Applications of Proof Theory to Limit Theorems and Stochastic Processes

Morenikeji Neri

A thesis submitted for the degree of Doctor of Philosophy

of the

University of Bath

Department of Computer Science

January 2025

Declarations

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Morenikeji Neri

I am the author of this thesis, and the work described therein was carried out by myself personally, with the exception of parts of Chapters 3, 4, 7 and 8, which were done in collaboration with my supervisor, Thomas Powell, and parts of Chapters 4 and 5 which were done in collaboration with Nicholas Pischke. Explicit details of these collaborations are given in Section 1.3.

Morenikeji Neri

I believe that mathematics is mostly about finding the right definitions, those that affect the way we see things.

Alessio Guglielmi

There is no answer to the Pythagorean theorem. Well, there is an answer, but by the time you figure it out, I got 40 points, 10 rebounds and then we're planning for the parade.

Shaquille O'Neal

Summary

This thesis represents a stepping stone in advancing the applications of proof-theoretic tools in probability theory and the theory of stochastic processes. Its contributions are both foundational and applied.

The applied aspect of this thesis presents quantitative versions of important results in probability theory. We give quantitative versions of various Strong Laws of Large Numbers, including the improvement of known bounds from the literature. We provide a quantitative version of Doob's seminal martingale convergence theorem, and in doing so, we generalise bounds on the stochastic fluctuations of martingales, found in the literature, to submartingales and supermartingales. We present improved stochastic fluctuation bounds in the pointwise ergodic theorem and bounds on local stability that generalise those found in the applied proof theory literature. Lastly, we provide a quantitative version of the celebrated Robbins-Siegmund theorem and various applications in stochastic approximation theory, including rates for a procedure of Dvoretzky.

The primary foundational contribution of the thesis is the development of abstract frameworks for studying the quantitative aspects of probability theory, with a particular focus on stochastic convergence. This includes introducing a formal system for reasoning about probability theory and a corresponding metatheorem guaranteeing the extractability of uniform quantitative data for a large class of results. Lastly, the thesis presents various proof-theoretic transfer principles that allow for the transformation of quantitative data from deterministic results to their probabilistic analogue.

We conclude this thesis with a discussion on the current work in progress, both foundational and applied, in proof mining in probability theory. We also present some open problems and conjectures, paving the way for future research and development in this exciting field.

Acknowledgements

I would like to express my deepest gratitude to the following individuals and groups for their unwavering support throughout my doctoral studies at Bath University:

First and foremost, I am incredibly grateful to my supervisor, Thomas Powell, for his patience and support. His constant encouragement and extensive academic discussions have been invaluable to the completion of my thesis and the research I conducted. I also want to thank the Mathematical Foundations of Computation group at Bath for their continuous academic support and friendship, as well as the EPSRC Centre for Doctoral Training in Digital Entertainment for funding my research.

I owe a great deal to Nicholas Pischke for his collaboration and friendship. Our meeting at the Logic Colloquium 2023 in Milan marked the beginning of a supportive and motivating academic relationship. I am especially thankful for his willingness to review my drafts and provide invaluable help with the presentation of my mathematics.

I also want to express my gratitude to Ulrich Kohlenbach for his encouragement and support, particularly during his visit to Bath in the summer of 2023. His advocacy for me to receive the OWLG grant to attend the Mathematical Logic: Proof Theory, Constructive Mathematics workshop at Oberwolfach was truly appreciated. I am also thankful to the broader proof mining community for their encouragement and to Pedro Pinto for inviting me to Oberwolfach for academic collaboration during his stay as a Leibniz fellow.

It was a great honour to have been examined by James Laird and Paulo Oliva, and I thank them for their many helpful comments.

In addition to my academic support network, I am deeply thankful to my family for their unwavering support and patience during my time at Bath. I want to thank my parents, Olagoke Neri and Olumide Neri, and my sisters, Subulade Fenojo and Elizabeth Neri, as well as my extended family. I am especially grateful to my maternal grandfather, the late Oba Samuel Kolapo Adegbite-Adedoyin, the Owa Ale of Ikare. Despite not being alive to see me start and complete my doctoral studies, he had always motivated me to strive for academic success, and I would not be in my current academic position without him.

I began my doctoral studies towards the end of the pandemic, a time that was lonely for many people around the world. Therefore, I want to express my gratitude to the many friends

I was fortunate to have during this time. Their friendship and support were crucial in helping me maintain good mental health throughout my doctoral studies. I would like to extend special thanks to Nils van de Berg, Ward van der Schoot, Sam Martin, Nathan Creighton, Tom Quilter, and Kip White-Saini. I am especially grateful to Amit Balter, who was always willing to listen, discuss, and learn about my research despite it not being directly related to his area of expertise.

Finally, to Lata Persson, I feel incredibly fortunate to have you in my life. Thank you for everything.

Contents

1 Introduction **8**
1.1 An abridged history of proof mining 8
1.2 The development of this thesis 13
1.3 A map of the thesis 15
1.4 List of publications and preprints 18

2 Preliminaries **19**
2.1 Weakly extensional Peano arithmetic 19
2.1.1 The formal system 19
2.1.2 Representing real numbers 21
2.1.3 Formalising analysis with abstract types and choice axioms 23
2.2 Logical metatheorems 24
2.2.1 The Dialectica interpretation, negative translation and Spector's bar recursion 25
2.2.2 A program extraction theorem for inner product spaces 28
2.3 Quantitative convergence 31
2.3.1 Quantitative notions of convergence: metastability, fluctuations and crossings 31
2.3.2 The computational hierarchy of the quantitative notions of convergence . 36
2.4 Probability theory 39
2.4.1 Basic notions 39
2.4.2 Martingale theory 44
2.4.3 Probability on Banach spaces 47

3 Non-stochastic proof mining: the computational content of recursive inequalities **50**
3.1 The computational content of a recursive inequality of Alber, Iusem and Solodov 51
3.2 Applications in convex optimization 58

4 Proof-theoretic aspects of probability theory 62
4.1 A formal system for probability theory amenable to program extraction 63
4.1.1 The formal system 64
4.1.2 The program extraction theorem 68
4.2 Quantitative notions of probabilistic convergence 70
4.2.1 Quantitative almost sure statements 70
4.2.2 Learnable rates 80
4.2.3 A proof-theoretic analysis of the relationship between pointwise and uniform metastability 83

5 Proof-theoretic transfer principles and Kronecker's lemma 87
5.1 Quantitative transfer principles 88
5.1.1 A principle for bounded random variables 89
5.2 The computational content of Kronecker's lemma 93
5.2.1 Rates for Kronecker's lemma 93
5.2.2 Computability of rates and the reverse mathematics of Kronecker's lemma 96
5.2.3 A transfer principle for almost surely finite random variables 100

6 Quantitative Laws of Large Numbers 106
6.1 Quantitative and computable aspects of the Laws of Large Numbers 108
6.1.1 Quantitative and computable aspects of the Laws of Large Numbers . . . 108
6.1.2 Quantitative Laws of Large Numbers in the literature 110
6.2 The computational content of Chung's Law of Large Numbers on Banach spaces 112
6.2.1 Rates for the probabilistic Kronecker's lemma 113
6.2.2 Rates for Chung's Law of Large Numbers on Banach spaces 116
6.3 Further quantitative Strong Laws of Large Numbers 122
6.3.1 A general theorem 127
6.3.2 Application I: Rates for pairwise independent random variables with bounded variance 132
6.3.3 Application II: Rates for the Chen-Sung Law of Large Numbers 136

7 Fluctuations in martingales and ergodic averages 141
7.1 Abstract results on probabilistic convergence 142
7.1.1 Crossings, fluctuations and pointwise convergence 143
7.1.2 Crossings, fluctuations and uniform convergence 146
7.2 The computational content of Doob's martingale convergence theorem 151
7.2.1 Learnable uniform rates for the martingale convergence theorem 152
7.2.2 Bounds on the fluctuations for martingales and optimality of rates 155

7.3 The computational content of Birkoff's pointwise ergodic theorem 158
7.3.1 Learnable uniform rates for Birkoff's pointwise ergodic theorem 158
7.3.2 Bounds on the fluctuations of ergodic averages 159
7.3.3 Rates via variational inequalities 162

8 The computational content of the Robbins-Siegmund theorem 165
8.1 The non-stochastic case 166
8.2 The main result 169
8.2.1 Preliminary lemmas 170
8.2.2 Rates for the Robbins-Siegmund theorem 172
8.3 Applications 177
8.3.1 Useful instantiations of the Robbins-Siegmund theorem 177
8.3.2 The Strong Law of Large Numbers 183
8.3.3 Rates for the Robbins-Monro algorithm 184
8.3.4 Rates for Dvoretzky's algorithm 188

9 Future work 193
9.1 Extending the logical foundations of proof mining in probability theory 193
9.2 Further investigation of the relationships between quantitative notions of stochastic convergence 194
9.3 The computational content of the Strong Laws of Large Numbers 195
9.4 Stochastic Fejér monotonicity 197

Chapter 1

Introduction

1.1 An abridged history of proof mining

Proof mining is a program in mathematical logic that aims to extract computational content from proofs, as found in mainstream mathematical literature. Throughout the years, many characters have contributed to the story of proof mining. However, three figures provided crucial paradigm shifts which brought the program to the maturity we enjoy today.

The first main character in the proof mining story is David Hilbert. On the 8th of August 1900, Hilbert presented twenty-three problems at the International Congress of Mathematicians in Paris [59]. The second of these problems, which was titled *Die Widerspruchslosigkeit der arithmetischen Axiome* (which translates to *The compatibility of the arithmetical axioms*), was interested in the axioms of real numbers in arithmetic and asked:

> *"To prove that they are not contradictory, that is, that a definite number of logical steps based upon them can never lead to contradictory results."*[1]

In modern terms, the problem asked for a proof of consistency for an axiomatisation of the real numbers. Hilbert had two formulations of the foundations of mathematics in mind, an uncontroversial *finitistic* system whose consistency was not in question and an extension of such a system that allowed for the use of *infinitary* reasoning, in which one could carry out mainstream mathematics. Hilbert's vision was to demonstrate the consistency of infinitary mathematics in the uncontroversial finitistic system.[2]

In 1931, Gödel published his second incompleteness theorem [50], which proved that any consistent system strong enough to formulate its own consistency (for example, able to formalise a theory of natural numbers) cannot prove its own consistency, let alone any finitistic subsystem. Gödel showed:

[1] This translation was due to Dr. Maby Winton Newson [60].

[2] In [58] Hilbert provides an introduction to his program and the rudiments of modern structural proof theory.

"If c be a given recursive, consistent class of formulae, then the propositional formula which states that c is consistent is not c-provable; in particular, the consistency of P is unprovable in P, it being assumed that P is consistent (if not, of course, every statement is provable)."[3]

The above result was a major blow to Hilbert's program, but it did not completely kill it, as noted by Gödel [50, 53]:

"It must be expressly noted that Proposition XI (and the corresponding results for M and A) represent no contradiction of the formalistic standpoint of Hilbert. For this standpoint presupposes only the existence of a consistency proof effected by finite means, and there might conceivably be finite proofs which cannot be stated in P (or in M or in A)."

However, Gödel's result caused a major emphasis shift in Hilbert's program. The focus of the program was still to obtain consistency proofs for infinitary systems of mathematics; however, the system this proof is carried out in was no longer a finitistic subsystem (and could not be, by Gödel's theorem) but in a system which would be deemed as more trustworthy and finitistic through heuristic and philosophical arguments. Notable examples of these relative consistency proofs came from Ackermann [1], Gentzen [46] and Gödel himself [52].

Though philosophically interesting, the shifted program lacked the precision and rigour of a proper mathematical question: what was meant by a finitistic proof? The second important character in the development of proof mining was Georg Kreisel, who set out to remedy this.

It was clear from his writings [58] that proving the consistency of mathematics was not the end game Hilbert had envisioned for his program. Hilbert was convinced that the use of infinite reasoning with so-called *ideal principles* in mathematics was just an artefact that could be eliminated and replaced by finitistic means (just as infinitesimal calculus was replaced by the more foundationally sound limit approach to calculus):

"It is, therefore, the problem of the infinite in the sense just indicated, which we need to resolve once and for all. Just as in the limit processes of the infinitesimal calculus, the infinite in the sense of the infinitely large and the infinitely small proved to be merely a figure of speech, so too we must realise that the infinite in the sense of an infinite totality, where we still find it used in deductive methods, is an illusion. Just as operations with the infinitely small were replaced by operations with the finite, which yielded exactly the same results and led to exactly the same elegant formal relationships, so in general, must deductive methods based on the infinite be replaced by finite procedures, which yield exactly the same results; i.e., which make

[3]This is a translation due to Meltzer [53] of Satz (Proposition) XI in Gödel's orginal paper [50].

possible the same chains of proofs and the same methods of getting formulas and theorems."[4]

Hilbert believed not only that the consistency of a system allowing for infinitary reasoning (as is done in ordinary mathematics) could be proven by a finitistic subsystem but that the infinitary system was a conservative extension[5] of the finitistic one. That is, a theorem proven by infinitary means could be proven by finitary ones. This is, of course, false by Gödel's theorem; furthermore, the relativised version of this problem (that is, finding philosophically finitistic systems that could replace infinitary reasoning) faced the same lack of precision as the relativised consistency program. Kreisel argued that although the precise notion of a *constructive* proof is in contention, that of a constructive theorem is not, and thus one can formulate the following, more mathematically enticing, program that keeps the spirit of Hilbert's original vision alive:

"To determine the constructive (recursive) content or the constructive equivalent of the nonconstructive concepts and theorems used in mathematics."[6]

What Kreisel meant by *the constructive equivalent of the nonconstructive concepts* is illustrated by the following example: If $A(x, y)$ is quantifier-free in some (suitable) system of arithmetic, then the constructive equivalent of $\forall x\, \exists y\, A(x, y)$ (x and y being variables taking natural numbers) will be $\forall x\, A(x, f(x))$ with f a computable. Here, f represents the constructive content of $\forall x\, \exists y\, A(x, y)$.

To make the notion of constructive equivalence precise, Kreisel introduced the no counter example interpretation (c.f [98, 99]). We omit the general interpretation but note that, in the previous example, f is said to be the solution to the no counterexample interpretation of $\forall x\, \exists y\, A(x, y)$ and if A is primitive recursive then such a solution can be found by unbounded search. Furthermore, if we were to ask about the constructive content of $\forall x\, \exists y\, \forall z\, B(x, y, z)$ with B quantifier-free and x, y, z variables taking natural numbers, a naive interpretation would be $B(x, f(x), z)$ for f computable. However, it is known that there are primitive recursive B in which such an f does not exist. The solution to the no counter example interpretation of $\forall x\, \exists y\, \forall z\, B(x, y, z)$ would instead be a functional $\Phi(x, g)$ satisfying

$$\forall x\, \forall g\, B(x, \Phi(x, g), g(\Phi(x, g))).$$

The above represents a realisation of the Herbrand normal form of $\forall x\, \exists y\, \forall z\, B(x, y, z)$ which can be shown to be equivalent. Furthermore, through the use of very heavy logical machinery,

[4]From a translation of [58] due to Erna Putnam and Gerald J. Massey, which is viewable online at `https://math.dartmouth.edu/~matc/Readers/HowManyAngels/Philosophy/Philosophy.html`

[5]A system T_1 is a conservative extension of a system T_2 if they share the same provable theorems.

[6]Take from the fifth paragraph of [100].

such as the use of epsilon substitution[7] Kreisel was able to show that if a suitable system of arithmetic could prove a formula, then, from the proof, one could compute solutions to the no counter example interpretation that were *ordinal recursive functionals* [99]. From this result, Kreisel was able to obtain a solution to the no counter example interpretation when $\forall x \exists y \forall z \, B(x, y, z)$ was a convergence statement in the case of a monotone bounded sequence. A similar computational interpretation of convergence and the solution for a monotone bounded sequence was later independently discovered by Tao [139, 140] without any logical motivations. Furthermore, Kreisel was able to demonstrate how one could, in principle, get bounds for the first sign change of Littlewood's theorem using the no counter example interpretation [98, 99].

As Kreisel's modification of Hilbert's program started to mature, it became clear that it had the potential to provide a greater impact than just from a foundational perspective. Now more commonly known as the *proof unwinding* program, its aims were summarised by the following quote:

> *"A positive counterpart to the inadequacy result: if not all true propositions about the ring of integers can be formally derived by means of given formal rules, we expect to formulate what more we know about a formally derived theorem F than if we merely know that F is true."*[8]

The aim of the program was to use tools from logic (typically used to prove relative consistency) to analyse proofs to get more information (usually in the form of numerical bounds). A few notable case studies from algebra, number theory and topological dynamics [49, 98, 99, 101] came from the unwinding program with a preference of logical tools being epsilon-substitution and cut-elimination, but the program never reached the heights Kreisel had imagined for it. This was until Ulrich Kohlenbach, the last essential piece of the proof mining story, entered the picture and was able to revive Kreisel's unwinding program by a shift of focus in both areas of application and logical methodology employed.

In 1958, Gödel produced a consistency proof of arithmetic in a quantifier-free system of functionals in higher types known as system T. Gödel's proof was to transform formulas in arithmetic to formulas in system T in such a way that provability was preserved, with $\bot$ remaining fixed under this transformation. Thus, the consistency of arithmetic is reduced to that of system T. This transformation of formulas is now known as the *Dialectica* interpretation (after the journal Gödel's consistency proof appeared in) and falls under the general notion of a recursive proof interpretation as introduced by Kriesel.

Kohlenbach aimed to continue Kreisel's unwinding program (now renamed *proof mining* due to the suggestion of Dana Scott) with an ingenious modification of Gödel's Dialectica

[7] coming from [1, 61].
[8] From page 110 of [101].

interpretation as the main logical tool employed.[9] Concretely, Kohlenbach combined Howard's notion of majorizability [63] and the Dialectica interpretation to obtain a new interpretation (known as the monotone functional interpretation) which asks for majorants (which can be seen as a generalisation of bounds for functionals) for solutions to the Dialectica interpretation. Thus, one can give a computational interpretation to theorems that may not have computable solutions to their Dialectica interpretation by constructing computable majorants, allowing for a wider scope of application. For example, the *weak Kőnig's leamm* (WKL) states that any binary tree (which we code as a set of 0-1 sequences) with infinitely many nodes contains an infinite path. This has a trivial monotone functional interpretation: although the path witnessing this result may be uncomputable, we know that it must be majorised by the constant 1 function. Furthermore, due to the modularity of the Dialectica interpretation, one can then treat any theorems that use WKL in their proofs, which includes various results in analysis as seen in the reverse mathematics program [135].

Since the monotone functional interpretation only extracts majorants and not witnesses, the method appears to be limited. However, Kohlenbach recognised that this was not a problem in analysis, as opposed to number theory and algebra, which were the main areas of interest of Kreisel's unwinding program:

> *"The monotonicity of many quantifiers occurring in analysis has always been one reason why I considered analysis as a particularly fruitful area of mathematics for 'proof mining' (whereas Kreisel's original 'unwinding' program mainly aimed at number theory and algebra)."*[10]

Typically, in analysis, bounds are just as good as exact witnesses, and this, combined with the fact that one could deal with compactness arguments due to the trivial interpretation of WKL made analysis an exciting prospect for Kohlenbach's proof mining program with this hope paying dividends in the extraction of computational data from results in approximation theory [77, 78, 90].

Kohlenbach continued to expand the logical tools available for the extraction of the computational content of proofs, such as the use of Spector's bar recursion [138] to treat proofs that made use of strong choice principles (typically associated with compactness arguments). However, a crucial observation made by Kohlenbach brought proof mining into the form we observe today.

During the development of the logical foundations of the proof mining program, Kohlenbach initially worked in the standard theory of Peano arithmetic in all finite types, and this meant that he could only work with spaces that could be represented in the system. Thus, the

[9] Unlike the formalisms originally used by Kreisel and his predecessors which were based on first, second and third order logic, Kohlenbach worked in systems of analysis in all finite types.

[10] From [83].

applications were restricted to such representable spaces, with Polish spaces as the most natural example. This obstacle was overcome through the introduction of so-called *abstract types* that could represent arbitrary spaces, and through the use of Bezem's model [14], Kohlenbach was able to obtain the first proof mining metatheorem [79]. This metatheorem showed that one could extract the computational content of a large class of theorems in analysis on nonrepresentable spaces whose proofs used strong (seemingly nonconstructive) principles. Another important aspect of Kohlenbach's metatheorem was that, for the first time, it could explain (logically) the empirical observation of the uniformity (in the spaces and certain parameters of the result) of the extracted computational content observed in analysis.

This initial work of Kohlenbach has been expanded in both applications and the development of foundational tools, resulting in the publication of hundreds of papers (we direct the reader to the proof mining bibliography, currently maintained by Nicholas Pischke[11]). More explicit details on the topics discussed in this brief introduction can be found in Kohlenbach's book [80] which provides a comprehensive overview of the state of proof mining up until 2008. One can turn to the survey papers [81, 82] for details on the recent progress of proof mining in nonlinear analysis and optimization.

1.2 The development of this thesis

Following the initial success of proof mining in the 1990s and early 2000s by Kohlenbach in approximation theory, this approach rapidly expanded into other areas of nonlinear analysis, with a particular emphasis on fixed point theory and optimization. Since the late 2000s, there have been sporadic case studies applying proof mining in areas outside of analysis, notably in algebra [134], Tauberian theory [123, 124], combinatorics [47, 102], and probability (measure) theory [3, 4, 5, 6, 7, 8]. Among these areas, probability theory showed the most promise due to the influential work initially championed by Avigad, Dean, Gerhardy, Rute, and Towsner. However, since that time, the progress of the proof mining program in probability theory has remained stagnant for about ten years. This changed with the publication of a paper by Arthan and Oliva [3], which reignited interest in the field. This thesis will detail my and my collaborators' efforts during my doctoral studies to further develop the proof mining program in probability theory.

Under the supervision of Thomas Powell, my initial research focused on applying the logical tools from the proof mining program to convex optimization in abstract spaces. The observation was made that many convergence results associated with this area relied on reasoning about the convergence of sequences of real numbers that satisfied recursive inequalities. This observation resulted in the joint work with Powell [115], in which we investigated the computational content

[11] `https://sites.google.com/view/nicholaspischke/proof-mining-bibliography/chronological`.

of a general recursive inequality that found itself in a vast number of applications.

The joint work with Powell, previously mentioned, aligned well with the survey paper by Franci and Grammatico [42], which offers a comprehensive overview of the application of recursive inequalities across various areas of analysis. This survey also includes an extensive review of stochastic recursive inequalities and their applications. Consequently, a natural progression from the initial project [115] was to explore the computational aspects of the convergence results related to these stochastic recursive inequalities, with the aim of furthering the proof mining program in stochastic optimization.

The most influential of the stochastic recursive inequality that appeared in [42] was the Robbins-Siegmund theorem [129], which was obtained from the celebrated martingale convergence theorem of Doob [33]. Thus, to make progress in proof mining in stochastic optimization, investigating the computational content of Doob's theorem was a clear priority, and we hoped that this would lead to a computational interpretation of the Robbin-Siegmund theorem.

The search for a computational interpretation of the Robbins-Siegmund theorem (which was obtained and presented in [116]) resulted in significant progress in proof mining in probability theory. In order to obtain a computational interpretation of the martingale convergence theorem (the key result in proving the Robbins-Siegmund theorem), Powell and I had to develop many of the ideas explored by Avigad and his collaborators. This effort culminated in the work presented in [117].

In addition, while addressing the problem of finding a computational interpretation of the Robbins-Siegmund theorem, I was already considering potential applications of this result. The first application of the Robbins-Siegmund theorem, presented in [129], was related to the Laws of Large Numbers. This prompted me to explore the computational properties of the Laws of Large Numbers, leading to the single-authored papers [112, 113].

I was not only interested in applications but also in developing the logical foundations of probability theory within the proof mining program. Specifically, I aimed to create a logical metatheorem, similar to those in [79], that would explain the success behind the increasing number of case studies extracting the computational content of results in probability theory.

During the summer of 2023, I was invited to speak at the European Summer Meeting of the Association of Symbolic Logic (Logic Colloquium). I presented my research on the Laws of Large Numbers and discussed progress made toward obtaining a computational interpretation of the Robbins-Siegmund theorem in collaboration with Powell. At this meeting, I had the opportunity to meet Nicholas Pischke, to whom I mentioned my interest in developing a metatheorem for probability theory. To my surprise, Pischke revealed that he had also been considering this topic. We decided to collaborate to tackle this problem together, which resulted in [114].

This thesis shall provide technical details of my and my collaborators' efforts in developing

proof mining in probability theory during my doctoral studies. We start with the motivating investigation of deterministic recursive inequalities and end with the computational interpretation of the Robbins-Siegmund theorem, detailing all the developments made along the way. We note that proof mining in probability theory continues to develop, even at the time of writing this thesis, and I am both proud and excited about the progress being made.

1.3 A map of the thesis

We now give a brief outline of the thesis, including details about the collaborations involved in its construction.

Chapter 2 consists of standard background material.

- Section 2.1 introduces the formal system of analysis in which the systems used in proof mining (including those in this thesis) are built on top. We closely follow [80].
- Section 2.2 provides an introduction to some of the logical tools used in proof mining, including Gödel's Dialectica interpretation and the statement of a metatheorem for inner product spaces. We closely follow [55, 79, 80].
- Section 2.3 introduces standard notions of quantitative deterministic convergence and the relationships between them. Most of the results we present are folklore in the applied proof theory literature with Proposition 2.3.20 a seemingly new result[12] appearing in a joint work with Thomas Powell for which a preprint is available in [117].
- Section 2.4 introduces the notions of probability theory we need in this thesis. The notions from probability theory we introduce are standard and our primary references are [35, 56]; for martingale theory we also refer to [142], and our primary reference for the theory of random variables taking values in Banach spaces is [105].

Chapter 3 consists of an illustrative example of non-stochastic proof mining. We investigate the computational content of a convergence result of real numbers due to Alber [2] as well as applications of this analysis to quantitative results in convex optimization. Lastly, we briefly discuss how our analysis falls under the proof mining metatheorem we introduced in Chapter 2. This chapter is part of a joint work with Thomas Powell and was published in [115].

Chapter 4 contains our theoretical contributions to proof mining in probability theory.

- Section 4.1 introduces a formal system for reasoning about probability contents and a metatheorem guaranteeing the existence and uniformity of computational content for a

[12] Although the result has been implicitly applied in the literature, namely [6, 70].

large class of results in probability theory. This section is part of a joint project with Nicholas Pischke and can be found in the preprint [114].[13]

- Section 4.2 presents an abstract framework for dealing with quantitative aspects of probability theory, and from our framework, we are able to motivate the notions of quantitative almost sure convergence introduced in the seminal papers of proof mining in probability theory [5, 6] as well as the mainstream quantitative probability theory literature [21, 70]. This part of the section was a joint work with Thomas Powell and can be found in the preprint [117].

  Later in the section, we provide a logical explanation for the complexity and uniformity found in the quantitative version of Egorov's theorem given in [5] by analysing the result in the formal system introduced earlier in the chapter. This analysis was part of a joint work with Nicholas Pischke and can be found in the preprint [114].

Chapter 5 presents a computational investigation of Kronecker's lemma, which is a crucial result in obtaining many results concerned with the Laws of Large Numbers. Motivated by Kronecker's lemma, we also provide general proof theoretical transfer results that allow for lifting computational content from deterministic theorems to their probabilistic analogue.

- Section 5.1 introduces a proof theoretical transfer principle that allows for lifting the computational content from results about sequences of real numbers to analogous probabilistic results. This section is part of a joint work with Nicholas Pischke and can be found in the preprint [114].

- Section 5.2 provides a generalisation of the transfer result presented in Section 5.1 motivated by a quantitative analysis of Kronecker's lemma. Furthermore, we investigate Kronecker's lemma from the perspective of the reverse mathematics program [135]. This section can be found in the single-authored preprint [112].

Chapter 6 contains our contributions to quantitative results concerning the Strong Law of Large Numbers.

- Section 6.2 presents our quantitative analysis of Chung's Law of Large Numbers [26] generalised to Banach space valued random variables [143]. This section can be found in the single-authored preprint [112].

[13] A lot of the technical details in the proof of the metatheorem was mainly due to Pischke, and we have chosen to omit the proof of the metatheorem in this thesis. However, the author was heavily involved in the development of the other results of [114], such as the logical explanation of the success and uniformities of the main result in [5], and we discuss some of these results in this thesis.

- Section 6.3 provides a quantitative generalisation of [29] from which many other results in the Strong Law of Large Numbers literature form special cases. Furthermore, we use our general quantitative result to improve the bound found in [110]. This section can be found in the single-authored preprint [113].

- Section 6.1 provides an example demonstrating the computational ineffectiveness of the Strong Laws of Large Numbers. This section can be found in the single-authored preprint [112].

Chapter 7 uses the theoretical framework developed in Chapter 4 to provide quantitative results for martingale and ergodic theory. This chapter forms part of a joint work with Thomas Powell found in the preprint [117].

- Section 7.1 provides abstract quantitative results concerning stochastic crossings, fluctuations and convergence from which our results for martingale theory and ergodic theory will follow.

- Section 7.2 contains our quantitative results for the martingale convergence theorem, including results that generalise those found in [21] and [70].

- Section 7.3 contains our quantitative results for the pointwise ergodic theorem, including results that generalise those found in [6, 70].

Chapter 8 provides a computational interpretation of the Robbins-Siegmund theorem [129] and applications. This chapter was done jointly with Thomas Powell, and many of the results can be found in the preprint [116] and in upcoming work with Thomas Powell and Nicholas Pischke.

- Section 8.1 provides a quantitative version of the deterministic version of the Robbins-Siegmund due to Qihou [126]. A similar result was already obtained in [85]; however, the simplified analysis we present lifts more naturally to the stochastic case.

- Section 8.2 presents our quantitative analysis of the Robbins-Siegmund theorem. This section includes many useful axillary lemmas.

- Section 8.3 provides applications of our quantitative Robbins-Siegmund theorem, including an illustrative example for in the Strong Law of Large Numbers (the rates obtained are not better than those presented in Chapter 6), a quantitative version of the Robbins-Monro procedure [128] and a quantitative version of Dvoretzky's Theorem [36].

Chapter 9 contains a brief discussion about current work in progress to extend many of the topics discussed in this thesis, including details about future collaborations with Pischke and Powell.

1.4 List of publications and preprints

This thesis contains research from the following publications and preprints:

- [115] **A computational study of a class of recursive inequalities**

  Journal of Logic and Analysis 5, 3 (2023), 1–48. (With Thomas Powell)

- [113] **Quantitative Strong Laws of Large Numbers**

  Electronic Journal of Probability 30 (2025), 1-22

- [112] **A finitary Kronecker's lemma and large deviations in the Strong Law of Large Numbers on Banach spaces**

  Annals of Pure and Applied Logic 176, 6 (2025), 103569

- [117] **On quantitative convergence for stochastic processes: Crossings, fluctuations and martingales**

  To be published in Transactions of the American Mathematical Society. Available at `https://arxiv.org/abs/2406.19979`*, 2025.* (With Thomas Powell)

- [114] **Proof mining and probability theory** *Preprint, available at* `https://arxiv.org/abs/2403.00659`*, 2024.* (With Nicholas Pischke)

- [116] **A quantitative Robbins-Siegmund theorem**

  Preprint, available at `https://arxiv.org/abs/2410.15986`*, 2024.* (With Thomas Powell)

Chapter 2

Preliminaries

This thesis explores the applications of proof theory to quantitative probability theory, so familiarity with these fields is necessary. This chapter provides an overview of the key concepts required to follow the discussions in the subsequent chapters.

The first two sections of this chapter outline the necessary logical preliminaries. These sections serve as motivation for many discussions in Chapter 4. The third section reviews quantitative concepts from nonstochastic analysis, which are extended to the stochastic context in this thesis. This section also serves to motivate various concepts and results presented in Section 4.2. The final section of this chapter addresses key notions from probability theory that we utilise throughout the thesis. Additionally, we briefly cover essential ideas from martingale theory and the theory of probability on Banach spaces, which will be crucial for Chapters 7 and 6, respectively.

We do not claim originality for any of the results presented in this chapter.

2.1 Weakly extensional Peano arithmetic

We begin by providing a brief overview of *weakly extensional Peano arithmetic* in all finite types. This system serves as the foundation for many formal systems utilised in proof mining, including the system for reasoning about probability theory that we introduce in Section 4.1. All the concepts we discuss are standard, and this section closely follows the work presented in [80].

2.1.1 The formal system

We start by fixing a set of types:

Definition 2.1.1. The set of types T is defined inductively via,

$$0 \in \mathrm{T}, \quad \rho, \tau \in \mathrm{T} \to \rho(\tau) \in \mathrm{T}.$$

Objects of type 0 are meant to be interpreted as natural numbers and objects of type $\rho(\tau)$ as mappings from objects of type τ to objects of type ρ. We use natural numbers to denote pure types. That is, we write $n+1 := 0(n)$.

Definition 2.1.2. The *degree*, $\deg(\rho)$, of a type ρ is defined by recursion as

$$\deg(0) := 0, \quad \deg(\rho(\tau)) := \max\{\deg(\tau), \deg(\rho)+1\}.$$

WE-PA$^\omega$ is built on top of many-sorted first-order classical logic (so we assume we have the standard logical connectives in our language), with a set of sorts (or types) T. So, for each $\rho \in \mathrm{T}$ we have variables, $x^\rho, y^\rho, z^\rho \ldots$ and quantifiers over them. The only primitive relation symbol is $=_0$, which represents equality at type 0, with equality for higher types defined as an abbreviation via,

$$x^{\tau(\xi)} =_{\tau(\xi)} y^{\tau(\xi)} := \forall z^\xi \, (xz =_\tau yz) \, .$$

Furthermore, for each $\sigma, \rho, \tau, \rho_1, \ldots, \rho_k, \in \mathrm{T}$ and $1 \leq i \leq k$ we have constants (here we write $\underline{\rho} := (\rho_1) \ldots (\rho_k)$ and $\underline{\rho}^t := (\rho_k) \ldots (\rho_1)$):

Constant	Type	Interpretation
0	0	Zero
S	$0(0)$	Successor
$\Pi_{\rho,\tau}$	$\rho(\tau)(\rho)$	Projector combinator, introduced by Schönfinkel [131]
$\Sigma_{\delta,\rho,\tau}$	$\tau\delta(\rho\delta)(\tau\rho\delta)$	Combinator introduced by Schönfinkel [131]
$(R_i)_{\underline{\rho}}$	$\rho_i(\rho_k 0 \underline{\rho}^t) \ldots (\rho_1 0 \underline{\rho}^t)\underline{\rho}^t 0$	Simultaneous primitive recursion [52, 80]

Now, as is standard in many-sorted first-order logic, terms are generated by variables x^ρ of type $\rho \in \mathrm{T}$, constants c^ρ of type $\rho \in \mathrm{T}$ and via the recursive construction that if $t^{\rho(\tau)}$ is a term of type $\rho(\tau) \in \mathrm{T}$ and s^τ is a term of type $\tau \in \mathrm{T}$, then $(ts)^\rho$ is a term of type $\rho \in \mathrm{T}$. Furthermore, for terms $t, s_1, \ldots s_n$ we usually write $t(s_1, \ldots, s_n)$ instead of $(\ldots(ts_1)\ldots s_n)$. Formulas are built from atomic formulas (formulas of the form $t =_0 s$ for terms t^0, s^0 of type 0) and using logical connectives as standard. Throughout the thesis, variables that are underlined will denote tuples of variables.

To get WE-PA$^\omega$, we extend our current system with axioms expressing that $=_0$ is an equivalence relation (reflexivity, symmetry and transitivity), the usual successor axioms and the

axiom scheme of complete induction:

$$A(0) \wedge \forall n^0(A(n) \to A(S(n))) \to \forall n^0 A(n) \tag{IA}$$

for all formulas $A(n^0)$. Furthermore, we include defining axioms for the combinators and the recursion constants. See [80] for explicit details of these axioms.

Remark 2.1.3. An important consequence of the inclusion of the combinators $\Sigma_{\delta,\rho,\tau}$ and $\Pi_{\rho,\tau}$ and their defining axioms in our system is that they allow for the definition of λ-abstraction. That is, for any term t of type τ and any variable x of type ρ, we can construct a term $\lambda x.t$ of type $\tau(\rho)$ such that the free variables of $\lambda x.t$ are exactly those of t without x. Furthermore,

$$\mathsf{WE\text{-}PA}^\omega \vdash (\lambda x.t)(s) =_\tau t[s/x]$$

for any term s of type ρ.

Lastly, we have the following rule of quantifier-free extensionality:

$$\frac{A_0 \to s =_\rho t}{A_0 \to r[s/x^\rho] =_\tau r[t/x^\rho]} \tag{QF-ER}$$

where A_0 is a quantifier-free formula, s and t are terms of type ρ and r is a term of type τ.

Remark 2.1.4. Crucially, we do not include the full extensionality axiom

$$\forall x^{\tau(\rho)}, y^\rho, y'^\rho \, (y =_\rho y' \to xy =_\tau xy') \,, \tag{$\mathsf{E}_{\rho,\tau}$}$$

as this would not allow for a result on program extraction, as that presented in Theorem 2.2.14.

Remark 2.1.5. *Weakly extensional Heyting arithmetic* in all finite types, $\mathsf{WE\text{-}HA}^\omega$, is defined similarly to $\mathsf{WE\text{-}PA}^\omega$ except it is built on top of intuitionistic many-sorted first-order logic (which is classical logic with the removal of the principle of excluded middle axiom $A \vee \neg A$, for all formulas A).

2.1.2 Representing real numbers

We access the real numbers in $\mathsf{WE\text{-}PA}^\omega$ through their representation as a Polish space in the system, as in Section 4 of [80].

By first representing natural numbers as objects of type 0, we represent rational numbers as codes for pairs of natural numbers using a canonical pairing function j. Concretely, we take

$$j(n^0, m^0) := \begin{cases} \min u \leq_0 (n+m)^2 + 3n + m[2u =_0 (n+m)^2 + 3n + m] & \text{if existent,} \\ 0^0 & \text{otherwise.} \end{cases}$$

Through terms that operate on such codes, we can primitively recursively define the usual operators $+_{\mathbb{Q}}, \cdot_{\mathbb{Q}}, |\cdot|_{\mathbb{Q}}$, etc., furthermore, the usual relations $=_{\mathbb{Q}}$, $<_{\mathbb{Q}}$, etc., are definable via quantifier-free formulas.

We now represent the reals via fast converging Cauchy sequences, with a fixed modulus of convergence. Concretely, using our coding of the rationals, we can interpret an object $f^{0(0)}$ of type $0(0)$ as a sequence of rational numbers, and we represent the reals as those sequences satisfying

$$\forall n^0 |f(n) -_{\mathbb{Q}} f(n+1)|_{\mathbb{Q}} \leq_{\mathbb{Q}} [2^{-n-1}]_{\mathbb{Q}}$$

where for a rational number r, $[r]_{\mathbb{Q}}$ represents the object of type 0 that codes for it. To improve readability, this is usually omitted when the context is clear.

To allow us to quantify over such fast converging sequences and, thus, the reals implicitly, we introduce the operator $\widehat{\cdot}$ turning f of type 1 into a fast-converging Cauchy sequence $\widehat{f}$ via

$$\widehat{f}(n) := \begin{cases} f(n) & \text{if } \forall k <_0 n \left(|f(k) -_{\mathbb{Q}} f(k+1)|_{\mathbb{Q}} <_{\mathbb{Q}} 2^{-k-1}\right), \\ f(k) & \text{for } k <_0 n \text{ least with } |f(k) -_{\mathbb{Q}} f(k+1)|_{\mathbb{Q}} \geq_{\mathbb{Q}} 2^{-k-1} \text{ otherwise.} \end{cases}$$

One can show that using $\widehat{\cdot}$ ensures that each type 1 object codes a unique real number. That is, it is the case that if f^1 represents a fast converging Cauchy sequence as defined above, then $\forall n^0(f(n) =_0 \widehat{f}(n))$. Unlike the rationals, the standard relations on the reals are not given by quantifier-free formulas. Instead, we have equality defined by the following Π^0_1 formula

$$f_1 =_{\mathbb{R}} f_2 :\equiv \forall n^0(|\widehat{f_1}(n+1) - \widehat{f_2}(n+1)| <_{\mathbb{Q}} 2^{-n}).$$

Similarly $<_{\mathbb{R}}$ and $\leq_{\mathbb{R}}$ are defined by Σ^0_1 and Π^0_1 formulas respectively. Furthermore, we can embed $\mathbb{N}$ and $\mathbb{Q}$ in $\mathbb{R}$ via constant sequences and the usual operations on $\mathbb{R}$ such as $+_{\mathbb{R}}$, $\cdot_{\mathbb{R}}$, $|\cdot|_{\mathbb{R}}$, etc., are primitively recursively definable.

We follow the standard convention that whenever the context is clear, we will omit the subscripts of the arithmetical operations for $\mathbb{R}$ and $\mathbb{Q}$. Furthermore, again, when the context is clear, we will omit types of variables, and we omit the operation $\cdot_{\mathbb{R}}$ altogether, as is standard mathematical practice.

Similarly, one can represent a general Polish space in $\mathsf{WE\text{-}PA}^{\omega}$. We omit the details as they do not serve any purpose in this thesis.

2.1.3 Formalising analysis with abstract types and choice axioms

To gain access to real analysis (or, more generally, analysis on Polish spaces), one can extend $\mathsf{WE\text{-}PA}^\omega$ with the following choice principles:

$$\forall \underline{x} \exists \underline{y} A_0(\underline{x}, \underline{y}) \to \exists \underline{Y} \forall \underline{x} A_0(\underline{x}, \underline{Y}\underline{x}) \tag{QF-AC}$$

$$\forall x^0, \underline{y}^{\underline{\rho}} \exists \underline{z}^{\underline{\rho}} A(x, \underline{y}, \underline{z}) \to \exists \underline{f}^{\underline{\rho}(0)} \forall x^0 A(x, \underline{f}(x), \underline{f}(S(x))) \tag{DC$^{\underline{\rho}}$}$$

The former is the *quantifier-free axiom of choice* schema in all types, with A_0 quantifier-free and the tuples of variables $\underline{x}$, $\underline{y}$ can take arbitrary types. The latter is the principle of *dependent choice* (we denote the collection $\mathsf{DC}^{\underline{\rho}}$ for all tuples of types $\underline{\rho}$ as DC) where $\underline{f}^{\underline{\rho}(0)}$ stands for $f_1^{\rho_1(0)}, \ldots, f_k^{\rho_k(0)}$ and A may now be arbitrary.

Denote by $\mathcal{A}^\omega := \mathsf{WE\text{-}PA}^\omega + \mathsf{QF\text{-}AC} + \mathsf{DC}$ the system $\mathsf{WE\text{-}PA}^\omega$ along with the quantifier-free axiom of choice schema and principle of dependent choice.

Remark 2.1.6. DC implies countable choice, so we have arbitrary comprehension over natural numbers. Therefore, full second-order arithmetic (in the sense of that used in reverse mathematics [135]) can be embedded in $\mathcal{A}^\omega$ (identifying subsets of $\mathbb{N}$ with their characteristic function).

As mentioned at the end of Section 2.1.2, we can represent general Polish spaces in $\mathcal{A}^\omega$ via fast converging Cauchy sequences, similar to the representation of the reals presented. It was the insight of Kohlenbach [48, 79, 80] that one could reason about more general spaces, not representable in $\mathcal{A}^\omega$ by introducing abstract types. This approach will be crucial in the system we present for probability spaces, which we introduce in Section 4.1. Here, we demonstrate how to represent arbitrary normed and inner product spaces in a system that extends $\mathcal{A}^\omega$ with abstract types.

We first introduce a system for reasoning about a normed space $(X, \|\cdot\|)$, which we denote by $\mathcal{A}^\omega[X, \|\cdot\|]$ (this system was first introduced in [79]). This is done by defining a new set of types, T^X, which is the extension of T with two ground types 0 and X. That is, we define T^X inductively via,

$$0, X \in \mathrm{T}^X, \quad \rho, \tau \in \mathrm{T}^X \to \rho(\tau) \in \mathrm{T}^X.$$

We then reformulate $\mathcal{A}^\omega$ over the new set of types T^X, where we have additional constants and additional axioms that now refer to the additional types. Over this new reformulation of $\mathcal{A}^\omega$, we add the constants:

Constant	Type	Interpretation
0_X	X	Zero vector
1_X	X	An arbitrary unit vector
$\|\cdot\|$	$1(X)$	The norm
$+_X$	$X(X)(X)$	Vector addition
$-_X$	$X(X)$	Additive inverse operator
$\cdot_X$	$X(X)(1)$	Scaler multiplication

In addition, we include axioms specifying that X with these operations specify a real normed vector space (we will have axioms specifying that the above constants represent what they are supposed to, as given in the 'Interpretation' column of the above table, the exact axioms can be found in [79]). To obtain a system for reasoning about inner product spaces, we do not need to introduce any further constants; we just need to note the characterising property of a normed space being an inner product space; that is, it satisfies the parallelogram identity. Concretely, the system $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ for real inner product spaces is $\mathcal{A}^\omega[X, \|\cdot\|]$ extended with the axiom

$$\forall x^X, y^X(\|x +_X y\|_X^2 +_{\mathbb{R}} \|x -_X y\|_X^2 =_{\mathbb{R}} 2(\|x\|_X^2 +_{\mathbb{R}} \|y\|_X^2))$$

and we define the inner product via the abbreviation

$$\langle x^X, y^X\rangle := \frac{\|x +_X y\|_X^2 -_{\mathbb{R}} \|x -_X y\|_X^2}{4}.$$

Furthermore, we define equality on X in $\mathcal{A}^\omega[X, \|\cdot\|]$ and $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ via the abbreviation

$$x^X =_X y^X :\equiv \|x -_X y\|_X =_{\mathbb{R}} 0.$$

Thus, equality is given by a Π_1^0 formula (recalling the definition of $=_{\mathbb{R}}$ given in Section 2.1.2). Lastly, one can prove that the operations defined by the constants in the above table and $\langle\cdot,\cdot\rangle$ are extensional with respect to $=_X$.

2.2 Logical metatheorems

The core logical foundations of proof mining are the so-called general logical metatheorems on bound extraction. These metatheorems employ established proof interpretations, such as Gödel's *functional (Dialectica) interpretation* [52], to offer broad results that quantify and enable the extraction of computational content from wide classes of theorems and proofs within their intended fields of application. Over the past three decades, proof mining, as supported by these metatheorems, has generated hundreds of new results across various application areas.

In this section, we briefly introduce the Dialectica interpretation and discuss its combination

with other logical techniques used in proof mining, in general, and this thesis. Our main reference will be [80], but an effort has been made to give further references when appropriate.

2.2.1 The Dialectica interpretation, negative translation and Spector's bar recursion

Extensions of the Dialectica interpretation of Gödel [52] are the main logical tools used in proof mining. The purpose of Gödel's original interpretation was to demonstrate the consistency of Heyting Arithmetic (HA, constructive arithmetic) relative to a system, known as system T,[1] that was argued to be more trustworthy. The Dialectica interpretation did this by transforming a formula provable in HA into one provable in T; in particular, the interpretation of $\bot$ remains itself. In the context of proof mining, the Dialectica interpretation is used to provide program extraction theorems, which guarantee the extraction of rates for theorems that are provable in strong classical theories, as well as a way to reformulate infinitary statements into finitary ones from which computational interpretations can be given.

We now present an extension of the Dialectica interpretation of formulas in $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$.

Definition 2.2.1 ([52, 80, 141]). The Dialectica interpretation $A^D = \exists \underline{x} \forall \underline{y} A_D(\underline{x}, \underline{y})$ of a formula A in the language of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ is defined via recursion on the structure of the formula:

1. $A^D := A_D := A$ for A being a prime formula.

If $A^D = \exists \underline{x} \forall \underline{y} A_D(\underline{x}, \underline{y})$ and $B^D = \exists \underline{u} \forall \underline{v} B_D(\underline{u}, \underline{v})$, we set:

2. $(A \wedge B)^D := \exists \underline{x}, \underline{u} \forall \underline{y}, \underline{v} (A \wedge B)_D$
   where $(A \wedge B)_D(\underline{x}, \underline{u}, \underline{y}, \underline{v}) := A_D(\underline{x}, \underline{y}) \wedge B_D(\underline{u}, \underline{v})$.

3. $(A \vee B)^D := \exists z^0, \underline{x}, \underline{u} \forall \underline{y}, \underline{v} (A \vee B)_D$
   where $(A \vee B)_D(z^0, \underline{x}, \underline{u}, \underline{y}, \underline{v}) := (z = 0 \to A_D(\underline{x}, \underline{y})) \wedge (z \neq 0 \to B_D(\underline{u}, \underline{v}))$.

4. $(A \to B)^D := \exists \underline{U}, \underline{Y} \forall \underline{x}, \underline{v} (A \to B)_D$
   where $(A \to B)_D(\underline{U}, \underline{Y}, \underline{x}, \underline{v}) := A_D(\underline{x}, \underline{Y}\underline{x}\underline{v}) \to B_D(\underline{U}\underline{x}, \underline{v})$.

5. $(\exists z^\tau A(z))^D := \exists z, \underline{x} \forall \underline{y} (\exists z^\tau A(z))_D$
   where $(\exists z^\tau A(z))_D(z, \underline{x}, \underline{y}) := A_D(\underline{x}, \underline{y}, z)$.

6. $(\forall z^\tau A(z))^D := \exists \underline{X} \forall z, \underline{y} (\forall z^\tau A(z))_D$
   where $(\forall z^\tau A(z))_D(\underline{X}, z, \underline{y}) := A_D(\underline{X}z, \underline{y}, z)$.

The Dialectica interpretation, as presented above, already allows for program extraction for theorems that can be proven constructively.

[1] System T is just the quantifier-free fragment of WE-HA$^\omega$.

Theorem 2.2.2. *Let A be a formula in the language of* WE-HA$^\omega$ *only containing the (potentially empty) tuple $\underline{a}$ as free variables. Then*

$$\textsf{WE-HA}^\omega \vdash A(\underline{a}) \text{ implies } \textsf{WE-HA}^\omega \vdash \forall a, y (A)_D(\underline{ta}, \underline{y}, \underline{a}).$$

In the above, $\underline{t}$ is a tuple of closed terms in the language of WE-HA$^\omega$ *that can be extracted from a proof of A. We call such a t a solution to the Dialectica interpretation of A.*

Remark 2.2.3. The above theorem does not hold for full classical arithmetic in all finite types, WE-PA$^\omega$. However, it does hold for certain semi-intuitionistic fragments of WE-PA$^\omega$; in particular, the equivalence of a formula to its Dialectica interpretation is provable in such a semi-intuitionistic fragment. Furthermore, the above soundness theorem also holds for suitable extensions of the language of WE-HA$^\omega$ (e.g. by any kind of new types and constants) together with any additional universal axioms in that language (this is because universal formulas have trivial Dialectica interpretation). In particular, the theorem holds for the intuitionistic fragment of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ (replacing WE-PA$^\omega$ by WE-HA$^\omega$ in the construction of $\mathcal{A}^\omega$) with the choice principles removed.

Example 2.2.4. If $A :\equiv \forall x^0 \exists y^0 B(x, y)$ (for a quantifier free formula B), then $A^D \equiv \exists F^1 \forall x^0 B(x, Fx)$. Thus, if A is provable in WE-HA$^\omega$, the (proof of) the soundness theorem provides an algorithm to extract a function $F : \mathbb{N} \to \mathbb{N}$ such that $\forall x^0 B(x, Fx)$.

To extend the Dialectica interpretation to classical arithmetic, we need a so-called negative translation. These translations take formulas provable in PA and output a formula (equivalent over PA) that is provable in HA (thus demonstrating the equiconsistency of PA and HA). The first of these translations was due to Kolmogorov [93] (with similar variants discovered independently by Gentzen [45] and Gödel [51]), and many such translations have been developed since the original. For our purposes, we use an extension of the negative translation of Kuroda [104].

Definition 2.2.5 ([104]). The *negative translation* of A is defined by $A' := \neg\neg A^*$ where A^* is defined by the following recursion on the structure of A:

1. $A^* := A$ for prime A.
2. $(A \circ B)^* := A^* \circ B^*$ for $\circ \in \{\wedge, \vee, \to\}$.
3. $(\exists x^\tau A)^* := \exists x^\tau A^*$.
4. $(\forall x^\tau A)^* := \forall x^\tau \neg\neg A^*$.

We have the following useful characterisation of the negative translation:

Theorem 2.2.6. *Let A be a formula in the language of* WE-HA$^\omega$. *Then*

$$\textsf{WE-PA}^\omega \vdash A \text{ implies } \textsf{WE-HA}^\omega \vdash A'.$$

Thus, combining the negative translation and the Dialectica interpretation allows for program extraction for theorems provable in WE-PA$^\omega$ as well as $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ with the choice principles removed (c.f. Remark 2.2.3).

Remark 2.2.7. If A is a formula in the language of WE-HA$^\omega$ with Dialectica interpretation $A^D = \exists \underline{x} \forall \underline{y} A_D(\underline{x}, \underline{y})$, then from the way one treats implication one has that

$$(\neg\neg A)^D \equiv \exists X\, \forall Y \neg\neg A_D(\underline{XY}, \underline{Y}(\underline{XY})) \tag{2.1}$$

with the above provably equivalent to

$$\exists X\, \forall Y\, A_D(\underline{XY}, \underline{Y}(\underline{XY})) \tag{2.2}$$

in WE-HA$^\omega$. Now, define,

$$P \equiv \forall x\, \exists y\, \forall z\, A(x, y, z).$$

with $A(x, y, z)$ a prime formula of WE-HA$^\omega$. We will have

$$P' \equiv \neg\neg\forall x\, \neg\neg\exists y\, \forall z\, \neg\neg A(x, y, z) \leftrightarrow \forall x\, \neg\neg\exists y\, \forall z\, A(x, y, z)$$

with the equivalence provable in WE-HA$^\omega$. Now

$$(\exists y\, \forall z\, A(x, y, z))^D \equiv \exists y\, \forall z\, A(x, y, z)$$

and so (2.1) implies

$$(\neg\neg\exists y\, \forall z\, A(x, y, z))^D \equiv \exists Y\, \forall Z\, A(x, Y(Z), Z(Y(Z))).$$

This yields

$$(P')^D \equiv \exists \Phi\, \forall x, Z\, A(x, \Phi(x, Z), Z(\Phi(x, Z))).$$

Therefore, Theorem 2.2.2 and Theorem 2.2.6 imply that if P is provable in WE-HA$^\omega$ then from such a proof we can extract a functional Φ such that

$$\forall x, Z\, A(x, \Phi(x, Z), Z(\Phi(x, Z))).$$

Such a Φ will represent the computational content of P. In the situation where P is a suitable formulation of Cauchy convergence $(P')^D$ gives rise to a notion known as metastable convergence

(c.f. Definition 2.3.4) and was independently discovered by Tao [139].

We now discuss how to obtain program extraction for theorems in analysis, that is, proofs that make use of compactness through DC. To do this, one needs (an extension of) a construction due to Spector [138], known as *bar recursion*. Just as the Dialectica interpretation was originally introduced to demonstrate the relative consistency of arithmetic, bar recursion was used to demonstrate the relative consistency of analysis by providing a Dialectica interpretation of DC. To this effect consider the extension of $\mathsf{WE\text{-}PA}^\omega$ where for each tuple of types $\underline{\rho} := \rho_1 \dots \rho_k$ and $\underline{\tau} := \tau_1 \dots \tau_k$ we add constants $B_i^{\underline{\rho},\underline{\tau}}$ for $i = 1, \dots k$ along with the axiom scheme:

$$(\mathrm{BR}_{\underline{\rho},\underline{\tau}}) :\equiv \begin{cases} y([\underline{x}, n]) <_0 n \to B_i^{\underline{\rho},\underline{\tau}}(y, \underline{z}, \underline{u}, n, \underline{x}) =_{\tau_i} z_i(n, [\underline{x}, n]) \\ y([\underline{x}, n]) \geq_0 n \to B_i^{\underline{\rho},\underline{\tau}}(y, \underline{z}, \underline{u}, n, \underline{x}) =_{\tau_i} u_i(\lambda \underline{D^\rho} \cdot \underline{B^{\rho,\tau}}(y, \underline{z}, \underline{u}, S(n), [\underline{x}, n] * \underline{D}), n, [\underline{x}, n]) \end{cases}$$

for $i = 1 \dots k$, where

$$[\underline{x}, n]_i(j) =_{\rho_i} \begin{cases} x_i(j) \text{ if } j < n \\ 0^{\rho_i}, \text{ otherwise} \end{cases}$$

and

$$([\underline{x}, n] * \underline{D})_i(j) =_{\rho_i} \begin{cases} x_i(j) \text{ if } j < n \\ D_i \text{ if } j = n \\ 0^{\rho_i}, \text{ otherwise} \end{cases}$$

(for each type ρ, we defined 0^ρ inductively with 0^0 defined as the constant 0^0 in $\mathsf{WE\text{-}PA}^\omega$ and if $\rho := \tau(\sigma)$, we set $0^\rho := \lambda x^\sigma \cdot 0^\tau$). We denote the collection of these axioms over all types in T by (BR).

(BR) allows us to provide the following program extraction theorem for $\mathcal{A}^\omega$:

Theorem 2.2.8 ([109, 138])**.** *Let $\mathcal{P}$ be a set of universal sentences and let $A(\underline{a})$ be an arbitrary formula (with only the variables $\underline{a}$ free) in the language of $\mathsf{WE\text{-}PA}^\omega$. Then*

$$\mathcal{A}^\omega + \mathcal{P} \vdash A(\underline{a}) \text{ implies } \mathsf{WE\text{-}PA}^\omega + (\mathrm{BR}) + \mathcal{P} \vdash \forall \underline{a}, \underline{y}(A')_D(\underline{t}\underline{a}, \underline{y}, \underline{a})$$

Here, $\underline{t}$ is a tuple of closed terms of $\mathsf{WE\text{-}PA}^\omega + (\mathrm{BR})$ which can be extracted from the respective proof.

2.2.2 A program extraction theorem for inner product spaces

As was the case for the soundness theorem for the Dialectica interpretation (c.f. Remark 2.2.3), Theorem 2.2.8 also holds for $\mathcal{A}^\omega[X, \langle \cdot, \cdot \rangle]$; more precisely:

Theorem 2.2.9 ([79]). *Let $\mathcal{P}$ be a set of universal sentences and let $A(\underline{a})$ be an arbitrary formula (with only the variables $\underline{a}$ free) in the language of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$. Then*

$$\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle] \vdash A(\underline{a}) \text{ implies } \mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle] + (\mathrm{BR}) \vdash \forall \underline{a}, \underline{y}(A')_D(\underline{t}a, \underline{y}, \underline{a})$$

Here, $\underline{t}$ is a tuple of closed terms of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]+(\mathrm{BR})$ which can be extracted from the respective proof.

Remark 2.2.10. The above result is given implicitly in the proof of Theorem 3.30 of [79].

Remark 2.2.11. The key aspect of the proof of the above theorem is that the additional axioms of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ are purely universal and thus have a trivial Dialectica interpretation. This then allows the proof of the above theorem to easily follow by an adaptation of Spector's original proof of Theorem 2.2.8 adapted to higher types [109].

Although the proof of the above theorem provides an algorithm to extract computational content from proofs in $\mathcal{A}^\omega$, the validity of such program extractions is in question. It is a well-known fact that bar recursion is not set-theoretically valid; more precisely, for an inner product space $(X, \|\cdot\|)$ the structure of all set-theoretic functionals $\mathcal{S}^{\omega,X}$, is defined via $\mathcal{S}_0 := \mathbb{N}$, $\mathcal{S}_X := X$ and

$$\mathcal{S}_{\tau(\xi)} := \mathcal{S}_\tau^{\mathcal{S}_\xi}.$$

This is the natural model of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ and it is known that it does not model (BR) (see the discussion at the start of Section 11.5 of [80]).

However, for a particular class of results provable in $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ we do obtain a program extraction theorem whose validity can be verified in the model of set-theoretic functionals. A particular feature of such formulas is that they contain variables with types that have low complexity, more precisely:

We say a type ρ is of degree n if $\rho \in \mathrm{T}$ and $\deg(\rho) \leq n$. Further, we call ρ small if it is of the form $\rho = \rho_0(0)\dots(0)$ for $\rho_0 \in \{0, X\}$ (including $0, X$) and call it admissible if it is of the form $\rho = \rho_0(\tau_k)\dots(\tau_1)$ where each τ_i is small and $\rho_0 \in \{0, X\}$ (also including $0, X$).

We now introduce a very important model of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ due to Kohlenbach, which is an extension of Bezem's [14] structure of hereditarily strongly majorizable functionals.

First, given $\tau \in \mathrm{T}^X$, by recursion, we define

$$\widehat{0} := 0,\ \widehat{X} := 0,\ \widehat{\tau(\xi)} := \widehat{\tau}(\widehat{\xi}).$$

Now, the majorizability relation $\gtrsim$ and the structure of all strongly majorizable functionals is defined as follows:

Definition 2.2.12 ([48, 79]). Let $(X, \langle\cdot,\cdot\rangle)$ be a non-empty inner-product space. The structure

$\mathcal{M}^{\omega,X}$ and the majorizability relation $\gtrsim_\rho$ are defined by

$$\begin{cases} \mathcal{M}_0 := \mathbb{N}, n \gtrsim_0 m := n \geq m \wedge n, m \in \mathbb{N}, \\ \mathcal{M}_X := X, n \gtrsim_X x := n \geq \|x\| \wedge n \in \mathcal{M}_0, x \in \mathcal{M}_X, \\ f \gtrsim_{\tau(\xi)} x := f \in \mathcal{M}_{\widehat{\tau}}^{\mathcal{M}_{\widehat{\xi}}} \wedge x \in \mathcal{M}_\tau^{\mathcal{M}_\xi} \\ \qquad\qquad \wedge \forall g \in \mathcal{M}_{\widehat{\xi}}, y \in \mathcal{M}_\xi (g \gtrsim_\xi y \to fg \gtrsim_\tau xy) \\ \qquad\qquad \wedge \forall g, y \in \mathcal{M}_{\widehat{\xi}} (g \gtrsim_{\widehat{\xi}} y \to fg \gtrsim_{\widehat{\tau}} fy), \\ \mathcal{M}_{\tau(\xi)} := \left\{ x \in \mathcal{M}_\tau^{\mathcal{M}_\xi} \mid \exists f \in \mathcal{M}_{\widehat{\tau}}^{\mathcal{M}_{\widehat{\xi}}} : f \gtrsim_{\tau(\xi)} x \right\}. \end{cases}$$

Remark 2.2.13. The previous definition originates from [79], where it is given implicitly in the proof of Theorem 3.30 of that article. An explicit presentation is given in Definition 9.1 of [48].

Furthermore, we define the following syntactical counterpart to $\gtrsim_\rho$ in the language of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ which we denote $\leq_\rho$:

1. $x \leq_0 y := x \leq_0 y$.
2. $x \leq_X y := \|x\| \leq_{\mathbb{R}} \|y\|$.
3. $x \leq_{\tau(\xi)} y := \forall z^\xi (xz \leq_\tau yz)$.

Here, we use the relations $\leq_{\mathbb{R}}$ and $\leq_0$ introduced in Section 2.1.2. We also have the obvious generalisation for $\leq_\rho$ to tuples, where, $\underline{x} \leq_{\underline{\sigma}} \underline{y}$ is an abbreviation for $x_1 \leq_{\sigma_1} y_1 \wedge \cdots \wedge x_k \leq_{\sigma_k} y_k$ where $\underline{x}$, $\underline{y}$ and $\underline{\sigma}$ are k-tuples of terms and types, respectively, such that x_i and y_i are of type σ_i.

Lastly, we introduce formulas of type Δ. Theorem 2.2.8 tells us the soundness theorem holds for extensions of $\mathcal{A}^\omega$ by any collection of universal axioms. This is because universal statements have trivial solutions to their Dialectica interpretation, formulas of type Δ were initially introduced in [75, 76] (and then lifted to abstract types in [55]) and represent a class of commonly occurring formulas with trivial monotone functional interpretations in the sense of Kohlenbach [80]. In our context, a formula of type Δ is any formula of the form

$$\forall \underline{a}^{\underline{\delta}} \exists \underline{b} \leq_{\underline{\sigma}} \underline{ra} \forall \underline{c}^{\underline{\gamma}} A_{qf}(\underline{a}, \underline{b}, \underline{c})$$

where A_{qf} is quantifier-free, the types in $\underline{\delta}$, $\underline{\sigma}$ and $\underline{\gamma}$ are admissible, $\underline{r}$ is a tuple of closed terms of appropriate type. We now have the following:

Theorem 2.2.14 ([55, 79])**.** *Let $\beth$ be a set of formulas of type Δ. Let τ be admissible, δ be of degree 1 and s be a closed term of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ of type $\sigma(\delta)$ for admissible σ and let*

$B_\forall(x,y,z,u)/C_\exists(x,y,z,v)$ be $\forall$-/$\exists$-formulas of $\mathcal{A}^\omega[X,\langle\cdot,\cdot\rangle]$ with only $x,y,z,u/x,y,z,v$ free. If

$$\mathcal{A}^\omega[X,\langle\cdot,\cdot\rangle]+\beth\vdash\forall x^\delta\forall y\leq_\sigma s(x)\forall z^\tau\left(\forall u^0 B_\forall(x,y,z,u)\to\exists v^0 C_\exists(x,y,z,v)\right),$$

then one can extract a partial functional $\Phi:\mathcal{S}_\delta\times\mathcal{S}_{\widehat{\tau}}\rightharpoonup\mathbb{N}$ which is total and (bar-recursively) computable on $\mathcal{M}_\delta\times\mathcal{M}_{\widehat{\tau}}$ and such that for all $x\in\mathcal{S}_\delta$, $z\in\mathcal{S}_\tau$, $z^*\in\mathcal{S}_{\widehat{\tau}}$, if $z^*\gtrsim z$, then

$$\mathcal{S}^{\omega,X}\models\forall y\leq_\sigma s(x)\left(\forall u\leq_0\Phi(x,z^*)B_\forall(x,y,z,u)\to\exists v\leq_0\Phi(x,z^*)C_\exists(x,y,z,v)\right)$$

holds whenever $\mathcal{S}^{\omega,X}\models\beth$ for $\mathcal{S}^{\omega,X}$ defined via any non-empty inner product space $(X,\|\cdot\|)$.

Further:

1. *If $\widehat{\tau}$ is of degree 1, then Φ is a total computable functional.*
2. *We may have tuples instead of single variables x,y,z,u,v.*
3. *If the claim is proved without DC, then τ may be arbitrary and Φ will be a total functional on $\mathcal{S}_\delta\times\mathcal{S}_{\widehat{\tau}}$ which is primitive recursive in the sense of Gödel [52] and Hilbert [58].*

Remark 2.2.15. The proof of the above theorem is rather involved, and we have chosen to omit it. Without the inclusion of the set $\beth$ of formulas of type Δ, the result is given as Theorem 3.24 of [79]. Dealing with $\beth$ follows by following [55].

2.3 Quantitative convergence

This thesis presents many quantitative results concerning the convergence of sequences of real numbers and random variables. This section presents well-known quantitative notions of deterministic convergence and their basic properties. This section is a starting point for developing the quantitative notions of probabilistic convergence we introduce in Section 4.2.

2.3.1 Quantitative notions of convergence: metastability, fluctuations and crossings

The first quantitative version of convergence we introduce can be seen as a direct computational interpretation of Cauchy convergence.

Definition 2.3.1. Suppose $\{x_n\}$ is a sequence of real numbers. We say the function $r:\mathbb{Q}^+\to\mathbb{N}$ is a *rate of (Cauchy) convergence* for $\{x_n\}$ if,

$$\forall\varepsilon\in\mathbb{Q}^+\,\forall n,m\geq r(\varepsilon)\left(|x_n-x_m|\leq\varepsilon\right).$$

Remark 2.3.2. The domain of a convergence rate, r, may change depending on preference and context. For example, if one wants to discuss the computability of specific rates, it may be easier to keep the domain as $\mathbb{Q}^+$; however, if one just wants a mathematical quantitative result, taking the domain to be $\mathbb{R}^+$ may work better. Sometimes, it may be convenient to take the domain of a rate to be some interval $(0, a]$ for some $a \in \mathbb{R}^+$ (typically 1).

In the context of the quantitative results of this thesis, if we are interested in the computability aspects of specific results, we will consider domains of $\mathbb{Q}^+$. Other than that, we shall pick domains that are most convenient for us in their respective contexts. Similar considerations will be given to all other quantitative notions we introduce in this thesis. We make the same consideration for the range of our quantitative notions. Note that all of these formulations are equivalent.

It is known that computable convergence rates do not generally exist, even if the sequence in question is computable.

Example 2.3.3 (Specker [137]). Fix a recursively enumerable set, A, that is not recursive (for example, the Halting set). Let $\{a_n\}$ be a recursive enumeration of A. That is, $\{a_n\}$ is a computable sequence of natural numbers containing all the elements of A exactly once. Let $\{s_n\}$ be the sequence defined as

$$s_n := \sum_{k=1}^{n} 2^{-a_k - 1}.$$

Then, it is clear that $\{s_n\}$ is a monotone increasing sequence that is bounded above by 1 (s_n is bounded by the sum of the reciprocals of the powers of 2). Now suppose $\{s_n\}$ has a computable rate of convergence. That is, suppose there is a computable function $\phi : \mathbb{Q}^+ \to \mathbb{N}$ satisfying

$$\forall \varepsilon \in \mathbb{Q}^+ \, \forall n, m \geq \phi(\varepsilon) \, |s_n - s_m| \leq \varepsilon. \tag{2.3}$$

We shall now produce an effective procedure that determines whether $k \in \mathbb{N}$ is in A or not, which will contradict the assumption that A is not a recursive set. Suppose $k \in \mathbb{N}$ is given. If $k = a_n$ for some $n \in \mathbb{N}$, then $n < \phi\left(2^{-k-2}\right) + 1$, if not, then $n - 1 \geq \phi\left(2^{-k-2}\right)$ which implies, by (2.3), that

$$2^{-a_n - 1} = |s_n - s_{n-1}| \leq 2^{-k-2}.$$

This implies $k < a_n$, contradicting the assumption that $k = a_n$.

Thus, to effectively determine if $k \in A$, it suffices to check if $k = a_n$ for $n < \phi\left(2^{-k-2}\right) + 1$ effectively, which can be done.

For statements that cannot always be given a direct computational interpretation, we can apply the proof interpretations introduced in Section 2.2. Observe that one can formulate

Cauchy convergence as

$$\forall \varepsilon \in \mathbb{Q}^+ \, \exists N \, \forall k \, \forall n, m \in [N; N+k] \, (|x_n - x_m| \leq \varepsilon) \tag{2.4}$$

where $[a; b] := \{a, a+1, \ldots, b\}$ if $a \leq b$ and empty otherwise. An application of the negative translation in combination with the Dialectica interpretation, in the same manner as the calculation in Remark 2.2.7, results in a formulation of Cauchy convergence equivalent to:

$$\forall \varepsilon \in \mathbb{Q}^+ \, \forall g : \mathbb{N} \to \mathbb{N} \, \exists N \, \forall n, m \in [N; N + g(N)](|x_n - x_m| \leq \varepsilon). \tag{2.5}$$

A proof of equivalence is readily obtained directly: (2.5) follows from (2.4) since if N is such that $|x_n - x_m| < \varepsilon$ for all $n, m \in [N; N+k]$, for all $k \in \mathbb{N}$, then we may set $k = g(N)$ and obtain (2.5). For the other direction, we argue by contradiction. If (2.4) does not hold then there exists some $\varepsilon \in \mathbb{Q}^+$ such that

$$\forall N \, \exists \, k \, \exists n, m \in [N, N+k](|x_n - x_m| \geq \varepsilon)$$

and therefore (by the axiom of choice), there exists some function $g : \mathbb{N} \to \mathbb{N}$ satisfying

$$\forall N \, \exists n, m \in [N, N + g(N)](|x_n - x_m| \geq \varepsilon)$$

which contradicts (2.5).

This new formulation changes the direct computational challenge of Cauchy convergence (that is, a rate of convergence) and, as discussed in Remark 2.2.7, if one uses classical logic to demonstrate the Cauchy convergence of a sequence, then one can hope to construct a realiser for (2.5). We thus have the following definition:

Definition 2.3.4. Suppose $\{x_n\}$ is a sequence of real numbers. We say the functional $\Phi : \mathbb{Q}^+ \times (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$ is a *rate of (Cauchy) metastability* for $\{x_n\}$ if,

$$\forall \varepsilon \in \mathbb{Q}^+ \, \forall g : \mathbb{N} \to \mathbb{N} \, \exists N \leq \Phi(\varepsilon, g) \, \forall n, m \in [N; N + g(N)](|x_n - x_m| \leq \varepsilon). \tag{2.6}$$

Although Example 2.3.3 demonstrates that one cannot obtain a general rate of convergence for nondecreasing, bounded sequences which just depends on a bound for the sequence, we can obtain such a rate of metastability:

Theorem 2.3.5 (Folklore, see essentially [80])**.** *Let $\{a_n\}$ be a monotone sequence of nonnegative numbers such that, for all $n \in \mathbb{N}$, we have $a_n < L$. Then*

$$\Phi(\varepsilon, g) := \tilde{g}^{(\lceil L/\varepsilon \rceil)}(0)$$

is a rate of metastable convergence for $\{a_n\}$*, where* $\tilde{g}(n) := n + g(n)$*, for all* $n \in \mathbb{N}$*.*

The systematic extraction of such rates, using proof-theoretic techniques, are standard results in proof mining with recent results including [43, 115, 124]. The idea of metastability was rediscovered in mainstream mathematics by Tao [139, 140], who was interested in finitizations of infinitary notions in mathematics and found nontrivial applications in several areas.

The next computational interpretation we introduce is a bound on fluctuations. We shall see, in the following subsection, that rates of convergence are computationally stronger than rates of metastability (that is, given a computable rate of convergence for a sequence, one can obtain a computable rate of metastability for the same sequence with the converse not possible by Example 2.3.3). We shall also see that the bounds on the fluctuations sit strictly in the middle of rates of convergence and metastability computationally, a result obtained in [92].

Definition 2.3.6. Suppose $\{x_n\}$ is a sequence of real numbers and $\varepsilon > 0$. We write $J_{N,\varepsilon}\{x_n\}$ for the total number of ε-fluctuations that occur in the initial segment $\{x_0, \ldots, x_{N-1}\}$ i.e. the maximal $k \in \mathbb{N}$ such that there exists

$$i_1 < j_1 \leq i_2 < j_2 \leq \ldots \leq i_k < j_k < N \text{ with } |x_{i_l} - x_{j_l}| \geq \varepsilon$$

for all $l = 1, \ldots, k$. We write

$$J_\varepsilon\{x_n\} := \lim_{N\to\infty} J_{N,\varepsilon}\{x_n\},$$

and this will be the total number of ε-fluctuations of the sequence $\{x_n\}$ (note that this could be infinite). A function $b : \mathbb{Q}^+ \to \mathbb{N}$ is a *bound on the fluctuations* of $\{x_n\}$ if for all $\varepsilon \in \mathbb{Q}^+$

$$J_\varepsilon\{x_n\} < b(\varepsilon).$$

The last computational interpretation we introduce is a bound on the crossings of a sequence.

Definition 2.3.7. Suppose $\{x_n\}$ is a sequence of real numbers and $\alpha < \beta$. We write $C_{N,[\alpha,\beta]}\{x_n\}$ for the total number of times $\{x_0, \ldots, x_{N-1}\}$ crosses the interval $[\alpha, \beta]$ i.e. the maximal $k \in \mathbb{N}$ such that there exists

$$i_1 < j_1 \leq i_2 < j_2 \leq \ldots \leq i_k < j_k < N \text{ with } x_{i_l} \leq \alpha \text{ and } \beta \leq x_{j_l} \text{ or vice-versa}$$

for all $l = 1, \ldots, k$. We write

$$C_{[\alpha,\beta]}\{x_n\} := \lim_{N\to\infty} C_{N,[\alpha,\beta]}\{x_n\}$$

for the total number of $[\alpha, \beta]$-crossings that occur in $\{x_n\}$. A function $b : \mathbb{Q}^+ \times \mathbb{Q}^+ \to \mathbb{N}$ is a

bound on the crossings of $\{x_n\}$ if for all $\alpha < \beta$

$$C_{N,[\alpha,\beta]}\{x_n\} < b(\alpha,\beta).$$

Remark 2.3.8. Crossings typically occur in the martingale theory literature. In this context, one generally encounters inequalities that deal specifically with *upcrossings* rather than crossings.

$U_{N,[\alpha,\beta]}\{x_n\}$ represents the number of times $\{x_0,\ldots,x_{N-1}\}$ upcrosses the interval $[\alpha,\beta]$ i.e. the maximal $k \in \mathbb{N}$ such that there exists

$$i_1 < j_1 \leq i_2 < j_2 \leq \ldots \leq i_k < j_k < N \text{ with } x_{i_l} \leq \alpha \text{ and } \beta \leq x_{j_l}$$

for all $l = 1,\ldots,k$. Similarly we define $D_{N,[\alpha,\beta]}\{x_n\}$ as the number of times $\{x_0,\ldots,x_{N-1}\}$ downcrosses the interval $[\alpha,\beta]$. Furthermore, we write

$$U_{[\alpha,\beta]}\{x_n\} := \lim_{N\to\infty} U_{N,[\alpha,\beta]}\{x_n\} \text{ and } D_{[\alpha,\beta]}\{x_n\} := \lim_{N\to\infty} D_{N,[\alpha,\beta]}\{x_n\}.$$

It is clear that between any two consecutive upcrossings, there has to be precisely one downcrossing, and therefore

$$C_{[\alpha,\beta]}\{x_n\} \leq 2U_{[\alpha,\beta]}\{x_n\} + 1 \text{ and } C_{[\alpha,\beta]}\{x_n\} \leq 2D_{[\alpha,\beta]}\{x_n\} + 1.$$

Although we have a strict computational hierarchy between rates of convergence, bounds on fluctuations and rates of metastability (which we shall see in the following subsection), a sequence possessing any of these is equivalent to the sequence converging.

Proposition 2.3.9 (Folklore)**.** *The following statements are equivalent to $\{x_n\}$ being convergent:*

(a) (Cauchy property) For all $\varepsilon > 0$ there exists some $n \in \mathbb{N}$ such that $i,j \geq n$ implies $|x_i - x_j| < \varepsilon$.

(b) (Finite crossings) $\{x_n\}$ is bounded and $C_{[\alpha,\beta]}\{x_n\} < \infty$ for all $\alpha < \beta$.

(c) (Finite fluctuations) $J_\varepsilon\{x_n\} < \infty$ for all $\varepsilon > 0$.

(d) (Metastability) For all $\varepsilon > 0$ and $g : \mathbb{N} \to \mathbb{N}$ there exists some $n \in \mathbb{N}$ such that $|x_i - x_j| < \varepsilon$ for all $i,j \in [n; n+g(n)]$.

Remark 2.3.10. If we have $C_{[\alpha,\beta]}\{x_n\} < \infty$ for all $\alpha < \beta$ then all we can conclude is that $\{x_n\}$ converges to some element of $\mathbb{R}^* := \mathbb{R} \cup \{\pm\infty\}$ and thus the boundedness condition forces convergence in $\mathbb{R}$.

2.3.2 The computational hierarchy of the quantitative notions of convergence

We start by presenting the relationship between rates of convergence and rates of metastabilities.

Theorem 2.3.11 (Folklore)**.** *$r : \mathbb{Q}^+ \to \mathbb{N}$ is a rate of convergence for a sequence $\{x_n\}$ iff r^M defined as $r^M(\varepsilon, g) = r(\varepsilon)$, for all $\varepsilon \in \mathbb{Q}^+, g : \mathbb{N} \to \mathbb{N}$, is a rate of metastability for $\{x_n\}$.*

Proof. For the forward direction, let $\varepsilon \in \mathbb{Q}^+$ and $g : \mathbb{N} \to \mathbb{N}$ be given. Then, taking $N = r^M(\varepsilon, g) = r(\varepsilon)$, we have (from the fact that r is a rate of convergence) $\forall n, m \geq N\,(|x_n - x_m| < \varepsilon)$. So, in particular $\forall n, m \in [N; N + g(N)](|x_n - x_m| < \varepsilon)$.

For the converse, let $\varepsilon \in \mathbb{Q}^+$ be given. Take $p, q \geq r(\varepsilon)$. Define $g : \mathbb{N} \to \mathbb{N}$ as, $g(n) = \max\{p, q\}$. Since r^M is a rate of metastability, there exists $N \leq r^M(\varepsilon, g) = r(\varepsilon) \leq p, q$ such that $\forall n, m \in [N; N + \max\{p, q\}](|x_n - x_m| < \varepsilon)$. Since it is clear that both $p, q \in [N; N + \max\{p, q\}]$ we have $|x_p - x_q| < \varepsilon$. □

The above demonstrates that given a computable rate of convergence, one can obtain a computable (in a suitable sense) rate of metastability. In particular, the function itself acts as a rate. Furthermore, if a rate of metastable convergence is independent of its function part, it can be regarded as a rate of convergence.

One can easily show that a rate of convergence is computationally stronger than a bound on the fluctuations.

Theorem 2.3.12 (Folklore)**.** *If $r : \mathbb{Q}^+ \to \mathbb{N}$ is a rate of convergence for a sequence of real numbers $\{x_n\}$, then r is a bound on the fluctuations for the same sequence.*

Proof. Given $\varepsilon \in \mathbb{Q}^+$, for all $i, j \geq r(\varepsilon)$ we must have $|x_i - x_j| < \varepsilon$. Therefore any ε-fluctuation must occur before $r(\varepsilon)$ and thus we must have $J_\varepsilon\{x_n\} < r(\varepsilon)$. □

The fact that a rate of convergence is strictly computationally stronger than a bound on the ε-fluctuation follows from the following example:

Example 2.3.13. Let $\{s_n\}$ be as in Example 2.3.3. Then, we have already shown that $\{s_n\}$ does not have a computable rate of convergence.

Since $\{s_n\}$ is a positive, increasing sequence of rationals, bounded above by 1 if we have

$$i_1 < j_1 \leq i_2 < j_2 \leq \ldots \leq i_k < j_k \text{ with } |s_{i_l} - s_{j_l}| \geq \varepsilon.$$

This implies $1 \geq s_{j_k} \geq k\varepsilon$ and so $b(\varepsilon) := \lceil 1/\varepsilon \rceil$ is a computable bound on the fluctuations.

Next, we observe that a bound on the fluctuations is computationally stronger than a rate of metastability. The construction of a computable sequence of rational numbers with a rate of metastability but without a computable bound on their fluctuations is given in [92]. On the other hand, we can show that a bound on the fluctuations of a sequence yields a rate of metastability of a particularly nice form.

Theorem 2.3.14 (Folklore)**.** *Suppose $b : \mathbb{Q}^+ \to \mathbb{N}$ is a bound on the fluctuations for a sequence of real numbers $\{x_n\}$. Then $\Phi(\varepsilon, g) := \tilde{g}^{(b(\varepsilon))}(0)$ is a rate of metastability for $\{x_n\}$.*

Proof. Suppose, for contradiction, that Φ defined above was not a rate of metastability for $\{x_n\}$, then we would have

$$\exists i, j \in [\tilde{g}^{(e)}(0); \tilde{g}^{(e+1)}(0)](|x_i - x_j| \geq \varepsilon)$$

for all $e = 0, \ldots, b(\varepsilon)$, and thus $J_\varepsilon\{x_n\} \geq b(\varepsilon) + 1$, a contradiction. □

The rate of metastability for a sequence of real numbers with a bound on their fluctuations has a particularly clean form, namely an iteration $\tilde{g}^{(e)}(0)$. We call such rates *learnable*, loosely following the terminology of [92] where the Cauchy property forms a simple instance of the class of effectively learnable formulas. It turns out that for such rates of metastability, it is always the case that their exponent of iteration provides a bound on the fluctuations. To see this, we need a lemma, which will be very helpful to us later.

Lemma 2.3.15. *Let $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ be sequences of natural numbers. Then we can define a function $g : \mathbb{N} \to \mathbb{N}$ in such a way that $\tilde{g}^{(i)}(0) = b_{i-1}$ for $i \geq 1$ and for any $n \in \mathbb{N}$ we have $[a_m; b_m] \subseteq [n; n + g(n)]$ for the least m such that $n \leq a_m$.*

Proof. Define $g(n) := b_{k(n)} - n$ where

$$k(n) := \min\{i \mid n \leq a_i\}$$

(this is well-defined, since $a_0 < a_1 < \ldots$ and $n \leq a_{k(n)} < b_{k(n)}$). By an easy induction we can show that $\tilde{g}^{(i)}(0) = b_{i-1}$, where in particular we have $g(b_{i-1}) = b_{k(b_{i-1})} - b_{i-1} = b_i - b_{i-1}$. Now for any $n \in \mathbb{N}$ we have $n + g(n) = b_{k(n)}$ and since $n \leq a_{k(n)}$ it follows that $[a_{k(n)}; b_{k(n)}] \subseteq [n; n + g(n)]$. □

Theorem 2.3.16. *$\Phi(\varepsilon, g) := \tilde{g}^{(b(\varepsilon))}(0)$ is a rate of metastability for a sequence $\{x_n\}$ if and only if $b : \mathbb{Q}^+ \to \mathbb{N}$ is a bound on the fluctuations for $\{x_n\}$.*

Proof. If $\Phi(\varepsilon, g) := \tilde{g}^{(b(\varepsilon))}(0)$ is a rate of metastability for $\{x_n\}$, then for all ε and $g : \mathbb{N} \to \mathbb{N}$,

$$\exists n \leq \tilde{g}^{(b(\varepsilon))}(0)\, \forall i, j \in [n; n + g(n)](|x_i - x_j| < \varepsilon). \tag{2.7}$$

The above is equivalent to the property that for any $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$

$$\exists n \leq b(\varepsilon)\, \forall i, j \in [a_n; b_n](|x_i - x_j| < \varepsilon) \tag{2.8}$$

To see that (2.8) implies (2.7) we just define $a_n := \tilde{g}^{(n)}(0)$ and $b_n := \tilde{g}^{(n+1)}(0)$ for all $n \leq b(\varepsilon)$, and some arbitrary increasing sequence from that point. Then if (2.7) is false, we have $a_n < b_n$ for all $n \leq b(\varepsilon)$ and (2.8) must also be false. Conversely, from $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ we define g as in Lemma 2.3.15, and then if (2.8) is false then by Lemma 2.3.15, for any $n \leq \tilde{g}^{(b(\varepsilon))}(0) = b_{b(\varepsilon)-1} \leq a_{\phi(\varepsilon)}$ we have $[a_m; b_m] \subseteq [n; n + g(n)]$ for some $m \leq b(\varepsilon)$, and thus (2.7) is also false.

Now, observe that $b(\varepsilon)$ satisfying (2.8) must be a bound for $J_\varepsilon\{x_n\}$. □

The above discussion allows us to obtain the following immediate corollary for monotone sequences:

Corollary 2.3.17. *Let $\{a_n\}$ be a monotone sequence of nonnegative numbers such that, for all $n \in \mathbb{N}$, we have $a_n < L$. Then*

$$\phi(\varepsilon) = \left\lceil \frac{L}{\varepsilon} \right\rceil$$

is a bound on the fluctuations for $\{a_n\}$.

Remark 2.3.18. One actually has the sharper bound of

$$J_\varepsilon\{a_n\} < \frac{L}{\varepsilon}$$

for $\{a_n\}$ in the previous result.

Similar equivalences will be presented in the stochastic setting in Section 4.2.

We now complete the computational picture by demonstrating how crossings fit. We first need the following definition, which will be important to us when discussing crossings moving on (specifically, in Section 4.2 and throughout Chapter 7).

Definition 2.3.19. Given some $M > 0$ and $l \in \mathbb{N}$, let $\mathcal{P}(M, l)$ denote the partition of $[-M, M]$ into l equally sized closed subintervals i.e.

$$\mathcal{P}(M, l) := \left\{ \left[-M + \frac{2Mi}{l}, -M + \frac{2M(i+1)}{l} \right] \,\middle|\, i = 0, \ldots, l-1 \right\}.$$

We now demonstrate that having finite crossings (and a bound) is computationally equivalent to having finite fluctuations.

Proposition 2.3.20. *Let $\{x_n\}$ be a sequence of real numbers:*

(i) If $J_\varepsilon\{x_n\} \leq \phi(\varepsilon)$ for all $\varepsilon > 0$ then $C_{[\alpha,\beta]}\{x_n\} \leq \phi(\beta - \alpha)$ for all $\alpha < \beta$.

(ii) If $C_{[\alpha,\beta]}\{x_n\} \leq \psi(\alpha, \beta)$ for all $\alpha < \beta$ and also $|x_n| \leq M$ for all $n \in \mathbb{N}$ then $J_\varepsilon\{x_n\} \leq \phi(\varepsilon)$ for all $\varepsilon > 0$ where

$$\phi(\varepsilon) := l \cdot \max\{\psi(\alpha, \beta) \,|\, [\alpha, \beta] \in \mathcal{P}(M, l)\} \quad \text{for } l := \left\lceil \frac{4M}{\varepsilon} \right\rceil.$$

Proof. Part (i) is immediate, so we focus on proving (ii). Fix $\varepsilon > 0$ and divide $[-M, M]$ into $l = \lceil 4M/\varepsilon \rceil$ equal subintervals, which we label $I_j = [\alpha_j, \beta_j]$ for $j = 1, \ldots, l$. Since $\beta_j - \alpha_j \leq \varepsilon/2$ and $\{x_n\}$ is contained in $[-M, M]$, a single ε-fluctuation of $\{x_n\}$ crosses at least one of the I_j. Therefore if $J_\varepsilon\{x_n\} \geq k$ then there must be some interval $[\alpha_j, \beta_j]$ with at least k/l crossings i.e.

$$C_{[\alpha_j,\beta_j]}\{x_n\} \geq \frac{k}{l}$$

for this particular j, and thus

$$k \leq l \cdot \psi(\alpha_j, \beta_j)$$

from which the bound follows. □

2.4 Probability theory

This section reviews the basic notions from probability theory on the reals and Banach spaces we need for this thesis.

2.4.1 Basic notions

We start with an introduction to the basic notions and results from probability theory we shall use freely throughout this thesis. We closely follow [56] and [35].

Probability theory aims to create a rigorous framework in which we can assign numerical values that capture how likely certain events are to occur. To do this, we must first make formal the space of events which we are discussing:

Definition 2.4.1 ($(\sigma$-)Algebra). Let Ω be a set. An *algebra of subsets* of Ω, $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, is a subset of $\mathcal{P}(\Omega)$ that satisfies the following:

(i) $\emptyset \in \mathcal{F}$.

(ii) $\forall A \in \mathcal{F}\, (A^c \in \mathcal{F})$.

(iii) $\forall A, B \in \mathcal{F}\, (A \cup B \in \mathcal{F})$.

An algebra, $\mathcal{F}$, is a *σ-algebra* if it additionally satisfies

$$\forall A_0, A_1, \ldots \in \mathcal{F} \left(\bigcup_{i=0}^{\infty} A_i \in \mathcal{F} \right).$$

Remark 2.4.2. In the context of probability theory, Ω in the previous definition is typically called the *sample space* and $\mathcal{F}$ the *space of events*. The elements of $\mathcal{F}$ are typically called *measurable*.

We now introduce the operators that assign numerical values expressing the likelihood of events occurring.

Definition 2.4.3 (Measures/contents). If $\mathcal{F}$ is an algebra of subsets of a set Ω, a map $\mathbb{P} : \mathcal{F} \to [0, 1]$ is called a *probability content* if:

(i) $\mathbb{P}(\emptyset) = 0$.

(ii) $\forall A, B \in \mathcal{F}(A \cap B = \emptyset \to \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B))$.

A probability content is called a *probability measure* if $\mathcal{F}$ is a σ-algebra and for all sequences of pairwise disjoint events $\{A_n\}$

$$\mathbb{P}\left(\bigcup_{n=0}^{\infty} A_n \right) = \sum_{n=0}^{\infty} \mathbb{P}(A_n).$$

For a fixed sample space Ω, if $\mathcal{F}$ is an algebra of subsets of Ω and $\mathbb{P}$ is a probability content, we call the tuple $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability content space*. If $\mathcal{F}$ is a σ-algebra and $\mathbb{P}$ is a probability measure we call $(\Omega, \mathcal{F}, \mathbb{P})$ a *probability space*.

The next notion from probability theory we present is that of a random variable. This definition tries to capture the idea of outcomes of experiments that contain levels of randomness.

Definition 2.4.4 (Random variables). The Borel σ-algebra on the reals, $\mathcal{B}(\mathbb{R})$, is the σ-algebra generated by the open intervals in $\mathbb{R}$. The Borel σ-algebra on the extended reals, $[-\infty, \infty]$, is defined as $\mathcal{B}([-\infty, \infty]) := \{A \subseteq [-\infty, \infty] \mid A \cap \mathbb{R} \in \mathcal{B}(\mathbb{R})\}$.

A random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \to [-\infty, \infty]$ such that for all sets $B \in \mathcal{B}([-\infty, \infty])$ we have

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

The *distribution* of a random variable X is a function $\mathbb{P}_X : \mathcal{B}([-\infty, \infty]) \to [0, 1]$ such that for all $B \in \mathcal{B}([-\infty, \infty])$ we have $\mathbb{P}_X(B) = \mathbb{P}(X^{-1}(B))$. One can show $([-\infty, \infty], \mathcal{B}([-\infty, \infty]), \mathbb{P}_X)$ is a probability space, for every random variable X.

Two random variables X and Y are said to be *identically distributed* if $\mathbb{P}_X \equiv \mathbb{P}_Y$. The random variables X and Y are said to be *independent* if for all $B_1, B_2 \in \mathcal{B}$ the events $A_1 := X^{-1}(B_1)$ and $A_2 := Y^{-1}(B_2)$ are independent, that is,

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2).$$

Remark 2.4.5. Throughout this thesis, we follow the standard convention of writing $\mathbb{P}(\psi(X))$ for $\mathbb{P}(\{\omega \in \Omega : \psi(X(\omega))\})$ whenever X is a random variable and ψ is a formula with $\{\omega \in \Omega : \psi(X(\omega))\} \in \mathcal{F}$. Furthermore, we say an event holds almost surely (a.s) if it holds with probability 1. For example, if X is a random variable, we say $X \geq 0$ a.s to mean $\mathbb{P}(X \geq 0) = 1$.

Remark 2.4.6. If we have a sequence of random variables $\{X_n\}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $f : \mathbb{R}^m \to [-\infty, \infty]$ is a Borel measurable function, with $m \in \mathbb{N}$, then one can show that $f(X_1, \ldots, X_m)$ is a random variable. This immediately tells us that sums, products, absolute values, maximums and minimums of random variables are random variables. Furthermore, one can show that $\inf_{n\in\mathbb{N}} X_n$, $\sup_{n\in\mathbb{N}} X_n$, $\liminf_{n\to\infty} X_n$, $\limsup_{n\to\infty} X_n$ are all random variables. An immediate consequence of the supremum of random variables being a random variable is that, if $\{X_n\}$ is nonnegative, then

$$\sum_{n=0}^{\infty} X_n$$

is a random variable. See [56] for details.

The last basic notion from probability theory we need is the *expected value*, $\mathbb{E}$ (also known as the expectation or mean), of a random variable. For the remainder of the subsection, fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Notation 2.4.7. *Throughout this thesis, we shall denote the* indicator function *for a set A by I_A.*[2]

We now introduce the notion of a simple function:

Definition 2.4.8 (Simple functions). A function $X : \Omega \to \mathbb{R}$ is called a *simple function* if there exists $a_0, \ldots a_n \in \mathbb{R}$ and $A_0, \ldots, A_n \in \mathcal{F}$, a partition of Ω, such that:

$$X \equiv \sum_{i=0}^{n} a_i I_{A_i}.$$

It is clear that simple functions are random variables. The expected value of a simple function is defined as follows:

[2]Typically A will be a subset of some set Ω and so $I_A : \Omega \to \{0, 1\}$. This will always be clear from the context.

Definition 2.4.9 (Expected value of simple functions). Suppose $X \equiv \sum_{i=0}^{n} a_i I_{A_i}$ is a simple function, with $a_0, \dots a_n \in \mathbb{R}$ and $A_0, \dots, A_n \in \mathcal{F}$, a partition of Ω. The expected value of X, written $\mathbb{E}(X)$, is defined as

$$\mathbb{E}(X) := \sum_{i=0}^{n} a_i \mathbb{P}(A_i).$$

Before defining the expected value for general random variables, we must define the notion of an integrable random variable. Intuitively, a random variable is integrable if it can be approximated by simple functions whose expected values are uniformly bounded.

Definition 2.4.10 (Integrable random variables). A random variable X on $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be *integrable* if any of the following hold:

(i) X is a simple function.

(ii) $X \geq 0$ and $\sup\{\mathbb{E}(g) : g \leq X, g \text{ is simple}\} < \infty$.

(iii) Both $X^+ := \max\{X, 0\}$ and $X^- := \max\{-X, 0\}$ satisfy (ii).

We can now define the expected value of an integrable random variable:

Definition 2.4.11 (Expected value of an integrable random variable). Suppose X is an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. If $X \geq 0$, then

$$\mathbb{E}(X) := \sup\{\mathbb{E}(g) : g \leq f, g \text{ is simple}\}.$$

If X is not assumed to be nonnegative, we define $\mathbb{E}(X) := E(X^+) - \mathbb{E}(X^-)$.

Remark 2.4.12. There are a few details to check to ensure that the expected value, as presented above, is well-defined and consistent. For example, one must check that the definition we gave for the expected value of a simple function in Definition 2.4.9 coincides with that of a general random variable given in Definition 2.4.11. We do not verify all of these details, but we refer the reader to Section 1.4 of [35].

One can show that the set of random variables and the set of integrable random variables, respectively, form real vector spaces. More generally, we have the following:

Definition 2.4.13 (L_p space). For $p \in (0, \infty)$, we write $L_p := L_p(\Omega, \mathcal{F}, \mathbb{P})$ for the set of all random variables X, such that $|X|^p$ is integrable. We write L_0 for the set of all random variables and L_∞ for the set of all *almost surely* bounded random variables.[3]

For $p \in [0, \infty]$, one can show that L_p is a real vector space and for $p \in [1, \infty]$, one can show

[3]By almost surely bounded, we mean there exists $M > 0$ such that $\mathbb{P}(|X| < M) = 1$.

that $(L_p, \|\cdot\|_p)$ is a real seminormed space,[4] with seminorm $\|\cdot\|_p$ defined as:

$$\|X\|_p := (\mathbb{E}(|X|^p))^{\frac{1}{p}}$$

for $p \in (0, \infty)$ and

$$\|X\|_\infty := \inf\{C > 0 \mid |X| < C \text{ a.s}\}.$$

Furthermore, one can show that if $p \leq q$, then $L_q \subseteq L_p$ and $\|X\|_p \leq \|X\|_q$.

Definition 2.4.14. If $X \in L_2$, then we define the variance as, $\mathrm{Var}(X) := \mathbb{E}([X - \mathbb{E}(X)]^2) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

We shall now state some properties of the expected value we shall use freely throughout this thesis:

Theorem 2.4.15 (Properties of the expected value c.f. [35, 56])**.** *Let X and Y be integrable random variables on $(X, \mathcal{F}, \mathbb{P})$. The following hold:*

(i) *If $X = 0$ a.s, then $\mathbb{E}(X) = 0$.*

(ii) $|X| < \infty$ a.s.

(iii) *If $\mathbb{E}(X) > 0$, then $\mathbb{P}(X > 0) > 0$.*

(iv) *For all $a, b \in \mathbb{R}$, $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.*

(v) *If $X = Y$ a.s, then $\mathbb{E}(X) = \mathbb{E}(Y)$.*

(vi) *If $X \leq Y$ a.s, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.*

(vii) *If XY is integrable, then X and Y are independent iff $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

(viii) *If $X, Y \in L_2$ and are independent, then $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.*

(ix) *If $X \in L_2$, then $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$*

(x) *If $X \in L_2$, then for all $a, b \in \mathbb{R}$, $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.*

(xi) *If X and Y are identically distributed then $\mathbb{E}(X) = \mathbb{E}(Y)$. Furthermore, if $p \in [0, \infty]$ and $X, Y \in L_p$, then $\|X\|_p = \|Y\|_p$.*

[4] $\|\cdot\|_p$ is not a norm as it is not positive definite, that is, there are nonzero random variables X satisfying $\|X\|_p = 0$. If $\|X\|_p = 0$, then $X = 0$ a.s. This observation inspires the relation on the set of random variables, L_0, that $X \sim Y$ iff $X - Y = 0$ a.s. One can easily verify that $(L_0/\sim, \|\cdot\|_p)$ forms a normed space, with the vector space operations and norm defined by application to class representatives.

(xii) Markov's inequality*: If $X \geq 0$, then for all $a > 0$ we have*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

(xiii) Jensen's inequality*: If X is a real-valued random variable and $f : \mathbb{R} \to \mathbb{R}$ is a convex function such that $f(X)$ is an integrable random variable, then*

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)).$$

(xiv) Fatou's Lemma*: If $\{X_n\}$ is a sequence of nonnegative, integrable random variables such that $\liminf_{n\to\infty} \mathbb{E}(X_n) < \infty$, then $\liminf_{n\to\infty} X_n$ is integrable and*

$$\mathbb{E}\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} \mathbb{E}(X_n).$$

(xv) If $\{X_n\}$ is a sequence of integrable random variables, then

$$\mathbb{E}\left(\sum_{n=0}^{\infty} X_n\right) = \sum_{n=0}^{\infty} \mathbb{E}(X_n)$$

if $\sum_{n=0}^{\infty} X_n$ is integrable, or $\sum_{n=0}^{\infty} \mathbb{E}(X_n) < \infty$.

2.4.2 Martingale theory

We shall now review the notions and results from discrete (real-valued) martingale theory we need in this thesis. We, again, closely follow [35, 56] as well as [142].

We typically call a sequence of random variables on some fixed probability space a *stochastic process* in the context of martingale theory.

Martingale theory is a foundational concept in probability theory, with profound implications in fields such as finance and statistical inference. At its core, a martingale is a model of a fair game, where future predictions are based solely on past knowledge, and no expected gain or loss can be anticipated.

Definition 2.4.16 (Filtration). A *filtration* on a σ-algebra, $\mathcal{F}$, of subsets of a sample space Ω is a family of σ-algebras $\{\mathcal{F}_n\}$ such that for all $n \leq m$ we have $\mathcal{F}_n \subseteq \mathcal{F}_m \subseteq \mathcal{F}$.

A stochastic process $\{X_n\}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is said to be *adapted* to a filtration $\{\mathcal{F}_n\}$ on $\mathcal{F}$ if for all $n \in \mathbb{N}$, X_n is $\mathcal{F}_n$-measurable, that is X_n is a random variable on the probability space $(\Omega, \mathcal{F}_n, \mathbb{P})$.

To state the definition of a martingale, we must introduce the *conditional expectation*, whose definition is motivated by the following result:

Theorem 2.4.17 (Kolmogorov 1933)**.** *Let X be an integrable random variable on a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{F}'$ be a sub-σ-algebra of $\mathcal{F}$. Then there exists a random variable Y, called a conditional expectation of X with respect to $\mathcal{F}'$, such that:*

(a) Y is $\mathcal{F}'$-measurable.

(b) Y is integrable.

(c) $\mathbb{E}(YI_A) = \mathbb{E}(XI_A)$ for any $A \in \mathcal{F}'$. Where I_A is the indicator function of A.

Further, if Y' is another such random variable, then $Y = Y'$ a.s.

The conditional expectation is defined as some choice of a conditional expectation:

Definition 2.4.18. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X an integrable random variable. Let $\mathcal{F}'$ be a sub-σ-algebra of $\mathcal{F}$. The conditional expectation of X with respect to $\mathcal{F}'$, written $\mathbb{E}(X \mid \mathcal{F}')$, is some choice of a conditional expectation of X with respect to $\mathcal{F}'$.

Remark 2.4.19. We note that the conditional expectation is only defined up to almost sure equivalence and so the standard convention is made that when a the conditional expectation is used in a relation (typically equalities and inequalities) these relations are understood to hold almost surely. We turn the reader to [142] for more detailed discussions.

Throughout this thesis, we freely use the following properties of the conditional expectation:

Theorem 2.4.20 ([56])**.** *Let X, Y be integrable random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{F}_1 \subseteq \mathcal{F}_2$ be sub-σ-algebras of $\mathcal{F}$, and $a, b, c \in \mathbb{R}$. The following hold:*

(i) $\mathbb{E}(\mathbb{E}(X \mid \mathcal{F}_1)) = \mathbb{E}(X)$.

(ii) $\mathbb{E}(aX + bY \mid \mathcal{F}_1) = a\mathbb{E}(X \mid \mathcal{F}_1) + b\mathbb{E}(Y \mid \mathcal{F}_1)$.

(iii) If X is $\mathcal{F}_1$-measurable, then $\mathbb{E}(X \mid \mathcal{F}_1) = X$.

(iv) $\mathbb{E}(c \mid \mathcal{F}_1) = c$.

(v) $\mathbb{E}(X \mid \{\emptyset, \Omega\}) = \mathbb{E}(X)$.

(vi) If $X \geq 0$ a.s*, then* $\mathbb{E}(X \mid \mathcal{F}_1) \geq 0$.

(vii) If XY is integrable and Y is $\mathcal{F}_1$-measurable, then $\mathbb{E}(XY \mid \mathcal{F}_1) = Y\mathbb{E}(X \mid \mathcal{F}_1)$.

(viii) $\mathbb{E}(\mathbb{E}(X \mid \mathcal{F}_1) \mid \mathcal{F}_2) = \mathbb{E}(X \mid \mathcal{F}_1) = \mathbb{E}(\mathbb{E}(X \mid \mathcal{F}_2) \mid \mathcal{F}_1)$.

(ix) If X is independent[5] *from $\mathcal{F}_1$, then* $\mathbb{E}(X \mid \mathcal{F}_1) = X$.

[5] X is independent from $\mathcal{F}_1$ means $\sigma(X)$ (the smallest σ-algebra such that X is measurable) is independent from $\mathcal{F}_1$. Meaning for $A \in \mathcal{F}_1$ and $B \in \sigma(X)$, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Note that random variables X and Y are independent iff $\sigma(X)$ and $\sigma(Y)$ are independent as σ-algebras.

(x) Conditional Jensen's inequality*: If X is a real-valued random variable and $f : \mathbb{R} \to \mathbb{R}$ is a convex function such that $f(X)$ is an integrable random variable, then*

$$f(\mathbb{E}(X \mid \mathcal{F}_1)) \leq \mathbb{E}(f(X) \mid \mathcal{F}_1).$$

We are now ready to define a martingale:

Definition 2.4.21 (Martingales). Let $\{\mathcal{F}_n\}$ be a filtration on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A stochastic process $\{X_n\}$ adapted to $\{\mathcal{F}_n\}$ is said to be a *martingale* if for all $n \in \mathbb{N}$:

(i) X_n is integrable.

(ii) $\mathbb{E}(X_{n+1} \,|\, \mathcal{F}_n) = X_n$.

$\{X_n\}$ is a *submartingale* (respectively *supermartingale*) if the equality in condition (ii) above is weakened to $\geq$ (respectively $\leq$).

Example 2.4.22 (Examples of Martingales). Let $\{X_n\}$ be a sequence of independent, identically distributed, integrable random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $X_0 :\equiv 0$. Then defining the filtration $\{\mathcal{F}_n\}$ with $\mathcal{F}_n := \sigma(X_0, \ldots, X_n)$ (that is, the smallest σ-algebra such that $X_0, \ldots X_n$ are measurable) ensures that $\{S_n\}$ defined as $S_n := \sum_{i=0}^{n} X_i$ is a martingale.

Definition 2.4.23 (Stopping times). A random variable τ with values in $\mathbb{N} \cup \{\infty\}$ is called a *stopping time* with respect to a filtration $\{\mathcal{F}_n\}$ if

$$\tau^{-1}([0, t]) \in \mathcal{F}_t$$

for all $t \in \mathbb{N}$. Furthermore, we define

$$\mathcal{F}_\tau = \{A \in \mathcal{F} \mid A \cap \tau^{-1}([0, t]) \in \mathcal{F}_t \text{ for all } t \in \mathbb{N}\}$$

and one can show that this forms a σ-algebra.

For a stochastic process $\{X_n\}$ adapted to $\{\mathcal{F}_n\}$ one can show that for any stopping time τ, the function X_τ is measurable with respect to $\mathcal{F}_\tau$.

Theorem 2.4.24 (The optional stopping theorem c.f. Theorem 10.10 of [142])**.** *Let $\rho \leq \tau$ (with probability 1) be bounded stopping times with respect to a filtration $\{\mathcal{F}_n\}$:*

1. *If $\{X_n\}$ is a martingale with respect to $\{\mathcal{F}_n\}$, then $\mathbb{E}(X_\tau \mid \mathcal{F}_\rho) = X_\rho$.*
2. *If $\{X_n\}$ is a submartingale with respect to $\{\mathcal{F}_n\}$, then $\mathbb{E}(X_\tau \mid \mathcal{F}_\rho) \geq X_\rho$.*
3. *If $\{X_n\}$ is a supermartingale with respect to $\{\mathcal{F}_n\}$, then $\mathbb{E}(X_\tau \mid \mathcal{F}_\rho) \leq X_\rho$.*

The last result from martingale theory, which we make frequent use of in this thesis, is a generalisation of Markov's inequality for supermartingales known as *Ville's inequality*.

Theorem 2.4.25 (Ville's inequality c.f. Exercise 4.8.2 of [35]). *Let $\{U_n\}$ be a nonnegative supermartingale. Then for any $a > 0$ we have*

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} U_n \geq a\right) \leq \frac{\mathbb{E}(U_0)}{a}.$$

2.4.3 Probability on Banach spaces

Fix a normed space $(\mathbb{B}, \|\cdot\|)$ and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We summarise the relevant parts of the theory of random variables taking values in a Banach space. We shall mainly follow the treatment given in [105]. If $\mathbb{B}$ is a Banach space, then the natural definition one would give a random variable taking values in $\mathbb{B}$ is a measurable map from $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathbb{B}$ endowed with the Borel σ-algebra generated by its open sets. However, as noted in [105], this definition is too general to develop a useful theory of probability (for example, the set of such random variables does not form a vector space. It is not even the case that this class is closed with respect to addition c.f. [111]). Therefore, it is standard to assume random variables X are *tight* (sometimes referred to as *Radon*), that is, for all $\varepsilon > 0$ there exists a compact set $K \subseteq \mathbb{B}$ such that

$$\mathbb{P}(X \in K) \geq 1 - \varepsilon.$$

Denote the set of tight Borel random variables in $\mathbb{B}$ by L_0. Working with such random variables ensures we can add random variables and multiply them by scalars without worrying about measurability. Furthermore, we also have the set $L_p := \{X \in L_0 \,|\, \mathbb{E}(\|X\|^p) < \infty\}$ is also a vector space. Lastly, we note that a random variable is tight if and only if it takes values on a separable subset of $\mathbb{B}$ (c.f. [105, Section 2.1]), and so it is standard to assume $\mathbb{B}$ is separable. We shall adopt this convention here.

We also introduce the notion of integration for random variables taking values in $\mathbb{B}$, attributed to Bochner, via the following theorem:

Theorem 2.4.26 (c.f. Theorem II.11 of [108]). *There exists a unique linear mapping $\mathbb{E} : L_1(\mathbb{B}) \to \mathbb{B}$ called the expectation such that:*

(a) $\mathbb{E}(X) = \sum_{i=0}^{n} \mathbb{P}(A_i)x_i$ for all $X = \sum_{i=0}^{n} 1_{A_i} x_i$ with $\{A_i\} \subseteq \mathcal{F}$, and $\{x_i\} \subseteq \mathbb{B}$.

(b) $\|\mathbb{E}(X)\| \leq \mathbb{E}(\|X\|)$ for all $X \in L_1(\mathbb{B})$.

For $1 \leq p \leq 2$, $\mathbb{B}$ is said to have (*Rademacher*) type p, if there exists a constant B such that, for every independent sequence of (real-valued) random variables $\{\varepsilon_n\}$ satisfying

$$\mathbb{P}(\varepsilon_n = \pm 1) = \frac{1}{2}$$

(such a sequence is sometimes known as a *Rademacher sequence*) and sequence of element $\{x_n\}$ in $\mathbb{B}$, we have

$$\mathbb{E}\left(\left\|\sum_{i=0}^{n} x_i\varepsilon_i\right\|^p\right) \leq B\sum_{i=0}^{n}\|x_i\|^p.$$

By the triangle inequality, every Banach space is type 1. Furthermore, if $\mathbb{B}$ is type p, then it is of type p' for all $p' \leq p$ (c.f. [105, Proposition 9.12]) and a Banach space is of type 2 if and only if it is isomorphic to a Hilbert space.

In [62], it is shown that for $1 \leq p \leq 2$, if $\{X_n\}$ is an independent identically distributed (iid) sequence of random variables taking values in $\mathbb{B}$, with $\mathbb{E}(X_n) = 0$ (where $\mathbb{E}$ is the Bochner integral introduced in Theorem 2.4.26) and

$$\sum_{n=0}^{\infty}\frac{\mathbb{E}(\|X_n\|^p)}{n^p} < \infty, \tag{2.9}$$

$\mathbb{B}$ being a type p Banach space is both a necessary and sufficient condition for the conclusion

$$\frac{S_n}{n} \to 0$$

almost surely to hold.

In [143], Woyczynski shows that one can weaken condition (2.9) to

$$\sum_{n=1}^{\infty}\frac{\mathbb{E}(\phi_n(|X_n|))}{\phi_n(n)} < \infty,$$

where $\phi_n : \mathbb{R}^+ \to \mathbb{R}^+$ satisfy that

$$\frac{\phi_n(t)}{t} \text{ and } \frac{t^p}{\phi_n(t)}$$

are nondecreasing, and still conclude that $S_n/n \to 0$ almost surely for $\mathbb{B}$ a type p Banach space. What Woyczynski actually showed, in [143], was that this result holds in spaces such that there exists a constant C satisfying,

$$\mathbb{E}\left(\left\|\sum_{i=0}^{n} Y_i\right\|^p\right) \leq C\sum_{i=0}^{n}\mathbb{E}(\|Y_i\|^p) \tag{2.10}$$

for all independent random variables taking values in $\mathbb{B}$, $Y_0, \ldots, Y_n$, with 0 expected value and finite pth moment.[6] If $\mathbb{B}$ is a type p Banach space with constant B, then (2.10) holds with $C = (2B)^p$ (c.f. [105, Proposition 9.11]). Therefore $\mathbb{B}$ is type p if and only if (2.10) holds.

[6]Woyczynski was working in so-called $\mathcal{G}_\alpha$ which are type $(\alpha - 1)$ spaces but they are smoother than general type $(\alpha - 1)$ spaces. However, they did not use further properties of such spaces other than that relation (2.10) was satisfied. So their result does indeed hold in general type $(\alpha - 1)$ spaces.

Remark 2.4.27. For the rest of the thesis, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and whenever we discuss random variables (real or otherwise), we shall always assume they are measurable with respect to this space.

Chapter 3

Non-stochastic proof mining: the computational content of recursive inequalities

Two central results of this thesis are the quantitative versions of the martingale convergence theorem (Chapter 7) and the Robbins-Siegmund theorem (Chapter 8), which are stochastic recursive inequalities that are central in establishing the convergence of stochastic algorithms. The author's motivation in studying these results was sparked by their initial interest in the convergence of sequences of real numbers satisfying deterministic recursive inequalities. This was in part due to the observation made in [42] that the way stochastic recursive inequalities were used in establishing the convergence of stochastic algorithms shared a striking resemblance to the use of deterministic recursive inequalities and deterministic algorithms.

The purpose of this chapter is to present part of the author's initial investigation of deterministic recursive inequalities, which instigated the development of proof mining in probability which is detailed in this thesis. This chapter also aims to present a nontrivial example of some of the core features that appear in (nonstochastic) proof mining in analysis, in particular, the constructions of rates of metastabilities and the justification of such rates through the construction of Specker sequences as in Example 2.3.3.

We shall start in Section 3.1 with a brief discussion on the role recursive inequalities play in establishing the convergence of deterministic algorithms in analysis. We shall then provide a computational investigation of the main recursive inequality found in [2], including the construction of rates of metastabilities and the construction of Specker sequences justifying such rates. Then, in Section 3.2, we provide an application of our computational results, establishing the convergence of a gradient decent algorithm for nonsmooth functions on Hilbert spaces. Furthermore, we demonstrate how Theorem 2.2.14 provides a logical explanation for the quantitative result we obtained.

3.1 The computational content of a recursive inequality of Alber, Iusem and Solodov

Recursive inequalities play an important role in (nonlinear) analysis. A common way they are used is to prove that sequences of points in some space, defined by an iterative algorithm, converge to a point satisfying some properties; in other words, establishing convergence relies on reasoning about the convergence of real numbers satisfying some recursive inequalities. A straightforward example of this is the Banach fixed point theorem. Suppose (X, d) is a non-empty complete metric space and $T : X \to X$ a contractive mapping with constant $c \in [0, 1)$, that is,

$$d(T(x), T(y)) \leq cd(x, y).$$

If x^* is a fixed point of T, then the sequence of elements in X, defined by $x_{n+1} := Tx_n$, with $x_0 \in X$ an arbitrary starting point, can be shown to converge to x^*. This is done by observing that,

$$d(x_{n+1}, x^*) = d(T(x_n), T(x^*)) \leq cd(x_n, x^*).$$

Therefore, $x_n \to x^*$ follows from the fact that any sequence of real numbers satisfying the recursive inequality

$$\mu_{n+1} \leq c\mu_n \tag{3.1}$$

converges to 0, and it is not hard to see that there is an explicit rate of convergence for sequences that satisfy such inequalities, namely,

Theorem 3.1.1. *Let $\{\mu_n\}$ be a sequence of real numbers satisfying (3.1), then $f : \mathbb{Q}^+ \to \mathbb{N}$ defined as*

$$f(\varepsilon) = \left\lceil log_c \left(\frac{\varepsilon}{\mu_0} \right) \right\rceil$$

is a rate of convergence for $\{\mu_n\}$ to 0.

The proof of this result follows by iterating (3.1).

More involved recursive inequalities have been handled in the proof mining literature. An early such inequality was:

$$\mu_{n+1} \leq (1 + \delta_n)\mu_n + \gamma_n$$

and one can show that $\{\mu_n\}$ converges to some limit under the conditions $\sum_{i=0}^{\infty} \gamma_i < \infty$ and $\sum_{i=0}^{\infty} \delta_i < \infty$. This is a special case of a result due to Qihou [126], which represents the deterministic version of the Robbins-Siegmund theorem, which we analyse in Chapter 8. Rates of metastability for $\{\mu_n\}$ have been extracted and then applied to obtain, for instance, bounds on the computation of approximate fixed points of asymptotically quasi-nonexpansive mappings in [85], or for nonexpansive mappings in uniformly convex hyperbolic spaces in [86].

Another important inequality considered in the proof mining literature is

$$\mu_{n+1} \leq (1-\alpha_n)\mu_n + \gamma_n \tag{3.2}$$

where one can show that $\mu_n \to 0$ if $\sum_{i=0}^{\infty} \alpha_i = \infty$ and either $\sum_{i=0}^{\infty} \gamma_i < \infty$ or $\gamma_n/\alpha_n \to 0$. Furthermore, quantitative results giving both direct and metastable rates of convergence have been crucial in numerous different contexts; for example, (3.2) was used to extract rates of asymptotic regularity for the Halpern iterations of nonexpansive self-mappings in [106], and a detailed discussion of variants of (3.2) is given in [87]. In recent years several instances of (3.2) with combined conditions have been analysed, including [24, 32, 107].

A more general case of (3.2), also considered in the literature, is

$$\mu_{n+1} \leq \mu_n - \alpha_n\beta_n + \gamma_n \tag{3.3}$$

for $\beta_n = \psi(\mu_n)$ or $\beta_n = \psi(\mu_{n+1})$ (with suitable continuity assumptions on ψ) with rates of convergence being applied in various contexts. The variant $\beta_n = \psi(\mu_n)$ was crucial for calculating rates of convergence for generalised asymptotically weakly contractive mappings [125], while the second variant $\beta_n = \psi(\mu_{n+1})$ was initially used in [84] to extract rates of convergence for pseudocontractive mappings. More recently, it has featured in [91, 136], in the context of obtaining quantitative results for algorithms involving set-valued accretive operators and jointly nonexpansive mappings respectively.

In [115], a detailed quantitative analysis of (3.3) in its full generality is given, from which many of the aforementioned results become special cases. We do not present the full analysis. However, we highlight a small section and give an application of our analysis to convex optimization. In this regard, we start with the following result of Alber, Iusem and Solodov [2]:

Theorem 3.1.2 (cf. Proposition 2 of [2])**.** *Suppose that $\{\alpha_n\}$ and $\{\beta_n\}$ are sequences of nonnegative real numbers with $\sum_{i=0}^{\infty} \alpha_i = \infty$ and $\sum_{i=0}^{\infty} \alpha_i\beta_i < \infty$. Then whenever there exists $\theta > 0$ such that the following condition holds:*

$$\beta_n - \beta_{n+1} \leq \theta\alpha_n \quad \textit{for all} \quad n \in \mathbb{N}$$

Then $\beta_n \to 0$.

Proof. Fix $\varepsilon > 0$ and let $N \in \mathbb{N}$ be such that $\sum_{i=N}^{\infty} \alpha_i\beta_i \leq \varepsilon^2/\theta$. We claim that $\beta_n \leq 2\varepsilon$ for all $n \geq N$, and then we are done. Suppose this were not the case and there exists $n \geq N$ with $\beta_n > 2\varepsilon$. First we note that there is some $m > 0$ with $\beta_{n+m} \leq \varepsilon$, otherwise we would have

$$\sum_{i=n+1}^{\infty} \alpha_i < \frac{1}{\varepsilon} \sum_{i=n+1}^{\infty} \alpha_i\beta_i < \infty$$

Now let $m > 0$ be the minimum such m, so that $\beta_n > 2\varepsilon$, $\beta_i > \varepsilon$ for $i = n, \ldots, n+m-1$ and $\beta_{n+m} \leq \varepsilon$. Then

$$\varepsilon < \beta_n - \beta_{n+m} = \sum_{i=n}^{n+m-1} (\beta_i - \beta_{i+1}) \leq \theta \sum_{i=n}^{n+m-1} \alpha_i < \frac{\theta}{\varepsilon} \sum_{i=n}^{n+m-1} \alpha_i\beta_i \leq \frac{\theta}{\varepsilon} \sum_{i=N}^{\infty} \alpha_i\beta_i \leq \varepsilon$$

a contradiction. □

Before we present a quantitative version of the above theorem, we give a computational interpretation of the sum of a sequence of nonnegative numbers diverging.

Definition 3.1.3. Suppose that $\{\alpha_n\}$ is a sequence of nonnegative real numbers such that $\sum_{i=0}^{\infty} \alpha_i = \infty$. A function $r : \mathbb{N} \times \mathbb{Q}^+ \to \mathbb{N}$ is a *rate of divergence* for $\sum_{i=0}^{\infty} \alpha_i = \infty$ if

$$\forall n \in \mathbb{N} \forall x \in \mathbb{Q}^+ \left(\sum_{i=n}^{r(n,x)} \alpha_i \geq x \right)$$

and the following monotonicity assumption is met:

$$m \leq n \to r(m, x) \leq r(n, x)$$

for all $m, n \in \mathbb{N}$ and $x \in \mathbb{Q}^+$.

Remark 3.1.4. Given a function, r, satisfying the first part of the definition of a rate of divergence, one can construct a rate of divergence by setting

$$\tilde{r}(n, x) := \max\{r(k, x) \mid k \leq n\}.$$

Then $\tilde{r}$ is a rate of divergence since for all n and x, $\tilde{r}(n, x) = r(k, x) \geq r(n, x)$ for some $k \leq n$, and so

$$\sum_{i=n}^{\tilde{r}(n,x)} b_i \geq \sum_{i=n}^{r(n,x)} b_i \geq x.$$

A rate of divergence is the most natural direct computational interpretation one can give to a diverging series of nonnegative real numbers, and this notion is used throughout the proof mining literature (see [23, 121] for some recent examples where this notion is used).

We can now give a quantitative version of Theorem 3.1.2:

Theorem 3.1.5. *Suppose that $\{\alpha_n\}$ and $\{\beta_n\}$ are sequences of nonnegative real numbers and r is a rate of divergence for $\sum_{i=0}^{\infty} \alpha_i = \infty$, and that there is some $\theta > 0$ such that $\beta_n - \beta_{n+1} \leq \theta\alpha_n$ for all $n \in \mathbb{N}$. Then:*

(a) If $\sum_{i=0}^{\infty} \alpha_i\beta_i < \infty$ with rate of metastability Φ, then $\beta_n \to 0$ with rate of metastability

$$\Psi(\varepsilon, g) := \Phi\left(\frac{\varepsilon^2}{4\theta}, h\right) \quad \textit{for} \quad h(n) := r\left(n + g(n), \frac{\varepsilon}{2\theta}\right) - n.$$

(b) If $\sum_{i=0}^{\infty} \alpha_i\beta_i < \infty$ with rate of convergence ϕ, then $\beta_n \to 0$ with rate of convergence

$$\psi(\varepsilon) := \phi\left(\frac{\varepsilon^2}{4\theta}\right).$$

Proof. Part (b) is immediate from part (a) and Theorem 2.3.11. For part (a), Fix $\varepsilon > 0$ and $g : \mathbb{N} \to \mathbb{N}$ and take $N \leq \Psi(\varepsilon, g) = \Phi(\varepsilon^2/4\theta, h)$ such that

$$\sum_{i=N}^{r(N+g(N),\varepsilon/2\theta)} \alpha_i\beta_i \leq \frac{\varepsilon^2}{4\theta}.$$

Suppose for contradiction that $\beta_n > \varepsilon$ for some $n \in [N, N + g(N)]$. We first show that there exists some $m \in [n, r(n, \varepsilon/2\theta)]$ with $\beta_m \leq \varepsilon/2$: If this were not the case then using monotonicity of r in its first component we would have

$$\frac{\varepsilon}{2\theta} \leq \sum_{i=n}^{r(n,\varepsilon/2\theta)} \alpha_i < \frac{2}{\varepsilon} \sum_{i=n}^{r(n,\varepsilon/2\theta)} \alpha_i\beta_i \leq \frac{2}{\varepsilon} \sum_{i=N}^{r(N+g(N),\varepsilon/2\theta)} \alpha_i\beta_i \leq \frac{\varepsilon}{2\theta}$$

which is a contradiction where, for the third inequality, we use that $n \in [N, N+g(N)]$ together with the assumption that r is monotone in its first argument. Now let $n < m \leq r(n, \varepsilon/4\theta)$ be the least such index such that $\beta_n > \varepsilon$, $\beta_i > \varepsilon/2$ for $i = n, \ldots, m-1$ and $\beta_m \leq \varepsilon/2$. Then since $N \leq n$ we have

$$\frac{\varepsilon}{2} < \beta_n - \beta_m \leq \sum_{i=n}^{m-1} (\beta_i - \beta_{i+1}) \leq \theta \sum_{i=n}^{m-1} \alpha_i < \frac{2\theta}{\varepsilon} \sum_{i=n}^{m-1} \alpha_i\beta_i \leq \frac{2\theta}{\varepsilon} \sum_{i=N}^{r(N+g(N),\varepsilon/2\theta)} \alpha_i\beta_i \leq \frac{\varepsilon}{2}$$

and so we have our contradiction. □

From the above and Theorem 2.3.5, we can obtain a rate of metastability for the conclusion of Theorem 3.1.2 given a bound for $\sum_{i=0}^{\infty} \alpha_i\beta_i$.

Corollary 3.1.6. *Suppose that $\{\alpha_n\}$ and $\{\beta_n\}$ are sequences of nonnegative real numbers and r is a rate of divergence for $\sum_{i=0}^{\infty} \alpha_i = \infty$, and that there is some $\theta > 0$ such that $\beta_n - \beta_{n+1} \leq \theta\alpha_n$ for all $n \in \mathbb{N}$. Then if $\sum_{i=0}^{\infty} \alpha_i\beta_i < L$ for some $L > 0$, then $\beta_n \to 0$ with rate of metastability*

$$\Psi(\varepsilon, g) := \tilde{h}^{(\lceil e \rceil)}(0) \quad \textit{for} \quad h(n) := r\left(n + g(n), \frac{\varepsilon}{2\theta}\right)$$

and

$$e := \frac{4L\theta}{\varepsilon^2}.$$

A natural question one could ask is whether one can obtain a computable rate of convergence for the conclusion of Theorem 3.1.2, given a bound for $\sum_{i=0}^{\infty} \alpha_i \beta_i$. We shall answer this in the negative via a construction akin to Example 2.3.3. We first need a couple of preliminary constructions.

Proposition 3.1.7. *Let $\{a_n\}$ be any strictly decreasing computable sequence of positive rationals that converges to 0. Define*

$$s_n := \begin{cases} a_m & \text{For the minimum } m \leq n \text{ such that } T_m \text{ halts on input } m \\ & \text{in exactly } n \text{ steps.} \\ 0 & \text{If no such } m \text{ exists.} \end{cases}$$

where T_m denotes the Turing machine with index m. Then $s_n \to 0$ but has no computable rate of convergence.

Proof. Fix $n \in \mathbb{N}$ and let N be such that any of the machines T_i which terminate on input i for $i = 0, \ldots, n-1$ do so in at most N steps. Then for any $k \geq N+1$, we have $s_k \leq a_n$: Were this not the case, then $s_k = a_m$ where $m \leq k$ is the least such that T_m halts on input m in exactly k steps. But since $\{a_n\}$ is strictly decreasing, we must have $m < n$ and therefore $k \leq N$. Since $a_n \to 0$, this therefore implies that $s_n \to 0$.

Now suppose for contradiction that $s_n \to 0$ with some computable rate of convergence ϕ. We argue that for any $k \in \mathbb{N}$, if T_k halts on input k, then it does so in less than $\psi(k) := \max\{k, \phi(a_{k+1})\}$ steps: If not and T_k halts in n steps for $\psi(k) \leq n$, then since $k \leq n$ then $s_n = a_m$ for some $m \leq k$. Thus $s_n = a_m \geq a_k > a_{k+1}$. But since ϕ is a rate of convergence for $s_n \to 0$ and $\phi(a_{k+1}) \leq n$ then $s_n \leq a_{k+1}$, a contradiction. Therefore, if ϕ were computable, then $\psi(k)$ forms a computable upper bound on the number of steps it takes T_k to halt on input k, contradicting the unsolvability of the halting problem. □

We also have the following:

Lemma 3.1.8. *There exists a computable sequence $\{s_n\}$ of rational numbers such that $\sum_{i=0}^{\infty} s_i < \infty$ and $s_n \to 0$ with no computable rate of convergence.*

Proof. Let $\{a_n\}$ and $\{s_n\}$ be as in Proposition 3.1.7 but with $\sum_{i=0}^{\infty} a_i < \infty$ (e.g. $a_n = 1/(n+1)^2$). Define $b_n := a_n$ if there exists some $k \in \mathbb{N}$ such that $s_k = a_n$, and 0 otherwise, and note that $\sum_{i=0}^{\infty} b_i \leq \sum_{i=0}^{\infty} a_i < \infty$. Let $\{s_{n_i}\}$ (respectively $\{b_{m_j}\}$), be the subsequence of $\{s_n\}$ (respectively $\{b_n\}$) consisting of the sequence's nonzero elements. Then for each index i there is exactly one j such that $s_{n_i} = b_{m_j}$, and vice-versa, where for uniqueness we note that if

$s_{n_i} = s_{n_{i'}} = b_{m_j}$, then $T_{m_j}(m_j)$ halts in exactly n_i and $n_{i'}$ steps, and thus $i = i'$. This means there is a bijection between $\{s_{n_i}\}$ and $\{b_{m_j}\}$, and therefore

$$\sum_{i=0}^{\infty} s_i = \sum_{i=0}^{\infty} s_{n_i} = \sum_{j=0}^{\infty} b_{m_j} = \sum_{j=0}^{\infty} b_j < \infty$$

where the first equality follows since $\{s_i\}$ is just $\{s_{n_i}\}$ padded out with zero elements (and similarly for the third), while the second equality follows from the fact that we can reorder the terms in series where all terms are positive. □

Remark 3.1.9. We want to thank Ulrich Kohlenbach for pointing out to us that if x_n is the Specker sequence in Example 2.3.3, then $s_n := x_{n+1} - x_n$ will be a summable computable sequence of nonnegative rationals that converges to 0 without a computable rate of convergence. Thus providing a simpler example for the previous lemma.

Theorem 3.1.10. *For any sequence of positive rationals $\{\alpha_n\}$ with $\sum_{i=0}^{\infty} \alpha_i = \infty$, together with $\theta \in \mathbb{Q}^+$, we can construct, computably in $\{\alpha_n\}$ and θ, a sequence of positive reals $\{\beta_n\}$ satisfying*

$$\beta_n - \beta_{n+1} \leq \theta\alpha_n,$$

$$\sum_{i=0}^{\infty} \alpha_i\beta_i < \infty$$

and $\beta_n \to 0$, but without a computable rate of convergence.

Proof. Take $\{s_n\}$ as in Lemma 3.1.8 and define $l : \mathbb{N} \to \mathbb{N}$ recursively with $l(0) := 0$, and $l(n+1) := k+1$ where $k \geq l(n)$ is the least number satisfying

$$\sum_{i=l(n)}^{k} \alpha_i \geq \frac{|s_n - s_{n+1}|}{\theta}$$

which is well defined by $\sum_{i=0}^{\infty} \alpha_i = \infty$. Now define $\{\beta_k\}$ as follows: $\beta_{l(n)} := s_n$, and if $l(n) < k < l(n+1)$ then

$$\beta_k := s_n + \theta\mathrm{sgn}(s_{n+1} - s_n) \sum_{i=l(n)}^{k-1} \alpha_i.$$

Since $l(n)$ is strictly increasing, $\{\beta_k\}$ is thereby defined for all $k \in \mathbb{N}$. We now show that $|\beta_k - \beta_{k+1}| \leq \theta\alpha_k$. There are two cases to deal with: If $l(n) \leq k < l(n+1) - 1$ then

$$|\beta_k - \beta_{k+1}| = |\mathrm{sgn}(s_{n+1} - s_n)| \cdot \theta\alpha_k = \theta\alpha_k$$

and if $k = l(n+1) - 1$ then :

$$\begin{aligned}
|\beta_{l(n+1)-1} - \beta_{l(n+1)}| &= \left| s_n + \operatorname{sgn}(s_{n+1} - s_n) \cdot \theta \sum_{i=l(n)}^{l(n+1)-2} \alpha_i - s_{n+1} \right| \\
&= |\operatorname{sgn}(s_{n+1} - s_n)| \cdot \left| \theta \sum_{i=l(n)}^{l(n+1)-2} \alpha_i - |s_n - s_{n+1}| \right| \\
&= |s_n - s_{n+1}| - \sum_{i=l(n)}^{l(n+1)-2} \alpha_i \\
&\leq \theta \sum_{i=l(n)}^{l(n+1)-1} \alpha_i - \sum_{i=l(n)}^{l(n+1)-2} \alpha_i \\
&= \theta \alpha_{l(n+1)-1}
\end{aligned}$$

here we use the defining property of $l(n+1)$.

To show that there is no computable rate of convergence for $\beta_n \to 0$, suppose for contradiction that ϕ is such a rate. Fixing $\varepsilon \in \mathbb{Q}^+$, we have $\beta_n \leq \varepsilon$ for all $n \geq \phi(\varepsilon)$. However, since $l(n)$ is strictly monotone, we have $l(n) \geq n$ for all $n \in \mathbb{N}$, and thus $s_n = \beta_{l(n)} \leq \varepsilon$ for all $n \geq \phi(\varepsilon)$. Therefore, ϕ is also a computable rate of convergence for $s_n \to 0$, which is not possible. Finally, we must show that $\sum_{i=0}^{\infty} \alpha_i \beta_i < \infty$. Let $c > 0$ be any upper bound on $\{s_n\}$. Using that $\beta_k \leq s_n + s_{n+1}$ for $l(n) \leq k < l(n+1)$, since for $l(n) \leq k < l(n+1)$ we have either

$$s_n \leq \beta_k \leq s_{n+1} \text{ or } s_{n+1} \leq \beta_k \leq s_n,$$

we have

$$\begin{aligned}
\sum_{i=l(n)}^{l(n+1)-1} \alpha_i \beta_i &\leq (s_n + s_{n+1}) \sum_{i=l(n)}^{l(n+1)-1} \alpha_i \\
&< (s_n + s_{n+1}) \left(\frac{|s_n - s_{n+1}|}{\theta} \right) \\
&\leq \frac{c(s_n + s_{n+1})}{\theta}.
\end{aligned}$$

Therefore, summing over the whole sequence:

$$\begin{aligned}\sum_{i=0}^{\infty}\alpha_i\beta_i &= \sum_{n=0}^{\infty}\sum_{i=l(n)}^{l(n+1)-1}\alpha_i\beta_i \\ &< (c/\theta)\sum_{n=0}^{\infty}(s_n + s_{n+1}) \\ &\leq \frac{2c}{\theta}\sum_{n=0}^{\infty}s_n < \infty.\end{aligned}$$

□

3.2 Applications in convex optimization

We now use our quantitative theorem, in the previous section, to give a quantitative version of a procedure by Alber et al. [2] for calculating the minimum of a continuous function on a convex subset of a Hilbert space.

Suppose that H is a real-valued Hilbert space, $Y \subseteq H$ a convex subset of H, and $f : H \to \mathbb{R}$ a convex and continuous function. The *ε-subdifferential* of f at $x \in H$ is defined by

$$\partial_\varepsilon f(x) := \{u \in H \mid f(y) - f(x) \geq \langle u, y - x\rangle - \varepsilon \quad \text{for all } y \in H\}$$

with the case $\varepsilon = 0$ coinciding with the usual subdifferential $\partial f(x)$. In addition, let $P_Y : H \to Y$ be the *orthogonal projection* of H into Y. In particular, we have the following properties (see [2, Proposition 3]):

$$\begin{aligned}&\|P_Y(x) - P_Y(y)\| \leq \|x - y\| \text{ for all } x, y \in H \\ &\langle x - y, x - P_Y(x)\rangle \geq 0 \text{ for all } x \in H \text{ and } y \in Y.\end{aligned} \tag{3.4}$$

For a sequence of stepsizes $\{\alpha_n\}$, satisfying

$$\sum_{i=0}^{\infty}\alpha_i = \infty \text{ and } \sum_{i=0}^{\infty}\alpha_i^2 < \infty$$

and a sequence of nonnegative error terms, $\{\varepsilon_n\}$, satisfying $\varepsilon_n \leq \mu\alpha_n$ for some $\mu > 0$, Alber et al. consider the following algorithm:

$$x_{n+1} = P_Y\left(x_n - \frac{\alpha_n}{\nu_n}u_n\right) \quad \text{for} \quad u_n \in \partial_{\varepsilon_n} f(x_n) \quad \text{with} \quad u_n \neq 0. \tag{3.5}$$

Here, $\nu_n := \max\{1, \|u_n\|\}$ (the algorithm halts if $0 \in \partial_{\varepsilon_n} f(x_n)$ at any point).

Let $x^* \in Y$ be a minimizer of f on Y, and suppose that $\{x_n\}$ is an infinite sequence generated by the above algorithm (3.5), whose components satisfy all of the properties outlined above. Suppose that $\rho > 1$ is such that $\|u_n\| \leq \rho$ for all $n \in \mathbb{N}$. Then $f(x_n) \to f(x^*)$. Quantitatively, we can obtain metastable rates for this convergence result.

Theorem 3.2.1. *Let $x^* \in Y$ be a minimizer of f on Y, and suppose that $\{x_n\}$ is an infinite sequence generated by the algorithm (3.5). Suppose that $\rho > 1$ is such that $\|u_n\| \leq \rho$ for all $n \in \mathbb{N}$, r is a rate of divergence for $\sum_{i=0}^{\infty} \alpha_i = \infty$ and $K_1, K_2 > 0$ are such that $\sum_{i=0}^{\infty} \alpha_i^2 \leq K_1$ and $\|x_0 - x^*\| \leq K_2$. Then for all $\varepsilon \in \mathbb{Q}^+$ and $g : \mathbb{N} \to \mathbb{N}$ we have*

$$\exists n \leq \Phi(\varepsilon, g)\, \forall k \in [n, n + g(n)]\, (|f(x_k) - f(x^*)| < \varepsilon)$$

where

$$\Psi(\varepsilon, g) := \tilde{h}^{(\lceil e \rceil)}(0) \quad \textit{for} \quad h(n) := r\left(n + g(n), \frac{\varepsilon}{2(\rho + \mu)}\right)$$

and

$$e := \frac{2(\rho + \mu)(\rho K_2^2 + K_1(5\rho + 2\mu))}{\varepsilon^2}.$$

Proof. Set $z_n := x_n - (\alpha_n/\nu_n)u_n$. Applying the first property of the projection map detailed in (3.4), we have, for all $n \in \mathbb{N}$,

$$\|x_{n+1} - x_n\| = \|P_Y(z_n) - P_Y(x_n)\| \leq \|z_n - x_n\| = \frac{\alpha_n}{\nu_n}\|u_n\| \leq \alpha_n. \tag{3.6}$$

Now, observe that,

$$\begin{aligned} &\frac{\alpha_n}{\nu_n}\langle u_n, x_n - x^*\rangle = \langle x_n - x^*, x_n - z_n\rangle \\ &= \langle x_n - x^*, x_n - x_{n+1}\rangle + \langle x_n - x^*, x_{n+1} - z_n\rangle \\ &= \langle x_n - x^*, x_n - x_{n+1}\rangle + \langle z_n - x_n, z_n - x_{n+1}\rangle + \langle x^* - z_n, z_n - x_{n+1}\rangle. \end{aligned} \tag{3.7}$$

By (3.4) and the fact that $P_Y(z_n) = x_{n+1}$ we have

$$\langle x^* - z_n, z_n - x_{n+1}\rangle = -\langle z_n - x^*, z_n - P_Y(z_n)\rangle \leq 0$$

and therefore, from (3.7) and $\nu_n \leq \max\{1, \rho\} \leq \rho$, we have

$$\langle \alpha_n u_n, x_n - x^*\rangle \leq \rho(\langle x_n - x^*, x_n - x_{n+1}\rangle + \langle z_n - x_n, z_n - x_{n+1}\rangle). \tag{3.8}$$

Furthermore, we see that.

$$\begin{aligned}\langle z_n - x_n, z_n - x_{n+1}\rangle &= \langle z_n - x_n, z_n - x_n\rangle + \langle z_n - x_n, x_n - x_{n+1}\rangle \\ &\leq \|z_n - x_n\|^2 + \|z_n - x_n\|\|x_n - x_{n+1}\| \\ &\leq 2\|z_n - x_n\|^2 \leq 2\alpha_n^2 \quad \text{by (3.6).}\end{aligned}$$

Substituting this into (3.8), we have

$$\langle \alpha_n u_n, x_n - x^*\rangle \leq \rho\langle x_n - x^*, x_n - x_{n+1}\rangle + 2\rho\alpha_n^2. \tag{3.9}$$

Now, setting $\beta_n := f(x_n) - f(x^*)$ and noting that $\beta_n \geq 0$ (since x^* is a minimiser of f), we have:

$$\begin{aligned}\beta_n - \beta_{n+1} &= f(x_n) - f(x_{n+1}) \\ &\leq \langle u_n, x_n - x_{n+1}\rangle + \mu\alpha_n \\ &\leq \|u_n\|\|x_n - x_{n+1}\| + \mu\alpha_n \\ &\leq (\rho + \mu)\alpha_n.\end{aligned}$$

The first inequality follows from the definition of the ε- subgradient and the fact that $\varepsilon_n \leq \mu\alpha_n$. The second inequality follows from the Cauchy-Schwartz inequality. The final inequality follows from (3.6). So the recursive inequality from Theorem 3.1.2 is satisfied with $\theta := \rho + \mu$. Now, we have

$$\begin{aligned}2\alpha_n\beta_n &\leq 2\langle \alpha_n u_n, x_n - x^*\rangle + 2\mu\alpha_n^2 \quad \text{by the definition of the } \varepsilon\text{- subgradient and } \varepsilon_n \leq \mu\alpha_n \\ &\leq 2\rho\langle x_n - x_{n+1}, x_n - x^*\rangle + 2(2\rho + \mu)\alpha_n^2 \quad \text{by (3.9)} \\ &= \rho(\|x_{n+1} - x_n\|^2 + \|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2) + 2(2\rho + \mu)\alpha_n^2 \\ &\leq \rho(\|x_n - x^*\|^2 - \|x_{n+1} - x^*\|^2) + (5\rho + 2\mu)\alpha_n^2 \quad \text{by (3.6),}\end{aligned} \tag{3.10}$$

so it follows that

$$\sum_{i=0}^{\infty} \alpha_i\beta_i \leq \frac{\rho}{2}\|x_0 - x^*\|^2 + \frac{5\rho + 2\mu}{2}\sum_{i=0}^{\infty}\alpha_i^2 \leq \frac{\rho K_2^2}{2} + \frac{K_1(5\rho + 2\mu)}{2}. \tag{3.11}$$

Thus, the result follows from Corollary 3.1.6. □

Remark 3.2.2. Note that in [2], the existence of some $\rho > 0$ satisfying $\|u_n\| \leq \rho$ for all $n \in \mathbb{N}$ follows by establishing that the $\{x_n\}$ are bounded, and then using an additional boundedness assumption for the subgradient, namely that $\partial_\varepsilon f$ is bounded on bounded sets.

Remark 3.2.3. Observe that the rate we obtain in the above theorem is independent of the space (in particular, the norm and inner product operators) and only depends on elements of the sequences in the theorem through bounds on their norms. We conjecture that this uniformity

(as well as the success of the extraction of quantitative data) can be explained by the fact that this result and proof can be formalised in an extension of $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ in which Theorem 2.2.14 still holds. We do not substantiate all the details of this claim here, but only provide a sketch of how this would be done. First, we expand our language with constants representing the key features of the theorem: $x^{X(0)}$ representing the converging sequence $\{x_n\}$ in the algorithm, $f^{1(X)}$ representing the function f we are minimising and $u^{X(0)}, \varepsilon^{X(0)}$ representing $\{u_n\}$ and $\{\varepsilon_n\}$ respectfully. Furthermore, we can introduce the subset Y as a characteristic function through a constant $Y^{0(X)}$. We could similarly do this for the ε-subdifferential. However, in the proof we only use $u_n \in \partial_{\varepsilon_n} f(x_n)$, that is

$$\forall n^0, y^X \, (f(y) - f(x(n)) \geq_{\mathbb{R}} \langle u(n), y - x(n)\rangle - \varepsilon(n))$$

which we introduce as an axiom (observing that it is universal and thus admissible as an axiom in an extension for which Theorem 2.2.14 holds). We can further introduce the projection operator as a constant $P^{X(X)}$ along with the axiom

$$\forall x^X \, (Y(P(x)) = 1)$$

which is universal and universal axioms expressing the crucial properties we use, namely (3.4).

We also add constants $\alpha^{1(0)}$ representing $\{\alpha_n\}$ and $r^{1(0)(0)}$, representing a rate of divergence for $\sum_{i=0}^{\infty} \alpha_i = \infty$ along with a universal axiom expressing this fact. In addition, we add constants K_1^0, K_2^0 and ρ^0.

The remaining assumptions of the theorem are then added as universal axioms. We then claim that Theorem 2.2.14 extended to the above sketched system, holds.

Chapter 4

Proof-theoretic aspects of probability theory

As the case studies of proof mining in probability increased, observations were made about the uniformity of the extracted computational content (such as their independence from the probability space and other parameters from the results being analysed). A key feature the proof mining metatheorems of the past shared was their ability to explain the uniformities in extracted computational content from abstract spaces and other parameters of theorems through the use of extensions of Bezem's majorization [14]. It was, therefore, natural to ask whether one could develop a formal system and corresponding metatheorem for probability theory, which could provide a logical explanation for the success of previous case studies, the observed uniformities of the extracted computational content, and guide future proof mining program extractions. Such a result was obtained by the author and Pischke in [114].

Another abstract observation that was made in the development of the author's work on proof mining in probability theory was the appearance of a systematic way in which one obtains quantitative stochastic concepts from deterministic notions. In particular, concerning the different modes of quantitative convergence detailed in Section 2.3. In collaboration with Powell [117], the author developed a systematic approach for obtaining stochastic notions from deterministic ones. This abstract investigation led to the development of notions of learnable uniform and learnable pointwise rates of convergence (c.f. Definition 4.2.21), which was crucial in obtaining our computational interpretations of the martingale convergence theorem (Chapter 7) and the Robbins-Siegmund theorem (Chapter 8).

This chapter aims to detail the author's and collaborators' theoretical contributions discussed above. We start in Section 4.1, where we present a formal system for reasoning about probability theory amenable to program extraction in the context of the proof mining program. Furthermore, we state the metatheorem obtained in [114]. We then continue in Section 4.2 and present an abstract approach to establish stochastic notions from deterministic ones. In par-

ticular, this section shall provide important quantitative stochastic notions (which we motivate through our general approach) that we shall need throughout this thesis. Some of these notions were known to logicians and probability theorists, and some are new and arise through our abstract approach. Lastly, we demonstrate how one can formalise the poof of the equivalence of pointwise and uniform metastability (introduced in [5] but also rediscovered via our abstract approach) proved in [5] and how the metatheorem of [114] explains the uniformities and bar recursive complexity of the extracted computational result.

Therefore, this chapter not only presents the author's theoretical contributions to proof mining in probability theory but also gives the crucial background required for the remaining chapters of this thesis. In particular, we introduce several definitions, which we then use freely throughout the thesis.

4.1 A formal system for probability theory amenable to program extraction

As we did for inner product spaces in Section 2.2.2, we present a formal system for reasoning about the quantitative aspects of probability theory and corresponding metatheorems that logically explain case studies that extract the computational content of results in probability theory.

As is common in proof mining, our system will be an extension of the system $\mathcal{A}^\omega$, which we presented in Section 2.1, with abstract types for reasoning about the notions from probability theory. The main problem one encounters when trying to develop such a system is that the defining axioms for countable unions and the countable additivity of the probability measure are not naturally admissible in such a system that allows for a metatheorem like that of Theorem 2.2.14. The key insight for the success of the development of such a system, which was mainly brought about by the increasing number of case studies in obtaining quantitative results, was that infinite unions are used in minimal ways. Furthermore, many of the quantitative results obtained in case studies directly apply to probability contents (c.f. Section 2.4). Thus, empirically, the theory of probability contents appears to be robust enough to handle the interesting aspects of quantitative probability theory.

In this section, we shall demonstrate that the entire theory of probability contents can be formalised in a system amenable to program extraction. This thus provides a logical explanation of the successes of the extraction of computational content of many results in probability theory. Furthermore, by extending Bezem's notion of majorization [14], as we did in Definition 2.2.12, the metatheorem we present also provides a logical explanation for the uniformity observed in the extracted computational content of results in probability theory.

4.1.1 The formal system

As previously mentioned, our formal system will be an extension of $\mathcal{A}^\omega$, which we presented in Section 2.1, and thus we will keep the same notations and definitions as in this section.

It will be convenient for us to be able to refer to intervals of real numbers intentionally. In this regard, we have the following:

Definition 4.1.1. We write $\mathcal{A}^\omega[\text{Int}]$ for the system resulting from $\mathcal{A}^\omega$ extended with the constants $[\cdot,\cdot]$ of type $0(1)(1)(1)$ and the axioms (we use the abbreviation $r \in [a,b]$ for $[a,b](r) = 1$):

Axiom	Interpretation
$\forall a^1, b^1, r^1\,([a,b](r) \leq_0 1)$	$[\cdot,\cdot]$ represents an indicator function
$\forall a^1, b^1, r^1\,(r \in [a,b] \to a \leq_{\mathbb{R}} r \leq_{\mathbb{R}} b)$	The points in the interval are in between the end points
$\forall a^1, b^1, r^1\,(a <_{\mathbb{R}} r <_{\mathbb{R}} b \to r \in [a,b])$	The points strictly in between the endpoints are in the interval
$\forall a^1, b^1\,(a, b \in [a,b])$	The endpoints are in the interval

Remark 4.1.2. Observe that $\mathcal{A}^\omega[\text{Int}]$ extends $\mathcal{A}^\omega$ by new constant and universal axioms (after expanding the hidden quantifiers in the relations on real numbers c.f. Section 2.1.2) and so we have theorem 2.2.8 holds for $\mathcal{A}^\omega[\text{Int}]$ (c.f. Remark 2.2.3).

We now present a system for reasoning about algebras. In this regard, we extend the set of types T, used to develop $\mathcal{A}^\omega$, by two new abstract types Ω and S and form the extended set of types $T^{\Omega,S}$ defined by

$$0, \Omega, S \in T^{\Omega,S}, \quad \rho, \tau \in T^{\Omega,S} \to \rho(\tau) \in T^{\Omega,S}.$$

Here, Ω is an abstract type representing the sample space, and S represents the algebra. As we did for normed spaces in Section 2.1.3, we then reformulate $\mathcal{A}^\omega[\text{Int}]$ over the new set of types $\text{T}^{\Omega,S}$, where we have additional constants and additional axioms that now refer to the additional types. Over this new reformulation of $\mathcal{A}^\omega[\text{Int}]$, we add the constants:

Constant	Type	Interpretation
eq	$0(\Omega)(\Omega)$	Equality on Ω
$\in$	$0(S)(\Omega)$	Element relation from between Ω and S
$\cup$	$S(S)(S)$	Union operator
$(\cdot)^c$	$S(S)$	Complement operator
$\emptyset$	S	Empty set
c_Ω	Ω	Witness of the nonemptiness of Ω

We follow obvious abbreviations to enhance readability. For example, we write:

- A^c for $(A)^c$

- $x \in A$ for $\in(x, A) =_0 1$
- $x \notin A$ for $\in(x, A) \neq_0 0$
- $A \cup B$ for $\cup(A, B)$
- $x =_\Omega y$ for $\mathrm{eq}(x, y) =_0 1$

In addition, we define $\Omega := \emptyset^c$ (this should not be confused with the type Ω, but it will be made clear from the context which is intended), and we introduce intersection via the following abbreviation:

$$A \cap B := (A^c \cup B^c)^c$$

for terms A^S, B^S.

We introduce arbitrary finite unions by further abbreviations. For a sequence of events $A^{S(0)}$ and two natural numbers $n^0 \leq_0 m^0$, we use the abbreviation,

$$\bigcup_{i=n}^{m} A(i) := R_S(m - n, A(n), \lambda B, x.(B \cup A(n + x + 1)))$$

where R_S is a (single) type S recursor constant. For $m <_0 n$, we set $\bigcup_{i=n}^{m} A(i) := \emptyset$. Furthermore, we write

$$\bigcap_{i=n}^{m} A(i) := \left(\bigcup_{i=n}^{m} (A(i))^c \right)^c.$$

We introduce equality on S via the following abbreviation: for A^S and B^s, we define

$$A =_S B :\equiv \forall x^\Omega (x \in A \leftrightarrow x \in B).$$

and we introduce the abbreviation

$$A \subseteq_S B :\equiv \forall x^\Omega \, (x \in A \to x \in B)$$

for A, B of type S.

Definition 4.1.3. We write $\mathcal{F}^\omega$ for the system resulting from $\mathcal{A}^\omega[\mathrm{Int}]$ over the augmented language including the types Ω, S (where all the respective constants and axioms now are allowed to also refer to these new types, if applicable) extended with the constants $\mathrm{eq}, \in, \cup, (\cdot)^c, \emptyset, c_\Omega$ (with types given in the above table) and the axioms

Axiom	Interpretation
$\forall x^\Omega, y^\Omega(\mathrm{eq}(x,y) \leq_0 1)$	eq represents an indicator function
$\forall x^\Omega, y^\Omega, z^\Omega\,(x =_\Omega x \wedge (x =_\Omega y \to y =_\Omega x) \wedge (x =_\Omega y \wedge y =_\Omega z \to x =_\Omega z))$	eq is an equivalence relation
$\forall x^\Omega \forall A^S(\in(x,A) \leq_0 1)$	$\in$ represents an indicator function
$\forall x^\Omega(x \notin \emptyset)$	The empty set is empty
$\forall x^\Omega \forall A^S, B^S(x \in A \cup B \leftrightarrow x \in A \vee x \in B)$	Characterising property of the union
$\forall x^\Omega \forall A^S(x \in A^c \leftrightarrow x \notin A)$	Characterising property of the complement

All of the basic properties of the above operations on algebras are provable in $\mathcal{F}^\omega$. In particular, $=_S$ is provably an equivalence relation and $\subseteq_S$ forms a partial order with respect to equality defined by $=_S$. Furthermore, the operations $\cup$ and $(\cdot)^c$ are provably extensional in $\mathcal{F}^\omega$, and it can be shown that the extensionality of the union extends to arbitrarily finite unions by induction.

We now introduce a system for reasoning about probability contents:

Definition 4.1.4. We write $\mathcal{F}^\omega[\mathbb{P}]$ for the system resulting from $\mathcal{F}^\omega$ extended with the constant $\mathbb{P}$ of type $1(S)$ and the axioms

Axiom	Interpretation
$\forall A^S(0 \leq_\mathbb{R} \mathbb{P}(A) \leq_\mathbb{R} 1)$	The probability of any event is always nonnegative and less than 1
$\mathbb{P}(\emptyset) =_\mathbb{R} 0$	The probability of the empty set is zeo
$\forall A^S, B^S(\mathbb{P}(A \cup B) =_\mathbb{R} \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B))$	Generalised additivity
$\forall A^S, B^S\,(A \subseteq_S B \to \mathbb{P}(A) \leq_\mathbb{R} \mathbb{P}(B))$	Monotonicity

Proposition 4.1.5. *The following properties of* $\mathbb{P}$ *are provable in* $\mathcal{F}^\omega[\mathbb{P}]$*:*

1. $\mathbb{P}$ *is extensional w.r.t.* $=_S$ *and* $=_\mathbb{R}$*, i.e.*

$$\forall A^S, B^S\,(A =_S B \to \mathbb{P}(A) =_\mathbb{R} \mathbb{P}(B))\,.$$

2. $\mathbb{P}$ *is definite on* $\emptyset$*, i.e.*

$$\forall A^S\,(\mathbb{P}(A) >_\mathbb{R} 0 \to A \neq_S \emptyset)\,.$$

3. $\mathbb{P}$ *is additive, i.e.*

$$\forall A^S, B^S(A \cap B =_S \emptyset \to \mathbb{P}(A \cup B) =_\mathbb{R} \mathbb{P}(A) + \mathbb{P}(B)).$$

4. $\mathbb{P}$ *respects the relative complements of subsets, i.e.*

$$\forall A^S, B^S(B \subseteq_S A \to \mathbb{P}(A \cap B^c) =_\mathbb{R} \mathbb{P}(A) - \mathbb{P}(B)).$$

In particular, we also have

$$\forall A^S(\mathbb{P}(A^c) =_\mathbb{R} 1 - \mathbb{P}(A)).$$

5. $\mathbb{P}$ *satisfies Boole's inequality, i.e.*

$$\forall A^{S(0)}, n^0 \left(\mathbb{P}\left(\bigcup_{i=0}^{n} A(i) \right) \leq_{\mathbb{R}} \sum_{i=0}^{n} \mathbb{P}(A(i)) \right).$$

Proof. 1. Assume $\mathbb{P}(A) > \mathbb{P}(B)$. By the monotonicity axiom, there exists an x such that $x \in A$ and $x \notin B$, i.e. $A \neq B$. Similarly we derive $A \neq B$ from $\mathbb{P}(A) < \mathbb{P}(B)$. Combined, we get that $A = B$ implies $\mathbb{P}(A) = \mathbb{P}(B)$.

2. Assume $\mathbb{P}(A) > 0 = \mathbb{P}(\emptyset)$. Then if $A = \emptyset$, we have $\mathbb{P}(A) = 0$ by the extensionality of $\mathbb{P}$ (part 1).

3. Let A, B be arbitrary with $A \cap B = \emptyset$. By the generalised additivity axiom, we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. As $\mathbb{P}$ is extensional, we get $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$ so that the above implies $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ as desired.

4. Let $E := A \cap B$ and $F := A \cap B^c$. Then $E \cap F = \emptyset$ (by the properties of algebras of sets). Thus, by additivity, $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$. We have that $E \cup F = A$ (again by the properties of algebras of sets). Thus, by extensionality of $\mathbb{P}$, we have $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$. Now, over $\mathcal{F}^\omega$, $B \subseteq A$ is equivalent to $A \cap B = B$, so the result follows from the extensionality of $\mathbb{P}$.

5. This follows via a simple induction and the generalised additivity axiom.

□

Contents on algebras enjoy certain continuity properties similar to continuity from above and below for measures but without the existence of limiting sets, i.e. infinite unions, etc. (see, e.g. [15]), and we now discuss how the system $\mathcal{F}^\omega[\mathbb{P}]$ recognizes Cauchy-variants of these properties.

For that, we introduce the following operation on terms of type $S(0)$ that allows for the implicit quantification over a disjoint countable family of sets: given $A^{S(0)}$, we set $(A\!\uparrow)(0) = A(0)$ and

$$(A\!\uparrow)(n+1) := A(n+1) \cap \left(\bigcup_{i=0}^{n} A(i) \right)^c.$$

This operation thus turns A into a sequence of disjoint sets $A\!\uparrow$ with the same (partial) union(s), and if A was already a disjoint family, then it is left unchanged by the operation.

We now begin with a Cauchy-type form of σ-additivity of $\mathbb{P}$ as a content. For this, note that for a given $A^{S(0)}$, the sequence of partial sums

$$\sum_{i=0}^{n} \mathbb{P}((A\!\uparrow)(i)) = \mathbb{P}\left(\bigcup_{i=0}^{n} (A\!\uparrow)(i) \right) = \mathbb{P}\left(\bigcup_{i=0}^{n} A(i) \right)$$

is a monotone and bounded sequence of real numbers and thus is Cauchy:

Lemma 4.1.6 (folklore, see essentially [80])**.** *The system* WE-PA$^\omega$ *proves that*

$$\begin{aligned}\forall a^{1(0)}\big(&\forall n^0\,(0 \leq_{\mathbb{R}} a(n) \leq_{\mathbb{R}} 1 \wedge a(n) \leq_{\mathbb{R}} a(n+1))\\ &\to \forall k^0 \exists N^0 \forall n^0, m^0 \geq_0 N\,\big(|a(n) - a(m)| <_{\mathbb{R}} 2^{-k}\big)\big).\end{aligned}$$

So, instantiating the above result with $a(n) = \sum_{i=0}^{n} \mathbb{P}((A\uparrow)(i))$, we can derive that $\mathcal{F}^\omega[\mathbb{P}]$ can prove the Cauchy-property of sequences of contents of increasing disjoint unions:

Proposition 4.1.7. *The system* $\mathcal{F}^\omega[\mathbb{P}]$ *proves*

$$\forall A^{S(0)} \forall k^0 \exists N^0 \forall n^0, m^0 \geq_0 N\left(\left|\sum_{i=0}^{n} \mathbb{P}((A\uparrow)(i)) - \sum_{i=0}^{m} \mathbb{P}((A\uparrow)(i))\right| <_{\mathbb{R}} 2^{-k}\right).$$

4.1.2 The program extraction theorem

We present a program extraction theorem for $\mathcal{F}[\mathbb{P}]$, similar to that of Theorem 2.2.14. In this regard, we first introduce the structure of strongly majorizable functionals that will be a model of $\mathcal{F}[\mathbb{P}] + (\mathrm{BR})$.[1]

We first define the operator $\widehat{\cdot}$ by recursion as,

$$\widehat{0} := 0,\ \widehat{\Omega} := 0,\ \widehat{S} := 0,\ \widehat{\tau(\xi)} := \widehat{\tau}(\widehat{\xi}).$$

Definition 4.1.8. Let Ω be a non-empty set, $S \subseteq 2^\Omega$ be an algebra and $\mathbb{P}$ be a probability content on S. The structure $\mathcal{M}^{\omega,\Omega,S}$ and the majorizability relation $\gtrsim_\rho$ are defined by

$$\begin{cases}\mathcal{M}_0 := \mathbb{N}, n \gtrsim_0 m := n \geq m \wedge n, m \in \mathbb{N},\\ \mathcal{M}_\Omega := \Omega, n \gtrsim_\Omega x := n \geq \mathbb{P}(\Omega) \wedge n \in \mathcal{M}_0, x \in \mathcal{M}_\Omega,\\ \mathcal{M}_S := S, n \gtrsim_S A := n \geq \mathbb{P}(A) \wedge n \in \mathcal{M}_0, A \in \mathcal{M}_S,\\ f \gtrsim_{\tau(\xi)} x := f \in \mathcal{M}_{\widehat{\tau}}^{\mathcal{M}_{\widehat{\xi}}} \wedge x \in \mathcal{M}_\tau^{\mathcal{M}_\xi}\\ \qquad\qquad \wedge \forall g \in \mathcal{M}_{\widehat{\xi}}, y \in \mathcal{M}_\xi (g \gtrsim_\xi y \to fg \gtrsim_\tau xy)\\ \qquad\qquad \wedge \forall g, y \in \mathcal{M}_{\widehat{\xi}} (g \gtrsim_{\widehat{\xi}} y \to fg \gtrsim_{\widehat{\tau}} fy),\\ \mathcal{M}_{\tau(\xi)} := \left\{x \in \mathcal{M}_\tau^{\mathcal{M}_\xi} \mid \exists f \in \mathcal{M}_{\widehat{\tau}}^{\mathcal{M}_{\widehat{\xi}}} : f \gtrsim_{\tau(\xi)} x\right\}.\end{cases}$$

Furthermore, we can also extend the concept of type Δ formulas (c.f. Section 2.2.2) to this context. A formula of type Δ is still any of the form

$$\forall \underline{a}^\delta \exists \underline{b} \leq_\sigma \underline{ra} \forall \underline{c}^\gamma A_{qf}(\underline{a}, \underline{b}, \underline{c})$$

[1] Here (BR) is now extended to the new set of abstract types $\mathrm{T}^{\Omega,S}$.

where A_{qf} is quantifier-free, the types in $\underline{\delta}$, $\underline{\sigma}$ and $\underline{\gamma}$ are admissible, $\underline{r}$ is a tuple of closed terms of the appropriate types, but now $\leq$ is defined by recursion on the type via:

1. $x \leq_0 y := x \leq_0 y$.
2. $x \leq_\Omega y := \mathbb{P}(\Omega) \leq_{\mathbb{R}} \mathbb{P}(\Omega)$.
3. $A \leq_S B := \mathbb{P}(A) \leq_{\mathbb{R}} \mathbb{P}(B)$.
4. $x \leq_{\tau(\xi)} y := \forall z^\xi (xz \leq_\tau yz)$.

In this context we call ρ small if it is of the form $\rho = \rho_0(0)\dots(0)$ for $\rho_0 \in \{0, \Omega, S\}$ (including $0, \Omega, S$) and call it admissible if it is of the form $\rho = \rho_0(\tau_k)\dots(\tau_1)$ where each τ_i is small and $\rho_0 \in \{0, \Omega, S\}$ (also including $0, \Omega, S$).

We now have the following program extraction theorem:

Theorem 4.1.9 ([114])**.** *Let $\beth$ be a set of formulas of type Δ. Let τ be admissible, δ be of degree 1 and s be a closed term of $\mathcal{C}^\omega$ of type $\sigma(\delta)$ for admissible σ and let $B_\forall(x,y,z,u)/C_\exists(x,y,z,v)$ be $\forall$-/$\exists$-formulas of $\mathcal{C}^\omega$ with only $x,y,z,u/x,y,z,v$ free. If*

$$\mathcal{C}^\omega + \beth \vdash \forall x^\delta \forall y \leq_\sigma s(x) \forall z^\tau \left(\forall u^0 B_\forall(x,y,z,u) \to \exists v^0 C_\exists(x,y,z,v)\right),$$

then one can extract a partial functional $\Phi : \mathcal{S}_\delta \times \mathcal{S}_{\widehat{\tau}} \rightharpoonup \mathbb{N}$ which is total and (bar-recursively) computable on $\mathcal{M}_\delta \times \mathcal{M}_{\widehat{\tau}}$ and such that for all $x \in \mathcal{S}_\delta$, $z \in \mathcal{S}_\tau$, $z^* \in \mathcal{S}_{\widehat{\tau}}$, if $z^* \gtrsim z$, then

$$\mathcal{S}^{\omega,\Omega,S} \models \forall y \leq_\sigma s(x) \left(\forall u \leq_0 \Phi(x,z^*) B_\forall(x,y,z,u) \to \exists v \leq_0 \Phi(x,z^*) C_\exists(x,y,z,v)\right)$$

holds whenever $\mathcal{S}^{\omega,\Omega,S} \models \beth$ for $\mathcal{S}^{\omega,\Omega,S}$ defined via any non-empty set Ω and any algebra $S \subseteq 2^\Omega$ together with any probability content $\mathbb{P}$ on S.

Further:

1. *If $\widehat{\tau}$ is of degree 1, then Φ is a total computable functional.*
2. *We may have tuples instead of single variables x,y,z,u,v and a finite conjunction instead of a single premise $\forall u^0 B_\forall(x,y,z,u)$.*
3. *If the claim is proved without DC, then τ may be arbitrary and Φ will be a total functional on $\mathcal{S}_\delta \times \mathcal{S}_{\widehat{\tau}}$ which is primitive recursive in the sense of Gödel [52] and Hilbert [58].*

The proof of the above result can be found in [114].

Remark 4.1.10. We noted in Remark 2.2.11 that a key aspect of the program extraction theorem for $\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ is that the norm space axioms added to $\mathcal{A}^\omega$ were purely universal. This is not

the case for $\mathcal{F}^\omega[\mathbb{P}]$. The monotonicity axiom

$$\forall A^S, B^S \, (A \subseteq_S B \to \mathbb{P}(A) \leq_{\mathbb{R}} \mathbb{P}(B))$$

is not purely universal (due to the hidden quantifiers in $\leq_{\mathbb{R}}$). However one can show this axiom is equivalent over $\mathcal{F}^\omega[\mathbb{P}]$ to

$$\forall A^S, B^S \exists x^\Omega \leq_\Omega c_\Omega(\mathbb{P}(A) >_{\mathbb{R}} \mathbb{P}(B) \to (x \in A \wedge x \notin B)),$$

thus, is of type Δ. This is a critical feature in the proof of the above program extraction theorem.

4.2 Quantitative notions of probabilistic convergence

In Section 2.3, we presented various computational interpretations for the convergence of real numbers.[2] We shall now look to extend these interpretations in the probabilistic setting. In this regard, we present a general abstract framework in which one can transfer quantitative deterministic notions into natural probabilistic analogues. Although the notions we introduce can naturally be formalised in $\mathcal{F}^\omega[\mathbb{P}]$ (justifying further the strength of the theory), to preserve some concreteness and allow for easier comparison with notions from Section 2.3, we opt to step away from the development of these notions in a formal system and present them in a normal mathematical setting.

In this section, we shall introduce a number of definitions, which we shall then use freely throughout the remainder of this thesis. Furthermore, the concepts of learnable uniform and pointwise rates (c.f. Definition 4.2.21) shall be crucial in obtaining our quantitative results in Chapters 7 and 8.

4.2.1 Quantitative almost sure statements

We start by outlining a general approach to providing quantitative versions of probabilistic statements. Our approach will allow us to rediscover known quantitative notions, such as stochastic analogues to convergence and fluctuations, as well as develop new concepts. Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 4.2.1. We say that a logical formula $\varphi(\omega, x_1, \ldots, x_n)$ with parameters $x_1, \ldots, x_n$ and ω, a variable taking values in Ω, is *measurable* if for all parameters $x_1, \ldots, x_n$, we have

[2]Some of these notions (rates of convergences and bounds of the fluctuations, for example) naturally lift to sequences taking values in general metric spaces.

$\{\omega \in \Omega : \varphi(\omega, x_1, \ldots, x_n)\} \in \mathcal{F}$. For such a measurable formula, we write $\varphi(x_1, \ldots, x_n) := \{\omega \in \Omega : \varphi(\omega, x_1, \ldots, x_n)\}$.

If $\varphi(\omega, n)$ is a measurable formula, with $n \in \mathbb{N}$, we define $\exists n\, \varphi(n), \forall n\, \varphi(n) \in \mathcal{F}$ in the expected way:

$$\exists n\, \varphi(n) := \bigcup_{n\in\mathbb{N}} \varphi(n) \quad \text{and} \quad \forall n\, \varphi(n) := \bigcap_{n\in\mathbb{N}} \varphi(n).$$

The following straightforward facts will be used repeatedly:

Lemma 4.2.2. *Let $p \in [0,1]$ and $\varphi(\omega, n)$ be a measurable formula satisfying $\varphi(n) \supseteq \varphi(n+1)$ for all $n \in \mathbb{N}$. Then:*

(i) $\mathbb{P}(\forall n\, \varphi(n)) \geq p \iff \forall n\, (\mathbb{P}(\varphi(n)) \geq p)$.

(ii) $\mathbb{P}(\forall n\, \varphi(n)) \leq p \iff \forall \lambda > 0\, \exists n\, (\mathbb{P}(\varphi(n)) < p + \lambda)$.

On the other hand, if $\varphi(n) \subseteq \varphi(n+1)$ for all $n \in \mathbb{N}$, then:

(iii) $\mathbb{P}(\exists n\, \varphi(n)) \leq p \iff \forall n\, (\mathbb{P}(\varphi(n)) \leq p)$.

(iv) $\mathbb{P}(\exists n\, \varphi(n)) \geq p \iff \forall \lambda > 0\, \exists n\, (\mathbb{P}(\varphi(n)) > p - \lambda)$.

Proof. Parts (i) and (ii) follow directly from the fact that $\{\mathbb{P}(\varphi(n))\}$ is a decreasing sequence of reals with

$$\lim_{n\to\infty} \mathbb{P}(\varphi(n)) = \mathbb{P}(\forall n\, \varphi(n)).$$

If $\mathbb{P}(\forall n\, \varphi(n)) \geq p$ then $\mathbb{P}(\varphi(n)) \geq \mathbb{P}(\forall n\, \varphi(n)) \geq p$ for any $n \in \mathbb{N}$, and conversely if $\mathbb{P}(\varphi(n)) \geq p$ for all $n \in \mathbb{N}$, we must have $\mathbb{P}(\forall n\, \varphi(n)) = \lim_{n\to\infty} \mathbb{P}(\varphi(n)) \geq p$. Similarly, for (ii), if $\mathbb{P}(\forall n\, \varphi(n)) = \lim_{n\to\infty} \mathbb{P}(\varphi(n)) \leq p$ then in particular, for any $\lambda > 0$ we have $\mathbb{P}(\varphi(n)) < p + \lambda$ for some $n \in \mathbb{N}$, and conversely if for any λ we have $\mathbb{P}(\varphi(n)) < p + \lambda$ for some $n \in \mathbb{N}$, since $\{\mathbb{P}(\varphi(n))\}$ is decreasing we have $\mathbb{P}(\forall n\, \varphi(n)) = \lim_{n\to\infty} \mathbb{P}(\varphi(n)) < p + \lambda$ for all $\lambda > 0$, and thus $\mathbb{P}(\forall n\, \varphi(n)) \leq p$. Parts (iii) and (iv) follow by negating both sides of the implications and applying (i) and (ii) to the complement of $\varphi(n)$. □

An important feature of the deterministic notions we study that allows us to obtain quantitative stochastic analogues is that they satisfy a monotonicity property:

Definition 4.2.3. A measurable formula $A(\omega, n, m)$ with $n, m \in \mathbb{N}$ is said to be *monotone* if $n \leq n'$ and $m' \leq m$ implies that $A(n', m') \subseteq A(n, m)$.

We are almost ready to prove a key general theorem to obtain quantitative stochastic analogues of important deterministic notions. We first need a lemma from [5].

Lemma 4.2.4. *For a sequence of events $\{B_n\}$ and any $\lambda > \lambda' > 0$, if for any $F : \mathbb{N} \to \mathbb{N}$ there exists an N such that*

$$\mathbb{P}\left(\forall n \leq N\, \exists k \in [n; F(n)]\, B_k\right) < \lambda'$$

then $\mathbb{P}(B_n) < \lambda$ for some n.

We delay the proof of this result to Section 4.2.3, where we present a formal proof of this result in $\mathcal{F}^{\omega}[\mathbb{P}]$ with our Theorem 4.1.9 allowing us to justify the complexity and uniformity of the bounds extracted in [5] for this result. We now have the following:

Theorem 4.2.5. *For a measurable monotone formula $A(\omega, n, m)$ with $n, m \in \mathbb{N}$, the following statements are equivalent:*

(a) Almost surely, there exists n such that $A(n, m)$ does not hold for any $m \geq n$, that is:

$$\mathbb{P}(\exists n\, \forall m\, A(n, n+m)^c) = 1.$$

(b) For any $\lambda > 0$ there exists $n \in \mathbb{N}$ such that for all $m \in \mathbb{N}$

$$\mathbb{P}(A(n, n+m)) < \lambda.$$

(c) For any $\lambda > 0$ and $g : \mathbb{N} \to \mathbb{N}$ there exists n such that

$$\mathbb{P}(A(n, n+g(n))) < \lambda.$$

(d) For any $\lambda > 0$ and $g : \mathbb{N} \to \mathbb{N}$ there exists N such that

$$\mathbb{P}(\forall n \leq N\, A(n, n+g(n))) < \lambda.$$

Proof. The equivalence of (a) and (b) follows through repeated applications of Lemma 4.2.2. Specifically, using the monotonicity property of A and Lemma 4.2.2 (iv) applied to $\varphi(n) := \forall m\, A(n, n+m)^c$ (and replacing $> 1 - \lambda$ with $\geq 1 - \lambda$), is equivalent to

$$\forall \lambda > 0\, \exists n\, \mathbb{P}(\forall m\, A(n, n+m)^c) \geq 1 - \lambda.$$

Now using (i) applied to $\varphi(m) := A(n, n+m)^c$, this is equivalent to

$$\forall \lambda > 0\, \exists n\, \forall m\, \mathbb{P}(A(n, n+m)^c) \geq 1 - \lambda$$

which is equivalent to

$$\forall \lambda > 0\, \exists n\, \forall m\, \mathbb{P}(A(n, m)) < \lambda.$$

Now, to demonstrate the equivalence between (b) and (c), (b) being false is equivalent to the existence of $\lambda > 0$ and $g : \mathbb{N} \to \mathbb{N}$ such that

$$\forall n \, \mathbb{P}(A(n, n + g(n))) \geq \lambda$$

which is then the negation of (c). That (c) implies (d) is clear. To prove (c) from (d), we apply Lemma 4.2.4. Fixing λ and g and defining

$$B_n^g := A(n, n + g(n))$$

it suffices to show that for some $0 < \lambda' < \lambda$, for all $F : \mathbb{N} \to \mathbb{N}$ there exists N such that

$$\mathbb{P}\left(\forall n \leq N \, \exists k \in [n; F(n)] \, B_k^g\right) < \lambda'.$$

But using monotonicity of A by which we have

$$\begin{aligned} \exists k \in [n; F(n)] \, B_k^g &= \exists k \in [n; F(n)] \, A(k, k + g(k)) \\ &\subseteq \exists k \in [n; F(n)] \, A(n, k + g(k)) \\ &\subseteq A(n, n + F^g(n)) \end{aligned}$$

for $F^g(n) := \max\{k - n + g(k) \mid k \in [n; F(n)]\}$, it suffices to show that

$$\mathbb{P}\left(\forall n \leq N \, A(n, n + F^g(n))\right) < \lambda'$$

for any F, and the existence of such an N then follows from (c). □

We now arrive at the following definitions, each giving a general quantitative meaning to the measurable formula $\exists n \, \forall m \geq n \, A(\omega, n, m)^c$ occurring almost surely.

Definition 4.2.6. Let $B := \exists n \, \forall m \geq n \, A(\omega, n, m)^c$ be a measurable formula. Then:

(a) A *(direct) rate* for B is any function $f : (0, 1] \to \mathbb{N}$ satisfying

$$\mathbb{P}(\exists m \geq f(\lambda) \, A(f(\lambda), m)) < \lambda$$

for all $\lambda \in (0, 1]$.

(b) A *uniform metastable rate* for B is any functional $\Phi : (0, 1] \times (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$ satisfying

$$\exists n \leq \Phi(\lambda, g) \, \mathbb{P}(A(n, n + g(n))) < \lambda$$

for all $\lambda \in (0, 1]$ and $g : \mathbb{N} \to \mathbb{N}$.

(c) A *pointwise metastable rate* for B is any function $\Phi : (0,1] \times (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$ satisfying

$$\mathbb{P}(\forall n \leq \Phi(\lambda, g)\, A(n, n + g(n))) < \lambda$$

for all $\lambda \in (0,1]$ and $g : \mathbb{N} \to \mathbb{N}$.

The above allows us to naturally introduce computational interpretations of properties of stochastic processes.

Let $\{X_n\}$ be a stochastic process. By Lemma 4.2.2 (iv), the property that $\{X_n\}$ is almost surely uniformly bounded i.e.

$$\sup_{n\in\mathbb{N}} |X_n| < \infty \quad \text{almost surely}$$

is equivalent to the statement that for any $\lambda > 0$ there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq N\right) < \lambda.$$

Uniform boundedness is related to the notion of *tightness*. In particular, it implies that the sequence $\{X_n\}$ is tight in the sense that for any $\lambda > 0$ there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}\left(|X_n| \geq N\right) < \lambda \quad \text{for all } n \in \mathbb{N}.$$

Tightness is strictly weaker than almost sure uniform boundedness. In particular, whenever

$$\sup_{n\in\mathbb{N}} \mathbb{E}(|X_n|) < \infty$$

then $\{X_n\}$ is tight by Markov's inequality but is not necessarily almost surely bounded:

Example 4.2.7. Define $\{X_n\}$ by

$$\begin{aligned} &X_0 = 1 \\ &X_1 = 2I_{[0,1/2]}, X_2 = 2I_{[1/2,1]} \\ &X_3 = 3I_{[0,1/3]}, X_4 = 3I_{[1/3,2/3]}, X_5 = 3I_{[2/3,1]} \\ &\dots \end{aligned}$$

Then $\mathbb{E}(|X_n|) = \mathbb{E}(X_n) = 1$ for all $n \in \mathbb{N}$, and thus $\{X_n\}$ is tight. On the other hand, for any $N \in \mathbb{N}$, we have

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} X_n \geq N\right) = 1$$

and so $\{X_n\}$ is almost surely *unbounded*.

Definition 4.2.8. Let $\{X_n\}$ be a stochastic process:

(a) Any function $\phi : (0,1] \to \mathbb{R}$ satisfying

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq \phi(\lambda)\right) < \lambda \quad \text{for all } \lambda \in (0,1]$$

is called a *modulus of uniform boundedness* for $\{X_n\}$.

(b) Any function $\phi : (0,1] \to \mathbb{R}$ satisfying

$$\mathbb{P}\left(|X_n| \geq \phi(\lambda)\right) < \lambda \quad \text{for all } \lambda \in (0,1] \text{ and } n \in \mathbb{N}$$

is called a *modulus of tightness* for $\{X_n\}$.

In particular, any modulus of uniform boundedness is also a modulus of tightness for the same stochastic process.

Remark 4.2.9. Defining the event $\sup_{n\in\mathbb{N}} |X_n| \geq \phi(\lambda)$ requires the use of infinite unions. However, the event is equivalent to $\forall m\, \exists n \leq m\, (|X_n| \geq \phi(\lambda))$ and so Lemma 4.2.2 implies that for all $\lambda \in (0,1]$, ϕ satisfying

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq \phi(\lambda)\right) < \lambda$$

is equivalent to ϕ satisfying

$$\forall m\, (\mathbb{P}\left(\exists n \leq m\, |X_n| \geq \phi(\lambda)\right) < \lambda)$$

and the latter only makes use of finite unions and is thus formalisable in $\mathcal{F}^{\omega}[\mathbb{P}]$. We, however, opt for the former (and as we continue the more *infinitary* versions of certain notions in this regard) to ease our ability to work with them in the informal context they are presented in.

By a simple application of Markov's inequality, we obtain the following:

Lemma 4.2.10. *Suppose that*

$$\sup_{n\in\mathbb{N}} \mathbb{E}(|X_n|) < M$$

for some $M > 0$. Then $\{X_n\}$ is tight with modulus $\phi(\lambda) := M/\lambda$.

Example 4.2.11. If $\{X_n\}$ is a nonnegative supermartingale with $\sup_{n\in\mathbb{N}} \mathbb{E}(|X_n|) < M$, then it is both tight and almost surely uniformly bounded, with a modulus $\phi(\lambda) = M/\lambda$ in both cases. The latter follows from Ville's inequality (Theorem 2.4.25), whereby

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} X_n \geq N\right) < \frac{M}{N}.$$

The property that, almost surely, the stochastic process $\{X_n\}$ converges is equivalent to

$$\mathbb{P}(\forall k \in \mathbb{N}\, \exists n\, \forall m\, \forall i, j \in [n; m](|X_i - X_j| < 2^{-k})) = 1$$

and Lemma 4.2.2 implies the above is equivalent to

$$\forall k \in \mathbb{N}\, \mathbb{P}(\exists n\, \forall m\, \forall i, j \in [n; m](|X_i - X_j| < 2^{-k})) = 1.$$

This has exactly the form of Theorem 4.2.5 (a) for

$$A(\omega, n, m) := \exists i, j \in [n; m]\, (|X_i(\omega) - X_j(\omega)| \geq 2^{-k})$$

where $A(\omega, n, m)$ will be a monotone measurable formula and also $A(n, m) = \emptyset$ for $m \leq n$. Therefore, Theorem 4.2.5 applies in this case, and we arrive at the following definitions:

Definition 4.2.12. Let $\{X_n\}$ be a stochastic process:

(a) Any function $\phi : (0, 1] \times (0, 1] \to \mathbb{R}$ satisfying

$$\mathbb{P}(\exists i, j \geq \phi(\lambda, \varepsilon)\, (|X_i - X_j| \geq \varepsilon)) < \lambda$$

for all $\lambda, \varepsilon \in (0, 1]$ is called a *rate of almost sure convergence* for $\{X_n\}$.

(b) Any functional $\Phi : (0, 1] \times (0, 1] \times (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$ such that for all $\lambda, \varepsilon \in (0, 1]$ and $g : \mathbb{N} \to \mathbb{N}$ there exists $n \leq \Phi(\lambda, \varepsilon, g)$ satisfying

$$\mathbb{P}(\exists i, j \in [n; n + g(n)](|X_i - X_j| \geq \varepsilon)) < \lambda$$

is called a *metastable rate of uniform convergence* for $\{X_n\}$.

(c) Any functional $\Phi : (0, 1] \times (0, 1] \times (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$ such that for all $\lambda, \varepsilon \in (0, 1]$ and $g : \mathbb{N} \to \mathbb{N}$

$$\mathbb{P}(\forall n \leq \Phi(\lambda, \varepsilon, g)\, \exists i, j \in [n; n + g(n)](|X_i - X_j| \geq \varepsilon)) < \lambda$$

is called a *metastable rate of pointwise convergence* for $\{X_n\}$.

As in the deterministic case, each of these definitions has their corresponding analogues regarding convergence to a fixed random variable.

Remark 4.2.13. The property that

$$\forall \varepsilon, \lambda > 0 \exists n\, \mathbb{P}(\exists i, j \geq n\, (|X_i - X_j| \geq \varepsilon))$$

is known as almost uniform convergence and the equivalence to almost sure convergence (a result we demonstrate in Theorem 4.2.5) is attributed to Egorov. Thus, our computational interpretation of almost sure convergence is actually one for almost uniform convergence.

The previously introduced definitions are not new. The first is a rate of convergence for

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{i,j\geq n}(|X_i - X_j| \geq \varepsilon)\right) = 0$$

and is used throughout the probability theory literature ([70, 110, 133], for example).

The two metastable notions originate from [5]. Specifically, a uniform metastable rate gives bounds on the λ-*uniform* ε-*metastable convergence of* $\{X_n\}$ for all $\lambda, \varepsilon > 0$ and a pointwise rate generates a λ-*uniform bound for the* ε-*metastable pointwise convergence of* $\{X_n\}$ for all $\lambda, \varepsilon > 0$.

For a stochastic process $\{X_n\}$ define

$$J_{N,\varepsilon}\{X_n\}(\omega) := J_{N,\varepsilon}\{X_n(\omega)\}^3$$

for each $\omega \in \Omega$, with $J_{N,\varepsilon}\{x_n\}$ defined as in Section 2.3.1. In other words, $J_{N,\varepsilon}\{X_n\}$ denotes the number of ε-fluctuations that occur in the initial segment $\{X_0, \ldots, X_{N-1}\}$ of the process. The stochastic analogue of total fluctuations $J_\varepsilon\{X_n\}$, along with those for $[\alpha, \beta]$-crossings $C_{N,[\alpha,\beta]}\{X_n\}$ and $C_{[\alpha,\beta]}\{X_n\}$, are defined in the same way.

The property that, almost surely, $\{X_n\}$ has finite ε-fluctuations for each $\varepsilon > 0$, is equivalent to

$$\mathbb{P}\left(\forall k\, J_{2^{-k}}\{X_n\} < \infty\right) = 1.$$

From Lemma 4.2.2 and monotonicity of $J_{2^{-k}}\{X_n\} < \infty$ in k the above equivalent to

$$\forall k\ \left(\mathbb{P}\left(J_{2^{-k}}\{X_n\} < \infty\right) = 1\right)$$

i.e. for any $\varepsilon > 0$, $\{X_n\}$ has finite ε-fluctuations almost surely. This leads to the following quantitative notion:

Definition 4.2.14. Let $\{X_n\}$ be a stochastic process. For fixed $\varepsilon > 0$, any function $\phi : (0, 1] \to \mathbb{R}$ satisfying

$$\mathbb{P}\left(J_\varepsilon\{X_n\} \geq \phi(\lambda)\right) < \lambda \quad \text{for all } \lambda \in (0, 1]$$

is called a *modulus of finite* ε-*fluctuations* for $\{X_n\}$. Any function $\phi : (0, 1] \times (0, 1] \to \mathbb{R}$ such that $\phi(\cdot, \varepsilon)$ is a modulus of finite ε-fluctuations for all $\varepsilon \in (0, 1]$ is simply called a *modulus of finite fluctuations* for $\{X_n\}$.

[3] One can easily show that this defines a random variable.

A modulus of finite fluctuations is just another way of formulating the rate of convergence of

$$\lim_{N\to\infty} \mathbb{P}(J_\varepsilon\{X_n\} \geq N) = 0$$

a quantitative notion that has been widely explored, particularly in the context of martingales (e.g. [70]).

To express the idea that a stochastic process, almost surely, has a finite number of $[\alpha, \beta]$ crossings for all intervals $[\alpha, \beta]$, we need to find a method to encode quantification over all such intervals in a monotonic manner. The encoding we choose, which informs our definition of the corresponding modulus, reflects how crossings are utilised in the convergence proofs we analyse. Specifically, having a finite number of crossings over arbitrary intervals, almost surely, is equivalent to the statement

$$\mathbb{P}\left(\forall k, M\, \forall [\alpha, \beta] \in \mathcal{P}(M, 2^{-k})\, C_{[\alpha,\beta]}\{X_n\} < \infty\right) = 1$$

where $\mathcal{P}(M, l)$ is as in Definition 2.3.19. Now applying Lemma 4.2.2, noting that the inner formula is monotone decreasing in both k and M, this is equivalent to

$$\forall k, M \left(\mathbb{P}\left(\forall [\alpha, \beta] \in \mathcal{P}(M, 2^{-k})\, C_{[\alpha,\beta]}\{X_n\} < \infty\right) = 1\right). \tag{4.1}$$

So, for any $\alpha < \beta$, picking k, M such that there exists $[\alpha', \beta'] \in \mathcal{P}(M, 2^{-k})$ with $[\alpha', \beta'] \subseteq [\alpha, \beta]$ establishes that $\mathbb{P}(C_{[\alpha,\beta]}\{X_n\} < \infty) = 1$. Thus, we have (4.1) is equivalent to

$$\forall \alpha, \beta > 0\, (\mathbb{P}(C_{[\alpha,\beta]}\{X_n\} < \infty) = 1).$$

The above discussion yields the following quantitative definitions:

Definition 4.2.15. Let $\{X_n\}$ be a stochastic process:

(a) For fixed $\alpha < \beta$, any function $\phi : (0, 1] \to \mathbb{R}$ satisfying

$$\mathbb{P}\left(C_{[\alpha,\beta]}\{X_n\} \geq \phi(\lambda)\right) < \lambda \quad \text{for all } \lambda \in (0, 1]$$

is called a *modulus of finite* $[\alpha, \beta]$*-crossings* for $\{X_n\}$.

(b) Any function $\phi : (0, 1] \times (0, \infty) \times \mathbb{N} \to \mathbb{R}$ satisfying

$$\mathbb{P}\left(\exists [\alpha, \beta] \in \mathcal{P}(M, l)\, C_{[\alpha,\beta]}\{X_n\} \geq \phi(\lambda, M, l)\right) < \lambda$$

for all $\lambda \in (0, 1]$, $M \in (0, \infty)$ and $l \in \mathbb{N}$ is called a *modulus of finite crossings* for $\{X_n\}$.

Lemma 4.2.16. *(i) If ϕ is a modulus of finite crossings and $\alpha < \beta$, then $\psi(\lambda) := \phi(\lambda, M, l)$ is a modulus of finite $[\alpha, \beta]$-crossings, for M, l such that there exists $[\alpha', \beta'] \subseteq [\alpha, \beta]$ with $[\alpha', \beta'] \in \mathcal{P}(M, l)$.*

(ii) If $\phi_{\alpha,\beta}$ is a modulus of finite $[\alpha, \beta]$-crossings for all $\alpha < \beta$, then

$$\psi(\lambda, M, l) := \max \left\{ \phi_{\alpha,\beta}\left(\frac{\lambda}{l}\right) \;\middle|\; [\alpha, \beta] \in \mathcal{P}(M, l) \right\}$$

is a modulus of finite crossings.

Proof. The first part is clear, for (ii) we observe that

$$\begin{aligned}
&\mathbb{P}\left(\exists [\alpha, \beta] \in \mathcal{P}(M, l)\, C_{[\alpha,\beta]}\{X_n\} \geq \psi(\lambda, M, l)\right) \\
&\leq \sum_{[\alpha,\beta]\in\mathcal{P}(M,l)} \mathbb{P}\left(C_{[\alpha,\beta]}\{X_n\} \geq \psi(\lambda, M, l)\right) \\
&\leq \sum_{[\alpha,\beta]\in\mathcal{P}(M,l)} \mathbb{P}\left(C_{[\alpha,\beta]}\{X_n\} \geq \phi_{\alpha,\beta}\left(\lambda/l\right)\right) \\
&< \sum_{[\alpha,\beta]\in\mathcal{P}(M,l)} \frac{\lambda}{l} = \lambda
\end{aligned}$$

where for the last step we recall that $\mathcal{P}(M, l)$ consists of l intervals by definition. □

Remark 4.2.17. Just as in Lemma 4.2.10, Markov's inequality gives us concrete moduli for the above when the expectation is bounded. For example, if $\tau : \mathbb{R} \times \mathbb{R} \to (0, \infty)$ is a function satisfying

$$\mathbb{E}(C_{[\alpha,\beta]}\{X_n\}) < \tau(\alpha, \beta) \tag{4.2}$$

for all $\alpha < \beta$, then $\phi_{[\alpha,\beta]}(\lambda) := \tau(\alpha, \beta)/\lambda$ is a modulus of finite $[\alpha, \beta]$-crossings for $\{X_n\}$, and similarly for ε-fluctuations bounded in mean.

Inequalities of the form (4.2) are known as upcrossing inequalities and are integral tools in establishing the convergence of stochastic process in Martingale theory [34]. Such inequalities offer further computational content than just moduli of finite crossings, which we exploit in Chapter 7. To this effect, we introduce the following definition.

Definition 4.2.18. Any function $\psi : (0, \infty) \times \mathbb{N} \to \mathbb{R}$ satisfying

$$\mathbb{E}\left[C_{[\alpha,\beta]}\{X_n\}\right] < \psi(M, l)$$

for all M, l and $[\alpha, \beta] \in \mathcal{P}(M, l)$ is called a *modulus of L_1-crossing* for $\{X_n\}$.

4.2.2 Learnable rates

Recall in the case of deterministic convergence, $\Phi(\varepsilon, g) := \tilde{g}^{(b(\varepsilon))}(0)$ is a rate of metastability for a sequence of real numbers $\{x_n\}$ iff b is a bound on the fluctuations for $\{x_n\}$ (see Section 2.3.2). It turns out that the metastable rates for almost sure statements in Definition 4.2.6 correspond to interesting measures of fluctuations for random variables.

Definition 4.2.19. Let $B := \exists n \forall m \geq n\, A(\omega, n, m)^c$ be a measurable formula. A function $\phi : (0, 1] \to \mathbb{N}$ is:

(a) a *uniform learnable rate* for B if

$$\exists n \leq \phi(\lambda)\, \mathbb{P}(A(a_n, b_n)) < \lambda$$

for any $\lambda \in (0, 1]$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$.

(b) a *pointwise learnable rate* for B if

$$\mathbb{P}(\forall n \leq \phi(\lambda)\, A(a_n, b_n)) < \lambda$$

for any $\lambda \in (0, 1]$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$.

Observe that a uniform learnable rate is a pointwise learnable rate, and we shall see in Example 4.2.24 that there are cases where this implication is strict. We shall now see that (uniform) pointwise learnable rates correspond directly to (uniform) pointwise metastable rates, analogous to the deterministic case with the correspondence between bounds on the fluctuations and rates of metastability (c.f. Theorem 2.3.16).

Lemma 4.2.20. *Let $B := \exists n \forall m \geq n\, A(\omega, n, m)^c$ and $\phi : (0, 1] \to \mathbb{N}$ be some function. Then $\Phi(\lambda, g) = \tilde{g}^{(\phi(\lambda))}(0)$ for $\tilde{g}(n) := n + g(n)$ is a (uniform) pointwise metastable rate for B iff $\phi(\lambda)$ a (uniform) pointwise learnable rate for B.*

Proof. For the uniform case, in one direction we define $a_n := \tilde{g}^{(n)}(0)$ and $b_n := \tilde{g}^{(n+1)}(0)$. Then if $\tilde{g}^{(\phi(\lambda))}(0)$ is not a uniform metastable rate then

$$\forall n \leq \phi(\lambda)\, \mathbb{P}(A(a_n, b_n)) \geq \lambda$$

so $\phi(\lambda)$ is not a uniform learnable rate. In the other direction, we define g in terms of $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ as in Lemma 2.3.15, and if $\phi(\lambda)$ is not a uniform learnable rate then since for any $n \leq \tilde{g}^{(\phi(\lambda))}(0) = b_{\phi(\lambda)-1} \leq a_{\phi(\lambda)}$ we have $n \leq a_m$ and $b_m = n + g(n)$ for some $m \leq \phi(\lambda)$, and since $A(a_m, b_m) \subseteq A(n, n + g(n))$ it follows that $\mathbb{P}(A(n, n + g(n))) \geq \lambda$ for all $n \leq \tilde{g}^{(\phi(\lambda))}(0)$.

The pointwise case is entirely analogous, with additional details needed to run the argument pointwise. For the first direction, defining a_n, b_n in the same way, we note that if $\omega \in A(n, n +$

$g(n))$ for all $n \leq \tilde{g}^{(\phi(\lambda))}(0)$, then in particular $\omega \in A(a_n, b_n)$ for $n \leq \phi(\lambda)$, and so if $g^{(\phi(\lambda))}(0)$ is not a pointwise metastable rate then

$$\mathbb{P}(\forall n \leq \phi(\lambda)\, A(a_n, b_n)) \geq \mathbb{P}(\forall n \leq \tilde{g}^{(\phi(\lambda))}(0)\, A(n, n + g(n))) \geq \lambda$$

where we must also note that for any $n \leq \phi(\lambda)$,

$$\mathbb{P}(A(a_n, b_n)) \geq \mathbb{P}(\forall n \leq \phi(\lambda)\, A(a_n, b_n)) \geq \lambda$$

and thus $a_n < b_n$. In the other direction, we note that if $\omega \in A(a_n, b_n)$ for all $n \leq \phi(\lambda)$, defining g in the same way as the uniform case, for any $n \leq \tilde{g}^{(\phi(\lambda))}(0)$ there exists $m \leq \phi(\lambda)$ such that $\omega \in A(a_m, b_m) \subseteq A(n, n + g(n))$, and so $\omega \in A(n, n + g(n))$ for all $n \leq \tilde{g}^{(\phi(\lambda))}(0)$, and thus if $\phi(\lambda)$ is not a pointwise learnable rate then

$$\mathbb{P}(\forall n \leq \tilde{g}^{(\phi(\lambda))}(0)\, A(n, n + g(n))) \geq \mathbb{P}(\forall n \leq \phi(\lambda)\, A(a_n, b_n)) \geq \lambda$$

from which we obtain our contradiction. □

We, therefore, have the following concrete definitions.

Definition 4.2.21. Let $\{X_n\}$ be a stochastic process:

(a) Any function $\phi : (0, 1] \times (0, 1] \to \mathbb{R}$ satisfying

$$\exists n \leq \phi(\lambda, \varepsilon)\, \mathbb{P}\left(\exists i, j \in [a_n; b_n]\, (|X_i - X_j| \geq \varepsilon)\right) < \lambda$$

for any $\varepsilon, \lambda \in (0, 1]$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ is called a *learnable rate of uniform convergence*.

(b) Any function $\phi : (0, 1] \times (0, 1] \to \mathbb{R}$ satisfying

$$\mathbb{P}\left(\forall n \leq \phi(\lambda, \varepsilon)\, \exists i, j \in [a_n; b_n]\, (|X_i - X_j| \geq \varepsilon)\right) < \lambda$$

for any $\varepsilon, \lambda \in (0, 1]$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ is called a *learnable rate of pointwise convergence*.

Remark 4.2.22. By Lemma 4.2.20, a learnable rate of uniform convergence $\phi(\lambda, \varepsilon)$ corresponds to the metastable rate of uniform convergence $\Phi(\lambda, \varepsilon, g) = \tilde{g}^{(\lceil \phi(\lambda, \varepsilon) \rceil)}(0)$ and similarly for pointwise convergence.

Remark 4.2.23. One can easily show that a modulus of finite fluctuations is a learnable rate of pointwise convergence since for all $\varepsilon \in (0, 1]$, $N > 0$ and $a_0 < b_1 \leq a_1 < b_1 \leq \ldots$,

$$\{\omega : \forall n \leq N\, \exists i, j \in [a_n; b_n]\, (|X_i(\omega) - X_j(\omega)| \geq \varepsilon)\} \subseteq \{\omega : J_\varepsilon\{X_n\}(\omega) \geq N\}.$$

However, it is currently unclear whether the converse holds. A pointwise learnable rate is not always a uniform learnable rate:

Example 4.2.24. Let $\mathcal{C}$ be the class of nonnegative stochastic processes $\{X_n\}$ that are monotone and uniformly bounded above by 1. These can experience at most $1/\varepsilon$ ε-fluctuations, and therefore, a modulus of finite fluctuations and hence learnable rate of pointwise convergence for any such process is given by

$$\phi(\lambda, \varepsilon) = \frac{1}{\varepsilon} + 1.$$

Suppose now that $\psi(\lambda, \varepsilon)$ is a learnable rate of uniform convergence that applies uniformly in $\mathcal{C}$, i.e. $\psi(\lambda, \varepsilon)$ is a learnable rate of uniform convergence for any $\{X_n\} \in \mathcal{C}$. Then we claim that

$$\frac{1}{\lambda\varepsilon} \leq \psi(\lambda, \varepsilon) \tag{4.3}$$

for all $\varepsilon, \lambda \in (0, 1]$. Suppose for contradiction that there exist $\varepsilon, \lambda \in (0, 1]$ on which (4.3) fails, where for simplicity we assume that $\varepsilon = 1/M$ and $\lambda = 1/N$ for some $M, N \in \mathbb{N}$. We define a stochastic process $\{X_n\}$ on the standard space $([0, 1], \mathcal{F}, \mu)$ and in terms of these parameters as follows: First, define the sequence of reals $\{x_n\}$ by

$$x_n := \begin{cases} 0 & \text{if } n = 0 \\ i/M & \text{if } (i-1)N < n \leq iN \text{ for } i = 1, \ldots, M \\ 1 & \text{if } n > MN \end{cases}$$

so that we have $x_{j+1} - x_j = 1/M$ for $j = iN$ and $i = 0, \ldots, M-1$, and $x_{j+1} - x_j = 0$ for all other $j \in \mathbb{N}$. Now letting $I_0, \ldots, I_{N-1}$ represent a division of $[0, 1]$ into N equal partitions, we define

$$X_n(\omega) := \begin{cases} 0 & \text{if } n < k \\ x_{n-k} & \text{otherwise} \end{cases} \quad \text{for } \omega \in I_k.$$

Then analogously to the situation with $\{x_n\}$, for $k = 0, \ldots, N-1$ and $\omega \in I_k$ we have $X_{j+1}(\omega) - X_j(\omega) = 1/M$ for $j = iN + k$ and $i = 0, \ldots, M-1$, and $X_{j+1}(\omega) - X_j(\omega) = 0$ for all other $j \in \mathbb{N}$. This means that for all $j \leq MN - 1$, there is exactly one $k = 0, \ldots, N-1$ for which $X_{j+1} - X_j = 1/M$ on I_k, and therefore

$$\mathbb{P}(|X_j - X_{j+1}| \geq \varepsilon) = \lambda$$

for all $\forall j \leq (1/\lambda\varepsilon) - 1$. But since (4.3) fails, we must have

$$\exists j < (1/\lambda\varepsilon)\, \mathbb{P}(|X_j - X_{j+1}| \geq \varepsilon) < \lambda$$

contradicting that this is a learnable rate of uniform convergence for $\{X_n\}$, and thus a rate that applies to all sequences in $\mathcal{C}$. This example also demonstrates that it is possible that a modulus of finite fluctuations is not a uniform learnable rate of convergence. Furthermore, in this example, it is the case that a uniform learnable rate is a modulus of finite fluctuations, but it is unclear whether this holds in general.

4.2.3 A proof-theoretic analysis of the relationship between pointwise and uniform metastability

We already know that metastable uniform and pointwise convergence are equivalent by Theorem 4.2.5. In this section, we investigate their quantitative relationship from a proof-theoretic perspective. It is clear that a rate of metastable uniform convergence is a rate of metastable pointwise convergence. However, obtaining a rate of metastable uniform convergence from a rate of metastable pointwise convergence appears not to be so straightforward. The central result of [5] was the following:

Theorem 4.2.25 (Avigad, Dean & Rute, Theorem 3.1 of [5])**.** *For every $\varepsilon > 0$, $\lambda > \lambda' > 0$ and functional $M_1 : (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$, there is a functional $\Gamma(\varepsilon, \lambda, \lambda', M_1) : (\mathbb{N} \to \mathbb{N}) \to \mathbb{N}$ such that whenever*

$$\mathbb{P}(\forall n \leq M_1(f_1)\, \exists i, j \in [n; f_1(n)](|X_i - X_j| \geq \varepsilon)) < \lambda' \tag{4.4}$$

for all $f_1 : \mathbb{N} \to \mathbb{N}$, then for any $f_2 : \mathbb{N} \to \mathbb{N}$ there exists some $n \leq \Gamma(\varepsilon, \lambda, \lambda', M_1)(f_2)$ such that

$$\mathbb{P}(\exists i, j \in [n; f_2(n)](|X_i - X_j| \geq \varepsilon)) < \lambda. \tag{4.5}$$

Theorem 4.2.25 immediately provides passage from metastable pointwise to uniform rates:

Corollary 4.2.26. *Suppose that Φ is a metastable rate of pointwise convergence for $\{X_n\}$ and let Γ be the construction from Theorem 4.2.25. Then, a metastable rate of uniform convergence is given by*

$$\Psi(\lambda, \varepsilon, g) := \Gamma\left(\varepsilon, \lambda, \tfrac{\lambda}{2}, M_1^{\lambda,\varepsilon}\right)(\tilde{g})$$

where $\tilde{g}(n) := n + g(n)$ and

$$M_1^{\lambda,\varepsilon}(f_1) := \Phi\left(\tfrac{\lambda}{2}, \varepsilon, \bar{f}_1\right)$$

for $\bar{f}_1(n) := f_1(n) - n$ if $f_1(n) \geq n$ and $\bar{f}_1(n) := 0$ otherwise. With Γ from Theorem 4.2.25.

Proof. Fixing $\lambda, \varepsilon > 0$, the functional $M_1^{\lambda,\varepsilon}$ satisfies (4.4) for $\lambda' := \lambda/2$, noting that $n + \bar{f}_1(n) = f_1(n)$ unless $f_1(n) < n$, in which case $\exists i, j \in [n; f_1(n)](|X_i - X_j| \geq \varepsilon)$ and $\exists i, j \in [n; n + \bar{f}_1(n)](|X_i - X_j| \geq \varepsilon)$ are both empty. Thus by Theorem 4.2.25, for any $g : \mathbb{N} \to \mathbb{N}$ there exists some $n \leq \Psi(\lambda, \varepsilon, g)$ satisfying (4.2.26) for $f_2 := \tilde{g}$, and since $\lambda, \varepsilon > 0$ are arbitrary, we are done. □

The precise construction of Γ in [5] uses bar recursion (as presented in Section 2.2.1). The complexity, as well as the observed uniformities of their bounds, can be justified because Theorem 4.2.25 can be formalised in $\mathcal{F}^\omega[\mathbb{P}]$.

Most of the heavy lifting of the proof of Theorem 4.2.25 is Lemma 4.2.4, from which a careful analysis of its proof reveals a lot.

We first need the following lemma:

Lemma 4.2.27. *The system $\mathcal{F}^\omega[\mathbb{P}]$ proves:*

$$\forall A^{S(0)}, k^0 \exists N^0 \forall n^0 \left(\mathbb{P}\left(\bigcup_{i=0}^{n} A(i) \cap \left(\bigcup_{i=0}^{N} A(i) \right)^c \right) <_{\mathbb{R}} 2^{-k} \right).$$

Proof. We reason in $\mathcal{F}^\omega[\mathbb{P}]$. Let $A^{S(0)}$ and k^0 be given. At first, note that Proposition 4.1.7 implies that

$$\exists N \forall n \left(n \geq N \to \left| \sum_{i=0}^{n} \mathbb{P}((A\uparrow)(i)) - \sum_{i=0}^{N} \mathbb{P}((A\uparrow)(i)) \right| < 2^{-k} \right). \tag{$*$}$$

Take such an N and let n be arbitrary. If $n < N$, then

$$\bigcup_{i=0}^{n} A(i) \cap \left(\bigcup_{i=0}^{N} A(i) \right)^c = \emptyset$$

and so by extensionality of $\mathbb{P}$, we get

$$\mathbb{P}\left(\bigcup_{i=0}^{n} A(i) \cap \left(\bigcup_{i=0}^{N} A(i) \right)^c \right) = 0$$

and are done. So suppose $n \geq N$. Then by $(*)$, we get

$$\left| \sum_{i=0}^{n} \mathbb{P}((A\uparrow)(i)) - \sum_{i=0}^{N} \mathbb{P}((A\uparrow)(i)) \right| < 2^{-k}$$

Since all the $(A\uparrow)(i)$ are disjoint (by definition of $A\uparrow$) and since we have

$$\bigcup_{i=0}^{j} (A\uparrow)(i) = \bigcup_{i=0}^{j} A(i)$$

for any j, we immediately derive

$$\sum_{i=0}^{j} \mathbb{P}((A\uparrow)(i)) = \mathbb{P}\left(\bigcup_{i=0}^{j} A(i) \right)$$

for any j by finite additivity and extensionality of $\mathbb{P}$. Thus, we, in particular, have

$$\left|\mathbb{P}\left(\bigcup_{i=0}^{n} A(i)\right) - \mathbb{P}\left(\bigcup_{i=0}^{N} A(i)\right)\right| < 2^{-k}$$

and since $n \geq N$ implies

$$\bigcup_{i=0}^{N} A(i) \subseteq \bigcup_{i=0}^{n} A(i),$$

we obtain

$$\mathbb{P}\left(\bigcup_{i=0}^{n} A(i) \cap \left(\bigcup_{i=0}^{N} A(i)\right)^c\right) = \mathbb{P}\left(\bigcup_{i=0}^{n} A(i)\right) - \mathbb{P}\left(\bigcup_{i=0}^{N} A(i)\right) < 2^{-k}$$

by Proposition 4.1.5. □

We now demonstrate that Lemma 4.2.4 can be formalised in $\mathcal{F}^{\omega}[\mathbb{P}]$. Observe this Lemma can be formalised as:

Theorem 4.2.28. *The system $\mathcal{F}^{\omega}[\mathbb{P}]$ proves:*

$$\forall A^{S(0)}, M^{0(1)}, u^0, v^0 >_0 u \exists n^0 \left(\forall F^1 \left(\mathbb{P}\left(\bigcap_{m=0}^{M(F)} \bigcup_{j=m}^{F(m)} A(j)\right) \leq_{\mathbb{R}} 2^{-v}\right) \to \mathbb{P}(A(n)) <_{\mathbb{R}} 2^{-u}\right).$$

Proof. Let $A^{S(0)}$, $M^{0(1)}$, u^0 and v^0 with $v > u$ be given and suppose

$$\forall F^1 \left(\mathbb{P}\left(\bigcap_{m=0}^{M(F)} \bigcup_{j=m}^{F(m)} A(j)\right) \leq 2^{-v}\right).$$

So, by the previous Lemma 4.2.27 applied to the sequence of events $f_m^{S(0)}$ defined by $f_m(k) = A(k+m)$, we have

$$\forall m \exists N \forall n \left(\mathbb{P}\left(\bigcup_{i=m}^{n+m} A(i) \cap \left(\bigcup_{i=m}^{N+m} A(i)\right)^c\right) < \frac{2^{-u} - 2^{-v}}{2^{m+1}}\right)$$

and so, in particular

$$\forall m \exists N \geq m \forall n \geq m \left(\mathbb{P}\left(\bigcup_{i=m}^{n} A(i) \cap \left(\bigcup_{i=m}^{N} A(i)\right)^c\right) < \frac{2^{-u} - 2^{-v}}{2^{m+1}}\right).$$

Thus, using AC (which follows from DC) there exists a function F^1 such that for all m and

$n \geq m$:

$$\mathbb{P}\left(\bigcup_{i=m}^{n} A(i) \cap \left(\bigcup_{i=m}^{F(m)} A(i)\right)^c\right) < \frac{2^{-u} - 2^{-v}}{2^{m+1}}.$$

It is now easy to see that for this function F, we have

$$\begin{aligned} A(M(F)) &\subseteq \bigcap_{m=0}^{M(F)} \bigcup_{i=m}^{M(F)} A(i) \\ &\subseteq \left(\bigcap_{m=0}^{M(F)} \bigcup_{j=m}^{F(m)} A(j)\right) \cup \bigcup_{m=0}^{M(F)} \left(\bigcup_{i=m}^{M(F)} A(i) \cap \left(\bigcup_{j=m}^{F(m)} A(j)\right)^c\right) \end{aligned}$$

and so, by the sub-additivity and monotonicity of $\mathbb{P}$, we derive

$$\mathbb{P}(A(M(F))) < 2^{-v} + \sum_{m=0}^{M(F)} \frac{2^{-u} - 2^{-v}}{2^{m+1}} < 2^{-u}$$

and so we can take $n := M(F)$ and the result follows. □

Remark 4.2.29. Theorem 4.1.9 tells us that since the above theorem can be formalised in $\mathcal{F}^\omega[\mathbb{P}]$, we can extract uniform computable bounds. In particular, the existence of a computable bound on the existential quantifier on n can be guaranteed to exist a priori. Furthermore, the bound can be guaranteed to be independent of the content space and the sequence of events, which matches exactly the properties of the bound explicitly calculated in [5]. Furthermore, an analysis of the above proof through Theorem 4.1.9 would result in a bound of bar recursive complexity due to the use of AC.

Chapter 5

Proof-theoretic transfer principles and Kronecker's lemma

In Chapter 6, we present various quantitative results concerning the Strong Law of Large Numbers. Our motivation for investigating such results was that a classical application of the Robbins-Siegmund theorem (which we give a quantitative version of in Chapter 8) is Kolmogorov's Strong Law of Large Numbers (c.f. [129]). A key result in obtaining many results concerning the Laws of Large Numbers is Kronecker's lemma. Thus, if one is to analyse such a result quantitatively, a computational interpretation of Kronecker's lemma is required. So, the author was led to investigate the computational content of Kronecker's lemma.

Kronecker's lemma is a statement about the convergence of sequences of real numbers, and the way it is applied in the context of the Strong Laws of Large Numbers is to lift this result to the probabilistic setting (we make precise what we mean by this in the coming section). Thus, to apply a computational interpretation of Kronecker's lemma to the Strong Laws of Large Numbers, one must lift the deterministic computational content to the stochastic setting. The author noticed that such lifting was not completely trivial and was only possible because the computational content of Kronecker's lemma for sequences of real numbers was incredibly uniform. So, the author generalised their quantitative result for the probabilistic analogue of Kronecker's lemma to other deterministic statements whose computational interpretations shared similar uniformities. This generalisation was first obtained in collaboration with Pischke in [114] for bounded random variables in the context of the formal system introduced in the previous chapter. The requirement for boundedness was then removed by the author in [112].

This chapter presents the author's computational investigation of Kronecker's lemma and the two transfer results for lifting the computational content of deterministic results to their stochastic analogue (one of himself and the other, in collaboration with Pischke). We start in section 5.1, where we introduce Kronecker's lemma and the problem of obtaining a computational interpretation of its stochastic analogue. We then present the transfer result of the

author and Pischke for bounded random variables, thus providing a strategy for obtaining the computational content for the stochastic analogue of Kronecker's lemma, with the additional requirement that the random variables were bounded. We then continue in Section 5.2, where we investigate the computational content of Kronecker's lemma. We solve the Dialectica interpretation of the statement of Kronecker's lemma and use this to obtain a result that converts rates of (metastable) convergence in the premise of the theorem to rates of (metastable) convergence in the conclusion. We then justify our metastable rates through a Specker construction, like that of Example 2.3.3, and investigate the proof-theoretic strength of Kronecker's lemma in light of the reverse mathematics program [135]. Lastly, motivated by our computational interpretation of Kronecker's lemma, we give a transfer result that does not require the random variables to be bounded.

We note that the computational result for the Laws of Large Numbers obtained in Section 6.2 shall rely on the computational interpretation we give to Kronecker's lemma in this chapter.

5.1 Quantitative transfer principles

A common step in establishing results in probability theory concerned with the probabilistic convergence of sequences of random variables is to use deterministic results about the convergence of sequences of real numbers and then, through pointwise arguments, obtain a probabilistic analogue of these deterministic results. An example of such a result is *Kronecker's lemma*, which states:

Theorem 5.1.1 (Kronecker's lemma)**.** *Let* $\{x_n\}$ *be a sequence of real numbers and* $0 < a_0 \leq a_1 \leq \ldots$ *be such that* $a_n \to \infty$*. If* $\sum_{i=0}^{\infty} x_i < \infty$*, then*

$$\frac{1}{a_n} \sum_{i=0}^{n} a_i x_i \to 0$$

as $n \to \infty$*.*

Through pointwise arguments, one can easily establish the probabilistic analogue of the above result, that is:

Theorem 5.1.2 (Probabilistic Kronecker's lemma)**.** *Let* $\{X_n\}$ *be a sequence of real-valued random variables and* $0 < a_0 \leq a_1 \leq \ldots$ *be such that* $a_n \to \infty$*. If* $\sum_{i=0}^{\infty} X_i < \infty$ *almost surely, then*

$$\frac{1}{a_n} \sum_{i=0}^{n} a_i X_i \to 0$$

almost surely.

A natural question is whether lifting computational content from the deterministic theorem is also as easy. It will turn out that this is not the case, and in this section, we shall investigate when this is possible.

5.1.1 A principle for bounded random variables

Here, we present a general condition on the form of the computational content of deterministic convergence results that allows us to obtain quantitative versions of their probabilistic analogue. Furthermore, we demonstrate that $\mathcal{F}^\omega[\mathbb{P}]$ supports such reasoning.

Concretely, to allow for a discussion of general modes of convergence for real numbers and random variables, we consider the following abstract formal setup: throughout this section, fix two Π_3-formulas

$$\tilde{P}(x^{1(0)}) = \forall a^0 \exists b^0 \forall c^0 P_0(a, b, c, x)$$

and

$$\tilde{Q}(x^{1(0)}) = \forall u^0 \exists v^0 \forall w^0 Q_0(u, v, w, x)$$

where P_0 and Q_0 are quantifier-free formulas which only have the indicated variables free. $\tilde{P}$ and $\tilde{Q}$ should be interpreted as abstract representations of modes of convergence for the parameter sequence, x, of real numbers.

For example, we may take

$$P_0(a, b, c, x) := \forall i^0, j^0 \left(b \leq_0 i, j \wedge i, j \leq_0 c \to |x(i) - x(j)| \in [0, 2^{-a}] \right), \tag{$**$}$$

using the intensional intervals (c.f. Definition 4.1.1), we can regard the above as a quantifier-free statement. In that case, P represents the usual Cauchy property for x.

To allow for a discussion of these modes applied to random variables, we extend the system $\mathcal{F}^\omega[\mathbb{P}]$ with four further constants

$$X^{1(\Omega)(0)},\ P^{S(0)(0)(0)},\ Q^{S(0)(0)(0)},\ \tau^{0(0)},$$

together with the axioms

$$\begin{aligned}
\forall a^0, b^0, c^0, z^\Omega (z \in P(a, b, c) \leftrightarrow P_0(a, b, c, \lambda n.X(n)(z))),\\
\forall a^0, b^0, c^0, z^\Omega (z \in Q(a, b, c) \leftrightarrow Q_0(a, b, c, \lambda n.X(n)(z))),\\
\forall n^0, z^\Omega (\tau(n) \leq_0 \tau(n+1) \wedge \tau(n) \geq_{\mathbb{R}} |X(n)(z)|_{\mathbb{R}}),
\end{aligned}$$

specifying that the properties P_0 and Q_0 induce measurable sets pointwisely relative to the sequence of random variables specified by X. Furthermore, we assume that these random variables are all bounded via a suitable monotone sequence of bounds (i.e. that X as a constant

is majorized by τ^1). Since the new axioms are purely universal,[2] Theorem 4.1.9 extends to this system, which we denote by $\mathcal{U}^\omega$.

In $\mathcal{U}^\omega$, we can provide a formula that represents the property P lifted to the sequence of random variables represented by X:

Definition 5.1.3. We say that X satisfies $\tilde{P}$ almost uniformly, and write $\tilde{P}(X)$ a.u., if

$$\forall k^0, a^0 \exists b^0 \forall c^0 \left(\mathbb{P}\left(P(a,b,c)^c\right) \leq_{\mathbb{R}} 2^{-k}\right).$$

Similarly, we define $\tilde{Q}(X)$ a.u.

If we consider the previous example for P_0 given in $(**)$, then by formulating $\tilde{P}(X)$ a.u. in this case, we recover the notion of almost uniform convergence (see Remark 4.2.13).

We now provide a relationship between statements of the form

$$\forall x^{1(0)}(\tilde{P}(x) \to \tilde{Q}(x))$$

and statements of the form

$$\tilde{P}(X) \text{ a.u.} \to \tilde{Q}(X) \text{ a.u.}$$

which will not only establish an upgrade-type theorem from relations between modes of convergence for sequences of reals to sequences of random variables but also allow for a transfer of the computational information obtainable for the implication in the premise to the implication in the conclusion.

Theorem 5.1.4. *Provably in $\mathcal{U}^\omega$, given functionals V, A, C such that*

$$\forall x \underbrace{x^*, B, u, w}_{\omega}(\forall n^0(x^*(n) \leq_0 x^*(n+1) \wedge x^*(n) \geq_{\mathbb{R}} |x(n)|_{\mathbb{R}}) \wedge P_0(A\omega, B(A\omega), C\omega, x) \\ \to Q_0(u, Vx^*Bu, w, x)),$$

we can construct V', A', C' such that

$$\forall \underbrace{B, k, u, w}_{\alpha} \left(\mathbb{P}(P(A'\alpha, Bk(A'\alpha), C'\alpha)^c) \leq 2^{-k} \to \mathbb{P}(Q(u, V'Bku, w)^c) \leq 2^{-k}\right).$$

[1] This would be the case if we treated the reals intensionally, via an abstract type. This notion of majorization will not be equivalent to the majorization of the reals as type 1 objects.

[2] As well as the fact that the new constants are majorizable.

Proof. Given such V, A, C and $\alpha = (B, k, u, w)$, we define

$$\begin{aligned} A'\alpha &:= A\tau(Bk)uw, \\ C'\alpha &:= C\tau(Bk)uw, \\ V'Bku &:= V\tau(Bk)u. \end{aligned}$$

Let z be arbitrary with $z \in Q(u, V'Bku, w)^c$. By the axioms of $\mathcal{U}^\omega$ and the definition of V', we have

$$\begin{aligned} z \in Q(u, V'Bku, w)^c &\leftrightarrow z \in Q(u, V\tau(Bk)u, w)^c \\ &\leftrightarrow \neg Q_0(u, V\tau(Bk)u, w, \lambda n.X(n)(z)) \end{aligned}$$

and the latter implies

$$\neg P_0(A\tau(Bk)uw, Bk(A\tau(Bk)uw), C\tau(Bk)uw, \lambda n.X(n)(z))$$

using the assumptions on V, A, C and that $\tau(n) \geq |X(n)(z)|$. This is, by definition of A', V', C', equivalent to

$$\neg P_0(A'\alpha, Bk(A'\alpha), C'\alpha, \lambda n.X(n)(z))$$

and thus to

$$z \in P(A'\alpha, Bk(A'\alpha), C'\alpha)^c.$$

Thus, we have

$$Q(u, V'Bku, w)^c \subseteq P(A'\alpha, Bk(A'\alpha), C'\alpha)^c$$

as z above was arbitrary, and therefore, we get

$$\mathbb{P}(Q(u, V'Bku, w)^c) \leq \mathbb{P}(P(A'\alpha, Bk(A'\alpha), C'\alpha)^c)$$

by the monotonicity of $\mathbb{P}$. This yields the claim. □

Remark 5.1.5. While this result initially looks rather technical and abstract, there are many concrete instances, such as Kronecker's lemma (which we shall see in the following section).

Observe that the conclusion of Theorem 5.1.4 is just a witnessed version of the Dialectica interpretation of

$$\tilde{P}(X) \text{ a.u.} \to \tilde{Q}(X) \text{ a.u.} \tag{+}$$

and therefore, this witnessed Dialectica interpretation, in particular, implies (+). In addition, the conclusion of Theorem 5.1.4 allows for the extraction of quantitative information in the sense that the functional V' transforms a rate for the premise $\tilde{P}(X)$ a.u. into a rate for the

conclusion $\tilde{Q}(X)$ a.u. Even further, V' can be constructed from V (from the proof of the result).

The premise of Theorem 5.1.4 is essentially the Dialectica interpretation of the statement

$$\forall x^{1(0)}(\tilde{P}(x) \to \tilde{Q}(x)) \tag{++}$$

in the sense that the functionals V, A, C represent realizers for this interpretation with the additional assumption that these realizers are suitably uniform, depending only on an upper bound for the sequence $x^{1(0)}$. Although one can construct examples where such uniformity of the realizers is not the case, in practice, for many theorems of the form (++) that have a semi-constructive proof, such uniform realizers can be given. In particular, this is true for Kronecker's Lemma.

We now present a counterexample demonstrating the necessity of the majorizability of the sequence of random variables in Theorem 5.1.4:

Remark 5.1.6. For the above transfer principle to hold, the assumption of the boundedness of the sequence of random variables is necessary, as the following example shows: Take $\Omega := \mathbb{N}$ and let S be the collection of all finite and co-finite subsets of $\mathbb{N}$, i.e.

$$S := \{A \subseteq \mathbb{N} : A \text{ is finite or } A^c \text{ is finite}\}.$$

Furthermore, define the probability content $\mathbb{P}$ by $\mathbb{P}(A) = 0$ if A is finite and $\mathbb{P}(A) = 1$ if A^c is finite, for all $A \in S$. Now, we consider the two properties

$$\tilde{P}(x) = P_0(x) \equiv 0 = 0 \text{ and } \tilde{Q}(x) \equiv \exists n \forall m Q_0(n, m, x) \equiv \exists n \forall m (n \geq_{\mathbb{Q}} [\widehat{x_0}](m))$$

for a sequence $x = (x_n)$ of real numbers. Clearly, both P and Q are Π^0_3-formulas and are trivially true for all sequences x. Therefore also $\tilde{P}(x) \to \tilde{Q}(x)$ is trivially true. Further, we can easily give V, A, C that satisfy the assumptions of Theorem 5.1.4. Now, for a fixed sequence of random variables $\{X_n\}$ taking rational values, we have the set $Q(n, m)$ corresponding to Q_0 is just

$$Q(n, m) = \{k \in \mathbb{N} \mid Q_0(n, m, \lambda l.X_n(k)\} = \{k \in \mathbb{N} \mid n \geq_{\mathbb{Q}} [X_0(k)]\}.$$

For each n, setting,

$$X_n : \mathbb{N} \to \mathbb{N}, \quad k \mapsto [k]_{\mathbb{Q}}$$

yields

$$Q(n, m) = \{k \in \mathbb{N} \mid n \geq k\}$$

which belongs to S as it is finite. P_0 is just represented by the full set $\mathbb{N}$. Therefore, X satisfies P almost uniformly and does not satisfy Q almost uniformly as any $Q(n, m)^c$ has measure 1.

5.2 The computational content of Kronecker's lemma

Kronecker's lemma is a key result in analysis, and its probabilistic analogue, Theorem 5.1.2, is a crucial result in probability theory, typically used to establish Strong Laws of Large Number (c.f. Chapter 6). In this section, we investigate the computational content of Kronecker's Lemma by extracting rates of convergence and metastability. Motivated by the uniformity observed in our computational interpretation of Kronecker's lemma, we generalise the transfer result presented in Theorem 5.1.4.

Furthermore, we explore the computability theory of Kronecker's lemma, investigating cases in which computable convergence rates for the conclusion of Kronecker's lemma are impossible and investigate the proof-theoretic strength of the result via the reverse mathematics program.

5.2.1 Rates for Kronecker's lemma

We start by giving a proof of Kronecker's lemma for sequences of elements in a general normed space $(\mathbb{B}, \|\cdot\|)$. The proof we present is a fleshed-out version of the proof of Theorem $A.6.2$ in [56].

Theorem 5.2.1 (Kronecker's lemma on $\mathbb{B}$)**.** *Let $\{x_n\}$ be a sequence of elements in $\mathbb{B}$ and $0 < a_0 \leq a_1 \leq \ldots$ be such that $a_n \to \infty$. If $\{\sum_{i=0}^{n} x_i\}$ is Cauchy, then*

$$\frac{1}{a_n}\sum_{i=0}^{n} a_i x_i \to 0$$

as $n \to \infty$.

Proof. Let $\varepsilon > 0$ be given. Define $s_n = \sum_{i=0}^{n} x_i$, by our hypothesis, $\{s_n\}$ is Cauchy. Take $M \in \mathbb{N}$ such that

$$\|s_n - s_M\| < \frac{\varepsilon}{4}$$

for all $n \geq M$. We first observe, by summation by parts, that for all $n \geq M$,

$$\left\| \frac{1}{a_n}\sum_{i=0}^{n} a_i x_i \right\| = \left\| s_n - \frac{1}{a_n}\sum_{i=0}^{n-1}(a_{i+1} - a_i)s_i \right\|$$

the right-hand side of the above becomes,

$$\begin{aligned}
&\left\| s_n - \frac{1}{a_n}\sum_{i=0}^{M-1}(a_{i+1}-a_i)s_i - \frac{1}{a_n}\sum_{i=M}^{n-1}(a_{i+1}-a_i)s_M - \frac{1}{a_n}\sum_{i=M}^{n-1}(a_{i+1}-a_i)(s_i-s_M)\right\| \\
&\leq \left\| s_n - (1-\frac{a_M}{a_n})s_M \right\| + \left\| \frac{1}{a_n}\sum_{i=0}^{M-1}(a_{i+1}-a_i)s_i \right\| + \left\| \frac{1}{a_n}\sum_{i=M}^{n-1}(a_{i+1}-a_i)(s_i-s_M) \right\| \\
&\leq \|s_n - s_M\| + \left\| \frac{a_M s_M}{a_n} \right\| + \left\| \frac{1}{a_n}\sum_{i=0}^{M-1}(a_{i+1}-a_i)s_i \right\| + \frac{1}{a_n}\sum_{i=M}^{n-1}(a_{i+1}-a_i)\|s_i - s_M\| \\
&< \frac{\varepsilon}{4} + \left\| \frac{a_M s_M}{a_n} \right\| + \left\| \frac{1}{a_n}\sum_{i=0}^{M-1}(a_{i+1}-a_i)s_i \right\| + \frac{\varepsilon}{4}\frac{1}{a_n}\sum_{i=M}^{n-1}(a_{i+1}-a_i) \\
&\leq \frac{\varepsilon}{2} + \left\| \frac{a_M s_M}{a_n} \right\| + \left\| \frac{1}{a_n}\sum_{i=0}^{M-1}(a_{i+1}-a_i)s_i \right\|.
\end{aligned}$$

Now since M is fixed and $\{a_n\}$ is an increasing sequence that tends to infinity, we can take n large enough to ensure $\|\frac{a_M s_M}{a_n}\|$ and $\|\frac{1}{a_n}\sum_{i=0}^{M-1}(a_{i+1}-a_i)s_i\|$ are both $< \frac{\varepsilon}{4}$. □

It turns out that the direct computational interpretation that one can give to Kronecker's lemma, that is, obtaining rates for the conclusion in terms of rates from the premise, is too weak to be able to get a computational interpretation of the probabilistic Kronecker's lemma (in particular to apply the transfer result presented in Theorem 5.1.4). We need a stronger result, which can be seen as a finitary quantitative formulation of Kronecker's lemma similar to results in [123, 124]. This result will also tell us information about what error in the premise is required to produce the error we want in the conclusion.

Theorem 5.2.2 (Finitary Kronecker's lemma)**.** *Let $\{x_n\}$ be a sequence of elements in $\mathbb{B}$ and $0 < a_0 \leq a_1 \leq \ldots$. For each $n \in \mathbb{N}$ and $x \geq 0$, define, $s_n := \sum_{i=0}^{n} x_i$ and $f_{\{a_n\}}(x) := \min\{n \in \mathbb{N} : a_n \geq x\}$.*

Now for every function $\gamma : \mathbb{Q}^+ \to \mathbb{N}$, sequence of natural numbers $\{z_n\}$, $\varepsilon \in \mathbb{Q}^+$ and $w \in \mathbb{N}$, if $M := \gamma(\frac{\varepsilon}{4})$ satisfies, $\forall i \leq M\,(z_M \geq \|s_i\|)$ and

$$\|s_n - s_M\| < \frac{\varepsilon}{4} \tag{5.1}$$

for all $n \in [M, w]$, then $N := \Gamma_{\{a_n\}}(\gamma, \{z_n\}, \varepsilon)$ satisfies,

$$\left\| \frac{1}{a_n}\sum_{i=0}^{n} a_i x_i \right\| < \varepsilon$$

for all $n \in [N, w]$. Where

$$\Gamma_{\{a_n\}}(\gamma, \{z_n\}, \varepsilon) := \max\left\{\gamma\left(\frac{\varepsilon}{4}\right), f_{\{a_n\}}\left(\frac{4a_{\gamma(\frac{\varepsilon}{4})}z_{\gamma(\frac{\varepsilon}{4})}}{\varepsilon}\right)\right\}.$$

Proof. We have for each $n \in [N, w]$,

$$\begin{aligned}\left\|\frac{1}{a_n}\sum_{i=0}^{n} a_i x_i\right\| &\leq \|s_n - s_M\| + \left\|\frac{a_M s_M}{a_n}\right\| + \left\|\frac{1}{a_n}\sum_{i=0}^{M-1}(a_{i+1}-a_i)s_i\right\| + \frac{1}{a_n}\sum_{i=M}^{n-1}(a_{i+1}-a_i)\|s_i - s_M\| \\ &< \frac{\varepsilon}{4} + \left\|\frac{a_M s_M}{a_n}\right\| + \left\|\frac{1}{a_n}\sum_{i=0}^{M-1}(a_{i+1}-a_i)s_i\right\| + \frac{\varepsilon}{4}\frac{1}{a_n}\sum_{i=M}^{n-1}(a_{i+1}-a_i) \\ &\leq \frac{\varepsilon}{2} + \frac{a_M z_M}{a_n} + \frac{a_M z_M}{a_n} \leq \varepsilon.\end{aligned}$$

The first line follows from precisely the first three lines in the calculation in the proof of Theorem 5.2.1. To get the second line, we use (5.1) to bound the first term and the fact that $[N, w] \subseteq [M, n-1]$ (since $N \geq M$ and $n \leq w$) and (5.1) allows us to bound the last term. The third line follows from the bounding condition of $\{z_n\}$ on $\{s_n\}$ and simplification. □

We can now obtain a quantitative version of Kronecker's lemma, which translates rates from the premise to rates for the conclusion.

Corollary 5.2.3. *Let $\{x_n\}$, $\{a_n\}$, $\{s_n\}$, $f_{\{a_n\}}$ be as in Theorem 5.2.2 and let $\{z_n\}$ be a sequence of nondecreasing natural numbers satisfying $z_n \geq \|s_n\|$ for all n.*

Suppose $\{s_n\}$ is Cauchy with rate of metastability Φ. Then

$$\frac{1}{a_n}\sum_{i=0}^{n} a_i x_i$$

converges to 0 with rate of metastability

$$\kappa_{\Phi,\{a_n\},\{z_n\}}(\varepsilon, g) := \max\left\{Q, f_{\{a_n\}}\left(\frac{4a_Q z_Q}{\varepsilon}\right)\right\}$$

where, $Q := \Phi(\frac{\varepsilon}{4}, h_{\varepsilon,g,\{a_n\},\{z_n\}})$ and

$$h_{\varepsilon,g,\{a_n\},\{z_n\}}(n) := \tilde{g}\left(\max\left\{n, f_{\{a_n\}}\left(\frac{4a_n z_n}{\varepsilon}\right)\right\}\right)$$

with $\tilde{g}(n) = n + g(n)$.

Proof. Let $\varepsilon > 0, g : \mathbb{N} \to \mathbb{N}$ be given. By definition $\exists M \leq Q = \Phi(\frac{\varepsilon}{4}, h_{\varepsilon,g,\{a_n\},\{z_n\}})$ such that

$$|s_n - s_M| < \frac{\varepsilon}{4}$$

for all $n \in [M, h_{\varepsilon,g,\{a_n\},\{z_n\}}(M)] \subseteq [M, M + h_{\varepsilon,g,\{a_n\},\{z_n\}}(M)]$. Now letting $\gamma(\sigma) = M$, for all $\sigma \in \mathbb{Q}^+$ gives, by Theorem 5.2.2,

$$N = \Gamma_{\{a_n\}}(\gamma, \{z_n\}, \varepsilon) = \max\left\{M, f_{\{a_n\}}\left(\frac{4a_M z_M}{\varepsilon}\right)\right\} \leq \max\left\{Q, f_{\{a_n\}}\left(\frac{4a_Q z_Q}{\varepsilon}\right)\right\},$$

(the last inequality follows since $M \leq Q$ and $f_{\{a_n\}}, \{a_n\}, \{z_n\}$ are all non-decreasing) and $w = h_{\varepsilon,g,\{a_n\},\{z_n\}}(M) = N + g(N)$ satisfies

$$\left\| \frac{1}{a_n} \sum_{i=0}^{n} a_i x_i \right\| < \varepsilon$$

for all $n \in [N, N + g(N)]$, so we are done. □

Remark 5.2.4. In light of Theorem 2.3.11, if Φ above is a rate of convergence, then we get a rate of convergence to 0 given by the above expression, but with

$$Q := \Phi(\varepsilon/4).$$

Remark 5.2.5. In both Theorem 5.2.2 and Corollary 5.2.3 we can replace $f_{\{a_n\}}$ by any nondecreasing function f^* bounding $f_{\{a_n\}}$, that is, satisfying $f^*(x) \geq f_{\{a_n\}}(x)$ for all $x \geq 0$.

5.2.2 Computability of rates and the reverse mathematics of Kronecker's lemma

We showed in Example 2.3.3 that there exist sequences of converging rational numbers that do not converge with a computable rate of convergence. We constructed a bounded monotone sequence of rational numbers that converge without a computable rate of Cauchy convergence, thus demonstrating that general rates of convergences cannot be extracted from any proof of the monotone convergence principle. A modification of this construction yields a similar result for Kronecker's lemma:

Example 5.2.6. We can construct a sequence of rational numbers $\{x_n\}$ such that $\sum_{i=0}^{\infty} x_i$ converges, but $\frac{1}{n+1}\sum_{i=0}^{n}(i+1)x_i$ converges to 0, without a computable rate of convergence.

Let A be a recursively enumerable set that is not recursive (e.g. the halting set). Let $\{a_n\}$ be a recursive enumeration of the elements in A. Let $x_i = 2^{-a_i}$. So $\{x_n\}$ is a positive sequence of rational numbers and we have,

$$\sum_{i=0}^{\infty} x_i \leq \sum_{i=0}^{\infty} 2^{-i} = 1.$$

Now suppose, for contradiction, $\frac{1}{n+1}\sum_{i=0}^{n}(i+1)x_i$ converges to 0 with a computable rate of convergence ϕ. We describe an effective procedure to determine whether $k \in A$, for all $k \in \mathbb{N}$. Suppose, $k = a_n$ with $n > \phi(2^{-k})$, then

$$\frac{1}{n+1}\sum_{i=0}^{n}(i+1)x_i \leq 2^{-k}$$

which implies,

$$\sum_{i=0}^{n}(i+1)x_i \leq (n+1)2^{-a_n}.$$

This is clearly a contradiction, as $n \geq 1$ and $\{x_n\}$ are positive. Thus, if $k = a_n$ then $n \leq \phi(2^{-k})$. So we can determine whether $k \in A$, by computably searching the first $\phi(2^{-k})$ terms in $\{a_n\}$.

The above construction demonstrates that a general rate of convergence for Kronecker's lemma must depend on a rate of convergence for $\sum_{i=0}^{\infty} x_i$.

Following [135], let RCA_0 be the standard base system of reverse mathematics (the subsystem of second order arithmetic containing only Σ_1^0 induction and Δ_1^0 comprehension) and the system ACA_0 which extends RCA_0 by arithmetic comprehension.[3]

Here, we shall be working with the language of second-order arithmetic, where we quantify over variables representing natural numbers and subsets of natural numbers. Furthermore, via the numerically defined paring function

$$(m, n) := (m + n)^2 + m$$

we can encode the integers, rationals and reals. In addition, as standard, for set variables $X, Y, f \in 2^{\mathbb{N}}$, $f : X \to Y$ is shorthand for

$$\forall l, n, m \in \mathbb{N}\,((l, n) \in f \wedge (l, m) \in f \to n = m)$$

and for a formula $\phi(f)$ in the language of second order arithmetic,

$$\forall f : X \to Y\,(\phi(f))$$

is shorthand for

$$\forall f \in 2^{\mathbb{N}}\,((f : X \to Y) \to \phi(f))$$

[3]Arithmetic comprehension is the scheme

$$\exists X\,\forall n\,(n \in X \leftrightarrow \phi(n))$$

for all arithmetic formulas ϕ (formulas with no bound set variables) without X occurring as a free variable.

with a similar convention for

$$\exists f : X \to Y\,(\phi(f)).$$

Moreover, we write $f(n) = m$ for $(n, m) \in f$.

In what follows, we need the standard fact that ACA_0 proves the axiom of choice on arithmetic formulas. More specifically:

Proposition 5.2.7 ([135])**.** *The axiom of choice for arithmetic formulas is provable in* ACA_0*. That is, the following holds:*

$$\forall n \in \mathbb{N}\,\exists m \in \mathbb{N}\,\phi(n, m) \to \exists f : \mathbb{N} \to \mathbb{N}\,\forall n\,\phi(n, f(n))$$

for all arithmetic formulas ϕ*, without* f *occurring free. Here* $\phi(n, f(n))$ *is shorthand for* $\exists m\,(f(n) = m \wedge \phi(n, m)))$.

Proof. By arithmetic comprehension, we have

$$\exists f \in 2^{\mathbb{N}}\,\forall n, m \in \mathbb{N}\,((n, m) \in f \leftrightarrow (\phi(n, m) \wedge (\forall m_0 \in \mathbb{N}\,(\phi(n, m_0) \to m \leq m_0)))).$$

Taking such an f, one can easily show $f : \mathbb{N} \to \mathbb{N}$ and satisfies the consequence of the implication in the statement of the result. □

Remark 5.2.8. A direct application of arithmetic comprehension in the proof of the previous proposition actually yields

$$\exists f \in 2^{\mathbb{N}}\,\forall n \in \mathbb{N}\,(n \in f \leftrightarrow (\exists i, j \in \mathbb{N}\,(n = (i, j) \wedge (\phi(i, j) \wedge (\forall j_0 \in \mathbb{N}\,(\phi(i, j_0) \to j \leq j_0))))))$$

and our stated formula follows.

Now, for sequences of real numbers $\{x_n\}, \{a_n\}$ let

$$\begin{aligned}\mathrm{KRON}(\{a_n\}, \{x_n\}) :\equiv{}& \forall n \in \mathbb{N}\,(0 < a_n \leq a_{n+1}) \wedge \forall m \in \mathbb{N}\,\exists k \in \mathbb{N}\,(a_k \geq m) \\ &\wedge \left(\left\{\sum_{i=0}^{n} x_n\right\} \text{ is Cauchy}\right) \to \frac{1}{a_n}\sum_{i=0}^{n} a_i x_i \text{ converges to } 0\end{aligned}$$

and

$$\begin{aligned}&\mathrm{RKRON}(\{a_n\}, \{x_n\}) :\equiv \forall n \in \mathbb{N}\,(0 < a_n \leq a_{n+1}) \wedge \forall m \in \mathbb{N}\,\exists k \in \mathbb{N}\,(a_k \geq m) \\ &\wedge \left(\left\{\sum_{i=0}^{n} x_n\right\} \text{ is Cauchy}\right) \to \exists g : \mathbb{N} \to \mathbb{N}\left(g \text{ is a rate of convergence to } 0 \text{ for } \frac{1}{a_n}\sum_{i=0}^{n} a_i x_i\right).\end{aligned}$$

Here we use the ' 2^{-k}' formulation of convergence, for example, by ' g is a rate of convergence

to 0 for $\frac{1}{a_n}\sum_{i=0}^{n} a_i x_i$' we mean

$$\forall k \in \mathbb{N}\,\forall n \in \mathbb{N}\left(n \geq g(k) \to \left|\frac{1}{a_n}\sum_{i=0}^{n} a_i x_i\right| \leq 2^{-k}\right).$$

We have the following:

Theorem 5.2.9. *In* RCA_0 *we have the following:*

(i) For all sequences of reals $\{x_n\}$ *and* $\{a_n\}$, $\mathrm{KRON}(\{a_n\},\{x_n\})$.

(ii) ACA_0 *implies, for all sequences of reals* $\{x_n\}$ *and* $\{a_n\}$, $\mathrm{RKRON}(\{a_n\},\{x_n\})$.

(iii) For all sequences of positive rationals $\{x_n\}$ $(\mathrm{RKRON}(\{n+1\},\{x_n\}))$ *implies* ACA_0.

Proof. Write $s_n := \sum_{i=0}^{n} x_i$.

For (i), suppose we have a sequence of reals $\{a_n\}$ that is increasing, positive, and satisfies $\forall m \in \mathbb{N}\,\exists k \in \mathbb{N}\,(a_k \geq m)$, as well as a sequence of reals $\{x_n\}$, such that $\{s_n\}$ is Cauchy. Now RCA_0 proves that Cauchy sequences are bounded so that we can take $S \in \mathbb{N}$ such that $\forall n \in \mathbb{N}\,(|s_n| < S)$. Let $k \in \mathbb{N}$ be given. By the Cauchy property of $\{s_n\}$, we may take $n \in \mathbb{N}$ such that, for all $m \geq n$ we have $|s_m - s_n| \leq 2^{-k-2}$. Now taking W such that $a_W \geq 2^{k+2} a_n S$ and following the proof of Theorem 5.2.1 implies that for $m \geq \max\{W, n\}$, $\left|\frac{1}{a_m}\sum_{i=0}^{m} a_i x_i\right| \leq 2^{-k}$.

For (ii), we have that for an increasing, positive sequence of reals $\{a_n\}$ satisfying $\forall m \in \mathbb{N}\,\exists k \in \mathbb{N}\,(a_k \geq m)$, and a sequence $\{x_n\}$ with $\{s_n\}$ Cauchy, part (i) implies that in RCA_0 we can prove $\frac{1}{a_n}\sum_{i=0}^{n} a_i x_i$ converges to 0, and the result follows from an application of the axiom of choice on arithmetic formulas.

For (iii), we demonstrate that we can construct the range of a given one-to-one function $f : \mathbb{N} \to \mathbb{N}$, and the result follows from [135, Lemma III.1.3]. Take such an $f : \mathbb{N} \to \mathbb{N}$ and let $x_i := 2^{-f(i)}$. Then, since RCA_0 proves that monotone bounded sequences are Cauchy, we have $\{s_n\}$ is Cauchy. Therefore, by $(\mathrm{RKRON}(\{n+1\},\{x_n\}))$, we have $g : \mathbb{N} \to \mathbb{N}$ satisfying

$$\forall k \in \mathbb{N}\,\forall n \in \mathbb{N}\left(n \geq g(k) \to \frac{1}{n+1}\sum_{i=0}^{n}(i+1)2^{-f(i)} \leq 2^{-k}\right).$$

One then has that for all $k \in \mathbb{N}$,

$$(\exists n\,(f(n) = k)) \leftrightarrow (\exists n \leq g(k)\,(f(n) = k)), \tag{5.2}$$

the backwards implication is clear. For the forward implication, if we do not have $(\exists n \leq g(k)\,(f(n) = k))$ but $(\exists n\,(f(n) = k))$, then $(\exists n > g(k)\,(f(n) = k))$, and for such an n, we have

$$\frac{1}{n+1}\sum_{i=0}^{n}(i+1)2^{-f(i)} \leq 2^{-k} = 2^{-f(n)}$$

a contradiction. Thus, (5.2) and Δ_1^0 comprehension implies $\exists X \in 2^{\mathbb{N}} \forall k \in \mathbb{N} (k \in X \leftrightarrow (\exists n (f(n) = k)))$ and the result follows. □

Remark 5.2.10. Example 5.2.6 demonstrates that when we assume that $\{x_n\}$ is nonnegative, a general computable rate of convergence for Kronecker's lemma, which depends on a computable bound for $\sum_{i=0}^{\infty} x_i$, cannot exist.

However, we have seen in Theorem 2.3.5 that if $\{a_n\}$ is a nondecreasing sequence of nonnegative numbers bounded above by $L > 0$, then

$$\Phi(\varepsilon, g) := g^{\lfloor \frac{L}{\varepsilon} \rfloor}(0)$$

is a rate of metastable convergence for $\{a_n\}$.

When $\{x_n\}$ is a nonnegative sequence, the partial sums will form a nondecreasing sequence. So, although we cannot hope to find a rate of convergence for $\frac{1}{a_n} \sum_{i=0}^{n} a_i x_i \to 0$ that just depends on a bound for $\sum_{i=0}^{\infty} x_i$, such a bound would give us a rate of metastability for the convergence of the partial sums of $\{\sum_{i=0}^{n} x_i\}$ and we can use Corollary 5.2.3 to obtain a rate of metastability for $\frac{1}{a_n} \sum_{i=0}^{n} a_i x_i \to 0$.

5.2.3 A transfer principle for almost surely finite random variables

One can weaken the boundedness assumption in Theorem 5.1.4. However, random variables in $\mathcal{F}^\omega[\mathbb{P}]$ are treated as objects of type $1(\Omega)$, and a constant majorizes such objects in this system. Therefore, Theorem 4.1.9, extended to $\mathcal{U}^\omega$, only guarantees bound extraction for theorems regarding bounded random variables. And so any improvement of Theorem 5.1.4 that weakens the boundedness assumption will result in no improvement if used alongside Theorem 4.1.9. However, such an improvement is still possible, and we shall present it in this section. For better clarity, we present this result in regular mathematical terms. That is, we do not formalise this result in an extension of $\mathcal{F}^\omega[\mathbb{P}]$. As such, we start by presenting notions from Section 5.1.1 in informal terms.

For a sequence of real numbers $\{x_n\}$ fix the two Π_3-formulas

$$P(\{x_n\}) = \forall a^0 \exists b^0 \forall c^0 P_0(a, b, c, \{x_n\})$$

and

$$Q(\{x_n\}) = \forall u^0 \exists v^0 \forall w^0 Q_0(u, v, w, \{x_n\})$$

where P_0 and Q_0 are quantifier-free.

Now, given a sequence of real valued random variables $\{X_n\}$, we say $\{X_n\}$ satisfies P almost

uniformly, and write $P(\{X_n\})$ a.u, if

$$\forall k^0, a^0 \exists b^0 \forall c^0 \left(\mathbb{P}\left(P_0(a,b,c,\{X_n\})^c\right) \leq 2^{-k}\right).$$

$Q(\{X_n\})$ a.u is defined similarly.

Now define the majorizability relation $\{\tau_n\} \gtrsim \{x_n\}$ by,

$$\{\tau_n\} \gtrsim \{x_n\} :\equiv \forall n \in \mathbb{N}\,(\tau_{n+1} \geq \tau_n \wedge \tau_n \geq |x_n|).$$

With these notations, Theorem 5.1.4 states:

Theorem 5.2.11. *Given functionals V, A, C such that for all sequences of real numbers $\{x_n\}$*

$$\forall \underbrace{\{\tau_n\}, B, u, w}_{\omega}\,(\{\tau_n\} \gtrsim \{x_n\} \wedge P_0(A(\omega), B(A(\omega)), C(\omega), \{x_n\}) \to Q_0(u, V(\{\tau_n\}, B, u), w, \{x_n\}))\,,$$

(where we quantify over all sequences of natural numbers $\{\tau_n\}$) then for all sequences of bounded random variables $\{X_n\}$, we can construct V', A', C' such that

$$\forall \underbrace{B, k, u, w}_{\alpha}(\mathbb{P}(P_0(A'(\alpha), B(k, A'(\alpha)), C'(\alpha), \{X_n\})^c) \leq 2^{-k} \to \mathbb{P}(Q_0(u, V'(B,k,u), w, \{X_n\})^c) \leq 2^{-k})$$

Furthermore, the functionals V', A', C' can be constructed from the proof, and depend on V, A, C and also on $\{X_n\}$, via a nondecreasing sequence of natural numbers $\{Z_n\}$ witnessing the boundedness of $\{X_n\}$, that is, satisfying for every $n \in \mathbb{N}$ and $\omega \in \Omega$, we have $Z_n \geq |X_n(\omega)|$.

If we take

$$P_0(a,b,c,\{x_n\}) := \forall m, n \in [b;c] \left|\sum_{k=0}^{n} x_k - \sum_{k=0}^{m} x_k\right| \leq 2^{-a}$$

and

$$Q_0(u,v,w,\{x_n\}) := \forall n \in [v;w] \left|\frac{1}{a_n}\sum_{k=0}^{n} a_k x_k\right| \leq 2^{-u}$$

with the sufficient assumptions on $\{a_n\}$, Kronecker's lemma becomes:

$$\forall \{x_n\}(P(\{x_n\}) \to Q(\{X_n\}).$$

and the probabilistic Kronecker's lemma is

$$P(\{X_n\}) \text{ a.u } \to Q(\{X_n\}) \text{ a.u }.$$

Thus, from the proof of Theorem 5.2.2, taking

$$\begin{aligned}
&A(\{\tau_n\}, B, u, w) := u + 2\\
&C(\{\tau_n\}, B, u, w) := w\\
&V(\{\tau_n\}, B, u) := \max\left\{B(u+2), f_{\{a_n\}}\left(2^{u+2}a_{B(u+2)}\sum_{i=0}^{B(u+2)}\tau_i\right)\right\}
\end{aligned}$$

with $f_{\{a_n\}}$ as defined in Theorem 5.2.2, allows the premise for Theorem 5.2.11 to be satisfied. The proof of Theorem 5.2.11 (given in the formal system as Theorem 5.1.4) can be used to obtain a quantitative version of the probabilistic Kronecker's lemma for bounded random variables. However, the computational solution to Kronecker's lemma we obtained contained further uniformities that allow us to reduce the boundedness condition; that is, the functional A, C and V are uniformly continuous in their first argument and, in particular, all have a modulus of continuity given by $M(B, u, w) := B(u+2)$, that is for all sequences $\{\tau_n^1\}$ and $\{\tau_n^2\}$, $B : \mathbb{N} \to \mathbb{N}$ and $u, w \in \mathbb{N}$ if we have $\tau_n^1 = \tau_n^2$ for all $n \leq M(B, u, w)$ then $V(\{\tau_n^1\}, B, u) = V(\{\tau_n^2\}, B, u)$ with the same holding for A and C (the modulus of continuity for A and C is the constant 0 function as these functionals are independent of $\{\tau_n\}$). This arises in Kronecker's lemma because the functionals A and C are independent of $\{\tau_n\}$ and thus are trivially uniformly continuous. Although such a uniform solution is not guaranteed to exist in general, this tends to be the case when one obtains quantitative results on the implications between two convergence statements. Typically, the error that one must apply the premise (A) and how far one must apply the premise (C) does not depend on the sequence. Furthermore, when the rate for the conclusion (in terms of the rate for the premise) does depend on the sequence, it only depends on a finite initial segment of the sequence, thus guaranteeing the uniform continuity requirement (see [91, Remark 3.5] and [125, Remark 3.2] for recent examples of this phenomenon in analysis.)

For such a solution, we have the following general transfer principle:

Theorem 5.2.12. *Suppose we have functionals V, A, C such that for all sequences of real numbers $\{x_n\}$*

$$\forall \underbrace{\{\tau_n\}, B, u, w}_{\omega}\,(\{\tau_n\} \gtrsim \{x_n\} \wedge P_0(A(\omega), B(A(\omega)), C(\omega), \{x_n\}) \to Q_0(u, V(\{\tau_n\}, B, u), w, \{x_n\}))\,,$$

and A, C and V are uniformly continuous in their first argument, each with a modulus of continuity $M(B, u, w)$. Then for all sequences of random variables $\{X_n\}$ with a function $Z : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ satisfying

$$\forall k, p \in \mathbb{N}\,\mathbb{P}\left(\bigcup_{i=0}^{p}\{|X_i| > Z(k, p)\}\right) \leq 2^{-k} \tag{5.3}$$

we can construct V', A', C' in terms of A, B, C, Z, M such that

$$\forall \underbrace{B, k, u, w}_{\alpha}(\mathbb{P}(P_0(A'(\alpha), B(k, A'(\alpha)), C'(\alpha), \{X_n\})^c) \leq 2^{-(k+1)}$$
$$\to \mathbb{P}(Q_0(u, V'(B, k, u), w, \{X_n\})^c) \leq 2^{-k}).$$

Proof. Take V, A, C and M satisfying the premise of the theorem and $\alpha = (B, k, u, w)$. Define

$$A'(\alpha) := A(\{z_n\}, B(k), u, w),$$
$$C'(\alpha) := C(\{z_n\}, B(k), u, w),$$
$$V'(B, k, u) := V(\{z_n\}, B(k), u).$$

With $z_n := Z(k+1, M(B, u, w))$ for all $n \in \mathbb{N}$. We have

$$\begin{aligned}
&\mathbb{P}(Q_0(u, V'(B, k, u), w, \{X_n\})^c) \\
&\leq \mathbb{P}\left(Q_0(u, V'(B, k, u), w, \{X_n\})^c \cap \bigcup_{i=0}^{M(B,u,w)} \{|X_i| > Z(k+1, M(B, u, w))\}\right) \\
&+ \mathbb{P}\left(Q_0(u, V'(B, k, u), w, \{X_n\})^c \cap \bigcap_{i=0}^{M(B,u,w)} \{|X_i| \leq Z(k+1, M(B, u, w))\}\right) \\
&\leq \mathbb{P}\left(Q_0(u, V'(B, k, u), w, \{X_n\})^c \cap \bigcap_{i=0}^{M(B,u,w)} \{|X_i| \leq Z(k+1, M(B, u, w))\}\right) \\
&+ 2^{-(k+1)}.
\end{aligned}$$

Now, take $\delta \in \Omega$ satisfying

$$\neg Q_0(u, V'(B, k, u), w, \{X_n(\delta)\}) \wedge \bigwedge_{i=0}^{M(B,u,w)} (|X_i(\delta)| \leq Z(k+1, M(B, u, w))).$$

For such a δ, define $\{\tau_n(\delta)\}$ to be an arbitrary increasing sequence of natural numbers such that as $\tau_n := Z(k+1, M(B, u, w))$ for $n \leq M(B, u)$ and $|X_n(\delta)| \leq \tau_n(\delta)$ for all n. Therefore, we have $\{\tau_n(\delta)\} \gtrsim \{X_n(\delta)\}$. Therefore, by unwinding the definition of V' and using the fact that for all $n \leq M(B, u, w)$, $\tau_n(\delta) = Z_n$ so, by the continuity of V we have $V(\{z_n\}, B(k), u) = V(\{\tau_n(\delta)\}, B(k), u)$ we must have

$$\neg Q_0(u, V(\{\tau_n(\delta)\}, B(k), u), w, \{X_n(\delta)\})$$

which implies

$$\neg P_0(A(\{\tau_n(\delta)\}, B(k), u, w), B(k, A(\{\tau_n(\delta)\}, B(k), u, w)), C(\{\tau_n(\delta)\}, B(k), u, w), \{X_n(\delta)\}).$$

Now, by the continuity of A and C, and the definition of A' and C' we must have,

$$\neg P_0(A'(\alpha), B(k, A'(\alpha)), C'(\alpha), \{X_n(\delta)\}).$$

Therefore,

$$\begin{aligned} Q_0(u, V'(B, k, u), w, \{X_n\})^c \cap \bigcap_{i=0}^{M(B,u)} \{|X_i| \leq Z(k+1, M(B, u))\} \\ \subseteq P_0(A'(\alpha), B(k, A'(\alpha)), C'(\alpha), \{X_n\})^c \end{aligned}$$

which implies,

$$\begin{aligned} \mathbb{P}\left(Q_0(u, V'(B, k, u), w, \{X_n\})^c \cap \bigcap_{i=0}^{M(B,u)} \{|X_i| \leq Z(k+1, M(B, u))\} \right) \\ \leq \mathbb{P}(P_0(A'(\alpha), B(k, A'(\alpha)), C'(\alpha), \{X_n\})^c) \leq 2^{-(k+1)} \end{aligned}$$

so we are done. □

Condition (5.3) can naturally be satisfied if we give a suitable computational interpretation to the finiteness of random variables.

Proposition 5.2.13. *A real-valued random variable Y is finite almost everywhere iff*

$$\mathbb{P}(|Y| \geq m) \to 0$$

as $m \to \infty$.

Proof. The events

$$A_m = \{\omega \in \Omega : |Y(\omega)| \geq m\}$$

form a decreasing sequence of events; thus, by the continuity of the probability measure, we have,

$$\mathbb{P}(A_m) \to \mathbb{P}\left(\bigcap_{m=0}^{\infty} A_m \right) = \mathbb{P}(\{\omega : |Y(\omega)| = \infty\})$$

as $m \to \infty$, and the result follows. □

The above result can also be obtained through an application of Lemma 4.2.2 by noting that the random variable Y being almost surely finite on the element $\omega \in \Omega$ is equivalent to

the formula $\exists N\,(|Y(\omega)| \leq N)$ which satisfies the monotonicity requirements. We now have the following definition:

Definition 5.2.14. $R : \mathbb{N} \to \mathbb{N}$ is a rate of almost sure finiteness if it is a rate of convergence for

$$\mathbb{P}(|Y| \geq m) \to 0,$$

that is, if it satisfies,

$$\forall k \in \mathbb{N}\,(\mathbb{P}(|Y| \geq R(k)) \leq 2^{-k})$$

Example 5.2.15. If a random variable, Y, is integrable then we have

$$\mathbb{P}(|Y| \geq m) \leq \frac{\mathbb{E}(|Y|)}{m}$$

for all $m > 0$, by Markov's inequality. Thus,

$$R(k) := 2^k E$$

is a rate of almost sure finiteness, for all $E \in \mathbb{N}$ satisfying $E \geq \mathbb{E}(|Y|)$.

Remark 5.2.16. Now we have if $\{X_n\}$ have respective rates of almost finiteness $\{R_n\}$, then for any $k \in \mathbb{N}$ we can take

$$Z(k,p) := \max_{i \leq p} R_i\left(\frac{2^{-k}}{p}\right)$$

in Theorem 5.2.12, as

$$\mathbb{P}\left(\bigcup_{i=0}^{p}\{|X_i| \geq Z(k,p)\}\right) \leq \sum_{i=0}^{p}\mathbb{P}\left(\{|X_i| \geq Z(k,p)\}\right) \leq \sum_{i=0}^{p}\mathbb{P}\left(\left\{|X_i| \geq R_i(2^{-k}/p)\right\}\right) \leq 2^{-k}.$$

Furthermore, if there exists a function $R : \mathbb{N} \to \mathbb{N}$ such that

$$\forall p \in \mathbb{N}\,\mathbb{P}\left(\bigcup_{i=0}^{p}\{|X_i| \geq R(k)\}\right) \leq 2^{-k},$$

we can take $Z(p,k) := R(k)$, for all $k, p \in \mathbb{N}$. Such an R can be seen as a modulus of uniform boundedness as defined in Definition 4.2.8.

Chapter 6

Quantitative Laws of Large Numbers

The Law of Large Numbers is an important concept in probability theory that makes mathematically rigorous the empirical fact that observed outcomes should get closer to the expected value as the sample size increases. Theorists such as Bernoulli, Markov, and Khintchine, among others, obtained such results.[1] However, Kolmogorov [94] is credited with the *Strong Law of Large Numbers*, which states:

Theorem 6.0.1 (Strong Law of Large Numbers)**.** *Suppose $X_0, X_1, \ldots$ are independent, identically distributed (iid) real-valued random variables with $\mathbb{E}(|X_0|) < \infty$. Then,*

$$\frac{1}{n}\sum_{i=0}^{n} X_i \to \mathbb{E}(X_0)$$

almost surely, that is, with probability 1*.*

In what follows, set $S_n := \sum_{i=0}^{n} X_i$. Variants of the above Strong Law of Large Numbers commonly studied in the literature are concerned with weakening the identical distribution assumption in Theorem 6.0.1, which, however, entails that other additional assumptions must be included if we still want to conclude $\frac{S_n}{n} \to 0$ almost surely. One such variant, concerned with the case when the $\{X_n\}$ are no longer identically distributed, is the following other classical result of Kolmogorov:

Theorem 6.0.2 (Kolmogorov's Strong Law of Large Numbers)**.** *Suppose $\{X_n\}$ is a sequence of independent real-valued random variables with $\mathbb{E}(X_n) = 0$ for all $n \in \mathbb{N}$ and*

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2} < \infty. \tag{6.1}$$

Then $\frac{S_n}{n} \to 0$ almost surely.

[1] For a detailed historical and technical explanation of the Laws of Large Numbers, one can refer to Seneta's work [132].

The direct computational interpretation one can attempt to give to the above theorem is obtaining computable rates of almost sure convergence to 0 from a suitable computational interpretation for (6.1). It was observed by the author (through a Specker construction c.f. Example 2.3.3) that this would not be possible given a computable bound for (6.1), and one actually requires the stronger property of a computable rate of convergence for (6.1). Assuming only a bound for (6.1) or, more generally, a rate of metastability for the sum (noting that the sequence of partial sums is a monotone bounded sequence, and thus, we obtain a rate of metastable convergence from Theorem 2.3.5) allows us to obtain a uniform metastable rate for the convergence in the conclusion of Kolmogorov's Strong Law of Large Numbers.

By application of the ideas presented in Section 4.2.1 one can show that $\frac{S_n}{n} \to 0$ almost surely is equivalent to the sequence of real numbers

$$P^*_{n,\varepsilon} := \mathbb{P}\left(\sup_{m\geq n}\left|\frac{S_m}{m}\right| > \varepsilon\right)$$

converging to 0 as $n \to \infty$, $\forall \varepsilon > 0$.

The quantitative content of the Strong Laws of Large Numbers has been studied extensively in the probability literature, in the form of studying large deviation probabilities, that is, finding a rate of convergence for the sequence of real numbers $\{P^*_{n,\varepsilon}\}$, for each $\varepsilon > 0$ (this is typically done implicitly through providing an asymptotic upper bound for $P^*_{n,\varepsilon}$).[2] When studying the quantitative content of the Strong Laws of Large Numbers, in this way, it is common to impose very strong conditions on the distributions of the random variables. For example, in [41, 133] Siegmund and Fill determine $P_{n,\varepsilon}$ up to asymptotic equivalence, under the assumptions the sequence of random variables are iid and that the moment-generating function $\mathbb{E}(e^{tX_0}) < \infty$ for t in a suitable interval. Furthermore, their rates depend heavily on the distribution of X_0.

Not much work has been done in studying the large deviations without strong conditions, such as the moment-generating function condition. This may be because, for weaker conditions, one cannot hope to calculate these probabilities up to asymptotic equivalence. The best we can hope for are bounds on the large deviation probabilities. In 2018, Luzia [110] showed that if $\{X_n\}$ are pairwise independent, identically distributed random variables with finite variance, then for all $\varepsilon > 0$ and $\beta > 1$

$$P^*_{n,\varepsilon} = O\left(\frac{log(n)^{\beta-1}}{n}\right).$$

Furthermore, Luzia gave exact rates for this bound, which were very uniform (they did not depend on the distribution of the random variables, for example). Luzia obtained his bound in a fairly ad-hoc manner but loosely following the elementary proof of the Strong Law of Large Numbers given by Etemadi [40]. We claim that a closer inspection of Etemadi's proof, in line

[2]It is clear that having a rate of convergence for $P^*_{n,\varepsilon}$ for each ε is equivalent to a rate of almost sure convergence.

with how one analyses proofs in the proof mining program, results in a tighter asymptotic upper bound. We do not present such an analysis here; however, we instead present an analysis of the proof of the Strong Law of Large Numbers given in [29], which also gives better bounds than those in [110], but further allows us to obtain general quantitative results applicable to a vast number of Strong Laws of Large Numbers in the literature.

This chapter shall detail the author's exploration of the computational content of the Strong Laws of Large Numbers. We start in Section 6.1, where we present our construction demonstrating that computable rates of almost sure convergence in the conclusion of Kolmogorov's Strong Law of Large Numbers do not exist given only a bound for the sum in the premise of the theorem. In addition, we provide a brief outline of the history of the literature on the quantitative aspects of the Strong Laws of Large Numbers.

Then, in Section 6.2, we provide a computational interpretation for a generalisation of Kolmogorov's Strong Law of Large Numbers on type p Banach spaces given in [143], through the construction of uniform metastable rates. Kronecker's lemma (introduced in Chapter 5) is a crucial result in the proof of this result. Thus, the computational interpretation for Kronecker's lemma and the transfer strategy of obtaining stochastic results from their deterministic analogue, which we gave in Chapter 5, shall be needed in this Section.

We conclude this Chapter in Section 6.3. Here, we provide our quantitative generalisation of the main result of [29], which results in a tighter asymptotic upper bound than that provided by Luzia in [110]. Furthermore, we demonstrate how our general quantitative result allows us to obtain quantitative versions of various Strong Laws of Large Numbers in the literature.

6.1 Quantitative and computable aspects of the Laws of Large Numbers

The purpose of this section is to present a picture of the problem of obtaining the computational interpretation of the Strong Laws of Large Numbers. We start by presenting the computational ineffectiveness of Kolmogorov's Strong Law of Large Numbers through a Specker construction. We then provide an outline of the current landscape of quantitative results concerning the Strong Laws of Large Numbers in the literature.

6.1.1 Quantitative and computable aspects of the Laws of Large Numbers

Through a modification of Specker's construction, Example 2.3.3, we justify that one cannot obtain a computable rate of almost sure convergence, for the conclusion of Kolmogorov's Strong

Law of Large Numbers given just a bound for the sum

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2}.$$

Example 6.1.1. Let us take a recursively enumerable set, A, that is not recursive. Let $\{a_n\}$ be a recursive enumeration of the elements in A.

Let $\{X_n\}$ be an independent sequence of discrete random variables, with distributions given by,

$$\mathbb{P}(X_n = x) := \begin{cases} 2^{-a_n-1} & \text{if } x = n - n2^{-a_n-1} \\ 1 - 2^{-a_n-1} & \text{if } x = -n2^{-a_n-1} \\ 0 & \text{Otherwise.} \end{cases}$$

Then, one can easily see that,

$$\mathbb{E}(X_n) = 0,\ \sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2} \leq \frac{5}{12} \text{ and, } \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}(|X_n|) \leq 1.$$

However, there is no computable function $\phi : \mathbb{Q}^+ \times \mathbb{Q}^+ \to \mathbb{N}$ such that

$$\forall \varepsilon, \lambda \in \mathbb{Q}^+ \, \forall n \geq \phi(\varepsilon, \lambda) \ \left(\mathbb{P}\left(\sup_{m \geq n} \left| \frac{S_m}{m} \right| > \varepsilon \right) \leq \lambda \right).$$

To show this, let

$$x_n := \sum_{k=1}^{n} k2^{-a_k-1}$$

and observe that we can write $S_n = K_n - x_n$, with $K_n \in \mathbb{N}$.

Suppose there is such a computable function ϕ, such that for all $\varepsilon, \lambda \in \mathbb{Q}^+$ and $n \geq \phi(\varepsilon, \lambda)$

$$\mathbb{P}\left(\max_{m \geq n} \left| \frac{1}{m} S_m \right| > \varepsilon \right) \leq \lambda.$$

This is equivalent to

$$\mathbb{P}\left(\forall m \geq n \ \left| \frac{1}{m} S_m \right| \leq \varepsilon \right) > 1 - \lambda,$$

which is equivalent to

$$\mathbb{P}\left(\forall m \geq n \ \left(-\varepsilon + \frac{1}{m} x_m \leq \frac{1}{m} K_m \leq \varepsilon + \frac{1}{m} x_m \right) \right) > 1 - \lambda. \tag{6.2}$$

We now describe an effective procedure to determine whether $k \in \mathbb{N}$ is in A, which will contradict the assumption that A is not a recursive set, leading us to the conclusion that no computable function ϕ can exist. Suppose $M \geq \phi(\frac{1}{2}, 2^{-k-1})$. We have, from (6.2),

$$\mathbb{P}\left(\forall m \geq \phi\left(\frac{1}{2}, 2^{-k-1}\right)\left(-\frac{1}{2}+\frac{1}{m}x_m \leq \frac{1}{m}K_m \leq \frac{1}{2}+\frac{1}{m}x_m\right)\right) > 1-2^{-k-1}.$$

Now observe

$$\frac{1}{M}x_M = \frac{1}{M}\sum_{k=1}^{M} k2^{-a_k-1} < \sum_{k=1}^{M} 2^{-a_k-1} < \frac{1}{2}. \tag{6.3}$$

We have that,

$$\forall m \geq \phi\left(\frac{1}{2}, 2^{-k-1}\right)\left(-\frac{1}{2}+\frac{1}{m}x_m \leq \frac{1}{m}K_m \leq \frac{1}{2}+\frac{1}{m}x_m\right)$$

implies

$$-\frac{1}{2}+\frac{1}{M}x_M \leq \frac{1}{M}K_M \leq \frac{1}{2}+\frac{1}{M}x_M.$$

This further implies $\frac{1}{M}K_M < 1$ by (6.3), which implies $X_M = -M2^{-a_M-1}$. Thus we have,

$$\begin{aligned}
&\mathbb{P}(X_M = -M2^{-a_M-1}) \\
&\qquad \geq \mathbb{P}\left(\forall m \geq \phi\left(\frac{1}{2}, 2^{-k-1}\right)\left(-\frac{1}{2}+\frac{1}{m}x_m \leq \frac{1}{m}K_m \leq \frac{1}{2}+\frac{1}{m}x_m\right)\right) \\
&\qquad > 1-2^{-k-1}.
\end{aligned}$$

So, $1-2^{-a_M-1} > 1-2^{-k-1}$ which implies $a_M > k$. Thus if $M \geq \phi(\frac{1}{2}, 2^{-k-2})$ then $k \neq a_M$. Thus to effectively determine if $k \in A$, it suffices to check if $k = a_m$ for $m < \phi(\frac{1}{2}, 2^{-k-1})$ effectively, which can be done.

6.1.2 Quantitative Laws of Large Numbers in the literature

The study of large deviations in the Strong Law of Large Numbers starts with Cramér's 1938 article [28], where he determined large deviation probabilities for the sums of iid random variables up to asymptotic equivalence. Furthermore, in this work, he introduced the moment-generating function condition (the moment-generating function of the random variables is finite on an interval), which has become a standard assumption in this area.

The subsequent notable work in this direction was in 1960 by Bahadur and Ranga Rao [9], where they built on Cramér's work to calculate large deviation probabilities for the weak law of large numbers up to asymptotic equivalence (again assuming the moment generating function condition from Cramér).

Then, in 1975, Siegmund [133] (see also [41]) was able to determine $P^*_{n,\varepsilon}$ up to asymptotic equivalence, again assuming the moment generating function condition. Thus, [133] provides the first quantitative interpretation of the Strong Law of Large Numbers. Furthermore, Siegmund's bounds heavily depend on the distribution of the random variables.

Not much work has been done to study the large deviations without strong conditions, such as the moment-generating function condition. This may be because, for weaker conditions, one cannot hope to calculate these probabilities up to asymptotic equivalence. The best we can hope for are bounds on the large deviation probabilities. As discussed already, in 2018, Luzia [110] obtained distribution-independent bounds under milder assumptions on the random variables.

Work has been done to study the large deviation probabilities for sequences of random variables that are not necessarily identically distributed. In 1943, Feller was able to generalise Cramér's 1938 article to random variables that are not necessarily identically distributed; however, his assumptions were too restrictive (he assumed the random variables only took values in finite intervals) that the result was not a complete generalisation of Cramér's. Petrov [118], in 1954, was able to provide a full generalisation of Cramér's result and has been able to strengthen this result (by relaxing the moment generating function condition) a further two times, with the most recent in 2006 [120] jointly with Robinson.

We also note that the pointwise ergodic theorem can be used to show that the Strong Law of Large Numbers also holds for stationary sequences of random variables and obtaining rates for $P^*_{n,\varepsilon}$, in this case, has been of great interest. For example, Gaposhkin [44] provides an asymptotic upper bound for $P^*_{n,\varepsilon}$ (which they demonstrate is optimal) for second-order stationary sequences of random variables with finite variance, with more recent work being done by Kachurovskii on this topic, see [71, 72].

Given a nonnegative sequence of real numbers $\{a_n\}$ that converges to 0, one way to measure the speed of convergence is to find $r \geq -1$ such that

$$\sum_{n=1}^{\infty} n^r a_n < \infty.$$

This form of a rate of convergence for Strong Laws of Large Numbers was first considered by Baum and Katz in [11]. In particular, in Theorem 2 of this paper, they show:

Theorem 6.1.2. *(Baum-Katz, cf. Theorem 2 of [11]) Suppose $\{X_n\}$ are iid random variables with, $\mathbb{E}(X_0) = 0$ and $Var(X_0) < \infty$. Then, for all $\varepsilon > 0$*

$$\sum_{n=1}^{\infty} P^*_{n,\varepsilon} < \infty.$$

Baum-Katz type rates have been obtained in the Strong Law of Large Numbers for non-negative random variables where both the independence and identical distribution conditions are weakened. In 2018, Korchevsky [96] obtained a Baum-Katz type rate for the Chen-Sung Strong Law of Large Numbers [22] under stronger assumptions. This result generalised the work of Kuczmaszewska [103], in 2016, who was able to obtain rates for a Strong Law of Large Numbers result of Korchevsky in [95], under stronger assumptions. No Baum-Katz type rates have been found for the full results in [95] and [22].

Lastly, Baum-Katz type results can be used to obtain results concerning large deviation probabilities. For example, if $\{X_n\}$ are iid random variables with, $\mathbb{E}(X_0) = 0$ and $\mathrm{Var}(X_0) < \infty$ then condition (iii) of Theorem 6.3.3 with $r = 0$ implies,

$$P^*_{n,\varepsilon} = o\left(\frac{1}{n}\right).$$

This result is ineffective in the sense that it does not explicitly tell you the constant C such that $P^*_{n,\varepsilon} \leq \frac{C}{n}$, in addition, one cannot determine, a priori, that such a constant is independent of the distribution of the random variables.

6.2 The computational content of Chung's Law of Large Numbers on Banach spaces

The following generalisation of Kolmogorov's Strong Law of Large Numbers is due to Chung:

Theorem 6.2.1 (Chung's Law of Large Numbers c.f. [26])**.** *Suppose $\{X_n\}$ is a sequence of independent real-valued random variables with $\mathbb{E}(X_n) = 0$ for all $n \in \mathbb{N}$. For each $n \in \mathbb{N}$, let $\phi_n : \mathbb{R}^+ \to \mathbb{R}^+$ be a function such that*

$$\frac{\phi_n(t)}{t} \text{ and } \frac{t^2}{\phi_n(t)} \tag{6.4}$$

are nondecreasing and assume

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}(\phi_n(|X_n|))}{\phi_n(n)} < \infty.$$

Then $\frac{S_n}{n} \to 0$ almost surely.

Chung's result was generalised in [143] to type p Banach spaces giving the following result:

Theorem 6.2.2. *Suppose $\{X_n\}$ is a sequence of independent random variables taking values in a type p Banach space $\mathbb{B}$ with $\mathbb{E}(X_n) = 0$ for all $n \in \mathbb{N}$. For each $n \in \mathbb{N}$, let $\phi_n : \mathbb{R}^+ \to \mathbb{R}^+$*

be a function such that

$$\frac{\phi_n(t)}{t} \text{ and } \frac{t^p}{\phi_n(t)} \tag{6.5}$$

are nondecreasing and assume

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}(\phi_n(|X_n|))}{\phi_n(n)} < \infty.$$

Then $\frac{S_n}{n} \to 0$ *almost surely.*

In this section, we shall give a computational interpretation to the above theorem by constructing rates of uniform metastability (which will be the obvious generalisation of Definition 4.2.12 to random variables taking values in separable Banach spaces) in the conclusion in terms of suitable computational interpretations given to the assumptions in the premise of the theorem.

Throughout the remainder of this chapter, we will only be concerned with uniform metastability. Thus, whenever we discuss metastable convergence in the context of almost sure convergence, we mean uniform metastable convergence.

6.2.1 Rates for the probabilistic Kronecker's lemma

A core component in the proofs of Theorem 6.2.1 and 6.2.2 is the probabilistic analogue of Kronecker's Lemma (which we introduced in Section 5.2). By combining Theorem 5.2.12 and Theorem 5.2.2, we can obtain a finitary quantitative probabilistic Kronecker's lemma (that is, a solution to the Dialectica interpretation of the probabilistic Kronecker's lemma) for real-valued random variables. We can obtain the same result for Banach space random variables by essentially the exact same arguments as in the proof of Theorem 5.2.12, but for clarity, we choose to explicitly show all the details in the specific case of Kronecker's lemma. For the rest of this chapter, fix a (separable) Banach space $(\mathbb{B}, \|\cdot\|)$.

Theorem 6.2.3 (Finitary probabilistic Kronecker's lemma)**.** *Let* $\{Y_n\}$ *be a sequence of* $\mathbb{B}$ *valued random variables, set* $Z_n := \sum_{i=0}^{n} Y_i$*. Let* $0 < a_0 \leq a_1 \leq \dots$ *be such that* $a_n \to \infty$ *and* $f_{\{a_n\}}$ *be as in Theorem 5.2.6.*

For every $\psi : (0,1] \times (0,1] \to \mathbb{N}$*, sequence of natural numbers* $\{z_n\}$*,* $\varepsilon, \lambda \in (0,1]$ *and* $k \in \mathbb{N}$*, if* $M := \psi(\frac{\lambda}{2}, \frac{\varepsilon}{4}) = \gamma_{\frac{\lambda}{2}}(\frac{\varepsilon}{4})$ *satisfies,*

$$\mathbb{P}\left(\bigcup_{i=0}^{M} \{\|Z_i\| \geq z_M\}\right) \leq \frac{\lambda}{2}$$

and

$$\mathbb{P}\left(\max_{M \leq m \leq k} \|Z_M - Z_m\| \geq \frac{\varepsilon}{4}\right) < \frac{\lambda}{2}$$

then $N := \Psi_{\{a_n\},\{z_n\}}(\psi, \varepsilon, \lambda)$ *satisfies,*

$$\mathbb{P}\left(\max_{N\leq n\leq k}\left\|\frac{1}{a_n}\sum_{i=0}^{n} a_i Y_i\right\| \geq \varepsilon\right) < \lambda.$$

Where,

$$\Psi_{\{a_n\},\{z_n\}}(\psi, \varepsilon, \lambda) := \max\left\{Q, f_{\{a_n\}}\left(\frac{4a_Q z_Q}{\varepsilon}\right)\right\} = \Gamma_{\{a_n\}}\left(\gamma_{\frac{\lambda}{2}}, \{z_n\}, \varepsilon\right)$$

(Γ as defined in Theorem 5.2.2) and,

$$Q := \psi\left(\frac{\lambda}{2}, \frac{\varepsilon}{4}\right)$$
$$\gamma_\lambda(\varepsilon) := \psi(\lambda, \varepsilon).$$

Proof. Let $\psi, \{z_n\}, \varepsilon, \lambda, k$ satisfying the premise of the theorem be given. We have,

$$\begin{aligned}
&\mathbb{P}\left(\max_{N\leq n\leq k}\left\|\frac{1}{a_n}\sum_{i=0}^{n} a_i Y_i\right\| \geq \varepsilon\right) \\
&= \mathbb{P}\left(\left\{\max_{N\leq n\leq k}\left\|\frac{1}{a_n}\sum_{i=0}^{n} a_i Y_i\right\| \geq \varepsilon\right\} \cap \bigcup_{i=1}^{M}\{\|Z_i\| \geq z_M\}\right) \\
&+ \mathbb{P}\left(\left\{\max_{N\leq n\leq k}\left\|\frac{1}{a_n}\sum_{i=0}^{n} a_i Y_i\right\| \geq \varepsilon\right\} \cap \bigcap_{i=1}^{M}\{\|Z_i\| < z_M\}\right) \\
&\leq \frac{\lambda}{2} + \mathbb{P}\left(\left\{\max_{N\leq n\leq k}\left\|\frac{1}{a_n}\sum_{i=0}^{n} a_i Y_i\right\| \geq \varepsilon\right\} \cap \bigcap_{i=1}^{M}\{\|Z_i\| < z_M\}\right).
\end{aligned}$$

Suppose

$$\max_{N\leq n\leq k}\left\|\frac{1}{a_n}\sum_{i=0}^{n} a_i Y_i\right\| \geq \varepsilon \wedge \bigwedge_{i=1}^{M}\left(\|Z_i\| < z_M\right),$$

but for all $m \in [M, k], |Z_m - Z_M| < \frac{\varepsilon}{4}$. Theorem 5.2.2 implies that,[3]

$$\forall n \in [N, k]\left\|\frac{1}{a_n}\sum_{i=0}^{n} a_i Y_i\right\| < \varepsilon,$$

(recalling $N = \Gamma_{\{a_n\}}(\gamma_{\frac{\lambda}{2}}, \{z_n\}, \varepsilon)$), which is a contradiction. So,

$$\max_{M\leq m\leq k}\|Z_M - Z_m\| \geq \frac{\varepsilon}{4}.$$

[3]Here we assume $\varepsilon \in (0, 1]$, instead of $\mathbb{Q}^+$, but it is clear that Theorem 5.2.2 holds for such ε as we did not use any properties of the rationals in the proof.

This implies,

$$\mathbb{P}\left(\left\{\max_{N\leq n\leq k}\left\|\frac{1}{a_n}\sum_{i=0}^{n}a_iY_i\right\|\geq\varepsilon\right\}\cap\bigcap_{i=1}^{M}\{\|Z_i\|<z_M\}\right)$$
$$\leq\mathbb{P}\left(\max_{M\leq m\leq k}\|Z_M-Z_m\|\geq\frac{\varepsilon}{4}\right)<\frac{\lambda}{2}.$$

So we are done. □

Theorem 6.2.3 allows us to immediately obtain a quantitative version of the probabilistic Kronecker's lemma, where we obtain rates in the conclusion given rates in the premise (with the natural lifting of metastability in the the deterministic and stochastic case to normed spaces).

Corollary 6.2.4. *Let $\{a_n\}$, $\{Z_n\}$ be as in Theorem 6.2.3 and for each $\lambda\in(0,1]$, let $\{z_n(\lambda)\}$ be a sequence of nondecreasing natural numbers satisfying, for all $n\in\mathbb{N}$,*

$$\mathbb{P}\left(\bigcup_{i=0}^{n}\{\|Z_i\|\geq z_n(\lambda)\}\right)\leq\lambda \tag{6.6}$$

for all $n\in\mathbb{N}$.

Now suppose, $\sum_{i=0}^{n}Y_i$ converges almost surely with a rate of metastable almost sure convergence Φ. Then

$$\frac{1}{a_n}\sum_{i=0}^{n}a_iY_i$$

converges to 0 almost surely, with rate of metastable almost sure convergence

$$\kappa^P_{\Phi,\{a_n\},\{z_n\}}(\lambda,\varepsilon,K):=\max\left\{Q,f_{\{a_n\}}\left(\frac{4a_Qz_Q(\lambda/2)}{\varepsilon}\right)\right\},$$

where,

$$Q:=\Phi\left(\frac{\lambda}{2},\frac{\varepsilon}{4},H\right)$$

and

$$H:=H_{\varepsilon,\lambda,K,\{a_n\},\{z_n\}}(n):=\tilde{K}\left(\max\left\{n,f_{\{a_n\}}\left(\frac{4a_nz_n(\lambda/2)}{\varepsilon}\right)\right\}\right),$$

with $\tilde{K}(n)=n+K(n)$.

Proof. Let $\varepsilon,\lambda\in(0,1]$ and $K:\mathbb{N}\to\mathbb{N}$ be given. There exists $M\leq Q:=\Phi(\frac{\lambda}{2},\frac{\varepsilon}{4},H)$ such that

$$\mathbb{P}\left(\max_{M\leq m\leq H(M)}\|Z_M-Z_m\|\geq\frac{\varepsilon}{4}\right)\leq\mathbb{P}\left(\max_{M\leq m\leq M+H(M)}\|Z_M-Z_m\|\geq\frac{\varepsilon}{4}\right)<\frac{\lambda}{2}$$

by the definition of a rate of metastability. Let $\psi(\delta_1, \delta_2) = M$, for all $\delta_1, \delta_2 \in (0, 1]$. Then

$$N = \Psi_{\{a_n\},\{z_n(\lambda/2)\}}(\psi, \varepsilon, \lambda) = \max\left\{M, f_{\{a_n\}}\left(\frac{4a_M z_M(\lambda/2)}{\varepsilon}\right)\right\} \leq \max\left\{Q, f_{\{a_n\}}\left(\frac{4a_Q z_Q(\lambda/2)}{\varepsilon}\right)\right\}$$

and $k = H(M) = N + K(N)$ satisfies,

$$\mathbb{P}\left(\max_{N\leq n\leq N+K(N)} \|\frac{1}{a_n}\sum_{i=0}^{n} a_i Y_i\| \geq \varepsilon\right) < \lambda$$

so we are done. □

Remark 6.2.5. As in Remark 5.2.16, condition (6.6) can naturally be satisfied given moduli of almost sure finiteness for the random variables, as defined in Definition 5.2.14.

Remark 6.2.6. Lastly, note that as in the deterministic case, if Φ above is a rate of convergence, then we get a rate of almost sure convergence to 0 given by the above expression, but with

$$Q := \Phi\left(\frac{\lambda}{2}, \frac{\varepsilon}{4}\right).$$

6.2.2 Rates for Chung's Law of Large Numbers on Banach spaces

Throughout this section, assume $\mathbb{B}$ is a type p Banach space with a constant $C \geq 1$ satisfying (2.10). In this section, we shall use the quantitative probabilistic version of Kronecker's lemma, presented in Corollary 6.2.4, to obtain a computational interpretation of the generalisation of Theorem 6.2.1 to Banach spaces given in [143]. The result follows from some lemmas, the first of which is a generalisation of *Kolmogorov's inequality*.

Lemma 6.2.7 (Kolmogorov's inequality, cf. Theorem 3.2.4B of [30])**.** *Let $\{X_n\}$ be a sequence of independent random variables taking values in $\mathbb{B}$, each with expected value 0. Setting $S_n := \sum_{i=0}^{n} X_i$, we have, for all $n \in \mathbb{N}$, $\varepsilon > 0$ and $r \geq 1$,*

$$\mathbb{P}\left(\max_{0\leq i\leq n} \|S_i\| > \varepsilon\right) \leq \frac{\mathbb{E}(\|S_n\|^r)}{\varepsilon^r}.$$

We omit the proof of this result. We now need quantitative versions of [143, Theorem 2 and 2a].

Theorem 6.2.8 (Quantitative version of Theorem 2 of [143])**.** *Let $\{X_n\}$ be a sequence of independent random variables taking values in $\mathbb{B}$, each with expected value 0. Suppose $\sum_{i=0}^{n} \mathbb{E}(\phi_0(\|X_i\|))$ converges with rate of Cauchy metastability Φ, where $\phi_0(t) = t^p$ for $0 \leq t \leq 1$ and $\phi_0(t) = t$ for $t > 1$. Then S_n converges almost surely with rate of uniform metastable almost sure convergence*

$$\Delta_\Phi(\lambda, \varepsilon, K) = \Phi(\tilde{\varepsilon}, K),$$

for all $\varepsilon, \lambda \in (0,1]$ *and* $K : \mathbb{N} \to \mathbb{N}$. *Where*

$$\tilde{\varepsilon} := \min\left\{\frac{\varepsilon\lambda}{6}, \frac{\lambda\varepsilon^p}{2^{3p-1}3C}, \left(\frac{\lambda\varepsilon^p}{2^{2p-1}3}\right)^{\frac{1}{p}}\right\} = \frac{\lambda\varepsilon^p}{2^{3p-1}3C}$$

and $S_n := \sum_{i=0}^n X_i$.

Proof. For each $i \in \mathbb{N}$, let $X_i' = X_i 1_{\{\|X_i\| \leq 1\}}$ and $X_i'' = X_i 1_{\{\|X_i\| > 1\}}$. Clearly $X_i = X_i' + X_i''$ and $\{X_n'\}$ and $\{X_n''\}$ are independent random variables taking values in $\mathbb{B}$. Let $S_n' = \sum_{i=0}^n X_i'$ and $S_n'' = \sum_{i=0}^n X_i''$. Suppose $\varepsilon > 0, \lambda > 0, K : \mathbb{N} \to \mathbb{N}$ are given, we have $N \leq \Phi(\tilde{\varepsilon}, K)$ such that

$$\sum_{i=N+1}^{N+K(N)} \mathbb{E}(\phi_0(\|X_i\|)) < \tilde{\varepsilon}.$$

Now setting $K := N + K(N)$ gives,

$$\begin{aligned}
\mathbb{P}\left(\max_{N\leq n\leq K}\|S_n - S_N\| > \varepsilon\right) &\leq \mathbb{P}\left(\max_{N\leq n\leq K}\|S_n' - S_N' + S_n'' - S_N''\| > \varepsilon\right) \\
&\leq \mathbb{P}\left(\max_{N\leq n\leq K}(\|S_n' - S_N'\| + \|S_n'' - S_N''\|) > \varepsilon\right) \\
&\leq \mathbb{P}\left(\max_{N\leq n\leq K}\|S_n' - S_N'\| > \frac{\varepsilon}{2} \vee \max_{N\leq n\leq K}\|S_n'' - S_N''\| > \frac{\varepsilon}{2}\right) \\
&\leq \mathbb{P}\left(\max_{N\leq n\leq K}\|S_n' - S_N'\| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\max_{N\leq n\leq K}\|S_n'' - S_N''\| > \frac{\varepsilon}{2}\right).
\end{aligned}$$

Now by Lemma 6.2.7 and the definition of ϕ_0, we have

$$\begin{aligned}
\mathbb{P}\left(\max_{N\leq n\leq K}\|S_n'' - S_N''\| > \frac{\varepsilon}{2}\right) &\leq \frac{2\mathbb{E}\left(\|\sum_{i=N+1}^K X_i''\|\right)}{\varepsilon} \\
&\leq \frac{2\mathbb{E}\left(\sum_{i=N+1}^K \|X_i''\|\right)}{\varepsilon} \leq \frac{2\sum_{i=N+1}^K \mathbb{E}(\phi_0(\|X_i\|))}{\varepsilon}
\end{aligned}$$

and we have

$$\begin{aligned}
\mathbb{P}\left(\max_{N\leq n\leq K}\|S'_n - S'_N\| > \frac{\varepsilon}{2}\right) &\leq \frac{2^p}{\varepsilon^p}\mathbb{E}\left(\left\|\sum_{i=N+1}^{K} X'_i\right\|^p\right) \\
&\leq \frac{2^p}{\varepsilon^p}\mathbb{E}\left(\left(\left\|\sum_{i=N+1}^{K}(X'_i - \mathbb{E}(X'_i))\right\| + \left\|\sum_{i=N+1}^{K}\mathbb{E}(X'_i)\right\|\right)^p\right) \\
&\leq \frac{2^{2p-1}}{\varepsilon^p}\mathbb{E}\left(\left\|\sum_{i=N+1}^{K}(X'_i - \mathbb{E}(X'_i))\right\|^p + \left\|\sum_{i=N+1}^{K}\mathbb{E}(X'_i)\right\|^p\right) \\
&\leq \frac{2^{2p-1}}{\varepsilon^p}\left(C\sum_{i=N+1}^{K}\mathbb{E}(\|X'_i - \mathbb{E}(X'_i)\|^p) + \left\|\sum_{i=N+1}^{K}\mathbb{E}(X''_i)\right\|^p\right) \\
&\leq \frac{2^{2p-1}}{\varepsilon^p}\left(C2^{p-1}\left(\sum_{i=N+1}^{K}\mathbb{E}(\|X'_i\|^p) + \sum_{i=N+1}^{K(N)}\mathbb{E}(\|X'_i\|^p)\right) + \left(\mathbb{E}\left(\left\|\sum_{i=N+1}^{K}X''_i\right\|\right)\right)^p\right) \\
&\leq \frac{2^{2p-1}}{\varepsilon^p}\left(C2^p\left(\sum_{i=N+1}^{K}\mathbb{E}(\phi_0(\|X_i\|))\right) + \left(\sum_{i=N+1}^{K}\mathbb{E}(\phi_0(\|X_i\|))\right)^p\right) \\
&= \frac{2^{3p-1}C}{\varepsilon^p}\sum_{i=N+1}^{K}\mathbb{E}(\phi_0(\|X_i\|)) + \frac{2^{2p-1}}{\varepsilon^p}\left(\sum_{i=N+1}^{K}\mathbb{E}(\phi_0(\|X_i\|))\right)^p.
\end{aligned}$$

The first inequality follows from Lemma 6.2.7, the second inequality by the triangle inequality, and the third inequality follows from the fact that if $a, b > 0$, then

$$\left(\frac{a+b}{2}\right)^p \leq \frac{a^p + b^p}{2}. \tag{6.7}$$

The fourth inequality follows from the fact that $\mathbb{B}$ is of type p and $0 = \mathbb{E}(X_i) = \mathbb{E}(X'_i) + \mathbb{E}(X''_i)$, for all $i \in \mathbb{N}$. The fifth inequality follows from the triangle inequality and (6.7). The sixth inequality follows from the definition of ϕ_0.

Putting all of this together and using the definition of $\tilde{\varepsilon}$, we get

$$\begin{aligned}
\mathbb{P}\left(\max_{N\leq n\leq K}\|S_n - S_N\| > \varepsilon\right) \leq \frac{2\sum_{i=N+1}^{K}\mathbb{E}(\phi_0(\|X_i\|))}{\varepsilon} + \\
\frac{2^{3p-1}C}{\varepsilon^p}\sum_{i=N+1}^{K}\mathbb{E}(\phi_0(\|X_i\|)) + \frac{2^{2p-1}}{\varepsilon^p}\left(\sum_{i=N+1}^{K}\mathbb{E}(\phi_0(\|X_i\|))\right)^p < \lambda.
\end{aligned}$$

□

From the above, we immediately get:

Corollary 6.2.9 (Quantitative version of Theorem 2a of [143]). *Let $\{X_n\}$ be a sequence of independent random variables taking values in $\mathbb{B}$, each with expected value 0. Let $0 < a_0 \leq a_1 \leq \ldots$ be such that $a_n \to \infty$. Further suppose we have a sequence of functions $\{\phi_n : \mathbb{R}^+ \to \mathbb{R}^+\}$ such that,*

$$\frac{\phi_n(t)}{t} \text{ and } \frac{t^p}{\phi_n(t)} \text{ are nondecreasing for all } n \in \mathbb{N}.$$

If we have,

$$\sum_{k=0}^{\infty} \frac{\mathbb{E}(\phi_k(\|X_k\|))}{\phi_k(a_k)} < \infty$$

and converges with rate of Cauchy metastability Φ,

then

$$\sum_{k=0}^{n} \frac{X_k}{a_k}$$

converges with rate of Cauchy metastability Δ_Φ, where Δ_Φ is defined as in Theorem 6.2.8.

Proof. For each $n \in \mathbb{N}$, let

$$\Gamma_n(t) = \frac{\phi_n(a_n t)}{\phi_n(a_n)}.$$

It is easy to see that for every function Γ with $\Gamma(1) = 1$ and both

$$\frac{\Gamma(t)}{t} \text{ and } \frac{t^p}{\Gamma(t)}$$

nondecreasing, we have $\Gamma(t) \geq \phi_0(t)$ for all $t \geq 0$. One can easily check that for each $n \in \mathbb{N}$, Γ_n satisfies this property. Thus, a rate of convergence for

$$\sum_{k=0}^{n} \frac{\mathbb{E}(\phi_k(\|X_k\|))}{\phi_k(a_k)} = \sum_{k=0}^{n} \mathbb{E}\left(\Gamma_k\left(\left\|\frac{X_k}{a_k}\right\|\right)\right) =$$

will be a rate of convergence for

$$\sum_{k=0}^{n} \mathbb{E}\left(\phi_0\left(\left\|\frac{X_k}{a_k}\right\|\right)\right)$$

and the result follows from Theorem 6.2.8. □

We can now prove a quantitative version of the main result of [143].

Theorem 6.2.10. *Let $\{X_n\}$ be a sequence of independent random variables taking values in $\mathbb{B}$, each with expected value 0. Let $0 < a_0 \leq a_1 \leq \ldots$ be such that $a_n \to \infty$ and $f_{\{a_n\}}(x) := \min\{n \in \mathbb{N} : a_n \geq x\}$. Further suppose we have a sequence of functions $\{\phi_n : \mathbb{R}^+ \to \mathbb{R}^+\}$ such*

that,

$$\frac{\phi_n(t)}{t} \text{ and } \frac{t^p}{\phi_n(t)} \text{ are nondecreasing for all } n \in \mathbb{N}. \tag{6.8}$$

For each $\lambda \in \mathbb{Q}^+$, let $\{z_n(\lambda)\}$ be a sequence of natural numbers satisfying,

$$\mathbb{P}\left(\bigcup_{i=0}^{n}\left\{\left\|\sum_{k=0}^{i}\frac{X_k}{a_k}\right\| \geq z_n(\lambda)\right\}\right) \leq \lambda$$

for all $n \in \mathbb{N}$. Suppose

$$\sum_{k=0}^{\infty}\frac{\mathbb{E}(\phi_k(\|X_k\|))}{\phi_k(a_k)} < \infty$$

and converges with rate of Cauchy metastability Φ. Then $\frac{S_n}{n}$ converges to 0 almost surely with a rate of metastable almost sure convergence

$$\kappa^P_{\Delta_\Phi,\{a_n\},\{z_n\}}$$

Proof.

$$\sum_{k=0}^{n}\frac{\mathbb{E}(\phi_k(\|X_k\|))}{\phi_k(a_k)}$$

converges with rate of Cauchy metastability Φ implies

$$\sum_{k=0}^{n}\frac{X_k}{a_k}$$

converges with rate of Cauchy metastability Δ_Φ, by Theorem 6.2.9. So, the result follows from Corollary 6.2.4. □

An instance of our general result above is the following:

Theorem 6.2.11. *Let $\{X_n\}$ is a sequence of independent real-valued random variables and $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ be such that $\mathbb{E}(X_n) = 0$, $\delta \geq \mathbb{E}\left(|X_n|\phi(|X_n|)\right)$ and $\tau \geq \mathbb{E}\left(|X_n|\right)$ for some $\delta, \tau > 0$ and for all $n \in \mathbb{N}$. Further, assume,*

$$\phi(t) \text{ and } \frac{t}{\phi(t)}$$

are non-decreasing. Lastly, suppose,

$$\sum_{n=1}^{\infty}\frac{1}{n\phi(n)} < \infty \tag{6.9}$$

and assume that its partial sums converge with a rate of Cauchy convergence Φ.

$$\Lambda(\lambda,\varepsilon) := \max\left\{\Phi\left(\frac{\lambda\varepsilon^2}{2^{10}3\delta}\right), \left\lceil \frac{16\tau}{\varepsilon}\Phi^2\left(\frac{\lambda\varepsilon^2}{2^{10}3\delta}\right)\log\left(e\Phi\left(\frac{\lambda\varepsilon^2}{2^{10}3\delta}\right)\right)\right\rceil\right\},$$

is a rate of almost sure convergence for $S_n/n \to 0$.

Proof. We apply Theorem 6.2.10 with $\mathbb{B} = \mathbb{R}$ (taking $p = 2$ and $C = 1$). Setting $a_n = n+1$ (so $f_{\{n\}}(x) = \lceil x - 1 \rceil$) and $\phi_n(t) = t\phi(t)$ implies (6.8) is satisfied. Furthermore, a rate of Cauchy convergence for the partial sums of

$$\sum_{k=0}^{\infty} \frac{\mathbb{E}(\phi_k(\|X_k\|))}{\phi_k(a_k)} < \infty$$

is given by,

$$\tilde{\Phi}(\varepsilon) := \Phi\left(\frac{\varepsilon}{\delta}\right).$$

By the triangle inequality, we have,

$$\mathbb{E}\left(\left\|\sum_{i=1}^{n}\frac{X_i}{i}\right\|\right) \leq \sum_{i=1}^{n}\frac{\mathbb{E}(\|X_i\|)}{i} \leq \tau\log(en).$$

This implies (by Remark 5.2.16 and Example 5.2.15),

$$\mathbb{P}\left(\bigcup_{i=0}^{n}\left\{\left\|\sum_{k=1}^{i}\frac{X_k}{k}\right\| \geq z_n(\lambda)\right\}\right) \leq \lambda$$

where

$$z_n(\lambda) := \frac{n\tau\log(en)}{\lambda}.$$

Therefore, by Remark 6.2.6, a rate of almost sure convergence for $S_n/n \to 0$ is given by,

$$\kappa^P_{\Delta_{\tilde{\Phi}},\{n\},\{z_n\}}(\lambda,\varepsilon,K) = \max\left\{\Phi\left(\frac{\lambda\varepsilon^2}{2^{10}3\delta}\right), \left\lceil\frac{4}{\varepsilon}\Phi\left(\frac{\lambda\varepsilon^2}{2^{10}3\delta}\right)z_Q(\lambda/2)\right\rceil\right\}$$

where,

$$Q := \Delta_{\tilde{\Phi}}\left(\frac{\lambda}{2},\frac{\varepsilon}{4}\right) = \Phi\left(\frac{\lambda\varepsilon^2}{2^{10}3\delta}\right)$$

and the result follows. □

The Hájek and Rényi theorem [57] states that if $\{X_n\}$ is a sequence of independent random variables, each with expected value 0, then we have

$$\mathbb{P}\left(\max_{n\leq k\leq m}\left|\frac{S_k}{k}\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2}\left(\frac{1}{n^2}\sum_{i=0}^{n}Var(X_k) + \sum_{k=n+1}^{m}\frac{Var(X_k)}{k^2}\right) \tag{6.10}$$

for all $\varepsilon > 0$ and $n < m$. From this, one easily obtains that

$$P^*_{n,\varepsilon} \leq \frac{2\sigma^2}{n\varepsilon^2}$$

if $\{X_n\}$ is assumed to also have respective variances bounded by σ^2. This corresponds to the case $\phi(t) = t$ in Theorem 6.2.11, and one quickly sees that our rates are not optimal in this case.

However, Theorem 6.2.11 can be used to obtain the convergence speeds for a wider class of random variables than is possible with the Hájek-Rényi inequality. For example, one obtains rates for sequences of independent random variables, $\{X_n\}$, each with expected value 0 and satisfying that $\mathbb{E}\left(|X_n||\log(|X_n|)|^{1+\kappa}\right)$ is uniformly bounded for some $\kappa > 0$. Such a moment condition is known to be optimal in the context of Strong Laws of Large Numbers for random variables that are not assumed to be identically distributed by [26, Theorem 2], that is, if $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ is any function such that (6.9) does not hold, there exists a sequence of independent random variables $\{X_n\}$, each with expected value 0 and $\mathbb{E}\left(|X_n|\phi(|X_n|)\right)$ uniformly bounded but where S_n/n diverges with probability 1.

6.3 Further quantitative Strong Laws of Large Numbers

In [40], Etemadi demonstrates that $\frac{S_n}{n} \to 0$ almost surely if we only assume the random variables are pairwise iid. Furthermore, Etemadi's proof is rather elementary compared to Kolmogorov's original proof of this result for iid random variables.

In [29] Csörgő et al. demonstrate that Kolmogorov's Strong Law of Large Numbers does not hold if we weaken the independence condition, in Theorem 6.0.1, to even pairwise independence (see [29, Theorem 3]). They instead obtain the following:

Theorem 6.3.1 (cf. Theorem 1 of [29])**.** *Suppose $\{X_n\}$ is a sequence of pairwise independent random variables, each with expected value 0, satisfying (6.1) and*

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}(|X_k|) = O(1). \tag{6.11}$$

Then $\frac{S_n}{n} \to 0$ almost surely.

This section will study the quantitative content of the Strong Law of Large Numbers when the iid assumption is weakened by calculating explicit rates of convergences for $P^*_{n,\varepsilon}$. We will prove a general technical theorem (in Section 6.3.1) whose quantitative content captures the key combinatorial idea in the proof of Theorem 6.3.1. Our general theorem will allow us to show:

Theorem 6.3.2. *Suppose $\{X_n\}$ is a sequence of pairwise independent random variables satisfying, $\mathbb{E}(X_n) = 0$, $\mathbb{E}(|X_n|) \leq \tau$ and $\mathrm{Var}(X_n) \leq \sigma^2$, for all $n \in \mathbb{N}$ and some $\tau, \sigma > 0$. There exists a universal constant $\kappa \leq 1536$ such that for all $0 < \varepsilon \leq \tau$,*

$$P^*_{n,\varepsilon} \leq \frac{\kappa\sigma^2\tau}{n\varepsilon^3}.$$

The above is an improvement of the asymptotic upper bound given by Luzia [110] (which was the best known, to the author's knowledge), who showed (with notation as in Theorem 6.3.2) that for all $\beta > 1$ and $0 < \varepsilon \leq \tau$, there exists $N(\beta, \varepsilon, \tau)$ such that, for all $n \geq N(\beta, \varepsilon, \tau)$,

$$P^*_{n,\varepsilon} \leq \frac{\sigma^2}{n\varepsilon^2}(C_\beta + D_\beta \log(n)^{\beta-1}),$$

for some $C_\beta, D_\beta > 0$ depending only on β. Thus, if we fix ε, σ, and τ the above tells us that for each $\beta > 1$,

$$P^*_{n,\varepsilon} = O\left(\frac{\log(n)^{\beta-1}}{n}\right)$$

as $n \to \infty$, for such a class of random variables, whereas the bound in Theorem 6.3.2 yields:

$$P^*_{n,\varepsilon} = O\left(\frac{1}{n}\right).$$

Observe we can take $\tau = \sigma$ by Jensen's inequality, so Theorem 6.3.2 in particular yields

$$P^*_{n,\varepsilon} \leq \frac{\kappa\sigma^3}{n\varepsilon^3}.$$

The above bound bears a resemblance to the bound one obtains in the case where the random variables are assumed to be independent since the Hájek and Rényi inequality (6.10) implies that if $\{X_n\}$ is assumed to be an independent sequence of random variables with a common bound on their variance σ^2,

$$P^*_{n,\varepsilon} \leq \frac{2\sigma^2}{n\varepsilon^2}. \tag{6.12}$$

However, the Hájek and Rényi inequality generalises Kolmogorov's inequality, which is known to fail for pairwise independent random variables; therefore, more work is needed to obtain a bound in this case.

In addition, through a simple construction, we demonstrate that $O(1/n)$ is the best power bound for $P^*_{n,\varepsilon}$ in the case that $\{X_n\}$ is iid with finite variance. That is, for each $\delta > 0$, we construct a sequence of random variables with finite variance, satisfying

$$\mathbb{P}\left(\sup_{m \geq n}\left|\frac{S_m}{m}\right| > \varepsilon\right) \geq \frac{\omega}{n^{1+\delta}}$$

for some $\omega > 0$ depending only on δ, for all $0 < \varepsilon \leq 1$. All of these results are in Section 6.3.2.

The following result is mainly attributed to Baum and Katz [11] as well as Chow [25]:

Theorem 6.3.3. *Let $\{X_n\}$ be a sequence of iid random variables satisfying $\mathbb{E}(X_0) = 0$ and let $r \geq -1$. Then for all $\varepsilon > 0$, the following are equivalent:*

(i) $\mathbb{E}(|X_0|^{r+2}) < \infty$,

(ii) $\sum_{n=1}^{\infty} n^r \mathbb{P}\left(\left|\frac{1}{n} S_n\right| > \varepsilon\right) < \infty$,

(iii) $\sum_{n=1}^{\infty} n^r \mathbb{P}\left(\sup_{m \geq n} \left|\frac{1}{m} S_m\right| > \varepsilon\right) < \infty$,

(iv) $\sum_{n=1}^{\infty} n^r \mathbb{P}\left(\max_{1 \leq m \leq n} |S_m| > n\varepsilon\right) < \infty$.

To prove this result, independence is crucial. Work has been done to extend this result to the case where the random variables are pairwise independent. It is clear that (iii) and (iv) both imply (ii) in the non-independent case. However, as noted in [10], it is possible that (iv) is strictly stronger than (ii) in the non-independent case, and work has been done in establishing the convergence of (iv) in the case where the random variables are pairwise iid. Many authors have established the following theorem, but the result goes back to Rio [127]:

Theorem 6.3.4. *Suppose $\{X_n\}$ are pairwise independent, identically distributed random variables with $\mathbb{E}(X_0) = 0$. For all $-1 \leq r < 0$: $\mathbb{E}(|X_0|^{2+r}) < \infty$ iff*

$$\sum_{n=1}^{\infty} n^r \mathbb{P}\left(\max_{1 \leq m \leq n} |S_m| > n\varepsilon\right) < \infty$$

for all $\varepsilon > 0$.

There does not appear to be any results in the literature for the convergence of the sum (iii), assuming the random variables are pairwise independent. However, a simple application of Theorem 6.3.2 gives the following:

Corollary 6.3.5. *Suppose $\{X_n\}$ are pairwise independent, identically distributed random variables with $\mathbb{E}(X_0) = 0$ and $\mathrm{Var}(X_0) < \infty$. Then, for all $\varepsilon > 0$ and $r < 0$:*

$$\sum_{n=1}^{\infty} n^r P_{n,\varepsilon}^* < \infty.$$

Proof. This result simply follows from the fact that $P_{n,\varepsilon}^* = O\left(\frac{1}{n}\right)$. □

Furthermore, it appears to be open whether it is the case that condition (*iii*) in Theorem 6.3.3 holds in the case $r = 0$ and if the random variables are only assumed to be pairwise independent, which is the case for iid random variables, by Theorem 6.3.3.

In [40], Etemadi's novel insight in demonstrating that $\frac{S_n}{n} \to 0$ almost surely for pairwise iid random variables was that one could first assume that the random variables were nonnegative, in which case one can take advantage of the monotonicity of the partial sums. The general case is then obtained by using the decomposition of a random variable into its positive and negative parts (that is, writing a random variable, X, as $X = X^+ - X^-$ where $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$). Due to this insight, there has been a lot of interest in studying when $\frac{S_n}{n} \to 0$, almost surely, for nonnegative random variables that are not assumed to be iid, as e.g. in Petrov [119], which was later generalised by Korchevsky et al. [97] and further generalised again by Korchevsky in [95]. In addition, Chandra et al. [20] generalised Theorem 6.3.1, with Chen and Sung [22] later producing a result which unified [20] and [95], as well as generalising results from [17, 66, 74, 130]. The proofs of all the results Chen and Sung generalised are adaptations of the proof of Theorem 6.3.1, and they established the following sufficient condition, which encompasses all the results mentioned:

Theorem 6.3.6 (cf. Theorem 2.1 of [22])**.** *Let $\{X_n\}$ be a sequence of nonnegative random variables with finite pth moment (for some fixed $p \geq 1$) and respective expected values $\{\mu_n\}$. Let $S_n := \sum_{k=1}^n X_k$ and $z_n := \sum_{i=1}^n \mu_i$. Suppose that*

$$\frac{z_n}{n} = O(1)$$

and that there exists a sequence of nonnegative real numbers $\{\gamma_n\}$ satisfying

- $\mathbb{E}(|S_n - z_n|^p) \leq \sum_{k=1}^n \gamma_k$,
- $\sum_{n=1}^\infty \frac{\gamma_n}{n^p} < \infty$.

Then

$$\frac{S_n}{n} - \frac{z_n}{n} \to 0$$

almost surely.

In this section, we will also produce a fully quantitative version of Theorem 6.3.6.

Theorem 6.3.7. *Let $\{X_n\}$ be a sequence of nonnegative random variables with finite pth moment (for some fixed $p \geq 1$) and respective expected values $\{\mu_n\}$. Let $S_n := \sum_{k=1}^n X_k$ and $z_n := \sum_{i=1}^n \mu_i$. Suppose there exists a sequence of nonnegative real numbers $\{\gamma_n\}$ satisfying*

$$\mathbb{E}(|S_n - z_n|^p) \leq \sum_{k=1}^n \gamma_k$$

and

$$\sum_{m=1}^{\infty} \frac{\gamma_m}{m^p} \leq \Gamma \tag{6.13}$$

for some $\Gamma \geq 1$*, with the partial sums of the above series converging to their limit with a strictly decreasing rate of convergence* Ψ*. Furthermore, assume for all* $n \in \mathbb{N}$*,*

$$\frac{z_n}{n} \leq W,$$

for some $W \geq 1$*. Then for all* $0 < \varepsilon \leq 1, \lambda > 0$ *and all*

$$n \geq A_p \left(\frac{W\Gamma}{\lambda \varepsilon^{p+1}} \right)^{\frac{1}{p}} \Psi \left(\frac{B_p \lambda \varepsilon^{p+1}}{W} \right),$$

it holds that

$$\mathbb{P}\left(\sup_{m \geq n} \left| \frac{S_m}{m} - \frac{z_m}{m} \right| > \varepsilon \right) \leq \lambda.$$

Here, A_p *and* B_p *are constants that only depend on* p*.*

From Theorem 6.3.7, one can obtain quantitative versions of many of the Strong Laws of Large Numbers discussed above. For example, we can easily obtain a quantitative version of Theorem 6.3.1:

Theorem 6.3.8. *Suppose* $\{X_n\}$ *is a sequence of pairwise independent random variables, each with expected value* 0 *and finite variance. Let* $S_n := \sum_{k=1}^{n} X_k$ *and* $z_n := \sum_{i=1}^{n} \mathbb{E}(|X_i|)$*. Further, assume*

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(X_n)}{n^2} \leq \Gamma \tag{6.14}$$

for some $\Gamma \geq 1$ *and that the partial sums of the above series converge to their limit with a strictly decreasing rate of convergence* Ψ*. Furthermore, assume for all* $n \in \mathbb{N}$*,*

$$\frac{z_n}{n} \leq W,$$

for some $W \geq 1$*. For all* $0 < \varepsilon \leq 1, \lambda > 0$ *and all*

$$n \geq A \left(\frac{W\Gamma}{\lambda \varepsilon^3} \right)^{\frac{1}{2}} \Psi \left(\frac{B\lambda \varepsilon^3}{W} \right),$$

it holds that

$$P^*_{n,\varepsilon} = \mathbb{P}\left(\sup_{m \geq n} \left| \frac{S_m}{m} \right| > \varepsilon \right) \leq \lambda.$$

Here, A *and* B *are universal constants.*

All of these results are in Section 6.3.3.

6.3.1 A general theorem

We shall state and prove the general quantitative theorem we alluded to earlier. This theorem will be a quantitative version of (a generalisation of) a critical step in proving [29, Theorem 1], which is a result that has been modified many times to obtain various Strong Laws of Large Numbers.

For this, we now first introduce the following definitions that are mostly as presented in [29]: Let $\{X_n\}$ be a sequence of nonnegative random variables with respective expected values $\{\mu_n\}$. Let $S_n := \sum_{k=1}^{n} X_k$, $z_n := \sum_{i=1}^{n} \mu_i$ and suppose there exists $W > 0$ such that

$$\frac{z_n}{n} \leq W$$

for all $n \in \mathbb{N}$. Further, we make use of the following definitions:

- For each $\delta > 0$, let $L_\delta := \left\lfloor \frac{W}{\delta} \right\rfloor$.

- For each $\delta > 0$, $\alpha > 1$ and natural numbers m and $0 \leq s \leq L_\delta$, let

$$C_{\alpha,s,\delta,m} := \left\{ \alpha^m \leq n < \alpha^{m+1} \mid \frac{z_n}{n} \in [s\delta, (s+1)\delta) \right\}.$$

- Let $k_s^-(m) := \min C_{\alpha,s,\delta,m}$ and $k_s^+(m) := \max C_{\alpha,s,\delta,m}$ if $C_{\alpha,s,\delta,m}$ is nonempty.

- Let $k_s^-(m) = k_s^+(m) := \lfloor \alpha^m \rfloor$ if $C_{\alpha,s,\delta,m}$ is empty.

One should note that $k_s^+(m)$ and $k_s^-(m)$ depend on δ, α but (following the convention of [29]) we hid this dependence to make the notation less cumbersome. We shall also adopt the convention (used in [29]) that $k_s^\pm(m)$ being used in a relationship (an equation, an inequality, a limit, etc.) is short-hand for that relationship holding for both $k_s^+(m)$ and $k_s^-(m)$.

Our general theorem is now the following:

Theorem 6.3.9. *For all $\varepsilon, \delta > 0$, $\alpha > 1$ and $0 \leq s \leq L_\delta$, if*

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left| \frac{S_{k_s^\pm(n)}}{k_s^\pm(n)} - \frac{z_{k_s^\pm(n)}}{k_s^\pm(n)} \right| > \varepsilon \right) < \infty, \tag{6.15}$$

then[4]

$$\frac{S_n}{n} - \frac{z_n}{n} \to 0$$

[4]Recall that the use of the $\pm$ notation means we are actually assuming the convergence of two sums in the premise.

almost surely.

Proof. The Borel-Cantelli Lemma and (6.15) implies that for all $\varepsilon, \delta > 0$, $\alpha > 1$ and all $0 \leq s \leq L_\delta$:

$$\frac{1}{k_s^{\pm}(n)} S_{k_s^{\pm}(n)} - \frac{1}{k_s^{\pm}(n)} z_{k_s^{\pm}(n)} \to 0 \tag{6.16}$$

almost surely. For all $m \in \mathbb{N}$, we can take a natural number $0 \leq s \leq L_\delta$ such that

$$\frac{1}{m} z_m \in [s\delta, (s+1)\delta) \tag{6.17}$$

since $z_n/n \leq W$ and $L_\delta = \left\lfloor \frac{W}{\delta} \right\rfloor$. Thus, if we take $p \in \mathbb{N}$ such that $\alpha^p \leq m < \alpha^{p+1}$, then $m \in C_{\alpha,s,\delta,p}$ by definition, so $C_{\alpha,s,\delta,p}$ is non-empty. Therefore, $k_s^-(p) \leq m \leq k_s^+(p)$ and, since $k_s^{\pm}(p) \in C_{\alpha,s,\delta,p}$, we have

$$\frac{1}{k_s^{\pm}(p)} z_{k_s^{\pm}(p)} \in [s\delta, (s+1)\delta)$$

which implies, by (6.17), that

$$\left| \frac{1}{m} z_m - \frac{1}{k_s^{\pm}(p)} z_{k_s^{\pm}(p)} \right| \leq \delta. \tag{6.18}$$

Now we have the following chain of inequalities,

$$\begin{aligned}
&-\delta - \left(1 - \frac{1}{\alpha}\right) W + \frac{1}{\alpha} \frac{1}{k_s^-(p)} \left(S_{k_s^-(p)} - z_{k_s^-(p)} \right) \\
&\quad \leq -\delta - \left(1 - \frac{1}{\alpha}\right) \frac{1}{k_s^-(p)} z_{k_s^-(p)} + \frac{1}{\alpha} \frac{1}{k_s^-(p)} \left(S_{k_s^-(p)} - z_{k_s^-(p)} \right) \\
&\quad \leq \frac{1}{m} S_{k_s^-(p)} - \frac{1}{m} z_m \\
&\quad \leq \frac{1}{m} (S_m - z_m) \\
&\quad \leq \frac{1}{m} S_{k_s^+(p)} - \frac{1}{k_s^+(p)} z_{k_s^+(p)} + \delta \\
&\quad \leq \frac{\alpha}{k_s^+(p)} \left(S_{k_s^+(p)} - z_{k_s^+(p)} \right) + (\alpha - 1) W + \delta.
\end{aligned} \tag{6.19}$$

Here, the first inequality follows since

$$\frac{1}{k_s^-(p)} z_{k_s^-(p)} < W.$$

The second inequality follows from expanding brackets, using (6.18) and the fact that $m \leq \alpha k_s^-(p)$ (since $m \in C_{\alpha,s,\delta,p}$, so by definition, $m < \alpha^{p+1}$ and $k_s^-(p) \in C_{\alpha,s,\delta,p}$, and so $\alpha^p \leq k_s^-(p)$). The third inequality follows from the fact that $\{S_n\}$ is monotone (since $\{X_n\}$ is nonnegative) and $k_s^-(p) \leq m$. The remaining inequalities are justified using similar reasoning to the above (see also [29]).

Thus, by (6.16) and the fact that $p \to \infty$ as $m \to \infty$, we have

$$-\delta - \left(1 - \frac{1}{\alpha}\right) W \leq \liminf_{n\to\infty} \frac{1}{m}(S_m - z_m) \leq \limsup_{n\to\infty} \frac{1}{m}(S_m - z_m) \leq (\alpha - 1)W + \delta$$

almost surely. So, taking $\delta \to 0$ and $\alpha \to 1$ gives our result. □

Remark 6.3.10. $\{X_n\}$ (not assumed to be nonnegative) is said to *converge completely* to 0 if

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > \varepsilon) < \infty$$

for all $\varepsilon > 0$. Hsu and Robbins first introduced this notion of convergence in [64], where they demonstrated that if $\{X_n\}$ were iid random variables with finite variance (again, not assumed to be nonnegative), then

$$\frac{S_n}{n} - \mathbb{E}(X_0)$$

converges to 0 completely. Furthermore, complete convergence implies almost sure convergence by the Borel-Cantelli Lemma, so Theorem 6.3.9 says that if specifically chosen sub-sequences of

$$\frac{S_n}{n} - \frac{z_n}{n}$$

converge completely to 0, then

$$\frac{S_n}{n} - \frac{z_n}{n}$$

converges to 0 almost surely.

Remark 6.3.11. To prove [29, Theorem 1], it is shown that

$$\sum_{n=1}^{\infty} \mathbb{E}\left(\left(\frac{S_{k_s^{\pm}(n)}}{k_s^{\pm}(n)} - \frac{z_{k_s^{\pm}(n)}}{k_s^{\pm}(n)}\right)^2\right) < \infty. \tag{6.20}$$

(6.15) in Theorem 6.3.9 follows from this by Chebyshev's inequality, so the result in [29] follows by our theorem. Therefore, Theorem 6.3.9 generalises the key step in proving [29, Theorem 1].

We now give a quantitative version of Theorem 6.3.9:

Theorem 6.3.12. *Suppose for each $\varepsilon, \delta > 0$, $\alpha > 1$ and $0 \leq s \leq L_\delta$:*

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k_s^{\pm}(n)}}{k_s^{\pm}(n)} - \frac{z_{k_s^{\pm}(n)}}{k_s^{\pm}(n)}\right| > \varepsilon\right) < \infty. \tag{6.21}$$

Furthermore, suppose that the partial sums of both sums converge to their respective limits with a rate of convergence $\Lambda_{\varepsilon,\delta,\alpha} : \mathbb{R} \to \mathbb{R}$, independent of s.[5]

[5]We can always obtain a rate independent of s by taking the maximum value of all such rates that depend

More explicitly, for each $\varepsilon, \delta > 0$, $\alpha > 1$, $0 \leq s \leq L_\delta$, $\lambda > 0$ and $p \geq \Lambda_{\varepsilon,\delta,\alpha}$, we have,

$$\sum_{n=p+1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k_s^-(n)}}{k_s^-(n)} - \frac{z_{k_s^-(n)}}{k_s^-(n)}\right| > \varepsilon\right) \leq \lambda \text{ and } \sum_{n=p+1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k_s^+(n)}}{k_s^+(n)} - \frac{z_{k_s^+(n)}}{k_s^+(n)}\right| > \varepsilon\right) \leq \lambda.$$

Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{m \geq n}\left|\frac{S_m}{m} - \frac{z_m}{m}\right| > \varepsilon\right) \to 0$$

with a rate of convergence given by

$$\Phi_{\varepsilon,\Lambda}(\lambda) := \alpha^{\Pi_\varepsilon(\lambda)},$$

where

$$\Pi_\varepsilon(\lambda) := \Lambda_{\frac{\varepsilon}{3\alpha},\frac{\varepsilon}{3},\alpha}\left(\frac{\lambda}{2}\right) + 1 \text{ and } \alpha := 1 + \frac{\varepsilon}{3W}.$$

Proof. First we observe that, for all $\delta, \lambda, \varepsilon > 0$, $\alpha > 1$, natural numbers $0 \leq s \leq L_\delta$ and $p \geq \Lambda_{\varepsilon,\delta,\alpha}(\lambda) + 1$:

$$\begin{aligned}\mathbb{P}\left(\sup_{q \geq p}\left|\frac{S_{k_s^\pm(q)}}{k_s^\pm(q)} - \frac{z_{k_s^\pm(q)}}{k_s^\pm(q)}\right| > \varepsilon\right) &= \mathbb{P}\left(\bigcup_{q=p}^{\infty}\left(\left|\frac{S_{k_s^\pm(q)}}{k_s^\pm(q)} - \frac{z_{k_s^\pm(q)}}{k_s^\pm(q)}\right| > \varepsilon\right)\right) \\ &\leq \sum_{q=p}^{\infty} \mathbb{P}\left(\left|\frac{S_{k_s^\pm(q)}}{k_s^\pm(q)} - \frac{z_{k_s^\pm(q)}}{k_s^\pm(q)}\right| > \varepsilon\right) \leq \lambda.\end{aligned}$$

Here, the last inequality follows from the fact that $p - 1 \geq \Lambda_{\varepsilon,\delta,\alpha}(\lambda)$ (and that Λ is a rate of convergence).[6]

Now, fix $\varepsilon, \lambda > 0$ and

$$n \geq \Phi_{\varepsilon,\Lambda}(\lambda) = \alpha^{\Lambda_{\frac{\varepsilon}{3\alpha},\frac{\varepsilon}{3},\alpha}(\frac{\lambda}{2})+1}.$$

We must show,

$$\mathbb{P}\left(\sup_{m \geq n}\left|\frac{S_m}{m} - \frac{z_m}{m}\right| > \varepsilon\right) \leq \lambda.$$

Set $\delta = \frac{\varepsilon}{3}$ and observe that having $\alpha = 1 + \frac{\varepsilon}{3W}$ ensures that

$$-\frac{\varepsilon}{3} \leq -(1 - \frac{1}{\alpha})W \text{ and } (\alpha - 1)W = \frac{\varepsilon}{3}. \tag{6.22}$$

on s, as s can only take the value of finitely many natural numbers. Furthermore, if both sums have different rates, we can obtain one that works for both by taking the maximum of the two rates.

[6]The above step can be seen as applying the computational interpretation of the Borel-Cantelli lemma from [3].

Take $p \in \mathbb{N}$ such that $\alpha^p \leq n < \alpha^{p+1}$. Then we have

$$\alpha^{\Lambda_{\frac{\varepsilon}{3\alpha},\frac{\varepsilon}{3},\alpha}(\frac{\lambda}{2})+1} \leq n < \alpha^{p+1}$$

which implies

$$p \geq \Lambda_{\frac{\varepsilon}{3\alpha},\frac{\varepsilon}{3},\alpha}\left(\frac{\lambda}{2}\right) + 1.$$

Thus, by the very first step of the proof, we have, for each $0 \leq r \leq L_\delta$,

$$\mathbb{P}\left(\sup_{q\geq p}\left|\frac{1}{k_r^{\pm}(q)}S_{k_r^{\pm}(q)} - \frac{1}{k_r^{\pm}(q)}z_{k_r^{\pm}(q)}\right| > \frac{\varepsilon}{3\alpha}\right) \leq \frac{\lambda}{2}. \tag{6.23}$$

Thus, it suffices to show that there exists $0 \leq r \leq L_\delta$ such that

$$\sup_{m\geq n}\left|\frac{S_m}{m} - \frac{z_m}{m}\right| > \varepsilon$$

implies that

$$\sup_{q\geq p}\left|\frac{1}{k_r^{-}(q)}S_{k_r^{-}(q)} - \frac{1}{k_r^{-}(q)}z_{k_r^{-}(q)}\right| > \frac{\varepsilon}{3\alpha}$$

or that

$$\sup_{q\geq p}\left|\frac{1}{k_r^{+}(q)}S_{k_r^{+}(q)} - \frac{1}{k_r^{+}(q)}z_{k_r^{+}(q)}\right| > \frac{\varepsilon}{3\alpha},$$

as then we would have, for such an r,

$$\begin{aligned}\mathbb{P}\left(\sup_{m\geq n}\left|\frac{S_m}{m} - \frac{z_m}{m}\right| > \varepsilon\right) &\leq \mathbb{P}\left(\sup_{q\geq p}\left|\frac{1}{k_r^{-}(q)}S_{k_r^{-}(q)} - \frac{1}{k_r^{-}(q)}z_{k_r^{-}(q)}\right| > \frac{\varepsilon}{3\alpha}\right)\\ &\quad + \mathbb{P}\left(\sup_{q\geq p}\left|\frac{1}{k_r^{+}(q)}S_{k_r^{+}(q)} - \frac{1}{k_r^{+}(q)}z_{k_r^{+}(q)}\right| > \frac{\varepsilon}{3\alpha}\right) \leq \lambda,\end{aligned}$$

which is what we are required to show (with the final inequality following from (6.23)).

Suppose, for contradiction, that the above was not the case. Then, for all $0 \leq r \leq L_\delta$, we have

$$\sup_{m\geq n}\left|\frac{1}{m}S_m - \frac{1}{m}z_m\right| > \varepsilon$$

and

$$\left|\frac{1}{k_r^{\pm}(q)}S_{k_r^{\pm}(q)} - \frac{1}{k_r^{\pm}(q)}z_{k_r^{\pm}(q)}\right| \leq \frac{\varepsilon}{3\alpha} \tag{6.24}$$

for all $q \geq p$. Take $m \geq n$ such that

$$\left|\frac{1}{m}S_m - \frac{1}{m}z_m\right| > \varepsilon. \tag{6.25}$$

We now use arguments similar to the proof of Theorem 6.3.9. We can find $0 \leq r \leq L_\delta$ such that

$$\frac{1}{m} z_m \in [r\delta, (r+1)\delta],$$

so, taking $q \in \mathbb{N}$ such that $\alpha^q \leq m < \alpha^{q+1}$, ensures that $m \in C_{\alpha,r,\delta,q}$. Furthermore, as $m \geq n$, we have $q \geq p$.

Now, since $k_r^{\pm}(q) \in C_{\alpha,r,\delta,q}$, we have

$$\frac{1}{k_r^{\pm}(q)} z_{k_r^{\pm}(q)} \in [r\delta, (r+1)\delta)$$

which implies

$$\left| \frac{1}{m} z_m - \frac{1}{k_r^{\pm}(q)} z_{k_r^{\pm}(q)} \right| \leq \delta.$$

Now, following the exact same reasoning as (6.19), we have

$$\begin{aligned} -\delta - \left(1 - \frac{1}{\alpha}\right) W + \frac{1}{\alpha} \frac{1}{k_r^-(q)} \left(S_{k_r^-(q)} - z_{k_r^-(q)} \right) \\ \leq \frac{1}{m} (S_m - z_m) \\ \leq \frac{\alpha}{k_r^+(q)} \left(S_{k_r^+(q)} - z_{k_r^+(q)} \right) + (\alpha - 1) W + \delta. \end{aligned}$$

So, (6.22) implies that (recalling that $\delta = \varepsilon/3$),

$$\begin{aligned} -\frac{2\varepsilon}{3} + \frac{1}{\alpha} \frac{1}{k_r^-(q)} \left(S_{k_r^-(q)} - z_{k_r^-(q)} \right) \leq \frac{1}{m} (S_m - z_m) \\ \leq \frac{\alpha}{k_r^+(q)} \left(S_{k_r^+(q)} - z_{k_r^+(q)} \right) + \frac{2\varepsilon}{3}. \end{aligned} \tag{6.26}$$

Now, the above and (6.24) (and the fact that $\alpha > 1$) implies $|1/m(S_m - z_m)| \leq \varepsilon$, which contradicts (6.25). □

6.3.2 Application I: Rates for pairwise independent random variables with bounded variance

This section will prove Theorem 6.3.2. First, we calculate a rate under the assumption that the random variables are nonnegative.

Fix a sequence of nonnegative, pairwise independent random variables $\{Y_n\}$ with, $\mathbb{E}(Y_n) \leq \mu \neq 0$, $\mathrm{Var}(Y_n) \leq \sigma_Y^2$ for all $n \in \mathbb{N}$ and some $\mu, \sigma_Y > 0$. Set $S_n^Y := \sum_{i=1}^n Y_i$ and $z_n^Y := \sum_{i=1}^n \mathbb{E}(Y_n)$.

Lemma 6.3.13. *For all $\varepsilon, \delta > 0$, $\alpha > 1$ and $0 \leq s \leq L_\delta$,*

$$R_{\varepsilon,\alpha}(\lambda) = \log_\alpha \left(\frac{2\sigma_Y^2}{\lambda \varepsilon^2 (\alpha - 1)} \right) - 1$$

is a rate of convergence for the partial sums of

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left| \frac{S^Y_{k_s^\pm(n)}}{k_s^\pm(n)} - \frac{z^Y_{k_s^\pm(n)}}{k_s^\pm(n)} \right| > \varepsilon \right) \tag{6.27}$$

to their respective limits.

Proof. Fix $\lambda, \varepsilon, \delta > 0$ and $\alpha > 1$ as well as $0 \leq s \leq L_\delta$, $Q \geq R_{\varepsilon,\alpha,}(\lambda)$. Then:

$$\begin{aligned}
\sum_{n=Q+1}^{\infty} \mathbb{P}\left(\left| \frac{1}{k_s^\pm(n)} S^Y_{k^\pm(n)} - \frac{1}{k_s^\pm(n)} z^Y_{k_s^\pm(n)} \right| > \varepsilon \right) &\leq \frac{1}{\varepsilon^2} \sum_{n=Q+1}^{\infty} \frac{\mathrm{Var}(S^Y_{k^\pm(n)})}{k_s^\pm(n)^2} \\
&\leq \frac{\sigma_Y^2}{\varepsilon^2} \sum_{n=Q+1}^{\infty} \frac{1}{k_s^\pm(n)} \\
&\leq \frac{2\sigma_Y^2}{\varepsilon^2} \sum_{n=Q+1}^{\infty} \alpha^{-n} \\
&\leq \frac{2\sigma_Y^2 \alpha^{-(Q+1)}}{\varepsilon^2(\alpha - 1)} \leq \lambda.
\end{aligned}$$

We get the first inequality from Chebyshev's inequality, the second inequality by pairwise independence, the third inequality by using $k^\pm(n) \geq \lfloor \alpha^n \rfloor > \alpha^n/2$, the fourth inequality by using the sum of an infinite geometric sequence and the last inequality from the assumption that $Q \geq R_{\varepsilon,\alpha}(\lambda)$. □

We can now apply Theorem 6.3.12 with the rate we obtained above, observing that R is independent of s (and δ), to easily obtain the following:

Lemma 6.3.14. *For all $\varepsilon, \lambda > 0$ and all*

$$n \geq \Delta_{\varepsilon,\mu,\sigma_Y}(\lambda) := \Phi_{\varepsilon,R}(\lambda) := \frac{36\alpha^2\sigma_Y^2}{\lambda\varepsilon^2(\alpha - 1)},$$

it holds that

$$\mathbb{P}\left(\sup_{m \geq n} \left| \frac{S_m^Y}{m} - \frac{z_{k_s^\pm(n)}}{k_s^\pm(n)} \right| > \varepsilon \right) \leq \lambda.$$

Here, $\alpha := 1 + \frac{\varepsilon}{3\mu}$, R is defined as in the previous lemma and Φ is defined as in Theorem 6.3.12.

Proof. This follows immediately from Theorem 6.3.12. Note we may take W to be μ. □

We can now obtain a rate where the random variables are not assumed to be nonnegative.

Proposition 6.3.15. *Let $\{X_n\}$ be a sequence of pairwise independent random variables with $\mathbb{E}(X_n) = 0$, $\mathrm{Var}(X_n) \leq \sigma^2$ and $\mathbb{E}(|X_n|) \leq \tau$ for all $n \in \mathbb{N}$ and some $\tau, \sigma > 0$. Furthermore, let $S_n := \sum_{i=1}^n X_i$. Then for all $\varepsilon, \lambda > 0$ and all $n \geq \Delta_{\frac{\varepsilon}{2},\frac{\tau}{2},\sigma}(\frac{\lambda}{2})$:*

$$P^*_{n,\varepsilon} = \mathbb{P}\left(\sup_{m \geq n} \left|\frac{S_m}{m}\right| > \varepsilon\right) \leq \lambda.$$

Here,

$$\Delta_{\frac{\varepsilon}{2},\frac{\tau}{2},\sigma}\left(\frac{\lambda}{2}\right) := \frac{288\alpha^2\sigma^2}{\lambda\varepsilon^2(\alpha - 1)}$$

as before with $\alpha := 1 + \frac{\varepsilon}{3\tau}$. Thus, $P^*_{n,\varepsilon}$ converges to 0 with a rate of convergence given above.

Proof. We have, for all $n \in \mathbb{N}$, $\sigma^2 \geq \mathrm{Var}(X_n) = \mathbb{E}(X_n^2) \geq \mathrm{Var}(X_n^+) + \mathrm{Var}(X_n^-)$ which implies $\sigma^2 \geq \mathrm{Var}(X_n^\pm)$. Furthermore, we have $\mathbb{E}(X_n^\pm) \leq \frac{\tau}{2}$ since $\mathbb{E}(X_n) = 0 = \mathbb{E}(X_n^+) - \mathbb{E}(X_n^-)$. Thus, if we take $\{Y_n\} = \{X_n^\pm\}$, we can set $\sigma_Y := \sigma$ and $\mu := \tau/2$. Furthermore, we can set

$$z_n := \sum_{i=1}^n \mathbb{E}(X_n^+) = \sum_{i=1}^n \mathbb{E}(X_n^-).$$

Thus, from the previous lemma:

$$\mathbb{P}\left(\sup_{m \geq n} \left|\frac{1}{m}S_m^\pm - \frac{1}{m}z_m\right| > \frac{\varepsilon}{2}\right) \leq \frac{\lambda}{2}$$

for all $\varepsilon, \lambda > 0$ and $n \geq \Delta_{\frac{\varepsilon}{2},\mu,\sigma_Y}(\frac{\lambda}{2})$. Thus, if $n \geq \Delta_{\frac{\varepsilon}{2},\frac{\tau}{2},\sigma}(\frac{\lambda}{2}) = \Delta_{\frac{\varepsilon}{2},\mu,\sigma_Y}(\frac{\lambda}{2})$, then

$$\begin{aligned}
\mathbb{P}\left(\sup_{m \geq n} \left|\frac{1}{m}S_m\right| > \varepsilon\right) &= \mathbb{P}\left(\sup_{m \geq n} \left|\left(\frac{1}{m}S_m^+ - \frac{1}{m}z_m\right) - \left(\frac{1}{m}S_m^- - \frac{1}{m}z_m\right)\right| > \varepsilon\right) \\
&\leq \mathbb{P}\left(\sup_{m \geq n} \left|\frac{1}{m}S_m^+ - \frac{1}{m}z_m\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\sup_{m \geq n} \left|\frac{1}{m}S_m^- - \frac{1}{m}z_m\right| > \frac{\varepsilon}{2}\right) \\
&\leq \lambda.
\end{aligned}$$

Here $S_n^\pm := \sum_{i=1}^n X_n^\pm$. □

This, in particular, allows us rather immediately to deduce Theorem 6.3.2:

Proof of Theorem 6.3.2. Using the above proposition, we have

$$P^*_{n,\varepsilon} \leq \frac{288\alpha^2\sigma^2}{n\varepsilon^2(\alpha - 1)}$$

for all $n \in \mathbb{N}$. So Theorem 6.3.2 follows by noting that if $\varepsilon \leq \tau$, we will have $\alpha \leq 4/3$. □

Remark 6.3.16. In the case $\varepsilon > \tau$, observe that $\alpha < 4\varepsilon/3\tau$ and so we can deduce

$$P^*_{n,\varepsilon} \leq \frac{1536\sigma^2}{n\varepsilon\tau}.$$

We shall now discuss the optimality of the bound we obtained.

Example 6.3.17. For $\delta > 0$, let $\{X_n\}$ be a sequence of integer-valued iid random variables such that

$$\mathbb{P}(X_0 = n) = \frac{c}{n^{3+\delta}} \text{ for } c = \left(\sum_{n=1}^{\infty} \frac{1}{n^{3+\delta}}\right)^{-1}$$

for all $n \in \mathbb{Z}^+$ (and probability 0 for all other integers). Then $\mathrm{Var}(X_0) < \infty$ and for all $1 \geq \varepsilon > 0$ and any $n \in \mathbb{N}$:

$$\mathbb{P}\left(\sup_{m \geq n} \left|\frac{S_m}{m} - \mu\right| > \varepsilon\right) \geq \frac{\omega}{n^{1+\delta}}$$

where μ is the mean of X_0, given by

$$\mu = \sum_{n=1}^{\infty} \frac{c}{n^{2+\delta}},$$

and

$$\omega = \frac{c}{2 \times 3^{2+\delta}(2+\delta)}.$$

Proof. These random variables clearly have finite variance. For any $1 \geq \varepsilon > 0$, we have

$$\mathbb{P}\left(\sup_{m \geq n} \left|\frac{S_m}{m} - \mu\right| > \varepsilon\right) \geq \mathbb{P}\left(\sup_{m \geq n} \left|\frac{S_m}{m} - \mu\right| \geq 1\right).$$

Observe that

$$\mu = \frac{\sum_{n=1}^{\infty} \frac{1}{n^{2+\delta}}}{\sum_{n=1}^{\infty} \frac{1}{n^{3+\delta}}} < \frac{\zeta(2)}{\zeta(4)} = \frac{15}{\pi^2} < 2.$$

Thus, we get

$$\begin{aligned}
\mathbb{P}\left(\sup_{m \geq n} \left|\frac{S_m}{m} - \mu\right| \geq 1\right) &\geq \mathbb{P}\left(\sup_{m \geq n} \frac{S_m}{m} \geq 3\right) \\
&\geq \mathbb{P}\left(\frac{S_n}{n} \geq 3\right) \\
&\geq \mathbb{P}(X_0 \geq 3n \cup \ldots \cup X_n \geq 3n) \\
&= 1 - \mathbb{P}(X_0 < 3n \cap \ldots \cap X_n < 3n) \\
&= 1 - (\mathbb{P}(X_0 < 3n))^n = 1 - (1 - \mathbb{P}(X_0 \geq 3n))^n.
\end{aligned}$$

We now have

$$\mathbb{P}(X_0 \geq 3n) = c \sum_{k=3n}^{\infty} \frac{1}{k^{3+\delta}} \geq c \int_{3n}^{\infty} \frac{1}{x^{3+\delta}}\, dx = \frac{c}{(3n)^{2+\delta}(2+\delta)} = \frac{w}{n^{2+\delta}},$$

where $w = \frac{c}{3^{2+\delta}(2+\delta)}$. This implies,

$$\mathbb{P}\left(\sup_{m\geq n}\left|\frac{S_m}{m} - \mu\right| > \varepsilon\right) \geq 1 - \left(1 - \frac{w}{n^{2+\delta}}\right)^n \geq$$
$$1 - \frac{1}{1 + \frac{w}{n^{1+\delta}}} \geq \frac{w}{2n^{1+\delta}},$$

where we used the inequality $(1+x)^n \leq \frac{1}{1+nx}$ and $w < 1$. This yields the result. □

Therefore, for every $\delta > 0$, by translation, we can obtain a sequence of iid random variables with expected values equal to 0 and finite variance, such that

$$\mathbb{P}\left(\sup_{m\geq n}\left|\frac{1}{m}S_m\right| > \varepsilon\right) \geq \frac{\omega}{n^{1+\delta}}.$$

This example demonstrates that $O\left(\frac{1}{n}\right)$ is an optimal general power of n bound for $P^*_{n,\varepsilon}$ in the case of finite variance. It however does not rule out the possibility that $P^*_{n,\varepsilon} = O\left(\frac{1}{n\log(n)}\right)$, for example.

6.3.3 Application II: Rates for the Chen-Sung Law of Large Numbers

Throughout this section, let $\{X_n\}$ be a sequence of random variables with finite pth moment (for some fixed $p \geq 1$) and respective means $\{\mu_n\}$. Let $S_n := \sum_{k=1}^n X_k$ and $z_n := \sum_{i=1}^n \mu_i$.

To use Theorem 6.3.12 to obtain a quantitative version of Theorem 6.3.6, we must find a rate of convergence for (6.15). To do this, we need some lemmas. The first is a technical lemma that resembles the Hájek and Rényi inequality.

Lemma 6.3.18. *Suppose $\{\gamma_n\}$ is a sequence of nonnegative real numbers satisfying*

$$\mathbb{E}(|S_n - z_n|^p) \leq \sum_{k=1}^{n} \gamma_k.$$

For all $\varepsilon, \delta > 0$, $\alpha > 1$ and $0 \leq s \leq L_\delta$:

$$\begin{aligned}\sum_{n=Q+1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k_s^\pm(n)}}{k_s^\pm(n)} - \frac{z_{k_s^\pm(n)}}{k_s^\pm(n)}\right| > \varepsilon\right) \\ \leq \frac{2^p\alpha^{2p}}{\varepsilon^p \left\lfloor \alpha^{Q+2} \right\rfloor^p (\alpha^p - 1)} \sum_{m=1}^{\left\lfloor \alpha^{Q+2} \right\rfloor} \gamma_m + \frac{2^p\alpha^{2p}}{\varepsilon^p(\alpha^p-1)} \sum_{m=\lfloor \alpha^{Q+1} \rfloor + 1}^{\infty} \frac{\gamma_m}{m^p}.\end{aligned} \tag{6.28}$$

Proof. Fix $M \in \mathbb{N}$. By the generalised Chebyshev's inequality, we have

$$\begin{aligned}\sum_{n=Q+1}^{M} \mathbb{P}\left(\left|\frac{S_{k_s^\pm(n)}}{k_s^\pm(n)} - \frac{z_{k_s^\pm(n)}}{k_s^\pm(n)}\right| > \varepsilon\right) &\leq \frac{1}{\varepsilon^p} \sum_{n=Q+1}^{M} \frac{\mathbb{E}\left(\left|S_{k_s^\pm(n)} - z_{k_s^\pm(n)}\right|^p\right)}{k_s^\pm(n)^p} \\ &\leq \frac{1}{\varepsilon^p} \sum_{n=Q+1}^{M} \frac{1}{k_s^\pm(n)^p} \sum_{m=1}^{k_s^\pm(n)} \gamma_m.\end{aligned}$$

Now, splitting the inner sum into two parts, observing that if $n > Q+1$ then $k_s^\pm(n) > k_s^\pm(Q+1)$, we have that the above sum is equal to

$$\frac{1}{\varepsilon^p} \sum_{n=Q+1}^{M} \frac{1}{k_s^\pm(n)^p} \sum_{m=1}^{k_s^\pm(Q+1)} \gamma_m + \frac{1}{\varepsilon^p} \sum_{n=Q+1}^{M} \frac{1}{k_s^\pm(n)^p} \sum_{m=k_s^\pm(Q+1)+1}^{k_s^\pm(n)} \gamma_m. \tag{6.29}$$

Now, interchanging summations in the first term and using $k_s^\pm(n) \geq \lfloor \alpha^n \rfloor > \alpha^n/2$, we can bound the first term from above by

$$\begin{aligned}\frac{2^p}{\varepsilon^p} \sum_{m=1}^{k_s^\pm(Q+1)} \gamma_m \sum_{n=Q+1}^{M} \alpha^{-pn} &\leq \frac{2^p\alpha^p}{\varepsilon^p\alpha^{p(Q+1)}(\alpha^p-1)} \sum_{m=1}^{k_s^\pm(Q+1)} \gamma_m \\ &\leq \frac{2^p\alpha^{2p}}{\varepsilon^p\alpha^{p(Q+2)}(\alpha^p-1)} \sum_{m=1}^{k_s^\pm(Q+1)} \gamma_m \\ &\leq \frac{2^p\alpha^{2p}}{\varepsilon^p \left\lfloor \alpha^{Q+2} \right\rfloor^p (\alpha^p-1)} \sum_{m=1}^{\left\lfloor \alpha^{Q+2} \right\rfloor} \gamma_m.\end{aligned}$$

We bound the first line by an infinite geometric series to get the second line, and we use $\alpha^{Q+2} > k_s^\pm(Q+1)$ to get from the penultimate line to the last line.

We now bound the second term in (6.29) from above. Again, interchanging summations and using similar manipulations as used to obtain the bound for the first term, we get that the

second is bounded above by

$$\frac{1}{\varepsilon^p}\sum_{m=k_s^{\pm}(Q+1)+1}^{k_s^{\pm}(M)}\gamma_m\sum_{\{M\geq n\geq Q+1:k_s^{\pm}(n)\geq m\}}\frac{1}{k_s^{\pm}(n)^p}$$
$$\leq\frac{2^p}{\varepsilon^p}\sum_{m=k_s^{\pm}(Q+1)+1}^{k_s^{\pm}(M)}\gamma_m\sum_{\{M\geq n\geq Q:\alpha^{n+1}\geq m\}}\alpha^{-pn}.$$

The inner sum is bounded by an infinite geometric series with the first term $\leq m^{-p}\alpha^p$. Thus, the above is again bounded above by

$$\frac{2^p\alpha^{2p}}{\varepsilon^p(\alpha^p-1)}\sum_{m=k_s^{\pm}(Q+1)+1}^{k_s^{\pm}(M)}\frac{\gamma_m}{m^p}\leq\frac{2^p\alpha^{2p}}{\varepsilon^p(\alpha^p-1)}\sum_{m=\lfloor\alpha^{Q+1}\rfloor+1}^{\lfloor\alpha^{M+1}\rfloor}\frac{\gamma_m}{m^p}.$$

Taking $M\to\infty$ gives the required result. □

To find a rate of convergence for (6.15), we must find one for the two terms on the right-hand side of (6.28). A rate for the second term can easily be calculated given one for $\sum_{m=1}^{\infty}\frac{\gamma_m}{m^p}$. To obtain a rate for the second term, we need a quantitative version of Kronecker's Lemma, which we give in very general form as Corollary 5.2.3. We spell out the specific instance of this theorem we need for our current purposes.

Lemma 6.3.19 (Quantitative Kronecker's lemma)**.** *Let $x_1, x_2, \ldots$ be a sequence of nonnegative real numbers such that $\sum_{i=1}^{\infty}x_i<\infty$ and let $0<a_1\leq a_2\leq\ldots$ be such that $a_n\to\infty$. Quantitatively, suppose $\sum_{i=1}^{\infty}x_i<S$ for some $S>0$ and that $s_n:=\sum_{i=1}^{n}x_i$ converges to $\sum_{i=1}^{\infty}x_i$ with rate of convergence ϕ. Further, suppose that there is a function $f:\mathbb{R}\to\mathbb{N}$ such that $a_{f(\omega)}\geq\omega$ for all $\omega>0$. Then*

$$\frac{1}{a_n}\sum_{i=1}^{n}a_ix_i\to 0$$

as $n\to\infty$ with rate of convergence

$$K_{\phi,f,\{a_n\},S}(\varepsilon)=\max\left\{\phi\left(\frac{\varepsilon}{4}\right),f\left(\frac{4a_{\phi(\frac{\varepsilon}{4})}S}{\varepsilon}\right)\right\}.$$

We can now calculate a rate of convergence for (6.15)

Lemma 6.3.20. *Suppose $\{X_n\}$ and $\{\gamma_n\}$ are as in Theorem 6.3.6. Suppose $\sum_{m=1}^{\infty}\frac{\gamma_n}{m^p}\leq\Gamma$ for some $\Gamma>0$ and that the partial sums converge to their limit with a strictly decreasing rate of convergence Ψ. For all $\varepsilon,\delta>0$, $\alpha>1$ and $0\leq s\leq L_\delta$, the function $\chi_{\varepsilon,\alpha,\Psi}$ is a rate of*

convergence for the partial sums of

$$\sum_{m=1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k_s^{\pm}(m)}}{k_s^{\pm}(m)} - \frac{z_{k_s^{\pm}(m)}}{k_s^{\pm}(m)}\right| > \varepsilon\right)$$

to their respective limits, where

$$\chi_{\varepsilon,\alpha,\Psi}(\lambda) = \max\left\{\log_\alpha\left(2\Psi\left(\frac{\lambda\varepsilon^p(\alpha^p-1)}{2^{p+1}\alpha^{2p}}\right)\right), \log_\alpha\left(2K_{\psi,f_p,\{n^p\},R}\left(\frac{\lambda}{2}\right)\right)\right\}$$

with

$$\psi(\lambda) = \Psi\left(\frac{\lambda\varepsilon^p(\alpha^p-1)}{2^p\alpha^{2p}}\right), \quad f_p(\omega) = \left\lceil \omega^{\frac{1}{p}} \right\rceil, \quad R = \frac{2^p\Gamma\alpha^{2p}}{\varepsilon^p(\alpha^p-1)}.$$

Proof. Let $\lambda, \varepsilon, \delta > 0$, $\alpha > 1$ and $0 \leq s \leq L_\delta$ as well as $n \geq \chi_{\varepsilon,\alpha,\Psi}(\lambda)$ be given. We have, by Lemma 6.3.18, that

$$\begin{aligned}
&\sum_{m=n+1}^{\infty} \mathbb{P}\left(\left|\frac{S_{k_s^{\pm}(m)}}{k_s^{\pm}(m)} - \frac{z_{k_s^{\pm}(m)}}{k_s^{\pm}(m)}\right| > \varepsilon\right) \\
&\qquad \leq \frac{2^p\alpha^{2p}}{\varepsilon^p \left\lfloor \alpha^{n+2} \right\rfloor^p (\alpha^p-1)} \sum_{m=1}^{\left\lfloor \alpha^{n+2} \right\rfloor} \gamma_m + \frac{2^p\alpha^{2p}}{\varepsilon^p(\alpha^p-1)} \sum_{m=\left\lfloor \alpha^{n+1} \right\rfloor+1}^{\infty} \frac{\gamma_m}{m^p}.
\end{aligned}$$

Now, $n \geq \chi_{\varepsilon,\alpha,\Psi}(\lambda)$ implies

$$n \geq \log_\alpha\left(2\Psi\left(\frac{\lambda\varepsilon^p(\alpha^p-1)}{2^{p+1}\alpha^{2p}}\right)\right)$$

and from this, we deduce

$$\left\lfloor \alpha^{n+1} \right\rfloor \geq \alpha^{n+1}/2 \geq \Psi\left(\frac{\lambda\varepsilon^p(\alpha^p-1)}{2^{p+1}\alpha^{2p}}\right),$$

which in turn implies

$$\frac{2^p\alpha^{2p}}{\varepsilon^p(\alpha^p-1)} \sum_{m=\left\lfloor \alpha^{n+1} \right\rfloor+1}^{\infty} \frac{\gamma_m}{m^p} \leq \frac{\lambda}{2}.$$

Now, observe that the partial sums of

$$\frac{2^p\alpha^{2p}}{\varepsilon^p(\alpha^p-1)} \sum_{m=1}^{\infty} \frac{\gamma_m}{m^p}$$

converge to their limit, with rate

$$\psi(\lambda) = \Psi\left(\frac{\lambda\varepsilon^p(\alpha^p-1)}{2^p\alpha^{2p}}\right).$$

Therefore, by Lemma 6.3.19, we have that

$$\frac{2^p\alpha^{2p}}{\varepsilon^p n^p(\alpha^p-1)}\sum_{m=1}^{n}\gamma_m$$

converges to 0 with rate $K_{\psi,f_p,\{n^p\},R}$. Now, $n \geq \chi_{\varepsilon,\alpha,\Psi}(\lambda)$ further implies

$$n \geq \log_\alpha\left(2K_{\psi,f_p,\{n^p\},R}\left(\frac{\lambda}{2}\right)\right)$$

and, arguing as above, we get

$$\left\lfloor \alpha^{n+2}\right\rfloor \geq K_{\psi,f_p,\{n^p\},R}\left(\frac{\lambda}{2}\right).$$

This allows us to conclude

$$\frac{2^p\alpha^{2p}}{\varepsilon^p\left\lfloor\alpha^{n+2}\right\rfloor^p(\alpha^p-1)}\sum_{m=1}^{\left\lfloor\alpha^{n+2}\right\rfloor}\gamma_m \leq \frac{\lambda}{2}$$

and we are done. □

This, in particular, allows us to deduce Theorem 6.3.7 and Theorem 6.3.8.

Proof of Theorem 6.3.7. In the context of the assumptions of Theorem 6.3.7, using the assumption that Ψ is strictly decreasing, the rate χ in the previous Lemma 6.3.20 simplifies to

$$\log_\alpha\left(\max\left\{2\Psi\left(\frac{\lambda\varepsilon^p(\alpha^p-1)}{2^{p+3}\alpha^{2p}}\right), 2\left\lceil\frac{2\alpha^2}{\varepsilon}\Psi\left(\frac{\lambda\varepsilon^p(\alpha^p-1)}{2^{p+3}\alpha^{2p}}\right)\left(\frac{8\Gamma}{\lambda(\alpha^p-1)}\right)^{\frac{1}{p}}\right\rceil\right\}\right).$$

We can now apply Theorem 6.3.12 with the rate we obtained above, observing that χ is independent of s (and δ) to deduce Theorem 6.3.7, noting that we can take $\alpha = 1+\frac{\varepsilon}{3W}$ and so the assumption that $\varepsilon \leq 1 \leq W$ implies $\alpha < 4/3$ and $\alpha^p - 1 \geq \alpha - 1$. Furthermore, the assumptions that $\Gamma, W \geq 1$ and $0 < \varepsilon \leq 1$, as well as the strictly decreasing condition on Ψ, allows us to conclude that the second argument in the max function above is bigger than the first argument. □

Proof of Theorem 6.3.8. We write $X_n = X_n^+ - X_n^-$ and apply Theorem 6.3.7 to each sequence $\{X_n^\pm\}$, with $p = 2$ and $\gamma_n = \mathrm{Var}(X_n^\pm) \leq \mathrm{Var}(X_n)$. We then obtain the result for $\{X_n\}$ by arguing as in Proposition 6.3.15. □

Chapter 7

Fluctuations in martingales and ergodic averages

The martingale convergence theorem is a fundamental result that establishes the convergence of stochastic algorithms in various contexts, as detailed in the survey [42]. In particular, it is crucial in proving the Robbins-Siegmund theorem (which we give a computational interpretation for in Chapter 8). Thus, understanding the computational content of the martingale convergence theorem is required to make significant progress in proof mining within stochastic processes.

A central result required to prove the martingale convergence theorem is a result of Doob that provides a bound on the expected value of the number of upcrossings made by martingales. In collaboration with Powell, the author investigated how such upcrossing inequalities could lead to the quantitative notions of stochastic convergence we introduced in Chapter 4. This exploration allows us to offer a computational interpretation of the martingale convergence theorem.

Initial progress in connecting upcrossing inequalities with uniform metastability was made by Avigad, Gerhardy and Towsner [6] in the context of the pointwise ergodic theorem with the assumption that the measurable functions (corresponding to sequences of random variables in the general context) were bounded. The author and Powell observed that one could generalise the uniform metastable rates of [6] to stochastic processes satisfying general upcrossing inequalities and weakening the boundedness requirement. The author and Powell strengthened this observation further by demonstrating that one could actually construct learnable uniform rates of convergence for a general class of stochastic processes, which included those in the pointwise ergodic theorem and the martingale convergence theorem, satisfying upcrossing inequalities. These results are noted in [117].

Quantitative convergence results concerning martingales and ergodic averages are of great interest and have been studied extensively in the probability theory and ergodic theory litera-

ture. Most notably, the very influential survey paper by Kachurovskii detailing the research of mostly Russian influences on quantitative aspects of the convergence of martingales and ergodic averages [70]. In addition, the collection of work by Jones and collaborators ([67, 68, 69], for example), which, through the use of deep results from ergodic theory, were able to independently obtain many of the results detailed in [70], among others. Through an abstract investigation of interactions between the notions of stochastic convergence, introduced in Section 4.2, the author and Powell [117] were able to generalise and improve some of the results coming from the aforementioned body of work.

The main result of this chapter will be a quantitative version of Doob's martingale convergence theorem (Theorem 7.2.4). We start in Section 7.1, where we investigate how the abstract modes of quantitative stochastic convergence, which we introduced in Section 4.2, interact. An abstract exploration of how moduli of finite crossing, moduli of finite fluctuations and pointwise learnable rates of convergence interact results in a generalisation of a result of Kachurovskii and an improvement of a bound attributed to Ivanov [70] (which we present in Section 7.3.2). We then investigate the interaction between moduli of L_1-crossing and uniform learnable rates, with this study crucial in obtaining our quantitative version of the martingale convergence theorem.

Then, continuing in Section 7.2, we present the quantitative martingale convergence theorem, in the form of uniform learnable rates, of the author and Powell [117]. We further present a detailed argument that the rates we obtain are indeed optimal (a result only sketched in [117]) and the new result that our quantitative martingale convergence theorem allows one to generalise the stochastic fluctuation bounds of Chashka [21] to submartingales and supermartingales.

In Section 7.3, we then conclude by detailing the generalisation of the metastable rates obtained for the pointwise ergodic theorem in [6] to unbounded functions. Furthermore, we present the improvement of Ivonov's bound on the fluctuations of ergodic averages, detailed in [70], which follows from our abstract analysis of the interactions between moduli of finite crossing, moduli of finite fluctuations and pointwise learnable rates of convergence. The aforementioned are from the author's collaboration with Powell [117]. We finish the section by briefly discussing the new observations between the framework of modes of convergence introduced in this thesis and the previously mentioned work championed by Jones and his collaborators.

7.1 Abstract results on probabilistic convergence

The proofs of the martingale convergence theorem and the pointwise ergodic theorem we analyse to obtain our quantitative versions of these respective theorems rely on so-called upcrossing inequalities. Such inequalities shall allow us to obtain the quantitative notions for finite crossings we introduced in Section 4.2. Thus, our quantitative results will follow from general results

on the interplay between the aforementioned quantitative notions of finite crossings, stochastic fluctuations, and learnable rates of almost sure convergence. In this section, we explore such results.

7.1.1 Crossings, fluctuations and pointwise convergence

We provide the stochastic analogue of Theorem 2.3.20, demonstrating the relationship between moduli of finite crossings and finite fluctuations. As in the deterministic case, going from fluctuations to crossings, quantitatively, is straightforward, and the converse is more interesting.

Theorem 7.1.1. *Let $\{X_n\}$ be a stochastic process. If ϕ is a modulus of finite crossings and f is a modulus of uniform boundedness for $\{X_n\}$ then*

$$\psi(\lambda, \varepsilon) := l \cdot \phi\left(\frac{\lambda}{2}, M, l\right) \quad \textit{for } \; l := \left\lceil \frac{4M}{\varepsilon} \right\rceil \; \textit{ and } \; M := f\left(\frac{\lambda}{2}\right)$$

is a modulus of finite fluctuations for the same process, and therefore also a learnable rate of pointwise convergence.

Proof. Fix $\lambda \in (0, 1]$ and note that for any event A and $M := f(\lambda/2)$:

$$\begin{aligned} \mathbb{P}(A) &\leq \mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| > M\right) + \mathbb{P}\left(A \cap \sup_{n\in\mathbb{N}} |X_n| \leq M\right) \\ &< \frac{\lambda}{2} + \mathbb{P}\left(A \cap \sup_{n\in\mathbb{N}} |X_n| \leq M\right). \end{aligned}$$

Now let $\varepsilon \in (0, 1]$ be given and set $N := \psi(\lambda, \varepsilon)$. It suffices to show

$$\mathbb{P}\left(J_\varepsilon\{X_n\} \geq N \cap \sup_{n\in\mathbb{N}} |X_n| \leq M\right) < \frac{\lambda}{2}. \tag{7.1}$$

For any fixed $\omega \in \Omega$, reasoning as in the proof of Proposition 2.3.20, if $J_\varepsilon\{X_n(\omega)\} \geq N$ and $\sup_{n\in\mathbb{N}} |X_n(\omega)| \leq M$, then for $l := \lceil 4M/\varepsilon \rceil$, any interval in $\mathcal{P}(M, l)$ has width $\leq \varepsilon/2$, and so any ε-fluctuation of $\{X_n(\omega)\}$ is also an $[\alpha, \beta]$-crossing for some $[\alpha, \beta] \in \mathcal{P}(M, l)$. By the pigeonhole principle there must therefore be some $[\alpha, \beta] \in \mathcal{P}(M, l)$ with $C_{[\alpha,\beta]}\{X_n(\omega)\} \geq N/l$, and so we have shown that

$$\mathbb{P}\left(J_\varepsilon\{X_n\} \geq N \cap \sup_{n\in\mathbb{N}} |X_n| \leq M\right) \subseteq \mathbb{P}\left(\exists [\alpha, \beta] \in \mathcal{P}(M, l)\, C_{[\alpha,\beta]}\{X_n(\omega)\} \geq \frac{N}{l}\right)$$

□

Remark 7.1.2. As discussed in Remark 4.2.17, given a crossing inequality, one can obtain a modulus of finite crossings by Markov's inequality. Thus, given such an inequality, we can

use the above theorem to obtain a modulus of finite fluctuations, which will be a pointwise learnable rate of convergence, and thus, we can get a rate of metastable pointwise convergence (see Remark 4.2.22). Furthermore, Corollary 4.2.26 also allows us to obtain a rate of metastable uniform convergence (although such a rate will have bar recursive complexity). We shall see in the next subsection that crossing inequalities contain further uniformity properties that allow us to directly obtain uniform learnable rates (and uniform metastable rates of low complexity) without the need to pass through fluctuations.

Remark 7.1.3. All of the quantitative definitions we have introduced are based on functions that reflect the logical structure of the underlying notion being captured. This is often at odds with the more traditional formulations. For example, we capture the uniform boundedness of a stochastic process with a function ϕ satisfying

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq \phi(\lambda)\right) < \lambda \quad \text{for all } \lambda \in (0,1]$$

rather than a function f satisfying

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq m\right) < f(m) \quad \text{for all } m \in \mathbb{N}.$$

It is precisely because they represent the underlying quantifier structure that these moduli are better suited to formulating the computational structure of proofs than traditional rates, which implicitly involve additional assumptions, such as

$$\lim_{m\to\infty} f(m) = 0$$

in the example above. In any case, we can always convert our moduli to traditional rates to facilitate comparison with known results.

Theorem 7.1.1 represents a generalisation of a result of Kachurovskii [70]. To make this more apparent, we must reformulate Theorem 7.1.1 in terms of traditional rates.

Corollary 7.1.4. *Let $\{X_n\}$ be a stochastic process such that:*

(a) $\mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq a\right) < g(a)$ *for all $a > 0$, where g is a strictly decreasing function satisfying $g(a) \to 0$ as $a \to \infty$.*

(b) $\mathbb{P}\left(C_{[\alpha,\beta]}\{X_n\} \geq a\right) < h_{\alpha,\beta}(a)$ *for all $\alpha < \beta$ such that $\mathbb{P}\left(C_{[\alpha,\beta]}\{X_n\} > 0\right) > 0$ and $a > 0$, where $h_{\alpha,\beta}$ is a strictly decreasing function satisfying $h_{\alpha,\beta}(a) \to 0$ as $a \to \infty$.*

Then for all $\varepsilon > 0$

$$\mathbb{P}\left(J_\varepsilon\{X_n\} \geq a\right) < G_\varepsilon^{-1}(a)$$

for any strictly decreasing function, G_ε satisfying

$$l \cdot H\left(\frac{\lambda}{2}, g^{-1}\left(\frac{\lambda}{2}\right), l\right) < G_\varepsilon(\lambda) \quad \textit{for} \quad l := \left\lceil \frac{4g^{-1}(\lambda/2)}{\varepsilon} \right\rceil$$

where H is any function such that

$$h^{-1}_{\alpha,\beta}\left(\frac{\lambda}{l}\right) \leq H(\lambda, M, l)$$

for any $\lambda \in (0,1], M > 0, l \in \mathbb{N}$ nonzero and $[\alpha,\beta] \in \mathcal{P}(M,l)$ with $\mathbb{P}\left(C_{[\alpha,\beta]}\{X_n\} > 0\right) > 0$.

Proof. By definition, g^{-1} is a modulus of uniform boundedness for $\{X_n\}$, and $h^{-1}_{\alpha,\beta}$ is a modulus of finite $[\alpha,\beta]$-crossings for all $\alpha < \beta$ with $\mathbb{P}\left(C_{[\alpha,\beta]}\{X_n\} > 0\right) > 0$. By Lemma 4.2.16 and the property of H (noting that the restriction to $[\alpha,\beta] \in \mathcal{P}(M,l)$ with $\mathbb{P}\left(C_{[\alpha,\beta]}\{X_n\} > 0\right) > 0$ does not affect Lemma 4.2.16), H must be a modulus of finite crossings for $\{X_n\}$. Now, by Theorem 7.1.1 (ii), any bound on

$$l \cdot H(\lambda/2, g^{-1}(\lambda/2), l)$$

for $l := \lceil 4g^{-1}(\lambda/2)/\varepsilon \rceil$ is a modulus of finite ε-fluctuations for $\{X_n\}$, and so by definition we have

$$\mathbb{P}\left(J_\varepsilon\{X_n\} \geq G_\varepsilon(\lambda)\right) < \lambda$$

for all $\lambda > 0$, from which the result follows. □

We now have the following:

Example 7.1.5. In the special case that $\{X_n\}$ satisfies:

(i) $\mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq a\right) < \frac{S}{a}$ for all $a > 0$.

(ii) $\mathbb{E}\left(U_{[\alpha,\beta]}\{X_n\}\right) < \frac{S+|\alpha|}{\beta-\alpha}$ for all $\alpha < \beta$.

then Corollary 7.1.4 gives us, for $S/\varepsilon \geq 1$,

$$\mathbb{P}\left(J_\varepsilon\{X_n\} \geq a\right) < \frac{c}{a^{1/4}}\left(1 + \frac{S}{\varepsilon}\right)$$

for a constant $c > 0$. To see this, we would set $g(a) := S/a$. Now, by Remark 2.3.8, we have

$$\mathbb{E}\left(C_{[\alpha,\beta]}\{X_n\}\right) < \frac{2(S+|\alpha|)}{\beta-\alpha} + 1$$

for all $\alpha < \beta$, and so, by Markov's inequality, we can set

$$h_{\alpha,\beta}(k) := \frac{2(S+|\alpha|)+\beta-\alpha}{k(\beta-\alpha)}.$$

for all $\alpha < \beta$, so that a suitable definition of the bounding function H becomes

$$h^{-1}_{\alpha,\beta}\left(\frac{\lambda}{l}\right) \leq \frac{2l(S+|\alpha|)+l(\beta-\alpha)}{\lambda(\beta-\alpha)} \leq \frac{2l(S+M)+2M}{2M\lambda/l} \leq \frac{l^2}{\lambda}\left(2+\frac{S}{M}\right) =: H(\lambda, M, l),$$

where the last inequality follows since $\lambda \in (0,1]$ and $l \in \mathbb{N}$. Now, Because $g^{-1}(\lambda/2) = 2S/\lambda$ and

$$l \leq \frac{16S}{\varepsilon\lambda}$$

we have,

$$l \cdot H(\lambda/2, g^{-1}(\lambda/2), l) = \frac{2l^3}{\lambda}\left(2+\frac{\lambda}{2}\right) \leq \frac{5l^3}{\lambda} < \frac{cS^3}{\varepsilon^3\lambda^4}$$

for suitable constant $c > 0$, and therefore (since $S/\varepsilon \geq 1$)

$$\mathbb{P}\left(J_\varepsilon\{X_n\} \geq a\right) < \frac{c}{a^{1/4}}\left(\frac{S}{\varepsilon}\right)^{3/4} < \frac{c}{a^{1/4}}\left(1+\frac{S}{\varepsilon}\right)$$

This particular case of Corollary 7.1.4 is already proven directly by Kachorovskii as Theorem 27 of [70], where the direct (though more ad-hoc) proof allows for a slightly better value of $c = 7$.

7.1.2 Crossings, fluctuations and uniform convergence

A particular case of Example 7.1.5 is the martingale convergence theorem (that is, the case where $\{X_n\}$ are martingales with uniformly bounded first moment), where conditions (i) and (ii) are met by Doob's maximal inequality and Doob's upcrossing inequality, respectively. However, in this case, sharper bounds are known to hold, with optimal bounds found by Chashka [21]. In the following sections of this chapter, amongst other results, we generalise Chashkas's result to submartingales and supermartingales, and a key component of these results will be an abstract theorem detailing how one obtains learnable uniform rates from moduli of L_1-crossings (c.f. Definition 4.2.18).

Our first step is to provide another analogue of Proposition 2.3.19,

Proposition 7.1.6. *Let $\{X_n\}$ be a stochastic process with modulus of L_1-crossings ψ and let $M > 0$. Then the formula*

$$Q_M(\varepsilon, n, m) := (\exists l, k \in [n; m](|X_l - X_k| \geq \varepsilon)) \cap (|X_n| \leq M)$$

satisfies

$$\sum_{i=0}^{\infty} \mathbb{P}(Q_M(\varepsilon, a_i, b_i)) \leq \omega_M(\varepsilon)$$

uniformly in $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ *where*

$$\omega_M(\varepsilon) := (p+2) \cdot \psi\left(M\left(1+\frac{2}{p}\right), p+2\right) \quad \textit{for } p := \left\lceil \frac{8M}{\varepsilon} \right\rceil.$$

Proof. Fix $\varepsilon > 0$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ and define the formula A_i and B_i for $i \in \mathbb{N}$ by

$$A_i := \exists k, l \in [a_i; b_i](|X_k - X_l| \geq \varepsilon) \quad \text{and} \quad B_i := |X_{a_i}| \leq M.$$

Divide $[-M, M]$ into $p = \lceil 8M/\varepsilon \rceil$ equal subintervals, which we label $[\alpha_j, \beta_j]$ for $j = 1, \ldots, p$, and add two further intervals of the same width on either side of $[-M, M]$, which we also label $[\alpha_j, \beta_j]$ for $j = 0$ and $j = p+1$. In other words

$$\{[\alpha_0, \beta_0], \ldots, [\alpha_{p+1}, \beta_{p+1}]\} = \mathcal{P}(M(1+2/p), p+2).$$

These intervals must have width $\leq \varepsilon/4$. Suppose that $\omega \in A_i \cap B_i$, so that there exists $k(\omega), l(\omega) \in [a_i; b_i]$ with $|X_{k(\omega)}(\omega) - X_{l(\omega)}(\omega)| \geq \varepsilon$, and also $|X_{a_i}(\omega)| \leq M$.

Then by the triangle inequality, either $|X_{a_i}(\omega) - X_{k(\omega)}(\omega)| \geq \varepsilon/2$ or $|X_{a_i}(\omega) - X_{l(\omega)}(\omega)| \geq \varepsilon/2$. Since $X_{a_i}(\omega) \in [-M, M]$ and we have the additional intervals $[\alpha_0, \beta_0]$ and $[\alpha_{p+1}, \beta_{p+1}]$, it follows that one of the intervals $[\alpha_j, \beta_j]$ for $j = 0, \ldots, p+1$ is crossed by $\{X_n(\omega)\}$ somewhere in $[a_i; b_i]$, and therefore defining

$$T_{i,j} := \{X_n\} \text{ crosses } [\alpha_j, \beta_j] \text{ somewhere in } [a_i; b_i]$$

we have shown that

$$A_i \cap B_i \subseteq \bigcup_{j=0}^{p+1} T_{i,j}.$$

Now suppose that $r < \sum_{i=0}^{\infty} \mathbb{P}(A_i \cap B_i)$ for some $r > 0$, which in particular means that for some $N \in \mathbb{N}$ we have

$$r < \sum_{i=0}^{N} \mathbb{P}(A_i \cap B_i) \leq \sum_{i=0}^{N} \mathbb{P}\left(\bigcup_{j=0}^{p+1} T_{i,j}\right) \leq \sum_{i=0}^{N} \sum_{j=0}^{p+1} \mathbb{P}(T_{i,j})$$

and so there is some $j \in \{0, \ldots, p+1\}$ such that

$$\begin{aligned} \frac{r}{p+2} &< \sum_{i=0}^{N} \mathbb{P}(T_{i,j}) \leq \sum_{i=0}^{\infty} \mathbb{P}(T_{i,j}) = \sum_{i=0}^{\infty} \mathbb{E}\left[I_{T_{i,j}}\right] \\ &= \mathbb{E}\left[\sum_{i=0}^{\infty} I_{T_{i,j}}\right] \leq \mathbb{E}\left[C_{[\alpha_j, \beta_j]}\{X_n\}\right] \leq \psi\left(M\left(1+\frac{2}{p}\right), p+2\right) \end{aligned}$$

and therefore

$$\sum_{i=0}^{\infty} \mathbb{P}(Q_M(\varepsilon, a_i, b_i)) = \sum_{i=0}^{\infty} \mathbb{P}(A_i \cap B_i) \leq (p+2) \cdot \psi\left(M\left(1+\frac{2}{p}\right), p+2\right)$$

and the result follows. □

We now present our main result on uniform metastability for stochastic processes.

Theorem 7.1.7. *Let $\{X_n\}$ be a stochastic process with a modulus of L_1-crossings ψ. Let $\omega_M(\varepsilon)$ be defined in terms of ψ as in Proposition 7.1.6. If $\{X_n\}$ has a modulus of tightness $h(\lambda)$ then $\{X_n\}$ converges almost surely with a learnable rate of uniform convergence given by:*

$$\phi(\lambda, \varepsilon) := \frac{2\omega_{h(\lambda/2)}(\varepsilon)}{\lambda}.$$

Proof. We define $Q_M(\varepsilon, n, m)$ as in the proof of Proposition 7.1.6. Fix $\lambda > 0$ and define $M_\lambda := h(\lambda/2)$. By Proposition 7.1.6, for any $\varepsilon > 0$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ we have

$$\sum_{i=0}^{\infty} \mathbb{P}(Q_{M_\lambda}(\varepsilon, a_i, b_i)) \leq \omega_{M_\lambda}(\varepsilon)$$

and so there exists some $n \leq 2\omega_{M_\lambda}(\varepsilon)/\lambda$ such that $\mathbb{P}(Q_{M_\lambda}(\varepsilon, a_n, b_n)) < \lambda/2$ i.e.

$$\mathbb{P}(\exists l, k \in [a_n; b_n](|X_l - X_k| \geq \varepsilon) \cap |X_{a_n}| \leq M_\lambda) < \frac{\lambda}{2}.$$

But, then it follows that

$$\mathbb{P}(\exists k, l \in [a_n; b_n](|X_k - X_l| \geq \varepsilon) + \mathbb{P}(|X_{a_n}| \leq M_\lambda) - 1 < \frac{\lambda}{2}$$

and therefore

$$\begin{aligned}\mathbb{P}(\exists k, l \in [a_n; b_n](|X_k - X_l| \geq \varepsilon) &< \frac{\lambda}{2} - \mathbb{P}(|X_{a_n}| \leq M_\lambda) + 1 \\ &= \frac{\lambda}{2} + \mathbb{P}(|X_{a_n}| > M_\lambda) < \lambda\end{aligned}$$

which completes the proof. □

Remark 7.1.8. It is currently open whether Theorem 7.1.1 can be improved by replacing the modulus of uniform boundedness with a modulus of tightness while still obtaining an explicit modulus of finite fluctuations. Theorem 7.1.7 demonstrates that we can weaken the condition of a modulus of uniform boundedness to a modulus of tightness if we strengthen the condition of a modulus of finite crossings to a modulus of L_1-crossings.

It is certainly the case that we can replace uniform boundedness with the weaker property of tightness to prove that finite crossings imply finite fluctuations. To see this, one can use a pointwise argument, namely that if $C_{[\alpha,\beta]}\{X_n(\omega)\} < \infty$ for all $\alpha < \beta$, then $\{X_n(\omega)\}$ converges to a limit in $\mathbb{R} \cup \{\pm\infty\}$, and therefore

$$X_\infty := \lim_{n\to\infty} X_n$$

exists almost surely in $\mathbb{R}\cup\{\pm\infty\}$. By tightness of $\{X_n\}$ we have that for any $\lambda > 0$ there exists $N \in \mathbb{N}$ such that

$$\mathbb{P}\left(|X_\infty| \geq N\right) = \mathbb{P}\left(\liminf_{n\to\infty} |X_n| \geq N\right) \leq \liminf_{n\to\infty} \mathbb{P}\left(|X_n| \geq N\right) < \lambda$$

by Fatou's lemma, and therefore $|X_\infty| < \infty$ almost surely, which in turn implies that $\{X_n\}$ converges and thus has finite fluctuations almost surely. However, converting this argument into a quantitative one, where one obtains a concrete modulus of finite fluctuations in terms of the corresponding moduli of tightness and finite crossing, is less obvious, as known proofs of Fatou's lemma are nonconstructive.

We now note a straightforward yet useful instance Theorem 7.1.7:

Theorem 7.1.9. *Suppose that the stochastic process $\{X_n\}$ is almost surely monotone and has a modulus of tightness $h(\lambda) \in [1,\infty)$, for all $\lambda > 0$. Then $\{X_n\}$ has a learnable rate of uniform convergence given by:*

$$\phi(\lambda,\varepsilon) := \frac{c}{\lambda\varepsilon} \cdot h\left(\frac{\lambda}{2}\right)$$

for a universal constant $c \leq 22$. In the special case that $\sup_{n\in\mathbb{N}}\|X_n\|_\infty < K$, for some $K \geq 1$, the rate becomes

$$\phi(\lambda,\varepsilon) := \frac{cK}{\lambda\varepsilon}.$$

Proof. Since $\{X_n(\omega)\}$ is almost surely monotone, we have $\mathbb{E}(C_{[\alpha,\beta]}\{X_n\}) \leq 1$ for any $\alpha < \beta$, and thus a modulus of uniformly bounded crossings is given by the constant function $\psi(M,l) = 1$. The result follows from Theorem 7.1.7, noting that in this case

$$\omega_M(\varepsilon) = \left\lceil \frac{8M}{\varepsilon} \right\rceil + 2 \leq \frac{11M}{\varepsilon}$$

and therefore

$$\frac{2\omega_{h(\lambda/2)}(\varepsilon)}{\lambda} \leq \frac{22}{\lambda\varepsilon} \cdot h\left(\frac{\lambda}{2}\right)$$

and we are done. □

Remark 7.1.10. Example 4.2.24 shows that, particularly for the special case $\sup_{n\in\mathbb{N}}\|X_n\|_\infty < K$, the bound on Theorem 7.1.9 above is optimal.

Remark 7.1.11. If $\{X_n\}$ is increasing almost surely and nonnegative, we can reduce the upper bound for the constant in the previous theorem from 22 to 2, even without the assumption that $h(\lambda) \geq 1$, through a more direct approach. Under these assumptions, we have,

$$\begin{aligned}\mathbb{P}\left(\exists i, j \in [a_n; b_n] \, |X_i - X_j| \geq \varepsilon\right) &= \mathbb{P}\left(X_{b_n} - X_{a_n} \geq \varepsilon\right) \\ &\leq \mathbb{P}\left(X_{b_n} - X_{a_n} \geq \varepsilon \cap \sup_{n \in \mathbb{N}} X_n < h(\lambda/2)\right) + \frac{\lambda}{2}.\end{aligned}$$

Setting $Q_n := X_{b_n} - X_{a_n} \geq \varepsilon$ and $R := \sup_{n \in \mathbb{N}} X_n < h(\lambda/2)$ it suffices to show that there exists some $n \leq \phi(\lambda, \varepsilon)$ such that $\mathbb{P}(Q_n \cap R) < \lambda/2$, where now

$$\phi(\lambda, \varepsilon) := \frac{2 \cdot h(\lambda/2)}{\lambda \varepsilon}.$$

If this were not the case, we would have

$$\frac{h(\lambda/2)}{\varepsilon} < \frac{\lambda}{2}\left(\phi(\lambda, \varepsilon) + 1\right) \leq \sum_{i=0}^{\phi(\lambda,\varepsilon)} \mathbb{P}(Q_i \cap R) = \mathbb{E}\left[I_R \sum_{i=0}^{\phi(\lambda,\varepsilon)} I_{Q_i}\right].$$

But for $\omega \in R$ we can have $\omega \in Q_n$ for strictly fewer than $h(\lambda/2)/\varepsilon$ distinct values of $n \in \mathbb{N}$, since whenever $\omega \in Q_{n_1} \cap \ldots \cap Q_{n_k}$ for $n_1 < \ldots < n_k$ then

$$h(\lambda/2) > X_{b_{n_k}}(\omega) \geq \sum_{j=1}^{k} \left(X_{b_{n_j}}(\omega) - X_{a_{n_j}}(\omega)\right) \geq k\varepsilon.$$

Therefore

$$\mathbb{E}\left[I_R \sum_{i=0}^{\infty} I_{Q_i}\right] < \mathbb{E}\left[I_R \cdot \frac{h(\lambda/2)}{\varepsilon}\right] \leq \frac{h(\lambda/2)}{\varepsilon}$$

a contradiction.

We have the following important instantiation of Theorem 7.1.7:

Theorem 7.1.12. *Let $\{X_n\}$ be a stochastic process and $p \in [1, \infty]$. Suppose that $K \geq 1$ is such that $\sup_{n \in \mathbb{N}} ||X_n||_p < K$ and for all $\beta > \alpha$*

$$\mathbb{E}(C_{[\alpha,\beta]}\{X_n\}) \leq \frac{2K}{\beta - \alpha} + 1.$$

Then $\{X_n\}$ has learnable rate of uniform convergence given by:

$$\phi_p(\lambda, \varepsilon) := \frac{cK^2}{\lambda \varepsilon^2} \cdot \left(\frac{2}{\lambda}\right)^{1/p}$$

for $p \in [1, \infty)$ and in the case $p = \infty$ we have a rate given by

$$\phi_\infty(\lambda, \varepsilon) := \frac{cK^2}{\lambda\varepsilon^2}$$

for a universal constant $c \leq 220$ (and independent of p).

Proof. Observing that for any $[\alpha, \beta] \in \mathcal{P}(M, l)$ we have $\beta - \alpha = 2M/l$, it follows that

$$\mathbb{E}(C_{[\alpha,\beta]}\{X_n\}) \leq \frac{lK}{M} + 1$$

and thus $\psi(M, l) := lK/M + 1$ is a modulus of crossings for $\{X_n\}$. Now if $p \in [1, \infty)$ then

$$\mathbb{P}(|X_n| \geq N) \leq \frac{\mathbb{E}(|X_n|^p)}{N^p} < \left(\frac{K}{N}\right)^p$$

and so $h(\lambda) := K\lambda^{-1/p}$ is a modulus of tightness for $\{X_n\}$. Applying Theorem 7.1.7, setting $q := \lceil 8M/\varepsilon \rceil \leq 9M/\varepsilon$ and assuming for now that $M \geq 1$, we have

$$\omega_M(\varepsilon) = \frac{(q+2)^2 K}{M(1+2/q)} + (q+2) \leq \frac{11 \cdot 9 \cdot MK}{\varepsilon^2} + \frac{11M}{\varepsilon} \leq \frac{11 \cdot 10 \cdot MK}{\varepsilon^2}$$

where we use that $K \geq 1$ and $\varepsilon < 1$ to get the last inequality. Instantiating $M := h(\lambda/2) = K(2/\lambda)^{1/p} \geq 1$ we have,

$$\frac{2\omega_{h(\lambda/2)}(\varepsilon)}{\lambda} \leq \frac{2 \cdot 11 \cdot 10K^2}{\lambda\varepsilon^2} \left(\frac{2}{\lambda}\right)^{1/p}$$

from which the main part of the result follows. On the other hand, for $p = \infty$, we have

$$\mathbb{P}(|X_n| \geq K) = 0$$

and thus $h(\lambda) := K$ is a modulus of tightness, so the above calculations can be simplified to give the stated rate. □

7.2 The computational content of Doob's martingale convergence theorem

Doob's martingale convergence theorem is the following central result in the study of stochastic processes, stating that L_1-bounded sub- or supermartingales converge almost surely to a random variable that is finite almost surely.

Here, we present optimal learnable uniform rates (and thus uniform metastable rates, c.f. Remark 4.2.22) for Doob's convergence theorem. Furthermore, we demonstrate how our results

generalise the main result of [21].

To obtain our quantitative results, we need the following result of Doob:

Theorem 7.2.1 (Doob's Upcrossing inequality, c.f. [34]). *Let $\{X_n\}$ be a submartingale, $N \in \mathbb{N}$, and $\alpha < \beta$ be real numbers. We have the following inequality on the expected value of the upcrossing:*

$$\mathbb{E}(U_{N,[\alpha,\beta]}\{X_n\}) \leq \frac{\mathbb{E}((X_N - \alpha)^+)}{\beta - \alpha}.$$

7.2.1 Learnable uniform rates for the martingale convergence theorem

To obtain optimal uniform learnable rates for Doob's convergence theorem, we first consider the case of nonnegative submartingales:

Theorem 7.2.2 (Quantitative positive submartingale convergence theorem). *Let $\{X_n\}$ be a nonnegative submartingale, $p \in [1, \infty]$, and suppose that $K \geq 1$ is such that*

$$\sup_{n \in \mathbb{N}} \|X_n\|_p < K.$$

Then $\{X_n\}$ has learnable rate of uniform convergence given by:

$$\phi_p(\lambda, \varepsilon) := \frac{cK^2}{\lambda \varepsilon^2} \cdot \left(\frac{2}{\lambda}\right)^{1/p}$$

for $p \in [1, \infty)$ and in the case $p = \infty$ we have a rate given by

$$\phi_\infty(\lambda, \varepsilon) := \frac{cK^2}{\lambda \varepsilon^2}$$

for a universal constant $c \leq 220$ (and independent of p).

Proof. Let $\alpha < \beta$. By Theorem 7.2.1 we have, for all $N \in \mathbb{N}$,

$$\mathbb{E}(U_{N,[\alpha,\beta]}\{X_n\}) \leq \frac{\mathbb{E}((X_N - \alpha)^+)}{\beta - \alpha}.$$

Now since $\{X_n\}$ is nonnegative, $U_{N,[\alpha,\beta]}\{X_n\} = 0$ if $\alpha < 0$, and if $\alpha \geq 0$ then $(X_N - \alpha)^+ \leq X_N$ and thus

$$\mathbb{E}(U_{N,[\alpha,\beta]}\{X_n\}) \leq \frac{\mathbb{E}(X_N)}{\beta - \alpha} \leq \frac{\sup_{n \in \mathbb{N}} \mathbb{E}(|X_n|)}{\beta - \alpha} \leq \frac{\sup_{n \in \mathbb{N}} \|X_n\|_p}{\beta - \alpha} < \frac{K}{\beta - \alpha}$$

which implies

$$\mathbb{E}(U_{[\alpha,\beta]}\{X_n\}) \leq \frac{K}{\beta - \alpha}.$$

Therefore Remark 2.3.8 implies that

$$\mathbb{E}(C_{[\alpha,\beta]}\{X_n\}) \leq \frac{2K}{\beta-\alpha} + 1$$

and the result follows form Theorem 7.1.12. □

Doob's upcrossing inequalities hold for general (not necessarily positive) sub- or supermartingales. So, one could adapt the proof of Theorem 7.1.12 directly to obtain learnable rates in those cases. However, in the nonnegative case, we would have to handle upcrossings where $\alpha < 0$, in which we would only have $(X_N - \alpha)^+ \leq |X_N| + |\alpha|$ and accordingly

$$\mathbb{E}(C_{[\alpha,\beta]}\{X_n\}) \leq \frac{2(K+|\alpha|)}{\beta-\alpha} + 1 \leq \frac{l(K+M)}{M} + 1$$

which due to the fact that M dominates K in the subsequent calculation results in a worse bound of $(cK^2/\lambda\varepsilon^2)(2/\lambda)^{2/p}$. This somewhat superficial complication can be circumvented by making use of standard decomposition theorems for martingales, exploiting the fact that learnable rates compose well for sums of stochastic processes as follows:

Lemma 7.2.3. *Let $\{X_n\}$ and $\{Y_n\}$ be stochastic processes with learnable rates of uniform convergence ϕ_1 and ϕ_2 respectively. Then $\{X_n + Y_n\}$ has a learnable rate of uniform convergence given by,*

$$\phi(\lambda,\varepsilon) := \phi_1(\lambda/2,\varepsilon/2) + \phi_2(\lambda/2,\varepsilon/2)$$

Proof. Fixing $\varepsilon, \lambda \in (0,1]$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$, suppose for contradiction that for all $n \leq \phi(\lambda,\varepsilon)$ we have

$$\mathbb{P}(\exists i,j \in [a_n;b_n](|X_i + Y_i - X_j - Y_j| \geq \varepsilon)) \geq \lambda.$$

For any $\omega \in \Omega$, if there exists $i(\omega), j(\omega) \in [a_n;b_n]$ such that $|X_{i(\omega)}(\omega) + Y_{i(\omega)} - X_{j(\omega)} - Y_{j(\omega)}| \geq \varepsilon$, by the triangle inequality we must have either $|X_{i(\omega)} - X_{j(\omega)}| \geq \varepsilon/2$ or $|Y_{i(\omega)} - Y_{j(\omega)}| \geq \varepsilon/2$. In other words, for each $n \leq \phi(\lambda,\varepsilon)$, we have

$$\begin{aligned}\lambda &\leq \mathbb{P}(\exists i,j \in [a_n;b_n](|X_i - X_j| \geq \varepsilon/2) \cup \exists i,j \in [a_n;b_n](|Y_i - Y_j| \geq \varepsilon/2)) \\ &\leq \mathbb{P}(\exists i,j \in [a_n;b_n](|X_i - X_j| \geq \varepsilon/2)) + \mathbb{P}(\exists i,j \in [a_n;b_n](|Y_i - Y_j| \geq \varepsilon/2))\end{aligned}$$

and so again, by the triangle inequality, we have either

$$\mathbb{P}(\exists i,j \in [a_n;b_n](|X_i - X_j| \geq \varepsilon/2)) \geq \lambda/2$$

or

$$\mathbb{P}(\exists i,j \in [a_n;b_n](|Y_i - Y_j| \geq \varepsilon/2)) \geq \lambda/2$$

for each $n \leq \phi(\lambda, \varepsilon)$. But then it follows that either there exists a subsequence $a_{n_0} < b_{n_0} \leq a_{n_1} < b_{n_1} \leq \ldots$ such that

$$\forall k \leq \phi_1(\lambda/2, \varepsilon/2) \left(\mathbb{P}(\exists i, j \in [a_{n_k}; b_{n_k}](|X_i - X_j| \geq \varepsilon/2)) \geq \lambda/2\right)$$

or a subsequence $a_{m_0} < b_{m_0} \leq a_{m_1} < b_{m_1} \leq \ldots$ such that

$$\forall k \leq \phi_2(\lambda/2, \varepsilon/2) \left(\mathbb{P}(\exists i, j \in [a_{m_k}; b_{m_k}](|Y_i - Y_j| \geq \varepsilon/2)) \geq \lambda/2\right)$$

which contradict the defining property of ϕ_1 and ϕ_2 respectively. □

We can now obtain learnable uniform rates for Doob's convergence theorem.

Theorem 7.2.4 (Quantitative Doob's convergence theorem)**.** *Let $\{X_n\}$ be a sub- or super martingale and suppose that $K \geq 1$ is such that*

$$\sup_{n \in \mathbb{N}} \mathbb{E}(|X_n|) < K.$$

Then $\{X_n\}$ has learnable rate of uniform convergence given by:

$$\phi(\lambda, \varepsilon) := c \left(\frac{K}{\lambda \varepsilon}\right)^2$$

for a universal constants $c \leq 2^{11} \cdot 3^2 \cdot c_1$ for $c_1 > 0$ as in Theorem 7.2.2.

Proof. We can assume WLOG that $\{X_n\}$ is a submartingale, since if $\{X_n\}$ is a supermartingale, then $\{-X_n\}$ is a submartingale which must converge with the same learnable rate. Let $X_n = M_n + A_n$ be the Doob decomposition in this case, i.e.

$$M_n := X_0 + \sum_{i=1}^{n} (X_i - \mathbb{E}(X_i \mid \mathcal{F}_{i-1})) \quad \text{and} \quad A_n := \sum_{i=1}^{n} (\mathbb{E}(X_i \mid \mathcal{F}_{i-1}) - X_{i-1}).$$

where it is easy to show that $\{M_n\}$ is a martingale and $\{A_n\}$ is almost surely nonnegative and increasing. Then by Lemma 7.2.3 $\{X_n\}$ has learnable rate of uniform convergence $\phi_1(\lambda/2, \varepsilon/2) + \phi_2(\lambda/2, \varepsilon/2)$ where ϕ_1 and ϕ_2 are learnable rates for $\{M_n\}$ and $\{A_n\}$ respectively.

Since $\{A_n\}$ is almost surely monotone and $\mathbb{E}(|A_n|) = \mathbb{E}(A_n) = \mathbb{E}(X_n - X_0) \leq \mathbb{E}(|X_n|) + \mathbb{E}(|X_0|) < 2K$ and so has modulus of tightness $2K/\lambda$, by Remark 7.1.11, we can define

$$\phi_2(\lambda, \varepsilon) := \frac{8K}{\lambda^2 \varepsilon}.$$

On the other hand, since $\{M_n\}$ is a martingale and $\mathbb{E}(|M_n|) = \mathbb{E}(|X_n - A_n|) \leq \mathbb{E}(|X_n|) + \mathbb{E}(|A_n|) < 3K$ for all $n \in \mathbb{N}$, we can write $M_n = M_n^+ - M_n^-$ where $\{M_n^+\}$ and $\{M_n^-\}$ are

both nonnegative submartingales (since $x \mapsto x^+$ and $x \mapsto x^-$ are convex functions, and any convex function applied to a martingale results in a submartingale), and have means uniformly bounded by $3K$. So, by Lemma 7.2.3 and Theorem 7.2.2, noting that a learnable rate of uniform convergence for $\{M_n^-\}$ is also one for $\{-M_n^-\}$, we can define, for $c_1 \leq 220$,

$$\phi_1(\lambda, \varepsilon) := 2\left(\frac{c_1(3K)^2}{(\lambda/2)(\varepsilon/2)^2}\left(\frac{4}{\lambda}\right)\right) = \frac{2^6 \cdot 3^2 \cdot c_1 K^2}{\lambda^2 \varepsilon^2}$$

and therefore

$$\phi_1(\lambda/2, \varepsilon/2) + \phi_2(\lambda/2, \varepsilon/2) \leq \frac{2^{10} \cdot 3^2 \cdot c_1 K^2}{\lambda^2 \varepsilon^2} + \frac{2^6 K}{\lambda^2 \varepsilon} \leq \frac{2^{11} \cdot 3^2 \cdot c_1 K^2}{\lambda^2 \varepsilon^2}$$

and the main result follows directly. □

7.2.2 Bounds on the fluctuations for martingales and optimality of rates

As mentioned in Remark 4.2.23, the precise relationship between learnable pointwise rates and moduli of finite fluctuations is open for general stochastic processes. However, in the case of martingales with uniformly bounded L_1 norm, these notions of convergence coincide.

Proposition 7.2.5. *Suppose ϕ is a pointwise learnable rate of convergence for all (super)submartingales $\{X_n\}$ with $\sup_n \mathbb{E}(|X_n|) \leq L$ for some $L > 0$. Then, ϕ is a modulus of finite fluctuations for all such stochastic processes.*

Proof. Fix $\varepsilon, \lambda \in (0, 1]$ and $N \in \mathbb{N}$. Define the stopping times $\{\tau_n\}$ as follows:

- $\tau_0 :\equiv 0$
- $\tau_j := \inf\{\tau_{j-1} \leq i \leq N : |X_{\tau_{j-1}} - X_i| \geq \varepsilon\}$
- $\tau_j := N$ if the above does not exist.

We have,

$$\mathbb{P}(J_{N,\varepsilon}\{X_n\} \geq \phi(\lambda, \varepsilon)) \leq \mathbb{P}(\forall i \leq \phi(\lambda, \varepsilon)\, |X_{\tau_i} - X_{\tau_{i+1}}| \geq \varepsilon).$$

Now, we have $\{X_{\tau_n}\}$ must be a (super)submrtingale by Theorem 2.4.24, with $\sup_n \mathbb{E}(|X_{\tau_n}|) \leq L$. Therefore, taking $a_i = i$ and $b_i := i + 1$ we must have

$$\mathbb{P}(\forall i \leq \phi(\lambda, \varepsilon)\, (|X_{\tau_i} - X_{\tau_{i+1}}| \geq \varepsilon)) = \mathbb{P}(\forall i \leq \phi(\lambda, \varepsilon)\, \exists k, l \in [i, i+1]\, (|X_{\tau_k} - X_{\tau_l}| \geq \varepsilon)) < \lambda.$$

and the result follows. □

From the above, we immediately get bounds on the fluctuations for Doob's convergence theorem:

Theorem 7.2.6 (Fluctuations in Doob's convergence theorem)**.** *Let $\{X_n\}$ be a sub- or super martingale and suppose that $K \geq 1$ is such that*

$$\sup_{n\in\mathbb{N}} \mathbb{E}(|X_n|) < K.$$

Then $\{X_n\}$ has a modulus of finite fluctuations given by:

$$\phi(\lambda, \varepsilon) := c\left(\frac{K}{\lambda\varepsilon}\right)^2$$

for a universal constants $c \leq 2^{11} \cdot 3^2 \cdot c_1$ for $c_1 > 0$ as in Theorem 7.2.2.

Proof. The result follows from Proposition 7.2.5, Theorem 7.2.4, and the fact that a learnable uniform rate of convergence is a learnable pointwise rate of convergence. □

Remark 7.2.7. A result of Chashka ([21] but see also [70, Theorem 28]) asserts that if $\{X_n\}$ is a martingale with $\sup_{n\in\mathbb{N}}\|X_n\|_1 \leq K$ then

$$\mathbb{P}(J_\varepsilon\{X_n\} \geq N) \leq \frac{CK}{N^{1/2}\varepsilon} \tag{7.2}$$

for some constant C. The above bound corresponds to the following modulus of finite fluctuations:

$$\phi(\lambda, \varepsilon) := \left(\frac{CK}{\lambda\varepsilon}\right)^2. \tag{7.3}$$

Thus, Theorem 7.2.6 provides a different proof of Chaska's result and extends the bound to submartingales and supermartingales.

In [21], Chashka provides a construction demonstrating that the bound in (7.2) is optimal. Therefore, Proposition 7.2.5 implies that the bound in Theorem 7.2.4 is optimal.

We provide a modification of Chaska's optimality construction demonstrating that the moduli of finite fluctuations given in (7.3) corresponds to an optimal pointwise learnable rate of convergence:

Example 7.2.8. Let us fix $\varepsilon, \lambda > 0$ and $p \in [1, \infty)$, where for simplicity we assume that $\varepsilon = 2^{-M}$ and $\lambda = 2^{-N}$ for some $M, N \in \mathbb{N}$. We define a stochastic process $\{X_n\}$ on the standard space $([0, 1], \mathcal{F}, \mu)$ and in terms of these parameters as follows: First, define the Rademacher functions $r_n : [0, 1] \to \mathbb{R}$ satisfying $r(x) = sgn(sin(2^{n+1}x\pi))$ for each $x \in [0, 1]$. Let

$$W := \left\lfloor \frac{2^{2M+1}2^{\frac{2N}{p}}}{3} \right\rfloor = \left\lfloor \frac{2}{3\varepsilon^2\lambda^2} \right\rfloor$$

. Let X_0 be the constant 0 random variable. For $n \leq W$ let

$$X_n(x) := \begin{cases} \sum_{i=0}^{n-1} 2^{-M} r_{i+N}(x) & \text{if } x \in [0, 2^{-N}) \\ \\ 0 & \text{otherwise.} \end{cases}$$

Let $X_n = X_W$ for $n > W$. As in the case of Chashka, one can easily check that $\{X_n\}$ will be a martingale with respect to the filtration $\{\mathcal{F}_n\}$, where $\mathcal{F}_n$ is the σ algebra generated by

$$\left\{ \left[\frac{i}{2^{n+N}}, \frac{i+1}{2^{n+N}} \right) : i \in \{0, .., 2^{n+N} - 1\} \right\}.$$

Furthermore, for each $n < W$ we have

$$\mathbb{P}(|X_n - X_{n+1}| \geq \varepsilon) = \lambda$$

and for $n \leq W$ (again arguing as Chashka does),

$$\mathbb{E}(|X_n|) = \int_0^{2^{-N}} \left| \sum_{i=0}^{n-1} 2^{-M} r_{i+N}(x) \right| dx = 2^{-N} \int_0^1 \left| \sum_{i=0}^{n-1} 2^{-M} r_i(x) \right| dx \leq$$
$$2^{-N} \left(\frac{3}{2} \right)^{\frac{1}{2}} \left(n 2^{-2M} \right)^{\frac{1}{2}}.$$

With the last inequality, a standard result for the Rademacher functions (see theorem 8.14 of [54]). So we have

$$\sup_{n \in \mathbb{N}} \mathbb{E}(|X_n|) \leq 2^{-N} \left(\frac{3}{2} \right)^{\frac{1}{2}} \left(W 2^{-2M} \right)^{\frac{1}{2}} \leq 1.$$

As in the case of monotone bounded sequences, suppose that some $\psi(\lambda, \varepsilon)$ is a learnable rate of *pointwise* convergence for all martingales whose first moment is uniformly bounded by 1. Then we must have

$$\left\lfloor \frac{2}{3\varepsilon^2 \lambda^2} \right\rfloor \leq \psi(\lambda, \varepsilon)$$

for all $\varepsilon, \lambda > 0$. Otherwise, defining $\{X_n\}$ in terms of fixed $\varepsilon, \lambda > 0$ as above, we get

$$\forall n \leq \psi(\lambda, \varepsilon) \mathbb{P}(\exists i, j \in [n, n+1](|X_i - X_j| \geq \varepsilon) \geq \lambda$$

contradicting that $\psi(\lambda, \varepsilon)$ is a learnable rate of *pointwise* convergence.

7.3 The computational content of Birkoff's pointwise ergodic theorem

Throughout this section, fix a measure preserving transformation $\tau : \Omega \to \Omega$ of our probability space $X := (\Omega, \mathcal{F}, \mathbb{P})$ and define the Koopman operator $T : L_1(X) \to L_1(X)$ by $Tf := f \circ \tau$. For $f \in L_1(X)$ we define

$$S_n f := \sum_{k=0}^{n-1} T^k f \quad \text{and} \quad A_n f := \frac{S_n f}{n}.$$

The Birkhoff pointwise ergodic theorem states that the ergodic averages $A_n f$ converge almost surely.

In this section, we shall produce uniform learnable rates of convergence for this theorem and improvements to the best-known moduli of finite fluctuations in the literature. As with Doob's convergence theorem, our quantitative results will follow from uppcrossing inequalities concerning such ergodic averages. The first such inequality is due to Bishop:

Theorem 7.3.1 (Bishop's upcrossing inequality [18])**.** *With the notation as above, for reals $\alpha < \beta$ we have the following inequality:*

$$\mathbb{E}(U_{[\alpha,\beta]}\{A_n f\}) \leq \frac{\mathbb{E}\left((f-\alpha)^+\right)}{\beta - \alpha}.$$

The second inequality, which we shall use to obtain our improvement on known moduli of finite fluctuations, is due to Ivanov:

Theorem 7.3.2 (Ivanov's downcrossing inequality [65])**.** *With the notation as above, for reals $0 < \alpha < \beta$ and $k > 0$,we have the following inequality:*

$$\mathbb{P}\left(D_{[\alpha,\beta]}\{A_n f\} \geq k\right) \leq \left(\frac{\alpha}{\beta}\right)^k.$$

7.3.1 Learnable uniform rates for Birkoff's pointwise ergodic theorem

We can now obtain uniform learnable rates for the pointwise ergodic theorem, as we did for Doob's convergence theorem. We first consider the nonnegative case, then through a decomposition and an application of Lemma 7.2.3, we get the general result.

Theorem 7.3.3 (Quantitative pointwise ergodic theorem)**.** *Let $p \in [1, \infty]$, and suppose that $K \geq 1$ is such that*

$$\|f\|_p < K$$

Then $\{A_n f\}$ has learnable rate of uniform convergence given by:

$$\phi_p(\lambda, \varepsilon) := \frac{16cK^2}{\lambda\varepsilon^2} \cdot \left(\frac{4}{\lambda}\right)^{1/p}$$

for $p \in [1, \infty)$, and in the case $p = \infty$ we have a rate given by

$$\phi_\infty(\lambda, \varepsilon) := \frac{16cK^2}{\lambda\varepsilon^2}$$

for a universal constant $c \leq 220$ (and independent of p).

Proof. First, assume the f is nonnegative. Since $\sup_{n\in\mathbb{N}} \mathbb{E}(|A_n f|) \leq \mathbb{E}(f) \leq \|f\|_p < K$, arguing as in Theorem 7.2.2, we have for each $\alpha > \beta$

$$\mathbb{E}(U_{[\alpha,\beta]}\{A_n f\}) \leq \frac{K}{\beta - \alpha}.$$

Therefore Remark 2.3.8 implies that

$$\mathbb{E}(C_{[\alpha,\beta]}\{A_n f\}) \leq \frac{2K}{\beta - \alpha} + 1.$$

and we obtain learnable uniform rates from Theorem 7.1.12. For the general case, we decomposing $f = f^+ - f^-$ where $f^+ := \max\{f, 0\}$ and $f^- := \max\{-f, 0\}$ are the positive and negative parts of f respectfully (noting $A_n(f) = A_n(f^+) - A_n(f^-)$), and apply Proposition 7.2.3. □

Remark 7.3.4. In [6], by assuming $f \in L_2(X)$, metastable rates of uniform convergence for $\{A_n f\}$ are given through a proof-theoretic analysis of a proof of Billingsley [16], and for bounded f they provide such rates through a more straightforward analysis of the relevant upcrossing inequality than that provided in Section 7.1.2. A comparison of the rates obtained through each approach is given (where, roughly speaking, uniform metastable rates obtained through the proof of Billingsley require asymptotically fewer iterations of a faster-growing function).

In the case $p = \infty$, in the previous theorem, we obtain (up to a constant) the same corresponding metastable uniform rate as in [6] for bounded f. However, Theorem 7.3.3 provides metastable uniform rates for $f \in L_1(X)$, which is the assumption required for the pointwise ergodic theorem. Hence, the aforementioned theorem generalises [6] and represents the first such rates for the full pointwise ergodic theorem.

7.3.2 Bounds on the fluctuations of ergodic averages

We now investigate moduli of finite fluctuations for the ergodic averages in the pointwise ergodic theorem.

It is shown by Kachorovskii in [70, Theorem 23] (though the result is attributed to Ivanov) that the following bound on the probabilistic fluctuations can be given:

$$\mathbb{P}\left(J_\varepsilon\{A_n f\} \geq a\right) < c\sqrt{\frac{\log(a)}{a}} \tag{7.4}$$

for all $f \in L_1(X)$ and $\varepsilon, a > 0$, with $c > 0$ a constant that depends on $\mathbb{E}(|f|)/\varepsilon$.

This result can be improved through Corollary 7.1.4. As in [70], we first assume $f \geq 0$. By the maximal ergodic theorem, we have for all $a > 0$,

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} |A_n f| \geq a\right) \leq \frac{\mathbb{E}(|f|)}{a},$$

so for any $S > \mathbb{E}(|f|)$ we can take $g(a) := S/a$ in Corollary 7.1.4. For crossings, we use the well-known result of Ivanov [65] which states that for $0 < \alpha < \beta$ and $k > 0$,

$$\mathbb{P}\left(D_{[\alpha,\beta]}\{A_n f\} \geq k\right) \leq \left(\frac{\alpha}{\beta}\right)^k$$

where $D_{[\alpha,\beta]}\{A_n f\}$ denotes the number of downcrossings of $[\alpha, \beta]$ made by $\{A_n f\}$. Thus, we have (by Remark 2.3.8)

$$\mathbb{P}\left(C_{[\alpha,\beta]}\{A_n f\} \geq k\right) < \left(\frac{\alpha}{\beta}\right)^{\frac{k-1}{4}}$$

(we divide the exponent by another factor of 2 to obtain a strict inequality), so we can take $h_{\alpha,\beta}(k) = (\alpha/\beta)^{\frac{k-1}{4}}$ in Corollary 7.1.4. We can restrict our attention $0 < \alpha < \beta$ (i.e. the situation $\mathbb{P}\left(C_{[\alpha,\beta]}\{A_n f\} > 0\right) > 0$), and in this case

$$h^{-1}_{\alpha,\beta}(\lambda) = \frac{4\log(1/\lambda)}{\log(\beta) - \log(\alpha)} + 1 \leq \frac{4\beta\log(1/\lambda)}{\beta - \alpha} + 1$$

where for the last step, we use

$$\log(\beta) - \log(\alpha) \geq \frac{\beta - \alpha}{\beta},$$

which follows from the mean value theorem. Thus for any $M, l > 0$ and $[\alpha, \beta] \in \mathcal{P}(M, l)$ with $0 < \alpha < \beta$, we have

$$h^{-1}_{\alpha,\beta}\left(\frac{\lambda}{l}\right) \leq \frac{4\beta\log(l/\lambda)}{\beta - \alpha} + 1 = 2l \cdot \log\left(\frac{l}{\lambda}\right) + 1 \leq 2l \cdot \log\left(\frac{2l}{\lambda}\right) =: H(\lambda, M, l)$$

and so the right-hand side defines a suitable bounding function H. We then observe that for

$$l = \left\lceil \frac{4g^{-1}(\lambda/2)}{\varepsilon} \right\rceil \leq \frac{9S}{\lambda\varepsilon}$$

it follows that

$$l \cdot H\left(\frac{\lambda}{2}, g^{-1}\left(\frac{\lambda}{2}\right), l\right) = 2l^2 \cdot \log\left(\frac{4l}{\lambda}\right) \leq c\left(\frac{S}{\lambda\varepsilon}\right)^2 \cdot \log\left(\frac{cS}{\lambda^2\varepsilon}\right) =: G_\varepsilon(\lambda)$$

for suitable constant $c \leq 200$. It remains to find the inverse of the function G_ε defined above. To this end, suppose that

$$a = c\left(\frac{S}{\lambda\varepsilon}\right)^2 \cdot \log\left(\frac{cS}{\lambda^2\varepsilon}\right).$$

Rearranging we obtain

$$\exp\left[\frac{a}{c}\left(\frac{\lambda\varepsilon}{S}\right)^2\right] = \frac{cS}{\lambda^2\varepsilon}.$$

Now letting $E(x) := x\exp(x)$ we have

$$E\left[\frac{a}{c}\left(\frac{\lambda\varepsilon}{S}\right)^2\right] = \frac{a}{c}\left(\frac{\lambda\varepsilon}{S}\right)^2 \cdot \frac{cS}{\lambda^2\varepsilon} = \frac{aS}{\varepsilon}$$

and therefore

$$\lambda = \frac{S}{\varepsilon} \cdot \sqrt{\frac{c}{a} \cdot W\left(\frac{aS}{\varepsilon}\right)}.$$

Where, W is the inverse of the E (i.e. the Lambert W-function), and so by Corollary 7.1.4 we have

$$\mathbb{P}\left(J_\varepsilon\{A_n f\} \geq a\right) < c'\sqrt{\frac{W\left(c'a\right)}{a}}$$

for $c' := S\sqrt{c}/\varepsilon$.

For general f, not assumed to be nonnegative (for different constant c_0), we can decompose f as the difference of two positive terms $f = f^+ - f^-$. Furthermore, we will have $\mathbb{E}(|f^\pm|) < S$, $A_n f = A_n(f^+) - A_n(f^-)$ and

$$\mathbb{P}\left(J_{\varepsilon/2}\{A_n f\} \geq a\right) \leq \mathbb{P}\left(J_{\varepsilon/2}\{A_n(f^+)\} \geq a/2\right) + \mathbb{P}\left(J_\varepsilon\{A_n(f^-)\} \geq a/2\right).$$

This leads to the following:

Theorem 7.3.5. *There exists a universal constant $c_0 > 0$ such that, for all $\varepsilon, a > 0$ and $S > 0$*

satisfying $\mathbb{E}(|f|) < S$ *we have*

$$\mathbb{P}\left(J_\varepsilon\{A_n f\} \geq a\right) < \frac{c_0 S}{\varepsilon}\sqrt{\frac{W\left(c_0 S a/\varepsilon\right)}{a}}.$$

The above implies Ivanov's bound (7.4) and improves it slightly in that $W(c_0 a) < \log(a)$ for $c_0 < \log(a)$.

Remark 7.3.6. The following bound is a conjecture attributed to Ivanov [70, Conjecture 5]:

$$\mathbb{P}\left(J_\varepsilon\{A_n f\} \geq a\right) < c\sqrt{\frac{1}{a}} \tag{7.5}$$

for all $f \in L_1(X)$ and $\varepsilon, a > 0$, with $c > 0$ a constant that depends on $\mathbb{E}(|f|)/\varepsilon$.

The above bound corresponds to a modulus of finite fluctuations given by

$$\phi(\lambda, \varepsilon) := \frac{c^2}{\lambda^2},$$

which is the precise form of the learnable uniform rate we presented in Theorem 7.3.3 for the $p = 1$ case. The exact relationship between learnable uniform rates and moduli of finite fluctuations is currently unknown. However, proving that any learnable uniform rate is a modulus of finite fluctuations would close Ivanov's Conjecture.

7.3.3 Rates via variational inequalities

We conclude this chapter by discussing how the quantitative results we have obtained in this chapter relate to some results in the ever-growing body of work championed by Jones and collaborators on investigating fluctuations in ergodic averages.

Let $\{x_n\}$ be a sequence of elements in an arbitrary normed space $(X, \|\cdot\|)$. A sufficient (but not necessary) condition for $\{x_n\}$ to be Cauchy is that there exists $Q, C > 0$ such that,

$$\sup_{\{n_k\}} \sum_{k=1}^{\infty} \|x_{n_{k+1}} - x_{n_k}\|^Q < C \tag{7.6}$$

where the above supremum is taken over sequences of indices $n_1 < n_2 < \ldots$. Furthermore, it is clear that if the above holds $J_\varepsilon\{x_n\} \leq C/\varepsilon^Q$ (where $J_\varepsilon\{x_n\}$ is the obvious generalisation of fluctuations to arbitrary normed spaces). Such a result was shown to hold for the mean ergodic theorem for nonexpansive maps on Hilbert space by Jones, Ostrovskii, and Rosenblatt [68] with later results concerning L_p averages obtained in [67] and [7], with the latter reference obtaining results on uniformly convex Banach spaces.

Such oscillation results have also been obtained for the pointwise ergodic theorem. If $\{X_n\}$

is a stochastic process, the probabilistic analogue of the above notion is

$$\sup_{\{n_k\}} \sum_{k=1}^{\infty} |X_{n_{k+1}} - X_{n_k}|^Q < \infty \text{ almost surely}$$

for some $Q > 0$. In light of Lemma 4.2.2, the above can be given a computational interpretation via a function $\phi(\lambda)$ satisfying

$$\mathbb{P}\left(\sup_{\{n_k\}} \sum_{k=1}^{\infty} |X_{n_{k+1}} - X_{n_k}|^Q \geq \phi(\lambda) \right) < \lambda \tag{7.7}$$

for all $\lambda > 0$. If we have such a $\phi(\lambda)$ one can easily show that $\psi(\lambda, \varepsilon) = \phi(\lambda)/\varepsilon^Q$ is a modulus of finite fluctuations for $\{X_n\}$.

In [67, Theorem B] it is shown that for $Q > 2$ there exists $C > 0$ such that for all $f \in L_1$ and $a > 0$

$$\mathbb{P}\left(\sup_{\{n_k\}} \left(\sum_{k=1}^{\infty} |A_{n_{k+1}} f - A_{n_k} f|^Q \right)^{1/Q} \geq a \right) \leq \frac{C\|f\|_1}{a}.$$

Or in the language of [67], the variational norm operator, V_Q, is weak type $(1,1)$. This implies we can find a modulus satisfying (7.7), and we further have the following bound on the fluctuations

$$\mathbb{P}\left(J_\varepsilon\{A_n f\} \geq a \right) \leq \frac{C\|f\|_1}{a^{1/Q}\varepsilon}.$$

Setting $Q = 2$ in the above expression would yield a bound conjectured by Ivanov [70, Conjecture 5]. However, the above only holds for $Q > 2$ and for all such Q, this results in a bound worse than that given by Ivanov [70, Theorem 23] and our improvement given in Theorem 7.3.3. Furthermore, it is unclear that such a result immediately gives learnable uniform rates. Nevertheless, such optimal rates are obtained in this Chapter.

In addition, in [67] it is shown that there exists $C > 0$ such that for all sequences of indices $n_1 < n_2 < \ldots$, $f \in L_1$, and $a > 0$

$$\mathbb{P}\left(\left(\sum_{k=1}^{\infty} \sup_{n_k \leq u \leq v < n_{k+1}} |A_u f - A_v f|^2 \right)^{1/2} \geq a \right) \leq \frac{C\|f\|_1}{a}.$$

Such a bound does not naturally give moduli of finite fluctuations. However, one can obtain learnable pointwise rates of almost sure convergence, as follows:

suppose $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ and ε are given. Then

$$\begin{aligned}&\mathbb{P}\left(\forall i \leq e \exists k, l \in [a_i, b_j] | A_k f - A_l f| \geq \varepsilon\right)\\ &\leq \mathbb{P}\left(\left(\sum_{k=1}^{\infty} \sup_{n_k \leq u \leq v < n_{k+1}} |A_u f - A_v f|^2\right)^{1/2} \geq \varepsilon \sqrt{\frac{e}{3}}\right) \leq \frac{C\|f\|_1}{a},\end{aligned}$$

where $n_{2k+1} = a_{3k}$ and $n_{2k+2} := b_{3k} + 1$. This yields that,

$$e(\lambda, \varepsilon) := \frac{3C^2\|f\|_1^2}{\varepsilon^2 \lambda^2}$$

is a pointwise learnable rate of almost sure convergence.

It is worth noting that the paper [67] appeared without the authors knowing of the work of [70, 65, 21]. Therefore, a few of the results in [67] were already known by the aforementioned authors (although [67] establish these results through very different methods); in particular, all the results of the final subsection of [67] were already obtained by Chaska in [21]. This is partially addressed by Jones, Rosenblatt and Wierdl in [69] (c.f. Remark 3.2 of [69]).

Chapter 8

The computational content of the Robbins-Siegmund theorem

The *Robbins-Siegmund theorem* is a convergence result for stochastic processes that has become fundamental in stochastic optimization. The theorem is the following:

Theorem 8.0.1 (Robbins-Siegmund [129])**.** *Let $\{X_n\}$, $\{A_n\}$, $\{B_n\}$ and $\{C_n\}$ be sequences of nonnegative integrable random variables on some arbitrary probability space and adapted to the filtration $\{\mathcal{F}_n\}$, with $\sum_{i=0}^{\infty} A_i < \infty$ and $\sum_{i=0}^{\infty} C_i < \infty$ almost surely and*

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq (1 + A_n)X_n - B_n + C_n$$

for all $n \in \mathbb{N}$. Then, almost surely, $\{X_n\}$ converges and $\sum_{i=0}^{\infty} B_i < \infty$.

In Chapter 3, we discussed the important role recursive inequalities play in establishing the convergence of iterative algorithms that appear in a variety of contexts in analysis. As noted in the survey paper [42], in a similar manner to the deterministic case, it is a common trend in stochastic analysis to use the Robbins-Siegmund theorem to establish the convergence of iterative stochastic algorithms.

Due to its fundamental nature in the area of stochastic optimization, to make progress in mining proofs in this area, obtaining a computational analogue of the Robbins-Siegmund theorem is of the utmost importance. Such a computational interpretation was obtained by the author, in collaboration with Powell [116], and it is the purpose of this chapter to present this result along with some applications.

The Robbins-Siegmund theorem can be seen as a generalisation of Doob's martingale convergence theorem for nonnegative supermartingales (indeed, we just take $A_n = B_n = C_n = 0$ in Theorem 8.0.1). Furthermore, Doob's convergence result is crucial in establishing the Robbins-Siegmund theorem. Thus, the quantitative result we obtained in collaboration with Powell

[117], which we present in Chapter 7, was a key stepping stone in establishing the quantitative Robbins-Siegmund theorem we present in this chapter.

This chapter will begin in Section 8.1, where we give a quantitative version of a result of Qihou [126], which can be seen as the nonstochastic version of the Robbins-Siegmund Theorem. The presentation of this result will serve two purposes. Firstly, analysing this deterministic result will serve as motivation for our approach to obtaining a quantitative version of the Robbins-Siegmund theorem since the strategy in obtaining our quantitative version of Qihou's result will mirror our approach to tackling the Robbins-Siegmund theorem. Secondly, we will need a quantitative version of Qihou's result when we discuss applications of the Robbins-Siegmund theorem later in the chapter.

Then, in Section 8.2, we give our quantitative version of the Robbins-Siegmund theorem. We start this section by investigating how uniform learnable rates and rates of uniform boundedness for random variables combine under arithmetic operations. We then use the preliminary lemmas and our computational interpretation of the martingale convergence theorem from Chapter 7 to obtain a quantitative version of the Robbins-Siegmund theorem.

We conclude this chapter, in Section 8.3, by discussing some applications of the Robbins-Siegmund theorem. As the Robbins-Siegmund theorem is a generalisation of the monotone convergence theorem, one can use Example 2.3.3 to demonstrate that general rates of almost sure convergence cannot be obtained for the convergence in the conclusion of the theorem. In this section, we investigate how imposing additional conditions on the result allows one to obtain rates of almost sure convergence. We use this general result to obtain rates of almost sure convergence for Kolmogorov's Strong Law of Large Numbers (although our rates are worse than those discussed in Chapter 6) and the celebrated result in Stochastic approximations of Dvoretsky [36]. Lastly, we obtain uniform metastable rates for the Robbins-Monro procedure [128], which is another celebrated result in stochastic approximation. The rates for the Robbins-Monro procedure were already sketched in [116]; however, the remaining applications, including the general condition that can be imposed on the Robbins-Siegmund theorem to allow rates of almost sure convergence, are new. Jointly with Powell and Pischke, the author is currently working to generalise these results.

8.1 The non-stochastic case

Consider the following result on sequences of real numbers:

Theorem 8.1.1 (Lemma 5.31 of [12]). *Let $\{x_n\}$, $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\gamma_n\}$ be sequences of non-negative reals with $\sum_{i=0}^{\infty} \alpha_i < \infty$ and $\sum_{i=0}^{\infty} \gamma_i < \infty$ such that*

$$x_{n+1} \leq (1 + \alpha_n)x_n - \beta_n + \gamma_n$$

for all $n \in \mathbb{N}$. *Then* $\{x_n\}$ *converges and* $\sum_{i=0}^{\infty} \beta_i < \infty$.

This result is essentially the deterministic version of the Robbin-Siegmund theorem. It represents a crucial lemma for numerous convergence proofs in fixed-point theory (see, e.g. [126] for the special case $\beta_n = 0$, and [12, 42] for further examples and references).

We shall give a computational interpretation for this theorem by computing bounds on the fluctuations of the converging sequence in the conclusion of the result (and a bound for the converging sum). A rate of metastability, of a form that corresponds to a fluctuation bound (c.f. Theorem 2.3.16), has already been obtained for this result by Kohlenbach and Lambov in [85, Lemma 16]. We deliberately analyse a slightly different proof of the result, which corresponds better to our approach to the Robbins-Siegmund theorem. Furthermore, our computational interpretation of this result will be needed when we discuss applications of the Robbins-Siegmund theorem in Section 8.3.

We first need elementary results on bounds on the fluctuations for sums and products of sequences of real numbers:

Lemma 8.1.2. *Suppose that* $\{x_n\}$ *and* $\{y_n\}$ *converge with bounds on their fluctuations given by* ϕ *and* ψ *respectively. Then:*

(a) A bound on the fluctuations of $\{x_n y_n\}$ *is given by*

$$\Delta(\varepsilon) := \phi(\varepsilon/2L) + \psi(\varepsilon/2K)$$

where $\sup_{n\in\mathbb{N}} |x_n| < K$ *and* $\sup_{n\in\mathbb{N}} |y_n| < L$.

(b) A bound on the fluctuations of $\{x_n + y_n\}$ *is given by*

$$\Delta(\varepsilon) := \phi(\varepsilon/2) + \psi(\varepsilon/2)$$

Proof. For part (a), assume for contradiction that there exists some $\varepsilon > 0$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ such that for all $n \leq \Delta(\varepsilon)$:

$$\exists i, j \in [a_n; b_n] \left(|x_i y_i - x_j y_j| \geq \varepsilon\right).$$

Then for each $n \leq \Delta(\varepsilon)$ we have

$$\exists i, j \in [a_n; b_n] \left(|x_i - x_j| \geq \varepsilon/2L\right) \quad \text{or} \quad \exists i, j \in [a_n; b_n] \left(|y_i - y_j| \geq \varepsilon/2K\right)$$

since

$$\varepsilon \leq |x_i y_i - x_j y_j| \leq |y_i||x_i - x_j| + |x_j||y_i - y_j| < L|x_i - x_j| + K|y_i - y_j|.$$

Form this we see that, there exists a subsequence $a_{n_0} < b_{n_0} \leq a_{n_1} < b_{n_1} \leq \ldots$ such that

$$\forall k \leq \phi(\varepsilon/2L)\, \exists i, j \in [a_{n_k}; b_{n_k}]\, (|x_i - x_j| \geq \varepsilon/2L)$$

or a subsequence $a_{m_0} < b_{m_0} \leq a_{m_1} < b_{m_1} \leq \ldots$ such that

$$\forall k \leq \psi(\varepsilon/2K)\, \exists i, j \in [a_{m_k}; b_{m_k}]\, (|y_i - y_j| \geq \varepsilon/2K)$$

contradicting the defining property of ϕ or ψ. Part (b) is proven similarly. □

We now give our quantitative nonstochastic Robbins-Siegmund theorem, where for simplicity, we replace the assumption $\sum_{i=0}^{\infty} \alpha_i < \infty$ with the equivalent property $\prod_{i=0}^{\infty}(1 + \alpha_i) < \infty$.

Theorem 8.1.3. *Let $\{x_n\}$, $\{\alpha_n\}$, $\{\beta_n\}$ and $\{\gamma_n\}$ be sequences of nonnegative reals with*

$$x_{n+1} \leq (1 + \alpha_n)x_n - \beta_n + \gamma_n$$

for all $n \in \mathbb{N}$. Suppose that $K, L, M > 0$ satisfy $x_0 < K$, $\prod_{i=0}^{\infty}(1 + \alpha_i) < L$ and $\sum_{i=0}^{\infty} \gamma_i < M$. Then

$$\phi(\varepsilon) := \frac{8L(K + M)}{\varepsilon}$$

is a bound on the fluctuations for $\{x_n\}$ and

$$\sum_{i=0}^{\infty} \beta_i < L(K + M)$$

Proof. Define

$$p_n := \prod_{i=0}^{n-1}(1 + \alpha_i), \quad \tilde{x}_n := \frac{x_n}{p_n}, \quad \tilde{\beta}_n := \frac{\beta_n}{p_{n+1}}, \quad \tilde{\gamma}_n := \frac{\gamma_n}{p_{n+1}}$$

with $p_0 = 1$ and let

$$u_n := \tilde{x}_n - \sum_{i=0}^{n-1} \tilde{\gamma}_n$$

with $u_0 := \tilde{x}_0$. Observe that

$$\tilde{x}_{n+1} \leq \tilde{x}_n - \tilde{\beta}_n + \tilde{\gamma}_n \leq \tilde{x}_n + \tilde{\gamma}_n,$$

which implies $\{u_n\}$ is a nonincreasing sequence and $\{u_n + M\}$ is nonincreasing and nonnegative. Furthermore, since $u_0 = \tilde{x}_0 \leq x_0 < K$, we have for all $n \in \mathbb{N}$, $u_n + M \leq u_0 + M < K + M$ and so Remark 2.3.18 implies $\phi_1(\varepsilon) := (K + M)/\varepsilon$ is a bound on the fluctuations for $\{u_n + M\}$, and so must be a bound on the fluctuations for $\{u_n\}$. Similarly, $\sum_{i=0}^{\infty} \tilde{\gamma}_i \leq \sum_{i=0}^{\infty} \gamma_i < M$ has a bound on their fluctuations given by $\phi_2(\varepsilon) := M/\varepsilon$.

Now part (b) of Lemma 8.1.2 implies

$$\phi_3(\varepsilon) := \phi_1(\varepsilon/2) + \phi_2(\varepsilon/2) = \frac{2K + 4M}{\varepsilon}$$

is a bound on the fluctuations for $\{\tilde{x}_n\}$. Furthermore, since $x_n = \tilde{x}_n p_n$ and $\phi_4(\varepsilon) := L/\varepsilon$ is a bound on the fluctuations for $\{p_n\}$ (by Remark 2.3.18), part (a) of Lemma 8.1.2 implies

$$\phi(\varepsilon) := \phi_3\left(\frac{\varepsilon}{2L}\right) + \phi_4\left(\frac{\varepsilon}{2K}\right)$$

is a bound on the fluctuations for $\{x_n\}$ and the first part follows. For the second part of the theorem, note that

$$\sum_{i=0}^{n} \tilde{\beta}_i = \tilde{x}_0 - \tilde{x}_{n+1} + \sum_{i=0}^{n} \tilde{\gamma}_i \leq x_0 + \sum_{i=0}^{n} \gamma_i < K + M$$

and therefore

$$\sum_{i=0}^{n} \beta_i = \sum_{i=0}^{n} \tilde{\beta}_i p_{n+1} < L \sum_{i=0}^{n} \tilde{\beta}_i < L(K + M)$$

and the theorem is proved. □

Remark 8.1.4. As previously mentioned, from [85, Lemma 16], we can get a bound on the fluctuations from the rate of metastability obtained. Their rate of metastability, where $L > 0$ instead satisfies $\sum_{i=0}^{\infty} \alpha_i < L$, corresponds to the bound on the fluctuations,

$$\psi(\varepsilon) := \frac{4L(K + M)e^L + 4M + (K + M)e^L}{\varepsilon}$$

whereas ours, noting that a bound L on the sum corresponds to the bound $\prod_{i=0}^{\infty}(1 + \alpha_i) < e^L$ on the product, is just

$$\phi(\varepsilon) := \frac{8(K + M)e^L}{\varepsilon}.$$

Asymptotically (in ε), the two bounds are equivalent. However, the strategy presented here extends more easily to the stochastic setting and is a direct reflection of standard proofs of the Robbins-Siegmund theorem.

8.2 The main result

In this section, we present uniform learnable rates for the Robbins-Siegmund theorem. Our proof strategy will mirror the deterministic case presented in Theorem 8.1.3. In particular, we need lemmas for combining uniform learnable rates and moduli of boundedness for sums

and products of random variables (recalling, we already needed such a result in obtaining our quantitative martingale convergence theorem c.f. Lemma 7.2.3).

8.2.1 Preliminary lemmas

We start by providing a stochastic version of Lemma 8.1.2:

Lemma 8.2.1. *Suppose that $\{X_n\}$ and $\{Y_n\}$ converge with learnable rates of uniform convergence ϕ and ψ, respectively. Then:*

(a) A learnable rate of uniform convergence for $\{X_n + Y_n\}$ is given by

$$\omega(\lambda, \varepsilon) := \phi(\lambda/2, \varepsilon/2) + \psi(\lambda/2, \varepsilon/2)$$

(b) A learnable rate of uniform convergence for $\{X_n Y_n\}$ is given by

$$\omega(\lambda, \varepsilon) := \phi\left(\frac{\lambda}{4}, \frac{\varepsilon}{2\sigma(\lambda/4)}\right) + \psi\left(\frac{\lambda}{4}, \frac{\varepsilon}{2\rho(\lambda/4)}\right)$$

where ρ and σ are moduli of uniform boundedness for $\{X_n\}$ and $\{Y_n\}$, respectively.

Proof. Part (a) is exactly Lemma 7.2.3. For part (b), fix $\lambda, \varepsilon \in (0, 1]$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$, and define the events

$$Q := \sup_{n \in \mathbb{N}} |X_n| < \rho\left(\frac{\lambda}{4}\right) \quad \text{and} \quad R := \sup_{n \in \mathbb{N}} |Y_n| < \sigma\left(\frac{\lambda}{4}\right)$$

Then for any $n \in \mathbb{N}$, we have

$$\begin{aligned} &\mathbb{P}\left(\exists i, j \in [a_n; b_n]\left(|X_i Y_i - X_j Y_j| \geq \varepsilon\right)\right) \\ &\leq \mathbb{P}\left(\exists i, j \in [a_n; b_n]\left(|X_i Y_i - X_j Y_j| \geq \varepsilon\right) \cap R \cap Q\right) + \frac{\lambda}{2} \end{aligned}$$

so it suffices to show that there exists some $n \leq \omega(\lambda, \varepsilon)$ such that

$$\mathbb{P}\left(\exists i, j \in [a_n; b_n]\left(|X_i Y_i - X_j Y_j| \geq \varepsilon\right) \cap R \cap Q\right) < \frac{\lambda}{2}$$

Take some ω in the set within the probability measure in (8.2.1). Then analogously to Lemma 8.1.2, we have either

$$\exists i, j \in [a_n; b_n]\left(|X_i(\omega) - X_j(\omega)| \geq \frac{\varepsilon}{2\sigma(\lambda/4)}\right)$$

or

$$\exists i, j \in [a_n; b_n] \left(|Y_i(\omega) - Y_j(\omega)| \geq \frac{\varepsilon}{2\rho(\lambda/4)} \right)$$

and therefore if (8.2.1) fails for all $n \leq \omega(\lambda, \varepsilon)$, then we have either

$$\mathbb{P}\left(\exists i, j \in [a_n; b_n] \left(|X_i - X_j| \geq \frac{\varepsilon}{2\sigma(\lambda/4)} \right) \right) \geq \frac{\lambda}{4}$$

or

$$\mathbb{P}\left(\exists i, j \in [a_n; b_n] \left(|Y_i - Y_j| \geq \frac{\varepsilon}{2\rho(\lambda/4)} \right) \right) \geq \frac{\lambda}{4}$$

and so again analogously to Lemma 8.1.2 there exists either a subsequence $a_{n_0} < b_{n_0} \leq a_{n_1} < b_{n_1} \leq \ldots$ such that

$$\forall k \leq \phi\left(\frac{\lambda}{4}, \frac{\varepsilon}{2\sigma(\lambda/4)}\right) \left[\mathbb{P}\left(\exists i, j \in [a_{n_k}; b_{n_k}] \left(|X_i - X_j| \geq \frac{\varepsilon}{2\sigma(\lambda/4)} \right) \right) \geq \frac{\lambda}{4} \right]$$

or a subsequence $a_{m_0} < b_{m_0} \leq a_{m_1} < b_{m_1} \leq \ldots$ such that

$$\forall k \leq \phi\left(\frac{\lambda}{4}, \frac{\varepsilon}{2\rho(\lambda/4)}\right) \left[\mathbb{P}\left(\exists i, j \in [a_{m_k}; b_{m_k}] \left(|Y_i - Y_j| \geq \frac{\varepsilon}{2\rho(\lambda/4)} \right) \right) \geq \frac{\lambda}{4} \right]$$

contradicting the defining properties of ϕ or ψ. □

The next technical results we need are how moduli of uniform boundedness combine under arithmetic operations:

Lemma 8.2.2. *Suppose that $\{X_n\}$ and $\{Y_n\}$ have moduli of uniform boundedness ρ and τ, respectively. Then:*

(a) A modulus of uniform boundedness for $\{X_n + Y_n\}$ is given by

$$\gamma(\lambda) := \rho\left(\lambda/2\right) + \tau\left(\lambda/2\right)$$

(b) A modulus of uniform boundedness for $\{X_n Y_n\}$ is given by

$$\gamma(\lambda) := \rho\left(\lambda/2\right) \cdot \tau\left(\lambda/2\right)$$

Proof. We just prove (b), as (a) is similar. Observe that for any $n, m \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n Y_n| \geq nm\right) &\leq \mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \sup_{n\in\mathbb{N}} |Y_n| \geq nm\right) \\ &\leq \mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq n \cup \sup_{n\in\mathbb{N}} |Y_n| \geq m\right) \\ &\leq \mathbb{P}\left(\sup_{n\in\mathbb{N}} |X_n| \geq n\right) + \left(\sup_{n\in\mathbb{N}} |Y_n| \geq m\right)\end{aligned}$$

and so the result follows immediately. □

Lastly, we note that $\sum_{i=0}^{\infty} A_i < \infty$ almost surely, where now the A_i are nonnegative random variables can be represented quantitatively through a modulus of uniform boundedness ρ, i.e. a function satisfying

$$\mathbb{P}\left(\sum_{i=0}^{\infty} A_i \geq \rho(\lambda)\right) < \lambda$$

for all $\lambda \in (0, 1]$, and this will be our preferred definition of the quantitative almost sure convergence for infinite series. Furthermore, since the partial sums form a monotone sequence of random variables, we can use Remark 7.1.11 to obtain learnable uniform rates.

8.2.2 Rates for the Robbins-Siegmund theorem

We are now ready to present a quantitative version of the Robbins-Siegmund theorem. Our strategy is to analyse the standard proof of the result (as in [129]), which is proven in a similar spirit to the implicit proof of Theorem 8.1.1 we presented in Theorem 8.1.3.

Theorem 8.2.3 (Quantitative Robbins-Siegmund theorem)**.** *Let $\{X_n\}$, $\{A_n\}$, $\{B_n\}$ and $\{C_n\}$ be nonnegative stochastic processes adapted to some filtration $\{\mathcal{F}_n\}$ such that*

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq (1 + A_n)X_n - B_n + C_n$$

for all $n \in \mathbb{N}$. Suppose that $K > \mathbb{E}(X_0)$ and that $\rho, \tau : (0, 1] \to [1, \infty)$ are nonincreasing and satisfy

$$\mathbb{P}\left(\prod_{i=0}^{\infty}(1 + A_n) \geq \rho(\lambda)\right) < \lambda \quad \text{and} \quad \mathbb{P}\left(\sum_{i=0}^{\infty} C_n \geq \tau(\lambda)\right) < \lambda$$

for all $\lambda \in (0, 1]$. Then $\{X_n\}$ converges almost surely, with learnable rate of uniform convergence

$$\phi_{K,\rho,\tau}(\lambda, \varepsilon) := \kappa \cdot \left(\frac{\rho\left(\frac{\lambda}{8}\right) \cdot \left(K + \tau\left(\frac{\lambda}{16}\right)\right)}{\lambda\varepsilon}\right)^2$$

where $0 < \kappa \leq 4096c + 272$ with c the constant from Theorem 7.2.4, and $\sum_{i=0}^{\infty} B_i$ is finite

almost surely, with modulus of uniform boundedness

$$\chi_{K,\rho,\tau}(\lambda) := \frac{16 \cdot \rho\left(\frac{\lambda}{2}\right) \cdot \left(K + \tau\left(\frac{\lambda}{4}\right)\right)}{\lambda}$$

Proof. Analogously to the nonstochastic case, we define

$$P_n := \prod_{i=0}^{n-1}(1 + A_i), \quad \tilde{X}_n := \frac{X_n}{P_n}, \quad \tilde{B}_n := \frac{B_n}{P_{n+1}}, \quad \tilde{C}_n := \frac{C_n}{P_{n+1}}$$

with $P_0 = 1$. Since $\{B_n\}$ is nonnegative we have

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq (1 + A_n)X_n + C_n$$

and therefore defining

$$U_n := \tilde{X}_n - \sum_{i=0}^{n-1} \tilde{C}_i$$

with $U_0 := \tilde{X}_0$ it follows that

$$\begin{aligned}
\mathbb{E}(U_{n+1} \mid \mathcal{F}_n) &= \mathbb{E}\left[\tilde{X}_{n+1} - \sum_{i=0}^{n} \tilde{C}_i \mid \mathcal{F}_n\right] \\
&= \frac{\mathbb{E}\left[X_{n+1} \mid \mathcal{F}_n\right]}{P_{n+1}} - \sum_{i=0}^{n} \tilde{C}_i \\
&\leq \frac{(1 + A_n)X_n + C_n}{P_{n+1}} - \sum_{i=0}^{n} \tilde{C}_i \\
&= \tilde{X}_n - \sum_{i=0}^{n-1} C_i = U_n.
\end{aligned}$$

Thus, $\{U_n\}$ is a supermartingale. Now, for $x > 0$ define the stopping time $T_x \in \mathbb{N} \cup \{\infty\}$ by

$$T_x := \inf\left\{n \,:\, \sum_{i=0}^{n} \tilde{C}_i > x\right\}.$$

The optional stopping theorem, Theorem 2.4.24, implies $\{U_{n\wedge T_x}\}$ (where $n \wedge m := \min\{n, m\}$) is also a supermartingale. Furthermore, $\{U_{n\wedge T_x} + x\}$ is a *nonnegative* supermartingale, since on $\{T_x < \infty\}$ we have

$$U_{n\wedge T_x} + x = \tilde{X}_{n\wedge T_x} - \sum_{i=0}^{n\wedge T_x - 1} \tilde{C}_i + x \geq X_{n\wedge T_x} - \sum_{i=0}^{T_x - 1} \tilde{C}_i + x \geq X_{n\wedge T_x} \geq 0$$

and on $\{T_x = \infty\}$ we have $\sum_{i=0}^{n-1} \tilde{C}_i \leq x$ for all $n \in \mathbb{N}$, and so also $U_{n\wedge T_x} + x \geq 0$. Noting that $\mathbb{E}(U_{0\wedge T_x} + x) = \mathbb{E}(\tilde{X}_0) + x < K + x$, by Theorem 7.2.4 a learnable rate of uniform convergence for $\{U_{n\wedge T_x} + x\}$ is given by

$$\phi_1^x(\lambda, \varepsilon) := c\left(\frac{K+x}{\lambda\varepsilon}\right)^2$$

for a constant $c > 0$ from that theorem. Furthermore, it is clear that the above also provides a learnable uniform rate for $\{U_{n\wedge T_x}\}$. We now find a rate for $\{U_n\}$. For $\lambda \in (0, 1]$, define the event

$$Q_\lambda := \sum_{i=0}^{\infty} \tilde{C}_i < \tau\left(\frac{\lambda}{2}\right)$$

noting that since $\tilde{C}_i \leq C_i$ we have $\mathbb{P}(Q_\lambda^c) < \lambda/2$. We now define

$$\phi_1(\lambda, \varepsilon) := \phi_1^{\tau(\lambda/2)}(\lambda/2, \varepsilon) = 4c\left(\frac{K+\tau(\lambda/2)}{\lambda\varepsilon}\right)^2.$$

Now, for any $\lambda, \varepsilon \in (0, 1]$ and $a_0 < b_0 \leq a_1 < b_1 \leq \ldots$ there exists some $n \leq \phi_1(\lambda, \varepsilon)$ such that

$$\mathbb{P}\left(\exists i, j \in [a_n; b_n]\left(|U_{i\wedge T_{\tau(\lambda/2)}} - U_{j\wedge T_{\tau(\lambda/2)}}| \geq \varepsilon\right)\right) < \lambda/2.$$

We have

$$\mathbb{P}\left(\exists i, j \in [a_n; b_n]\left(|U_i - U_j| \geq \varepsilon\right)\right) \leq \mathbb{P}\left(\exists i, j \in [a_n; b_n]\left(|U_i - U_j| \geq \varepsilon\right) \cap Q_\lambda\right) + \lambda/2$$

and if $\omega \in Q_\lambda$ then $T_{\tau(\lambda/2)}(\omega) = \infty$ so $U_{n\wedge T_{\tau(\lambda/2)}(\omega)}(\omega) = U_n(\omega)$. Hence

$$\begin{aligned}&\mathbb{P}\left(\exists i, j \in [a_n; b_n]\left(|U_i - U_j| \geq \varepsilon\right) \cap Q_\lambda\right)\\ &\leq \mathbb{P}\left(\exists i, j \in [a_n; b_n]\left(|U_{i\wedge T_{\tau(\lambda/2)}} - U_{j\wedge T_{\tau(\lambda/2)}}| \geq \varepsilon\right)\right) < \lambda/2\end{aligned}$$

from which it follows that ϕ_1 is a learnable rate of uniform convergence for $\{U_n\}$.

Similarly, we argue to obtain a rate of uniform boundedness for $\{U_n\}$. By Ville's inequality, Theorem 2.4.25, a modulus of uniform boundedness for the positive supermartingale $\{U_{n\wedge T_x} + x\}$ is given by

$$\chi_1^x(\lambda) := \frac{K+x}{\lambda}.$$

From this, we have

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} |U_{n\wedge T_x}| \geq \frac{2(K+x)}{\lambda}\right) \leq \mathbb{P}\left(\sup_{n\in\mathbb{N}} (U_{n\wedge T_x} + x) \geq \frac{2(K+x)}{\lambda}\right) < \frac{\lambda}{2}$$

where for the first inequality we note that for any $y > x$, if $|U_{n\wedge T_x(\omega)}(\omega)| \geq y$ then $|U_{n\wedge T_x(\omega)}(\omega)| =$

$U_{n\wedge T_x(\omega)}(\omega)$ since $-U_{n\wedge T_x(\omega)}(\omega) \leq x$, and so in particular $U_{n\wedge T_x(\omega)}(\omega) + x \geq U_{n\wedge T_x(\omega)}(\omega) \geq y$, and since $2(K+x)/\lambda > x$ the inequality on suprema holds. Finally, we have

$$\begin{aligned}\mathbb{P}\left(\sup_{n\in\mathbb{N}}|U_n| \geq \frac{2(K+\tau(\lambda/2))}{\lambda}\right) &\leq \mathbb{P}\left(\left[\sup_{n\in\mathbb{N}}|U_n| \geq \frac{2(K+\tau(\lambda/2))}{\lambda}\right] \cap Q_\lambda\right) + \frac{\lambda}{2} \\ &\leq \mathbb{P}\left(\sup_{n\in\mathbb{N}}|U_{n\wedge T_{\tau(\lambda/2)}}| \geq \frac{2(K+\tau(\lambda/2))}{\lambda}\right) + \frac{\lambda}{2} < \lambda\end{aligned}$$

and so

$$\chi_1(\lambda) := \frac{2(K+\tau(\lambda/2))}{\lambda}$$

is a modulus of uniform boundedness for $\{U_n\}$.The proof is then concluded through repeated applications of Lemmas 8.2.1 and 8.2.2.

Since $\tilde{X}_n = U_n + \sum_{i=0}^{n-1}\tilde{C}_i$ and τ is also a modulus of uniform boundedness for $\{\tilde{C}_n\}$, Remark 7.1.11 and Lemma 8.2.1 (a) yield that a learnable rate of uniform convergence for $\{\tilde{X}_n\}$ is given by

$$\phi_2(\lambda,\varepsilon) := 64c\left(\frac{K+\tau(\lambda/4)}{\lambda\varepsilon}\right)^2 + \frac{8\tau(\lambda/4)}{\lambda\varepsilon}$$

Furthermore, by Lemma 8.2.2 (a), a modulus of uniform boundedness for $\{\tilde{X}_n\}$ is given by

$$\chi_2(\lambda) := \frac{4(K+\tau(\lambda/2))}{\lambda} + \tau(\lambda/2).$$

Now, since $X_n = \tilde{X}_n P_n$ and ρ is by definition a modulus of uniform boundedness for $\{P_n\}$, by Remark 7.1.11 and monotonicity of $\{P_n\}$ a learnable rate of uniform convergence for the sequence is given by $2\rho(\lambda/2)/\lambda\varepsilon$. Therefore, Lemma 8.2.1 (b) yields that a learnable rate of uniform convergence for $\{X_n\}$ is given by any bound on

$$\phi_2\left(\frac{\lambda}{4}, \frac{\varepsilon}{2\rho(\lambda/4)}\right) + \frac{16\cdot\chi_2(\lambda/4)\cdot\rho(\lambda/8)}{\lambda\varepsilon}.$$

The simplified bound in the statement of the theorem then follows in a crude way by bringing together terms and using the assumptions $\lambda,\varepsilon \in (0,1]$ and that ρ,τ are nonincreasing taking values in $[1,\infty)$.

To obtain the modulus of uniform boundedness for $\sum_{i=0}^{\infty} B_i$, we define

$$V_n := \tilde{X}_n - \sum_{i=0}^{n-1}(\tilde{C}_i - \tilde{B}_i) = U_n + \sum_{i=0}^{n-1}\tilde{B}_i$$

with $V_0 := \tilde{X}_0$. By an essentially identical argument to that for $\{U_n\}$, we can show that $\{V_n\}$ is a supermartingale, and defining the stopping time T_x just as before and observing that because

$\{B_n\}$ is nonnegative it is also the case that $\{V_{n\wedge T_x} + x\}$ is a nonnegative supermartingale with $\mathbb{E}(V_{0\wedge T_x} + x) < K + x$ and has modulus of uniform boundedness χ_1^x defined previously. Then just as before, for $\omega \in Q_\lambda$ we have $T_x(\omega) = \infty$ and thus $V_{n\wedge T_x(\omega)}(\omega) = V_n(\omega)$, and so by an identical argument, χ_1, as previously defined, is a modulus of uniform boundedness for $\{V_n\}$. We now apply Lemma 8.2.2 several times: Since $\sum_{i=0}^{n-1} \tilde{B}_i = V_n + (-U_n) \leq V_n + \sum_{i=0}^{n-1} C_i$, we have that

$$\chi_3(\lambda) := \frac{5(K + \tau(\lambda/2))}{\lambda} \geq \chi_1(\lambda/2) + \tau(\lambda/2)$$

is a modulus of uniform boundedness for $\{\sum_{i=0}^{n-1} \tilde{B}_i\}$. Finally, since

$$\sum_{i=0}^{n-1} B_i = \sum_{i=0}^{n-1} \tilde{B}_i P_{i+1} \leq P_n \sum_{i=0}^{n-1} \tilde{B}_i$$

a modulus of uniform boundedness for $\{\sum_{i=0}^{n-1} B_i\}$ is given by one for the product $P_n \sum_{i=0}^{n-1} \tilde{B}_i$, and so by Lemma 8.2.2 we may take

$$\chi(\lambda) := \rho(\lambda/2) \cdot \chi_3(\lambda/2)$$

as such a modulus. The second part of the theorem follows after simplification. □

A very useful corollary from the proof of the quantitative Robbins-Siegmund theorem we have just presented is that we can obtain a modulus of uniform boundedness for $\{X_n\}$:

Corollary 8.2.4. *Suppose we have the same assumptions as in Theorem 8.2.3. Then $\{X_n\}$ has a modulus of uniform boundedness given by:*

$$\nu_{K,\rho,\tau}(\lambda) := \frac{9\rho(\lambda/2)(K + \tau(\lambda/8))}{\lambda}$$

Proof. By the proof of Theorem 8.2.3

$$\chi_2(\lambda) := \frac{4(K + \tau(\lambda/2))}{\lambda} + \tau(\lambda/2).$$

is a modulus of uniform boundedness for $\{\tilde{X}_n\}$. Furthermore, since $X_n = P_n\tilde{X}_n$ and ρ is a modulus of uniform boundedness for $\{P_n\}$, Lemma 8.2.2 implies that $\chi_2(\lambda/2) \cdot \rho(\lambda/2)$ is a modulus of uniform boundedness for $\{X_n\}$ and the result follows since $\nu_{K,\rho,\tau}(\lambda) \geq \chi_2(\lambda/2) \cdot \rho(\lambda/2)$. □

8.3 Applications

In this section, we discuss some applications of the Robbins-Siegmund theorem. We start by giving a quantitative version of a common instantiation of the Robbins-Siegmund theorem, and by imposing additional assumptions on this result, we can obtain rates of almost sure convergence.

We then discuss applications of the Robbins-Siegmund theorem. The first application is an illustrative example, demonstrating how one obtains rates for Kolmogorov's Strong Law of Large Numbers from the Robbins-Siegmund theorem. However, the rates we present here are worse than those given in Chapter 6. We then discuss how our quantitative Robbins-Siegmund theorem can be used to obtain a computational interpretation of a generalisation of the Robbins-Monro procedure introduced in [129]. Lastly, we give rates of almost sure convergence for Dvoretsky's theorem [36].

8.3.1 Useful instantiations of the Robbins-Siegmund theorem

This section will present how our quantitative Robbins-Siegmund theorem can be used to obtain quantitative versions of various important results in probability theory.

Our applications all follow from the following corollary of the Robbin-Siegmund theorem:

Corollary 8.3.1. *Let $\{X_n\}$, $\{A_n\}$, $\{U_n\}$, $\{V_n\}$ and $\{C_n\}$ be nonnegative stochastic processes, adapted to some filtration $\{\mathcal{F}_n\}$, satisfying*

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq (1 + A_n)X_n - U_nV_n + C_n$$

for all $n \in \mathbb{N}$. Suppose further that, almost surely,

$$\sum_{i=0}^{\infty} A_n < \infty, \sum_{i=0}^{\infty} C_n < \infty \text{ and } \sum_{i=0}^{\infty} U_i = \infty.$$

Then $\liminf_{n\to\infty} V_n = 0$ almost surely. Furthermore, if $\liminf_{n\to\infty} X_n = 0$ almost surely, then $X_n \to 0$ almost surely.

Proof. By the Robbins-Siegmund theorem we have $\sum_{i=0}^{\infty} U_iV_i < \infty$ almost surely and since $\sum_{i=0}^{\infty} U_i = \infty$ almost surely, we must have $\liminf_{n\to\infty} V_n = 0$ almost surely. If $\liminf_{n\to\infty} X_n = 0$ almost surely then, since $\{X_n\}$ almost surely converges to some limit by the Robbins-Siegmund theorem, it must therefore converge to 0 almost surely. □

We give two quantitative versions of the above theorem. The first will be a direct computational interpretation of the above theorem, where we will get metastable uniform rates in the conclusion. In the second, we shall see that strengthening the assumption that $\liminf_{n\to\infty} X_n =$

0 almost surely to $\liminf_{n\to\infty} \mathbb{E}(X_n) = 0$ will allow us to obtain direct rates of almost sure convergence.[1]

To give the above theorem a computational interpretation, we must first make the assumptions of the theorem quantitative, specifically a series diverging almost surely, and the liminf of a sequence being 0.

Definition 8.3.2. A function $\Phi : (0,1] \times (0,1] \times \mathbb{N} \to \mathbb{N}$ is a *liminf-modulus* for $\{X_n\}$ if for all $\lambda, \varepsilon \in (0,1]$ and $n \in \mathbb{N}$:

$$\mathbb{P}\left(\forall k \in [n; \Phi(\lambda, \varepsilon, n)](X_k \geq \varepsilon)\right) < \lambda.$$

As in Definition 3.1.3, we will assume the following monotonicity property

$$\forall \lambda, \varepsilon \in (0,1]\, \forall m, n \in \mathbb{N}\, (n \leq m \to \Phi(\lambda, \varepsilon, n) \leq \Phi(\lambda, \varepsilon, m)).$$

Furthermore, we assume

$$\forall \lambda, \varepsilon \in (0,1]\, \forall n \in \mathbb{N}\, (n \leq \Phi(\lambda, \varepsilon, n)).$$

Definition 8.3.3. A function $r : (0,1] \times (0,\infty) \times \mathbb{N} \to \mathbb{N}$ is a *modulus of almost sure divergence* for $\sum_{i=0}^{\infty} U_i$ if for all $\lambda \in (0,1]$, $x > 0$ and $n \in \mathbb{N}$:

$$\mathbb{P}\left(\sum_{i=n}^{r(\lambda,x,n)} U_i < x\right) < \lambda.$$

Again, we will assume the following monotonicity property

$$\forall \lambda, \varepsilon \in (0,1]\, \forall m, n \in \mathbb{N}\, (n \leq m \to r(\lambda, \varepsilon, n) \leq r(\lambda, \varepsilon, m))$$

and,

$$\forall \lambda, \varepsilon \in (0,1]\, \forall n \in \mathbb{N}\, (n \leq r(\lambda, \varepsilon, n)).$$

Remark 8.3.4. The previous definition is a stochastic analogue of Definition 3.1.3.

We now give a computational interpretation to a main idea used in the proof of Theorem 8.3.1.

Lemma 8.3.5. *Suppose that $\{U_n\}$ and $\{V_n\}$ are nonnegative stochastic processes where ψ is a a modulus of uniform boundedness for $\sum_{i=0}^{\infty} U_i V_i$ and r is a modulus of divergence for $\sum_{i=0}^{\infty} U_i$. Then*

$$\Phi(\lambda, \varepsilon, n) := r(\lambda/2, \psi(\lambda/2)/\varepsilon, n)$$

is a liminf-modulus for $\{V_n\}$.

[1] The fact that this is a stronger assumption can be seen from an application of Fatou's lemma.

Proof. Fix $\lambda, \varepsilon \in (0,1]$ and $n \in \mathbb{N}$. Let A denote the event we are interested in, namely,

$$\forall k \in [n; \Phi(\lambda, \varepsilon, n)](V_k \geq \varepsilon).$$

Now define the following events

$$B := \sum_{i=0}^{\infty} U_i(\omega)V_i(\omega) < \psi(\lambda/2) \quad \text{and} \quad C := \sum_{i=n}^{\Phi(\lambda,\varepsilon,n)} U_i < \psi(\lambda/2)/\varepsilon.$$

Suppose that $\omega \in A \cap C^c$. Then we would have

$$\psi(\lambda/2) \leq \varepsilon \sum_{i=n}^{\Phi(\lambda,\varepsilon,n)} U_i(\omega) \leq \sum_{i=n}^{\Phi(\lambda,\varepsilon,n)} U_i(\omega)V_i(\omega) \leq \sum_{i=0}^{\infty} U_i(\omega)V_i(\omega)$$

which implies $\omega \in B^c$. Therefore $A \cap C^c \subseteq B^c$.

This implies

$$\mathbb{P}\left(\forall k \in [n; \Phi(\lambda, \varepsilon, n)](V_k \geq \varepsilon)\right) = \mathbb{P}(A \cap C^c) + \mathbb{P}(A \cap C) < \mathbb{P}(B^c) + \lambda/2 < \lambda$$

and the result is proven. □

We now have the following quantitative version of Theorem 8.3.1.

Theorem 8.3.6. *Let $\{X_n\}$, $\{A_n\}$, $\{U_n\}$, $\{V_n\}$ and $\{C_n\}$ be nonnegative stochastic processes adapted to some filtration $\{\mathcal{F}_n\}$ such that*

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq (1 + A_n)X_n - U_nV_n + C_n$$

for all $n \in \mathbb{N}$. Suppose that $K > \mathbb{E}(X_0)$ and $\rho, \tau : (0,1] \to [1,\infty)$ are nonincreasing, representing moduli of uniform boundedness for $\prod_{i=0}^{\infty}(1 + A_i)$ and $\sum_{i=0}^{\infty} C_i$ respectively. In addition, let $r : (0,1] \times (0,\infty) \times \mathbb{N} \to \mathbb{N}$ be a modulus of divergence for $\sum_{i=0}^{\infty} U_i$. Let $\phi_{K,\rho,\tau}$ and $\chi_{K,\rho,\tau}$ be defined in terms of K, ρ and τ as in Theorem 8.2.3. Then

$$\Phi(\lambda, \varepsilon, n) := r(\lambda/2, \chi_{K,\rho,\tau}(\lambda/2)/\varepsilon, n)$$

is a liminf-modulus for $\{V_n\}$.

Moreover, if Ψ is a liminf modulus for $\{X_n\}$. Then

$$\Gamma(\lambda, \varepsilon, g) := \tilde{f}^{\phi_{K,\rho,\tau}(\lambda/2,\varepsilon/2)}(0),$$

with

$$\tilde{f}(j) := j + \max\{g(j), \Psi(\lambda/2, \varepsilon/2, j) - j\},$$

is a rate of metastable uniform convergence for $\{X_n\}$ *to* 0.

Proof. By Theorem 8.2.3, $\chi_{K,\rho,\tau}$ is a modulus of uniform boundedness for $\sum_{i=0}^{\infty} U_i V_i$, and therefore by Lemma 8.3.5, Φ as defined in the statement of the theorem is a liminf-modulus for $\{V_n\}$. For the second part, by Theorem 8.2.3 we have that $\phi_{K,\rho,\tau}$ is a learnable rate of uniform convergence for $\{X_n\}$, and so fixing $\lambda, \varepsilon \in (0, 1]$ and defining $f(j) := \max\{g(j), \Psi(\lambda/2, \varepsilon/2, j) - j\}$, there exists some $n \leq \tilde{f}^{\phi_{K,\rho,\tau}(\lambda/2,\varepsilon/2)}(0)$ such that

$$\mathbb{P}\left(\exists i, j \in [n; n + f(n)](|X_i - X_j| \geq \varepsilon/2)\right) < \lambda/2.$$

Let A be the event inside the probability above. For this particular n, since Ψ is a liminf-modulus for $\{X_n\}$ we also have

$$\mathbb{P}\left(\forall k \in [n; \Psi(\lambda/2, \varepsilon/2, n)](X_k \geq \varepsilon/2)\right) < \lambda/2.$$

Let B be the event inside this probability. Fix $\omega \in \Omega$ and suppose that there exists some $k(\omega) \in [n; n + g(n)] \subseteq [n; n + f(n)]$ such that $X_{k(\omega)}(\omega) \geq \varepsilon$. Either $\omega \in B$, or there exists some $j(\omega) \in [n; \Psi(\lambda/2, \varepsilon/2, n)] \subseteq [n; n + f(n)]$ such that $X_{j(\omega)}(\omega) < \varepsilon/2$. But this then implies that $|X_{j(\omega)}(\omega) - X_{k(\omega)}(\omega)| \geq \varepsilon/2$, and so $\omega \in A$. Therefore

$$\mathbb{P}\left(\exists k \in [n; n + g(n)](X_k \geq \varepsilon)\right) \leq \mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B) < \lambda$$

and the theorem is proved. □

We have already seen through Example 3.1.7 that a general sequence of nonnegative numbers converging to 0 does not necessarily have a computable rate of convergence to 0, even if the sequence is a computable sequence of computable numbers. However, for a nonnegative nonincreasing sequence $\{a_n\}$ that converges to 0, if we impose the condition that for each $n \in \mathbb{N}$ and $\varepsilon \in \mathbb{Q}^+$ there is a computable process to determine whether or not $a_n < \varepsilon$ then such a sequence does possess a computable rate of convergence to 0 by an unbounded search. Another condition that would guarantee a computable rate of convergence for such a class of sequence is the existence of a computable liminf-modulus, that is, a computable function $\Phi : \mathbb{Q}^+ \times \mathbb{N} \to \mathbb{N}$ satisfying,

$$\forall \varepsilon \in \mathbb{Q}^+ \, \forall n \in \mathbb{N} \, \exists k \in [n; n + \Phi(\varepsilon, n)](a_k < \varepsilon).$$

In such a case, a rate of convergence will be given by $r(\varepsilon) := \Phi(\varepsilon, 0)$. Thus, we introduce the following definition:

Definition 8.3.7. A function $\Phi : (0,1] \times \mathbb{N} \to \mathbb{N}$ is a liminf-modulus for a sequence of reals $\{a_n\}$ if:

$$\forall \varepsilon \in (0,1] \, \forall n \in \mathbb{N} \, \exists k \in [n; \Phi(\varepsilon, n)](a_k < \varepsilon).$$

We will assume the following monotonicity property

$$\forall \lambda, \varepsilon \in (0,1] \, \forall m, n \in \mathbb{N} \, (n \leq m \to \Phi(\varepsilon, n) \leq \Phi(\varepsilon, m)).$$

This definition of a liminf-modulus is clearly the the deterministic analogue of the previously introduced Definition 8.3.2.

Supermartingales can be seen as a stochastic generalisation of nonincreasing sequences of real numbers. Indeed, a nonincreasing sequence of reals is a supermartingale, and the inability to obtain computable rates of convergences carries over to such stochastic processes. However, as in the deterministic case, providing additional computational information on the liminf of the process allows us to obtain rates:

Proposition 8.3.8. *Let $\{X_n\}$ be a nonnegative supermartingale and let $K > 0$ satisfy $\mathbb{E}(X_0) < K$. If $\Phi : (0,1] \times \mathbb{N} \to \mathbb{N}$ is a liminf-modulus for the sequence of real numbers $\{\mathbb{E}(X_n)\}$, then $\{X_n\}$ converges to 0 with a rate of convergence*

$$\phi(\lambda, \varepsilon) := \Phi(\varepsilon\lambda, 0).$$

Proof. Let $\varepsilon, \lambda \in (0,1]$ be given and let $N := \phi(\lambda, \varepsilon)$. By Ville's inequality (Theorem 2.4.25),

$$\mathbb{P}(\exists n \geq N \, (X_n \geq \varepsilon) \leq \frac{\mathbb{E}(X_N)}{\varepsilon} < \lambda.$$

Where the last inequality follows by the fact that Φ is a liminf modulus and $\{X_n\}$ is a supermartingale so $\mathbb{E}(X_N) \leq \mathbb{E}(X_k)$ for all $k \in [0; N]$. □

We have the following generalisation of the above result, where we now incorporate error terms:

Theorem 8.3.9. *Let $\{X_n\}, \{C_n\}$ and $\{V_n\}$ be nonnegative integrable stochastic processes adapted to some filtration $\{\mathcal{F}_n\}$ and let $\{\alpha_n\}, \{u_n\}$ be sequences of nonnegative real numbers such that for all $n \in \mathbb{N}$*

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq (1 + \alpha_n)X_n - u_n V_n + C_n$$

and

$$\prod_{i=0}^{\infty}(1 + \alpha_i) < L, \quad \forall \varepsilon \in (0,1] \left(\sum_{i=\phi(\varepsilon)}^{\infty} \mathbb{E}(C_i) < \varepsilon \right) \quad \text{and} \quad \forall n \in \mathbb{N} \forall x > 0 \left(\sum_{i=n}^{r(n,x)} u_i \geq x \right)$$

for some $L \geq 1/2$, $\phi : (0,1] \to \mathbb{N}$ and $r : \mathbb{N} \times (0,\infty) \to \mathbb{N}$ (that is, r is a rate of divergence as in Definition 3.1.3). Furthermore, take $K, M > 0$ such that $\mathbb{E}(X_0) < K$ and $\sum_{i=0}^{\infty} \mathbb{E}(C_i) < M$. Then

$$\Phi(\varepsilon, n) := r(n, L(K+M)/\varepsilon)$$

is a liminf-modulus for $\{\mathbb{E}(V_n)\}$.

Moreover, if Ψ is a liminf-modulus for $\{\mathbb{E}(X_n)\}$, then $X_n \to 0$ almost surely with a rate of almost sure convergence

$$\Gamma(\lambda, \varepsilon) := \Psi(\lambda\varepsilon/2L, \phi(\lambda\varepsilon/2L)).$$

Proof. Taking expectations on both sides of the recurrence, we have

$$\mathbb{E}(X_{n+1}) \leq (1+\alpha_n)\mathbb{E}(X_n) - u_n\mathbb{E}(V_n) + \mathbb{E}(C_n)$$

and we can now apply Theorem 8.1.3 with $x_n := \mathbb{E}(X_n)$, $\beta_n := u_n\mathbb{E}(V_n)$ and $\gamma_n := \mathbb{E}(C_n)$, where in particular we obtain $\sum_{i=0}^{\infty} u_i\mathbb{E}(V_i) < L(K+M)$. Now observe for any $\varepsilon \in (0,1]$ and $n \in \mathbb{N}$, if $\mathbb{E}(V_i) \geq \varepsilon$ for all $i \in [n; \Phi(\varepsilon, n)]$, then we have

$$L(K+M) \leq \varepsilon \sum_{i=n}^{\Phi(\varepsilon,n)} u_i \leq \sum_{i=n}^{\Phi(\varepsilon,n)} u_i\mathbb{E}(V_i) \leq \sum_{i=0}^{\infty} u_i\mathbb{E}(V_i). \tag{8.1}$$

which is a contradiction and establishes the first part. For the second part, define

$$Z_n := \tilde{\alpha}_n X_n + \sum_{i=n}^{\infty} \tilde{\alpha}_{i+1}\mathbb{E}(C_i|\mathcal{F}_n)$$

for

$$\tilde{\alpha}_n := \prod_{i=n}^{\infty}(1+\alpha_i) < L.$$

We show that $\{Z_n\}$ is a supermartingale. To see this, note that

$$\begin{aligned}
\mathbb{E}(Z_{n+1} \mid \mathcal{F}_n) &= \tilde{\alpha}_{n+1}\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) + \mathbb{E}\left(\sum_{i=n+1}^{\infty} \tilde{\alpha}_{i+1}\mathbb{E}(C_i|\mathcal{F}_{n+1}) \mid \mathcal{F}_n\right) \\
&\leq \tilde{\alpha}_n X_n - \tilde{\alpha}_{n+1}u_n V_n + \tilde{\alpha}_{n+1}C_n + \sum_{i=n+1}^{\infty} \tilde{\alpha}_{i+1}\mathbb{E}(C_i|\mathcal{F}_n) \\
&\leq Z_n - \tilde{\alpha}_{n+1}u_n V_n \leq Z_n.
\end{aligned}$$

Furthermore

$$\mathbb{E}(Z_n) \leq L\left(\mathbb{E}(X_n) + \sum_{i=n}^{\infty} \gamma_i\right).$$

Now from the fact that Ψ is a liminf modulus for $\{\mathbb{E}(X_n)\}$ and the defining property of ϕ (and the monotonicity of Φ), we have,

$$\Delta(\varepsilon, n) := \Psi(\varepsilon/2L, \max\{n, \phi(\varepsilon/2L)\})$$

is a liminf modulus for $\{\mathbb{E}(Z_n)\}$. Now since $Z_n \geq X_n$ we have for each $N \in \mathbb{N}$ and $\varepsilon > 0$

$$\mathbb{P}(\exists n \geq N\,(X_n \geq \varepsilon)) \leq \mathbb{P}(\exists n \geq N\,(Z_n \geq \varepsilon))$$

thus a rate of almost sure convergence to 0 for $\{Z_n\}$ must be one for $\{X_n\}$ and so the result follows from Proposition 8.3.8. □

8.3.2 The Strong Law of Large Numbers

The first application of the Robbins-Siegmund theorem we shall discuss will be Kolmogorov's Strong Law of large numbers:

Through (6.10) and the quantitative Kronecker's lemma (in the form of Lemma 6.3.19), one can obtain rates for Kolmogorov's strong law of large numbers given a rate of convergence for the sum (8.2). Theorem 8.3.9 also allows us to obtain rates.

Theorem 8.3.10. *Let $\{Z_n\}$ be a sequence of independent random variables, each having 0 expected value. Suppose*

$$\sum_{n=1}^{\infty} \frac{\mathrm{Var}(Z_n)}{n^2} < M \tag{8.2}$$

for some $M > 0$ and we have a function ϕ such that for all $\varepsilon > 0$

$$\sum_{n=\phi(\varepsilon)}^{\infty} \frac{\mathrm{Var}(Z_n)}{n^2} < \varepsilon.$$

If $\mathbb{E}(Z_0^2) < K$, for some $K > 0$, then

$$\frac{1}{n}\sum_{i=1}^{n} Z_n \to 0$$

almost surely, with rate of almost sure convergence given by

$$\Delta(\lambda, \varepsilon) := \exp\left(\frac{8(K+M)}{\varepsilon^2\lambda}\right)\phi\left(\frac{\lambda\varepsilon^2}{4}\right).$$

Proof. Let $\mathcal{F}_n$ be the σ-algebra generated by $Z_0, \ldots, Z_n$. Then by independence of the random variables, we have $\mathbb{E}(Z_{n+1}|\mathcal{F}_n) = \mathbb{E}(Z_{n+1}) = 0$. We would like to to apply Theorem 8.3.9, so in

this context, setting

$$X_n := \frac{1}{n^2}\left(\sum_{i=1}^{n} Z_n\right)^2,\ u_n := \frac{(n+1)^2 - n^2}{(n+1)^2},\ C_n := \frac{1}{(n+1)^2}\mathbb{E}(Z_{n+1}^2|\mathcal{F}_n)$$

as in [129] (with $X_0 := Z_0$ and $\alpha_n :\equiv 0$), gives

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n - u_n X_n + C_n.$$

Furthermore, we can take $L := 2$ and

$$\sum_{n=1}^{\infty} \mathbb{E}(C_n) = \sum_{n=1}^{\infty} \frac{\mathbb{E}(Z_{n+1}^2)}{(n+1)^2} < M.$$

Moreover, for each $\varepsilon > 0$

$$\sum_{n=\phi(\varepsilon)-1}^{\infty} \mathbb{E}(C_n) < \varepsilon.$$

Now we take $r(n,x) := (n+1)\exp(x)$ then we have

$$\sum_{i=n}^{r(n,x)} u_i \geq \sum_{i=n}^{r(n,x)} \frac{1}{i+1} \geq \int_n^{r(n,x)} \frac{1}{x+1}dx = \log\left(\frac{r(n,x)+1}{n+1}\right) \geq x.$$

So Theorem 8.3.9 implies

$$\Phi(\varepsilon, n) := r(n, 2(K+M)/\varepsilon)$$

is a liminf-modulus for $\{\mathbb{E}(X_n)\}$ (since, in this case $V_n = X_n$). Therefore, X_n converges to 0 with rate given by

$$\Gamma(\lambda, \varepsilon) := \Phi(\lambda\varepsilon/4, \phi(\lambda\varepsilon/4) - 1).$$

Thus, $S_n/n = \sqrt{X_n}$ converges to 0 with a rate of almost sure convergence given by $\Gamma(\lambda, \varepsilon^2)$ and the result follows by simplifying. □

Remark 8.3.11. The above theorem demonstrates the versatility of Theorem 8.3.9; however, the exponential rates obtained are worse than those one can calculate through (6.10) and Kronecker's lemma.

8.3.3 Rates for the Robbins-Monro algorithm

Suppose $\{Y(x) : x \in \mathbb{R}\}$ is a family of random variables with finite mean, and the function $M(x) := \mathbb{E}(Y(x))$ has a root at θ. If we assume further that for all $x < \theta$, we have $M(x) < 0$ and for $x > \theta$, we have $M(x) > 0$; then we can approximate θ by means of an interval

bisection method, which will be computable from M. Suppose, however, we do not have access to M. The *Robbin-Monro procedure* [128] was the first in a class of stochastic algorithms to solve this problem. In [128], the random variables $Y(x)$ are assumed to be uniformly (in x) bounded, and thus M is also assumed to be bounded. The assumption of boundedness was then weakened by Blum [19]. Blum showed that one only needed that the variances of $Y(x)$ were uniformly bounded and that M could be bounded by a linear function, as long as the additional requirement that M could not get arbitrarily close to 0 on intervals of arbitrarily long lengths, more precisely:

$$\inf_{\varepsilon<|x-\theta|<\varepsilon^{-1}} |M(x)| > 0$$

for all $\varepsilon > 0$.

One of the applications of the Robbins-Siegmund theorem, given in [129], is a generalisation of Blum's result that weakens the requirement of uniformly bounded variance. We demonstrate how Theorem 8.3.6 allows us to obtain rates for this generalisation:

Theorem 8.3.12. *Let $\{Y(x) : x \in \mathbb{R}\}$ be a family of random variables with $M(x) := \mathbb{E}(Y(x))$ and $\sigma(x) := \mathrm{Var}(Y(x))$ finite for all $x \in \mathbb{R}$ and measurable. Assume further that there exists $a, b > 0$ such that for all $x \in \mathbb{R}$,*

$$\sigma(x) + |M(x)| \leq a + b|x|,$$

and suppose there exists $\theta \in \mathbb{R}$ and $F : (0, \infty) \to (0, \infty)$ such that for all $\varepsilon > 0$

$$\inf_{\varepsilon<|x-\theta|<\varepsilon^{-1}} |M(x)| > F(\varepsilon). \tag{8.3}$$

Furthermore, assume we have $x < \theta$ implies $M(x) < 0$ and $x > \theta$ implies $M(x) > 0$.

Let $\{a_n\}$ be a sequence of nonnegative random variables, such that

$$\sum_{n=0}^{\infty} a_n^2 < \infty$$

almost surely with rate of uniform boundedness β and

$$\sum_{n=0}^{\infty} a_n = \infty$$

almost surely with rate of divergence r (as in Definition 8.3.3).

Define $\{x_n\}$ recursively via

$$x_{n+1} = x_n - a_n y_n$$

with x_0 arbitrary and $\{y_n\}$ a sequence of independent random variables with respective distri-

butions the same as $\{Y(x_n)\}$. *In addition, suppose we have* $K > 0$ *such that* $\mathbb{E}((x_0-\theta)^2) < K$. *Then* $\{x_n\}$ *converges to* θ *almost surely, with rate of uniform metastable convergence*

$$\Delta(\lambda, \varepsilon, g) := \tilde{f}^{\phi_{K,\rho,\tau}(\lambda/2,\varepsilon^2/2)}(0)$$

with

$$\tilde{f}(j) := j + \max\{g(j), \Phi(\lambda/4, 2\delta F(\delta), j) - j\},$$

for

$$\delta := \min\left\{\frac{\sqrt{\varepsilon}}{2\sqrt{2}}, \frac{1}{\sqrt{\nu_{K,\rho,\tau}(\lambda/4)}}\right\},$$

and

$$\Phi(\lambda, \varepsilon, n) := r(\lambda/2, \omega_{K,\rho,\tau}(\lambda/2)/\varepsilon, n).$$

Here, $\phi_{K,\rho,\tau}, \chi_{K,\rho,\tau}$ *and* $\nu_{K,\rho,\tau}$ *are as defined in Theorem 8.2.3 and Corollary 8.2.4 with*

$$\rho(\lambda) := \exp(4b^2\beta(\lambda))$$
$$\tau(\lambda) := 2(a^2 + 2b^2\theta^2)\beta(\lambda).$$

Proof. As in [129], defining $\mathcal{F}_n$ to be the σ-algebra generated by $x_0, y_0, \ldots, x_{n-1}, y_{n-1}$, $X_n := (x_n - \theta)^2$, $A_n := 4b^2a_n^2$, $C_n := 2a_n^2(a^2 + 2b^2\theta^2)$, $U_n := a_n$ and $V_n := 2|x_n - \theta||M(x_n)|$ yields,

$$\mathbb{E}(X_{n+1} \mid \mathcal{F}_n) \leq (1 + A_n)X_n - U_nV_n + C_n \quad \text{for all } n \in \mathbb{N}.$$

Moreover

$$\mathbb{P}\left(\sum_{n=0}^{\infty} C_n \geq \tau(\lambda)\right) < \lambda$$

and

$$\mathbb{P}\left(\prod_{n=0}^{\infty}(1 + A_n) \geq \rho(\lambda)\right) < \lambda,$$

with the former inequality following from the fact that $\exp(x) \geq 1 + x$ for all $x \in \mathbb{R}$.

Therefore, Theorem 8.3.6 implies

$$\Phi(\lambda, \varepsilon, n) := r(\lambda/2, \chi_{K,\rho,\tau}(\lambda/2)/\varepsilon, n)$$

is a liminf modulus for $\{V_n\}$, so

$$\mathbb{P}\left(\forall k \in [n; \Phi(\lambda, \varepsilon, n)](2|x_k - \rho||M(x_k)| \geq \varepsilon)\right) < \lambda$$

and Corollary 8.2.4 implies

$$\mathbb{P}\left(\sup_{n\in\mathbb{N}} X_n \geq \nu_{K,\rho,\tau}(\lambda)\right) < \lambda.$$

Now for

$$\Psi(\lambda,\varepsilon,n) := \Phi(\lambda/2, 2\tilde{\delta}F(\tilde{\delta}), n)$$

with

$$\tilde{\delta} := \min\left\{\frac{\sqrt{\varepsilon}}{2}, \frac{1}{\sqrt{\nu\left(\lambda/2\right)}}\right\}$$

we have,

$$\begin{aligned}
&\mathbb{P}\left(\forall k \in [n;\Psi(\lambda,\varepsilon,n)](X_k \geq \varepsilon)\right)\\
&\leq \mathbb{P}\left(\forall k \in [n;\Psi(\lambda,\varepsilon,n)](|x_k-\theta| > \sqrt{\varepsilon}/2) \wedge \sup_{n\in\mathbb{N}}|x_k-\theta| < \sqrt{\nu\left(\frac{\lambda}{2}\right)}\right) + \frac{\lambda}{2}\\
&\leq \mathbb{P}\left(\forall k \in [n;\Psi(\lambda,\varepsilon,n)](\tilde{\delta}^{-1} > |x_k-\theta| > \tilde{\delta})\right) + \frac{\lambda}{2}\\
&\leq \mathbb{P}\left(\forall k \in [n;\Psi(\lambda,\varepsilon,n)](2|x_k-\theta||M(x_k)| \geq 2\tilde{\delta}F(\tilde{\delta}))\right) + \frac{\lambda}{2} < \lambda.
\end{aligned}$$

Thus, Ψ is a liminf modulus for $\{X_n\}$. Therefore, Theorem 8.3.6 implies that $\{X_n\}$ converges to 0 almost surely with rate of uniform metastable convergence given by

$$\Gamma(\lambda,\varepsilon,g) := \tilde{f}^{\phi_{K,\rho,\tau}(\lambda/2,\varepsilon/2)}(0) \quad \text{for} \quad \tilde{f}(j) := j + \max\{g(j), \Psi(\lambda/2,\varepsilon/2,j) - j\}$$

and so $\{x_n\}$ converges to θ almost surely with rate of uniform metastable convergence given by $\Gamma(\lambda,\varepsilon^2,g)$ and the result follows by simplification. □

Remark 8.3.13. Condition (8.3) is a computational interpretation of the condition, due to Blum [19], that

$$\inf_{\varepsilon<|x-\theta|<\varepsilon^{-1}} |M(x)| > 0$$

for all $\varepsilon > 0$.

Remark 8.3.14. It does not seem possible to obtain rates of almost sure convergence for the previous presentation of the Robbins-Monro algorithm, using Theorem 8.3.9 (as we did for the Strong Law of Large Numbers in Section 8.3.2). However, our present procedure is one of many generalisations of the original stochastic approximation procedures produced by Robbins and Monro [128]. We anticipate that rates of almost sure convergence can be computed for a weakened version of the Robbins-Monron procedure than that presented in [129], through a modification of Theorem 8.3.9. This is current work in progress with Pischke and Powell.

8.3.4 Rates for Dvoretzky's algorithm

A different stochastic approximation method that follows from the Robbins-Siegmund theorem is a generalisation of a method due to Dvoretzky [36].

Dvoretzky's theorem is a key result in the area of stochastic approximations. The theorem generalised the very influential Kiefer-Wolfowitz procedure [73] and the Robbins-Monro procedure [128].

Robbins and Siegmund extended Dvoretzky's result to random variables taking values in Hilbert spaces. This does not easily follow from Dvoretzky's original proof in [36] as he uses crucial properties of the real numbers (for example, the use of the sign of real numbers), and we now present a quantitative version of this result:

Theorem 8.3.15. *Suppose X is a Hilbert space. Writing X^n as the the n times Cartesian product of X, suppose we have nonnegative sequences of real numbers $\{a_n\}, \{b_n\}, \{c_n\}$ and a Borel measurable function $T_n : X^n \to X$ (for $n = 1, 2, 3 \ldots$), satisfying:*

$$\|T_n(x_0, \ldots, x_{n-1})\| \leq \max\{a_{n-1}, (1 + b_{n-1})\|x_{n-1}\| - c_{n-1}\}. \tag{8.4}$$

Furthermore, suppose we have random variables $\{y_n\}$ and x_0, taking values from X satisfying

$$\mathbb{E}(y_n|\mathcal{F}_n) = 0 \text{ for each } n$$

where $\mathcal{F}_n := \sigma(x_0, y_0, \ldots, y_{n-1})$ is the Borel σ-algebra generated by the random variables $x_0, y_0, \ldots, y_{n-1}$ (with $\mathcal{F}_0 := \sigma(x_0)$). Dvoretzky's iterative algorithm sets

$$x_{n+1} := T_{n+1}(x_0, \ldots, x_n) + y_n. \tag{8.5}$$

Suppose we have ϕ_0, r satisfying

$$a_n \to 0 \text{ with a rate of convergence } \phi_0$$

$$\sum_{n=0}^{\infty} 2(1 + b_n)c_n = \infty \text{ with rate of divergence } r.$$

Furthermore, take $\{K_n\}$ and for each $\delta > 0$ take $L_\delta \geq 1/2, M_\delta > 0$ and ϕ_δ satisfying,

$$\forall n \in \mathbb{N}\,(\mathbb{E}(\|x_n\|^2) < K_n),\ \forall \delta > 0 \left(\prod_{i=0}^{\infty}[(1 + \delta b_n)\,(1 + b_n)^2] < L_\delta\right),$$

$$\forall \delta > 0 \left(\sum_{n=0}^{\infty}[(1 + \delta b_n)\,c_n^2 + \delta b_n\,(1 + \delta b_n) + \mathbb{E}(\|y_n\|^2)] < M_\delta\right)$$

and

$$\forall \delta > 0, \varepsilon \in (0,1] \left(\sum_{i=\phi_\delta(\varepsilon)}^{\infty} \left[(1+\delta b_n)\, c_n^2 + \delta b_n \,(1+\delta b_n) + \mathbb{E}(\|y_n\|^2) \right] < \varepsilon \right).$$

Then $\{x_n\}$ *converges to* 0 *almost surely, with rate of almost sure convergence*

$$\Gamma(\lambda, \varepsilon) := r\left(\left(\phi_\delta \left(\frac{\lambda^2 \varepsilon}{16 L_\delta} \right) - N \right)^+ + N, \frac{8 L_\delta \sqrt{L_\delta}(K_N + M_\delta)}{\lambda \varepsilon} \right)$$

with $N := \phi_0(\varepsilon/2)$ *and* $\delta := \varepsilon/2$.

Proof. We start with modifying arguments given in [129]. Fix $\delta > 0$ and first let us assume that for all $n \in \mathbb{N}$, $a_n \leq \delta$. Then, let

$$X_n := \left[(\|x_n\| - \delta)^+ \right]^2$$

and $T_n := T_n(x_0, \ldots, x_{n-1})$, (this will imply that T_n will be $\mathcal{F}_{n-1}$-measurebale). For each $n \in \mathbb{N}$ define the $\mathcal{F}_n$-measurable random variable

$$u_n := T_{n+1} I_{\{\|T_{n+1}\| \leq \delta\}} + \delta \frac{T_{n+1}}{\|T_{n+1}\|} I_{\{\|T_{n+1}\| > \delta\}}.$$

By definition, we have $\|u_n\| \leq \delta$ which implies for each $n \in \mathbb{N}$

$$X_{n+1} \leq \left[(\|x_{n+1} - u_n\| + \|u_n\| - \delta)^+ \right]^2 \leq \|x_{n+1} - u_n\|^2.$$

Therefore,

$$\begin{aligned}
\mathbb{E}(X_{n+1}|\mathcal{F}_n) &\leq \mathbb{E}(\|x_{n+1} - u_n\|^2 | \mathcal{F}_n) = \mathbb{E}(\|T_{n+1} + y_n - u_n\|^2 | \mathcal{F}_n) \\
&= \mathbb{E}(\|T_{n+1} - u_n\|^2 | \mathcal{F}_n) + \mathbb{E}(\|y_n\|^2 | \mathcal{F}_n) + 2\mathbb{E}(\langle y_n, T_{n+1} - u_n \rangle | \mathcal{F}_n) \\
&= \|T_{n+1} - u_n\|^2 + \mathbb{E}(\|y_n\|^2 | \mathcal{F}_n) = \left[(\|T_{n+1}\| - \delta)^+ \right]^2 + \mathbb{E}(\|y_n\|^2 | \mathcal{F}_n).
\end{aligned}$$

Now, from (8.4) we have

$$\|T_{n+1}\| - \delta \leq \max\{0, (1+b_n)\|x_n\| - c_n - \delta\}\}$$

and since the right hand side of the above inequality is nonnegative, we have:

$$\begin{aligned}
(\|T_{n+1}\| - \delta)^+ &\leq \max\{0, (1+b_n)\|x_n\| - c_n - \delta\}\} \\
&= \left((1+b_n)(\|x_n\| - \delta) - c_n + b_n \delta \right)^+.
\end{aligned}$$

Therefore,

$$\left[(\|T_{n+1}\| - \delta)^+\right]^2 \leq \left[(1+b_n)(\|x_n\| - \delta)^+ - c_n + b_n\delta\right]^2.$$

Now since for all $y \geq 0$ and $x \in \mathbb{R}$ we have $(x+y)^2 \leq (1+y)x^2 + y(1+y)$, taking $x := (1+b_n)(\|x_n\| - \delta)^+ - c_n$ and $y := b_n\delta$ yields,

$$\left[(\|T_{n+1}\| - \delta)^+\right]^2 \leq (1+\delta b_n)(1+b_n)^2 X_n - 2(1+b_n)c_n X_n^{\frac{1}{2}} + (1+\delta b_n)c_n^2 + \delta b_n(1+\delta b_n).$$

Therefore

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq (1+\alpha_n)X_n - u_n\sqrt{X_n} + C_n$$

with,

$$\begin{aligned}
\alpha_n &:= (1+\delta b_n)(1+b_n)^2 - 1 \\
u_n &:= 2(1+b_n)c_n \\
C_n &:= (1+\delta b_n)c_n^2 + \delta b_n(1+\delta b_n) + \mathbb{E}(\|y_n\|^2|\mathcal{F}_{n-1}).
\end{aligned}$$

Furthermore, for all $\varepsilon > 0$

$$\sum_{n=\phi_\delta(\varepsilon)}^{\infty} \mathbb{E}(C_n) < \varepsilon \text{ and } \sum_{n=0}^{\infty} \mathbb{E}(C_n) < M_\delta.$$

So applying the first part of Theorem 8.3.9 (taking $V_n := \sqrt{X_n}$) implies a liminf modulus for $\mathbb{E}(\sqrt{X_n})$ is given by

$$\Phi_\delta(\varepsilon, n) := r(n, L_\delta(K_0 + M_\delta)/\varepsilon).$$

As in the proof of Theorem 8.3.9, set

$$Z_n := \tilde{\alpha}_n X_n + \sum_{i=n}^{\infty} \tilde{\alpha}_{i+1}\mathbb{E}(C_i|\mathcal{F}_{n-1})$$

for

$$\tilde{\alpha}_n := \prod_{i=n}^{\infty}(1+\alpha_i) < L.$$

Then, by the same calculation as the proof of Theorem 8.3.9 $\{Z_n\}$ is a nonnegative supermartingale. Furthermore, as the square root function is increasing and concave, which implies $\{\sqrt{Z_n}\}$ must also be a nonnegative supermartingale. Moreover,

$$\mathbb{E}(\sqrt{Z_n}) = \mathbb{E}\left(\sqrt{\tilde{\alpha}_n X_n + \sum_{i=n}^{\infty} \tilde{\alpha}_{i+1}\mathbb{E}(C_i|\mathcal{F}_{n-1})}\right) \leq \sqrt{L}\left(\mathbb{E}(\sqrt{X_n}) + \mathbb{E}\left(\sqrt{\sum_{i=n}^{\infty}\mathbb{E}(C_i|\mathcal{F}_{n-1})}\right)\right).$$

Now, by Jensen's inequality, the above is bounded by

$$\sqrt{L}\left(\mathbb{E}(\sqrt{X_n})+\sqrt{\sum_{i=n}^{\infty}\mathbb{E}(C_n)}\right).$$

Therefore, a lim inf-modulus for $\{\mathbb{E}(\sqrt{Z_n})\}$ is given by

$$\Psi_\delta(\varepsilon,n):=\Phi_\delta\left(\frac{\varepsilon}{2\sqrt{L_\delta}},\max\left\{n,\phi\left(\frac{\varepsilon^2}{4L_\delta}\right)\right\}\right).$$

So, Proposition 8.3.8 implies $\{\sqrt{Z_n}\}$ converges to 0 with rate of almost sure convergence

$$\Delta_\delta(\lambda,\varepsilon):=\Psi_\delta(\varepsilon\lambda,0).$$

This implies that $\{Z_n\}$ converges to 0 with rate of almost sure convergence

$$\Delta_\delta(\lambda,\sqrt{\varepsilon})=r\left(\phi\left(\frac{\varepsilon\lambda^2}{4L_\delta}\right),\frac{2L_\delta\sqrt{L_\delta}(K_0+M_\delta)}{\lambda\sqrt{\varepsilon}}\right)$$

which will be the same rate as $\{X_n\}$ to 0 as $Z_n\geq X_n$. Now, for each $N\in\mathbb{N}$, shift all of the sequences in the theorem by N (so x_n now becomes x_{n+N} and so on) and denote the new sequences with a superscript N. This would mean that if $a_n^N\leq\delta$, then the sequence $\{X_n^N\}:=\{X_{n+N}\}$ converges to 0 almost surely with rate

$$\Delta_\delta^N(\lambda,\varepsilon)=r\left(\left(\phi_\delta\left(\frac{\lambda^2\varepsilon}{16L_\delta}\right)-N\right)^+ +N,\frac{2L_\delta\sqrt{L_\delta}(K_N+M_\delta)}{\lambda\sqrt{\varepsilon}}\right)-N.$$

Now, let $\varepsilon,\lambda>0$ be given and set $N:=\phi_0(\varepsilon/2)$ then for all $n\in\mathbb{N}$, $a_n^N\leq\delta:=\varepsilon/2$ and thus $\{X_n^N\}$ converges to 0 with the above rate. Therefore, we have

$$\mathbb{P}\left(\sup_{n\geq\Delta_\delta^N(\lambda,\varepsilon^2/4)+N}\|x_n\|\geq\varepsilon\right)\leq\mathbb{P}\left(\sup_{n\geq\Delta_\delta^N(\lambda,\varepsilon^2/4)}\left[(\|x_{n+N}\|-\varepsilon/2)^+\right]^2\geq\varepsilon^2/4\right)<\lambda$$

and thus a rate of almost sure convergence for $\{\|x_n\|\}$ to 0 is given by

$$\Delta_\delta^N(\lambda,\varepsilon^2/4)+N.$$

and the result follows from simplification. □

Remark 8.3.16. The above result is not the typical way Dvoretsky's procedure is presented. For example, in the original paper the procedure appears [36], one assumes $\sum_{n=0}^{\infty}b_n<\infty$,

$\sum_{n=0}^{\infty} c_n = \infty$ and $\sum_{n=0}^{\infty} \mathbb{E}(\|y_n\|^2) < \infty$ instead of

$$\begin{aligned}
&\prod_{n=0}^{\infty}[(1+\delta b_n)\,(1+b_n)^2] < \infty \\
&\sum_{n=0}^{\infty}[(1+\delta b_n)\,c_n^2 + \delta b_n\,(1+\delta b_n) + \mathbb{E}(\|y_n\|^2)] < \infty \\
&\sum_{n=0}^{\infty} 2(1+b_n)c_n = \infty
\end{aligned} \tag{8.6}$$

for each $\delta > 0$. However, as in the proof presented in [129], we may replace the sequence $\{c_n\}$ with one (which can be explicitly calculated in terms of $\{c_n\}$) satisfying (8.4) and $\sum_{i=0}^{\infty} c_i = \infty$ but also $\sum_{i=0}^{\infty} c_i^2 < \infty$ and for this new sequence and the original conditions from [36] we obtain the conditions (8.6). Furthermore, given suitable computational interpretations for the conditions from [36], one can construct the required L_δ, M_δ and ϕ_δ in the quantitative theorem we present above.

Remark 8.3.17. We note that analysing different proofs of Dvoretsky's theorem may lead to better rates than those presented in this thesis. It was communicated to the author, before they obtained their rate for Dvoretsky's theorem, by Arthan and Oliva that rates of almost sure convergence could be obtained through an analysis of a proof due to Derman and Sacks [31]. We anticipate such rates to be better than those presented here. However, we note that the rates we present here are more general as they hold for random variables taking values in Hilbert spaces, whereas the Derman-Sacks result is for real-valued random variables.

Chapter 9

Future work

In this concluding chapter, we outline the general wider research effort that the author is involved in to expand proof mining in probability theory as of the end of 2024. Furthermore, we identify some open questions and conjectures that naturally arise from many of the discussions in this thesis.

9.1 Extending the logical foundations of proof mining in probability theory

In section 4.1, we present a formal system for reasoning about probability contents and a metatheorem for guaranteeing the ex-tractability of very uniform computational content. This section comes from a joint work with Pischke [114]. As well as the work presented in this section [114] also presents novel intensional approaches to σ-algebras, measurable functions and integration.

Many of the quantitative theorems in this thesis were not obtained through the formal application of the metatheorem of [114] (which is a generalisation of Theorem 4.1.9); in fact, this metatheorem only guarantees the extraction of computational content for theorems concerning bounded random variables (due to the way majorizability is defined), and since we provide results with weaker assumptions (such as bounds on the pth moment on the random variables) the metatheorem does not explain many of the results in this thesis.

However, the quantitative results in this thesis represent new examples where highly uniform quantitative information was possible. In these results, the use of infinite unions is still very limited and can be handled by the intensional methods of [114]; however, it appears that although boundedness is not required, some bounds on the moments of the random variables are required. These observations suggest that the methods of [114] can be extended to explain the quantitative results of this thesis:

Conjecture 9.1.1. *We conjecture that the uniformities present in this thesis can be formally explained by expanding the system from [114]. Concretely, we believe one can treat random variables as intensional objects using an abstract type. Furthermore, by investigating different notions of majorizability coming from the integrability of the random variables, we believe one can then explain those uniformities as instances of a general logical metatheorem.*

Investigating the above conjecture is current work in progress with Powell and Pischke.

9.2 Further investigation of the relationships between quantitative notions of stochastic convergence

An open question at this time is to determine the precise relationship between learnable rates of uniform convergence, learnable rates of pointwise convergence and moduli of finite fluctuations. Both learnable rates of uniform convergence and moduli of finite fluctuations are learnable rates of pointwise convergence. Furthermore, Example 4.2.24 demonstrates that there are cases in which learnable rates of pointwise convergence and moduli of finite fluctuations are not learnable rates of uniform convergence. However, all other relationships between these notions are currently open. In this regard, we make the following conjectures:

Conjecture 9.2.1. *We believe the following holds:*

(I) There is an example of a stochastic process that has a learnable rate of pointwise convergence, which is not a modulus of finite fluctuations. Therefore, in light of example 4.2.24, a modulus of finite fluctuations is a strictly stronger notion than that of a learnable rate of pointwise convergence.

(II) A learnable rate of uniform convergence is a modulus of finite fluctuations. Therefore, in light of example 4.2.24, a learnable rate of uniform convergence is a strictly stronger notion than that of a modulus of finite fluctuations.

In other words, we have the following series of strict implications

Learnable rates of uniform convergence
$\rightarrow$ *Moduli of finite fluctuations*
$\rightarrow$ *Learnable rates of pointwise convergence*

If we were able to prove, Conjecture 9.2.1 (II), then an immediate consequence of Theorem 7.3.3, for $p = 1$, would be the existence of a modulus of finite fluctuations of the form

$$\phi(\lambda, \varepsilon) := C \left(\frac{\|f\|_1}{\lambda \varepsilon} \right)^2$$

with a numerical constant $C > 0$, for the ergodic averages $\{A_n f\}$ with $f \in L_1(X)$, and this would resolve a conjecture of Ivanov in [70, Conjecture 5].

We further anticipate that the three aforementioned notions are computationally equivalent, that is, if one notion holds for a stochastic process there exists a computable process to get the other notions for the same stochastic process that is independent of the stochastic processes. Thomas Powell has sketched in personal communications how one can adapt the bar-recursive construction of a rate of metastable uniform convergence from a rate of metastable pointwise convergence in [5] to obtain a learnable rate of uniform convergence from a learnable rate of pointwise convergence without the need of bar recursion. Powell conjectures that the relationship between these notions is exponential.

9.3 The computational content of the Strong Laws of Large Numbers

Baum-Katz type rates are typically given by showing some series converges through the bounding of the series. As already seen in Example 2.3.3, given a bound on a sum, there is no general computable process to extract a rate of convergence. Thus, one avenue of study is to try to extract rates for these Baum-Katz type results or demonstrate that such rates do not exist through constructions akin to Example 6.1.1. We will gain a more descriptive picture of how these large deviation probabilities behave if such a rate can be found, and computability theory provides us with the tools to demonstrate the negative result.

This problem appears to have already been considered in passing by Erdős in [37]. In the case $r = 0$ of Theorem 6.3.3, Erdős provides an elementary proof that condition (i) implies condition (ii) (this was first demonstrated by Hsu and Robbins [64] by techniques involving Fourier analysis) as well as the converse implication. Erdős's approach to demonstrating that (i) implies (ii) was to split the sum (ii) into three parts. For two parts, Erdős calculates explicit rates of convergence that are independent of the distribution of the random variables; however, for the last part, Erdős bounds the sum, and it is unclear how one obtains a rate from this bound (that is independent of the distribution of the random variables). Thus, getting a rate from Erdős's proof is not computationally viable.

Uniform rates of convergence (uniform in that they do not depend on the distribution of the random variables) have been found for the Central Limit Theorem:

Theorem 9.3.1 (Berry [13] and Esseen [39])**.** *Let $\{X_n\}$ be iid random variables satisfying $\mathbb{E}(X_0) = 0, Var(X_0) = \sigma^2 > 0, \mathbb{E}(|X_0|^3) = \rho < \infty$. Let*

$$S_n = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}},$$

F_n be the cumulative distribution function of S_n and Φ the cumulative distribution function of the standard normal distribution. Then for all $n \in \mathbb{N}$ and $x \in \mathbb{R}$

$$|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sqrt{n}\sigma^3}$$

For some $C > 0$.

Suppose we do not assume the random variables have a finite third moment but do have finite variance. In that case, we can still deduce that $F_n(x) \to \Phi(x)$ by the Central Limit Theorem, but we do not get such uniform rates of convergence. Berry initially attempted to do this but could not and made note of his failure to do so:

> *"The author originally developed the early sections of this paper for the case of bounded variates, and is indebted to W. Feller who urged the study, in these sections, of the case of finite third order absolute moments"*[1]

A similar phenomenon appears to occur in the Strong Law of Large Numbers, as the result holds if we assume the random variables have a finite first moment. It is also unclear if one can obtain rates (independent of the distribution) by only assuming a finite first moment. However, taking one moment higher (so finite variance) allows one to obtain rates independent of the distribution of the random variables.

These observations lead us to make the following conjectures:

Conjecture 9.3.2. *We believe the following to be true:*

(I) There do not exist general rates of almost sure convergence that are independent of the distributions of the random variable for the conclusion of the Strong Law of Large Numbers if one only assumes the random variables have a finite first moment.

(II) There do not exist general rates of convergence that are independent of the distributions of the random variables for the conclusion of the Central limit theorem if one only assumes the random variables have a finite second moment.

(III) There do not exist general rates of convergence that are independent of the distributions of the random variables for the Hsu-Robbins-Erdős sum [37, 64] if one assumes the random variables have finite second moment. However, such uniform rates exist if we assume the random variables have finite third moment.

(IV) There is a general proof-theoretic explanation of when increasing the moment condition of a stochastic convergence result yields uniform computational data.

[1] From the fifth footnote of [13].

9.4 Stochastic Fejér monotonicity

Quasi-Fejér monotonicity is a crucial property leveraged in proofs of convergence of algorithms in optimization and quantitative results concerning sequences of elements in abstract spaces satisfying quasi-Fejér monotonicity properties have been heavily investigated by Kohlenbach and collaborators [88, 89] and Pischke [122].

The main intended applications for Theorem 8.2.3 is to extend the work in the deterministic case by providing quantitative results for stochastic quasi-Fejér monotone sequences.[2]

Following the presentation of [27], we say that a stochastic process $\{x_n\}$ in some Hilbert space X is quasi-Fejér monotone with respect to the set Z if for all $z \in Z$ there exist random sequences $\{A_n(z)\}$, $\{B_n(z)\}$ and $\{C_n(z)\}$ adapted to $\{\mathcal{F}_n\}$ such that

$$\mathbb{E}(\phi(\|x_{n+1} - z\|) \mid \mathcal{F}_n) \leq (1 + A_n(z))\phi(\|x_n - z\|) - B_n(z) + C_n(z)$$

for all $n \in \mathbb{N}$, where $\phi : [0, \infty) \to [0, \infty)$ is a strictly increasing function with $\lim_{t\to\infty} \phi(t) = \infty$ and $\prod(1 + A_n(z)), \sum C_n(z) < \infty$ almost surely. In [27], the Robbins-Siegmund theorem plays a central role in establishing the convergence of $\{x_n\}$ under additional assumptions, which can then, in turn, be used to prove convergence of several concrete iterative algorithms in Hilbert spaces. We make the following conjectures:

Conjecture 9.4.1. *We believe the following to be true:*

(I) There exist general quantitative convergence results (in the form of uniform rates of metastability) for stochastic quasi-Fejér monotone sequences. Furthermore, such quantitative results will be immediately applicable to concrete stochastic algorithms, including, for example, the block-coordinate fixed point algorithms of [27].

(II) There are rates of almost sure convergence for the Robbins-Monro procedure we gave uniform metastable rates for in Section 8.3.3.

(III) Obtaining the right extension of so-called moduli of regularity (c.f. [89]) will allow us to extract rates of almost sure convergence for the main result of [27] and its applications.

Investigating the above conjectures is current work in progress with Powell and Pischke.

[2]This notion was studied in a simplified form in the context of generalised gradient descent methods in [38] and reintroduced in a more general setting in [27].

Concluding remarks

This thesis provides several contributions to the quantitative aspects of stochastic processes and Laws of Large Numbers from a proof-theoretic perspective. The deep analysis of results related to sequences of real numbers that satisfy recursive inequalities inspired a pursuit for a computational interpretation of the Robbins-Siegmund theorem; a key result in the convergence of stochastic algorithms across various contexts.

The standard proof of the Robbins-Siegmund theorem is nonconstructive and relies heavily on advanced concepts from martingale theory, with Doob's martingale convergence theorem being essential for establishing the result. Since the Robbins-Siegmund theorem generalises the monotone convergence theorem, Specker's construction [137] demonstrates that the theorem itself is computationally ineffective. Specifically, we cannot determine rates of almost sure convergence, which are important computational interpretations often referenced in probability theory literature.

The concept of rates of metastability for sequences of real numbers has been identified as a natural computational interpretation of convergence that can be extracted from large classes of sequences whose convergence is established through nonconstructive methods. This notion of metastability has also been extended to the stochastic setting by Avigad and his collaborators [5, 6]. Therefore, we hoped to establish such rates for the Robbins-Siegmund theorem.

The first step in analysing the Robbins-Siegmund theorem involved studying the computational aspects of Doob's convergence result. The use of upcrossing inequalities in Doob's convergence theorem and the pointwise ergodic theorem led us to explore the quantitative analysis presented by Avigad, Gerhardy, and Towsner [6]. A detailed examination of their work allowed us to develop many concepts from it.

Additionally, investigating the ideas from [6] and [5] in an abstract way proved to be very beneficial. This study facilitated the development of the concepts of uniform and pointwise learnability, which serve as natural measures of stochastic fluctuations.

Ultimately, this approach enabled us to address the computational aspects of the martingale convergence theorem. Moreover, through our abstract analysis, we were able to generalise and improve upon known bounds discussed in the survey by [70].

The success of our abstract analysis in obtaining a computational interpretation of the mar-

tingale convergence theorem led us to believe that examining probability theory more abstractly could deepen our understanding of the field. Consequently, we began a formal investigation into logical systems for reasoning about the computational aspects of probability theory, in collaboration with Pishchke. A key observation from this investigation was that significant progress in probability theory can be achieved using only finite unions. Specifically, all the quantitative results in the proof mining literature rely solely on finite unions, while the quantitative findings in the broader probability theory literature utilize infinite unions in a tameable way from a proof-theoretic perspective. Additionally, a novel extension of Bezem's majorization provides an explanation for the empirical observation that quantitative data derived from proofs in probability theory tends to be independent of the sample space and probability measure.

Equipped with our computational interpretation of the martingale convergence theorem and a deeper understanding of the quantitative aspects of probability theory from our formal work with Pischke, we set out to derive a computational interpretation of the Robbins-Siegmund theorem. While exploring potential applications of our quantitative Robbins-Siegmund theorem, we undertook an in-depth computational investigation of the Strong Laws of Large Numbers.

As our research progressed, it became evident that the literature surrounding quantitative Strong Laws of Large Numbers was extensive. It was important to comprehend this body of work to effectively position the rates we could potentially derive for the Strong Law of Large Numbers, as facilitated by our quantitative Robbins-Siegmund theorem, within the larger context of quantitative investigations into limit theorems.

As a result of our comprehensive study on the Strong Laws of Large Numbers, we were able to make several contributions, including improving existing bounds in the literature. Moreover, we provided insights into the computability theory related to the Strong Laws of Large Numbers.

In the end, we successfully developed a quantitative version of the Robbins-Siegmund theorem. Additionally, the rate we obtained is in a particularly clean form, which was unexpected to me, given the complexity of the proof. This rate can also be demonstrated to be optimal in a suitable sense. Achieving a quantitative version of the Robbins-Siegmund theorem opens the door to numerous applications in stochastic optimization. In this thesis, we presented several of these applications, including the Robbins-Monro procedure and Dvoretsky's theorem. However, we anticipate many more applications arising from our promising collaboration with Pischke and Powell.

Probability theory is a fascinating branch of mathematics that models our understanding of chance in the world around us and presents many intriguing theoretical challenges. Proof mining in probability theory has shown significant potential, highlighted by influential papers from Avigad and his collaborators. However, this research area has unfortunately stagnated for about a decade.

This thesis addresses several critical practical challenges that must be overcome to advance

proof mining in stochastic optimization and the Strong Laws of Large Numbers. It also tackles conceptual obstacles necessary for developing a formal framework to reason about the computational aspects of probability theory.

The initial work presented in this thesis is expected to be significantly developed and improved through our promising collaboration with Powell and Pischke. Additionally, we anticipate that this initial endeavour will inspire further proof-theoretic investigations in other aspects of probability theory that are not addressed in this thesis, which greatly excites me.

Index

Symbols

(BR) **28**
$C_{\alpha,s,\delta,m}$ **127**
I_A **41**
L_δ **127**
L_p **42**
$P^*_{n,\varepsilon}$ **107**
P_Y **58**
DC **23**
$\mathbb{E}$ **41**
$\mathbb{P}$ **40**
$\mathbb{P}_X$ **40**
WE-PA$^\omega$ **20**
$C_{[\alpha,\beta]}\{x_n\}$ **34**
$C_{N,[\alpha,\beta]}\{x_n\}$ **34**
$D_{[\alpha,\beta]}\{x_n\}$ **35**
$D_{N,[\alpha,\beta]}\{x_n\}$ **35**
$J_\varepsilon\{x_n\}$ **34**
$J_{N,\varepsilon}\{x_n\}$ **34**
$\mathcal{A}^\omega$ **23**
$\mathcal{A}^\omega[X, \langle\cdot,\cdot\rangle]$ **24**
$\mathcal{A}^\omega[X, \|\cdot\|]$ **23**
$\mathcal{A}^\omega[\mathrm{Int}]$ **64**
$\mathcal{B}(\mathbb{R})$ **40**
$\mathcal{F}^\omega$ **65**
$\mathcal{F}^\omega[\mathbb{P}]$ **66**
$\mathcal{M}^{\omega,X}$ **30**
$\mathcal{M}^{\omega,\Omega,S}$ **68**
$\mathcal{P}(M,l)$ **38**
$\mathcal{S}^{\omega,X}$ **29**
$\mathcal{U}^\omega$ **90**
ACA_0 **97**
RCA_0 **97**
QF-AC **23**
$\|\cdot\|_p$ **43**
∂_ε **58**
σ-**algebra** **40**
$U_{[\alpha,\beta]}\{x_n\}$ **35**
$U_{N,[\alpha,\beta]}\{x_n\}$ **35**
ε-**subdifferential** **58**
$k^+_s(m)$ **127**
$k^-_s(m)$ **127**

A

algebra of subsets **39**

B

bar recursion **28**
Bishop's upcrossing inequality **158**
bound on the crossings **35**
bound on the fluctuations **34**

C

complete convergence **129**
conditional expectation **44**
Conditional Jensen's inequality **46**

D

degree **20**
dependent choice **23**
distribution **40**
Doob's Upcrossing inequality **152**
Dvoretzky's theorem **188**

E

expected value 41

F

Fatou's Lemma 44
filtration 44
functional (Dialectica) interpretation 24

I

identically distributed 41
independent 41
indicator function 41
integrable 42
Ivanov's downcrossing inequality 158

J

Jensen's inequality 44

K

Kolmogorov's inequality 116
Kronecker's lemma 88

L

learnable rate of pointwise convergence 81
learnable rate of uniform convergence 81

M

Markov's inequality 44
martingale 46
measurable 40
metastable rate of pointwise convergence 76
metastable rate of uniform convergence 76
modulus of almost sure divergence 178
modulus of finite crossings 78
modulus of finite fluctuations 77
modulus of tightness 75
modulus of uniform boundedness 75

N

negative translation 26

P

probability content 40
probability content space 40
probability measure 40
probability space 40

Q

quantifier-free axiom of choice 23

R

Rademacher sequence 48
Rademacher type 47
Radon 47
rate of (Cauchy) convergence 31
rate of (Cauchy) metastability 33
rate of almost sure convergence 76
rate of divergence 53
Robbin-Monro procedure 185
Robbins-Siegmund theorem 165

S

sample space 40
simple function 41
space of events 40
stopping time 46
Strong Law of Large Numbers 106
submartingale 46
supermartingale 46

T

tight 47

U

upcrossings 35

V

Ville's inequality 47

W

Weakly extensional Heyting arithmetic 21
weakly extensional Peano arithmetic 19

Bibliography

[1] Ackermann, W. Zur widerspruchsfreiheit der zahlentheorie. *Mathematische Annalen 117*, 1 (1940), 162–194.

[2] Alber, Y., Iusem, A., and Solodov, M. On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. *Mathematical Programming 81*, 1 (1998), 23–35.

[3] Arthan, R., and Oliva, P. On the Borel-Cantelli Lemmas, the Erdős-Rényi Theorem, and the Kochen-Stone Theorem. *Journal of Logic and Analysis 13*, 6 (2021), 23pp.

[4] Avigad, J. The metamathematics of ergodic theory. *Annals of Pure and Applied Logic 157*, 2 (2009), 64–76.

[5] Avigad, J., Dean, E., and Rute, J. A metastable dominated convergence theorem. *Journal of Logic and Analysis 4*, 3 (2012). 19pp.

[6] Avigad, J., Gerhardy, P., and Towsner, H. Local stability of ergodic averages. *Transactions of the American Mathematical Society 362*, 1 (2010), 261–288.

[7] Avigad, J., and Rute, J. Oscillation and the mean ergodic theorem for uniformly convex Banach spaces. *Ergodic theory and dynamical systems 35*, 4 (2014), 1009–1027.

[8] Avigad, J., and Towsner, H. Metastability in the furstenberg-zimmer tower. *Fundamenta Mathematicae 210*, 3 (2010), 243–268.

[9] Bahadur, R., and Ranga, R. On deviations of the sample mean. *The Annals of Mathematical Statistics 31*, 4 (1960), 1015–1027.

[10] Bai, P., Chen, P., and Sung, S. On complete convergence and the strong law of large numbers for pairwise independent random variables. *Acta Mathematica Hungarica 142* (2014), 502–518.

[11] Baum, L., and Katz, M. Convergence rates in the law of large numbers. *Transactions of the American Mathematical Society 120*, 1 (1965), 108–123.

[12] Bauschke, H., and Combettes, P. *Convex Analysis and Montone Operator Theory in Hilbert Spaces*. Springer, 2017.

[13] Berry, A. The accuracy of the Gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society 49*, 1 (1941), 122–136.

[14] Bezem, M. Strongly majorizable functionals of finite type: a model for bar recursion containing discontinuous functionals. *The Journal of Symbolic Logic 50* (1985), 652–660.

[15] Bhaskara Rao, K., and Bhaskara Rao, M. *Theory of Charges*, vol. 109 of *Pure and Applied Mathematics*. Elsevier, 1983.

[16] Billingsley, P. *Ergodic Theory and Information*. Robert E. Krieger Publishing Co., Huntington, N.Y., 1978. Reprint of the 1965 original.

[17] Birkel, T. A note on the strong law of large numbers for positively dependent random variables. *Statistics & Probability Letters 7*, 1 (1988), 17–20.

[18] Bishop, E. *Foundations of Constructive Analysis*. Mcgraw-Hill, New York, NY, USA, 1967.

[19] Blum, J. R. Approximation methods which converge with probability one. *Annals of Mathematical Statistics 25* (1954), 382–386.

[20] Chandra, T., and Goswami, A. Cesaro uniform integrability and the strong law of large numbers. *Sankhyā: The Indian Journal of Statistics, Series A* (1992), 215–231.

[21] Chashka, A. Fluctuations in martingales. *Uspekhi Matematicheskikh Nauk 49*, 2 (1994), 179–180. English translation in Russian Math. Surveys 49:2 (1994).

[22] Chen, P., and Sung, S. A strong law of large numbers for nonnegative random variables and applications. *Statistics & Probability Letters 118* (2016), 80–86.

[23] Cheval, H., Kohlenbach, U., and Leuştean, L. On modified halpern and tikhonov–mann iterations. *Journal of Optimization Theory and Applications 197*, 1 (2023), 233–251.

[24] Cheval, H., and Leuştean, L. Quadratic rates of asymptotic regularity for the Tikhonov-Mann iteration. *Optimization Methods and Software* (2022). Electronic publication ahead of print.

[25] Chow, Y. Delayed sums and Borel summability of independent, identically distributed random variables. *Bulletin of the Institute of Mathematics, Academia Sinica 1*, 2 (1973), 207–220.

[26] Chung, K.-L. Note on Some Strong Laws of Large Numbers. *American Journal of Mathematics 69*, 1 (1947), 189–192.

[27] Combettes, P. L., and Pesquet, J.-C. Stochastic Quasi-Fejér Block-Coordinate Fixed Point Iterations with Random Sweeping. *SIAM Journal on Optimization 25*, 2 (2015), 1221–1248.

[28] Cramér, H. On a new limit theorem of probability theory. *Actualités scientifiques et industrielles 736* (1938), 5–23.

[29] Csörgő, S., Tandori, K., and Totik, V. On the strong law of large numbers for pairwise independent random variables. *Acta Mathematica Hungarica 42* (1983), 319–330.

[30] Delporte, J. Almost surely continuous random functions on a closed interval. In *Annals of the IHP Probabilities and Statistics* (1964), pp. 111–215.

[31] Derman, C., and Sacks, J. On Dvoretsky's stochastic approximation theorem. *The Annals of Mathematical Statistics 30*, 2 (1959), 601–606.

[32] Dinis, B., and Pinto, P. Quantitative Results on the Multi-Parameters Proximal Point Algorithm. *Journal of Convex Analysis 28*, 3 (2021), 729–750.

[33] Doob, J. *Stochastic Processes*. New York: Wiley, 1953.

[34] Doob, J. Notes on Martingale Theory. In *Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, 1961, pp. 95–102.

[35] Durrett, R. *Probability: theory and examples*, vol. 49. Cambridge university press, 2019.

[36] Dvoretsky, A. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (1956), vol. 1, University of California Press, pp. 39–56.

[37] Erdős, P. On a Theorem of Hsu and Robbins. *The Annals of Mathematical Statistics 20*, 2 (1949), 286 – 291.

[38] Ermol'ev, Y. M. On the method of generalized stochastic gradients and quasi-Fejér sequences. *Cybernetics 5* (1969), 208–220.

[39] Esseen, C.-G. On the Liapunov limit error in the theory of probability. *Arkiv för matematik, astronomi och fysik 28* (1942), 1–19.

[40] Etemadi, N. An elementary proof of the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 55*, 1 (1981), 119–122.

[41] Fill, J. Convergence Rates Related to the Strong Law of Large Numbers. *The Annals of Probability 11*, 1 (1983), 123 – 142.

[42] Franci, B., and Grammatico, S. Convergence of sequences: A survey. *Annual Reviews in Control 53* (2022), 161–186.

[43] Freund, A., and Kohlenbach, U. Bounds for a nonlinear ergodic theorem for Banach spaces. *Ergodic Theory and Dynamical Systems 43*, 5 (2023), 1570–1593.

[44] Gaposhkin, V. Some Examples of the Problem of ε-Deviatations for Stationary Sequences. *Theory of Probability & Its Applications 46*, 2 (2002), 341–346.

[45] Gentzen, G. The consistency of pure number theory. *Mathematical Annals 112*, 1 (1936), 493–565.

[46] Gentzen, G. Die widerspruchsfreiheit der reinen zahlentheorie. *Mathematische annalen 112*, 1 (1936), 493–565.

[47] Gerhardy, P. Proof mining in topological dynamics. *Notre Dame Journal of Formal Logic 49*, 4 (2008).

[48] Gerhardy, P., and Kohlenbach, U. General logical metatheorems for functional analysis. *Transactions of the American Mathematical Society 360*, 5 (2008), 2615–2660.

[49] Girard, J.-Y. *Proof theory and logical complexity. Volume I*, vol. 1 of *Stud. Proof Theory, Monogr.* Bibliopolis, Edizioni di Filosofia e Scienze, Napoli, 1987.

[50] Gödel, K. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für mathematik und physik 38* (1931), 173–198.

[51] Gödel, K. On intuitionistic arithmetic and number theory. *Results of a mathematical colloquium 4*, 1933 (1933), 34–38.

[52] Gödel, K. Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes. *Dialectica 12*, 3–4 (1958), 280–287.

[53] Gödel, K. On formally undecidable propositions of principia mathematica and related systems i 1 (1931). In *Godel's Theorem in Focus*. Routledge, 2012, pp. 17–47.

[54] Golubov, B., Efimov, A., and Skvortsov, V. *Walsh series and transforms: theory and applications*, vol. 64. Springer Science & Business Media, 2012.

[55] Günzel, D., and Kohlenbach, U. Logical metatheorems for abstract spaces axiomatized in positive bounded logic. *Advances in Mathematics 290* (2016), 503–551.

[56] Gut, A. *Probability: A Graduate Course*. Springer Texts in Statistics. Springer New York, New York, 2013.

[57] Hájek, J., and Rényi, A. Generalization of an inequality of Kolmogorov. *Acta Mathematica Hungarica 6*, 3-4 (1955), 281–283.

[58] Hilbert, D. Über das unendliche. *Mathematische Annalen 95*, 1 (1926), 161–190.

[59] Hilbert, D. *Mathematische probleme*. Springer, 1935.

[60] Hilbert, D. Mathematical problems. *Bulletin-American Mathematical Society 37*, 4 (2000), 407–436.

[61] Hilbert, D., and Bernays, P. *Grundlagen der Mathematik, vol. II*. Springer Berlin, 1939.

[62] Hoffmann-Jørgensen, J., and Pisier, G. The law of large numbers and the central limit theorem in Banach spaces. *The Annals of Probability* (1976), 587–599.

[63] Howard, W. Hereditarily majorizable functionals of finite type. In *Metamathematical Investigation of Intuitionistic Arithmetic and Analysis*, A. Troelstra, Ed., vol. 344 of *Lecture Notes in Mathematics*. Springer, New York, 1973, pp. 454–461.

[64] Hsu, P., and Robbins, H. Complete convergence and the law of large numbers. *Proceedings of the national academy of sciences 33*, 2 (1947), 25–31.

[65] Ivanov, V. Oscillations of averages in the ergodic theorem. *Doklady Akademii Nauk 347*, 6 (1996), 736–738.

[66] Jabbari, H. On almost sure convergence for weighted sums of pairwise negatively quadrant dependent random variables. *Statistical Papers 54* (2013), 765–772.

[67] Jones, R., Kaufman, R., Rosenblatt, J., and Wierdl, M. Oscillation in ergodic theory. *Ergodic Theory and Dynamical Systems 18*, 4 (1998), 889–935.

[68] Jones, R., Ostrovskii, I., and Rosenblatt, J. Square functions in ergodic theory. *Ergodic Theory & Dynamical Systems 16* (1996), 267–305.

[69] Jones, R., Rosenblatt, J., and Wierdl, M. Counting in ergodic theory. *Canadian Journal of Mathematics 51*, 5 (1999), 996–1019.

[70] Kachurovskii, A. The rate of convergence in ergodic theorems. *Russian Mathematical Surveys 51*, 4 (1996), 653–703.

[71] Kachurovskii, A., and Podvigin, I. Measuring the rate of convergence in the Birkhoff ergodic theorem. *Mathematical Notes 106* (2019), 52–62.

[72] Kachurovskii, A., Podvigin, I., and Svishchev, A. The maximum pointwise rate of convergence in Birkhoff's ergodic theorem. *Journal of Mathematical Sciences 255*, 2 (2021).

[73] Kiefer, J., and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics* (1952), 462–466.

[74] Kim, T.-S., and Baek, J. The strong laws of large numbers for weighted sums of pairwise quadrant dependent random variables. *Journal of the Korean Mathematical Society 36*, 1 (1999), 37–49.

[75] Kohlenbach, U. *Theorie der majorisierbaren und stetigen Funktionale und ihre Anwendung bei der Extraktion von Schranken aus inkonstruktiven Beweisen: Effektive Eindeutigkeitsmodule bei besten Approximationen aus ineffektiven Eindeutigkeitsbeweisen.* PhD thesis, Goethe-Universität Frankfurt am Main, 1990.

[76] Kohlenbach, U. Effective bounds from ineffective proofs in analysis: an application of functional interpretation and majorization. *The Journal of Symbolic Logic 57* (1992), 1239–1273.

[77] Kohlenbach, U. Effective moduli from ineffective uniqueness proofs. An unwinding of de La Vallée Poussin's proof for Chebycheff approximation. *Annals of Pure and Applied Logic 64*, 1 (1993), 27–94.

[78] Kohlenbach, U. New effective moduli of uniqueness and uniform a priori estimates for constants of strong unicity by logical analysis of known proofs in best approximation theory. *Numerical Functional Analysis and Optimization 14*, 5-6 (1993), 581–606.

[79] Kohlenbach, U. Some logical metatheorems with applications in functional analysis. *Transactions of the American Mathematical Society 357*, 1 (2005), 89–128.

[80] Kohlenbach, U. *Applied Proof Theory: Proof Interpretations and their Use in Mathematics.* Springer Monographs in Mathematics. Springer, 2008.

[81] Kohlenbach, U. Recent progress in proof mining in nonlinear analysis. *IFCoLoG Journal of Logics and their Applications 10*, 4 (2017), 3361–3410.

[82] Kohlenbach, U. Proof-theoretic methods in nonlinear analysis. In *Proceedings of the International Congress of Mathematicians 2018*, vol. 2. World Scientific, 2019, pp. 61–82.

[83] Kohlenbach, U. Kreisel's 'shift of emphasis' and contemporary proof mining. *Studies in Logic 16*, 3 (2023), 1–35.

[84] Kohlenbach, U., and Körnlein, D. Effective rates of convergence for Lipschitzian pseudocontractive mappings in general Banach spaces. *Nonlinear Analysis 74* (2011), 5253–5267.

[85] Kohlenbach, U., and Lambov, B. Bounds on iterations of asymptotically quasi-nonexpansive mappings. In *Proceedings of the International Conference on Fixed Point Theory and Applications* (2004), Yokohama Publishers, pp. 143–172.

[86] Kohlenbach, U., and Leuştean, L. Asymptotically nonexpansive mappings in uniformly convex hyperbolic spaces. *Journal of the European Mathematical Society 12*, 1 (2010), 71–92.

[87] Kohlenbach, U., and Leuştean, L. Effective metastability of Halpern iterates in CAT(0) spaces. *Advances in Mathematics 321* (2012), 2526–2556.

[88] Kohlenbach, U., Leuştean, L., and Nicolae, A. Quantitative results on Fejér monotone sequences. *Communications in Contemporary Mathematics 20* (2018), 42pp.

[89] Kohlenbach, U., López-Acedo, G., and Nicolae, A. Moduli of regularity and rates of convergence for Fejér monotone sequences. *Israel Journal of Mathematics 232* (2019), 261–297.

[90] Kohlenbach, U., and Oliva, P. Proof mining in L1-approximation. *Annals of Pure and Applied Logic 121*, 1 (2003), 1–38.

[91] Kohlenbach, U., and Powell, T. Rates of convergence for iterative solutions of equations involving set-valued accretive operators. *Computers and Mathematics with Applications 80* (2020), 490–503.

[92] Kohlenbach, U., and Safarik, P. Fluctuations, effective learnability and metastability in analysis. *Annals and Pure and Applied Logic 165* (2014), 266–304.

[93] Kolmogorov, A. O principe tertium non datur. *mathematicheskij sbornik 32* (1925), 646–667.

[94] Kolmogorov, A. Sur la loi forte des grands nombres. *Comptes rendus de l'Académie des Sciences 191* (1930), 910–912.

[95] Korchevsky, V. A generalization of the Petrov strong law of large numbers. *Statistics & Probability Letters 104* (2015), 102–108.

[96] Korchevsky, V. On the Rate of Convergence in the Strong Law of Large Numbers for Nonnegative Random Variables. *Journal of Mathematical Sciences 229*, 6 (2018), 719–727.

[97] Korchevsky, V., and Petrov, V. On the strong law of large numbers for sequences of dependent random variables. *Vestnik St. Petersburg University: Mathematics 43* (2010), 143–147.

[98] Kreisel, G. On the Interpretation of Non-Finitist Proofs, Part I. *Journal of Symbolic Logic 16* (1951), 241–267.

[99] Kreisel, G. On the Interpretation of Non-Finitist Proofs, Part II: Interpretation of Number Theory. *Journal of Symbolic Logic 17* (1952), 43–58.

[100] Kreisel, G. Mathematical significance of consistency proofs. *The Journal of Symbolic Logic 23*, 2 (1958), 155–182.

[101] Kreisel, G. What have we learnt from Hilbert's second problem. *Browder [8: 1, pp. 93-130] 67* (1976), 8–12.

[102] Kreuzer, A. P. *Proof mining and combinatorics: Program extraction for Ramsey's theorem for pairs*. PhD thesis, Technische Universität Darmstadt, 2012.

[103] Kuczmaszewska, A. Convergence rate in the Petrov SLLN for dependent random variables. *Acta Mathematica Hungarica 148* (2016), 56–72.

[104] Kuroda, S. Intuitionistische Untersuchungen der formalistischen Logik. *Nagoya Mathematical Journal 2* (1951), 35–47.

[105] Ledoux, M., and Talagrand, M. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[106] Leuştean, L. Rates of Asymptotic Regularity for Halpern Iterations of Nonexpansive Mappings. *Journal of Universal Computer Science 13*, 11 (2007), 1680–1691.

[107] Leuştean, L., and Pinto, P. Quantitative results on a Halpern-type proximal point algorithm. *Computational Optimization and Applications 79*, 1 (2021), 101–125.

[108] Li, D., and Queffélec, H. *Introduction to Banach Spaces: Analysis and Probability: Volume 1*, vol. 166. Cambridge University Press, 2017.

[109] Luckhardt, H. *Extensional Gödel Functional Interpretation.* Springer Verlag, 1973.

[110] Luzia, N. A simple proof of the strong law of large numbers with rates. *Bulletin of the Australian Mathematical Society 97*, 3 (2018), 513–517.

[111] Masani, P. Measurability and Pettis integration in Hilbert spaces. In *Measure Theory* (1976), A. Bellow and D. Kölzow, Eds., Springer Berlin Heidelberg, pp. 69–105.

[112] Neri, M. A finitary Kronecker's lemma and large deviations in the Strong Law of Large Numbers on Banach spaces. *Annals of Pure and Applied Logic 176*, 6 (2025), 103569.

[113] Neri, M. Quantitative strong laws of large numbers. *Electronic Journal of Probability 30* (2025), 1–22.

[114] Neri, M., and Pischke, N. Proof mining and probability theory. Preprint, available at `https://arxiv.org/abs/2403.00659`, 2024.

[115] Neri, M., and Powell, T. A computational study of a class of recursive inequalities. *Journal of Logic and Analysis 15*, 3 (2023), 1–48.

[116] Neri, M., and Powell, T. A quantitative Robbins-Siegmund theorem. Preprint, available at `https://arxiv.org/abs/2410.15986`, 2024.

[117] Neri, M., and Powell, T. On quantitative convergence for stochastic processes: Crossings, fluctuations and martingales. *Transactions of the American Mathematical Society* (2025). To appear.

[118] Petrov, V. Generalization of Cramér's limit theorem. *Uspekhi Matematicheskikh Nauk 9*, 4 (1954), 195–202.

[119] Petrov, V. On the Strong Law of Large Numbers for Nonnegative Random Variables. *Theory of Probability & Its Applications 53*, 2 (2009), 346–349.

[120] Petrov, V., and Robinson, J. Large deviations for sums of independent non identically distributed random variables. *Communications in Statistics—Theory and Methods 37*, 18 (2008), 2984–2990.

[121] Pinto, P. Nonexpansive maps in nonlinear smooth spaces. *Transactions of the American Mathematical Society* (2024).

[122] Pischke, N. Generalized Fejér monotone sequences and their finitary content. *Optimization* (2024). To appear.

[123] Powell, T. A note on the finitization of Abelian and Tauberian theorems. *Mathematical Logic Quarterly 66*, 3 (2020), 300–310.

[124] Powell, T. A finitization of Littlewood's Tauberian theorem and an application in Tauberian remainder theory. *Annals of Pure and Applied Logic 174*, 4 (2023). 103231, 28pp.

[125] Powell, T., and Wiesnet, F. Rates of convergence for asymptotically weakly contractive mappings in normed spaces. *Numerical Functional Analysis and Optimization 42*, 15 (2021), 1802–1838.

[126] Qihou, L. Iterative sequences for asymptotically quasi-nonexpansive mappings with error member. *Journal of Mathematical Analysis and Applications 259* (2001), 18–24.

[127] Rio, E. Convergence speeds in the strong law for dependent sequences. *Proceedings of the Academy of Sciences. Series 1, Mathématics 320*, 4 (1995), 469–474.

[128] Robbins, H., and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics 22*, 3 (1951), 400–407.

[129] Robbins, H., and Siegmund, D. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*. Elsevier, 1971, pp. 233–257.

[130] Sani, H. N., Azarnoosh, H., and Bozorgnia, A. The strong law of large numbers for pairwise negatively dependent random variables. *Iranian Journal of Science 28*, 2 (2004), 211–217.

[131] Schönfinkel, M. Über die Bausteine der mathematischen Logik. *Mathematische annalen 92*, 3 (1924), 305–316.

[132] Seneta, E. A Tricentenary history of the Law of Large Numbers. *Bernoulli 19*, 4 (2013), 1088 – 1121.

[133] Siegmund, D. Large deviation probabilities in the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 31*, 2 (1975), 107–113.

[134] Simmons, W., and Towsner, H. Proof mining and effective bounds in differential polynomial rings. *Advances in Mathematics 343* (2019), 567–623.

[135] Simpson, S. *Subsystems of second order arithmetic*, vol. 1. Cambridge University Press, 2009.

[136] Sipoş, A. Revisiting jointly firmly nonexpansive families of mappings. *Optimization* (2021). Electronic publication ahead of print.

[137] Specker, E. Nicht Konstruktiv Beweisbare Sätze der Analysis. *Journal of Symbolic Logic 14* (1949), 145–158.

[138] Spector, C. Provably recursive functionals of analysis: a consistency proof of analysis by an extension of principles in current intuitionistic mathematics. In *Recursive Function Theory: Proc. Symposia in Pure Mathematics* (1962), F. D. E. Dekker, Ed., vol. 5, American Mathematical Society, pp. 1–27.

[139] Tao, T. Soft analysis, hard analysis, and the finite convergence principle. Essay posted 23 May 2007, 2007. Appeared in: 'T. Tao, Structure and Randomness: Pages from Year One of a Mathematical Blog. AMS, 298pp., 2008'.

[140] Tao, T. Norm convergence of multiple ergodic averages for commuting transformations. *Ergodic Theory and Dynamical Systems 28* (2008), 657–688.

[141] Troelstra, A., Ed. *Metamathematical Investigation of Intuitionistic Arithmetic and Analysis*, vol. 344 of *Lecture Notes in Mathematics*. Springer, Berlin, 1973.

[142] Williams, D. *Probability with Martingales*. Cambridge University Press, 1991.

[143] Woyczynski, W. Random series and laws of large numbers in some Banach spaces. *Theory of Probability & Its Applications 18*, 2 (1974), 350–355.