle \\ \vdots\end{pmatrix} = \begin{pmatrix}\langle f, \psi_1\rangle \\ \langle f, \psi_2\rangle \\ \langle f, \psi_3\rangle \\ \vdots\end{pmatrix} = \zeta(f), \tag{3.22}$$

with $f \in PW_{\pi w}$. By Parseval's identity it holds that $\|\alpha\|^2 = \|f\|^2$. $\|\zeta(f)\|^2$ is furthermore equal to $\|f\|^2$ as the set $\{\psi_i\}_{i\in\mathbb{N}}$ is a tight frame with unit frame bound. Hence $\|U\alpha\|^2 = \|\alpha\|^2$ and U is an isometry.

3.2.3 *Infinite Dimensional Compressed Sensing with PSWF basis*

3.2.3.1 *Approximation on the real line*

We estimate an upper bound for the coherence of matrix U defined in Eq. (3.14) with the PSWF basis $\{\phi_i\}$ and the tight frame $\{\psi_i\}$ from Eq. (3.21). All PSWF have unit energy on the real line. A standard result from the theory of RKHS (see [52]) yields that the maximum value attainable by a function from the RKHS at an arbitrary point $t_0 \in \mathbb{R}$, $\mathbb{R}$ being the domain, assuming that the function has energy E, is given by $\max_{\|f\|^2\leq E} |f(t_0)|^2 = Ek(t_0, t_0)$, $k(t,s)$ being the reproducing kernel. It follows that the coherence of the infinite matrix U is bounded from above by $v(U) \leq \sqrt{\frac{w}{w'}}$. We conclude that the coherence of the infinite matrix U in Eq. (3.14) depends on the bandwidth πw and on the sampling rate $w' > w$. By oversampling suitably, the coherence of U can be made as small as required.

Let us now study the term $N \cdot v^2(U)$ from Eq. (3.17). As discussed above, an upper bound for the squared coherence $v^2(U)$ of matrix U in Eq. (3.14) is equal to w divided by w'. Furthermore, the following is valid:

$$f \in PW_{\pi w} \Rightarrow f \to 0 \quad \text{as} \quad t \to \pm\infty. \tag{3.23}$$

Therefore, for every function in $PW_{\pi w}$ there exists a $T > 0$ such that almost all of the energy of the function is located in $\left[-\frac{T}{2}, \frac{T}{2}\right]$. We consider equispaced sampling points in $\left[-\frac{T}{2}, \frac{T}{2}\right]$, spaced apart by $\frac{1}{w'}$. N, the number of sampling points in $\left[-\frac{T}{2}, \frac{T}{2}\right]$, is therefore linear in w'. The matrix $P_N U P_M$ from which m rows are drawn with uniform distribution is therefore close to an isometry, implying that the theorems from [46] apply. The product of N and $v^2(U)$ reduces to a constant value which depends on w. Disregarding log-terms, the number of measurements needed scales then linearly with $|\Delta|$ as can be seen from Eq. (3.17):

$$m \geq C \cdot T \cdot w \cdot |\Delta| \cdot \left(\log\left(\epsilon^{-1}\right) + 1\right) \cdot \log\left(\frac{MN\sqrt{|\Delta|}}{m}\right). \tag{3.24}$$

In the case of sampling the function *globally*, the number of measurements would therefore be proportional to the support $|\Delta|$ and the time-bandwidth product $T \cdot w$. For a general bandlimited function, one cannot assume a priori knowledge on Δ. By assuming sparsity, however, a bandlimited function can be fully reconstructed from finitely many samples in a stable way. Without the assumption of sparsity, infinitely many samples are necessary for a sampling set to be stable [21]. A stable sampling scheme in this sense is then also a set of uniqueness, i.e. there is only one bandlimited function whose values at the sampling set correspond to the sampled values.

Remark 1. In principle, one could choose a different orthonormal basis for $PW_{\pi w}$ than the one set up by the PSWF. Since bandlimited functions are contained in $L^2(\mathbb{R})$ on the real line, any orthonormal basis for $L^2(\mathbb{R})$ would suffice. The reproducing kernel for $PW_{\pi w}$, however, acts as a lowpass filter, implying that any non-bandlimited basis element from such a hypothetical orthonormal basis would first have to be ideally lowpass-filtered before being evaluated at a specific point. Since the PSWF are bandlimited themselves, pointwise evaluation suffices.

Remark 2. The size of N (or equivalently of w') influences the irregularity of the sampling pattern. The larger N becomes, the more the sampling process resembles truly uniform sampling on the interval $\left[-\frac{T}{2}, \frac{T}{2}\right]$. In fact, a larger N implies a greater possible sampling pattern irregularity in $\left[-\frac{T}{2}, \frac{T}{2}\right]$. This can be seen as follows: Set $C' = C \cdot N \cdot v^2(U) \cdot \left(\log\left(\epsilon^{-1}\right) + 1\right)$ (C' is a constant for fixed ϵ), and $|\Delta| = M$. We obtain the following inequality from Eq. (3.17):

$$e^m m^{C'M} \geq C'' N^{C'M}, \tag{3.25}$$

with $C'' = \left(M\sqrt{M}\right)^{C'M}$. In the case of a growing N, the required m in order to fulfil Eq. (3.25) (and hence Eq. (3.17)) will grow slower than N. This can be seen by comparing the derivative of the left hand side of Eq. (3.25) with respect to m with the derivative of the right hand side with respect to N. For the left hand side one obtains $e^m m^{C'M} e^{-1}\left(C'M + m\right)$ and for the right hand side $C'' N^{C'M} N^{-1} C'M$. If N and m are chosen in such a way as to fulfil Eq. (3.25) it follows that

$$\begin{aligned} & e^m m^{C'M} e^{-1}\left(C'M + m\right) \geq C'' N^{C'M} e^{-1}\left(C'M + m\right) \\ & \geq C'' N^{C'M} N^{-1} C'M, \end{aligned} \tag{3.26}$$

provided that $C'M + m \geq eN^{-1}C'M$. The latter inequality is certainly fulfilled as N is always larger than e.

3.2.3.2 *Approximation in intervals*

Assume now that we are only interested in the approximation of a bandlimited function in an interval. In general, finitely many samples in an interval do not determine a bandlimited function uniquely. By restricting a bandlimited function to an interval, the resulting function space ceases to be a RKHS, that is, pointwise sampling is no longer continuous. After normalization, the PSWF form an orthonormal basis for the interval of interest. Within the interval, an upper bound for the values of the PSWF cannot be obtained from RKHS techniques as in Section 3.2.3.1. It is known, however, that for large enough integers $n \geq 0$ the largest absolute value of the normalized PSWF can be found at $-\frac{T}{2}$ and $\frac{T}{2}$ [53]. Furthermore, an upper bound for this largest value is proportional to $\sqrt{n}$ [53]. Hence the approach from Section 3.2.3.1 cannot be used, as no upper bound for the coherence can be given. Following Corollary 7.1 in [54], we use a weighted minimization scheme to solve the interpolation problem in the interval by introducing weights $\{w_i\}$ which grow as fast as the maximum value of the PSWF in our interval of interest, i.e. with rate $\sqrt{i}$. Following the line of argument given in [54], it transpires that the number of measurements needed in the interval of interest is proportional to M^2, with M the largest integer for which a coefficient that is nonzero is expected. As discussed above in Section 3.2.1, the worst case error of functions from the energy-bounded Paley-Wiener space expanded in a subspace spanned by the first n PSWF basis elements is equal to $\sqrt{E\lambda_n}$. Given that λ decays rapidly for $n > \frac{2c}{\pi}$, in general one will need more than $\frac{2c}{\pi}$ PSWF basis elements for an acceptable worst case approximation error. Equating n with M and setting $\Delta = \{1, \ldots, M\}$ (in general, all coefficients $\langle f, \phi_j \rangle$ for $j \in \{1, \ldots, M\}$ will be nonzero), one can conclude that one has to oversample locally, as $\frac{2c}{\pi}$ corresponds to the number of Nyquist-rate samples in an interval $[-\frac{T}{2}, \frac{T}{2}]$.

It is interesting to observe the qualitative difference between the sampling of bandlimited functions on the real line and on intervals assuming sparsity. In the former case, the number of sampling points sufficient scales linearly with $|\Delta|$, in the latter case quadratically. As mentioned above, one of the reasons for this behavior is the lack of continuity in the sampling process in the time-limited case since the space is no longer a RKHS.

Remark 3. A different approach to approximate a bandlimited function locally from uniformly distributed samples can be based on previous work showing how to determine the Restricted Isometry Property (RIP) for finitely many measurements in potentially infinite-dimensional Hilbert spaces [55].

It is known (see [5]) that for sampling times $\{t_n\}_{n\in\mathbb{Z}}$ with a maximum separation $\delta = \sup_{n\in\mathbb{Z}} (t_{n+1} - t_n) < \frac{1}{w}$ the following is true for any $f \in PW_{\pi w}$:

$$(1-\delta w)^2 \|f\|^2 \leq \sum_{n\in\mathbb{Z}} \omega_n |f(t_n)|^2 \leq (1+\delta w)^2 \|f\|^2 \tag{3.27}$$

with $\omega_n = \frac{t_{n+1}-t_{n-1}}{2}$. We assume that nearly all of the energy of f is located in the interval of interest. This assumption is in contrast to the first part of Section 3.2.3.2 in which it was not assumed that the interval of interest contains most of the signal energy. By suitable time-windowing that leaves the function bandlimited (albeit with a potentially different bandlimit), however, one can enforce this energy condition. By sampling only in the interval of interest, Eq. (3.27) will not be strictly fulfilled; we disregard this error from now on as it can be made arbitrarily small by increasing our interval. Using Theorem II.2 from [55] and Eq. (3.27), we now show that uniformly distributed sampling points satisfy the RIP with a large probability. Define the continuous linear map L that operates on the coefficients η of $f \in PW_{\pi w}$ in the PSWF basis and that returns m pointwise evaluations of f distributed uniformly in the interval $\left[-\frac{T}{2}, \frac{T}{2}\right]$, each sample being rescaled by $\sqrt{\omega_n}$. Let μ be the probability measure which leads to uniformly distributed rescaled sampling points in $\left[-\frac{T}{2}, \frac{T}{2}\right]$. Then $\mathbb{E}_\mu \|L(\eta)\|^2 = \omega \sum_{i=1}^{m} \left| f\left(-\frac{T}{2} + i\Lambda\right)\right|^2$ with $\Lambda = \frac{T}{m+1}$ and $\omega = \Lambda = \frac{1}{w'}$. Hence, if m is chosen such that the induced $w' \geq w$, it follows that $\mathbb{E}_\mu \|L(\eta)\|^2 = \|f\|^2$. As in [55], define $\delta_{S,\mu,2} = \sup_{x\in S} \left| \|L(\eta)\|^2 - \mathbb{E}_\mu \|L(\eta)\|^2 \right| = (2\delta w + \delta^2 w^2) \|f\|^2$ using Eq. (3.27), with S being the set of $2|\Delta|$-sparse coefficient vectors η with unit energy $\|f\|^2 = 1$. Since the RIP constant $\delta_{\mathrm{RIP}} \geq \delta_{S,\mu,2}$, it is necessary to oversample for a RIP constant below one to be feasible: $(2\delta w + \delta^2 w^2) < 1 \Leftrightarrow \delta < \frac{\sqrt{2}-1}{w}$. If, as above, $\Delta \subset \{1, \ldots, M\}$, then S has a finite upper box-counting dimension (cf. [55] for definitions). Additionally, for Theorem II.2 from [55] to work, the following probability must be bounded from above:

$$\mathbb{P}\Big\{ \Big| \|L(\eta_1)\|^2 - \mathbb{E}_\mu \|L(\eta_1)\|^2 - \|L(\eta_2)\|^2 + \mathbb{E}_\mu \|L(\eta_2)\|^2 \Big| \geq \\ \lambda \|\eta_1 - \eta_2\| \Big\}, \tag{3.28}$$

with η_1 and $\eta_2 \in S$ or zero in all its entries and $\lambda \geq 0$. For η_1 and η_2 either all zero or identical, a bound is trivial, i.e. statement (3.28) has probability one. For η_1 and η_2 distinct, we can use Hoeffding's inequality to obtain an upper bound. Using Eq. (3.27), we derive:

$$\mathbb{P}\left\{\left|\|L(\eta_1)\|^2 - \mathbb{E}_\mu\|L(\eta_1)\|^2 - \|L(\eta_2)\|^2 + \mathbb{E}_\mu\|L(\eta_2)\|^2\right| \geq \right.$$
$$\left.\lambda\|\eta_1 - \eta_2\|\right\} \leq 2\exp\left(-\frac{\lambda^2}{8\delta^2 w^2}\right). \tag{3.29}$$

Equation (3.29) is valid for all λ. It is worth emphasizing that Eq. (3.29) is only a correct statement insofar as Eq. (3.27) is true which necessitates oversampling. In case m points are chosen with uniform probability independently in an interval $\left[-\frac{T}{2}, \frac{T}{2}\right]$ and sorted in ascending order, the distribution for the greatest distance between any two consecutive points is given by $F_\delta(z) = \left(1 - \frac{(T-z)^m}{T^m}\right)^{m-1}$ with z ranging from zero to T. $F_{\delta^2}(z)$ is then given by $F_{\delta^2}(z) = \left(1 - \frac{\left(T-z^{\frac{1}{2}}\right)^m}{T^m}\right)^{m-1}$. Hence it follows that $\delta^2 < \Lambda$ (assuming $\delta < 1$) with overwhelming probability for even slight oversampling (factor two); Eq. (3.29) can then be manipulated to yield:

$$\mathbb{P}\left\{\left|\|L(\eta_1)\|^2 - \mathbb{E}_\mu\|L(\eta_1)\|^2 - \|L(\eta_2)\|^2 + \mathbb{E}_\mu\|L(\eta_2)\|^2\right| \geq \right.$$
$$\left.\lambda\|\eta_1 - \eta_2\|\right\} \leq 2\exp\left(-\frac{\lambda^2 m}{8Tw^2}\right). \tag{3.30}$$

Given that the continuous linear map L has a finite upper box-counting dimension and that inequality (3.30) holds, we conclude by invoking Theorem II.2 from [55]: The RIP holds with probability $1-\xi$, i.e. $\delta_{S,\mu,2} \leq \delta_{\text{RIP}}$ for any $\xi, \delta_{\text{RIP}} \in (0,1)$ if

$$m \geq \frac{8CTw^2}{\delta_{\text{RIP}}^2} max\left\{(2|\Delta|+1)\log\left(\frac{1}{\epsilon_S}\right), \log\left(\frac{6}{\xi}\right)\right\}, \tag{3.31}$$

with $C > 0$ a constant independent of all other parameters and ϵ_S given in [55]. The bound shown in Eq. (3.31) is structurally similar to the one derived in Eq. (3.24).

3.3 IMPLEMENTATION AND RESULTS

The PSWF are generated by using a freely available numerical software package [56]. Since the generation of PSWF with a large parameter c is numerically difficult, we have restricted our practical investigation to $c = 30$.

To test the algorithm, a speech signal is bandpass-filtered and the resulting bandpass-filtered signal is represented in its equivalent baseband form. This equivalent baseband form is then sampled uniformly in an interval. An illustrative example is shown in Fig. 3.1. In this example, T is set to 2, while $c = 30$. 90 sampling points are uniformly distributed within the interval $\left[-\frac{T}{2}, \frac{T}{2}\right]$. The approximation via weighted l_1-minimization in the interval is essentially perfect. It is worth pointing out that the actual number of measurements needed is smaller than the number required by theory. The upper bound discussed in Section 3.2.3.2 for the normalized PSWF element i (proportional to $\sqrt{i}$) seems to be a conservative estimate in practice. The homotopy method is used in our implementation for the minimization of the CS problem. In the case of noisy measurements, a weighted l_1-minimization is able to recover a solution with an error proportional to the noise term. Fig. 3.2 depicts the same signal as Fig. 3.1; contrary to the case in Fig. 3.1, however, uniform i.i.d. noise (SNR ≈ 4) is added to the perfect sampling values.

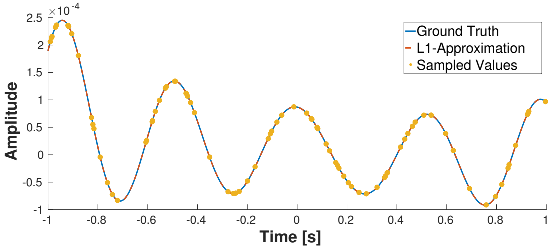

FIGURE 3.1: Example of a reconstructed equivalent baseband signal (real part): M = 30; 90 uniformly distributed sampling points; $c = 30$ © 2018 IEEE

3.4 CONCLUSION

We have discussed and derived methods for approximating bandlimited functions on the real line or in finite intervals from finitely many samples generated by a uniformly distributed sampling process, assuming sparsity in the PSWF basis. In the case that nearly all of the signal energy is concentrated in the interval of interest, the number of sampling points necessary is proportional to the sparsity in the PSWF basis. In the case that a significant portion of the signal energy is outside the interval of interest, we have derived results showing that a lower bound for the number of sampling

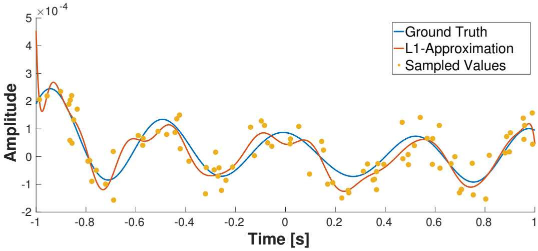

Figure 3.2: Example of a reconstructed equivalent baseband signal (real part): M = 30; 90 uniformly distributed sampling points; $c = 30$; SNR ≈ 4, i.i.d. uniformly distributed noise on sampling values © 2018 IEEE

points sufficient is proportional to M^2, assuming all coefficients up to M are to be recovered. Future work includes investigating the reason for this qualitative change in behavior when moving from the real line to finite intervals.

4

FILTERING OF NONUNIFORMLY SAMPLED BANDLIMITED FUNCTIONS

Most signals in the natural world are continuous-time in nature. An operation often applied to such continuous-time functions f is continuous-time convolution with some filter g:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau. \tag{4.1}$$

Due to the ubiquity of digital computers, it is desirable to find methods that enable the realization of Eq. (4.1) on such devices. A mapping from the space in which the continuous-time functions and filters reside to a sequence space is hence required. For the two operations in the two domains to be equivalent, the mapping should constitute an isomorphism. In principle arbitrary many possible mappings exist. A useful restriction placed on mappings is to demand that the time-invariant nature of filtering in continuous-time is preserved for the mapped discrete sequence, i.e. that the discrete filtering is shift-invariant. A classical example for a mapping with such a property is the uniform sampling of square-integrable bandlimited functions at an appropriate sampling rate: in this case, time-invariant continuous-time filters are mapped onto shift-invariant discrete-time filters. Uniform sampling is not the only mapping scheme with such a desirable property. It was recognized early on that a suitably modified version of the bilinear transform (which is often used in filter design to map continuous-time onto discrete-time filters and vice versa) preserves the convolution property as well [58]. The possible set of functions is furthermore larger than for uniform sampling as the continuous-time function only needs to be square-integrable and not bandlimited any more. Necessary and sufficient conditions were then given to ensure that the isomorphisms linking the continuous-time to the discrete space preserve the convolution property, thereby in principle allowing arbitrarily many different mappings [59]. A review by Oppenheim and Johnson discusses and summarizes the conditions that mappings need to fulfill for convolution to be preserved [60].

[3] Published in [57] © 2019 IEEE

In the case of dense enough irregular sampling of bandlimited square-integrable functions, the mapping from the continuous-time space to the sequence space is isomorphic as well. Instead of orthonormal bases, however, Riesz bases and frames are now required. Shift-invariant linear filtering, i.e. convolution, cannot be carried out any more directly on the irregular samples. Different algorithms are needed for the filtering on the sampled values. In the rest of this chapter, a possible approach is discussed that enables filtering. The scheme is iterative in nature, and each iteration is composed of discrete convolutions. The cumulative effect of the iterations is an approximation of Eq. (4.1) on the set of sampling times.

4.1 INTRODUCTION

The filtering of irregularly sampled bandlimited functions has been studied far less than the uniform counterpart, even though nonuniform sampling patterns arise often in practice, either as defects in uniform sampling or deliberately, such as in MRI (to accelerate measurements, for example) and seismology [61]. In the uniform sampling case, a wide body of literature describing the design and analysis of digital filters is available [62]. Irregular sampling of functions (albeit without the assumption of bandlimitedness) has recently become popular within the area of event-based signal processing, but has been mostly addressed from a hardware perspective [63]. The IIR filtering of functions sampled with a level-crossing scheme has been discussed as well [64]. It is an interesting problem to study under which conditions on the filters filtering can be performed on irregularly sampled bandlimited functions and which conditions the irregular sampling pattern has to fulfill for such filtering to be feasible. In the uniform sampling case, most attention has been focused on the design and analysis of linear time-invariant resp. linear shift-invariant systems (LTI resp. LSI), for those systems allow the filtering of a function with a time-invariant filter via convolutions. If the sampling pattern is nonuniform, however, no such time-invariant filter exists. Instead, filters will depend on the specific structure of the sampling pattern and will hence have to change over time.

Continuous convolution is mapped to discrete convolution for uniformly sampled bandlimited functions. Continuous filtering can hence be fully carried out on a discrete representation of the uniformly sampled bandlimited function. The mapping of continuous convolution onto operations performed on irregularly sampled bandlimited functions is feasible as well [65, 66]. The specific structure of this mapping is described in Section 4.2. The operations on the sampled values as described in these references, however, do not correspond to a discrete convolution. Since all operations, however, only require the sampled values, a filtering operation is performed that is in some sense equivalent to classical discrete-time filtering.

4.1.1 *Previous Work*

Previous work on the filtering of nonuniformly sampled functions focused either on the design of time-varying finite impulse response (FIR) filters via the minimization of some cost term [67] or on filtering via Projections onto Convex Sets (POCS) type algorithms [68]. For both approaches, no analysis of sampling pattern requirements is given to ensure equivalence of numerical filtering to continuous-time filtering. The first approach additionally requires the estimation of the Fourier transform of the function to be filtered. In the second approach it is not possible to use arbitrary filter functions due to the nature of the POCS algorithm employed.

The algorithm discussed in this chapter enables the filtering of bandlimited functions with bandlimited filters to a desired precision, in case effects such as truncation, aliasing and time jitter errors are not taken into account. All operations are furthermore carried out at the sampling instants, i.e. no uniform grid is required as in [68]. Contrary to [67], the filtering is in principle exact (disregarding error terms found in the filtering of uniform sampling as well) as it does not require the minimization of some surrogate cost function.

4.1.2 *Notation*

The version of the Fourier transform used is $\hat{f}(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t) e^{-iut} dt$. All considered functions live in the space $L^1(\mathbb{R})$, and are additionally bandlimited: $\operatorname{supp} \hat{f} \subseteq Q$, where supp denotes the support of a function and Q is any bounded set in $\mathbb{R}$. A sampling pattern $(t_n)_{n\in\mathbb{Z}}$ is associated with partitions of unity $\Psi = (\psi_n)_{n\in\mathbb{Z}}$. Characteristic functions on Voronoi regions are an example for a partition of unity. A sampling pattern is γ-dense, $\gamma > 0$, if $\cup_{n\in\mathbb{Z}}\, t_n + [-\gamma, \gamma] = \mathbb{R}$. δ_{t_n} denotes a time-shifted Dirac delta function.

4.2 FILTERING

In Section 4.2.1, the mapping of continuous-time filtering onto operations carried out using only irregularly sampled function values is briefly described. Thereafter the filtering algorithm based on such an approximation is described in Section 4.2.2.

4.2.1 *Approximation of Convolutions*

Assume that f is a bandlimited function in $L^1(\mathbb{R})$ with $\operatorname{supp} \hat{f} \subseteq Q$ and define the operator $D_\Psi^+(f)$ acting on f as follows:

$$D_\Psi^+(f) = \sum_{n\in\mathbb{Z}} f(t_n) \|\psi_n\|_1 \delta_{t_n}. \tag{4.2}$$

The partition of unity Ψ reweights the sampling values $(f(t_n))_{n\in\mathbb{Z}}$ to account for local irregularities in the sampling pattern. Given the operator Eq. (4.2), the following approximation to a convolution $C_g f := f * g$ can be defined, where g is a bandlimited function in $L^1(\mathbb{R})$:

$$A_g f := \left(D_\Psi^+(f)\right) * g. \tag{4.3}$$

The function f is to be convolved with a bandlimited function g with $\operatorname{supp} \hat{g} \subseteq Q$. If a function h with the properties $\hat{h} \equiv 1$ on Q and $\operatorname{supp} \hat{h} \subseteq Q_0$ (with $Q \subseteq Q_0$ and Q_0 a bounded set) is introduced, then the following equivalence can be shown (Proposition 6.1 in [66]), with $C_h f := f * h$:

$$C_g f = C_g A_h \left(\sum_{k=0}^{\infty} (C_h - A_h)^k \right) f. \tag{4.4}$$

Expanding the expression $\left(\sum_{k=0}^{\infty} (C_h - A_h)^k\right) f$ in Eq. (4.4), the following iterative procedure is obtained:

$$f_0 := f \tag{4.5}$$

$$f_{k+1} := f_k * h - D_\Psi^+(f_k) * h. \tag{4.6}$$

Eq. (4.6) then directly implies that

$$C_g f_k = C_g C_h f_k = C_g D_\Psi^+(f_k) + C_g f_{k+1}, \quad k \geq 0, \tag{4.7}$$

yielding by induction

$$f * g = \sum_{k=0}^{m} \left(D_\Psi^+(f_k) * g\right) + f_{m+1} * g, \quad m \in \mathbb{N}. \tag{4.8}$$

A requirement for the validity of Eq. (4.4) is that $C_h - A_h$ is a contraction on the space of bandlimited functions with support on Ω_0. $C_h - A_h$ is a contraction if the sampling density is sufficiently high. For lowpass functions h, for example, that have frequency support $\operatorname{supp} \hat{h} \subseteq [-\pi w, \pi w]$ with $w > 0$, the sampling pattern must necessarily be γ-dense with γ smaller than $1/w$ [69]. If $C_h - A_h$ is a contraction, then the terms f_{m+1} in Eq. (4.8) decay at a geometric rate. The specific decay properties depend on the sampling pattern. For regular equidistant sampling of lowpass functions, for example, Theorem 6.11 in [21] shows that $C_g f = C_g A_h f$ for f, g and h as above, hence one iteration is sufficient. Irregular sampling patterns require in general more iterations. The results shown in Section 4.3 indicate that a small value of m is often sufficient for good numerical performance. Since Eqs. (4.5) and (4.6) contain only linear operators, an error analysis for practical implementation errors (e.g. due to truncation, aliasing, and time jitter) can be carried out (cf. a related analysis for the case of reconstruction [70]). If these error types are present, then Eq. (4.8), truncated to the first m terms, will only result in an approximate convolution, even if $m \to \infty$.

4.2.2 *Filtering Algorithm*

The algorithm given in Section 4.2.1 can be directly implemented on the sampling values. Given the definition of the Fourier transform used in this letter and the requirements on h described in Section 4.2.1, the convolution of f_k and h in Eq. (4.6) results in $\sqrt{2\pi} f_k$, while $D_{\Psi}^{+}(f_k) * h$ reduces to $\sum_{n\in\mathbb{Z}} f_k(t_n) \|\psi_n\|_1 h(t - t_n)$. The k-th term f_k is then given by

$$\begin{aligned} f_k = (2\pi)^{\frac{k}{2}} f &- \sum_{n\in\mathbb{Z}} f(t_n) \|\psi_n\|_1 h^{k*}(t - t_n) \\ &- \sum_{n\in\mathbb{Z}} f_1(t_n) \|\psi_n\|_1 h^{(k-1)*}(t - t_n) - \cdots \\ &- \sum_{n\in\mathbb{Z}} f_{k-1}(t_n) \|\psi_n\|_1 h(t - t_n), \end{aligned} \tag{4.9}$$

with h^{l*} defined as

$$h^{l*} := \underbrace{h * h * \cdots * h}_{l \text{ times}}. \tag{4.10}$$

Eq. (4.9) can be directly evaluated on the sampling pattern $(t_n)_{n\in\mathbb{Z}}$ for each k. Then Eq. (4.8), if the correction term $f_{m+1} * g$ is discarded, can be evaluated on $(t_n)_{n\in\mathbb{Z}}$ as well, yielding an approximate convolution, or, in case of uniform sampling, perfect convolution in the case where practical implementation errors are not present. The partition of unity Ψ used in Eqs. (4.9) and (4.8) can be chosen in such a way as to be numerically beneficial.

Assume now that $Q = [-\pi w, \pi w]$ with $w > 0$. h can then be chosen as

$$\hat{h_\lambda}(u) = \begin{cases} 1, & |u| \leq \pi w \\ 1 - \frac{|u| - \pi w}{\lambda \pi w}, & \pi w \leq |u| \leq (1+\lambda)\pi w \\ 0, & |u| \geq (1+\lambda)\pi w, \end{cases} \tag{4.11}$$

with $\lambda > 0$. The larger λ is, the faster the decay of h_λ. A large λ therefore implies that in the evaluation of Eq. (4.6) at sampling time t_j, only nearby sampling times matter as $h_\lambda\left(t_j - t_n\right)$ decays quickly in t_n. With increasing λ, the filtering becomes more local in nature. The downside of a larger λ is the requirement of a denser sampling pattern. Such a statement therefore directly echoes similar statements for uniform sampling in case of oversampling. Higher iteration orders, however, require a larger neighborhood of sampling time t_j as the filters h^{l*} tend to broaden with increasing l. If Q was shifted to $[u_0 - \pi w, u_0 + \pi w]$ with $u_0 \in \mathbb{R}$ and u_0 known, then the algorithm remains unchanged, provided h and g are suitably modulated in frequency space.

If h, g and f are chosen such that supp $\hat{g} \subseteq$ supp $\hat{f} \subseteq$ supp $\hat{h}$, then the required sampling density is determined by h. For defining an approximation of convolutions on irregularly sampled functions, however, only supp $\hat{g} \subseteq$ supp $\hat{h}$ is required. The requirement of supp $\hat{f} \subseteq$ supp $\hat{h}$ is necessary such that Eqs. (4.5) and (4.6) can be evaluated directly on the sampling values without knowing f itself as $\hat{h} \equiv 1$ on Q results in all sampling values just being rescaled upon convolution. The downside of such a choice of h is that the sampling pattern is not tailored anymore to the filter g itself. To understand this point, it is worthwhile to consider the case of uniform sampling. For uniform sampling, it is possible to undersample a bandlimited function and still filter it perfectly if the frequency support of the filter is contained only within the nonoverlapping regions of the spectrum of the undersampled function. Sampling at or above the Nyquist rate is therefore not always necessary for filtering a bandlimited function. An extrapolation of such a statement to filtering performed on irregularly sampled functions is not possible within the framework discussed in this letter as supp $\hat{f} \subseteq$ supp $\hat{h}$ is required. γ must therefore necessarily be smaller than $1/(1+\lambda)\,w$ if $Q = [-\pi w, \pi w]$ and h is chosen according to Eq. (4.11) [69].

4.3 NUMERICAL EXAMPLES

To illustrate the algorithm, two differently sampled bandlimited functions are lowpass filtered. The first example involves a uniformly sampled bandlimited function with missing samples, while the second example is based on a bandlimited function that is sampled by an event-based sampling mechanism. In both cases, time-domain forms of Eq. (4.11) as well as of h_λ convolved with itself are required. These forms are given in Eqs. (4.12)-(4.13), with $h_\lambda(t) = h_\lambda^{1*}(t)$, for $1 \leq l \leq 2$.

$$h_\lambda^{1*}(t) = \frac{2\sqrt{2}\sin(\pi w t(1+\lambda/2))\sin(\pi w t\lambda/2)}{\pi^{3/2}\lambda w t^2} \tag{4.12}$$

$$h_\lambda^{2*}(t) = \frac{2\sqrt{2}\left(\pi w\lambda t\cos(\pi w t)+\sin(\pi w t)-\sin(\pi w t(1+\lambda))\right)}{\pi^{5/2}t^3\lambda^2 w^2} \tag{4.13}$$

Closed-form expressions of h_λ^{l*} for the case of $l > 2$ can be easily derived as well. The first lowpass filter used is bandlimited (therefore not causal) and given by

$$\hat{g}_1(u) = \begin{cases} \dfrac{e^{-\pi^2 w_f^2} - e^{u^2-2\pi^2 w_f^2}}{e^{u^2-\pi^2 w_f^2}}, & |u| < \pi w_f \\ 0, & |u| \geq \pi w_f \end{cases} \tag{4.14}$$

and

$$g_1(t) = \sqrt{\frac{2}{\pi}}e^{-\pi^2 w_f^2}\Im\left\{e^{i\pi w_f t}D_+\left(\frac{t}{2}+i\pi w_f\right)\right\} - \sqrt{\frac{2}{\pi}}e^{-\pi^2 w_f^2}\frac{\sin\left(\pi w_f t\right)}{t}, \tag{4.15}$$

where i is the imaginary unit, $\Im\{\cdot\}$ the imaginary part of its argument, w_f the cutoff frequency of the filter and D_+ the Dawson function. The second lowpass filter used is a first order Butterworth filter. This filter is causal but not bandlimited. Its use therefore results in an unavoidable aliasing error. Its time-domain form is given by

$$g_2(t) = \pi v e^{-\pi v t}u(t), \tag{4.16}$$

where u is the step function and $v > 0$ determines the cutoff frequency. Eq. (4.15) is additionally chosen as the first bandlimited function to be filtered, i.e. $\forall t f_1(t) = g_1(t)$ with $w_f = w$. The ground truth filtered signal is then readily obtainable if f_1 is convolved with g_1:

$$\begin{aligned}(f_1 * g_1)(t) = {} & \sqrt{\frac{2}{\pi}} e^{-\pi^2\left(w_f^2+w^2\right)} \frac{\sin\left(\pi w_f t\right)}{t} - \\ & \sqrt{\frac{2}{\pi}} \Im\left\{ e^{i\pi w_f t - 2\pi^2 w_f^2} D_+\left(\frac{t}{2} + i\pi w_f\right)\right\} - \\ & \sqrt{\frac{2}{\pi}} \Im\left\{ e^{i\pi w_f t - \pi^2\left(w_f^2+w^2\right)} D_+\left(\frac{t}{2} + i\pi w_f\right)\right\} + \\ & \frac{1}{\sqrt{\pi}} \Im\left\{ e^{i\pi w_f t - 2\pi^2 w_f^2} D_+\left(\frac{1}{\sqrt{2}}\left(\frac{t}{2} + 2i\pi w_f\right)\right)\right\}. \end{aligned} \tag{4.17}$$

The second bandlimited function to be filtered is given by

$$\hat{f}_2(u) = \begin{cases} 1, & |u| \leq \pi \frac{w}{1+\mu} \\ 1 - \frac{(1+\mu)|u| - \pi w}{\mu \pi w}, & \pi \frac{w}{1+\mu} \leq |u| \leq \pi w \\ 0, & |u| \geq \pi w, \end{cases} \tag{4.18}$$

with $\mu > 0$. The filtered function $(f_2 * g_2)(t)$ is then obtained as

$$\begin{aligned}(f_2 * g_2)(t) = {} & \frac{1}{2\mu\sqrt{2\pi}wt} e^{-\pi v t} v u(t) \Bigg(2\mu\pi^2 w t + \\ & 4 e^{\pi v t}(1+\mu)\left(\cos(\pi w t) - \cos\left(\frac{\pi w t}{1+\mu}\right)\right) + \\ & 4\pi t (1+\mu) \Re\left\{(w+iv)\operatorname{Si}(\pi t (w+iv))\right\} + \\ & 4\pi t (1+\mu) \Re\left\{i(w+iv)\operatorname{Ci}(\pi t (w+iv))\right\} + \\ & 4\pi t \Re\left\{ i(w + i(v+\mu v))^* \operatorname{Ci}^*\left(\pi t\left(\frac{w}{1+\mu} + iv\right)\right)\right\} - \\ & 4\pi t \Re\left\{ (w + i(v+\mu v)) \operatorname{Si}\left(\pi t\left(\frac{w}{1+\mu} + iv\right)\right)\right\}\Bigg) \end{aligned} \tag{4.19}$$

with $\Re(\cdot)$ the real part of its argument, Ci the cosine integral, Si the sine integral and $(\cdot)^*$ the complex conjugate of its argument.

For the numerical filtering examples, the following parameter choices are made: f_1 and f_2 have a cutoff frequency of $\pi w = 1$, i.e. $w = 1$, $w_f = 0.5$ is used in g_1, and $v = 0.01$ and $v = 0.5$ are chosen for g_2. These two choices of v lead to different amounts of aliasing error in the eventual filtering. For f_2, $\mu = 1$ is chosen. For the auxiliary function Eq. (4.11), $\lambda = 1$ is taken. Given the choices of w and λ, a maximum distance of 0.5 between consecutive samples is then required for the filtering to be feasible. This requirement on the sampling pattern, however, is only sufficient for $f_1 * g_1$. Since g_2 is not bandlimited, this maximum distance between consecutive samples is not sufficient if the filtering of $f_2 * g_2$ is be carried out on the irregular samples - an aliasing error is unavoidable.

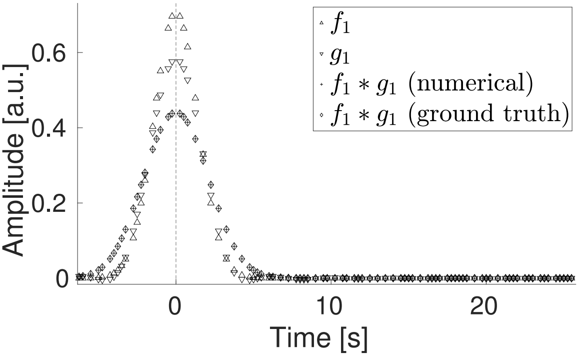

FIGURE 4.1: Result of $f_1 * g_1$; f_1 and g_1 as well as parameter choices are given in Section 4.3. The ground truth filtered function $f_1 * g_1$ is on top of the numerically computed filtered function. © 2019 IEEE

The first nonuniform sampling pattern considered is obtained by dropping samples from a uniform sampling pattern (with a distance between consecutive samples of 0.25 before removal of samples). The sample deletion is done such that the maximum distance between consecutive samples of the resulting irregular sampling pattern is 0.5, i.e. if a sample is dropped, the next one will certainly be kept. Three iterations of Eq. (4.6) are used in all numerical examples. Examples of filtering functions sampled with such a sampling pattern are shown in Figs. 4.1, 4.2 and 4.3.

The second nonuniform sampling pattern considered is obtained by the application of the Send-on-Delta sampling scheme on a bandlimited function [16]. If the function was sampled at t_j, then a new sample would be generated at the first time point t_{j+1} for which the following condition holds: $|f(t_{j+1}) - f(t_j)| = \delta$, with $\delta > 0$. An example of filtering a bandlimited function sampled by Send-on-Delta is shown in Fig. 4.4.

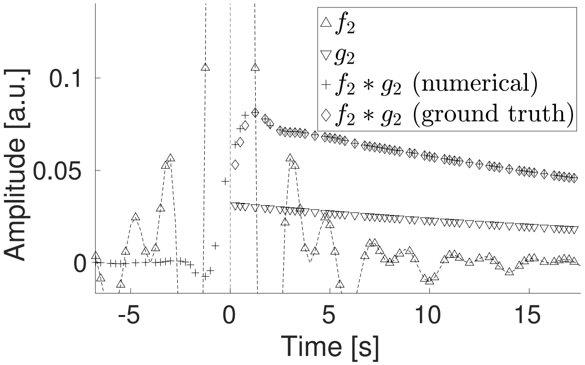

FIGURE 4.2: $f_2 * g_2$; f_2 and g_2 are given in Section 4.3; $v = 0.01$. © 2019 IEEE

4.4 DISCUSSION

The examples from Section 4.3 show that - provided both the function to be filtered and the filter are bandlimited - the numerically computed solution is close to ground truth. Fig. 4.3 as compared to Fig. 4.2 illustrates the impact of aliasing on the numerical filtering. The presence of aliasing is unavoidable as the Butterworth filter used is not bandlimited. To avoid aliasing (at the cost of distorting the filter, however), it would be necessary to warp the filter in the frequency domain, as is commonly done in uniform discrete-time filtering via the bilinear transform. Fig. 4.4 shows that in regions of sufficient sampling density, the filtering is better than in regions of insufficient density (left and right parts of Fig. 4.4).

Even in the case when the filter g is chosen to be causal, the overall algorithm is noncausal as h is noncausal. If it is desired to filter a function at time point t_j, then nearby sampling times t_n need to be kept for which $h_\lambda\left(t_j - t_n\right)$ is nonzero. With increasing iteration number, the auxiliary functions tend to get broader, implying that the amount of buffering required depends strongly on the total number of iterations.

The filtering algorithm discussed in this chapter is simple to implement and flexible. As all operations are carried out directly on the nonuniform samples themselves, the discrete-time equivalent of continuous convolution is achieved, even though the algorithm itself does not correspond to a convolution operation. Future investigations could target hardware implementations to enable real-time filtering of nonuniformly sampled functions.

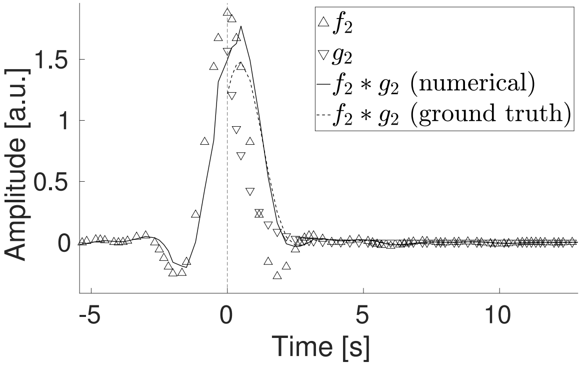

FIGURE 4.3: $f_2 * g_2$; f_2 and g_2 are given in Section 4.3; $v = 0.5$. © 2019 IEEE

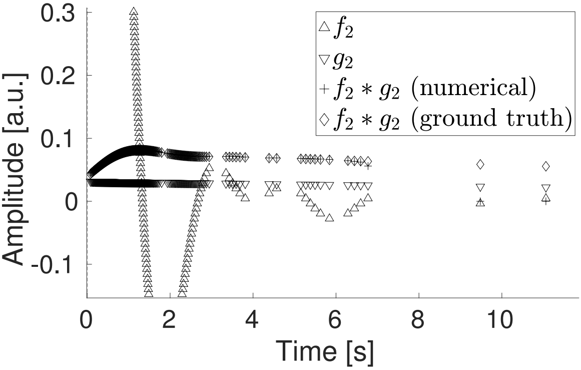

FIGURE 4.4: $f_2 * g_2$; f_2 and g_2 are given in Section 4.3. Send-on-Delta with $\delta = 0.008$ and $v = 0.01$. The numerical result is close to the ground truth - a significant difference can only be seen at the last samples. © 2019 IEEE

5 OPTIMAL SAMPLING OF PARAMETRIC FAMILIES

Inference from data is an old problem in statistics. Classical approaches following Fisher's work deal with the parametric case: it is assumed that a parametric family is known which contains the data-generating distribution; the task of inference then consists in estimating an optimal set of parameters via the maximum likelihood method from the data in order to identify the data-generating distribution or a distribution close to it. For this approach to be successful, is is necessary that the parametric family is not too wide and that a large number of samples is given. A different line of work dealing with the nonparametric case is connected to Kolmogorov, Glivenko and Cantelli: Glivenko and Cantelli showed that for any probability distribution function of a random variable ξ, $F(z) = P\{\xi < z\}$, the empirical distribution function $F_l(z) = \frac{1}{l}\sum_{i=1}^{l}\theta(z - z_i)$ (with $\theta(u) = 1$ if $u \geq 0$ and $\theta(u) = 0$ for $u < 0$ and $z_1, \ldots, z_l$ i.i.d. according to the unknown distribution function F) converges in probability to $F(z)$, i.e. for any $\epsilon > 0$ [71]

$$\lim_{l\to\infty} P\left\{\sup_z |F(z) - F_l(z)| > \epsilon\right\} = 0. \tag{5.1}$$

The rate of convergence was given by Kolmogorov. The approach pioneered by this latter approach formed the starting point for the development of statistical learning theory. Here we briefly sketch the main elements of this theory as it forms the basis for understanding the generalization behavior of supervised machine learning methods. The discussion follows standard expositions of the field [71, 72].

Assume that some generator G is given that independently draws data $x \in \mathbb{R}^n$ from a fixed but unknown probability distribution function F, as well as some supervisor S returning an output y for every input x according to a distribution function $F(y|x)$. Additionally, a learning machine is given that can implement a set of functions $f(x,\alpha)$, $\alpha \in \Lambda$, where Λ is some set of parameters. Given i.i.d. data $(x_1, y_1), \ldots, (x_l, y_l)$, the goal of learning consists in minimizing the risk functional

$$R(\alpha) = \int L(y, f(x,\alpha))\, dF(x,y), \tag{5.2}$$

[4] Accepted for publication by Neural Computation

with Eq. (5.2) being minimized for some $f(x,\alpha_0)$ and $L(\cdot,\cdot)$ being some loss function. $(x_1,y_1),\ldots,(x_l,y_l)$ is the only information given about the unknown F. Instead of using data (x_i,y_i), the risk functional Eq. (5.2) can be written more abstractly as

$$R(\alpha) = \int Q(z,\alpha)\, dF(z), \quad \alpha \in \Lambda, \tag{5.3}$$

for F defined on some space Z and the only information provided about F being the i.i.d. samples $z_1,\ldots,z_l$. In this form, risk minimization can be seen as encompassing many problems such as pattern recognition, regression and density estimation. Eq. (5.3) can be replaced by the empirical risk functional

$$R_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha), \tag{5.4}$$

which can be minimized over $Q(z,\alpha)$, $\alpha \in \Lambda$. Assume that the minimum of Eq. (5.4) for l samples z_i is given by $Q(z,\alpha_l)$. Statistical learning theory deals primarily with the two following questions: what are the conditions placed on the learning machine $Q(z,\alpha)$, $\alpha \in \Lambda$, that guarantee that a minimization of Eq. (5.4) converges in probability to a minimum of Eq. (5.3) for $l \to \infty$, and what conditions can be placed on the function set of the learning machine to guarantee that the convergence is fast. The first demand is known as consistency of the learning method. It turns out that empirical risk minimization is consistent if the following condition holds [71]

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left(R(\alpha) - R_{\text{emp}}(\alpha)\right) > \epsilon\right\} = 0, \quad \forall\epsilon > 0. \tag{5.5}$$

Eq. (5.5) is similar in structure to Eq. (5.1). For consistency and fast convergence to hold simultaneously for all possible distribution functions F, a necessary and sufficient condition is that the learning machine $Q(z,\alpha)$, $\alpha \in \Lambda$, has finite VC dimension [71, 72]. The VC dimension intuitively measures the maximum number of data samples that can be shattered by the set $Q(z,\alpha)$, $\alpha \in \Lambda$. For pattern recognition problems, the VC dimension is given by the maximum number h of samples $z_1,\ldots,z_h$ which can be described in all 2^h possible ways by the learning machine. Generalizations of the concept exist for real-valued functions as well. Minimizing Eq. (5.4) for a particular l yields as solution some $Q(z,\alpha_l)$ which in general induces a generalization error in the following sense with probability at least $1-\eta$

$$R(\alpha_l) \leq R_{\text{emp}}(\alpha_l) + \Phi(h,\eta,l). \tag{5.6}$$

$\Phi(h, \eta, l)$ increases with growing VC dimension h and decreases in l. To minimize the generalization error, it is therefore advisable to choose a learning machine $Q(z, \alpha)$, $\alpha \in \Lambda$, with small VC dimension, while the learning machine should still be flexible enough to find a small empirical risk Eq. (5.4). These two goals are in general contradictory.

5.1 INTRODUCTION

The main problems in machine learning are density estimation, regression and classification based on samples drawn according to an unknown but fixed probability distribution function F. To assess the quality of a machine learner, the notion of generalization was introduced, most prominently in statistical learning theory [71, 72]. Statistical learning theory describes conditions on the hypothesis space of the learning algorithm and the number of samples drawn from F such that the empirical risk is close in probability to the expected risk. For generalization to be defined in this framework, it is crucial that the expected risk is calculated with respect to the same probability distribution function that generated the samples used for the evaluation of the empirical risk. A change in the probability distribution function cannot be directly incorporated into statistical learning theory.

Recent findings have shown, however, that even slight changes in the probability distribution function that generates the data (i.e. different distribution functions for the training/test set) lead to decreases in performance of the learned model [73]. This problem can be partially circumvented by including data drawn from different possible probability distribution functions (which are allowed to possess different functional forms) in the training set, effectively demanding that a joint solution is found for all subproblems [74]. In the limit, it is possible that infinitely many probability distribution functions could have generated the data. One possible way of modeling the infinitely many data-generating probability distribution functions is by grouping them into a parametric family.

In this chapter, we assume that the data-generating process is itself parametric. Data are then drawn from the whole parametric family: the task that a learning algorithm has to solve is to learn a model for the entire parametric family. Without further prior information on the specific probabilistic structure of the test set, it is a natural requirement to demand that a learned model is equally good for all members of the parametric family. The central question studied in this chapter is therefore how training sets containing a finite number of samples can be constructed such that the training set represents the entire parametric family optimally. The tools needed for the analysis carried out in this chapter mostly stem from information theory, specifically universal coding theory, and not from machine learning [75, 76].

For the sake of clarity and in order to derive quantitative statements, we focus on a specific stochastic process, the Ornstein-Uhlenbeck process. Being both a Gaussian and Markovian process, this stochastic process is rich in structure while still being analytically tractable. Most of the results presented in this chapter, however, will apply to more general problem classes.

As alluded to above, the problem of how to optimally sample from a parametric family is tightly connected to universal coding theory. Some universal coding inequalities described in Section 5.2 directly correspond to the problem of sequential prediction in the case of an Ornstein-Uhlenbeck process as shown in Section 5.3. The specific stochastic process chosen therefore yields a task (sequential prediction – having observed a time series up to sample n, sample $n+1$ is predicted) which directly corresponds to questions of how to sample a parametric family optimally in the sense of universal coding theory. The chapter concludes by empirically studying the generalization behavior shown by a deep network trained on the Ornstein-Uhlenbeck parametric family in an autoregressive manner. We empirically find that a simple model trained on optimally constructed training sets generalizes better to changes in the test set distribution than if the model is trained on suboptimally generated training sets.

NOTATION Let $x^n = (x_1, x_2, \cdots, x_n)$ be a sequence of real-valued elements and $X^n = (X_1, X_2, \cdots, X_n)$ a sequence of random variables on $\mathbb{R}^n$. In this work, X^n will denote strictly stationary stochastic processes. Define a set of probability density functions (PDF) $\{P_\lambda, \lambda \in \Omega\}$ on $\mathbb{R}^n$ with Ω a compact subset of $\mathbb{R}^m$, assuming there are m free parameters. $|\cdot|$ denotes the operation of taking the determinant of a square matrix. $\log(\cdot)$ is the natural logarithm.

5.2 REVIEW ON UNIVERSAL CODING

We give a brief description of ideas from the universal coding literature which are crucial for this work. Assume that a family of PDFs $\{P_\lambda, \lambda \in \Omega\}$ on $\mathbb{R}^n$ and an observed sequence $x^n = (x_1, x_2, \cdots, x_n)$ (which is generated by one of the densities in the family) is given. If the specific PDF P_λ generating x^n is known, then the entropy rate $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}_\lambda\left[-\log\left(P_\lambda\left(X^n\right)\right)\right] = H\left(\lambda\right)$, with $\mathbb{E}_\lambda\left[\cdot\right]$ the expectation with respect to P_λ, corresponds to the best compression of the source. Such a compression statement follows from the asymptotic equipartition property (AEP) [76]. For the sampled strictly stationary Ornstein-Uhlenbeck process which is discussed in Section 5.3 in more detail, the AEP holds [77]. If P_λ is not known, however, the question arises of whether it is still possible (asymptotically in n) to reach the entropy rate of the stochastic process, provided that the parametric family $\{P_\lambda, \lambda \in \Omega\}$ is known. Universal coding theory answers this question in the affirmative for a wide class of parametric families [78]. To show this, a mixture source $P\left(x^n\right) = \int_\Omega w\left(\lambda\right) \cdot P_\lambda\left(x^n\right) d\lambda$ is introduced, with w a PDF (we do not consider cases in which w might be discrete) on Ω. This mixture source can then be used as a replacement for the unknown P_λ. A natural question associated with such a mixture source is how w should be chosen. It is intuitively clear that mixture sources $P\left(x^n\right)$ set up by different w will behave differently. It turns out that a particular choice of w carries with it a notion of channel capacity. Let Λ denote a random variable with PDF w on Ω. The parameters λ indexing P_λ are realizations of Λ. The prior w^* which reaches channel capacity $C_n = \sup_w I_w\left(\Lambda; X^n\right)$ with channel input Λ and channel output X^n, where $I_w\left(\Lambda; X^n\right)$ denotes mutual information induced by $w\left(\lambda\right) P_\lambda\left(x^n\right)$, maximizes the mutual information between Λ and X^n. If Λ is distributed as w^*, then observations x^n generated by P_λ contain most information about the m parameters in Ω. Additionally, w^* has the further property of being the prior that induces maximin redundancy [78]. The channel capacity C_n is furthermore a lower bound on the Kullback-Leibler divergence between the true data generating distribution P_λ and any other PDF $Q\left(x^n\right)$ [79]:

$$D\left(P_\lambda || Q\right) > \left(1-\epsilon\right) C_n. \tag{5.7}$$

Inequality (5.7) holds for all $\epsilon > 0$ and for all $\lambda \in \Omega$ except for some λ in a subset $B \subset \Omega$ whose size under w^* vanishes at an exponential rate with C_n. For $w = w^*$, $D\left(P_\lambda || P^*\right) = C_n$, with P^* the mixture source with capacity-achieving prior w^*. Hence for w^* nearly all sources P_λ lie on or close to a hypersphere centered at P^* with Kullback-Leibler divergence equal to C_n, as can be inferred from the previous discussion and inequality (5.7). It is crucial to emphasize that this statement only holds for the capacity-achieving prior w^*. Other mixture sources based on different priors w will in general be closer to some subset of sources in the parametric family

$\{P_\lambda, \lambda \in \Omega\}$, and have larger Kullback-Leibler divergence than C_n to other sources in the parametric family.

It is interesting to note that for the parametric family introduced in Section 5.3 (sampled strictly stationary Ornstein-Uhlenbeck processes) an asymptotically accurate form of the channel capacity can be deduced [80]:

$$C_n = \frac{m}{2}\log\left(\frac{n}{2\pi}\right) + \log\int_\Omega \sqrt{|I(\lambda)|}d\lambda + o(1), \tag{5.8}$$

with $o(1)$ tending to zero for $n \to \infty$ and $I(\lambda)$ the Fisher information matrix of the stochastic process:

$$I_{ij}(\lambda_*) = \lim_{n\to\infty}\frac{1}{n}\left\{\frac{\partial^2}{\partial\lambda_i\partial\lambda_j}\mathbb{E}_{\lambda_*}\left[-\log P_\lambda(X^n)\right]\right\}_{\lambda_*}, \tag{5.9}$$

with i and j ranging from 1 to m and λ_* in Ω.

An additional interpretation of C_n can be given in terms of the number of distributions in $\{P_\lambda, \lambda \in \Omega\}$ that are distinguishable based on the observation of a sequence of length n [75, 81]. It is intuitively clear that different sources in the parametric family $\{P_\lambda, \lambda \in \Omega\}$ are not necessarily distinguishable after observing n samples. This notion can be made more precise by using the language of hypothesis testing. For the parametric family discussed in this paper, this analysis is described in Section 5.3. Note that Eq. (5.8) is a consequence of choosing Jeffreys' prior in the mixture source $P(x^n)$ which is given by the following expression [82]:

$$w_{\text{Jeffreys}}(\lambda) = \frac{\sqrt{|I(\lambda)|}}{\int_\Omega \sqrt{|I(\lambda')|}d\lambda'}, \tag{5.10}$$

which is asymptotically equal to the capacity-achieving prior w^* for the parametric family considered in this paper. The number of distinguishable distributions after observing a sequence of length n is roughly equal to e^{C_n}. Since Jeffreys' prior (Eq. (5.10)) is asymptotically capacity inducing, the maximal number of distinguishable distributions is reached for Jeffreys' prior. More precisely, if Λ is distributed according to w_{Jeffreys}, then the sampled stochastic processes P_λ are maximally distinguishable on average. Any other prior w would (at least asymptotically) lead to a smaller number of distinguishable distributions. This argument can be strengthened by appealing to the analogue of Eq. (5.7) for arbitrary priors [79]. It can be shown that $D(P_\lambda||Q)$ is larger than $(1-\epsilon)C_R$, with $\epsilon > 0$ and C_R equal to the logarithm of the maximal number of random sources chosen under the prior w that can be distinguished in the sense of having a bounded error probability [79]. Q is an arbitrary distribution on x^n as in Eq. (5.7). The inequality holds again for all parameters λ except in a set $B' \subset \Omega$ whose size measured by w tends to zero for $n \to \infty$ under certain conditions.

The previous ideas, although formulated in terms of probabilities (equivalently, in terms of log-loss) can be directly applied to the case of sequential prediction under the MSE loss, at least for the Gauss-Markov processes used in this chapter. This idea is described in Section 5.3.

5.3 LOWER BOUNDS ON THE SEQUENTIAL PREDICTION ERROR

In this section, we first introduce the parametric family which is studied in this chapter. Thereafter we derive lower bounds on the sequential prediction error under the MSE loss for different priors w from which the strictly stationary sampled Ornstein-Uhlenbeck processes are drawn.

5.3.1 *Some Results on the Ornstein-Uhlenbeck Process*

The Ornstein-Uhlenbeck process is defined as

$$dX_t = \theta \left(\mu - X_t\right) dt + \sigma dW_t, \tag{5.11}$$

with $\theta > 0$, $\mu \in \mathbb{R}$, $t \geq 0$, $\sigma > 0$ and W_t the standard Wiener process. For the process to be strictly stationary, the first value x_0 at time $t = 0$ is drawn from a Gaussian distribution with mean μ and variance $\frac{\sigma^2}{2\theta}$. In the strictly stationary case, the Ornstein-Uhlenbeck process can be alternatively written as follows:

$$X_t = \mu + \frac{\sigma}{\sqrt{2\theta}} e^{-\theta t} W_{e^{2\theta t}}, \tag{5.12}$$

with $\{W_{e^{2\theta t}}\}$ a time-scaled Wiener process. We next derive some bounds on the growth of strictly stationary Ornstein-Uhlenbeck processes. These bounds are needed in the explicit construction of the RNN that implements the asymptotically optimal solution of the sequential prediction problem described in Section 5.4.1. To understand the growth behavior of the strictly stationary Ornstein-Uhlenbeck process, the law of the iterated logarithm is invoked:

$$\limsup_{t \to \infty} \frac{|W_t|}{\sqrt{2t\log\left(\log\left(t\right)\right)}} = 1 \quad \text{a.s.} \tag{5.13}$$

By applying the law of the iterated logarithm to the time-scaled Wiener process, the denominator of Eq. (5.13) is changed to $\sqrt{2e^{2\theta t}\log\left(\log\left(e^{2\theta t}\right)\right)}$, while the numerator is replaced by $|W_{e^{2\theta t}}|$. Multiplying the denominator by $\frac{\sigma}{\sqrt{2\theta}} e^{-\theta t}$, one obtains $\frac{\sigma}{\sqrt{\theta}} \sqrt{\log(2\theta t)}$. Hence one can conclude the following about the second term of Eq. (5.12):

$$\limsup_{t\to\infty} \frac{\sigma}{\sqrt{2\theta}} e^{-\theta t} \left|W_{e^{2\theta t}}\right| = \frac{\sigma}{\sqrt{\theta}} \sqrt{\log(2\theta t)}. \tag{5.14}$$

For a finite $t > 0$, there will in general exist a constant $C > 0$ such that the strictly stationary Ornstein-Uhlenbeck process in $[0, t]$ will be a.s. contained within the interval

$$\left[\mu - C \frac{\sigma}{\sqrt{\theta}} \sqrt{\log(2\theta t)}, \mu + C \frac{\sigma}{\sqrt{\theta}} \sqrt{\log(2\theta t)}\right]. \tag{5.15}$$

5.3.2 *Sampling the Ornstein-Uhlenbeck Process*

We consider Ornstein-Uhlenbeck processes drawn from a parametric family. The two free parameters are $\mu \in (c, d)$ with $c, d \in \mathbb{R}$, $d > c$ and $\theta \in (a, b)$ with $a, b \in \mathbb{R}^+$, $b > a$. $\sigma \in \mathbb{R}^+$ is arbitrary but fixed. The uniformly sampled Ornstein-Uhlenbeck process amounts to an autoregressive AR(1)-process

$$X_{n\delta} = e^{-\theta\delta} X_{(n-1)\delta} + \mu\left(1 - e^{-\theta\delta}\right) + \epsilon_n, \tag{5.16}$$

with $\epsilon_n \sim N\left(0, \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta\delta}\right)\right)$ independent over time, $\delta > 0$ the distance between consecutive samples and $X_{n\delta}$ the n-th sample. $(X_\delta, X_{2\delta}, \ldots, X_{n\delta})^\top$ is distributed according to a multivariate normal distribution with mean vector $(\mu, \mu, \ldots, \mu)^\top$ and covariance matrix

$$\Sigma = \frac{\sigma^2}{2\theta} \begin{pmatrix} 1 & e^{-\theta\delta} & \cdots & e^{-\theta(n-1)\delta} \\ e^{-\theta\delta} & 1 & \cdots & e^{-\theta(n-2)\delta} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\theta(n-1)\delta} & e^{-\theta(n-2)\delta} & \cdots & 1 \end{pmatrix}. \tag{5.17}$$

We next derive the asymptotic Kullback-Leibler divergence between two strictly stationary Ornstein-Uhlenbeck processes as well as the Fisher information matrix of this stochastic process. Both are needed for the subsequent discussion of distinguishability as well as for the explicit construction of Jeffreys′ prior. The inverse of covariance matrix (5.17) is given by

$$\Sigma^{-1} = \frac{2\theta}{\sigma^2\left(1-e^{-2\theta\delta}\right)} \cdot \begin{pmatrix} 1 & -e^{-\theta\delta} & 0 & \cdots & 0 \\ -e^{-\theta\delta} & 1+e^{-2\theta\delta} & -e^{-\theta\delta} & \cdots & 0 \\ 0 & -e^{-\theta\delta} & 1+e^{-2\theta\delta} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -e^{-\theta\delta} \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \tag{5.18}$$

which is a symmetric tridiagonal matrix. From Eq. (5.18) we obtain the determinant of Σ

$$|\Sigma| = \frac{1}{|\Sigma^{-1}|} = \frac{\sigma^{2n}\left(1-e^{-2\theta\delta}\right)^{n-1}}{(2\theta)^n}. \tag{5.19}$$

The asymptotic Kullback-Leibler divergence is then equal to

$$\begin{aligned} & D\left(\mu_1, \theta_1 || \mu_0, \theta_0\right) = \lim_{n\to\infty} \frac{1}{n} D\left(P_{(\mu_1,\theta_1)} || P_{(\mu_0,\theta_0)}\right) = \\ & \frac{1}{2}\frac{\theta_0}{\theta_1}\frac{1}{1-e^{-2\theta_0\delta}}\left(1-2e^{-(\theta_0+\theta_1)\delta}+e^{-2\theta_0\delta}\right)+ \\ & (\mu_1-\mu_0)^2\frac{\theta_0}{\sigma^2\left(1-e^{-2\theta_0\delta}\right)}\left(1-e^{-\theta_0\delta}\right)^2 - \\ & \frac{1}{2}+\frac{1}{2}\cdot\log\left(\frac{1-e^{-2\theta_0\delta}}{1-e^{-2\theta_1\delta}}\right)+\frac{1}{2}\cdot\log\left(\frac{\theta_1}{\theta_0}\right). \end{aligned} \tag{5.20}$$

Evaluating the Fisher information matrix Eq. (5.9) for the strictly stationary sampled Ornstein-Uhlenbeck process, we find

$$I\left(\mu_*, \theta_*\right) = \begin{pmatrix} \frac{\left(\left(e^{2\theta_*\delta}-1\right)-2\theta_*\delta\right)^2}{2\theta_*^2\left(e^{2\theta_*\delta}-1\right)^2}+\delta^2\frac{1}{e^{2\theta_*\delta}-1} & 0 \\ 0 & \frac{2\theta_*}{\sigma^2}\frac{e^{\theta_*\delta}-1}{e^{\theta_*\delta}+1} \end{pmatrix}, \tag{5.21}$$

where we first differentiate with respect to θ and then with respect to μ. Eq. (5.20) can be locally approximated as follows:

$$D\left(\mu_1,\theta_1||\mu_0,\theta_0\right) \approx \\ \frac{1}{2}\begin{pmatrix}\theta_1-\theta_0 & \mu_1-\mu_0\end{pmatrix} I\left(\theta_0,\mu_0\right)\begin{pmatrix}\theta_1-\theta_0\\ \mu_1-\mu_0\end{pmatrix}. \tag{5.22}$$

Eq. (5.22) is a quadratic approximation to Eq. (5.20), i.e. it corresponds to a Taylor expansion truncated after the second expansion coefficient. Eq. (5.21), plugged into Eq. (5.10), yields Jeffreys' prior for the parametric family composed of sampled Ornstein-Uhlenbeck processes. Jeffreys' prior is visualized in Fig. 5.1 for $\delta = 10$.

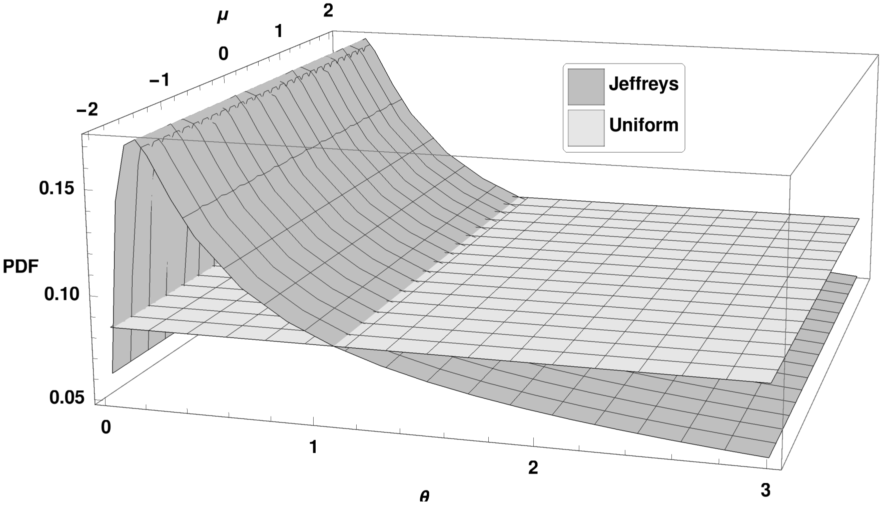

FIGURE 5.1: Jeffreys' and uniform prior for the Ornstein-Uhlenbeck process

5.3.3 *Lower Bounds*

In Section 5.2, various lower bounds under log-loss were discussed that pertain to representing a parametric family by some mixture source. Here we discuss lower bounds under MSE loss for the task of sequential prediction tailored to the sampled Ornstein-Uhlenbeck process.

Theorem 1. Consider Gaussian ARMA processes with compact parameter space $\Omega \subset \mathbb{R}^m$, $m > 0$, and p autoregressive terms, $p < m$. Given any prior w on Ω with corresponding random coding capacity C_R and any $\epsilon > 0$, the following lower bound is valid for all parameters λ except in a set $B' \subset \Omega$ whose size measured by w tends to zero for $n \to \infty$:

$$\frac{1}{n-p}\mathbb{E}_\lambda\left[\sum_{i=p+1}^{n}\left(X_i - h_i\left(X^{i-1}\right)\right)^2\right] \geq \sigma^2\left(\lambda\right)\left[1 + (1-\epsilon)\frac{2C_R}{n-p}\right], \tag{5.23}$$

with $\sigma^2\left(\lambda\right)$ the variance of the stationary Wold decomposition of the stochastic process and $\hat{x}_i = h_i\left(x^{i-1}\right)$ any measurable prediction function.

Proof. The random coding theorem [79] holds for Gaussian ARMA processes. In case P_λ and Q from Eq. (5.7) as well as its extension to the random coding case are both Gaussian distributions, the random coding theorem leads directly to a lower bound on the MSE loss. $P_\lambda\left(x^n\right)$ is the probability of data sequence x^n induced by the Gaussian ARMA model, while $Q\left(x^n\right)$ is obtained by converting the arbitrary prediction function $\hat{x}_i = h_i\left(x^{i-1}\right)$ into a PDF:

$$Q\left(x_i|x^{i-1}\right) = \sqrt{\frac{1}{2\pi\sigma^2\left(\lambda\right)}}e^{-\frac{\left(x_i - h_i\left(x^{i-1}\right)\right)^2}{2\sigma^2(\lambda)}}. \tag{5.24}$$

The prediction begins after observing p initial values. We then find that

$$\mathbb{E}_\lambda\left[\log\frac{P_\lambda\left(X^n\right)}{Q\left(X^n\right)}\right] = -\frac{1}{2}\left(n-p\right) + \frac{1}{2\sigma^2\left(\lambda\right)}\mathbb{E}_\lambda\left[\sum_{i=p+1}^{n}\left(X_i - h_i\left(X^{i-1}\right)\right)^2\right], \tag{5.25}$$

which upon rearranging and insertion into the random coding theorem and division by $n-p$ yields Eq. (5.23). □

Corollary 1. For a strictly stationary sampled Ornstein-Uhlenbeck process with sampling interval $\delta > 0$, the following lower bound is obtained:

$$\frac{1}{n-1}\mathbb{E}_{(\mu,\theta)}\left[\sum_{i=2}^{n}\left(X_{i\delta} - h_i\left(X^{(i-1)\delta}\right)\right)^2\right] \geq \frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}\left[1 + (1-\epsilon)\frac{2C_R}{n-1}\right], \tag{5.26}$$

Proof. By choosing $\sigma^2(\lambda) = \frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}$ and $p = 1$ according to the Ornstein-Uhlenbeck process specifications Eq. (5.16), the desired result is obtained. □

Remark 1. If the prior w is chosen as Jeffreys' prior, then the random coding capacity C_R can be replaced by C_n from Eq. (5.8) in the case of Gaussian ARMA processes.

Theorem 5.23 is a generalization of a well known lower bound obtained for a uniform prior w [83]. The greatest lower bound results from choosing Jeffreys' prior. In the case of a uniform prior w, the number of distinguishable distributions is proportional to $n^{\frac{m}{2}}$, provided that some parameter estimators exist that converge sufficiently fast (cf. [79]). The conditions hold for the strictly stationary sampled Ornstein-Uhlenbeck process. In that case, C_R in Inequality (5.23) has to be replaced by $\frac{m}{2}\log(n)$ with $m = 2$ in our case on account of the number of free parameters in the Ornstein-Uhlenbeck parametric family. Note that if w was chosen such that only one distribution could be effectively distinguished, the lower bound would be equal to $\frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}$. The same lower bound would be reached if the two free parameters θ and μ were known and would not have to be estimated first. The second part of Eq. (5.23) $(1-\epsilon)\frac{2C_R}{n-p}$ hence measures the additional complexity of having unknown free parameters.

The lower bound in Eq. (5.23) for Jeffreys' prior and the lower bound for the uniform prior can be reached asymptotically. By estimating the AR-coefficient $\psi_1 = e^{-\theta\delta}$ and $\psi_2 = \mu\left(1-e^{-\theta\delta}\right)$ with ordinary least squares (OLS), which for the Ornstein-Uhlenbeck process coincides with a maximum likelihood (ML) estimation of the two parameters conditioned on the first observation, and using these estimates to predict the next sample $\hat{x_{i\delta}} = \hat{\psi_1}x_{(i-1)\delta} + \hat{\psi_2}$, one obtains: $\mathbb{E}_{(\mu,\theta)}\left[\left(X_{i\delta}-\hat{X}_{i\delta}\right)^2\right] = \frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}\left(1+\frac{2}{i}\right) + O\left(i^{-\frac{3}{2}}\right)$ [84, 85]. Summing the previous expression from $i = 2$ to n and dividing by $n-1$, one obtains

$$\begin{aligned}&\frac{1}{n-1}\sum_{i=2}^{n}\mathbb{E}_{(\mu,\theta)}\left[\left(X_{i\delta}-\hat{X}_{i\delta}\right)^2\right]\\ &=\frac{\sigma^2\left(1-e^{-2\theta\delta}\right)}{2\theta}\left(1+\frac{2\left(H_n-1\right)}{n-1}\right)+\\ &O\left(\frac{H_n^{\left(\frac{3}{2}\right)}-1}{n-1}\right),\end{aligned} \tag{5.27}$$

with H_i being the $i-$th harmonic number and $H_i^{(m)}$ the $i-$th generalized harmonic number. For $n \to \infty$, H_n can be approximated by $\log(n)$, while the second term tends to zero. Hence the lower bound in Eq. (5.23) can be reached asymptotically in the Ornstein-Uhlenbeck case as can be seen by inspecting the asymptotic behavior of the term $\frac{C_n}{n-1}$ with C_n given by Eq. (5.8).

5.3.4 *Distinguishability Of Processes From The Ornstein-Uhlenbeck Parametric Family*

Explicit regions of indistinguishability for the Ornstein-Uhlenbeck parametric family are now constructed. If only a finite number of samples are given, then distinct strictly stationary Ornstein-Uhlenbeck processes will not be distinguishable if their parameters (θ_0, μ_0) and (θ_1, μ_1) are too close to one another in a suitable sense. To make this notion more precise, we construct regions of indistinguishability around (θ_0, μ_0) such that, given n samples, the process corresponding to parameters (θ_0, μ_0) and a process corresponding to parameters drawn from the region of indistinguishability around (θ_0, μ_0) will not be effectively distinguishable. The analysis is based on a related investigation of distinguishability for i.i.d. stochastic processes [81]. Let us therefore assume that a realization of the random vector $(X_\delta, \dots, X_{n\delta})^\top$ has been observed. $\mathrm{P}_{(\theta_0,\mu_0)}$ corresponds to the null hypothesis, while $\mathrm{P}_{(\theta_1,\mu_1)}$ is the alternative hypothesis. The observed random vector is drawn from either $\mathrm{P}_{(\theta_0,\mu_0)}$ or $\mathrm{P}_{(\theta_1,\mu_1)}$. Assuming that the type-I error probability α_n is bounded from above by a constant $\epsilon \in (0,1)$, $\alpha_n \le \epsilon$, the minimum type-II error probability

$$\beta_n^\epsilon = \inf_{\substack{\mathrm{A}_n \subseteq \mathbb{R}^n \\ \alpha_n \le \epsilon}} \beta_n, \tag{5.28}$$

with A_n an acceptance region for the null hypothesis, is given asymptotically (via a generalized Stein's Lemma [86]) as

$$\lim_{n\to\infty} -\frac{1}{n}\log\left(\beta_n^\epsilon\right) = D\left(\mu_1, \theta_1 || \mu_0, \theta_0\right). \tag{5.29}$$

For a fixed number of samples n, we then find the following region of indistinguishability around (θ_0, μ_0)

$$\frac{\kappa}{n} \geq D\left(\mu_1, \theta_1 || \mu_0, \theta_0\right) \approx$$
$$\frac{1}{2}\begin{pmatrix}\theta_1-\theta_0 & \mu_1-\mu_0\end{pmatrix} I\left(\theta_0, \mu_0\right)\begin{pmatrix}\theta_1-\theta_0 \\ \mu_1-\mu_0\end{pmatrix}, \tag{5.30}$$

with $\kappa = -\log\left(\beta^*\right) + \log\left(1-\epsilon\right)$ and β^* a constant between 0 and 1. For sufficiently large n, β^* will be smaller than β_n^ϵ, showing that the type-II error will be greater than a certain constant. Eq. (5.30) shows that the regions of indistinguishability around (θ_0, μ_0) are given by ellipses whose major axes depend on the local value of the Fisher Information Matrix. Starting with such regions of indistinguishability, a covering of parameter space can be carried out. An illustration of such a procedure is given in Fig. 5.2 with parameters $\beta^* = 0.95$, $\epsilon = 0.01$ and $\delta = 0.1$ for two different sequence lengths $n = 50$ and $n = 100$.

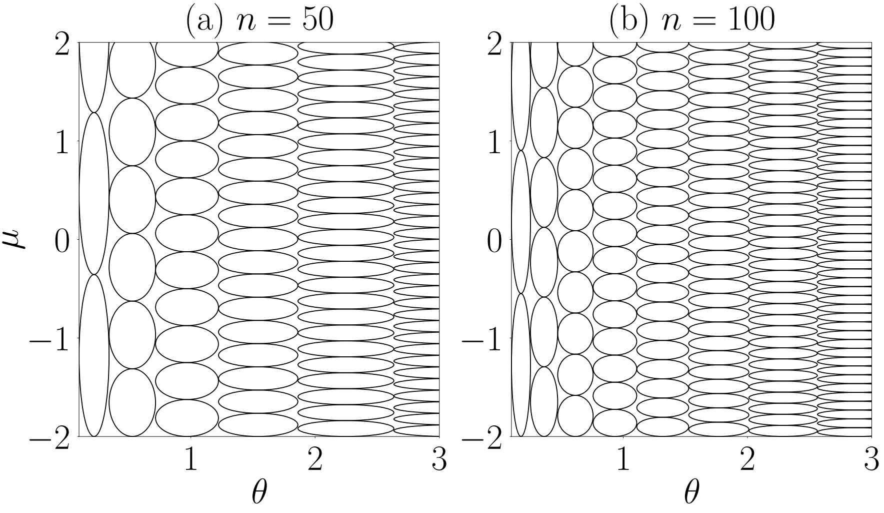

FIGURE 5.2: Coarse illustrative partition of parameter space by regions of indistinguishability

5.4 EMPIRICAL RESULTS WITH DEEP NETWORKS

The results described in Sections 5.2 and 5.3 are intrinsic properties of parametric families. We first recapitulated general results of universal coding theory and derived specific results for the Ornstein-Uhlenbeck parametric family thereafter. By an empirical analysis, we show in this Section that the previously made statements have repercussions for machine learning as well. The choice of the specific learning algorithm is to some extent arbitrary for this task. We have hence chosen standard RNN architectures with LSTM units [87], as these are state of the art for time series prediction.

We first describe a constructive scheme to approximate the optimal solution from Section 5.3.3 within the hypothesis space of a RNN. The approximation methods described in Section 5.4.1 are used to verify that the chosen RNN architecture described in Section 5.4.2 can in principle approximate closely the optimal solution. To carry out the approximations, the results from Eq. (5.15) and the appendix of this chapter are required as the domain of the input to the RNN needs to be known.

5.4.1 *Approximating The Optimal Solution Through Explicit Construction*

A RNN with a single hidden layer with LSTM units is used for the sequential prediction task. In order to approximate the solution based on the OLS equations discussed in Section 5.3.3 (cf. [85] for the OLS equations), each sub-expression in the OLS equations is approximated through one of the units in the recurrent layer. In order to approximate the expression $x^2 + y$, for example, we first approximate x and y through two of the recurrent units, x^2 with another unit and finally $x^2 + y$ with a fourth unit. The OLS equations contain both polynomial terms of second order as well as reciprocal terms. Three main ideas are used for the approximation of the equations with the LSTM layer. The first idea is to rescale the input to the approximately linear region of the corresponding tanh/sigmoid non-linearity. This step requires a careful analysis of the growth behavior of the individual terms in the OLS equations. Eq. (5.15) provides an upper and lower bound within finite time intervals for the strictly stationary Ornstein-Uhlenbeck process, with $C \approx 1$ from numerical simulations. From this as well as a more thorough analysis of the growth behavior of terms in the OLS equations detailed in the appendix of this chapter, it is possible to obtain scaling factors that ensure that the rescaled input is within the linear region for some finite time horizon. The second idea is to approximate the multiplication operation required in the OLS equations by the use of Hadamard multiplication in the LSTM update equation for the cell state. The last idea is to approximate the division operation by first approximating the inverse of the divisor and by then using the multiplication approximation to multiply the dividend and the inverse of the divisor. For the approximation of the inverse, we can either train a sub-network to approximate the operation within our range of interest or we can use a constructive approximation scheme closely based on previous work [88].

5.4.2 *Training On Jeffreys' Prior And Uniform Prior*

To elucidate the importance of sampling of the parameter space on the performance of the RNN, we train two networks with the same configuration and training conditions, one where the process parameters are sampled according to Jeffreys' prior and the other where the sampling is carried out according to a uniform prior. We choose a network with a single layer of 100 units, followed by a linear transformation to a single dimension for the prediction. This network can approximate the optimal solution closely. The network is trained with stochastic gradient descent with a learning rate of 0.001 with early stopping. The range of the parameter μ for the process is $(-2, 2)$, while the range for the parameter θ is $(0.01, 3)$. The sampling interval δ is set to 10, while n is arbitrarily set to 500.

Each of the two trained models are tested on sequences drawn from the two priors - Jeffreys' and Uniform. The results for the case of 50 parameters sampled during training are shown in Table 5.1. The results are averaged over 5 draws of parameter sampling and 10 random initializations of the network for each draw.

It is observed that with an increasing number of parameter samples drawn from the parameter space, the difference in the performance of the models trained on the two priors gets smaller. This can be seen in Fig. 5.3, in which the performance of the models trained on stochastic process realizations drawn from the two priors (Jeffreys' and Uniform) and tested on Jeffreys' prior is plotted against the number of stochastic process realizations drawn.

5.5 DISCUSSION

Classical machine learning theory investigates the learnability of relationships from i.i.d. samples drawn from a fixed but unknown probability

		Test Prior	
		Uniform	Jeffreys'
Train Prior	Uniform	2.91 ± 0.4	3.83 ± 0.25
	Jeffreys'	2.94 ± 0.32	3.4 ± 0.2
	Optimal	0.79	1.11

Table 5.1: Comparing the performance (MSE) of models trained on the two priors and tested on the two priors. 'Optimal' is related to the lower bounds from Section 5.3.3.

distribution, as alluded to in Section 5.1. For the non-i.i.d. case, extensions of statistical learning theory type guarantees have been developed (cf. [89, 90] as well as references therein). Generalization is always understood to refer to the same distribution generating the training/test set.

If multiple distributions are to be learned, it is natural to require the model to do equally well on all of them. This requirement can be directly translated into the language of universal coding theory. The number of independent realizations of stochastic processes p drawn independently according to some prior w on the compact parameter space as well as the length n of each stochastic process realization are, as is intuitively clear, crucial for any required theory of generalization in the parametric family context. In classical statistical learning theory it is n, as well as the complexity of the hypothesis space, which is the main focus of investigation. For finite n, only finitely many stochastic processes are distinguishable. Asymptotically in n, for the stochastic processes considered in this paper, the capacity-inducing prior will be given by Jeffreys' prior. Since the maximum number of distinguishable models is close to e^{C_n}, p will have to be at least equal to e^{C_n}. In fact, since C_n is in general growing with increasing n, the minimum number of required stochastic process realizations p will depend on n. The dependence of p on n therefore implicitly reflects the fact that the number of distinguishable distributions in a parametric family grows with increasing n. Since the capacity-inducing prior w^* is the prior under which the maximum number of distributions in the parametric family are distinguishable, it follows that p adapted to this prior is sufficient for any other prior. Finding a p adapted to w^* is therefore a necessary requirement if one attempts to learn the entire parametric fam-

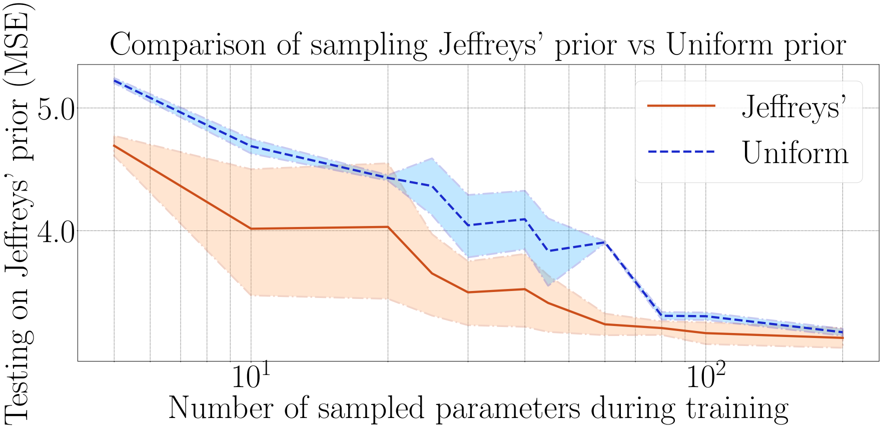

FIGURE 5.3: Comparing the performance (MSE) of the models trained on two priors, tested on Jeffreys' prior, with increasing number of sampled parameters during training.

ily. The empirical counterpart of this statement for the case of MSE loss is found in Fig. 5.3 as well as Table 5.1. Training on stochastic process realizations drawn from Jeffreys' prior ensures that testing on a different prior (here the uniform prior was chosen) does not lead to an increased MSE loss. Training on the uniform prior and testing on Jeffreys' prior, however, leads to a marked increase in MSE loss.

The capacity used in the lower bound Eq. (5.7) as well as in the lower bound Eq. (5.23) is the capacity of the parametric family and not the capacity of the hypothesis space of the machine learner. Notions of capacity for the machine learner reflect the richness of the class of functions that such a learner can approximate. The capacity C_n, on the other hand, measures the richness of the parametric family.

Assume that it was only known that a set of observations could be modelled by a parametric family with m free parameters, while the specific form of the parametric family was not known. In such a case, it would not be possible to obtain p such that, uniformly for all possible parametric families with m free parameters, p would be sufficient to guarantee that any parametric family could be fully learned (in the sense that the solution found should be close to a mixture source induced by the capacity-achieving prior). If the form of the parametric family was not known, it seems reasonable to use stochastic process realizations drawn uniformly from the space of parameters. If the capacity-inducing prior, however, was very different from the uniform prior, then most of the obtained realizations from the uniform prior would not facilitate learning the parametric family fully. The ill-adapted sampling mechanism would prohibit an optimal learning of the parametric family. The testing error in Fig. 5.3, with testing performed by drawing stochastic process realizations from Jeffreys' prior and training carried out either by using Jeffreys' or the Uniform prior, converges to the same error for increasing p. This behavior is expected in view of the fact that the two priors are positive everywhere within the parameter space, as can be seen in Fig. 5.1. A more subtle analysis of this fact can be carried out by noting that the number of distinguishable distributions under both priors is not too different from one another as discussed in Section 5.3.3 for the parametric family considered in this paper.

Eq. (5.23) provides a lower bound on the sequential prediction error for the MSE loss, assuming that the form of the parametric family was known. The empirical results obtained in Section 5.4, on the other hand, do not require knowledge of the specific form. By the explicit construction detailed in Section 5.4.1, it is shown that a solution close to an optimal solution lies in the hypothesis space of the chosen network architecture. It is hence guaranteed that the chosen deep network is in principle well specified. The results shown in Table 5.1 indicate that the empirical solution found by the network does not reach the lower bounds, here denoted by 'Optimal', implying that an inefficiency exists in the optimization procedure. A thorough analysis is outside of the scope of this chapter, however, as it would necessitate an investigation of the loss landscape of the chosen deep network with stochastic process realizations drawn according to some prior w as input as well as of the optimization algorithm used.

Empirically it was observed in the experiments that if one first trains the deep network with observations drawn from some prior w_1 until convergence and thereafter changes the prior to some w_2 and continues training, the previously found solution changes. This behavior is expected in view of the previous discussion, as a changed prior induces a different optimal solution. It follows that there is a close link between optimal solutions and the sampling of parameter space.

Most of the previous statements hold for more general families of distributions and not only for parametric families. Eq. (5.7) as well as the statements on the capacity-achieving prior hold in particular in more general contexts [79]. The simple form of the capacity Eq. (5.8) as well as the fact that Jeffreys' prior is asymptotically capacity-inducing are, however, not correct in a more general context. To achieve optimality, however, the sampling mechanism should still be matched to w^*.

APPENDIX

We derive some results which are needed for the explicit construction of the RNN used to implement the aymptotically optimal solution for the sequential prediction of the sampled strictly stationary Ornstein-Uhlenbeck process. Let us study the time-integral of the strictly stationary Ornstein-Uhlenbeck process:

$$Y_t = \int_0^t X_s ds. \tag{5.31}$$

$\{Y_t\}$ is a Gaussian process implying that it is fully characterized by its mean and covariance function. For the mean as a function of t one obtains

$$\mathbb{E}\left[Y_t\right] = \mathbb{E}\left[\int_0^t \mu + \frac{\sigma}{\sqrt{2\theta}} e^{-\theta s} W_{e^{2\theta s}} ds\right] = \int_0^t \mathbb{E}\left[\mu + \frac{\sigma}{\sqrt{2\theta}} e^{-\theta s} W_{e^{2\theta s}}\right] ds = \mu t, \tag{5.32}$$

with the exchange of integration and expectation order justified by Fubini's theorem, while the covariance function is given by

$$\begin{aligned} \mathrm{Cov}\left(Y_t, Y_s\right) &= \mathbb{E}\left[Y_t Y_s\right] - \mu^2 ts = \mathbb{E}\left[\int_0^s \int_0^t X_a X_b da db\right] - \mu^2 ts \\ &= \frac{\sigma^2}{2\theta^3}\left(e^{-\theta s} + e^{-\theta t} - e^{-\theta|t-s|} + 2\theta \min\left(s,t\right) - 1\right). \end{aligned} \tag{5.33}$$

Let us next analyze the time-integral of the squared strictly stationary Ornstein-Uhlenbeck process:

$$Z_t = \int_0^t X_s^2 ds. \tag{5.34}$$

The expectation of $\{Z_t\}$ is given by

$$\begin{aligned} \mathbb{E}\left[Z_t\right] &= \mathbb{E}\left[\int_0^t \mu^2 + \sqrt{2}\mu \frac{\sigma}{\sqrt{\theta}} e^{-\theta s} W_{e^{2\theta s}} + \frac{\sigma^2}{2\theta} e^{-2\theta s} W^2_{e^{2\theta s}} ds\right] \\ &= \left(\mu^2 + \frac{\sigma^2}{2\theta}\right) t, \end{aligned} \tag{5.35}$$

while the covariance function is

$$\begin{aligned}\mathrm{Cov}\left(Z_t, Z_s\right) &= \mathbb{E}\left[\int_0^s \int_0^t X_a^2 X_b^2 dadb\right] - \left(\mu^2 + \frac{\sigma^2}{2\theta}\right)^2 ts \\ &= \frac{\sigma^4}{8\theta^4}\left(e^{-2\theta s} + e^{-2\theta t} - e^{-2\theta|t-s|} + 4\theta \mathrm{min}\left(t,s\right) - 1\right) + \\ &\frac{2\mu^2\sigma^2}{\theta^3}\left(e^{-\theta s} + e^{-\theta t} - e^{-\theta|t-s|} + 2\theta \mathrm{min}\left(t,s\right) - 1\right). \end{aligned} \tag{5.36}$$

$\{Z_t\}$ is not a Gaussian process. Let us study sums of the form $\sum_{i=1}^n X_{(i-1)\delta}$ with a sampling interval δ and $\{X_t\}$ a strictly stationary Ornstein-Uhlenbeck process.

$\left(X_0, X_\delta, \ldots, X_{(n-1)\delta}\right)^\top$ is distributed according to a multivariate normal distribution with mean vector $(\mu, \mu, \ldots, \mu)^\top$ and covariance matrix

$$\Sigma = \frac{\sigma^2}{2\theta}\begin{pmatrix} 1 & e^{-\theta\delta} & \ldots & e^{-\theta(n-1)\delta} \\ e^{-\theta\delta} & 1 & \ldots & e^{-\theta(n-2)\delta} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\theta(n-1)\delta} & e^{-\theta(n-2)\delta} & \ldots & 1 \end{pmatrix}. \tag{5.37}$$

Hence it follows that the sum $\sum_{i=1}^n X_{(i-1)\delta}$ is distributed according to a Gaussian distribution with mean $n\mu$ and variance

$$\mathrm{Var}\left(\sum_{i=1}^n X_{(i-1)\delta}\right) = \frac{\sigma^2}{2\theta}\frac{2e^{-\theta(n-1)\delta} - 2e^{\theta\delta} + n\left(e^{2\theta\delta} - 1\right)}{\left(e^{\theta\delta} - 1\right)^2}. \tag{5.38}$$

Next sums of the form $\sum_{i=1}^n X_{(i-1)\delta}^2$ are studied. We find $\mathbb{E}\left[\sum_{i=1}^n X_{(i-1)\delta}^2\right] = \left(\mu^2 + \frac{\sigma^2}{2\theta}\right)$ and

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^n X_{(i-1)\delta}^2\right) = \frac{\sigma^2}{2\theta}\Bigg(&\frac{8\theta\mu^2\left(e^{-\theta(n-1)\delta} - e^{\delta\theta} + ne^{\delta\theta} - n\right)}{\left(e^{2\delta\theta} - 1\right)^2} n\sigma^2 + \\ &\frac{2\sigma^2\left(e^{-2(n-1)\delta\theta} - e^{2\delta\theta} + ne^{2\delta\theta} - n\right)}{\left(e^{2\delta\theta} - 1\right)^2} + \\ &\frac{8\theta\mu^2\left(e^{-\theta(n-1)\delta} - e^{\delta\theta} + ne^{\delta\theta} - n\right)}{\left(e^{2\delta\theta} - 1\right)^2}\Bigg). \end{aligned} \tag{5.39}$$

Given that $\sum_{i=1}^{n} X_{(i-1)\delta}$ is a Gaussian random variable, $\left(\sum_{i=1}^{n} X_{(i-1)\delta}\right)^2$ will be a noncentral χ^2 distribution. $\frac{\left(\sum_{i=1}^{n} X_{(i-1)\delta}\right)^2}{\frac{\sigma^2}{2\theta}\frac{2e^{-\theta(n-1)\delta}-2e^{\theta\delta}+n\left(e^{2\theta\delta}-1\right)}{\left(e^{\theta\delta}-1\right)^2}}$ is hence distributed as

$$\chi^2\left(1, \frac{n^2\mu^2}{\left(\frac{\sigma^2}{2\theta}\frac{2e^{-\theta(n-1)\delta}-2e^{\theta\delta}+n\left(e^{2\theta\delta}-1\right)}{\left(e^{\theta\delta}-1\right)^2}\right)^2}\right). \tag{5.40}$$

6 SUMMARY AND OUTLOOK

The benefits of randomness and irregularity have been studied in this thesis. In Chapter 3 it has been shown that the use of randomness in sampling in addition to some sparsity assumptions has enabled the stable recovery of a bandlimited function from finitely many samples. Such a result cannot be obtained for deterministic sampling schemes in general. Chapter 4 has then shown that operations such as filtering which are commonly performed on uniformly sampled functions can be extended to nonuniformly sampled ones, provided the maximum distance between consecutive samples is bounded by a constant which depends on the bandwidth of the sampled function. To preserve some properties needed to define an approximation to a convolution operation, some restriction to the possible amount of irregularity is therefore necessary. Chapter 5 has then shown that the specific nature of randomness used in sampling (in the statistical sense) has implications on the amount of information transmitted by the sampled data. Randomness/irregularity therefore has to be matched to specific problems to be optimal or to provide benefits. Chapter 5 has additionally shown that solutions found by black-box machine learning methods strongly depend on sampling procedures, i.e. on the way training data are generated.

This thesis can be extended in multiple ways. In Chapter 2, event-based sampling of bandlimited functions was studied. An extension of such an analysis to other function spaces would be useful as many real-world signals are not bandlimited or possess an unknown bandlimit. For a non-bandlimited function to be reconstructible from its samples, different conditions are in general required than for the bandlimited case. Shift-invariant function spaces are commonly encountered as models that are more flexible than bandlimited function spaces. Necessary and sufficient conditions on the sampling pattern which ensure reconstructability have been investigated for these spaces [91–94]. Instead of starting with a function space and identifying necessary and sufficient sampling set requirements, it would be interesting to reverse this process: given an event-based sampling scheme, determine and describe the function space whose members can be uniquely identified from samples obtained by the event-based sampling scheme. A related investigation has been carried out for zero-crossing sampling [95]. A subclass of bandpass signals is described in this publication for which the zero-crossing samples determine the original function. It would be a fruitful endeavor to extend these results to more general event-based sampling schemes. Chapter 4 studied the filtering of irregularly sampled bandlimited functions. An algorithm for filtering was given. Efficient numerical implementations could be investigated next which could be similar to fast reconstruction algorithms (cf. [96]) as the underlying mathematical structure is similar in both cases. The impact of different error types such as truncation, jitter, quantization and aliasing errors on the quality of the filter algorithm could be investigated as well. Due to the linearity of all involved operations in the algorithm, such a study could be carried out both analytically as well as numerically.

Chapter 5 could be extended in multiple different directions. Explicit schemes could be developed that implement minimax optimal cumulative Kullback-Leibler risk prediction for Ornstein-Uhlenbeck parametric families. The basic ingredients for such an approach are already developed as an explicit formula for Jeffreys' prior has been derived in this chapter. Such a prediction scheme could then be compared to Kalman filters, for example, which can be derived for a Gaussian prior on the parameter space of the parametric family. A different extension of this chapter could be a study on how to make black-box machine learning algorithms more data efficient. The results shown in Chapter 5 indicate that on average more information is provided to a learning algorithm if data are sampled according to the capacity-inducing prior than according to a different prior. If fast and data-efficient learning of a parametric family is desired and if only few realizations can be drawn from a parametric family, it is certainly beneficial to draw these realizations according to the capacity-inducing prior. Machine learning approaches are particularly suitable for the case that it is easier to derive the capacity-inducing prior of a problem than to solve the ensuing analytical integrals needed for the determination of the minimax optimal cumulative Kullback-Leibler risk prediction scheme. The latter part could be substituted by a machine learning algorithm. In principle it would be possible to use similar results as those derived in Chapter 5 for the task of incremental learning. If only few example across categories and within categories can be kept, then it would be preferential to choose them according to the capacity-inducing prior, as this would ensure that these examples contain on average the maximum amount of information about the underlying classification task. The main difficulty that such a method would face is the general lack of knowledge about the probability distributions which describe different categories. It might even be argued that the determination of these distributions from the data (either explicitly or implicitly) is the task that a machine learning algorithm should solve. Without some estimates of the distributions of the categories that are to be learned, however, no capacity-inducing prior can be derived. Future investigations could therefore preferentially identify problems for which it is easy to obtain some estimates of the underlying probability distributions, but for which it is hard to derive optimal solutions. For this type of problems, the analysis presented in Chapter 5 could provide a good starting point.

BIBLIOGRAPHY

1. Landau, H. Necessary Density Conditions for Sampling and Interpolation of Certain Entire Functions. *Acta Mathematica* **117**, 37 (1967).
2. Hruščëv, S., Nikol'skii, N. & Pavlov, B. in *Complex Analysis and Spectral Theory* 214 (Springer, 1981).
3. Kadec, M. Bases and their Spaces of Coefficients. *Dopov. Akad. Ukr. RSR* **9**, 1139 (1964).
4. Jaffard, S. *A Density Criterion for Frames of Complex Exponentials.* 1991.
5. Feichtinger, H. G. & Gröchenig, K. Theory and Practice of Irregular Sampling. *Wavelets: Mathematics and Applications*, 305 (1994).
6. Kay, S. *Modern Spectral Estimation: Theory and Application* (PTR Prentice Hall, 1988).
7. Beutler, F. Alias-Free Randomly Timed Sampling of Stochastic Processes. *IEEE Transactions on Information Theory* **16**, 147 (1970).
8. Shapiro, H. S. & Silverman, R. A. Alias-Free Sampling of Random Noise. *Journal of the Society for Industrial and Applied Mathematics* **8**, 225 (1960).
9. Masry, E. Random Sampling and Reconstruction of Spectra. *Information and Control* **19**, 275 (1971).
10. Masry, E. Alias-Free Sampling: An Alternative Conceptualization and its Applications. *IEEE Transactions on Information Theory* **24**, 317 (1978).
11. Masry, E. Poisson Sampling and Spectral Estimation of Continuous-Time Processes. *IEEE Transactions on Information Theory* **24**, 173 (1978).
12. Srivastava, R. & Sengupta, D. Effect of Inter-Sample Spacing Constraint on Spectrum Estimation with Irregular Sampling. *IEEE Transactions on Information Theory* **57**, 4709 (2011).
13. Srivastava, R. & Sengupta, D. Nonparametric Spectrum Estimation under the Constraint of a Minimum Inter-Sample Spacing. *Statistica Sinica*, 691 (2013).
14. Baccelli, F. & Woo, J. O. *On the Entropy and Mutual Information of Point Processes* in *2016 IEEE International Symposium on Information Theory (ISIT)* (2016), 695.

15. Sayiner, N. A Level-Crossing Sampling Scheme for A/D Conversion. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing* **43**, 335 (1996).

16. Miskowicz, M. Send-On-Delta Concept: An Event-Based Data Reporting Strategy. *Sensors* **6**, 49 (2006).

17. Huber, A. E. & Liu, S.-C. *On Send-on-Delta Sampling of Bandlimited Functions* in *2017 International Conference on Sampling Theory and Applications (SampTA)* (2017), 422.

18. Steele, R. *Delta Modulation Systems* (Pentech Press & Halsted Press, 1975).

19. Yang, M., Liu, S. C. & Delbruck, T. A Dynamic Vision Sensor With 1% Temporal Contrast Sensitivity and In-Pixel Asynchronous Delta Modulator for Event Encoding. *IEEE Journal of Solid-State Circuits* **50**, 2149 (2015).

20. Yang, M., Chien, C.-H., Delbruck, T. & Liu, S.-C. A 0.5 V 55 μW 64x2 Channel Binaural Silicon Cochlea for Event-Driven Stereo-Audio Sensing. *IEEE Journal of Solid-State Circuits* **51**, 2554 (2016).

21. Higgins, J. R. *Sampling Theory in Fourier and Signal Analysis: Foundations* (Oxford University Press on Demand, 1996).

22. Seip, K. An Irregular Sampling Theorem for Functions Bandlimited in a Generalized Sense. *SIAM Journal on Applied Mathematics* **47**, 1112 (1987).

23. Boche, H. & Mönich, U. J. Convergence Behavior of Non-Equidistant Sampling Series. *Signal Processing* **90**, 145 (2010).

24. Mönich, U. J. & Boche, H. Non-Equidistant Sampling for Bounded Bandlimited Signals. *Signal Processing* **90**, 2212 (2010).

25. Bar-David, I. An Implicit Sampling Theorem for Bounded Bandlimited Functions. *Information and Control* **24**, 36 (1974).

26. Bond, F. & Cahn, C. On Sampling the Zeros of Bandwidth Limited Signals. *IEEE Transactions on Information Theory* **4**, 110 (1958).

27. Lazar, A. A. & Tóth, L. T. Perfect Recovery and Sensitivity Analysis of Time Encoded Bandlimited Signals. *IEEE Transactions on Circuits and Systems I: Regular Papers* **51**, 2060 (2004).

28. Gontier, D. & Vetterli, M. Sampling Based on Timing: Time Encoding Machines on Shift-Invariant Subspaces. *Applied and Computational Harmonic Analysis* **36**, 63 (2014).

29. Feichtinger, H. G., Príncipe, J. C., Romero, J. L., Singh Alvarado, A. & Velasco, G. A. Approximate Reconstruction of Bandlimited Functions for the Integrate and Fire Sampler. *Advances in Computational Mathematics* **36**, 67 (2012).

30. Partington, J. R. *Interpolation, Identification, and Sampling* **17** (Oxford University Press, 1997).

31. Ferreira, P. J. S. G. & Kempf, A. Superoscillations: Faster than the Nyquist Rate. *IEEE Transactions on Signal Processing* **54**, 3732 (2006).

32. Schmidt, J. W. Univariate Strip Interpolations by Nonlinear Parametric Splines. *Computing* **65**, 323 (2000).

33. Kempf, A. Fields with Finite Information Density. *Physical Review D* **69** (2004).

34. Hao, Y. & Kempf, A. On a Non-Fourier Generalization of Shannon Sampling Theory. *10th Canadian Workshop on Information Theory, CWIT 2007*, 193 (2007).

35. Huber, A. E. & Liu, S.-C. *On Approximation of Bandlimited Functions with Compressed Sensing* in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 4009.

36. Donoho, D. Compressed Sensing. *IEEE Transactions on Information Theory* **52**, 1289 (2006).

37. Candès, E. J., Romberg, J. & Tao, T. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. *IEEE Transactions on Information Theory* **52**, 489 (2006).

38. Foucart, S. & Rauhut, H. *A Mathematical Introduction to Compressive Sensing* (Springer New York, 2013).

39. Helms, H. D. & Thomas, J. B. Truncation Error of Sampling-Theorem Expansions. *Proceedings of the IRE* **50**, 179 (1962).

40. Brown, J. Bounds for Truncation Error in Sampling Expansions of Band-Limited Signals. *IEEE Transactions on Information Theory* **15**, 440 (1969).

41. Knab, J. System Error Bounds for Lagrange Polynomial Estimation of Band-Limited Functions. *IEEE Transactions on Information Theory* **21**, 474 (1975).

42. Radzyner, R. & Bason, P. T. An Error Bound for Lagrange Interpolation of Low-Pass Functions. *IEEE Transactions on Information Theory* **18**, 669 (1972).

43. Klamer, D. & Masry, E. Polynomial Interpolation of Randomly Sampled Bandlimited Functions and Processes. *SIAM Journal on Applied Mathematics* **42**, 1004 (1982).

44. Strohmer, T. & Tanner, J. Fast Reconstruction Methods for Bandlimited Functions From Periodic Nonuniform Sampling. *SIAM J. Numer. Anal.* **44**, 1073 (2006).

45. Selva, J. Functionally Weighted Lagrange Interpolation of Band-Limited Signals from Nonuniform Samples. *IEEE Transactions on Signal Processing* **57**, 168 (2009).

46. Adcock, B. & Hansen, A. Generalized Sampling and Infinite-Dimensional Compressed Sensing. *Foundations of Computational Mathematics* **16**, 1263 (2016).

47. Gosse, L. Compressed Sensing with Preconditioning for Sparse Recovery with Subsampled Matrices of Slepian Prolate Functions. *Annali dell'Universita di Ferrara* **59**, 81 (2013).

48. Slepian, D. & Pollak, H. Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - I. *Bell System Technical Journal, The* **40**, 43 (1961).

49. Landau, H. J. & Pollak, H. O. *Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - II* 1961.

50. Landau, H. J. & Pollak, H. O. Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - III: The Dimension of the Space of Essentially Time- and Band-Limited Signals. *Bell System Technical Journal* **41**, 1295 (1962).

51. Kowalski, M. On Approximation of Band-Limited Signals. *Journal of Complexity* **5**, 283 (1989).

52. Yao, K. Applications of Reproducing Kernel Hilbert Spaces - Bandlimited Signal Models. *Information and Control* **11**, 429 (1967).

53. Osipov, A., Rokhlin, V. & Xiao, H. *Prolate Spheroidal Wave Functions of Order Zero: Mathematical Tools for Bandlimited Approximation* 379 (Springer US, 2013).

54. Adcock, B. Infinite-Dimensional Compressed Sensing and Function Interpolation. *Foundations of Computational Mathematics* (2017).

55. Puy, G., Davies, M. & Gribonval, R. Recipes for Stable Linear Embeddings from Hilbert Spaces to $\mathbb{R}^m$. *IEEE Transactions on Information Theory* **63**, 2171 (2017).

56. Adelman, R., Gumerov, N. A. & Duraiswami, R. Software for Computing the Spheroidal Wave Functions Using Arbitrary Precision Arithmetic. *ArXiv e-prints* (2014).

57. Huber, A. & Liu, S.-C. Filtering of Nonuniformly Sampled Bandlimited Functions. *IEEE Signal Processing Letters*, 1036 (2019).

58. Steiglitz, K. The Equivalence of Digital and Analog Signal Processing. *Information and Control* **8**, 455 (1965).

59. Masry, E., Steiglitz, K. & Liu, B. Bases in Hilbert Space Related to the Representation of Stationary Operators. *SIAM Journal on Applied Mathematics* **16**, 552 (1968).

60. Oppenheim, A. V. & Johnson, D. H. Discrete Representation of Signals. *Proceedings of the IEEE* **60**, 681 (1972).

61. Marvasti, F. *Nonuniform Sampling: Theory and Practice* (Springer Science & Business Media, 2001).

62. Hamming, R. W. *Digital Filters* (Courier Corporation, 1998).

63. Chen, Y. & Tsividis, Y. *Signal Processing in Continuous Time using Asynchronous Techniques* in *2017 51st Asilomar Conference on Signals, Systems, and Computers* (2017), 1605.

64. Fesquet, L. & Bidégaray-Fesquet, B. IIR Digital Filtering of Non-Uniformly Sampled Signals via State Representation. *Signal Processing* **90**, 2811 (2010).

65. Feichtinger, H. G. in *Progress in Approximation Theory* 333 (Academic Press, Boston, MA, 1991).

66. Feichtinger, H. G. & Gröchenig, K. Irregular Sampling Theorems and Series Expansions of Band-Limited Functions. *J. Math. Anal. Appl* **167**, 530 (1992).

67. Tarczynski, A., Valimaki, V. & Cain, G. D. *FIR Filtering of Nonuniformly Sampled Signals* in *IEEE International Conference on Acoustics, Speech, and Signal Processing* **3** (1997), 2237.

68. Kose, K. & Cetin, A. E. Low-Pass Filtering of Irregularly Sampled Signals using a Set Theoretic Framework [Lecture Notes]. *IEEE Signal Processing Magazine* **28**, 117 (2011).

69. Landau, H. Sampling, Data Transmission, and the Nyquist Rate. *Proceedings of the IEEE* **55**, 1701 (1967).

70. Butzer, P., G. Feichtinger, H. & Grōchenig, K. Error Analysis in Regular and Irregular Sampling Theory. *Applicable analysis* **50**, 167 (1993).

71. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer Science & Business Media, 2013).

72. Vapnik, V. *Statistical Learning Theory* (John Wiley & Sons, 1998).

73. Recht, B., Roelofs, R., Schmidt, L. & Shankar, V. Do CIFAR-10 Classifiers Generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451* (2018).

74. Caruana, R. Multitask Learning. *Machine Learning* **28**, 41 (1997).

75. Rissanen, J. *Information and Complexity in Statistical Modeling* (Springer Science & Business Media, 2007).

76. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, 2012).

77. Barron, A. R. The Strong Ergodic Theorem for Densities: Generalized Shannon-McMillan-Breiman Theorem. *The Annals of Probability* **13**, 1292 (1985).

78. Merhav, N. & Feder, M. Universal Prediction. *IEEE Transactions on Information Theory* **44**, 2124 (1998).

79. Merhav, N. & Feder, M. A Strong Version of the Redundancy-Capacity Theorem of Universal Coding. *IEEE Transactions on Information Theory* **41**, 714 (1995).

80. Rissanen, J. Fisher Information and Stochastic Complexity. *IEEE Transactions on Information Theory* **42**, 40 (1996).

81. Balasubramanian, V. A Geometric Formulation of Occam's Razor for Inference of Parametric Distributions. *arXiv preprint adap-org/9601001* (1996).

82. Jeffreys, H. *The Theory of Probability* (OUP Oxford, 1998).

83. Rissanen, J. Universal Coding, Information, Prediction, and Estimation. *IEEE Transactions on Information Theory* **30**, 629 (1984).

84. Fuller, W. A. & Hasza, D. P. Properties of Predictors for Autoregressive Time Series. *Journal of the American Statistical Association* **76**, 155 (1981).

85. Fuller, W. A. & Hasza, D. P. Predictors for the First-Order Autoregressive Process. *Journal of Econometrics* **13**, 139 (1980).

86. Vajda, I. *Theory of Statistical Inference and Information* (Kluwer Academic Pub, 1989).

87. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9**, 1735 (1997).

88. Jones, L. K. Constructive Approximations for Neural Networks by Sigmoidal Functions. *Proceedings of the IEEE* **78**, 1586 (1990).

89. Kuznetsov, V. & Mohri, M. *Learning Theory and Algorithms for Forecasting Non-Stationary Time Series* in *Advances in Neural Information Processing Systems* (2015), 541.

90. McDonald, D. J., Shalizi, C. R. & Schervish, M. Nonparametric Risk Bounds for Time-Series Forecasting. *Journal of Machine Learning Research* **18**, 1 (2017).

91. Chen, W., Han, B. & Jia, R.-Q. Maximal Gap of a Sampling Set for the Exact Iterative Reconstruction Algorithm in Shift Invariant Spaces. *IEEE Signal Processing Letters* **11**, 655 (2004).

92. Aldroubi, A. & Gröchenig, K. Beurling-Landau-Type Theorems for Non-Uniform Sampling in Shift Invariant Spline Spaces. *Journal of Fourier Analysis and Applications* **6**, 93 (2000).

93. Aldroubi, A. Non-Uniform Weighted Average Sampling and Reconstruction in Shift-Invariant and Wavelet Spaces. *Applied and Computational Harmonic Analysis* **13**, 151 (2002).

94. Aldroubi, A. & Gröchenig, K. Nonuniform Sampling and Reconstruction in Shift-Invariant Spaces. *SIAM review* **43**, 585 (2001).

95. Logan Jr, B. F. Information in the Zero Crossings of Bandpass Signals. *Bell System Technical Journal* **56**, 487 (1977).

96. Feichtinger, H. G., Gröchenig, K. & Strohmer, T. Efficient Numerical Methods in Non-Uniform Sampling Theory. *Numerische Mathematik* **69**, 423 (1995).

CURRICULUM VITAE

PERSONAL DATA

Name	Adrian Emmanuel Georg Huber
Date of Birth	March 06, 1990
Place of Birth	Düsseldorf, Germany
Citizen of	Germany

EDUCATION

2008 – 2013	RWTH Aachen University Aachen, Germany *Final degree:* M.Sc. Electrical Engineering, Information Technology and Computer Engineering
2006 – 2008	Malvern College Malvern, England *Final degree:* International Baccalaureate

EMPLOYMENT

2016 – 2019	Scientific Researcher Institute of Neuroinformatics, ETH and University Zürich Zürich, Switzerland
2014 – 2016	Scientific Researcher Institute for Biomedical Engineering, ETH Zürich Zürich, Switzerland

PUBLICATIONS

Articles in peer-reviewed journals:

1. Huber, A. & Liu, S.-C. Filtering of Nonuniformly Sampled Bandlimited Functions. *IEEE Signal Processing Letters*, 1036 (2019).
2. Huber, A. E., Anumula, J. & Liu, S.-C. Optimal Sampling of Parametric Families: Implications for Machine Learning. *Neural Computation*, accepted.

Conference contributions:

3. Huber, A. E. & Liu, S.-C. *On Send-on-Delta Sampling of Bandlimited Functions* in *2017 International Conference on Sampling Theory and Applications (SampTA)* (2017), 422.
4. Huber, A. E. & Liu, S.-C. *On Approximation of Bandlimited Functions with Compressed Sensing* in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), 4009.
5. Anumula, J., Ceolini, E., He, Z., Huber, A. & Liu, S.-C. *An Event-Driven Probabilistic Model of Sound Source Localization using Cochlea Spikes* in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)* (2018), 1.
6. Ceolini, E., Anumula, J., Huber, A., Kiselev, I. & Liu, S.-C. *Speaker Activity Detection and Minimum Variance Beamforming for Source Separation* in *Proc. Interspeech 2018* (2018), 836.