Are we there yet?

Advances in Anytime-Valid Methods for Hypothesis Testing and Prediction

Muriel F. Pérez-Ortiz

Pérez-Ortiz **Are we there yet?**

Advances in Anytime-Valid Methods for Hypothesis Testinhg and Prediction

Are we there yet?
Advances in anytime-valid methods for hypothesis testing and prediction

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op donderdag 6 juli 2023
klokke 15.00 uur

door

Muriel Felipe Pérez Ortiz
geboren te Bogotá, Distrito Capital, Colombia
in 1993

Promotores:

Prof.dr. P.D. Grünwald (Universiteit Leiden and
Centrum Wiskunde & Informatica)

Prof.dr. W.M. Koolen (University of Twente and
Centrum Wiskunde & Informatica)

Promotiecommissie:

Prof.dr. J. Ziegel (Universität Bern)

Prof.dr. V. Vovk (Royal Holloway University of London)

Prof.dr. M. Larsson (Carnegie Mellon University)

Prof.dr. H.J. Hupkes

Prof.dr. E.A. Verbitskiy

This work was funded by the Dutch Research Council NWO and was carried out at Centrum Wiskunde & Informatica in Amsterdam.

Universiteit Leiden
The Netherlands

CWI
Research institute for mathematics &
computer science in the Netherlands

¿Cómo fue que se pasó todo este tiempo? ¡Qué vergüenza con ustedes!

Nicolás y los Fumadores

Why do I bother over and over again trying the wrong way when the right way was staring at me all the time? I don't know.

Herbert Robbins

—It gets easier
—Huh?
—Everyday it gets a little easier
—Yeah?
—But you gotta do it every day. That's the hard part, but it gets easier
—Ok.

Bojack Horseman

Preface

This dissertation is the culminating document of my doctoral studies at the Machine Learning group of the *Centrum Wiskunde & Informatica*, in Amsterdam. It presents a number of mathematical results on statistical methods for sequential experimentation and prediction, where the decisions about future observations depend on what has been done before. The present-day interest in anytime-valid methods stems perhaps from two reasons. First, these methods are an answer to modern applications in forecasting and online experimentation that require the continuous monitoring of data—this renders classic, fixed-sample methods inapplicable. Second, and in relation to the first point, sequential methods offer a principled methodological alternative to fixed-sample methods under peeking, the common practice in scientific laboratories of checking for statistical significance during the data collection process—another barrier posed by fixed-sample methods. The results contained in this work are crossed by three intersecting axes: time, invariance and robustness.

Time. With the advent of the coronavirus disease (COVID-19) pandemic, large research efforts were driven towards finding new treatments for it. In the early days of the pandemic—before any disease-specific vaccines were available—multiple medical centers around the world were carrying randomized controlled trials on the use of the Bacillus Calmette–Guérin (BCG) vaccine, typically used against tuberculosis, to treat COVID. Remarkably, Judith ter Schure, then also a Ph.D. student at CWI, convinced several of these medical centers to perform a live meta-analysis of their data using anytime-valid methods. In order to carry this task, it was needed to develop and implement new sequential methodology for the analysis of time-to-event data, one of the classic topics in statistics since the work of David A. Cox in 1972. The ensuing work with Judith ter Schure, Alexander Ly, and Peter Grünwald is the subject Chapter 3 in this thesis; it contains methodology for the continuous montoring of time-to-event data when the survival times of two groups are being compared. This work took place predominantly at home, given the restrictions of the pandemic.

Invariance. Principles of invariance have turned out to be a very valuable tool in statistics. The t-test, perhaps the most used test on Earth, is the prototypical example of a scale-invariant test, a test that does not depend on the units of measurement of the observations. A large part of the introductory statistical theory for the inference of location parameters can be summed up in the single statement that the likelihood ratio test for the t-statistic is the overall—invariant or not—most powerful fixed-sample test. In Chapter 2, with Rianne de Heide, Tyron Lardy and Peter Grünwald, we tackle the anytime-valid counterpart of this problem, where power maximization is no longer meaningful, under more general invariances. The main results in this line of work were found during the world-wide lock-downs of 2020, but the final form of the results took much longer to reach their present form.

Robustness Will it rain tomorrow? Prediction is at the center of many tasks of modern applied research. In this line of research it is asked whether predictors can be built that perform well in the worst case—that are robust—, and work even better when data is "easy". With Wouter Koolen, we studied the simplest problem of prediction with expert advice, one of the fundamental problems in computational learning theory.

Acknowledgements

I am grateful to all of those who accompanied me in the long journey that led to the contents of this dissertation; in particular, to my supervisors Peter Grünwald and Wouter Koolen for their enthusiasm, patience, and support; to Rianne de Heide, Judith ter Schure, Alexander Ly, and Tyron Lardy, for being my coauthors and sharing so many scientific discussions; to Sébastien Gerchinovitz, for his invitation to Toulouse; to Yunda Hao, Udo Böhm, Tom Sterkenburg, and the members of the Machine Learning group, for their company and ping-pong sessions after lunch; to Vera Sarkol, Bikkie Aldeias, and Rob van Rooijen, for providing access to a remarkable mathematical library; to Nada Mitrovic, Susanne van Dam, Erik Baquedano, Duda Tepsic and Minnie Middelberg, for their support to the researchers at CWI; to Mohamed Berdouni, Luz van Fredereci, and Rob, for always serving food with a smile; to Esteban Landerreche, Jens Klooster, Giovanni Puccetti, Reneé Rietsma, Fatou Bangoura, Johana Maasen, Ismani Nieuweboer, Marina Dietrich, Pawan Gupta, Felipe Vargas, Nedime Gökmen, David Arcila and Martha Agudelo for their friendship and company; to Eduardo Pareja and Carolina Moscoso, for their affection and support; to Marina Arias, for taking care of my mother; to Uriel Pérez, my father, for his advice and serenity; and to my family. Without them, none of this would have been possible.

This work is specially dedicated to Esperanza Ortiz, my mother, who could not see this day, but would have been incredibly happy and proud holding this book in her hands.

Contents

1. Introduction 1

1.1. Statistical Hypothesis Testing 1
1.2. Sequential analysis 4
1.3. Anytime-valid testing 6
1.4. Decision, evidence, and prediction 8
1.5. Outline 10

2. E-statistics, Group Invariance, and Anytime-Valid Testing 13

2.1. Introduction 13
2.2. Technical outline 22
2.3. Assumptions 25
2.4. Main Result 28
2.5. Invariance and Sufficiency 31
2.6. Anytime-valid testing under group invariance 33
2.7. Testing multivariate normal distributions under group invariance 34
2.8. Composite invariant hypotheses 37
2.9. Discussion, Related and Future Work 40
2.10. Proof of the main theorem, Theorem 2.4.2 42
2.11. Acknowledgements 47

3. The Anytime-Valid Logrank Test 49

3.1. Introduction 49
3.2. Proportional hazards model and Cox' partial likelihood 52
3.3. The AV logrank test 54
3.4. A Gaussian approximation to the AV logrank test 62
3.5. Optional continuation and live meta-analysis 65
3.6. Anytime-valid confidence sequences 66
3.7. Power and sample size 67
3.8. Discussion, Conclusion and Future Work 67

4. Luckiness in Multiscale Online Learning 73

4.1. Introduction 73
4.2. The Muscada Multiscale Online Learning Algorithm 77
4.3. Multiscale Stochastic Luckiness 82
4.4. Optimism 83
4.5. Computation 84
4.6. Experiments on Synthetic Data 84
4.7. Discussion 85

5. Exponential Stochastic Inequality **89**
5.1. Introduction . . . 89
5.2. Basic ESI Properties . . . 95
5.3. When does a family of RVs satisfy an ESI? . . . 102
5.4. PAC-Bayes . . . 108
5.5. ESI with random η . . . 111
5.6. Non-iid Sequences . . . 114
5.7. Discusion . . . 116
5.8. Acknowledgements . . . 117

6. Discussion **119**
6.1. Anytime-valid methods . . . 119
6.2. Individual-sequence prediction . . . 121
6.3. Concentration Inequalities . . . 122

A. Appendix to Chapter 2 **125**
A.1. Computations . . . 125
A.2. Importance of the filtration for randomly stopped E-Statistics . . . 128

B. Appendix to Chapter 3 **129**
B.1. Omitted Proofs and Details . . . 129
B.2. Covariates: the full Cox Proportional Hazards E-Variable . . . 135
B.3. Gaussian AV logrank test . . . 138

C. Appendix to Chapter 4 **143**
C.1. Saddle-Point Computation in Multiscale Games . . . 143
C.2. Tuning 3 . . . 146
C.3. Algorithm Analysis . . . 148
C.4. Optimism, proof of Proposition 4.4.1 . . . 151
C.5. Luckiness . . . 152
C.6. Technical Lemmas . . . 157
C.7. Proof of Theorem 4.1.2 . . . 162

D. Appendix to Chapter 5 **167**
D.1. Proofs for Section 5.2 . . . 167
D.2. Proofs for Section 5.3.1 . . . 168
D.3. Proofs for Section 5.3.2 . . . 171
D.4. Proofs for Section 5.5 . . . 175

Samenvatting **177**

Summary **179**

Resumen **181**

Curriculum Vitae **183**

List of Publications 185

Bibliography 187

1. Introduction

Presently, data collection and processing costs are the lowest they have been in history. Both public and private organizations have, through the intensive use of data, accessed competitive advantages such as the optimization of their value chains or the capacity to offer personalized services to their clients. These advances have also changed how science is made; it is more common every day to make discoveries using data analysis. To that end, models for performing statistical hypothesis testing and prediction are crucial. This thesis presents a number of mathematical results in the theory of anytime-valid analysis for statistical hypothesis testing and prediction. Here, "anytime-valid" makes reference to the ability of continuing or stopping experimentation at any moment in time.

The purpose of this introductory chapter is twofold. First, we introduce the main topics of this thesis: hypothesis testing and prediction. To that end, in Section 1.1, we introduce the classic problem of statistical hypothesis testing as a decision problem between two probabilistic models for data. We will introduce anytime-valid methods by opposing them to classic sequential methods. Our discussion will be centered around the role played by sampling plans in experimentation: anytime-valid methods do not require them; sequential methods do. At the risk of using nonstandard terminology, we will refer collectively to anytime-valid and sequential methods as *serial.* In Section 1.2, we recall the historical context in which the field of sequential analysis appeared, during World War II, and we introduce the Sequential Probability Ratio and its sampling plan. There, because of its applications to lot inspection, the objective is that of minimizing the length of the sampling plan. In Section 2.6 we contrast the Sequential Probability Ratio Test to anytime-valid tests. This serves as an introduction to anytime-valid testing in general using sequential analysis as a point of entrance. As a connecting thread, we use a simple example that has become classic in the statistical community: testing whether a coin is biased. In Section 1.4 we briefly touch upon the issues that arise when regarding hypothesis testing as a decision problem, and the interpretation of serial procedures as prediction strategies; in particular, for gambling. Lastly, fulfilling the second purpose of this chapter, Section 1.5 outlines the ensuing chapters of this dissertation.

1.1. Statistical Hypothesis Testing

In order to illustrate the type of problems that concern us, let us first consider the standard problem of statistical hypothesis testing. Suppose that an experiment is designed to collect n observations $X_1, \dots, X_n$, and we are interested in quantifying whether the data is consistent with one of two hypotheses, $\mathcal{H}_0$ or $\mathcal{H}_1$, about the probabilistic distribution of the data. The hypothesis $\mathcal{H}_0$ may correspond to the prediction of a scientific theory or a model of the system under consideration; the alternative hypothesis $\mathcal{H}_1$, to a deviation from $\mathcal{H}_0$. For example, a collaboration of scientists investigating the effect of a new treatment for a disease may formulate their problem as that of comparing survival rates between two groups of people: a group that received the new treatment and another that received a placebo. Such a problem can be formulated in terms of a statistical hypothesis test between a baseline—no effect—and a baseline-plus-effect model for the subjects' survival rates. The process of formulating a statistical hypothesis test may require ample deliberation, and the validity of any ensuing statistical procedure will depend very much on the specifics of the experimental design, the skill of the scientists in choosing the data appropriately, and their ability to connect the observations to the principles or conceptual theories of their field. Nevertheless, we study the abstract problem of statistical hypothesis testing. By this, we refer to the problem of quantifying to what degree the collected data $X_1, \ldots, X_n$ conform to the probabilistic model hypothesized by $\mathcal{H}_0$ in comparison to $\mathcal{H}_1$. The mathematical study of such problems can be carried out with relative independence of their deployment, but their motivation is typically driven by applications. In this work, we study models that are used when data are collected serially, that is, when the decision to either continue or stop making observations may depend on what has been observed previously.

On first thought, one may try to use a fixed-sample test in a serial situation. It is well known that if the decision to continue or stop making observations is completely independent of the data that has been observed so far, no problems arise; in the presence of dependencies, this idea may fail dismally [Anscombe, 1954]. We illustrate this fact with a classic example: testing whether a coin is biased, that is, whether a coin is not equally likely to land on "heads" or "tails" when tossing it repeatedly. Under the null hypothesis, both outcomes are equally likely; under the alternative, they are not—in both cases the coin tosses are assumed to be independent of each other. In a fixed-sample experiment a coin is tossed a predetermined number n of times and the outcome of each toss $X_1, \ldots, X_n$ is registered. In introductory statistics courses it is shown that the decision to reject the null hypothesis whenever the fraction $\hat{p}_n = \frac{1}{n}\#\{\text{heads in } X_1, \ldots, X_n\}$ is outside the range $1/2 \pm 0.98/\sqrt{n}$ falsely rejects the null hypothesis—this is called a type-I error—with probability approximately 5%. The type of reasoning behind this procedure is the foundation of most statistical analysis used currently in scientific research; its correctness is important. Nevertheless, by tossing the coin according to a sampling plan that is not accounted for, the standard fixed-sample statistical procedure fails spectacularly. As an extreme example, when the coin is unbiased, if the data is collected according to the sampling plan "keep tossing the coin until $\hat{p}_n$ is outside $1/2 \pm 0.98/\sqrt{n}$", the standard fixed sample test always rejects the null hypothesis. Consequently, the probabilistic statement that "$\mathcal{H}_0$ will be falsely rejected with probability 5%" is false: the experiment executed with this sampling plan falsely rejects the hypothesis $\mathcal{H}_0$ with 100% probability. In this work, we study tests that retain type-I error control—tests that falsely reject $\mathcal{H}_0$ with small probability—irrespective of the sampling plan that is employed. The above is a simplified example of a researcher sampling until significance is reached, which is known in the literature as “sampling to reach a foregone conclusion” [Anscombe, 1954]. Relatedly, in more complicated situations, even a Bayesian approach may fail [De Heide and Grünwald, 2021].

Currently, the standard methods for error control under serial data collection are collectively known as *sequential methods*; they have become part of the statistical canon since their inception during World War II and their ensuing development [Wald, 1947, Siegmund, 1985, Lai, 2001]. Nevertheless, in this work, we make a distinction between *sequential methods*, whose experimental designs include fixed rules for continuing or stopping data collection, and *anytime-valid* methods, which do not require such rules. Thus, instead of focusing on sequential methods, we focus on anytime-valid methods, which are closely related and have received renewed attention during the last lustrum [see Ramdas et al., 2022a]. As we will see, the sequential-analytic requirement of a sampling plan is motivated by their first applications to statistical quality control. There are, however, practical reasons why one might want to focus on anytime-valid methods instead.

No sampling plan can be enforced. Despite the great success of sequential analysis, there are practical situations in which no sampling plans can be enforced. A prominent situation where this is commonplace is in meta-analysis, the task of aggregating evidence from multiple studies performed with a common goal. For instance, several small studies have been conducted on the effect of music on insomnia in adults. Even though each of the studies may not be conclusive, after gathering their findings, stronger conclusions may be reached [Jespersen et al., 2022]. In statistical terms, this is the problem of combining either the observations or other summary statistics from several studies. Crucially, the studies under analysis may not have been carried out independently; the outcomes of the first may have caused the existence of the following ones in ways that are impossible to know or model mathematically. For instance, a follow-up study may have been performed only because the first one showed a significant result. Hence, even though each study may be carefully designed and executed, the pooled observations from multiple studies do not obey any explicit sampling plan. Assuming erroneously that the pooled sample forms a statistically independent set of observations may lead, similarly to the earlier coin-tossing example, to invalid conclusions. In the absence of a sampling plan, anytime-valid methods are necessary for meta-analysis [Ter Schure and Grünwald, 2022]. More generally, anytime-valid methods open the possibility of performing statistical analysis on data that was gathered with an unknown sampling plan, provided that the serial data collection model is in accordance with how data was obtained, and that the analysis that is performed is chosen independently of the data [see Ramdas et al., 2022a, Section 6.4].

Data is serial. Currently, there exist applications where data are streamed continously and assuming that they are collected according to some plan is unrealistic. Data may be either monitored continously or analysis may be carried at moments in time that are not prespecified. To name a few examples, this includes hypothesis testing for online experimentation [Urban et al., 2021, Lindon and Malek, 2022], weather forecasting [Henzi and Ziegel, 2021], financial forecasting [Wang et al., 2022], and, as mentioned earlier, meta-analysis of medical data [Ter Schure and Grünwald, 2019, 2022]. For instance, in weather prediction, multiple forecasts can be compared to the ground-truth every day. Since anytime-valid procedures can be continuously monitored in time, valuable time-domain information can be observed and acted upon in real time without invalidating the type-I error guarantees of the tests. For instance, using anytime-valid methods one may observe that some weather forecasts are better than others during certain moments of the year [Choe and Ramdas, 2022].

It is with these applications in mind that the results of this dissertation are to be understood. Sequential analysis also appeared in a concrete context, that of the United States during World War II, were the efforts of the statistical community were directed towards problems related to the war. The strengths and limitations of sequential methods can only be understood in relation to those applications. In the next section we will introduce the central test in sequential analysis, the Sequential Probability Ratio Test, in historical context. We will then introduce its anytime-valid counterpart with the objective of placing this dissertation in context.

1.2. Sequential analysis

The fact that sequential-analyitic methods require fixed sampling rules is not a deficiency, it is part of their design. This design and subsequent success is, in turn, related to their origin and first applications. Sequential analysis has its roots in the Statistical Research Group (SRG), a unit of statisticians formed in 1942, during World War II, at Columbia University in New York with the aim of working on military problems [Wallis, 1980]. One of their most important contributions is the design and analysis of the Sequential Probability Ratio test (SPRT). Without compromising on the probability of error, if a sequential experiment that follows the SPRT's sampling plan is repeated multiple times, it will require fewer observations on average than the best possible fixed-sample test[1] [Wald, 1947]. This is crucial when, as is frequently the case in war applications, tests are destructive or the cost of testing is higher than the cost of production [Tukey, 1947]. The main figure in the development of sequential analysis is Abraham Wald, a Jewish mathematician born in today Romania in 1902 who had moved to the United States before the start of the war [Wolfowitz, 1952]. One of the most prominent uses of the SPRT is in the problem of lot inspection, where an operator must accept or reject each lot of industrally manufactured parts depending on whether or not a fraction of the lot is defective. To that end, workers are instructed to follow a prespecified sequential sampling plan, and decide on the defectiveness of each lot depending on their observations. In his summary after the war, in 1946, Warren

[1] Milton Friedman, also one of the members of the SRG recalls in an interview: "[...] we stated the problem in such a way that statisticians found it difficult to accept. We said, 'we know how to construct a test that's more powerful than the uniformly most powerful test.' They said, 'That's mathematically impossible, you can't do that, we've proved that this is the most powerful test.' And so statisticians wouldn't have anything to do with it. Then, we talked to Abraham Wald, and he initially had the same reaction. But then he went home and a day later he called and said, 'you are right and I know how to do it and I know what the answer is.'" [Taylor, 2001, p.114]

Weaver, founder of the Statistical Research Group, wrote about the importance of this application.

> In March 1945, the Quartermaster General wrote to the War Department liaison officer for NDRC [National Defense Research Committee] a letter containing the following statement: "[...] With thousands of contractors producing approximately billions of dollars worth of equipment each year, even a 1% reduction in defective merchandise would result in a great saving to the Government. Based on our experience with sequential sampling in the past year, it is the considered opinion of this office that savings of this magnitude can be made through wide dissemination of sequential sampling procedures." [...] The Quartermaster Corps imported in October 1945 that at least 5,000 separate installations of sequential sampling plans have been made and that in the few months prior to the end of the war new installations were being made at the rate of 500 per month. [Office of Scientific Research and Development, 1946]

Given their success, the developments in sequential analysis were unclassified after the war, and their adoption became widespread in statistical quality control. The sequential methods inspired by these initial developments [see Siegmund, 1985, Tartakovsky et al., 2014] are used routinely, for instance, in monitoring clinical trials [Proschan et al., 2006]. A comprehensive recount of the ensuing developments and challenges was written by Lai [2001].

A simplified version of the lot-inspection problem is equivalent to the coin-tossing example from before; we now describe the SPRT for it. Symbolically, consider the problem

$$\mathcal{H}_0 : X_i \text{ is heads with prob. } p \quad \text{vs.} \quad \mathcal{H}_1 : X_i \text{ is heads with prob. } q,$$

for a fixed value of $q \neq p$. Here, the problem is the same as testing a very large lot: upon testing an item of a lot with a fraction p of defective items, each observation will be defective—analogously, each coin toss will land heads—with probability approximately p. The coin-tossing experiment is a special case with $p = 1/2$. The SPRT is based on sequentially monitoring the likelihood ratio $L_k(q)$, that is, the ratio of the probabilities of having observed the outcomes under $\mathcal{H}_1$ and $\mathcal{H}_0$. If H_k is the number of heads in $X_1, \ldots, X_k$ and $T_k = k - H_k$ is the number of tails in the same data, the likelihood ratio L_k is

$$L_k(q) = \frac{q^{H_k}(1-q)^{T_k}}{p^{H_k}(1-p)^{T_k}}.$$

If data are more likely under the alternative model $\mathcal{H}_1$, L_k will take large values; lower values in the oposite case, when data is more likely under $\mathcal{H}_0$. In the lot-inspection example, there are two types of errors that can be made: either acceptable lots are rejected (a type-I error) or faulty lots are accepted (a type-II error). The first kind of error corresponds to a false rejection of $\mathcal{H}_0$; the second, to a false acceptance of $\mathcal{H}_0$. Despite the confusing terminology, given two tolerable values α and γ for the type-I and type-II errors—with $\alpha \leq \gamma$—the SPRT yields the shortest sampling plan on average

[Wald and Wolfowitz, 1948]. For a fixed q—we will treat the case that multiple values of q are of interest—, the SPRT's sampling plan monitors whether $L_k(q)$ lies between or outside two limits: an upper limit $B = \gamma/\alpha$ and a lower limit $A = (1-\gamma)/(1-\alpha)$. These limits are chosen so that, if the procedure carried to completion—until a decision is reached—, the target error probabilities α and γ are met. The SPRT sampling plan is as follows: at $k = 1$, a first observation is made, $L_k(q)$ is computed, and one of the following three decisions is made

$$\begin{aligned}
&\text{if } L_k(q) \geq B, && \text{choose } \mathcal{H}_1;\\
&\text{if } L_k(q) \leq A, && \text{choose } \mathcal{H}_0;\\
&\text{if } L_k(q) \in (A, B), && \text{make one more observation and repeat.}
\end{aligned}$$

This sequential sampling procedure can be easily implemented in a factory as the decision boundaries can be written in a table that only depends on the number of defective items observed upon inspection. A composite alternative hypothesis $\mathcal{H}_1 : q \neq p$ can be handled by using a mixture of likelihood ratios with a "prior" probability distribution π on q, that is, by using a statistic $L_i(\pi) = \int L_i(q) \mathrm{d}\pi(q)$. In that case, an integrated version of the type-II error is controlled instead [see Wald, 1945, Section 6]. We remark that the guarantees on the average errors—over repetitions of the experiment—depend crucially on using the specific sampling plan from previous display. As we noted earlier, such sampling plans cannot always be implemented: similarly as in "sampling to a foregone conclusion", if data that is not gathered with the SPRT's sampling plan is analized as if had been, the error probabilities may be very different from what was initially intended. To prevent that, anytime-valid methods offer type-I error control irrespective of the sampling plan that is used.

1.3. Anytime-valid testing

There exists a canonical construction of anytime-valid tests [Ramdas et al., 2020], which uses ideas already present in tests of power one [Darling and Robbins, 1968b]. This construction is extendable to large nonparametric tests, where no likelihood ratios may be available. In the coin-tossing example, where the null hypothesis is a single distribution, this canonical construction corresponds to rejecting $\mathcal{H}_0$ when a mixture of likelihood ratios is large, similarly to the SPRT. The main difference with the SPRT is that only the sequential properties of the likelihood ratio are used—not the fact that it is a likelihood ratio. More precisely, the key fact is that, under the null hypothesis, the mixture $L_i(\pi) = \int L_i(q) \mathrm{d}\pi(q)$ with respect to the "prior" π is a nonnegative martingale starting at one, an object called test martingale [Shafer et al., 2011]. An anytime-valid test for $\mathcal{H}_0$ would, similarly to the SPRT, reject the null hypothesis when the likelihood ratio takes large values. The difference is that uniform statistical type-I error control over all sampling plans—anytime validity—is now obtained with standard maximal inequalities for martingales. Indeed, an equality of Ville [1939]—also known as Doob's maximal inequality—implies that, for the threshold B',

$$\mathbf{P}\{\sup_i L_i(\pi) \geq B'\} \leq \frac{\mathbf{E}_{\mathbf{P}}[L_1(\pi)]}{B'} = \frac{1}{B'}.$$

This implies with the particular choice $B' = 1/\alpha$, that, for any random time ν,

$$\mathbf{P}\{L_\nu(\pi) \geq 1/\alpha\} \leq \alpha.$$

If we interpret ν to be the final sample size of an unknown sampling plan, the last display implies that the test that rejects $\mathcal{H}_0$ if $L_\nu(\pi) \geq 1/\alpha$ has type-I error bounded by α. This error guarantee holds irrespective of the sampling plan used—the test is anytime valid. This even covers the most aggressive sampling plan possible, the one that stops as soon as the statistic $L_k(\pi)$ crosses the threshold $B' = 1/\alpha$. Notice the differences with the SPRT: by using a more conservative threshold for rejecting $\mathcal{H}_0$ ($1/\alpha$ instead of γ/α), uniform type-I error control over all sampling plans is achieved. This construction already expands significantly the scope of application of anytime-valid methods: test martingales may be constructed for problems in which a likelihood ratio is not available [Shafer and Vovk, 2001, Howard et al., 2018b].

The main advances in this line of research, in which this thesis is inscribed, pertain the extension of this framework to more general composite statistical hypotheses problems. These problems are written symbolically as

$$\mathcal{H}_0 : X_1, \ldots, X_\nu \sim \mathbf{P} \in \mathcal{P} \quad \text{vs.} \quad \mathcal{H}_1 : X_1, \ldots, X_\nu \sim \mathbf{Q} \in \mathcal{Q}, \tag{1.1}$$

where $\mathcal{Q}$ and $\mathcal{P}$ are two families of distributions, and ν is an arbitrary random sample size. Two chapters of this dissertation are dedicated to such problems: Chapter 2, to the case in which both families of distributions are symmetric under a group of transformations; Chapter 3, to the nonparametric problem of comparing the survival-time distributions of two groups of subjects. An additional challenge posed by this family of problems is that, if type-I error control is sought over all possible sampling plans, power maximization is no longer a meaningful criterion; an alternative theory of optimality has to be established [Grünwald et al., 2020]. Although it is often the case that anytime-valid tests can be built with standard techniques, their optimality is harder to assess. This assessment is one of the main themes of the results about anytime-valid tests contained in this dissertation.

The testing problems that are treated in this dissertation can be solved using test martingales. As a side note—we do not treat this question in this dissertation—a natural question that arises is whether all anytime-valid tests use necessarily the test-martingale construction, that is, whether there is a test martingale behind every anytime-valid test. Interestingly, this turns out not to be the case. For instance, when the null hypothesis is "too big", the only test martingale available is the constant one, but anytime-valid tests still be constructed. In general, all anytime-valid tests are constructed using a more general type of statistics called E-processes—a composite-null generalization of test martingales—[Ramdas et al., 2022b].

Thus far, we have described the problem of statistical hypothesis testing as a decision problem. Yet, the statistics underlying anytime-valid tests—and hence the results of this dissertation—can also be interpreted outside the framework of acceptance-rejection procedures; in particular, as prediction strategies in certain games of chance. This provides a link to the study of prediction problems in this dissertation (Chapter 4).

1.4. Decision, evidence, and prediction

In this section we establish a bridge between the three concepts in the title of the section and serves both to interpret our results in anytime-valid testing and as a motivation to Chapter 4. We now outline our argument. We will see that multiple criticisms have been made to using pure decision procedures that divide results into “significant” and “nonsignificant” in scientific practice. Thus, in some applications the role of statisticians should not be limited to decision-making. As a response to this criticism, an alternative role of statistics can be that of quantifying evidence—for one model against another. This section will be centered in accepting the proposition that prediction quality can be a surrogate to evidence quantification. In the coin-tossing example, one could regard both alternatives $\mathcal{H}_0$ and $\mathcal{H}_1$ as models that make predictions about the data. Under that lens, there is evidence for the alternative $\mathcal{H}_1$ against the null $\mathcal{H}_0$ if the model signified by $\mathcal{H}_1$ makes better predictions than that of $\mathcal{H}_0$. If prediction is a worthwhile enterprise, one may take the view that absolutely no probabilistic assumptions should be made about the origin of the data. Prediction in this lawless setting is the subject of Chapter 4. We now develop these ideas.

The use of acceptance-rejection procedures in scientific practice in general (not only for sequential tests), has been criticized since their very inception. Indeed, even Ronald A. Fisher, an English biologist and statistician who is credited with laying the foundations of modern statistics, was critical of the decision-theoretic approach to hypothesis testing.

> [...] Acceptance procedures are of great importance in the modern world. [...] but the logical differences between such an operation and the work of scientific discovery by physical or biological experimentation seems to me so wide that the analogy between them is not helpful and the identification of the two sorts of operation is decidedly misleading. [Fisher, 1955]

Fisher points at the fact that acceptance procedures are well suited for quality control when one tests repeatedly batches of goods—just as in the lot-inspection problem—, but that they were not appropriate for testing scientific hypotheses. This follows from the fact that decision problems are crafted with very specific objectives in mind. For instance, in the lot-inspection problem, decisions must be made while minimizing the average length of the sampling plan—subject to error-control constraints—for economic reasons. In contrast, scientific studies sometimes have more diffuse objectives; a coarse classification of results into “significant” and “nonsignificant” may obscure the interpretation of the data, the estimation of the effects, and the uncertainty of the estimates [Greenland et al., 2016]. For example, in clinical trials, different groups may reach diverging decisions because of factors additional to whether the new treatment is significantly better than the previous one. Considerations such as costs, side effects, ease of administration and the assessment of other studies are paramount [Armitage, 1985]. Some even call for abandoning the decisional aspects of statistical testing from scientific practice altogether [Amrhein et al., 2019]. Rather than making decisions, we could regard the task of statisticians as that of quantifying evidence against a specific statistical hypothesis relative to an alternative [Berger and Wolpert, 1984, Royall, 1997]. To accommodate the view of statistics as evidence quantification, it can be useful to consider narratives about statistical procedures that depart from pure acceptance-rejection and its parallel with lot-inspection.

Several narratives about testing have emerged over the last decades in order to accommodate the evidentiary aspects of sequential testing. Some of these narratives place prediction instead of decision at the center of the stage. These narratives include prequential statistics [Dawid, 1984], where serial testing and probabilistic forecasting are identified; minimum description length [Grünwald, 2007, Grünwald and Roos, 2019], where the identification is between prediction and data compression; and gambling [Shafer, 2019], where the parallel is between test statistics and betting strategies. Loosely speaking, a statistical model is better than another if it predicts better, compresses data more, or makes us richer. All of these can be used as evidence that one model describes the data better. Among these, it is the gambling narrative that has become dominant in the anytime-valid testing community.

Many of the intuitions behind anytime-valid testing can be explained using a parallel between the monetary gains made in a casino and evidence—there exist deep historical connections between gambling and probability theory [Shafer, 2021]. Fundamental results in standard courses on stochastic processes are usually interpreted through their implications for gambling. For the purpose of anytime-valid testing, the main example is the optional stopping theorem, a central theorem in martingale theory. It expresses the fact that in a fair betting game there is no strategy that will consistently make a player rich. Thus, if monetary gains—which quantify our ability to make predictions—are viewed as a surrogate for evidence, we can picture a game whose payoffs are determined by the observed data. If the game is fair under the null hypothesis, just as there is no strategy that will consistently make a player rich, there is no sampling plan that will accumulate evidence on average against the null hypothesis. Hence, accumulating a large amount of money—evidence—is a sign that the game is not fair or, in other words, that the null hypothesis must be disqualified.

If one accepts the idea that there is a connection—by analogy or otherwise— between hypothesis testing and prediction, one may take the extreme position that no assumptions whatsoever can be made about the data sequence that is observed [Cesa-Bianchi and Lugosi, 2006]. In the coin-tossing example, when comparing whether the coins land heads or tails with probability p or q, each of these probabilities may be viewed as the probabilistic forecast of two "experts" about the next outcome of the coin. Accepting the absence of assumptions on the data sequence entails that the analysis of prediction algorithms must be carried in the worst case over all possible data sequences [Vovk, 1990, Freund and Schapire, 1997]. The goal becomes to try and predict as well as the best expert, no matter what data sequence is observed during experimentation. One of the contributions of this dissertation, contained in Chapter 4, is to this game of prediction with expert advice.

This dissertation dwells in the intersection where techniques for decision, testing and prediction meet. In the next section we outline its contents.

1.5. Outline

The rest of this work consists of four chapters. The first two chapters are contributions to the theory of anytime-valid testing; the third, to the theory of individual sequence prediction; the fourth, to concentration inequalities. We now sketch each of them briefly.

1.5.1. Group-invariant tests

When formulating models for physical observations, it is usually desirable to choose them to be invariant under certain arbitrary choices. For instance, when measuring a physical quantity, the choice of measurement units, frame of reference, and starting time are all arbitrary. This desire translates into invariances of the probabilistic models under consideration. For instance, if the units of the observations and of the parameters of a model are changed simultaneously, probability assignments should remain unchanged. In mathematical terms, this is described as an invariance under the action of a group of transformations. For instance, in the case of "invariance under change of measurement units" this corresponds to an invariance under an action of $(\mathbb{R}^+, \cdot\,)$, the group of positive real numbers with multiplication.

Chapter 2 studies optimum—in a sense to be defined—anytime-valid tests under general groups of transformations. We show, for testing problems, that if the families of probability distributions are invariant under a common group of transformations, the best overall anytime-valid test is guaranteed to reside within the family of invariant procedures. In many interesting cases, optimal testing boils down to monitoring the likelihood ratio for certain invariant functions of the data. Loosely speaking, these invariant functions must loose no information about the invariant component of the data. Our result is an anytime-valid counterpart of the celebrated theorem of Hunt and Stein [Lehmann and Romano, 2005], which states that, when the sample size is fixed, an overall most powerful test exists within the smaller family of invariant tests. Applications of these results include the t-test and linear regression, among many others.

1.5.2. Tests for survival data

A prominent topic in medical statistics is the analysis of time-to-event data. For instance, when studying the effect of a new treatment—a vaccine, for example—, patients may be given either a placebo or the treatment. Researchers record the time that it takes each patient to become ill from the target disease and, if vaccination is effective, it is expected that the treatment group takes longer on average to become ill. Survival-time models aim to mathematize these situations. Using these models, the most prominent test for comparing two groups of subjects is the logrank test [Slud, 1984], a simple but fundamental application of the proportional hazards model of Cox [1972], a cornerstone of statistical science. In Chapter 3, we develop an anytime-valid counterpart of the logrank test. This test allows for the continuous monitoring of medical data and the resulting statistics can be easily combined to perform meta-analysis. The combined statistics can be themselves monitored in real time, going beyond the realm of conventional meta-analysis. The author made a contribution to the `safestats` R package [Turner et al., 2022], which is available on the Comprehensive R Archive Network (CRAN).

1.5.3. Multiscale worst-case prediction

If we regard hypothesis testing as a problem of prediction, there is no reason to stop at the assumption that data are generated according to a probabilistic distribution; data might be generated adversarially. This setting is known as worst-case or individual sequence prediction [Cesa-Bianchi and Lugosi, 2006]. A simple but fundamental problem in this line of work is that of prediction with expert advice [Vovk, 1998]. This is a game played in rounds where we are tasked with aggregating the advice of K experts. At each round, the quality of each expert's advice is judged with a numerical loss. The remarkable feature of this problem is that, aside from a range restriction on their possible values, absolutely no assumptions are made about how these losses are assigned. The objective is to perform, after a number of rounds, as well as the best expert in hindsight—the one whose advice has the lowest cumulative losses. Despite the apparent simplicity of this problem, efficient solutions to this problem have profound consequences to other areas in computational learning theory, which include convex optimization [Hazan, 2021], statistical learning theory [Freund and Schapire, 1997], and probabilistic maximal inequalities [Foster et al., 2017], to name a few. In Chapter 4, we study this problem under multiscale range restrictions on the losses of the experts. We formulate Muscada, a multiscale and computationally efficient algorithm that is safe in the worst case, and performs much better when data is not completely adversarial.

1.5.4. Concentration Inequalities

In Chapter 5, we introduce a notational device, the exponential stochastic inequality (ESI), that provides an ordering of random variables. It captures the situation when two random variables are ordered both in expectation and with high probability—it is possible to construct random variables that are ordered in one sense but not in the other. This notation was originally introduced by Koolen et al. [2016] and Grünwald and Mehta [2020]. Our interest in such statements come from arguments used to derive excess-risk bounds for machine learning algorithms. The ESI is particularly well suited to deriving PAC-Bayesian bounds. We show how the ESI is useful when deriving bounds for sums and averages of random variables, and how its use can yield improvements over conventional union bounds. We characterize the random variables that satisfy an ESI under weak moment conditions in terms of existing tail conditions.

2. E-statistics, Group Invariance, and Anytime-Valid Testing[1]

We study worst-case-growth-rate-optimal (GROW) E-statistics for hypothesis testing between two dominated group models. If the underlying group G acts freely on the observation space, there exists a maximally invariant statistic of the data. We show that among all E-statistics, invariant or not, the likelihood ratio of the maximally invariant statistic is GROW and that an anytime-valid test can be based on it. By virtue of a representation theorem of Wijsman, the GROW E-statistic is equivalent to a Bayes factor with a right Haar prior on G. Such Bayes factors are known to have good frequentist and Bayesian properties. We show that reductions through sufficiency and invariance can be made in tandem without affecting optimality. A crucial assumption on the group G is its amenability, a well-known group-theoretical condition, which holds, for instance, in general scale-location families. Our results also apply to finite-dimensional linear regression.

2.1. Introduction

Classically, hypothesis tests for group-invariant problems have been studied in great detail both for fixed-sample-size and sequential experiments [Cox, 1952, Hall et al., 1965, Eaton, 1989, Lehmann and Romano, 2005]. Nevertheless, due to methodological concerns about combining evidence from multiple experiments using classical methods [Royall, 1997, Wagenmakers, 2007, Benjamin et al., 2018, Grünwald, 2023], a theory of testing based on E-statistics[2] has been developed [Vovk and Wang, 2021, Shafer, 2019, Grünwald et al., 2020, Ramdas et al., 2020]. The main concern that is successfully addressed by testing with E-statistics is that of error control in two common situations: when experiments are optionally stopped, and when aggregating the evidence of interdependent experiments that may themselves have been optionally stopped [Wang and Ramdas, 2022, Vovk and Wang, 2021]. The latter of these situations is sometimes referred to as optional continuation [Grünwald et al., 2020], and tests that remain valid under optional stopping are called anytime valid. As a consequence of the ability

[1] This chapter is based on M. F. Pérez-Ortiz, T. Lardy, R. de Heide, and P. Grünwald. E-Statistics, Group Invariance and Anytime Valid Testing, Aug. 2022. URL `http://arxiv.org/abs/2208.07610`. arXiv:2208.07610 [math, stat], under submission

[2] E-statistics are mostly known as E-variables or, in analogy to p-values, E-values; we call them E-statistics here to emphasize that they are, in fact, statistics of the data.

to control type-I errors under both optional stopping and continuation, a wide interest in E-statistics has kindled in recent years [as a small sample, we mention Shafer, 2019, Henzi and Ziegel, 2021, Ramdas et al., 2022b, Wang and Ramdas, 2022, Ren and Barber, 2023]. As a contribution to this line of work, we characterize optimal E-statistics in group-invariant testing problems. As we will see, such problems include testing under linear-model and Gaussian assumptions.

We concern ourselves with testing dominated composite hypotheses where both null and alternative models remain unchanged under a group of transformations. In particular, we study the case where the parameter of interest is a function $\delta = \delta(\theta)$ of the model parameter θ that is invariant under such transformations. For example, in the Gaussian case, the coefficient of variation is invariant under scale changes; the correlation coefficient, under affine transformations; and the variance of the principal components, under rotations around the origin. Roughly speaking, by replacing the data $X^n = (X_1, \dots, X_n)$ by an invariant function $M_n = m_n(X^n)$, one discards all information that is not relevant to the parameter δ. Through the lens of the invariance-reduced data M_n, the hypotheses about the parameter of interest δ may simplify. In particular, in the simplest but central special case, the null then expresses that the parameter of interest is equal to some fixed δ_0, and some other fixed δ_1 under the alternative. Then the test applied to the reduced data M_n becomes a simple-vs.-simple test. One may now base a sequential test on the likelihood ratio statistics of the M_n, an idea going back to Rushton [1952], Cox [1952] and developed by Hall et al. [1965]. By a representation theorem due to Wijsman [1967]—we use the version of Andersson [1982]—, under mild regularity conditions on the group, this likelihood ratio is equivalent to the (formal) Bayes factor obtained by equipping both the null and the alternative with a (usually improper) right Haar prior; such Bayes factors have been studied in detail within the Bayesian community [Dawid et al., 1973, Berger et al., 1998].

In this chapter we characterize E-statistics that are growth rate optimal in the worst case (GROW), as defined by Grünwald et al. [2020] (GHK from now on); see Section 2.1.2 for the definition of GROW), and we show how to use these for anytime-valid testing. Informally, among all tests based on E-statistics, those satisfying GROW have the fastest (in terms of sample size) expected logarithmic growth rate under the alternative, thereby accumulating evidence against the null as fast as possible in expectation. GROW approaches are a worst-case version of the Kelly (1956) betting criterion, which has been advocated within information theory and economics since the 1950s. These approaches have become central in the nascent field of e-processes, anytime valid testing and confidence sequences.

We need to distinguish between the main *statistical result* (statement with direct repercussions for statistical practice) and the main underlying *technical result* of this chapter. Our main *statistical* result is the following: under regularity conditions, when the test about the invariance-reduced data M_n becomes a simple-vs.-simple test, among all E-statistics, whether a function of M_n or not, the GROW E-statistic is simply the aforementioned likelihood ratio statistic for M_n. What is remarkable here is the 'whether a function of M_n or not' part: that the GROW E-statistic is a function of M_n (and hence itself 'invariant') is a consequence, rather than than input to, the analysis. The main consequence is that the aforementioned existing classical sequential tests based on M_n and Bayesian tests based on Bayes factors with a right Haar prior can be trivially modified to become ‘anytime valid’, and that they are then optimal for the testing problem at hand.

This main statistical result arises as a corollary (Corollary 2.4.3) obtained from combining Theorem 1 of GHK and our main *technical* result. Theorem 1 of GHK shows that finding a GROW E-statistic is equivalent to performing joint minimization of the Kullback-Leibler (KL) divergence between the convex hulls of the alternative and null sets of distributions. The main *technical* contribution of this chapter is computing the value of the joint KL divergence minimization problem: Theorem 2.4.2, shows that, under regularity assumptions, this value coincides with the KL divergence between the distributions, under each hypothesis, of a maximally invariant function M_n of the data. A maximally invariant function, informally, looses as little information as possible about the invariant component of the data. The central assumption in our results is the amenability of the group G, a well known group-theoretical condition [Bondar and Milnes, 1981]. This condition also plays a key role in the celebrated theorem of Hunt and Stein [Lehmann and Romano, 2005, Section 8.5], which relates tests that are max-min optimal for statistical power to group-invariant tests. We show that, just as in the result of Hunt and Stein, the amenability of G is a sufficient condition, but not a necessary one (see Section 4.7). We remark that the concepts of power and GROW are, to some mild extent, related: one may view GROW as the analogue of power in an optional stopping and continuation setting (Section 3.3.1). Despite the ensuing analogy to Hunt-Stein, the proof techniques that we develop to prove our results are significantly different (see Section 2.1.4).

Besides these main statistical and technical contributions we provide two additional novel results: Proposition 2.4.4 and Proposition 2.8.1. Proposition 2.4.4 investigates E-statistics that are relatively GROW (abbreviated to REGROW by GHK), an optimality criterion closely related to GROW (see Section 3.3.1). We show that, as opposed to the general case (where GROW E-statistics can be very different from relatively GROW ones), in our group invariant setting, any GROW E-statistic is also relatively GROW. In Proposition 2.8.1 we extend the main technical result Theorem 2.4.2 to settings where the parameters δ_0 and δ_1 may take values in sets Δ_0 and Δ_1, respectively, including the case when prior distributions on Δ_0 and Δ_1 are available, relating our work to testing with Bayes factors as in [Jeffreys and Jeffreys, 1998, Berger et al., 1998].

Finally, we provide some results for which we do not claim novelty—they are rather a rephrasing, within our group invariant context, of existing results. These include Proposition 2.1.2, which shows that if data are gathered sequentially, then the sequence of GROW E-statistics can be used for anytime-valid sequential testing, the reason being that it becomes a test martingale, the mathematical object that forms the basis for anytime-valid testing [Shafer, 2019, Grünwald et al., 2020]. We also describe when the optionally stopped optimal E-statistic remains an E-statistic, which is important in an optional continuation context. Finally, in Proposition 2.5.2 we show how data can be further reduced if a sufficient statistic for the invariant parameter is available. For the latter purpose, a result of C. Stein, reported by Hall et al. [1965], is instrumental.

We illustrate all our results with several examples.

The rest of this introduction gives an overview of the whole chapter. It is organized in the following manner. In Section 2.1.1, we introduce formally our setup for hypothesis testing under group invariance and we introduce our running example, the t-test. In Section 2.1.2, we define E-statistics, our main objects of study, and in Section 3.3.1 we define our optimality criteria. In Section 2.1.4, we give an informal exposition of our main statistical result, Corollary 2.4.3, and our main technical result, Theorem 2.4.2. In Section 2.1.5 we highlight previous work made in group-invariant testing and in Section 2.1.6 we introduce notation. Finally, in Section 2.1.7 we outline the rest of the chapter.

2.1.1. Group invariance

In this section we describe the group-invariant hypotheses that are of our current interest. Assume that a group G acts freely on both the observation space $\mathcal{X}$ and the parameter space Θ. Denote the action of G on $\mathcal{X}$ by $(g, X) \mapsto gX$ for $g \in G$ and $X \in \mathcal{X}$. For samples of size n, we extend the action of G on $\mathcal{X}$ to $\mathcal{X}^n$ componentwise, that is, by $(g, X^n) \mapsto gX^n := (gX_1, \dots, gX_n)$ for $g \in G$ and $X^n \in \mathcal{X}^n$. By invariance of a probabilistic model $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ on $\mathcal{X}^n$ we understand that, for any $g \in G$ and measurable $B \subseteq \mathcal{X}^n$ and parameter $\theta \in \Theta$, the distribution $\mathbf{P}_\theta$ satisfies

$$\mathbf{P}_\theta\{X^n \in B\} = \mathbf{P}_{g\theta}\{X^n \in gB\}, \tag{2.1}$$

where $gB = \{gb : b \in B\}$ is the left translate of the set B by g. In particular, we study situations where the parameter of interest $\delta = \delta(\theta)$ indexes the orbits in the parameter space Θ under the action of G. More formally, we assume that δ is a maximally invariant function of the parameter θ, meaning that, for any pair $\theta, \theta' \in \Theta$, there exists g such that $g\theta = \theta'$ any time that $\delta(\theta) = \delta(\theta')$. In that case, we say that δ is a maximally invariant parameter. We are prepared to state the main statistical hypothesis testing problem in this work. For two possible values δ_1, δ_0 of δ, we consider the composite vs. composite testing problem

$$\mathcal{H}_0 : \delta(\theta) = \delta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \delta(\theta) = \delta_1. \tag{2.2}$$

As is known, many classical parametric problems can be cast in this shape. Let us call maximally invariant any G-invariant function $M_n = m_n(X^n)$ that indexes the orbits of the action of G on $\mathcal{X}^n$. The distribution of M_n depends on θ only through the maximally invariant parameter δ, and, under this reduction, the problem (2.2) becomes simple. It is with the optimality of this reduction that we are concerned. In Section 2.8, we study cases in which, even after the invariance reduction, the problem under study remains composite.

Example 2.1.1 (t-test under Gaussian assumptions). Consider an i.i.d. sample $X^n = (X_1, \dots, X_n)$ of size n from an unknown Gaussian distribution $N(\mu, \sigma)$, and testing whether $\mu/\sigma = \delta_0$ or $\mu/\sigma = \delta_1$. The parameter space Θ consists of all pairs $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ and the Gaussian model is invariant under scale transformations. The group $G = (\mathbb{R}^+, \cdot\ )$ of positive real numbers with multiplication acts on Θ by $(c, (\mu, \sigma)) \mapsto$

$(c\mu, c\sigma)$ for each $c \in \mathbb{R}^+$ and $(\mu, \sigma) \in \Theta$. The parameter of interest is the ratio $\delta = \mu/\sigma$ between the mean μ and the standard deviation σ. The parameter δ is scale-invariant and indexes the orbits of the action of G on Θ. The group G acts on the observation space $\mathcal{X} = \mathbb{R}^n$ by coordinatewise multiplication. A maximally invariant statistic is $M_n = m_n(X^n) = (X_1/|X_1|, \ldots, X_n/|X_1|)$, and its distribution only depends on the maximally invariant parameter $\delta = \mu/\sigma$.

2.1.2. The family of E-statistics and their use in optional continuation and stopping

We now define E-statistics, our measure of evidence for the alternative over the null hypothesis. Given two subsets Θ_0, Θ_1 of the parameter space Θ, interpreted as the null and an alternative hypothesis, the family of E-statistics comprises all nonnegative functions of the data $X^n \in \mathcal{X}^n$ whose expected value is bounded by one under all elements of the null, that is, all statistics $T_n(X^n) \geq 0$ such that

$$\sup_{\theta_0 \in \Theta_0} \mathbf{E}^{\mathbf{P}}_{\theta_0}[T_n(X^n)] \leq 1. \tag{2.3}$$

To make this concrete, we first remark that, as is immediately seen by evaluating the expectation, the likelihood ratio in any simple-vs-simple testing problem is an E-statistic (see e.g. Section 1 of GHK). In particular, in the setting above,

$$T_n^* = \frac{p^{M_n}_{\delta_1}(m_n(X^n))}{p^{M_n}_{\delta_0}(m_n(X^n))}, \tag{2.4}$$

where, for $j = 0, 1$, $p^{M_n}_{\delta_j}$ are densities of any given maximally invariant function $M_n = m_n(X^n)$ under $\mathcal{H}_j$ relative to a common underlying measure, is an E-statistic. The suitability of E-statistics in optional continuation contexts is due to the following two properties, which readily follow from (2.3):

1. The type-I error of the test that rejects the null hypothesis anytime that $T_n(X^n) \geq 1/\alpha$ is smaller than α, a direct consequence of Markov's inequality and the definition of E-statistic.

2. Suppose that X^n and X^m are independent, representing the outcomes of two subsequent experiments, and let $T_n(X^n)$ be an E-statistic for X^n and for all ϕ in some set Φ, let $T_m(X^m; \phi)$ be an E-statistic for X^m. Suppose that, after observing the first sample X^n, the $T_m(X^m; \phi)$ to be used to measure evidence for the second sample may be chosen as a function of X^n. That is, we use $T_m(X^m; \hat{\phi}(X^n))$ where $\hat{\phi}(X^n)$ is some function of X^n. Then $T_{n+m}(X^n, X^m) := T_n(X^n)T_m(X^m; \hat{\phi}(X^m))$ is also an E-statistic, irrespective of the definition of $\hat{\phi}$.

Together, these two properties imply that the test based on T_{n+m} that rejects the null if $T_{n+m}(X^{n+m}) \geq 1/\alpha$, has Type-I error bounded by α, no matter the definition of $\hat{\phi}$—the details of which definition may be unknown or even unknowable to the statistician. For example, it may be that upon observing some X^n, it is decided not to consider a second sample at all (this amounts to using a ϕ° such that $T_m(X^m, \phi^\circ) \equiv 1$, independently of data X^m). It may also be that one decides to consider a second sample after seeing X^n but one does not really know if one would have continued as well had X^n been different—still, one obtains a valid Type-I error guarantee. This property can be extended recursively to more than two samples and sample sizes depending on the past: we get the general result that, in a sequence of studies of (say) a new medication, we can combine the results of each study as measured by an E-statistic by multiplication; it may be decided by the statistician (or by external factors such as availability of funding), after each study, and depending on previous study outcomes, whether or not to consider an additional study and if so, what sample size, what study-protocol and what E-statistics to use for the next study—the total number of studies is unlimited in advance, and still we have Type-I error control if we multiply the individual E-statistics: we can safely engage in optional continuation, something which is not easily done with p-values (Section 1.3 of GHK).

The use of E-statistics in itself is sufficient to allow for optional continuation with fixed sample size studies. Some—not all—types of E-statistics can also be used in two additional related settings: *optional stopping*, when there is a single sequence of data $X_1, X_2, \dots$ and we want to do a sequential test with Type-I error guarantees irrespective of when we stop; and optional continuation as above, but with individual E-statistics whose sample size is itself not fixed but determined by some stopping rule. Proposition 2.1.2 below shows that the E-statistics (2.4) are of this kind. Suppose then that data $X_1, X_2, \dots$ are gathered one by one. Here, a sequential test is a sequence of zero-one-valued statistics $\xi = (\xi_n)_{n\in\mathbb{N}}$ adapted to the natural filtration generated by $X_1, X_2, \dots$. We consider the test defined by $\xi_n = \mathbf{1}\{T_n^* \geq 1/\alpha\}$ for some value α, whose anytime validity we prove. Additionally, we show that, for certain stopping times $\tau \leq \infty$, the optionally stopped E-statistic T_τ^* remains an E-statistic, which validates the use of the stopped T_τ^* for optional continuation: we can multiply such that T_τ^* across studies as explained above while retaining Type-I error control.

Proposition 2.1.2. *Let $T^* = (T_n^*)_{n\in\mathbb{N}}$, where, for each n, T_n^* is the likelihood ratio for the maximally invariant function $M_n = m_n(X^n)$ for the action of G on $\mathcal{X}^n$. Let $\xi = (\xi_n)_{n\in\mathbb{N}}$ be the sequential test given by $\xi_n = \mathbf{1}\{T_n^* \geq 1/\alpha\}$. Then, the following two properties hold:*

1. *The sequential test ξ is anytime valid at level α, that is,*

$$\text{for any random time } N, \ \sup_{\theta_0\in\Theta_0} \mathbf{P}_{\theta_0}\{\xi_N = 1\} \leq \alpha.$$

2. *Suppose that $\tau \leq \infty$ is a stopping time with respect to $M = (M_n)_{n\in\mathbb{N}}$. Then the optionally stopped* E*-statistic T_τ^* is also an* E*-statistic, that is,*

$$\sup_{\theta_0\in\Theta_0} \mathbf{E}^{\mathbf{P}}_{\theta_0}[T_\tau^*(X^\tau)] \leq 1. \tag{2.5}$$

The mechanism of the proof of this proposition—showing that $T^* = (T^*_n)_{n\in\mathbb{N}}$ is a nonnegative martingale with expected value one—is, by now, standard; we perform it in Section 2.6. The main ingredient, where invariance plays a role, pertains how the maximally invariant statistic at step n contains all information about the invariant component of the data at previous steps. An inequality of Ville [1939] and standard optional stopping theorems give the desired results. It is natural to ask whether (2.5) also holds for stopping times that are adapted to the full data $(X^n)_{n\in\mathbb{N}}$ instead of the reduced $(M_n)_{n\in\mathbb{N}}$ can be allowed. In our t-test example, this could be a stopping time τ^* such as "$\tau^* := 1$ if $|X_1| \notin [a,b]$; $\tau^* = 2$ otherwise" for some $0 < a < b$. The answer is negative: in Appendix A.2, we show that, for appropriate choice of a and b, this τ^* provides a counterexample. This means that such nonadapted τ^* cannot be safely used in an optional continuation context.

2.1.3. Optimality criteria for E-statistics

The conventional optimality criterion for hypothesis tests satisfying a type-I error guarantee is their fixed-sample-size or fixed-stopping-rule worst-case power maximization. This criterion cannot be used in a context with unknown stopping rules because this knowledge is required by the very definition of power. Similarly, the E-statistic which optimizes power for a given study with given stopping time will take on value 0 with positive probability, making it useless for optional continuation by multiplication. As GHK point out, a much more sensible criterion in both optional stopping and continuation settings is growth rate optimality in the worst case. Should it exist, an E-statistic T^*_n is GROW if it maximizes the worst-case expected logarithmic value under the alternative hypothesis, that is, if it maximizes

$$T_n \mapsto \inf_{\theta_1\in\Theta_1} \mathbf{E}_{\theta_1}[\ln T_n(X^n)] \tag{2.6}$$

over all E-statistics. The objective here is to gather evidence, measured by $T_n(X^n)$, as fast as possible. To this end, it is sensible to maximize expectation of $f(T_n(X^n))$ under the alternative, for some increasing function f. Shafer [2019] and GHK argue extensively why it makes sense to take f as the logarithm, an idea also known as *Kelly betting* [Kelly Jr., 1956]. Relatedly, this criterion produces tests with the smallest expected sample size until the null can rejected in a specific testing setting [Breiman, 1961].

Given its worst-case nature, GHK explain that the GROW E-statistic is too conservative in some scenarios and cannot be used if the alternative can be arbitrarily close to the null. For example, in the t-test this would mean that the value of $\delta = \mu/\sigma$ under the alternative is unknown. As a response to this issue, GHK propose to instead maximize a relative form of (2.6) to obtain less conservative E-statistics outside the worst-case regime (see also Turner et al. [2021] who, in their contingency table setting, achieve good results in practice with this relative criterion, but not with the absolute criterion). With this in mind, we say that an E-statistic T^*_n is relatively GROW if it maximizes the gain in expected logarithmic value relative to an oracle that is given the particular distribution in the alternative hypothesis from which data are generated, that is, if T_n^* maximizes, over all E-statistics,

$$T_n \mapsto \inf_{\theta_1 \in \Theta_1} \left\{ \mathbf{E}_{\theta_1}^{\mathbf{Q}}[\ln T_n(X^n)] - \sup_{T_n' \text{ E-stat.}} \mathbf{E}_{\theta_1}^{\mathbf{Q}}[\ln T_n'(X^n)] \right\}. \tag{2.7}$$

As we we will see and contrary to the general case, in our group-invariant setting, any GROW E-statistic is also relatively GROW. Hence, we can avoid discussing which of the two is more appropriate. While acknowledging that there may be situations in which an E-statistic optimality property distinct from being GROW is more relevant, in the remainder of this chapter we will simply take the goal of finding (relatively) GROW E-statistics for granted, without further motivation.

2.1.4. Main results

We now informally outline the main results of this chapter. The main result of *statistical* interest is Corollary 2.4.3, a characterization of the GROW E-statistic for the group-invariant problem defined in (2.2). This corollary is a consequence of Theorem 2.4.2, our main *technical* contribution. Recall that once data are reduced through a maximally invariant function $M_n = m_n(X^n)$ for the action of G on $\mathcal{X}^n$, the testing problem (2.2) becomes simple. We extend our results to situations when the invariance-reduced problem is still composite in Section 2.8. Sidestepping technicalities, the main statistical result is as follows:

Corollary 2.1.3 (Informal statement of Corollary 2.4.3)**.** *Under a number of technical conditions on the group G, among all possible* E*-statistics, G-invariant or not, the likelihood ratio $T_n^* = p_{\delta_1}^{M_n} / p_{\delta_0}^{M_n}$ for any maximally invariant function $M_n = m_n(X^n)$ is GROW for (2.2).*

We show further in Proposition 2.4.4 that, in our group-invariant setting, any GROW E-statistic is also relatively GROW, as defined in Section 2.1.2. With this theorem at hand, we characterize optimal E-statistics for group-invariant problems in fixed-sample experiments.

In Section 2.5 we further relate this result to sufficiency: we utilize the invariance and sufficiency reductions of Hall et al. [1965] to conclude that monitoring the likelihood ratio for $M_1, M_2, \ldots$ is equivalent to monitoring the likelihood ratio of a sufficient statistic for the maximally invariant parameter δ (see Proposition 2.5.2). Besides our running t-test example, in Section 2.7 we show two applications to testing under multivariate Gaussian assumptions: testing whether the population mean is zero, and testing whether a linear regression coefficient is zero. In Section 2.8 we further extend Corollary 2.1.3 to cases where the null and alternative hypotheses are still composite even after an invariance reduction of the data (see Proposition 2.8.1).

Technical contributions

Our main technical result is Theorem 2.4.2, a computation of the infimum value of the Kullback-Leibler (KL) divergence between elements in the convex hulls of the null and alternative models in (2.2). In Section 2.2, we show in detail how our approach operates in the simpler case when G is finite or compact. The main contribution in this chapter is the extension of this result to a large class of noncompact groups for which almost-right-invariant probability measures exist. The existence of such measures on G is known as amenability [Bondar and Milnes, 1981], and it is the key assumption in our results. The amenability condition, as will be stated Definition 2.2.1, is the same that is used in the classical theorem of Hunt and Stein [Lehmann and Romano, 2005, Section 8.5]. The proof techniques that are needed for the results of this work are, however, distinct. Hunt-Stein's theorem shows that, when looking for a test that is max-min optimal in the sense of power, it is enough to look among group-invariant tests. At the core of the proof of the Hunt-Stein theorem lays the fact that the power is a linear function of the test under consideration. In its proof, an approximate symmetrization of the test is carried using almost-right-invariant priors without affecting power guarantees. This line of reasoning cannot be directly translated to our setting because of the nonlinearity of the objective function that characterizes GROW E-statistics.

Besides the main technical contribution, Theorem 2.4.2, additional novel mathematical results are in Proposition 2.4.4, relating GROW to relatively GROW E-statistics, and the propositions in Section 2.8, extending Theorem 2.4.2 to settings in which $\mathcal{H}_0$ and $\mathcal{H}_1$ refer to composite sets of δ's and may be equipped with prior distributions.

2.1.5. Previous work

Invariance, as data-reduction method, has a long tradition in statistics [Eaton, 1989]. Perhaps the closest result to the ones we present is the classical theorem of Hunt and Stein [see Lehmann and Romano, 2005, Section 8.5]. It establishes that, in group-invariant models like the ones we treat here, there is no loss in considering only group-invariant tests when searching for most powerful tests at a fixed sample size. The relation of data reductions based in invariance and sufficiency are well understood [Hall et al., 1965]. In the Bayesian literature, group-invariant inference with right Haar priors has been thoroughly studied [Dawid et al., 1973, Berger et al., 1998]. It has been shown that, in contrast to some other improper priors, inference based on right Haar priors yields admissible procedures in a decision-theoretical sense [Eaton and Sudderth, 2002, 1999].However, there have also been concerns in the Bayesian literature [Sun and Berger, 2007, Berger and Sun, 2008] that in some situations, the right Haar prior is not uniquely defined, and different choices lead to different conclusions. Interestingly, as we discuss in Section 4.7, in our setting this issue cannot arise. Finally, we mention the work of Liang and Barron [2004], who provide exact min-max procedures for predictive density estimation for general location and scale families under Kullback-Leibler loss. Although there are clearly some similarities, the precise min-max result they prove is quite different; we provide a more detailed comparison, also in Section 4.7.

2.1.6. Notation

We use letters $\mathbf{P}$ and $\mathbf{Q}$ to refer to distribuitions of X^n. For a measurable function $T = T(X^n)$, we write $\mathbf{P}^T$ for the image measure of $\mathbf{P}$ under T, that is, $\mathbf{P}^T\{T \in B\} = \mathbf{P}\{T(X) \in B\}$. When writing conditional expectations, we write $\mathbf{E}^{\mathbf{P}}[f(X)|Y]$, and write $\mathbf{P}^{X|y}$ for the conditional distribution of X given $Y = y$. We only deal with situations where such conditional distributions exist. For a prior distribution $\mathbf{\Pi}$ on some parameter space Θ—with a suitable measurable structure—, we write $\mathbf{\Pi}^\theta\mathbf{P}_\theta$ for the marginal distribution that assigns probability $\mathbf{\Pi}^\theta\mathbf{P}_\theta\{X \in B\} = \int \mathbf{P}_\theta\{X \in B\}\mathrm{d}\mathbf{\Pi}(\theta)$ to any measurable set B. For the posterior distribution of θ given X we write $\mathbf{\Pi}^{\theta|X}$. Given two subsets H, K of a group G we write $HK = \{hk : h \in H, k \in K\}$ for the set of all possible products between an element of H and an element of K. Similarly, for an element $g \in G$ and a subset K of G, we define $gK = \{gk : k \in K\}$, the translation of K by g, and $K^{-1} = \{k^{-1} : k \in K\}$, the set of inverses of K. We say that K is symmetric if $K = K^{-1}$.

2.1.7. Outline

The rest of this chapter is structured as follows. We begin by describing our approach for finite and compact groups in Section 2.2. There, we also describe the challenges that are encountered when dealing with general groups and introduce the main group-theoretical condition, amenability. Next, in Section 2.3, we lay down formally the conditions necessary for our main results. In Section 2.4, we state the main results of this chapter in full. We continue in Section 2.5 by discussing our approach in the presence of a sufficient statistic for the models under consideration. We show, under regularity conditions, that there is no loss in further reducing the data through a sufficient statistic. With regards to anytime-valid testing, the subject of Section 2.6 is to show Proposition 2.1.2. In Section 2.7 we apply our results to two examples. In Section 2.8 we extend our results to cases in which, even after an invariance reduction of the data, the hypotheses at hand remain composite. We end this chapter with Section 4.7, where we discuss our results; and Section 2.10, where we give the proofs omitted from the rest of the text.

2.2. Technical outline

This section shows our techniques in the simple case when the group G in question is finite, and is intended to delineate our general approach. Next, we describe how we generalize the result to noncompact amenable groups, and point at the difficulties that are found. We start by reparametrizing the problem described in (2.2). Using that the action of the group on the parameter space is free, we can reparametrize each orbit in Θ/G with G. Indeed, we can pick an arbitrary but fixed element in the orbit $\theta_0 \in \delta_0$ and, for any other element $\theta \in \delta_0$, we can identify θ with the group element $g(\theta) \in G$ that transports θ_0 to θ, that is, such that $g(\theta)\theta_0 = \theta$. Hence, with a slight abuse of notation, we can identify $\theta \in \delta_0$ with $g = g(\theta) \in G$ and identify $\mathbf{P}_\theta = \mathbf{P}_{g(\theta)\theta_0}$ with $\mathbf{P}_g$. With analogous definitions, for a fixed $\theta_1 \in \delta_1$, the same identification can carried in the alternative model by an analogous choice θ_1. In order to make notation more succinct, we use $\mathcal{Q} = \{\mathbf{Q}_g\}_{g\in G}$ to denote the alternative hypothesis to $\mathcal{P} = \{\mathbf{P}_g\}_{g\in G}$. We assume that each member of $\mathcal{Q}$ is absolutely continuous with respect to each member of $\mathcal{P}$. With these remarks at hand, the starting problem (2.2) can be rewritten in the form

$$\mathcal{H}_0 : X^n \sim \mathbf{P}_g,\ \text{for some } g \in G, \text{ vs. } \mathcal{H}_1 : X^n \sim \mathbf{Q}_g,\ \text{for some } g \in G. \tag{2.8}$$

As will follow from our discussion, our results are insensitive to the choices of θ_0 and θ_1. Using the invariance of the models, we show in Proposition 2.4.4 that, in our setting, an E-statistic is GROW if and only if it is relatively GROW (see Section 2.1.2 for definitions).

2.2.1. Finite groups

Start by assuming that G is a finite group. For instance, a group of permutations. Then, if $M_n = m_n(X^n)$ is a maximally invariant function of X^n—a function that identifies in which orbit X^n is—, the distribution of M_n can be computed by averaging over the group. Indeed, by the invariance of $\mathcal{P}$ and $\mathcal{Q}$, a uniform distribution along each orbit is induced and each orbit is isomorphic to G because its action on $\mathcal{X}^n$ is free. In the general—possibly noncompact—case, we will use a measure-decomposition theorem of Wijsman [Andersson, 1982, Eaton, 1989]. Since M_n is G-invariant, its distribution does not depend on g. We call $\mathbf{P}^{M_n}$ and $\mathbf{Q}^{M_n}$ the distributions of M_n under any member of $\mathcal{P}$ and $\mathcal{Q}$, respectively. We call p^{M_n} and q^{M_n} their respective densities. Then, the so far hypothesized GROW E-statistic T_n^*, the likelihood ratio for the maximal invariant $M_n = m_n(X^n)$, satisfies

$$T_n^*(X^n) = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))} = \frac{\frac{1}{|G|}\sum_{g\in G} q_g(X^n)}{\frac{1}{|G|}\sum_{g\in G} p_g(X^n)}. \tag{2.9}$$

For finite parameter spaces, as in our current case, Theorem 1 of GHK takes a simple form: the value of the max-min problem that defines a GROW E-statistic coincides with that of a KL divergence minimization problem, that is,

$$\max_{T_n \text{ E-stat.}} \min_{g\in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n(X^n)] = \min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{g_1}\mathbf{Q}_{g_1}, \mathbf{\Pi}_0^{g_0}\mathbf{P}_{g_0}), \tag{2.10}$$

where $\mathrm{KL}(\mathbf{Q},\mathbf{P}) = \mathbf{E}^{\mathbf{Q}}[\ln(q/p)]$ is the KL divergence, and the minimum on the right hand side is taken over all pairs of distributions on the group G—we state a more general form of their result in Section 2.4. The crucial observation is that, if $\mathbf{\Pi}_{\mathrm{U}(G)}$ is the uniform distribution on the group G, then

$$\mathbf{E}_g^{\mathbf{Q}}[\ln T_n^*(X^n)] = \mathrm{KL}(\mathbf{\Pi}_{\mathrm{U}(G)}^{g_1}\mathbf{Q}_{g_1}, \mathbf{\Pi}_{\mathrm{U}(G)}^{g_0}\mathbf{P}_{g_0}) = \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}).$$

An application of the information processing inequality [Cover and Thomas, 2006, Section 2.8] implies that, for any pair $(\mathbf{\Pi}_0, \mathbf{\Pi}_1)$ of probability distributions on G,

$$\mathrm{KL}(\mathbf{\Pi}_1^{g_1}\mathbf{Q}_{g_1}, \mathbf{\Pi}_0^{g_0}\mathbf{P}_{g_0}) \geq \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}) = \min_{g\in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n^*(X^n)],$$

where the last equality follows from the fact that T_n^* from (2.9) only depends on X^n through the invariant $M_n = m_n(X^n)$ and consequently its distribution is independent of $g \in G$. Thus, (2.9) shows that the minimum KL of the right hand side of (2.10) is achieved for the particular choice of two uniform priors on G. Consequently, T_n^*, defined in (2.9), is a GROW E-statistic, that is,

$$\min_{g \in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n^*(X^n)] = \max_{T_n \text{ E-stat.}} \min_{g \in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n(X^n)]. \tag{2.11}$$

We now turn to the challenges encountered when dealing with infinite groups.

2.2.2. Noncompact groups

As we will see, a similar reasoning to that of the previous section can be carried out for compact groups—where there exists an invariant probability distribution—, but difficulties arise in the noncompact case. Luckily, these these dificulties can be circumvented under additional assumptions; among others, the assumption that the group G is amenable. Anytime that G is a locally compact topological group, there exist left- and right-invariant measures λ and ρ, respectively, on G [see Bourbaki, 2004, VII, §1,n° 2]. This means that, for any $g \in G$ and any $B \subseteq G$ measurable, $\lambda\{gB\} = \lambda\{B\}$ and $\rho\{Bg\} = \rho\{B\}$. The left and right Haar measures will take the place that the uniform distribution took on finite groups. For simplicity of exposition, let us assume that both probabilistic models are dominated by a left-invariant measure ν on $\mathcal{X}$. In that case, the invariance assumption (2.1) implies that the densities w.r.t. ν take the form $p_g(X^n) = p_1(g^{-1}X^n)$ and $q_g(X^n) = q_1(g^{-1}X^n)$, where 1 makes reference to the unit element of the group G. Using disintegration-of-measure results from Bourbaki [2004, VIII.27], Andersson [1982] argues that, in analogy to (2.9), for any locally compact group acting on $\mathcal{X}^n$—under mild regularity conditions on the action, which we will see—, the likelihood ratio for the maximal invariant $M_n = m_n(X^n)$ can be computed by integration over the group G, that is,

$$T_n^*(X^n) = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))} = \frac{\int_G q_g(X^n)\mathrm{d}\rho(g)}{\int_G p_g(X^n)\mathrm{d}\rho(g)}. \tag{2.12}$$

This is known as Wijsman's representation theorem [see also Eaton, 1989, Theorem 5.9]. If the right Haar measure ρ could always be chosen to be a probability measure, we could carry out the same computations that we made in the finite case of Section 2.2.1 to conclude that T_n^* is indeed GROW. However, the right Haar measure ρ is finite if and only if the group G at hand is compact [see Reiter and Stegeman, 2000, Proposition 3.3.5]. This is a severe limitation; it would not even cover our guiding example, the t-test, because the group $(\mathbb{R}^+, \cdot\ )$ is not compact (see Example 2.1.1). The main technical contribution of this chapter is the extension of this result to noncompact *amenable* groups, defined next, for which there exist *almost-right-invariant* probability measures.

Definition 2.2.1 (Amenability). A group G is amenable if there exists a sequence of *almost-right-invariant* probability distributions, that is, a sequence $\mathbf{\Pi}_1, \mathbf{\Pi}_2, \ldots$ such that, for any measurable set $B \subseteq G$ and group element $g \in G$,

$$\lim_{k \to \infty} |\mathbf{\Pi}_k \{B\} - \mathbf{\Pi}_k \{Bg\}| = 0.$$

Amenable groups have been thoroughly studied [Paterson, 2000] and include, among others, all finite, compact, commutative, and solvable groups. An example of a nonamenable group is the free group in two elements and any group containing it. A prominent example of a nonamenable group is that of invertible $d \times d$ matrices with matrix multiplication. Under the amenability of G and additional regularity conditions, we will show that, for a sequence $(\mathbf{\Pi}_k)_{k \in \mathbb{N}}$ of almost-right-invariant probability distributions on G,

$$\lim_{k \to \infty} \mathrm{KL}(\mathbf{\Pi}_k^g \mathbf{Q}_g, \mathbf{\Pi}_k^g \mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}) = \min_{g \in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n^*(X^n)],$$

where the last equality follows from the fact that $T_n^*(X^n)$ depends on X^n only through the maximal invariant M_n and, consequently, its distribution does not depend on g. From this, via Theorem 1 of GHK, the analogue of (2.11) holds and, consequently, as in the finite case of Section 2.2.1, T_n^* from (2.12) is GROW.

Example 2.1.1 (continued). The group $G = (\mathbb{R}^+, \cdot)$ of the t-test setting is amenable (it is a commutative group). The right Haar measure ρ on G is given by $\mathrm{d}\rho(\sigma) = \mathrm{d}\sigma/\sigma$, and the rightmost expression of (2.12) becomes, with $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$,

$$T_n^*(X^n) = \frac{\int_{\sigma > 0} \frac{1}{\sigma^n} \exp\left(-\frac{n}{2}\left[\left(\frac{\bar{X}_n}{\sigma} - \delta_1\right)^2 + \frac{1}{n}\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2\right]\right) \frac{\mathrm{d}\sigma}{\sigma}}{\int_{\sigma > 0} \frac{1}{\sigma^n} \exp\left(-\frac{n}{2}\left[\left(\frac{\bar{X}_n}{\sigma} - \delta_0\right)^2 + \frac{1}{n}\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2\right]\right) \frac{\mathrm{d}\sigma}{\sigma}}. \tag{2.13}$$

The expression (2.13) was obtained by Cox [1952] who realized that it was equivalent to the likelihood ratio of the maximal invariant. Lai [1976] also used it in an anytime-valid context (essentially exploiting that it gives an E-statistic). Our results establish, for the first time, that (2.13) is also GROW and relatively GROW. Lai also considered placing a proper prior distribution on δ_1; the same is done in *Jeffreys' Bayesian t-test* [Jeffreys and Jeffreys, 1998, Rouder et al., 2009]. We return to this idea in Section 2.8.

Consider now the sufficient statistic $s_n(X^n) = (\hat{\mu}_n, \hat{\sigma}_n)$, where $\hat{\mu}_n$ is the maximum likelihood estimator for the mean μ; and $\hat{\sigma}_n$, for the standard deviation σ. The t-statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(X^n) \propto \hat{\mu}_n/\hat{\sigma}_n$ is a maximally invariant function of the sufficient statistic. Our results imply that T_n^* also equals the likelihood ratio for $M_{\mathcal{S},n}$ is also relatively GROW, and that the test $\xi = (\xi_n)_{n \in \mathbb{N}}$ given by $\xi_n = \mathbf{1}\{T_n^* \geq 1/\alpha\}$ satisfies the conclusions of Proposition 2.1.2.

2.3. Assumptions

In this section we describe the assumptions made in our main results, their part in the proofs, and discuss their role for the purpose of parametric inference. We gather all assumptions below, in Assumption 1, for ease of reference. We start by laying out the conditions on the spaces involved, followed by those on the probabilistic models under consideration. Our additional assumptions on the group G, the parameter space Θ and the observation space $\mathcal{X}$ are topological in nature. They have two purposes. The first, in relation to the discussion of Section 2.2, is to ensure that (2.12), the representation theorem of Wijsman [Eaton, 1989, Theorem 5.9], holds (see Remark 2.3.3). The second purpose of our assumptions is to ensure that the observation space $\mathcal{X}^n$ can be put in bijective and bimeasurable[3] correspondence with a subset of $G \times \mathcal{X}^n / G$, where the group G acts naturally by multiplication on the first component. To this end, a theorem of Bondar [1976] is instrumental (see Remark 2.3.2).

We assume that G is a topological group, that is, a group equipped with a topology whose operation, seen as a function $G \times G \to G$, is continuous. We assume that all topological spaces under consideration are equipped with their Borel σ-algebra, the one generated by their topology. As topological spaces, we assume that both G and $\mathcal{X}$ are Polish—separable and completely metrizable—and locally compact. We assume that the action of G on $\mathcal{X}^n$ is continuous and proper; the latter means that the map $G \times \mathcal{X}^n \to \mathcal{X}^n \times \mathcal{X}^n$ defined by

$$(g, x^n) \mapsto (gx^n, x^n)$$

is proper, that is, the inverse of compact sets is compact. Properness ensures that the induced topology on the orbits $\mathcal{X}^n / G$ is Hausdorff, locally compact, and σ-finite [see Andersson, 1982]. We further assume that both probabilistic models are dominated by a common relatively left-invariant measure μ on $\mathcal{X}^n$ with some multiplier χ, that is, a measure μ such that, for any measurable set $B \subseteq \mathcal{X}^n$ and any group element $g \in G$, satisfies $\mu\{gB\} = \chi(g)\mu\{B\}$. We gather these assumptions for ease of reference.

Assumption 1. Let G be a topological group acting on $\mathcal{X}^n$, a topological space. The group G, the observation space $\mathcal{X}^n$, and the probabilistic models under consideration satisfy the following three properties:

1. As topological spaces, G and $\mathcal{X}^n$ are Polish—separable and completely metrizable—and locally compact.

2. The action of G on $\mathcal{X}^n$ is free, continuous and proper.

3. The models $\{\mathbf{P}_g\}_{g \in G}$ and $\{\mathbf{Q}_g\}_{g \in G}$ are invariant and have densities with respect to a common measure μ on $\mathcal{X}^n$ that is relatively left invariant with some multiplier χ.

Remark 2.3.1. Assumption 1 holds in most cases of interest for the purpose of parametric inference. We summarize some situations in which Assumption 1 holds, which will be helpful in Section 2.7, where we apply our results in two examples, a test for multivariate location and linear regression. Let $\mathcal{X} = \mathbb{R}^d$ and identify $\mathcal{X}^n$ with set of $d \times n$ matrices. Here we quote the properness of the action of two nonamenable groups on $\mathcal{X}^n$, which are consequences of the more general results of Wijsman [1985]. The relevant groups to Section 2.7 are closed amenable subgroups of those presented next, so that their actions on $\mathcal{X}^n$ are also proper.

[3] We call an invertible map bimeasurable if both the map and its inverse are measurable.

1. The general linear group in d dimensions $\mathrm{GL}(d)$, consisting of all $d \times d$ invertible real matrices with multiplication, acts continuously on $\mathcal{X}^n$ by left matrix multiplication. The continuous action of $\mathrm{GL}(d)$ on the restriction of $\mathcal{X}^n$ to matrices of rank d is free and proper any time that $n \geq d$. Seen as a subset of $\mathbb{R}^{d \times n}$, the restriction of the Lebesgue measure to $\mathcal{X}^n$ is relatively left-invariant with multiplier $\chi(g) = |\det(g)|^n$, for $g \in \mathrm{GL}(d)$.

2. The affine linear group $\mathrm{AL}(d)$, all pairs (A, b) with $A \in \mathrm{GL}(d)$ and $b \in \mathbb{R}^d$ with group operation $(A, v)(B, u) = (AB, Au + v)$, also acts continuously on $\mathcal{X}^n$. An action is given by $((A, b), X^n) \mapsto [Ax_1 + b, \ldots, Ax_n + b]$, where $x_1, \ldots, x_n$ are the columns of $X^n \in \mathcal{X}^n$, and the square brackets make reference to the matrix with the given columns. This action is proper on the restriction of $\mathcal{X}^n$ to matrices of rank d any time that $n \geq d + 1$. Seen as a subset of $\mathbb{R}^{d \times n}$, the restriction of the Lebesgue measure to $\mathcal{X}^n$ is relatively left-invariant with multiplier $\chi(g) = |\det(A)|^n$ for $g = (A, v) \in \mathrm{AL}(d)$.

Remark 2.3.2. We use in the proof of the main theorem that, under these assumptions, the space $\mathcal{X}^n$ can be put in one-to-one bimeasurable correspondence with a subset of $G \times \mathcal{X}^n/G$, where G acts naturally by multiplication in the first component. More explicitly, under assumptions 1 and 2, Theorem 2 of Bondar [1976] guarantees the existence of a one-to-one map $r : \mathcal{X}^n \to G \times \mathcal{X}^n/G$ such that both r and its inverse are measurable, and, anytime that $x^n \mapsto (h(x^n), m(x^n))$, then, for any $g \in G$, the image of gx^n under r is $(gh(x^n), m(x^n))$.

Remark 2.3.3. The topological conditions under which Wijsman's representation theorem [Eaton, 1989, Theorem 5.9] holds are weaker than those presented in Assumption 1 (see also the previous remark). For the representation theorem to hold, it is only necessary that $\mathcal{X}^n$ and G are locally compact and that the action of G is both continuous and proper. Notice also that this representation theorem holds for nonamenable groups.

Remark 2.3.4. In our proofs, it will be useful to use, without loss of generality, the following modification to 3 in Assumption 1:

3' The models $\{\mathbf{P}_g\}_{n \in \mathbb{N}}$ and $\{\mathbf{Q}_g\}_{n \in \mathbb{N}}$ are invariant and have densities with respect to a common measure ν on $\mathcal{X}^n$ that is left invariant.

The reason that there is no loss in generality is that from any relatively left-invariant measure μ with multiplier χ, a left-invariant measure ν can be constructed. Indeed, Bourbaki [2004, Chap. 7, §2 Proposition 7] shows that, under our assumptions, for any multiplier χ there exists a function $\varphi : \mathcal{X}^n \to \mathbb{R}$ with the property that $\varphi(gx) = \chi(g)\varphi(x)$ for any $x \in \mathcal{X}$ and $g \in G$. With this function at hand, one can define the measure $\mathrm{d}\nu(x) = \mathrm{d}\mu(x)/\varphi(x)$, which is left invariant. After multiplication by φ, probability densities with respect to μ are readily transformed into probability densities with respect to ν.

Remark 2.3.5. On any locally compact group G there exists a left-invariant measure λ, called left Haar measure. It can be shown that λ is relatively right invariant with a multiplier Δ, that is, for any measurable $B \subseteq G$ and $g \in G$ it holds that $\lambda^h\{Bg\} =$

$\Delta(g)\lambda^h\{B\}$ for any $g \in G$. This multiplier is called the (right) modulus of the group G. A computation shows that the measure ρ defined by $\rho^h\{B\} = \lambda^h\{B^{-1}\}$ for each measurable $B \subseteq G$, is right invariant, in other words, ρ is a right Haar measure. In the following, we always refer to right and left Haar measures that are related to each other by that identity. In our proofs we will use that for any integrable function f defined on G, the identities $\int f(h)\mathrm{d}\rho(h) = \int f(h)/\Delta(h)\mathrm{d}\lambda(h)$ and $\int f(h^{-1})\mathrm{d}\lambda(h) = \int \mathrm{d}f(h)\mathrm{d}\rho(h)$ hold [see Eaton, 1989, Section 1.3].

2.4. Main Result

In this section, we state in full detail the main result of this chapter, Corollary 2.4.3, a characterization of the GROW statistic for the statistical hypothesis testing problem (2.8). In Corollary 2.4.5 we will show that any GROW E-statistic is also relatively GROW in our group-invariant setting. Our main result stems from an application of the main technical contribution of this chapter, Theorem 2.4.2, which shows that the infimum Kullback-Leibler (KL) divergence between the elements of the convex hulls of the null and alternative hypotheses is exactly equal to the KL divergence between the distributions of the maximal invariant under both models. Theorem 2.4.2 will allow us to directly apply GHK's Theorem 1, which provides a general recipe for constructing the GROW E-statistic in terms of the KL minimization problem (or joint information projection in information theoretic terminology). For simplicity and completeness, we present here a special case of GHK's Theorem 1 that will be used in our group-invariant setting.

Theorem 2.4.1 (Theorem 1 of GHK, most general version given in their Section 4.3)**.** *Let $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta_0}$ and $\mathcal{Q} = \{\mathbf{Q}_\theta\}_{\theta\in\Theta_1}$ be two families of probability distributions on $\mathcal{X}^n$ that are dominated by a common measure. Suppose that there exists a random variable $V_n = v_n(X^n)$ such that*

$$\inf_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{\theta_1}\mathbf{Q}_{\theta_1}, \mathbf{\Pi}_0^{\theta_0}\mathbf{P}_{\theta_0}) = \min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{\theta_1}\mathbf{Q}_{\theta_1}^{V_n}, \mathbf{\Pi}_0^{\theta_0}\mathbf{P}_{\theta_0}^{V_n}) < \infty, \tag{2.14}$$

where the minimum and the infimum are over all pairs of proper probability distributions on Θ_0 and Θ_1. Let $\mathbf{\Pi}_0^\star$ and $\mathbf{\Pi}_1^\star$ be the pair of probability distributions—on Θ_0 and Θ_1, respectively—-where the previous minimum is achieved, that is,

$$\min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{\theta_1}\mathbf{Q}_{\theta_1}^{V_n}, \mathbf{\Pi}_0^{\theta_0}\mathbf{P}_{\theta_0}^{V_n}) = \mathrm{KL}(\mathbf{\Pi}_1^{\star\theta_1}\mathbf{Q}_{\theta_1}^{V_n}, \mathbf{\Pi}_0^{\star\theta_0}\mathbf{P}_{\theta_0}^{V_n}).$$

Then

$$\max_{T_n \text{ E-}stat.} \inf_{\theta_1\in\Theta_1} \mathbf{E}_{\theta_1}^{\mathbf{Q}}[\ln T_n(X^n)] = \mathrm{KL}(\mathbf{\Pi}_1^{\star\theta_1}\mathbf{Q}_{\theta_1}^{V_n}, \mathbf{\Pi}_0^{\star\theta_0}\mathbf{P}_{\theta_0}^{V_n}).$$

In that case, the maximum on the left hand side of the previous display is achieved by the E*-statistic $T_n^\star$ given by*

$$T_n^\star(X^n) := \frac{\int q_{\theta_1}^{V_n}(v_n(X^n))\mathrm{d}\mathbf{\Pi}_1^\star(\theta_1)}{\int p_{\theta_0}^{V_n}(v_n(X^n))\mathrm{d}\mathbf{\Pi}_0^\star(\theta_0)},$$

that is, $T_n^\star$ is GROW for testing $\mathcal{P}$ against $\mathcal{Q}$.

Once the connection between GROW E-statistics and KL divergence minimization is established, our next step is Theorem 2.4.2. In this section, we only treat the case in which, after an invariance reduction of the data, both null and alternative hypothesis become simple so that the minimum in (2.14) trivializes. Theorem 2.4.2 establishes that, under our assumptions, (2.14) does indeed hold where V_n plays the role of the maximally invariant statistic M_n and $\Theta_0 = \Theta_1 = G$ refer to the group. In Section 2.8 we investigate the case when the hypotheses are still composite after the invariance reduction. Theorem 2.4.2 below immediately implies that the likelihood ratio for the maximal invariant is GROW; we delay its proof to Section 2.10.

Theorem 2.4.2 (Main technical result). *Let $M_n = m_n(X^n)$ be a maximally invariant function of the data X^n under the action of the group G on $\mathcal{X}^n$. Under Assumption 1, assume further that the group G is amenable as in Definition 2.2.1, and that there is $\varepsilon > 0$ such that*

$$\mathbf{E}_1^{\mathbf{Q}}\left[\left|\ln\frac{q_1(X^n)}{p_1(X^n)}\right|^{1+\varepsilon}\right], \mathbf{E}^{\mathbf{Q}}\left[\left|\ln\frac{q^{M_n}(M_n)}{p^{M_n}(M_n)}\right|^{1+\varepsilon}\right] < \infty, \tag{2.15}$$

where the subindex in $\mathbf{Q}_1$ refers to the unit element of G, and the second expected value is with respect to the distribution of M_n under any of the members of $\{\mathbf{Q}_g\}_{g \in G}$. Then,

$$\inf_{\mathbf{\Pi}_0, \mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^g \mathbf{Q}_g, \mathbf{\Pi}_0^g \mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}),$$

where the infimum is taken over all pairs $(\mathbf{\Pi}_0, \mathbf{\Pi}_1)$ probability distributions on the group G.

From our previous discussion and with Theorem 2.4.2 at hand, the main result of this chapter follows.

Corollary 2.4.3 (Main 'statistical' result). *Under the assumptions of Theorem 2.4.2, a GROW* E*-statistic T^* for (2.8) is given by*

$$T_n^*(X^n) = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))},$$

the likelihood ratio for any maximally invariant statistic $M_n = m_n(X)$.

Example 2.1.1 (continued). We had already established that the group $G = (\mathbb{R}^+, \cdot)$ of the t-test setting is amenable and satisfies Assumption 1. The condition (2.15) is also readily verified; we can apply Corollary 2.4.3 to conclude that (2.13) is a GROW E-statistic.

We end by showing that, in our group-invariant setting, any statistic that is GROW is also relatively GROW, meaning that any E-statistic that maximizes (2.7) also maximizes (2.6). This is not true in general; the result relies crucially on the invariant structure of the models under consideration. For example, for contingency tables, the two E-statistics are wildly different (GHK). We give the proof of the following proposition at the end of the section.

Proposition 2.4.4. *Suppose that the models* $\{\mathbf{P}_g\}_{g\in G}$ *and* $\{\mathbf{Q}_g\}_{g\in G}$ *satisfy 3 of Assumption 1 and suppose that, for each* $g \in G$*, there exists* $h \in G$ *such that* $\mathrm{KL}(\mathbf{Q}_g, \mathbf{P}_h)$ *is finite. Then the map defined by*

$$g \mapsto \sup_{T_n \text{ E-}stat.} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n(X^n)]$$

is constant. Consequently, any maximizer of (2.7) also maximizes (2.6), that is, an E*-statistic is relatively GROW if and only if it is also GROW for the hypothesis testing problem (2.8).*

After inspecting that Proposition 2.4.4 indeed applies under the assumptions of Corollary 2.4.3, we can conclude the following corollary, the main objective of this section.

Corollary 2.4.5. *Not only is* T^* *from Corollary 2.4.3 GROW, it is also relatively GROW.*

Proof. It is only left to check that, under the assumptions of Corollary 2.4.3, Proposition 2.4.4 applies. This is indeed the case because, by the invariance of the model and Hölder's inequality,

$$\mathrm{KL}(\mathbf{Q}_g, \mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}_1, \mathbf{P}_1) \leq \left(\mathbf{E}_1^{\mathbf{Q}}\left[\left|\ln \frac{q_1(X^n)}{p_1(X^n)}\right|^{1+\varepsilon}\right]\right)^{\frac{1}{1+\varepsilon}},$$

which was assumed to be finite. □

Proof of Proposition 2.4.4. Let g be a fixed group element of G. Recall from Remark 2.3.4 that we may assume that both models are dominated by a left invariant measure ν on $\mathcal{X}$. By Theorem 1 of GHK (its simplest instantiation in their Section 2), any time that $\inf_{h\in G} \mathrm{KL}(\mathbf{Q}_g, \mathbf{P}_h) < \infty$, there exists a subprobability density $\bar{p}$ on $\mathcal{X}^n$ relative to the left-invariant measure ν with two key properties: first, the function $T_n^\star(X^n) = q_g(X^n)/\bar{p}(X^n)$ is an E-statistic; second, $T_n^\star$ achieves the supremum in (2.7). Moreover, the theorem implies that

$$\sup_{T \text{ E-stat.}} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n(X^n)] = \mathbf{E}_g^{\mathbf{Q}}\left[\ln \frac{q_g(X^n)}{\bar{p}(X^n)}\right] = \inf_{\mathbf{\Pi}_0} \mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}_0^{g'} \mathbf{P}_{g'}), \tag{2.16}$$

where the infimum is over all distributions $\mathbf{\Pi}_0$ on G. We will show that for any pair $g, h \in G$ and any prior $\mathbf{\Pi}$ on G, there exists a prior $\tilde{\mathbf{\Pi}}$ such that

$$\mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}_0^{g'} \mathbf{P}_{g'}) = \mathrm{KL}(\mathbf{Q}_h, \tilde{\mathbf{\Pi}}^{g'} \mathbf{P}_{g'}). \tag{2.17}$$

From this, our claim will follow: by symmetry, the previous display implies that $g \mapsto \sup_{T_n \text{ E-stat.}} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n(X^n)]$ is constant over G because of its relation to the KL minimization in (2.16). Let $\bar{p} = \int p_g \mathrm{d}\mathbf{\Pi}(g)$, use both the invariance of ν and of $\mathcal{Q}$, and compute

$$\begin{aligned} \mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}^{g'}\mathbf{P}_{g'}) &= \mathbf{E}_g^{\mathbf{Q}}\left[\ln \frac{q_g(X^n)}{\bar{p}(x^n)}\right] \\ &= \int q_g(x^n) \ln \frac{q_g(x^n)}{\bar{p}(x^n)} \mathrm{d}\nu(x^n) \\ &= \int q_h(hg^{-1}x^n) \ln \frac{q_h(hg^{-1}x^n)}{\bar{p}(x^n)} \mathrm{d}\nu(x^n). \end{aligned}$$

Next, define $\tilde{\mathbf{\Pi}}$ as the probability distribution on G that assigns $\tilde{\mathbf{\Pi}}\{B\} = \mathbf{\Pi}\{gh^{-1}B\}$ for any measurable set $B \subseteq G$. Then

$$\bar{p}(x^n) = \int p_g(x^n)\mathrm{d}\mathbf{\Pi}(g) = \int p_{gh^{-1}g}(x^n)\mathrm{d}\tilde{\mathbf{\Pi}}(g) = \int p_g(hg^{-1}x^n)\mathrm{d}\tilde{\mathbf{\Pi}}(g).$$

Define $\tilde{p} = \int p_g \mathrm{d}\tilde{\mathbf{\Pi}}(g)$. The two last displays together imply that

$$\mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}^{g'}\mathbf{P}_{g'}) = \int q_h(hg^{-1}x^n) \ln \frac{q_h(hg^{-1}x^n)}{\tilde{p}(hg^{-1}x^n)} \mathrm{d}\nu(x^n).$$

After a change of variable and using the invariance of ν, the right hand side of this equation equals $\mathrm{KL}(\mathbf{Q}_g, \tilde{\mathbf{\Pi}}^{g'}\mathbf{P}_{g'})$. Thus, this last equation is nothing but (2.17), as was our objective. By our previous discusion, the result follows. □

2.5. Invariance and Sufficiency

The relationship between invariance and sufficiency has been thoroughly investigated [Hall et al., 1965, 1995, Berk, 1972, Nogales and Oyola, 1996]. Consider a G-invariant hypothesis testing problem such that a sufficient statistic is available. If the action of G on the original data space induces a free action on the sufficient statistic, there must be a maximally invariant function of the sufficient statistic. With this structure in mind, the results presented thus far suggest two approaches for solving the hypothesis testing problem. The first is to reduce the data using the sufficient statistic, and to test the problem using the maximally invariant function of the sufficient statistic. The second approach is to use the maximally invariant function of the original data. These two approaches yield two potentially different growth-optimal E-statistics, and one question arises naturally: are both approaches equivalent? In this section we show that this is indeed the case, under certain conditions.

We now introduce the setup formally. At the end of this section we revisit our guiding example, the t-test, and show how the results of this section apply to it. Let Θ be the parameter space, and let $\delta = \delta(\theta)$ be a maximally invariant function of θ for the action of G on Θ. Let $s_n : \mathcal{X}^n \to \mathcal{S}_n$ be a sufficient statistic for $\theta \in \Theta$. Consider again the hypothesis testing problem in the form presented in (2.2). Assume further that G acts freely and continuously on the image space $\mathcal{S}_n$ of the sufficient statistic

$S_n = s_n(X^n)$, and assume that s_n is compatible with the action of G in the sense that, for any $X^n \in \mathcal{X}^n$ and any $g \in G$, the identity $gs_n(X^n) = s_n(gX^n)$ holds, where $(g, s) \mapsto gs$ makes reference to the action of G on $\mathcal{S}_n$. Let $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n)$ and $M_{\mathcal{S},n} = m_{\mathcal{S},n}(S_n)$ be two maximally invariant functions for the actions of G on $\mathcal{X}^n$ and $\mathcal{S}_n$, respectively. Because of their invariance, the distributions of $M_{\mathcal{X},n}$ and $M_{\mathcal{S},n}$ depend only on the maximally invariant parameter δ. Hall et al. [1965, Section II.3] proved that, under regularity conditions, if $S_{\mathcal{X},n} = s_{\mathcal{X},n}(X^n)$ is sufficient for $\theta \in \Theta$, then the statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is sufficient for δ. In that case, we call $M_{\mathcal{S},n}$ invariantly sufficient. Here we state the version of their result, attributed by Hall et al. [1965] to C. Stein, that suits best our purposes[4].

Theorem 2.5.1 (C. Stein). *If there exists a Haar measure on the group G, the statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is invariantly sufficient, that is, it is sufficient for the maximally invariant parameter δ.*

With this theorem at hand, and the fact that the KL divergence does not decrease by the application of sufficient transformations, we obtain the following proposition.

Proposition 2.5.2. *Let $s_n : \mathcal{X}^n \to \mathcal{S}_n$ be sufficient statistic for $\theta \in \Theta$. Assume that G acts freely on $\mathcal{S}_n$ and that $s_n(gX^n) = gs_n(x^n)$ for all $X^n \in \mathcal{X}^n$ and $g \in G$. Let $m_{\mathcal{S},n}$ be a maximal invariant for the action of G on $\mathcal{S}_n$, and let $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$. Then,*

$$\mathrm{KL}\left(\mathbf{P}_{\delta_1}^{M_{\mathcal{X},n}}, \mathbf{P}_{\delta_0}^{M_{\mathcal{X},n}}\right) = \mathrm{KL}\left(\mathbf{P}_{\delta_1}^{M_{\mathcal{S},n}}, \mathbf{P}_{\delta_0}^{M_{\mathcal{S},n}}\right).$$

Proof. The function $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is invariant, and consequently its distribution only depends on the maximally invariant parameter δ. Since $M_{\mathcal{X},n}$ is maximally invariant for the action of G on $\mathcal{X}^n$, there is a function f such that $M_{\mathcal{S},n} = f(M_{\mathcal{X},n})$. By Stein's theorem, Theorem 2.5.1, $M_{\mathcal{S},n}$ is sufficient for δ. Consequently, f is a sufficient transformation. Hence, from the invariance of the KL divergence under sufficient transformations, the result follows. □

Via the factorization theorem of Fisher and Neyman, the likelihood ratio for the maximal invariant $M_{\mathcal{X},n}$ coincides with that of the invariantly sufficient $M_{\mathcal{S},n}$. As a consequence, we obtain the answer to the motivating question of this section: performing an invariance reduction on the original data and on the sufficient statistic are equivalent.

Corollary 2.5.3. *Under the assumptions of Proposition 2.5.2, if $S_n = s_n(X^n)$,*

$$T_n^*(X^n) = \frac{p_{\delta_1}^{M_{\mathcal{X},n}}(m_{\mathcal{X},n}(X^n))}{p_{\delta_0}^{M_{\mathcal{X},n}}(m_{\mathcal{X},n}(X^n))} = \frac{p_{\delta_1}^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}{p_{\delta_0}^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}.$$

Hence, if assumptions of Corollary 2.4.3 also hold, the likelihood ratio for the invariantly sufficient statistic $M_{\mathcal{S},n}$ is (relatively) GROW.

[4] The assumption that there exists an invariant measure on G implies what Hall et al. [1965] call Assumption A. [see Hall et al., 1965, discussion in p. 581]

Example 2.1.1 (continued). We have seen that a maximally invariant function of the data is $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n) = (X_1/|X_1|,\dots,X_n/|X_1|)$ while the t-statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(X^n) \propto \hat{\mu}_n/\hat{\sigma}_n$ is a maximally invariant function of the sufficient statistic $s_n(X^n) = (\hat{\mu}_n, \hat{\sigma}_n)$. Stein's theorem (Theorem 2.5.1) shows that the t-statistic $M_{\mathcal{S},n}$ is sufficient for the maximally invariant parameter $\delta = \mu/\sigma$. Corollary 2.5.3 shows that the likelihood ratio for the t-statistic is relatively GROW.

2.6. Anytime-valid testing under group invariance

The main objective of this section is to prove Proposition 2.1.2, the main consequence of our results pertaining testing under optional stopping and continuation. We now assume that the observations are made sequentially. At the end of the section we describe the consequences to our main example, the t-test. We begin by defining our working model for this scenario. Let $X = (X_n)_{n\in\mathbb{N}}$ be a random process, where each X_n is an observation that takes values on a space $\mathcal{X}$. Let $(M_n)_{n\in\mathbb{N}}$ be a sequence where, for each n, $M_n = m_n(X^n)$ is a maximally invariant function for the action of G on $\mathcal{X}^n$. If, at each sample size n, the assumptions of Corollary 2.4.3 hold, we have shown that

$$T_n^*(X^n) = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))}, \tag{2.18}$$

the likelihood ratio for the maximal invariant $M_n = m_n(X^n)$, defines a sequence $T^* = (T_n^*)_{n\in\mathbb{N}}$ of relatively GROW E-statistics for (2.8). With an eye towards proving Proposition 2.1.2, in the next proposition we show that $T^* = (T_n^*)_{n\in\mathbb{N}}$ is a martingale.

Proposition 2.6.1. *If $M = (M_n)_{n\in\mathbb{N}}$ is a sequence of maximally invariant statistics $M_n = m_n(X^n)$ for the action of G on $\mathcal{X}^n$, the process $T^* = (T_n^*)_{n\in\mathbb{N}}$ given by (2.18) is a nonnegative martingale with respect to M under any of the elements of the null hypothesis $\{\mathbf{P}_g\}_{g\in G}$.*

Proof. Let $g \in G$ be arbitrary but fixed. We start by showing that T_n^* equals the likelihood ratio for $M^n = (M_1,\dots,M_n)$ between $\mathbf{P}_g$ and $\mathbf{Q}_g$. For each $t > 1$, the maximally invariant statistic at $n-1$, $M_{n-1} = m_{n-1}(X^{n-1})$ is invariant if seen as a function of X^n. Hence, by the maximality of m_n, M_{n-1} can be written as a function of M_n. Repeating this reasoning $n-1$ times yields that M_n contains all information about the value of $M^{n-1} = (M_1,\dots,M_{n-1})$, all the maximally invariant statistics at previous times. Two consequences fall from these observations. First, no additional information about T_n^* is gained by knowing the value of $M^{n-1} = (M_1,\dots,M_{n-1})$ with respect to only knowing M_{n-1}, that is, $\mathbf{E}_g^{\mathbf{P}}\left[T_n^*|M_{n-1}\right] = \mathbf{E}_g^{\mathbf{P}}\left[T_n^*|M^{n-1}\right]$. Second, the likelihood ratio between $\mathbf{P}_g$ and $\mathbf{Q}_g$ for the sequence $M_1,\dots,M_n$ equals the likelihood ratio for M_n alone, that is,

$$T_n^*(X^n) = \frac{q^{M_1,\dots,M_n}(m_1(X^1),\dots,m_n(X^n))}{p^{M_1,\dots,M_n}(m_1(X^1),\dots,m_n(X^n))}.$$

The previous two consequences, and a computation, together imply that T^* is an M-martingale under $\mathbf{P}_g$, that is, $\mathbf{E}_g^{\mathbf{P}}\left[T_n^*|M^{n-1}\right] = T_{n-1}^*$. Since $g \in G$ was arbitrary, the result follows. □

With this result at hand, we are in the position to prove Proposition 2.1.2 from Section 2.1.4, the main result in this work pertaining sequential testing. We end this section with the implications to the t-test.

Proof of Proposition 2.1.2. From Proposition 2.6.1, we know that $T^* = (T_n)_{n\in\mathbb{N}}$ is a nonnegative martingale with expected value equal to one. Let $\xi = (\xi_n)_n$ be the sequential test given by $\xi_n = \mathbf{1}\left\{T_n^* \geq 1/\alpha\right\}$. The anytime-validity at level α of ξ, is a consequence of Ville's inequality, and the fact that the distribution of each T_n^* does not depend on g. Indeed, these two, together, imply that

$$\sup_{g\in G} \mathbf{P}_g\{T_n^* \geq 1/\alpha \text{ for some } n \in \mathbb{N}\} \leq \alpha.$$

This implies the first statement. Now, let $\tau \leq \infty$ be a stopping time with respect to M. If the stopping time τ is almost surely bounded, T_τ^* is an E-statistic by virtue of the optional stopping theorem. However, since T^* is a nonnegative martingale, Doob's martingale convergence theorem implies the existence of an almost sure limit T_∞^*. Even when τ might be infinite with positive probability, Theorem 4.8.4 of Durrett [2019] implies that T_τ^* is still an E-statistic. □

Example 2.1.1 (continued). In the previous section we saw that T_n^*, the likelihood ratio for the t-statistic is a GROW E-statistic. This, in conjunction with Proposition 2.1.2 implies that the test $\xi = (\xi_n)_{n\in\mathbb{N}}$ defined by $\xi_n = \mathbf{1}\left\{T_n^* \geq 1/\alpha\right\}$ is anytime-valid at level α and that the randomly stopped E-statistic T_τ^* remain one as long as the stopping time τ is with respect to the sequence of maximally invariant statistics. In Appendix A.2 we show a situation where the optionally stopped E-statistic is not an E-statistic if we take a stopping time that depends on the full data.

2.7. Testing multivariate normal distributions under group invariance

We show how the theory developed in the previous sections can be applied to hypothesis testing under normality assumptions. The family of d-dimensional normal distributions carries a natural invariance under scale-location transformations. The group of interest is the affine linear group $\mathrm{AL}(d)$, the group consisting of all pairs (v, A) with $v \in \mathbb{R}^d$, and A an invertible $d \times d$ matrix, and group operation $(v, A)(u, B) = (v + Au, AB)$. By considering amenable subgroups of $\mathrm{AL}(d)$, we obtain useful examples to which our results apply. We develop two in detail. The first is an alternative to Hotelling's T^2 for testing whether the mean of the distribution is identically zero, and results from the consideration $A \in \mathrm{LT}^+(d)$, the group of lower triangular matrices with positive entries on the diagonal, and $v = 0$. This test is in direct relation with the step-down procedure of Roy and Bargmann [1958][5] [see also Subbaiah and Mudholkar, 1978]. The second example that we consider is, in the setting of linear regression, a test for whether or not a specific regression coefficient is identically zero. It results from the restriction $A = cI$, a multiple of the $d \times d$ identity matrix.

2.7.1. The lower triangular group

Consider data $X^n = (X_1, \ldots, X_n)$ where $X_i \in \mathcal{X} = \mathbb{R}^d$. We assume each X_i to have a Gaussian distribution $N(\mu, \Sigma)$ with unknown mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ. We consider a test for whether the mean μ of the distribution is zero. Before stating explicitly our hypothesis testing problem, we first reparametrize the Gaussian model using Cholesky's decomposition. Indeed, for a positive definite matrix Σ, its Cholesky decomposition is $\Sigma = \Lambda\Lambda'$ for a unique $\Lambda \in \mathrm{LT}^+(d)$. Consequently, $\mathrm{LT}^+(d)$ can be used to parametrize all covariance matrices. Hence, we may take the parameter space Θ to be $\Theta = \mathbb{R}^d \times \mathrm{LT}^+(d)$. In this parametrization, the likelihood of the original data $X^n = (X_1, \ldots, X_n)$ takes the form

$$p^{X^n}_{\Lambda,\delta}(X^n) = \frac{1}{(2\pi)^n (\det \Lambda)^n} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left\| \Lambda^{-1} X_i - \delta \right\|^2 \right).$$

Consider the following hypothesis testing problem, which generalizes the t-test to dimensions $d \geq 1$:

$$\mathcal{H}_0 : \Lambda^{-1}\mu = \delta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \Lambda^{-1}\mu = \delta_1, \tag{2.19}$$

from which a test for whether μ is zero can be obtained by setting $\delta_0 = 0$. We now apply our results to this testing problem. Recall that the group $\mathrm{LT}^+(d)$ is amenable and acts on Θ by

$$(L, (\mu, \Lambda)) \mapsto (L\mu, L\Lambda) \tag{2.20}$$

for each $(\mu, \Lambda) \in \Theta$ and $L \in \mathrm{LT}^+(d)$, and a maximally invariant parameter is $\delta = \Lambda^{-1}\mu$. The group $\mathrm{LT}^+(d)$ acts on $\mathcal{X}^n$ by componentwise matrix multiplication, and the Gaussian model is invariant under this action. With the help of Remark 2.3.1, the assumptions of Corollary 2.4.3 are readily checked anytime that $n \geq d$, and we can conclude that, for any maximally invariant function $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n)$ of the data, the likelihood ratio $T_n^* = p_{\delta_1}^{M_{\mathcal{X},n}} / p_{\delta_0}^{M_{\mathcal{X},n}}$ is GROW. However, from our discussion in Section 2.5, this likelihood ratio coincides with that of a invariantly sufficient statistic for δ. We now proceed to compute one such a statistic. Recall that the pair $S_n = s_n(X^n) = (\bar{X}_n, \bar{V}_n)$, consisting of the unbiased estimators $\bar{X}_n$ for the mean and the covariance matrix $\bar{V}_n$ is a sufficient statistic for (μ, Σ). We can apply to the sufficient statistic the same considerations that we applied to the parameter space. For $n \geq d$, we can perform the Cholesky decomposition of the empirical covariance matrix $\bar{V}_n =$

[5] Even though not explicitly in group-theoretic terms, the test of Roy and Bargmann [1958] test is based on a different maximally invariant function of the data. The fact that the test statistic of Roy and Bargmann [1958] is maximally invariant under the action of $\mathrm{LT}^+(d)$ is shown by Subbaiah and Mudholkar [1978]

$L_n L_n'$. The statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(S_n) = \sqrt{\frac{n}{n-1}} L_n^{-1} \bar{Y}_n$ is maximally invariant under the action (2.20), and, by our discussion from Section 2.5, invariantly sufficient. In other words, $M_{\mathcal{S},n}$ is sufficient for δ. Hence, the GROW E-statistic can be written as $T_n^* = p_{\delta_1}^{M_{\mathcal{S},n}} / p_{\delta_0}^{M_{\mathcal{S},n}}$. For the purposes of sequential testing, Proposition 2.1.2 shows that the sequential test $(\xi_{n,\alpha}^* : n \in \mathbb{N})$ with $\xi_{n,\alpha}^* = \mathbf{1}\{T_n^* \geq 1/\alpha\}$ is anytime-valid. For completeness, we give an explicit expression for the likelihood ratio $T_{\mathcal{S},n}^*$ when $\delta_0 = 0$. From this expression, the likelihood ratio for other values of δ_0 can be computed. We show the computations in Proposition A.1.1.

Lemma 2.7.1. For the maximally invariant statistic $M_{\mathcal{S},n} = \sqrt{\frac{n}{n-1}} L_n^{-1} \bar{Y}_n$, we have

$$\frac{p_\delta^{M_{\mathcal{S},n}}(M_{\mathcal{S},n})}{p_0^{M_{\mathcal{S},n}}(M_{\mathcal{S},n})} = \mathrm{e}^{-\frac{n}{2}\|\delta\|^2} \int \mathrm{e}^{n\langle \delta, T A_n^{-1} M_{\mathcal{S},n}\rangle} \mathrm{d}\mathbf{P}_{n,I}(T), \tag{2.21}$$

where A_n is the lower triangular matrix resulting from the Cholesky decomposition $I + M_{\mathcal{S},n} M_{\mathcal{S},n}' = A_n A_n'$, and $\mathbf{P}_{n,I}^T$ is the distribution according to which $nTT' \sim W(n, I)$, a Wishart distribution.

Proof. Use Proposition A.1.1 with $\gamma = \sqrt{n}\delta$, $X = \sqrt{n}\bar{X}_n$, $m = n-1$, and $S = \bar{V}_n$. □

2.7.2. A subset of the affine group $\mathrm{AL}(d)$: linear regression

Consider the problem of testing whether one of the coefficients of a linear regression is zero under Gaussian error assumptions. Assume that the observations are of the form $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$, where, for each i, $X_i, Y_i \in \mathbb{R}$ and $Z_i \in \mathbb{R}^d$. We consider the the linear model given by

$$Y_i = \gamma X_i + \beta' Z_i + \sigma \varepsilon_i,$$

where $\gamma \in \mathbb{R}$, $\beta \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}^+$ are the parameters, and $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. errors with standard Gaussian distribution $N(0,1)$. We are interested in testing

$$\mathcal{H}_0 : \gamma/\sigma = \delta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \gamma/\sigma = \delta_1. \tag{2.22}$$

A test for whether $\gamma = 0$ is readily obtained by taking $\delta_0 = 0$. This problem is invariant under the action of the subgroup G of $\mathrm{AL}(d)$ that results from the restriction to the pairs (A, v) where $A = cI$, a multiple of the $d \times d$ identity matrix, and $v \in \mathbb{R}^d$ [Kariya, 1980, Eaton, 1989]. This group is amenable. On the observation space, G acts by $((cI, v), (X, Y, Z)) \mapsto (X, cY + v'Z, Z)$; on the parameter space, by $((cI, v), (\gamma, \beta, \sigma)) \mapsto (c\gamma, c\beta + v, c\sigma)$. A maximally invariant parameter is $\delta = \gamma/\sigma$. With this parametrization, the conditional density of Y becomes

$$p_{\delta,\beta,\sigma}(Y|X,Z) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - \beta' Z - \sigma\delta X)^2\right).$$

Define the vectors $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)'$ and $\boldsymbol{X}_n = (X_1, \ldots, X_n)'$, and the $n \times d$ matrix $Z^n = [Z_1, \ldots, Z_n]'$ whose rows are the vectors $Z_1, \ldots, Z_n$. Assume that Z has full rank. A maximally invariant function of the data is given by $M_n = \left(\frac{A^{n\prime}\boldsymbol{Y}_n}{\|A^{n\prime}\boldsymbol{Y}_n\|}, \boldsymbol{X}_n, Z_n\right)$, where A^n is a $(n-d) \times n$ matrix whose columns form an orthonormal basis for the orthogonal complement of the column space of Z^n. It follows that $A^{n\prime}A^n = I^{n-d}$ and $A^n A^{n\prime} = I^n - Z^n(Z^{n\prime}Z^n)^{-1}Z^{n\prime}$, where and I^n is the $n \times n$ identity matrix [Kariya, 1980, Bhowmik and King, 2007]. In order to compute the likelihood of the maximally invariant statistic M_n, we assume that the mechanism that generates $\boldsymbol{X}_n$ and Z^n is the same under both hypotheses. It only remains to compute the distribution of $\boldsymbol{U}_n = \frac{A^{n\prime}\boldsymbol{Y}_n}{\|A^{n\prime}\boldsymbol{Y}_n\|}$ conditionally on $\boldsymbol{X}_n$ and Z_n. Bhowmik and King [2007] show that for arbitrary effect size δ, the density of this distribution is given by

$$p_\delta^{\boldsymbol{U}_n}(u|\boldsymbol{X}_n, Z^n) = \frac{1}{2}\Gamma\left(\frac{k}{2}\right)\pi^{-\frac{k}{2}}e^{c(\delta)}\Bigg[{}_1F_1\left(\frac{k}{2}, \frac{1}{2}, \frac{a^2(u,\delta)}{2}\right) + \sqrt{2}a(u,\delta)\,\frac{\Gamma((1+k)/2)}{\Gamma(k/2)}\,{}_1F_1\left(\frac{1+k}{2}, \frac{3}{2}, \frac{a^2(u,\delta)}{2}\right)\Bigg],$$

where $k = n - d$, u is a unit vector in $\mathbb{R}^k$, a is the function $a(u,\delta) = \delta\boldsymbol{X}_n' A^n u, c(\delta) = -\frac{1}{2}\delta^2\boldsymbol{X}_n' A^n A^{n\prime}\boldsymbol{X}_n$, and ${}_1F_1$ is the confluent hypergeometric function. This can be used to compute the relatively GROW E-statistic in Theorem 2.4.2 for (2.22). In fact, they compute in more generality the density of the maximally invariant statistic when X is allowed to have a non-linear effect on Y. This does not impact the group invariance structure of the model, so that our results can also be used in this semilinear setting if the hypotheses are adjusted accordingly.

2.8. Composite invariant hypotheses

Until now we have considered null and alternative hypotheses that become simple when viewed through the lens of the maximally invariant statistic. As we saw, in the t-test this corresponds to testing simple hypotheses about the effect size δ. However, there are situations where it is desirable to contemplate hypotheses that are composite in the maximally invariant parameter. Later in this section, we will revisit the t-test, and view Hotelling's T^2 test through this lens in Section 4.7. We also consider problems in which a fixed prior is placed on the maximally invariant parameter δ, in Corollary 2.8.3, thereby implementing the *method of mixtures*, a standard method to combine test martingales going back to Wald [1945] and Darling and Robbins [1968a]. It was already used in the context of our t-test example by Lai [1976].

Consider, as in the previous section, Θ to be the parameter space on which G acts freely and continuously. Let δ be a maximally invariant parameter. Suppose that the parameter space Θ can be decomposed as $\Theta \cong G \times \Theta/G$. Consider the testing problem

$$\mathcal{H}_0 : X^n \sim \mathbf{P}_{g,\delta},\ \ \delta \in \Delta_0,\ g \in G \ \text{ vs. }\ \mathcal{H}_1 : X^n \sim \mathbf{Q}_{g,\delta},\ \ \delta \in \Delta_1,\ g \in G, \tag{2.23}$$

where Δ_0, Δ_1 are two sets of possible values of the maximally invariant parameter $\delta = \delta(\theta)$. Recall that the distribution of a maximally invariant function of the data $M_n = m_n(X^n)$ depends on the parameter θ only through δ. Consequently, the alternatives in the previous hypothesis testing problem are not simple when data are reduced through invariance. The main objective of this section is to show that when searching for a GROW E-statistic for (2.23) it is enough to do so for the invariance-reduced problem

$$\mathcal{H}_0 : M_n \sim \mathbf{P}_\delta^{M_n}, \quad \delta \in \Delta_0 \ \text{ vs. } \ \mathcal{H}_1 : M_n \sim \mathbf{Q}_\delta^{M_n}, \quad \delta \in \Delta_1. \tag{2.24}$$

We follow the same steps that we followed in Section 2.4, and begin by showing that if there exists a minimizer for the KL minimization problem associated to (2.24), then it has the same value as that associated to (2.23).

Proposition 2.8.1. *Assume that there exists a pair of probability distributions $\boldsymbol{\Pi}_0^\star, \boldsymbol{\Pi}_1^\star$ on Δ_0 and Δ_1 that satisfy*

$$\mathrm{KL}(\boldsymbol{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \boldsymbol{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}) = \min_{\boldsymbol{\Pi}_0,\boldsymbol{\Pi}_1} \mathrm{KL}(\boldsymbol{\Pi}_1^{\delta}\mathbf{Q}_\delta^{M_n}, \boldsymbol{\Pi}_0^{\delta}\mathbf{P}_\delta^{M_n}). \tag{2.25}$$

For each $g \in G$, define the probability distributions $\mathbf{P}_g^\star = \boldsymbol{\Pi}_0^{\star\delta}\mathbf{P}_{g,\delta}$ and $\mathbf{Q}_g = \boldsymbol{\Pi}_1^{\star\delta}\mathbf{Q}_{g,\delta}$ on $\mathcal{X}^n$. If the models $\{\mathbf{P}_g^\star\}_{g\in G}$ and $\{\mathbf{Q}_g^\star\}_{g\in G}$ satisfy the assumptions of Theorem 2.4.2, then

$$\inf_{\boldsymbol{\Pi}_0,\boldsymbol{\Pi}_1} \mathrm{KL}(\boldsymbol{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}, \boldsymbol{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) = \min_{\boldsymbol{\Pi}_0,\boldsymbol{\Pi}_1} \mathrm{KL}(\boldsymbol{\Pi}_1^{\delta}\mathbf{Q}_\delta^{M_n}, \boldsymbol{\Pi}_1^{\delta}\mathbf{P}_\delta^{M_n}).$$

Proof. Let $\boldsymbol{\Pi}_0^{g,\delta}, \boldsymbol{\Pi}_1^{g,\delta}$ be two probability distributions on $G \times \Delta_0$ and $G \times \Delta_1$, respectively. If we call $\boldsymbol{\Pi}_0^\delta$ and $\boldsymbol{\Pi}_1^\delta$ their respective marginals on Δ_0 and Δ_1, then, the information processing inequality implies that

$$\mathrm{KL}(\boldsymbol{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}, \boldsymbol{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) \geq \mathrm{KL}(\boldsymbol{\Pi}_1^{\delta}\mathbf{Q}_\delta^{M_n}, \boldsymbol{\Pi}_0^{\delta}\mathbf{P}_\delta^{M_n}) \geq \mathrm{KL}(\boldsymbol{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \boldsymbol{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}).$$

This means that the right-most member of the previous display is a lower bound on our target infimum, that is,

$$\inf_{\boldsymbol{\Pi}_0,\boldsymbol{\Pi}_1} \mathrm{KL}(\boldsymbol{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}\boldsymbol{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) \geq \mathrm{KL}(\boldsymbol{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \boldsymbol{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}). \tag{2.26}$$

To show that this is indeed an equality, it suffices to prove it when taking the infimum over a smaller subset of probability distributions $\boldsymbol{\Pi}_0, \boldsymbol{\Pi}_1$. We proceed to build such a subset. Let $\mathcal{P}(\boldsymbol{\Pi}_0^{\star\delta})$ be the set of probability distributions on $G \times \Delta_0$ with marginal distribution $\boldsymbol{\Pi}_0^{\star\delta}$. Define analogously the set of probability distributions $\mathcal{P}(\boldsymbol{\Pi}_1^{\star\delta})$ on $G \times \Delta_1$. By our assumptions, Theorem 2.4.2 can be readily used to conclude that

$$\inf_{(\boldsymbol{\Pi}_0,\boldsymbol{\Pi}_1)\in\mathcal{P}(\boldsymbol{\Pi}_0^{\star\delta})\times\mathcal{P}(\boldsymbol{\Pi}_1^{\star\delta})} \mathrm{KL}(\boldsymbol{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}, \boldsymbol{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) = \mathrm{KL}(\boldsymbol{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \boldsymbol{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}) \tag{2.27}$$

holds; (2.26) and (2.27) together imply the result that we were after. □

From the previous proposition, using Theorem 2.4.1 and the steps used for Corollaries 2.4.3 and 2.4.5, we can conclude that the ratio of the Bayes marginals for the invariance-reduced data M_n using the optimal priors $\boldsymbol{\Pi}_0^\star$ and $\boldsymbol{\Pi}_1^\star$ is a relatively GROW E-statistic for (2.23). We now state the corollary and apply it to to our running example, the t-test.

Corollary 2.8.2. *Under the assumptions of Proposition 2.8.1, the statistic given by*

$$T^{\star}(X^n) = \frac{\int q_{\delta}^{M_n}(m_n(X^n))\mathrm{d}\mathbf{\Pi}_1^{\star}(\delta)}{\int p_{\delta}^{M_n}(m_n(X^n))\mathrm{d}\mathbf{\Pi}_0^{\star}(\delta)}$$

is a GROW E*-statistic for (2.23).*

Example 2.1.1 (continued). Suppose, in the t-test setting, that we are now interested in testing

$$\mathcal{H}_0 : \delta \in (-\infty, \delta_0] \quad \text{vs.} \quad \mathcal{H}_1 : \delta \in [\delta_1, \infty) \tag{2.28}$$

for some $\delta_0 < \delta_1$, where, recall, $\delta = \mu/\sigma$ is the maximally invariant parameter. Corollary 2.8.2 shows that no loss is incurred if we only look among E-statistics that are a function of the maximally invariant function M_n, the t-statistic. Because the density of t-statistic is monotone in δ, we readily conclude that the minimum in (2.25) is achieved by the probability distributions $\mathbf{\Pi}_0^{\star}$ and $\mathbf{\Pi}_1^{\star}$ that put all of their mass on δ_0 and δ_1, respectively. Corollary 2.8.2 yields that $T_n^{\star} = p_{\delta_1}^{M_n}/p_{\delta_0}^{M_n}$ is GROW among all possible E-statistics of the original data (not only the scale-invariant ones). This result can be extended to other families with this type of monotonicity property.

A standard approach to deal with unknown parameter values, both with Bayesian statistics and with E-statistics, is to employ proper prior distributions on the unknown parameters. In our setting, we may want to use specific priors $\tilde{\mathbf{\Pi}}_0$ and $\tilde{\mathbf{\Pi}}_1$ on Δ_0 and Δ_1. If we define for each g the probability distributions $\tilde{\mathbf{P}}_g = \tilde{\mathbf{\Pi}}_0^{\delta}\mathbf{P}_{g,\delta}$ and $\tilde{\mathbf{Q}}_g = \tilde{\mathbf{\Pi}}_1^{\delta}\mathbf{Q}_{g,\delta}$, and the resulting models $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ and $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$ also satisfy the conditions of Corollary 2.4.3, the proof of Proposition 2.8.1 also shows the following corollary.

Corollary 2.8.3. *Let $\tilde{\mathbf{\Pi}}_0$ and $\tilde{\mathbf{\Pi}}_1$ be two probability distributions on Δ_0 and Δ_1, respectively. Let $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ and $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$ be two probability models defined by $\tilde{\mathbf{P}}_g = \tilde{\mathbf{\Pi}}_0^{\delta}\mathbf{P}_{g,\delta}$ and $\tilde{\mathbf{Q}}_g = \tilde{\mathbf{\Pi}}_1^{\delta}\mathbf{Q}_{g,\delta}$. If $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ and $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$ satisfy the conditions of Corollary 2.4.3 (or more precisely, the conditions of Theorem 2.4.2 with $\tilde{\mathbf{P}}_g$ in the role of $\mathbf{P}_g$ and $\tilde{\mathbf{Q}}_g$ in the role of $\mathbf{Q}_g$), then*

$$\tilde{T}_n(X^n) = \frac{\int q_{\delta}(m_n(X^n))\mathrm{d}\tilde{\mathbf{\Pi}}_1(\delta)}{\int p_{\delta}(m_n(X^n))\mathrm{d}\tilde{\mathbf{\Pi}}_0(\delta)} \tag{2.29}$$

is a (relatively) GROW E*-statistic for testing $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ against $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$.*

Example 2.1.1 (continued). Jeffreys and Jeffreys [1998] proposed a Bayesian version of the t-test based on Bayes factor (2.13), setting δ_0 to 0 and putting a Cauchy prior on δ_1 centered at 0. Popularized as the *Bayesian t-test* [Rouder et al., 2009], it is an instance of (2.29) with $\tilde{\Pi}_1$ set to aforementioned Cauchy prior and $\tilde{\Pi}_0$ putting mass 1 on $\delta_0 = 1$. It is itself an E-statistic (see GHK), but if we check the conditions of Theorem 2.4.2, we see that condition (2.15) does not hold, due to the Cauchy distribution not having any moments. Thus, we cannot verify whether (2.29) has the relative GROW property. However, as soon as we replace the Cauchy prior by any prior centered at 0 for which, for some $\epsilon > 0$, the $2 + \epsilon$th moment exists (such as e.g. a normal distribution centered at 0, as has also been proposed for this problem), we can use Lemma 2.7.1 (applied with dimension $d = 1$) to infer that assumption (2.15) holds. Proposition 2.8.3 can then be applied after all to conclude that the corresponding Bayes factor does have the relative GROW property.

2.9. Discussion, Related and Future Work

In this concluding section we bring up an issue that deserves further discussion and may inspire future work. It also highlights the differences between our work and related work in a Bayesian and information-theoretic context.

2.9.1. Amenability is not always necessary

We have shown that if a hypothesis testing problem is invariant under a group G and our assumptions are satisfied, then amenability of G is a sufficient condition for the likelihood ratio of the maximal invariant to be GROW. A natural question is whether amenability is also a necessary condition for the latter to hold. Not only is this of theoretical relevance: some groups that are important for statistical practice are not amenable. For instance, $\mathrm{GL}(d)$, the relevant group in Hotelling's test, is nonamenable. The setup of this test is similar to that in Section 2.7.1, except that the hypotheses are given by

$$\mathcal{H}_0 : \|\Lambda^{-1}\mu\|^2 = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \|\Lambda^{-1}\mu\|^2 = \gamma. \tag{2.30}$$

A maximally invariant statistic is the T^2-statistic $n\bar{X}_n'\bar{V}_n^{-1}\bar{X}_n$, where, as in Section 2.7.1, $\bar{X}_n$ and $\bar{V}_n$ are the unbiased estimators of the mean and the covariance matrix, respectively. Notice that this test is equivalent to (2.19) with the alternative expanded to $\Delta = \{\delta : \|\delta\|^2 = \gamma\}$, but that T^2 is not a maximal invariant under the lower triangular group. However, Giri et al. [1963] have shown that for $d = 2$ and $n = 3$, the likelihood ratio of the T^2-statistic can be written as an integral over the likelihood ratio in (2.21) with a proper prior on $\delta \in \Delta$ as defined there. It follows as a result of Proposition 2.8.1 that the likelihood ratio of the T^2-statistics is also GROW in the case that $d = 2$ and $n = 3$. These results can be extended to the case that $d = 2$ with arbitrary n by the work of Shalaevskii [1971]. As future work, it may be interesting to investigate whether amenability can be more generally replaced by a weaker condition, and/or whether a counterexample to Theorem 2.4.2 for nonamenable groups can be given.

2.9.2. Comparison to Sun and Berger [2007] and Liang and Barron [2004]: two families vs. one

As the above example illustrates, it is sometimes possible to represent the same $\mathcal{H}_0$ and $\mathcal{H}_1$ via (at least) two different groups, say G_a and G_b. Group G_a is combined with parameter of interest in some space Δ_a and priors $\mathbf{\Pi}_j^{*\delta_a}$ on Δ_a achieving (2.25) relative to group G_a, for $j = 0, 1$; group G_b has parameter of interest in Δ_b and priors $\mathbf{\Pi}_j^{*\delta_b}$ achieving (2.25) relative to group G_b; yet the tuples $\mathcal{T}_a = (G_a, \Delta_a, \{\mathbf{\Pi}_j^{*\delta_a}\}_{j=0,1})$

and $\mathcal{T}_b = (G_b, \Delta_b, \{\mathbf{\Pi}_j^{*\delta_b}\}_{j=0,1})$ define the same hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. That is, the set of distributions $\{\mathbf{P}_g^*\}_{g\in G_a}$ obtained by applying Proposition 2.8.1 with group G_a (representing $\mathcal{H}_0$ defined relative to group G_a) coincides with the set of distributions $\{\mathbf{P}_g^*\}_{g\in G_b}$ obtained by applying Proposition 2.8.1 with group G_b (representing $\mathcal{H}_0$ defined relative to group G_b); and analogously for the set of distributions $\{\mathbf{P}_g^*\}_{g\in G_a}$ and the set of distributions $\{\mathbf{P}_g^*\}_{g\in G_b}$. In the example above, G_a was $\mathrm{GL}(d)$ and the priors $\mathbf{\Pi}_0^{*\delta_a}, \mathbf{\Pi}_1^{*\delta_a}$ were degenerate priors on 0 and γ as in (2.30), respectively; G_b was the lower triangular group with a specific prior as indicated above. In such a case with multiple representations of the same $\mathcal{H}_0$ and $\mathcal{H}_1$, using the fact that the notion of 'GROW' does not refer to the underlying group, Corollary 2.8.2 can be used to identify the GROW E-statistic as soon as the assumptions of Proposition 2.8.1 hold for at least one of the tuples $\mathcal{T}_a$ or $\mathcal{T}_b$. Namely, if the assumptions hold for just one of the two tuples, we use Corollary 2.8.2 with that tuple; then $T^*(X^n)$ as defined in the corollary must be GROW, irrespective of whether $T^*(X^n)$ based on the other tuple is the same (as it was in the example above) or different. If the assumptions hold for *both* groups, then, using the fact that the GROW E-statistic is 'essentially' unique (see Theorem 1 of GHK for definition and proof), it follows that $T^*(X^n)$ as defined in Corollary 2.8.2 must coincide for both tuples.

Superficially, this may seem to contradict Sun and Berger [2007] who point out that in some settings, the right Haar prior is not uniquely defined, and different choices for right Haar prior give different posteriors. To resolve the paradox, note that, whereas we always formulate two models $\mathcal{H}_0$ and $\mathcal{H}_1$, Sun and Berger [2007] start with a single probabilistic model, say $\mathcal{P}$, that can be written as in (2.1) for some group G. Their example shows that the same $\mathcal{P}$ can sometimes arise from two different groups, and then it is not clear what group, and hence what Haar prior to pick, and their quantity of interest—the Bayesian posterior, i.e. a ratio between Bayes marginals for the same model $\mathcal{P}$ at different sample sizes n and $n-1$—can depend on the choice. In contrast, our quantity of interest, the GROW e-statistic $T_n^*(X^n)$, a ratio between Bayes marginals for different models $\mathcal{H}_0$ and $\mathcal{H}_1$ at the same sample size, is uniquely defined as soon as there exists one group G with $\mathcal{H}_0$ and $\mathcal{H}_1$ as in (2.2) for which the assumptions of Theorem 2.4.2 hold; or more generally, as soon as there exists one tuple $\mathcal{T} = (G, \Delta, \{\mathbf{\Pi}_j^{*\delta}\}_{j=0,1})$ for which the assumptions of Proposition 2.8.1 hold, even if there exist other such tuples.

The consideration of two families $\mathcal{H}_0$ and $\mathcal{H}_1$ vs. a single $\mathcal{P}$ is also one of the main differences between our setting and the one of Liang and Barron [2004], who provide exact min-max procedures for predictive density estimation for general location and scale families under Kullback-Leibler loss. Their results apply to any invariant probabilistic model $\mathcal{P}$ as in (2.1) where the invariance is with respect to location or scale (and more generally, with respect to some other groups including the subset of the affine group that we consider in Section 2.7.2). Consider then such a $\mathcal{P}$ and let $p^{M_n}(m_n(X^n))$ be as in (2.12). As is well-known, provided that n' is larger than some minimum value, for all $n > n'$, $r(X_{n'+1}, \ldots, X_n \mid X_1, \ldots, X_{n'}) := p^{M_n}(m_n(X^n))/p^{M_{n'}}(m_{n'}(X^{n'}))$ defines a conditional probability density for $X_{n'+1}, \ldots, X_n$; this is a consequence of the formal-Bayes posterior corresponding to the right Haar prior becoming proper after

n' observations, a.s. under all $\mathbf{P} \in \mathcal{P}$. For example, in the t-test setting, $n' = 1$. Liang and Barron [2004] show that the distribution corresponding to r minimizes the $\mathbf{P}^{n'}$-expected KL divergence to the conditional distribution $\mathbf{P}^n \mid X^{n'}$, in the worst case over all $\mathbf{P} \in \mathcal{P}$. Even though their optimal density r is defined in terms of the same quantities as our optimal statistic T_n^*, it is, just as Berger and Sun [2008], considered above, a ratio between likelihoods for the same model at different sample sizes, rather than, as in our setting, between likelihoods for different models, both composite, at the same sample sizes. Our setting requires a joint KL minimization over two families, and therefore our proof techniques turn out quite different from their information- and decision-theoretic ones.

2.10. Proof of the main theorem, Theorem 2.4.2

For the proof of the main result, we use an equivalent definition of amenability to the one that was already anticipated in Section 2.2.2. We take the one that suits our purposes best [see Bondar and Milnes, 1981, p. 109, Condition A_1].

Assumption 2 (Amenability of G). There exists a increasing sequence of symmetric compact subsets $C_1 \subseteq C_2, \dots \subset G$ such that, for any compact set $K \subseteq G$,

$$\frac{\rho^h\{h \in C_i\}}{\rho^h\{h \in C_i K\}} \to 1$$

as $i \to \infty$.

In this formulation, amenability is the existence of *almost invariant* symmetric compact subsets of the group G. We use these sets to build a sequence of *almost invariant* probability measures when G is noncompact.

Proof of Theorem 2.4.2. Under our assumptions, Theorem 2 of Bondar [1976] implies the existence of a bimeasurable one-to-one map $\mathcal{X}^n \to G \times \mathcal{X}^n/G$ such that $r(x^n) = (h(x^n), m(x^n))$ and $r(gx^n) = (gh(x^n), m(x^n))$ for $h(x^n) \in G$ and $m(x^n) \in \mathcal{X}^n/G$ (see Remark 2.3.2). Hence, by a change of variables, we can assume that the densities are with respect to the image measure μ under r on $G \times \mathcal{X}^n/G$. Call the random variables $M = m(X^n)$ and $H = h(X^n)$. We can therefore assume, without loss of generality, that the data is of the form (H, M), that the group G acts canonically by multiplication on the first component, and that the measures are with respect to a G-invariant measure $\nu = \lambda \times \beta$ where λ is the Haar measure on G and β is some measure on $\mathcal{X}^n/G$ (see Remark 2.3.4). For each $g \in G$, write $\mathbf{P}_g^{H|m}$ and $\mathbf{Q}_g^{H|m}$ for the conditional probabilities $\mathbf{P}_g^H[\,\cdot\,|M = m]$ and $\mathbf{Q}_g^H[\,\cdot\,|M = m]$, which can be obtained through disintegration [see Chang and Pollard, 1997], and write $p_g(\,\cdot\,|m)$ and $q_g(\,\cdot\,|m)$ for their respective conditional densities with respect to the left Haar measure λ. Recall, we write $\mathbf{P}_1$ and $\mathbf{Q}_1$ where 1 is the unit element of the group G.

We turn to our KL minimization objective. The chain rule for the KL divergence implies that, for any probability distribution $\mathbf{\Pi}$ on G,

$$\mathrm{KL}(\mathbf{\Pi}^g\mathbf{Q}_g, \mathbf{\Pi}^g\mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}^M, \mathbf{P}^M) + \int \mathrm{KL}(\mathbf{\Pi}^g\mathbf{Q}_g^{H|m}, \mathbf{\Pi}^g\mathbf{P}_g^{H|m})\mathrm{d}\mathbf{Q}(m). \tag{2.31}$$

In order to prove our claim, we will build a sequence $\{\mathbf{\Pi}_i\}_{i\in\mathbb{N}}$ of probability distributions on G such that the term in (2.31) pertaining the conditional distributions given M—the second term on the right hand side—goes to zero, that is, such that

$$\int \mathrm{KL}(\mathbf{\Pi}_i^g \mathbf{Q}_g^{H|m}, \mathbf{\Pi}_i^g \mathbf{P}_g^{H|m}) \mathrm{d}\mathbf{Q}(m) \to 0 \quad \text{as} \quad i \to \infty. \tag{2.32}$$

We define the distributions $\mathbf{\Pi}_i$ as the normalized restriction of the right Haar measure ρ to carefully chosen compact sets $C_i \subset G$, that we describe in brief. In other words, for $B \subseteq G$ measurable, we define $\mathbf{\Pi}_i$ by

$$\mathbf{\Pi}_i^g\{g \in B\} := \frac{\rho^g\{g \in B \cap C_i\}}{\rho^g\{g \in C_i\}}, \tag{2.33}$$

Next, the choice of compact sets C_i. For technical reasons that will become apparent later, we pick $C_i = J_i K_i L_i$, where J_i, K_i, and L_i are increasing compact symmetric neighborhoods of the unity of G with the growth condition that J_i is not much bigger—measured by ρ–than C_i. More precisely, we choose C_i according to the following lemma.

Lemma 2.10.1. Under the amenability of G there exist sequences $\{J_i\}_{i\in\mathbb{N}}$, $\{K_i\}_{i\in\mathbb{N}}$ and $\{L_i\}_{i\in\mathbb{N}}$ of compact symmetric neighborhoods of the unity of G, each increasing to cover G, such that

$$\frac{\rho^h\{h \in J_i\}}{\rho^h\{h \in J_i K_i L_i\}} \to 1$$

as $i \to \infty$.

There is no risk of dividing by ∞ in (2.33): by the continuity of the group operation each C_i is compact, hence $\rho\{C_i\} < \infty$. Lemma 2.10.1 ensures that $\mathbf{\Pi}_i^g\{g \in J_i\} \to 1$ as $i \to \infty$, a fact that will be useful later in the proof. Write $\mathbf{Q}_i^{H|m} := \mathbf{\Pi}_i^g \mathbf{Q}_g^{H|m}$, and $\mathbf{P}_i^{H|m} := \mathbf{\Pi}_i^g \mathbf{P}_g^{H|m}$, and $q_i(h|m)$ and $p_i(h|m)$ for their respective densities. We use a change of variable and split the integral in our quantity of interest from (2.32). To this end, notice that for any function $f = f(h,m)$, the expted value $\mathbf{E}_g^{\mathbf{Q}}[f(H,M)] = \mathbf{E}_1^{\mathbf{Q}}[f(gH,M)]$. Indeed,

$$\begin{aligned}\int f(h,m) q_g(h,m) \mathrm{d}\lambda(g) \mathrm{d}\beta(m) &= \int f(h,m) q_1(g^{-1}h,m) \mathrm{d}\lambda(g) \mathrm{d}\beta(m) \\ &= \int f(gh,m) q_1(h,m) \mathrm{d}\lambda(g) \mathrm{d}\beta(m).\end{aligned}$$

Use this fact to obtain that

$$\begin{aligned}\int \mathrm{KL}(\mathbf{\Pi}_i^g \mathbf{Q}_g^{H|m}, \mathbf{\Pi}_i^g \mathbf{P}_g^{H|m})\mathrm{d}\mathbf{Q}(m) &= \int \mathbf{E}_1^{\mathbf{Q}}\left[\ln \frac{q_i(gH|M)}{p_i(gH|M)}\right]\mathrm{d}\mathbf{\Pi}_i(g) \\ &= \underbrace{\int \mathbf{E}_1^{\mathbf{Q}}\left[\mathbf{1}\{gH \in J_iK_i\} \ln \frac{q_i(gH|M)}{p_i(gH|M)}\right]\mathrm{d}\mathbf{\Pi}_i(g)}_{\mathrm{A}} + \\ &\quad \underbrace{\int \mathbf{E}_1^{\mathbf{Q}}\left[\mathbf{1}\{gH \notin J_iK_i\} \ln \frac{q_i(gH|M)}{p_i(gH|M)}\right]\mathrm{d}\mathbf{\Pi}_i(g)}_{\mathrm{B}}.\end{aligned} \tag{2.34}$$

We separate the rest of the proof in two steps, one for bounding each term in (2.34). These steps use two technical lemmas whose proof we give after showing how they help at achieving our goals.

Bound for A in (2.34)**:** Recall that that

$$\ln \frac{q_i(gh|m)}{p_i(gh|m)} = \ln \frac{\int \mathbf{1}\{g' \in J_iK_iL_i\}\, q_{g'}(gh|m)\mathrm{d}\rho(g')}{\int \mathbf{1}\{g' \in J_iK_iL_i\}\, p_{g'}(gh|m)\mathrm{d}\rho(g')}$$

Use $N = J_iK_i$—not necessarily symmetric—and $L = L_i$ in the following lemma.

Lemma 2.10.2. Let N and L be compact subsets of G. Assume that L is symmetric. Then, for each $m \in \mathcal{M}$ it holds that

$$\sup_{h\in N} \ln \frac{\int \mathbf{1}\{g \in NL\}\ \ q_g(h|m)\mathrm{d}\rho(g)}{\int \mathbf{1}\{g \in NL\}\ \ p_g(h|m)\mathrm{d}\rho(g)} \le -\ln \mathbf{P}_1\{H \in L \mid M = m\}.$$

With this lemma at hand, conclude that, for all $gh \in J_iK_i$, and $m \in \mathcal{M}$

$$\ln \frac{q_i(gh|m)}{p_i(gh|m)} \le -\ln \mathbf{P}_1\{H \in L_i \mid M = m\}.$$

At the same time this implies that A in (2.34) is smaller than

$$-\int \ln \mathbf{P}_1\{H \in L_i \mid M = m\}\mathrm{d}\mathbf{Q}(m).$$

Since the sets L_i were chosen to satisfy $L_i \uparrow G$, the probability $\mathbf{P}\{H \in L_i \mid M = m\} \to 1$ monotonically for each value of m. Consequently the quantity in last display tends to 0 by the monotone convergence theorem, and so does A in (2.34). This ends the first step of the proof. Now, we turn to the second term in (2.34).

Bound for B in (2.34)**:** Our strategy at this point is to show that, as $i \to \infty$,

$$\int \mathbf{Q}_1^h\{gh \notin J_iK_i\}\,\mathrm{d}\mathbf{\Pi}_i(g) \to 0, \tag{2.35}$$

and to use (2.15) to show our goal, that B in (2.34) tends to zero. To show (2.35), notice that if $g \in J_i$ and $h \in K_i$, then $gh \in J_iK_i$, which implies that

$$\int \mathbf{Q}_1 \left\{ gH \in J_iK_i \right\} \mathrm{d}\mathbf{\Pi}_i(g) \geq \mathbf{\Pi}_i^g \left\{ g \in J_i \right\} \mathbf{Q}_1 \left\{ H \in K_i \right\}.$$

Since the sets K_i increase to cover G, we have $\mathbf{Q}\left\{H \in K_i\right\} \to 1$ as $i \to \infty$, and by our initial choice of sets J_i, K_i, L_i, the probability $\mathbf{\Pi}_i^g \left\{g \in J_i\right\} \to 1$, as $i \to \infty$. Hence (2.35) holds.

To bound the second term, we use the following lemma with $\mathbf{\Pi} = \mathbf{\Pi}_i$.

Lemma 2.10.3. Let $\mathbf{\Pi}$ be a distribution on G. Then, for each $h \in G$ and $m \in \mathcal{M}$, it holds that

$$\ln \frac{\int q_g(h|m) \mathrm{d}\mathbf{\Pi}(g)}{\int p_g(h|m) \mathrm{d}\mathbf{\Pi}(g)} \leq \int \ln \frac{q_g(h|m)}{p_g(h|m)} \mathrm{d}\mathbf{\Pi}(g|h,m).$$

where $\mathrm{d}\mathbf{\Pi}(g|h,m) = \frac{q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}{\int_g q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}$.

After invoking the previous lemma, apply Hölder's and Jensen's inequality consecutively to bound B in (2.34) by

$$\begin{aligned}
&\iint \left[\mathbf{1}\left\{ gh \notin J_iK_i \right\} \int \left[\ln \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)} \right] \mathrm{d}\mathbf{\Pi}_i(g'|h,m) \right] \mathrm{d}\mathbf{Q}_1(h,m)\mathrm{d}\mathbf{\Pi}_i(g) \\
&\leq \underbrace{\left(\int \mathbf{Q}_1 \left\{ gH \notin J_iK_i \right\} \mathrm{d}\mathbf{\Pi}_i(g) \right)^{1/q}}_{\to 0 \text{ as } i \to \infty \text{ by (2.35)}} \times \\
&\left(\iint \left| \int \left[\ln \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)} \right] \mathrm{d}\mathbf{\Pi}_i(g'|h,m) \right|^p \mathrm{d}\mathbf{Q}_1(h,m)\mathrm{d}\mathbf{\Pi}_i(g) \right)^{1/p}
\end{aligned} \tag{2.36}$$

where $p = 1 + \varepsilon$ and q is p's Hölder conjugate, that is, $1/p + 1/q = 1$. Next, we show that the second factor in (2.36) remains bounded as $i \to \infty$. By Jensen's inequality, this quantity is smaller than

$$\left(\iiint \left| \ln \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)} \right|^p \mathrm{d}\mathbf{\Pi}_i(g'|h,m)\mathrm{d}\mathbf{Q}_1(h,m)\mathrm{d}\mathbf{\Pi}_i(g) \right)^{1/p}.$$

After a series of rewritings and using our Assumption (2.15), we will show that this quantity is bounded. Now, we use again the change of variable that we used to obtain (2.34)—but now in the opposite direction—to deduce that

$$\begin{aligned}
&\iiint \left| \ln \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)} \right|^p \mathrm{d}\mathbf{\Pi}_i(g'|h,m)\mathrm{d}\mathbf{Q}_1(h,m)\mathrm{d}\mathbf{\Pi}_i(g) = \\
&\iiint \left| \ln \frac{q_{g'}(h|m)}{p_{g'}(h|m)} \right|^p \mathrm{d}\mathbf{\Pi}_i(g'|h,m)\mathrm{d}\mathbf{Q}_g(h,m)\mathrm{d}\mathbf{\Pi}_i(g) = \\
&\iint \left| \ln \frac{q_{g'}(h|m)}{p_{g'}(h|m)} \right|^p \mathrm{d}\mathbf{\Pi}_i(g'|h,m)\mathrm{d}\mathbf{Q}_i(h,m).
\end{aligned}$$

At this point, Bayes' theorem implies that this last quantity is equal to

$$\iint \left|\ln \frac{q_{g'}(h|m)}{p_{g'}(h|m)}\right|^p \mathrm{d}\mathbf{Q}_{g'}(h,m)\mathrm{d}\mathbf{\Pi}_i(g') = \mathbf{E}_1^{\mathbf{Q}} \left|\ln \frac{q_1(H|M)}{p_1(H|M)}\right|^p$$

Hence, as

$$\left(\mathbf{E}_1^{\mathbf{Q}}\left[\left|\ln \frac{q(H|M)}{p(H|M)}\right|^p\right]\right)^{1/p} \leq$$
$$\left(\mathbf{E}_1^{\mathbf{Q}}\left[\left|\ln \frac{q(H,M)}{p(H,M)}\right|^p\right]\right)^{1/p} + \left(\mathbf{E}^{\mathbf{Q}}\left[\left|\ln \frac{q(M)}{p(M)}\right|\right]^p\right)^{1/p} < \infty$$

by (2.15). We have shown that (2.36) tends to 0 as $i \to \infty$ and that consequently B in (2.34) tends to 0 in the same limit.

After completing these two steps, we have shown that both A and B in (2.34) tend to 0 as $i \to \infty$, and that consequently the claim of the theorem follows. All is left is to prove lemmas 2.10.1, 2.10.2, and 2.10.3. □

2.10.1. Proof of technical lemmas 2.10.1, 2.10.2, and 2.10.3

Proof of Lemma 2.10.1. Let $\{\varepsilon_i\}_i$ be a sequence of positive numbers decreasing to zero. Let $\{K_i\}_{i\in\mathbb{N}}$ and $\{L_i\}_{i\in\mathbb{N}}$ be two arbitrary sequences of compact symmetric subsets that increase to cover G. Fix $i \in \mathbb{N}$. The set K_iL_i is compact and by our assumption there exists a sequence $\{J_l\}_{l\in\mathbb{N}}$ and such that $\rho\{J_l\}/\rho\{J_lK_iL_i\} \to 0$ as $l \to \infty$. Pick $l(i)$ to be such that $\rho\{J_{l(i)}\}/\rho\{J_{l(i)}K_iL_i\} \geq 1-\varepsilon_i$. The claim follows from a relabeling of the sequences. □

Proof of Lemma 2.10.2. Let $h \in N$. Then we can write

$$\begin{aligned}\int \mathbf{1}\{g \in NL\}\ q_g(h|m)\mathrm{d}\rho(g) &= \int \mathbf{1}\{g \in NL\}\ q_1(g^{-1}h|m)\mathrm{d}\rho(g)\\ &= \int \mathbf{1}\left\{g \in (NL)^{-1}\right\}\ q_1(gh|m)\mathrm{d}\lambda(g)\\ &= \Delta(h^{-1})\int \mathbf{1}\left\{g \in (NL)^{-1}h\right\}\ q_1(g|m)\mathrm{d}\lambda(g)\\ &= \Delta(h^{-1})\mathbf{Q}_1\{H \in (NL)^{-1}h \mid M = m\}\end{aligned}$$

The same computation can be carried out for p. Consequently

$$\begin{aligned}\ln \frac{\int \mathbf{1}\{g \in NL\}\ q_g(h|m)\mathrm{d}\rho(g)}{\int \mathbf{1}\{g \in NL\}\ p_g(h|m)\mathrm{d}\rho(g)} &= \ln \frac{\mathbf{Q}_1\{H \in (NL)^{-1}h \mid M = m\}}{\mathbf{P}_1\{H \in (NL)^{-1}h \mid M = m\}}\\ &\leq -\ln \mathbf{P}_1\{H \in (NL)^{-1}h \mid M = m\}.\end{aligned}$$

By our assumption that $h \in N$, we have that $(NL)^{-1}h = L^{-1}N^{-1}h \supseteq L^{-1} = L$. This implies that the last quantity of the previous display is smaller than $-\ln \mathbf{P}_1\{H \in L \mid M = m\}$. The result follows. □

Proof of Lemma 2.10.3. The result follows from a rewriting and an application of Jensen's inequality. Indeed,

$$
\begin{aligned}
-\ln \frac{\int p_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}{\int q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)} &= -\ln \frac{\int q_g(h|m)\frac{p_g(h|m)}{q_g(h|m)}\mathrm{d}\mathbf{\Pi}(g)}{\int q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)} \\
&= -\ln \int \left[\frac{p_g(h|m)}{q_{g'}(h|m)}\right]\mathrm{d}\mathbf{\Pi}(g|h,m) \\
&\leq -\int \ln \frac{p_g(h|m)}{q_g(h|m)}\mathrm{d}\mathbf{\Pi}(g|h,m) \\
&= \int \ln \frac{q_g(h|m)}{p_g(h|m)}\mathrm{d}\mathbf{\Pi}(g|h,m),
\end{aligned}
$$

as it was to be shown. □

2.11. Acknowledgements

Peter D. Grünwald is also affiliated with the Mathematical Institute of Leiden University. This work is part of the research program with project number 617.001.651, which is financed by the Dutch Research Council (NWO). We thank Wouter Koolen for useful conversations.

3. The Anytime-Valid Logrank Test: Error Control Under Continuous Monitoring with Unlimited Horizon[1]

We introduce the anytime-valid (AV) logrank test, a version of the logrank test that provides type-I error guarantees under optional stopping and optional continuation. The test is sequential without the need to specify a maximum sample size or stopping rule, and allows for cumulative meta-analysis with type-I error control. The method can be extended to define anytime-valid confidence intervals. The logrank test is an instance of the martingale tests based on E-variables that have been recently developed. We demonstrate type-I error guarantees for the test in a semi-parametric setting of proportional hazards and show how to extend it to ties, Cox' regression and confidence sequences. Using a Gaussian approximation on the logrank statistic, we show that the AV logrank test (which itself is always exact) has a similar rejection region to O'Brien-Fleming α-spending but with the potential to achieve 100% power by optional continuation. Although our approach to *study design* requires a larger sample size, the *expected* sample size is competitive by optional stopping.

3.1. Introduction

The logrank test is arguably the most important tool for the statistical comparison of time-to-event data between two groups of participants. Our main focus is when the two groups refer to the treatment and control groups in a randomized controlled trial; the outcome of interest are event times, that is, the time elapsed until an outcome of interest. The logrank test, in turn, uses a simplified version of the proportional hazard ratio model of Cox [1972]. For a fixed sample size and under this model, Cox gave a simple but profound insight: inference can be performed using the partial likelihood of having observed the events in the particular order that they were observed. To this end, the logrank test [Mantel, 1966, Peto and Peto, 1972], the score test associated to the Cox' partial likelihood, is optimal for fixed sample size and a restricted alternative. Large-sample properties of the logrank test are known in very general settings [Tsiatis, 1981, Schoenfeld, 1981, Andersen et al., 1993]. Nevertheless, even shortly after the

[1] This chapter is based on J. ter Schure, M. F. Pérez-Ortiz, A. Ly, and P. Grünwald. The Safe Logrank Test: Error Control under Continuous Monitoring with Unlimited Horizon, July 2021. URL `http://arxiv.org/abs/2011.06931`. arXiv:2011.06931 [math, stat], under submission

publication of the groundbreaking article of Cox, it became clear that the fixed-sample assumption can be overly restrictive. Indeed, due to ethical and practical constraints in human survival-time medical trials, interim analyses may be performed to terminate the study earlier than planned if needed. Consequently, it has been of fundamental importance to develop methods for the sequential analysis of time-to-event data in general; for the logrank test, in particular.

In order to legitimate the use of sequential boundary decisions, uniform asymptotic approximations over the study period have been developed for the logrank statistic [Tsiatis, 1982, Sellke and Siegmund, 1983, Slud, 1984]. The results in this line of work show the convergence of the sequentially computed logrank statistic to a rescaled Brownian motion under very general censoring and participant-arrival patterns. When interim analyses are only performed at discrete times, the decision boundaries based on continuously monitoring the logrank statistic are known to be overly conservative. This deficiency is addressed by group-sequential and α-spending methods, which, using knowledge of the interim analysis times relative to a predefined maximum number of events, allow for tighter decision boundaries [Pocock, 1977, O'Brien and Fleming, 1979, Kim and DeMets, 1987]. These sequential methods allow several interim looks at the data to stop for efficacy (if the treatment shows to be beneficial) or futility (if the study is no longer likely to reach statistical significance).

Despite the profound impact that these methods have had in statistical practice, the requirement of a maximum sample size limits the utility of a promising but non-significant study once the maximum sample size is reached. Because of their design, extending such a trial makes it impossible to control their type-I error. Moreover, the evidence gathered in new—possibly unplanned—trials cannot be added in a typical retrospective meta-analysis, when the number of trials or timing of the meta-analysis are dependent on the trial results. Such dependencies introduce accumulation bias and invalidate the assumptions of conventional statistical procedures in meta-analysis [ter Schure and Grünwald, 2019]. In order to address these deficiencies, we look for flexible anytime-valid methods that provide type-I error control in two situations: (1) optional stopping, which refers to halting the experiment earlier or later than planned under arbitrary stopping rules, and (2) meta-analysis and optional continuation, which refers to the aggregation of evidence of possibly interdependent studies. Just as the existing methods, our approach is connected to early work by H. Robbins and collaborators [Darling and Robbins, 1967, Lai, 1976]. Most notably, existing approaches come with fixed stopping rules, which are not desirable in the use cases that are of our present interest. The details of the present approach are very different, and to some extent, as we will see, more straightforward.

The main result of this work is the anytime-valid (AV) logrank test, an anytime-valid test for the statistical comparison of time-to-event data from two groups of participants. The AV logrank test uses the exact ratio of the sequentially computed Cox partial likelihood as test statistic. The advantage of having an exact test manifests, for instance, in the case of unbalanced allocation, when both control and treatment groups start with different numbers of participants. In this case, α-spending approaches do not provide strong type-I error guarantees due to the approximations involved [Wu and Xiong, 2017]. The basic version of the AV logrank test is, however, exact; unbalanced allocation presents no difficulties. Additionally, the AV logrank test can be used for meta-analysis and optional continuation while preserving the same type-I error guarantees.

From a technical point of view, we show, under general patterns of incomplete observations, that under the composite null hypothesis our test statistic is a continuous-time martingale with expected value equal to one. Statistics with this sequential property are referred to as test martingales; they form the basis of anytime-valid tests [Ramdas et al., 2020]. The AV logrank test is a concrete instance of such a test martingale derived from the recent theory of anytime-valid hypothesis testing based on E-processes [Henzi and Ziegel, 2021, Grünwald et al., 2020, Shafer, 2021, Wang and Ramdas, 2020]. In contrast to p-values, an analysis based on E-processes can extend existing trials as well as inform the decision to start new trials and meta-analyses, while still controlling type-I error rate. Type-I error control is retained even (i) if the E-process is monitored continuously and the trial is stopped early whenever the evidence is convincing, (ii) if the evidence of a promising trial is increased by extending the experiment and (iii) if a trial result spurs a new trial with the intention to combine them in a meta-analysis. Even with dependence between the trials, the test based on the multiplication of the values of these E-processes retains type-I error control, as long as all trials test the same (i.e., global) null hypothesis. This becomes especially interesting if we want to combine the results of several trials in a bottom-up retrospective meta-analysis, where no top-down stopping rule can be enforced. It is even possible to combine interim results of ongoing trials by multiplication, stepping beyond the realm of existing sequential approaches.

3.1.1. Contributions and outline

We begin with Section 3.2, where we review the special instance Cox' proportional hazards model for the two-group setting. There, we set the assumptions and notation used in the rest of the chapter. The definitions presented there are standard. In Section 3.3, we define and prove that the AV logrank test is indeed anytime valid. We first do this for (a) the case with only a group indicator (no other covariates) and without simultaneous events (ties). There, we also discuss its optimality properties and extend it to (b) the case with ties and to (c) the case when one wants to learn the actual effect size of the data and/or use prior knowledge about the effect size into the method via a Bayesian prior. This presents no technical difficulties. The resulting version of the test keeps providing nonasymptotic type-I error control even if the priors are wildly misspecified, that is, if they predict very different data from the data we actually observe. These results hinge on showing that the likelihood underlying Cox' proportional hazards model can be used to define E-variables and test martingales. In Section 3.4, we show a Gaussian approximation to the AV logrank statistic that is useful in the common situation when only summary statistics are available. We then provide extensive computer simulations to compare the AV logrank test to the classic logrank test and α-spending approaches. In Section 3.4.1, we show that the exact AV logrank test has a similar rejection region to O'Brien-Fleming α-spending for those designs and hazard ratios where it is well-approximated by a Gaussian AV logrank test. While always needing a small amount of extra data in the design phase (the price for indefinite optional continuation), the expected sample size needed for true rejections remain very competitive. During the design phase of a study, we might want to design for a maximum sample size in order to achieve a certain power, but need a smaller sample size on average during the study since we can safely engage in optional stopping. In Section 3.5, we show that AV-logrank-type tests can be combined through multiplication to perform meta-analysis, and in Section 3.6, we show how the test can be used to derive confidence sequences for the hazard ratio. In Section 3.7, we compare the sample sizes that are needed during the design phase in order to achieve a targetted power. Lastly, in Section 3.8 we make concluding remarks and discuss future research directions.

We remark that once the definitions are in place, the technical results are mostly straightforward consequences from earlier work; in particular, of the work of Cox [1975], Slud [1992] and Andersen et al. [1993]. The novelty of the present work is thus mainly in *defining* the AV logrank test and showing by computer simulation that, while being substantially more flexible, it is competitive with existing approaches—the classic logrank test with fixed design and in combination with α-spending.

Next to the main body of this chapter, we provide two appendices. We delegate to Appendix B.1 proofs and remarks that, while important, are not needed to follow the main development. Most importantly, the particular E-variable we design is *growth-rate optimal in the worst case*, GROW (see Section 3.3.1). Grünwald et al. [2020] provide several motivations for this criterion; we provide an additional one using an argument of Breiman [1961], which does not seem to be widely known. This argument shows a connection between growth-rate optimality and tests with minimal expected stopping time. In Appendix B.2, we provide an extension to the case when covariates other than group membership are present. This extension, based on the full Cox model, requires solving a challenging optimization problem and its implementation is therefore deferred to future work.

3.2. Proportional hazards model and Cox' partial likelihood

We begin by describing the hypothesis that is being tested, the data that are available, and Cox' proportional hazards model. We are interested in comparing the survival rates between two groups of participants, Group A and Group B. In a randomized controlled trial, Group A would signify the control group; Group B, the treatment group. We assume that the available data about m participants are of the form $\{(X^i, g^i, \delta^i) : i = 1, \ldots, m\}$, where $X^i = \min\{T^i, C^i\}$ is the minimum between the event time T^i and the (possibly infinite) censoring time C^i; g^i is a zero-one covariate depending on group membership ($g^i = 0$ signifies that $i \in A$; $g^i = 1$, that $i \in B$); and $\delta^i = \mathbf{1}\left\{X^i = T^i\right\}$ is the indicator of whether the event was witnessed before censoring or not. Let m^A be the number of members of Group A and m^B the number of members of Group B—then $m^A + m^B = m$. Define $\mathbf{g} = (g^1, \ldots, g^n)$, the vector of group memberships. We assume that $T^1, \dots, T^n, C^1, \dots, C^n$ are independent and have continuous distribution functions. The continuity assumption precludes tied observations; we relax this assumption later on, in Section 3.3.3. For $i = 1, \dots, m$, the survival rates are quantified by the hazard functions $\lambda^i = (\lambda^i_t)_{t \geq 0}$ for T_i, given by

$$\lambda^i_t = -\frac{\mathrm{d}}{\mathrm{d}t} \ln \mathbf{P}\{T^i \geq t\}. \tag{3.1}$$

As is customary, the hazard function λ^i at t can be interpreted via the conditional probability of witnessing an event in a short time span provided that the event has not been witnessed up to t, that is,

$$\mathbf{P}\{t \leq T^i < t + \Delta t \mid t \leq T^i\} = \lambda^i_t \Delta t + o(\Delta t) \quad \text{as} \quad \Delta t \to 0. \tag{3.2}$$

Given our interest in comparing the survival rates between the two groups, suppose that all participants i of Group A have a common hazard function $\lambda^i_t = \lambda^A_t$; members i of Group B, $\lambda^i_t = \lambda^B_t$. Using the data, we wish to test proportional hazards hypotheses. Concretely, we test the hypotheses $\mathcal{H}_0$ that the hazard function of the members of both groups satisfy $\lambda^A_t = \theta_0 \lambda^B_t$, against an alternative hypothesis $\mathcal{H}_1$ that $\lambda^B_t = \theta \lambda^A_t$ for a $\theta \neq \theta_0$. As a first application of the methods that we develop, we consider the statistical hypothesis testing problem between the null hypothesis that the hazard functions of the two groups are the same against the left-sided alternative, that is,

$$\begin{gathered} \mathcal{H}_0 : \lambda^B_t = \theta_0 \lambda^A_t \quad \text{vs.} \quad \mathcal{H}_1 : \lambda^B_t = \theta \lambda^A_t \\ \text{for some} \quad \theta \leq \theta_1 < \theta_0 \text{ and all } t, \end{gathered} \tag{3.3}$$

where θ is known as the hazard ratio and is the main quantity of statistical interest, and θ_1 would be, in a clinical trial, a minimal clinically relevant effect size. The alternative is what we hope for in case of negative events, such as death, with treatments that are set out to lower (relative to the control condition) the hazard rate. Notice that the hypotheses in (3.3) are, in fact, nonparametric. Similarly, if the event is positive, e.g., recovery from an infection, we would typically set a right-sided alternative, which can be also be treated with the present methods.

Right-sided, two-sided and the full alternative hypothesis $\mathcal{H}'_1 : \theta \neq 1$ are also amenable to the methods that will follow. We remark, however, that all the methods retain their type-I error guarantees irrespective of the specific alternative that we use. We now turn to defining Cox' partial likelihood PL_t, which is at the center of our approach. To that end, we need a battery of standard definitions—we lay them out to establish the notation. Let $y^i_t = \mathbf{1}\left\{X^i \geq t\right\}$ be the at-risk process, that is, the indicator of whether participant i is still at risk at time t, and let $\bar{y}^A_t = \sum_{i \in A} y^i_t$ and $\bar{y}^B_t = \sum_{i \in B} y^i_t$ be the number of participants at risk in each of the groups at time t. Define $\mathbf{y}_t = (y^1_t, \dots, y^m_t)$, the vector of at-risk processes, and $\mathcal{R}_t = \{j : y^j_t = 1\}$, the set of participants at risk at time t. Let $T^{(1)} < T^{(2)} < \dots < T^{(\bar{N}_\infty)}$ be the set of ordered events times that were witnessed (not censored). Note that, if all participants witness the event and censoring is absent, $\bar{N}_\infty = m$. For each $k = 1, \dots, \bar{N}_\infty$, let $I_{(k)}$ be the index of the individual that witnessed the event at time $T^{(k)}$. This means, for example, that if participant with label three was the fifth to witness the event, then $I_{(5)} = 3$. Abbreviate by $y^i_{(k)}, \bar{y}^A_{(k)}, \bar{y}^B_{(k)}, \mathcal{R}_{(k)}$ the corresponding quantities at event time $T^{(k)}$, and define $g^{(k)} := g^{I_{(k)}}$. Cox' partial likelihood $\mathrm{PL}_{\theta,t}$ can be sequentially computed by

$$\mathrm{PL}_{\theta,t} = \prod_{k:T^{(k)}\leq t} \frac{\theta^{g^{(k)}}}{\sum_{l\in\mathcal{R}_{T^{(k)}}} \theta^{g^l}} = \prod_{k:T^{(k)}\leq t} \frac{\theta^{g^{(k)}}}{\bar{y}^A_{(k)} + \theta\bar{y}^B_{(k)}}. \tag{3.4}$$

Cox' likelihood evaluated at the event times $T^{(1)}, T^{(2)}, \ldots$ coincides to that of a sequence of multinomial trials where, at event time $T^{(k)}$, each of the participants $i \in \mathcal{R}_{(k)}$ witnesses the event with probability

$$\begin{aligned} p_{\theta,(k)}(\ i\ ) &:= \mathbf{P}\{I_{(k)} = i \mid \mathbf{y}_{(l)}, \mathbf{g};\ l = 1, \ldots k\}, \\ p_{\theta,(k)}(\ i\ ) &= \frac{\theta^{g^i}}{\bar{y}^A_{(k)} + \theta\bar{y}^B_{(k)}}. \end{aligned} \tag{3.5}$$

Cox showed that, indeed, conditionally on all the information accrued strictly before $T^{(k)}$, the probability that participant i observes an event at time $T^{(k)}$ is exactly $p_{\theta,(k)}(\ i\ )$ as long as the hazard ratio is θ. With these likelihood computations at hand, we are in place to show the main contribution of this chapter, the AV logrank test, which uses the partial likelihood ratio as the test statistic.

3.3. The AV logrank test

In this section the AV logrank test for (3.3) is introduced; its type-I error guarantees and optimality properties are investigated. We give a solution to the first of the purposes laid down in the introduction: we show that the AV logrank test is anytime valid—its type-I error guarantees are not affected by optional stopping. The fact that it is also type-I-error-safe under optional continuation, our second purpose, is proven in Section 3.5. Without further ado, we define the AV logrank statistic $S^{\theta_1}_{\theta_0,t}$, typically, $\theta_0 = 1$, for (3.3) as the partial likelihood ratio

$$S^{\theta_1}_{\theta_0,t} = \frac{\mathrm{PL}_{\theta_1,t}}{\mathrm{PL}_{\theta_0,t}} = \prod_{k:T^{(k)}\leq t} \frac{p_{\theta_1,(k)}(I_{(k)})}{p_{\theta_0,(k)}(I_{(k)})}. \tag{3.6}$$

Here, $p_{\theta,(k)}$ is as defined in (3.5); the product that defines our statistic $S^{\theta_1}_{\theta_0,t}$ runs over the events that have been witnessed up to and including time t, and the empty product is taken to be equal to one. As is conventional with likelihood ratios, high values of $S^{\theta_1}_{\theta_0,t}$ are indicative that the alternative hypothesis is better than the null hypothesis at the describing the data. Given a tolerable type-I error bound α and an arbitrary random time τ, the AV logrank test is the test that rejects the null hypothesis if $S^{\theta_1}_{\theta_0\tau}$ is above the threshold $1/\alpha$, that is,

$$\xi^{\theta_1}_{\theta_0,\tau} = \mathbf{1}\left\{S^{\theta_1}_{\theta_0,\tau} \geq 1/\alpha\right\} := \begin{cases} 1 & \text{if } S^{\theta_1}_{\theta_0,\tau} \geq 1/\alpha \\ 0 & \text{otherwise.} \end{cases} \tag{3.7}$$

As we will see, by its sequential properties, $S^{\theta_1}_{\theta_0,t}$ takes large values with small probability under the null hypothesis uniformly over time, which translates into type-I error control for the test $\xi^{\theta_1}_{\theta_0,\tau}$. This observation is behind the any-time validity of the AV logrank test, and of anytime-valid tests in general (more details and general constructions to the effect of anytime-valid sequential testing can be found in the work of Ramdas et al. [2020]). We shown in the following proposition that the test $\xi^{\theta_1}_{\theta_0,\tau}$ has the desired type-I error control.

Proposition 3.3.1. *Let $\mathbf{P}_0$ be any distribution under which the hazard ratio is equal to θ_0, and let τ be any random time. The test $\xi^{\theta_1}_{\theta_0,\tau} = \mathbf{1}\left\{S^{\theta_1}_{\theta_0,\tau} \geq 1/\alpha\right\}$, where $S^{\theta_1}_{\theta_0,t}$ is as in (3.6), has level α, that is,*

$$\mathbf{P}_0\{\xi^{\theta_1}_{\theta_0,\tau} = 1\} \leq \alpha.$$

This result can be readily obtained using the sequential-multinomial interpretation of Cox' likelihood ratio. As we will see, in Section 3.3.1, this result can be interpreted in terms of E-variables and E-processes [Grünwald et al., 2020]. Define the process $(S^{\theta_1}_{\theta_0,(k)})_{k=1,2,\ldots}$ as the value of the AV logrank statistic at the event times $T^{(k)}$, that is, $S^{\theta_1}_{\theta_0,(k)} := S^{\theta_1}_{\theta_0,T^{(k)}}$. In this time discretization, the AV logrank statistic is the product of random variables

$$R^{\theta_1}_{\theta_0,(k)} = p_{\theta_1,(k)}(I_{(k)})/p_{\theta_0,(k)}(I_{(k)}), \tag{3.8}$$

the one-outcome partial likelihood ratio for the kth event, where $p_{\theta_0,(k)}$ is as in (3.5) and $k = 1, 2, \ldots$.

Proof of Proposition 3.3.1. Under any distribution under which the hazard ratio is θ_0, the fact that the likelihood of observing $I_{(k)}$ conditionally on $\{\mathbf{y}_{(l)} : l = 1, \ldots, k\}$ equals $p_{\theta_0,(k)}(I_{(K)})$ implies that

$$\mathbf{E}[R^{\theta_1}_{\theta_0,(k)} \mid \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}] = \sum_{j \in \mathcal{R}_{(k)}} p_{\theta_0,(k)}(j) \frac{p_{\theta_1,(k)}(j)}{p_{\theta_0,(k)}(j)} = 1. \tag{3.9}$$

This immediately shows that $S^{\theta_1}_{\theta_0,(k)} = \prod_{i \leq k} R^{\theta_1}_{\theta_0,(k)}$ is a test martingale, a nonnegative martingale with expected value equal to one, with respect to the filtration $\mathcal{F}_- = (\mathcal{F}_{(k)^-})_{k=1,2,\ldots}$ of sigma-algebras $\mathcal{F}_{(k)^-} = \sigma(\mathbf{y}_{(k)} : k = 1, \ldots, k)$. Next, the type-I error control for the the test $\xi^{\theta_1}_{\theta_0}$ follows from Ville's inequality, which asserts that, under the null hypothesis, the test martingale $S^{\theta_1}_{\theta_0,(k)}$ takes large values with small probability. Ville's inequality [Ville, 1939] (also known as Doob's maximal inequality) implies that

$$\mathbf{P}\left\{\sup_{k=1,2,\ldots} S^{\theta_1}_{\theta_0,(k)} \geq 1/\alpha\right\} \leq \mathbf{E}[S^{\theta_1}_{\theta_0,(1)}]\alpha = \alpha.$$

The previous display is a bound on ever making a type-I error when using the AV logrank test $\xi^{\theta_1}_{\theta_0,\tau}$. □

Under general patterns of incomplete observation—like independent censoring or independent left truncation—, the AV logrank test provides the same type-I error guarantees. To proof this, we give an alternative proof of Proposition 3.3.1 in Appendix B.1 using the counting-process formalism [Andersen et al., 1993]. There, we show that if the compensators of the underlying counting processes have a certain general product structure—which is the case under complete observation—, the AV logrank test is anytime-valid. We then refer to Andersen et al. [1993], who show that this structure is preserved under said patterns of incomplete observation.

The AV-logrank test is optimal—in a sense to be defined in the next section—among a large family of statistics. A second look at the proof of Proposition 3.3.1 suggests a generalization of the AV logrank statistic given in (3.6). Let, for each k, $q_{(k)}$ be a probability distribution on participants in the risk set $\mathcal{R}_{(k)}$ which is only allowed to depend on $\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(k)}$. Analogously to (3.8), we define the one-outcome ratio $R^{q}_{\theta_0,(k)} := q_{(k)}(I_{(k)})/p_{\theta_0,(k)}(I_{(k)})$—we now use $q_{(k)}$ instead of p_{θ_1}—, and

$$S^{q}_{\theta_0,t} := \prod_{k:T^{(k)} \leq t} R^{q}_{\theta_0,(k)} = \prod_{k:T^{(k)} \leq t} \frac{q_{(k)}(I_{(k)})}{p_{\theta_0,(k)}(I_{(k)})}. \tag{3.10}$$

A modification of the previous argument shows, for any random time τ, a type-I error guarantee for the test $\xi^{q}_{\theta_0,\tau}$ based on the value of $S^{q}_{\theta_0,\tau}$, that is, $\xi^{q}_{\theta_0,\tau} := \mathbf{1}\left\{S^{q}_{\theta_0,\tau} \geq 1/\alpha\right\}$ (see Proposition 3.3.1). Any such test is also anytime valid as long as each $q_{(k)}$ depends on the data only through $\mathbf{y}_{(1)}, \dots, \mathbf{y}_{(k)}$. In Section 3.3.2, we use this generalization to provide tests when no value of θ_1 is available. This generalization raises a natural question about the optimality of the AV logrank test based on (3.6) among test statistics of the form (3.10). This is the subject of the next section.

3.3.1. E-variables and optimality

The random variables $\{R^{\theta_1}_{\theta_0,(k)}\}_{k=1,2\dots}$ from (3.8) and $\{R^{q}_{\theta_0,(k)}\}_{k=1,2\dots}$ from (3.10) are examples of (conditional) E-variables—nonnegative random variables whose (conditional) expected value is below 1 uniformly over the null hypothesis. E-variables and E-processes are the “correct” generalization of likelihood ratios to the case that either or both $\mathcal{H}_0$ and $\mathcal{H}_1$ are composite and can be interpreted in terms of gambling [Grünwald et al., 2020, Shafer, 2021, Ramdas et al., 2020]. Under this gambling interpretation, a test martingale, a product of conditional E-variables, is the total profit made in a sequential gambling game where no earnings are expected under the null hypothesis. The analogy is thus between profit and evidence: no evidence can be gained against the null hypothesis if it is true. Just as p-values, the definition of E-variables and test martingales does not need any mention of an alternative hypothesis. However, if a composite set of alternative distributions is available, a gambler who is skeptical of the null distribution might want to maximize the speed of evidence accumulation (or of capital growth) under the alternative hypothesis. The worst-case growth rate is defined (conservatively) as the smallest expectation of the logarithm of the E-variable under the alternative. Consequently, any E-variable achieving it is called GROW, for

Growth-Rate Optimal in the Worst case (see the work of Grünwald et al. [2020] and Shafer [2021] for additional reasons to use this optimality criterion).

We instantiate this reasoning to our present problem. For the left-sided alternative (3.3), the choice $R^{\theta_1}_{\theta_0,(k)}$ is conditionally GROW because it maximizes the worst-case conditional growth rate

$$R^q_{\theta_0,(k)} \mapsto \min_{\theta\leq\theta_1} \mathbf{E}_\theta[\ln R^q_{\theta_0,(k)} | \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}],$$

over all valid choices of $q_{(k)}$ (which can only depend on the data through $\mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}$), that is,

$$\begin{aligned} &\min_{\theta\leq\theta_1} \mathbf{E}_\theta[\ln R^{\theta_1}_{\theta_0,(k)} | \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}] \\ &\quad = \max_q \min_{\theta\leq\theta_1} \mathbf{E}_\theta[\ln R^q_{\theta_0,(k)} | \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}]. \end{aligned}$$

In Appendix B.1.1, we show that in the limit that the risk sets are much larger than the number of events that are witnessed, this worst-case growth criterion yields a test that minimizes the worst-case expected stopping time—under the alternative hypothesis—among the tests that stop as soon as $S^q_{\theta_0,t} \geq 1/\alpha$. Thus, among all possible AV logrank tests of the form (3.10), there are strong reasons to choose $\xi^{\theta_1}_{\theta_0,\tau}$.

In a similar fashion, a test can be constructed for two sided alternatives. Indeed, consider a testing problem of the form

$$\begin{aligned} &\mathcal{H}_0 : \lambda^B = \lambda^A \quad \text{vs.} \\ &\mathcal{H}_1 : \lambda^B = \theta\lambda^A \quad \text{for some} \quad \theta \leq \theta_1 \quad \text{or} \quad \theta \geq 1/\theta_1, \end{aligned} \tag{3.11}$$

where $\theta_1 < 1$. For this problem, we can create a weighted, conditionally GROW, E-variable by using $R^{\text{2-sided}} = \frac{1}{2} R^{\theta_1}_{\theta_0,(k)} + \frac{1}{2} R^{1/\theta_1}_{\theta_0,(k)}$.

3.3.2. Learning the hazard ratio from data

So far, the alternative hypotheses that we have studied are of the form $\mathcal{H}_1 : \theta \leq \theta_1$ for some value of $\theta_1 < 1$. In some cases, such a value of θ_1 is available from the context of the analysis. For instance, θ_1 can correspond to a minimal clinically relevant effect that is satisfactory in a medical trial. However, sometimes it is not clear which value θ_1 to chose. Still, statistics of the form (3.10) are useful to test a null hypothesis $\mathcal{H}_0$ as in (3.3). Indeed, for each k, we can use conditional probability mass functions $q_{(k)}$ that depend on data observed on $t < T^{(k)}$ and enable us to implicitly learn the hazard ratio θ. We describe two such alternatives: a prequential plug-in likelihood and Bayes predictive distribution.

Prequential plugin test approach

Using only the data observed in $t < T^{(k)}$, let $\hat{\theta}_{(k)}$ be the smoothed maximum likelihood estimator

$$\hat{\theta}_{(k)} = \arg\max_{\theta\geq 0} \left(p_{\theta,0} \times \prod_{k:T^{(k)}<t} p_{\theta,(k)}(I_{(k)}) \right),$$

where $p_{\theta,0}$ is a smoothing based on the likelihood of having observed two "virtual" data points prior to the observed data, that is, $p_{\theta,0} = 1/(\bar{y}_0^A + 1 + \theta(\bar{y}_0^B + 1)) \times \theta/(\bar{y}_0^A + \theta(\bar{y}_0^B + 1))$. The statistic $S_{\theta_0,t}^{\text{preq}}$ is (3.10) with $q_{(k)} = p_{\hat{\theta}_{(k)},(k)}$, and it can also be used to define an anytime-valid test. With this choice, the process $q_{(1)}, q_{(2)} \ldots$, is a typical instance of a *prequential plug-in* likelihood [Dawid, 1984], that is often based on suitable smoothed likelihood-based estimators [Grünwald and Roos, 2019]. The rationale behind this method is the following. Suppose the data are actually sampled from a distribution according to which the hazard ratio is θ. For sufficiently large initial risk sets, that is, if $\bar{y}_0^A$ and $\bar{y}_0^B$ are not too small, by the law of large numbers, the smoothed maximum likelihood estimate $\hat{\theta}_{(k)}$ will with high probability be close to θ. Therefore, $p_{\hat{\theta},(k)}$ will behave more and more like the real $p_{\theta,(k)}$ from which data are sampled. Thus, the process $S_{\theta_0}^{\text{preq}}$, will behave more and more similarly to the "correct" partial likelihood ratio (3.6).

Bayesian approach

Instead of $q_{(k)}$ based on a plug-in estimate of θ, it is also possible to use a Bayes predictive distribution based on a prior $\mathbf{W}$ on θ. If $\mathbf{W}_{(k)} = \mathbf{W} \mid \mathbf{y}_{(1)}, \ldots, \mathbf{y}_{(k)}$ is the Bayes posterior on θ based on a prior $\mathbf{W}$ and the data up to time $t < T^{(k)}$, then

$$q_{(k)} = p_{\mathbf{W},(k)} := \int p_{\theta,(k)} \mathrm{d}\mathbf{W}_{(k)}(\theta),$$

where $\mathbf{W}_{(1)} = \mathbf{W}$. Hence, $p_{\mathbf{W},(k)}$ is the Bayesian predictive distribution. The resulting statistic $S_t^{\mathbf{W}}$ is the result of multiplying the conditional probability mass functions $p_{\mathbf{W},(k)}$, and we obtain that

$$S_{\theta_0,t}^{\mathbf{W}} = \prod_{k:T^{(k)} \le t}^{n} \frac{p_{\mathbf{W},(k)}(I_{(k)})}{p_{\theta_0,(k)}(I_{(k)})} \tag{3.12}$$

is a Bayes factor between the Bayes marginal distribution based on $\mathbf{W}$ and θ_0. This technique has been employed in sequential analysis; it is known as the method of mixtures [Darling and Robbins, 1967, Robbins and Siegmund, 1970]. We do not know of a prior for which (3.12) or the constituent products have an analytic expression, but it can certainly be implemented using, for example, Gibbs sampling.

As shown in Section 3.3, the use of any $S_{\theta_0,t}^{q}$ instead of $S_{\theta,t}^{\theta_1}$ does not compromise on safety: a test based on monitoring $S_{\theta_0}^{q}$ is anytime-valid, whether q makes reference to plug-in estimators or Bayes predictive distributions, no matter what prior $\mathbf{W}$ was chosen. The type-I error guarantee always holds, also when the prior is "misspecified", putting most of its mass in a region of the parameter space far from the actual θ from which the data were sampled. Thus, our set-up is intimately related to the concept of *luckiness* in the machine learning theory literature [Grünwald and Mehta, 2019] rather than to "pure" Bayesian statistics. Indeed, given a target value θ_1—a minimal clinically relevant effect size—the *worst-case* logarithmic growth rate of $S_{\theta_0,t}^{q}$ will in general be smaller than that of the GROW $S_{\theta_0,t}^{\theta_1}$. Nevertheless, $S_{\theta_0,t}^{q}$ can come close to the optimal for a whole range of potentially data-generating θ and may thus sometimes be preferable over choosing $S^{\theta_1}_{\theta_0,t}$. More precisely, the use of a prior allows us to exploit favorable situations in which θ is even smaller (more extreme) than θ_1. In such situations, the GROW $S^{\theta_1}_{\theta_0,t}$ is effectively misspecified. By using $S^{q}_{\theta_0,t}$ that learn from the data, we may actually obtain a test martingale that grows faster than the GROW $S^{\theta_1}_{\theta_0,t}$, which is fully committed to detecting the worst-case θ_1.

In Figure 3.1, we illustrate such a situation where we start with 1000 participants in both groups. We generated data using different hazard ratios, and used a 'misspecified' $S^{\theta_1}_{\theta_0,t}$ that always used $\theta_1 = 0.8$. Note that while this is still the GROW (minimax optimal) test martingale for $\mathcal{H}_1 : \theta \leq \theta_1 \leq 0.8$. If we knew the true θ, we could use the test martingale $S^{\theta}_{\theta_0,t}$—it grows faster. We will call the test based on this latter martingale the *oracle* exact AV logrank test because it is based on inaccessible (oracle) knowledge. We estimated the number of events needed to reject the null with 80% power for $S^{0.8}_{\theta_0,t}$, the oracle $S^{\theta}_{\theta_0,t}$, and the prequential plug-in $S^{\text{preq.}}_{\theta_0,t}$. In all cases, we used the aggressive stopping rule that stops as soon as the statistic in question crosses the threshold $1/\alpha = 20$. We see that, as the true θ gets smaller than 0.8, we need fewer events using the GROW test $S^{0.8}_{\theta_0,t}$ (the data are favorable to us), but using the oracle exact AV logrank test we get a considerable additional reduction. The prequential plug-in $S^{\text{preq.}}_{\theta_0}$ 'tracks' the oracle $S^{\theta}_{\theta_0,t}$ by learning the true θ from the data: for θ near 0.8, it behaves worse (more data are needed) than $S^{0.8}_{\theta_0,t}$ (which knows the right θ from the start), but for $\theta < 0.6$ it starts to behave better. For comparison we also added the methods discussed in Section 3.4.1. Notably, the O'Brien-Fleming procedure, even though unsuitable for optional continuation, needs even more events than the misspecified AV logrank test $S^{0.8}_{\theta_0,t}$ as soon as θ goes below 0.8. The simulations were performed using exactly the same algorithms as for Figure 3.4 so the y-axis at $\theta = 0.8$ coincides with that of Figure 3.4, but now with absolute rather than relative numbers; details are described in Appendix B.1.4.

3.3.3. Tied observations

Here, we propose a sequential test for applications where events are not monitored continuously, but only at certain observation times. In this case, more than one event may be witnessed in the time interval between two observation moments. Since the order in which these observations are made would be unknown, our previous approaches fail to offer a satisfactory sequential test. Assume that we make observations at times $t_0 < t_1 < t_2 < \ldots$ that are fixed before the start of the study. Even though we assume the absence of censoring in this section, this approach can be adapted to its presence under an additional common assumption: that the events reported between two observation times t_{k-1} and t_k precede any censorings, so that censored patients contribute fully to the risk sets under consideration. We assume that the available data are of the form $(O^A_1, O^B_1), (O^A_2, O^B_2), \ldots$, where O^A_k and O^B_k are the number of events witnessed in each group in the time interval $(t_{k-1}, t_k]$, and $O_k = O^A_k + O^B_k$ is the total. Notice that since the observation times are discrete, we can index the observations by k instead of t_k. For each k, let $\bar{y}^A_k = \sum_{j \in A} y^j_{t_k}$ the number of participants

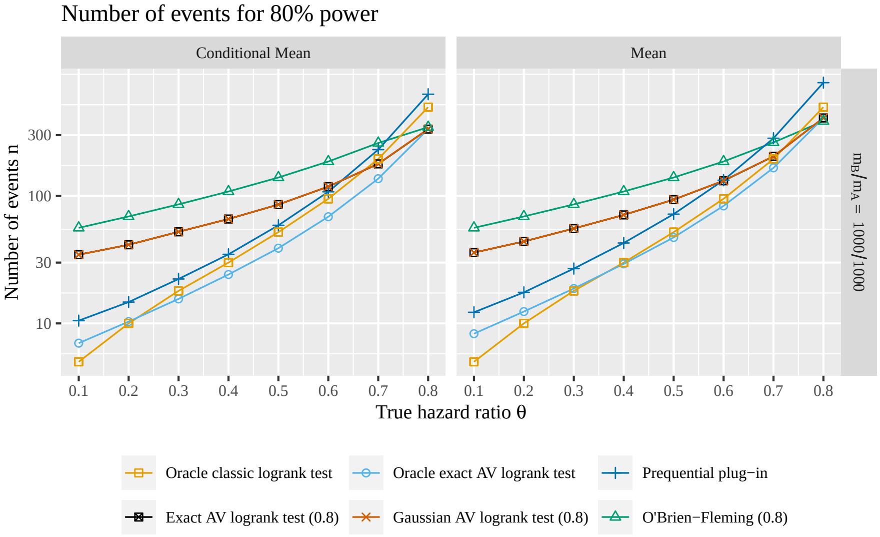

Figure 3.1.: We show the number of events at which one can stop retaining 80% power at $\alpha = 0.05$ using the process $S^{\theta_1}_{\theta_0,t}$ with $\theta_0 = 1$ and $\theta_1 = 0.80$ when the true hazard ratio θ generating the data are different from θ_1. "Oracle" means that the method is specified with knowledge of the true θ, which in reality is unknown. Note that the y-axis is logarithmic.

at risk at time t_k, define similarly $\bar{y}^B_k$, and let $\bar{y}_k = \bar{y}^A_k + \bar{y}^B_k$ be the total. We derive an anytime-valid test—a test valid at any observation time—for the problem (3.3), where the hazard ratio under the null hypothesis is $\theta_0 = 1$. The reason for this restriction in the null hypothesis—only $\theta_0 = 1$ is allowed—will soon become clear. Observe that, at time t_k, conditionally on $(\bar{y}^A_{k-1}, \bar{y}^B_{k-1})$ and the total number of events O_k, the number of events O^B_k in group B follows a hypergeometric distribution. This implies that, conditionally on $(\bar{y}^A_{k-1}, \bar{y}^B_{k-1}, O_k)$, the conditional likelihood of observing O^B_k is $p_k(O^B_k) = p_{\text{Hyper}}(O^B_k;\ \bar{y}_{k-1}, \bar{y}^B_{k-1}, O_k)$, where p_{Hyper} is the probability mass function of a hypergeometric random variable, that is,

$$p_{\text{Hyper.}}(o^B;\ \bar{y}, \bar{y}^B, o) = \frac{\binom{\bar{y}^B}{o^B}\binom{\bar{y}-\bar{y}^B}{o-o^B}}{\binom{\bar{y}}{o}}.$$

With this observation at hand, we can build, analogously to (3.10) from the continuous-monitoring case, anytime-valid tests based on partial likelihood ratios,

$$S^q_{1,k} = \prod_{l \leq k} \frac{q_l(O^B_l)}{p_l(O^B_l)}, \tag{3.13}$$

where each q_k is a conditional distribution on the possible values of O^B_k that only depends on the data up to time t_{k-1}. Following the same steps as in Section 3.3, a sequential test based on monitoring whether $S^{q}_{1,k}$ crosses the threshold $1/\alpha$ is also anytime valid at level α.

Lemma 3.3.2. Let $t_\kappa \in \{t_1, t_2, \dots\}$ be an arbitrary random time. The test $\xi^{q}_{1,\kappa}$ given by $\xi^{q}_{1,\kappa} = \mathbf{1}\left\{S^{q}_{1,\kappa} \geq 1/\alpha\right\}$, where $S^{q}_{1,\kappa}$ is as in (3.13), has type-I error bounded by α, that is,

$$\mathbf{P}_0\{\xi^{q}_{1,\kappa} = 1\} \leq \alpha,$$

under any distribution $\mathbf{P}_0$ such that the hazard ratio is $\theta = 1$.

Just as in the proof of Proposition 3.3.1, this lemma is shown by a combination of the martingale property of $S^{\theta_1}_{1,k}$ and Doob's maximal inequality. Therefore, we omit the proof of Lemma 3.3.2.

In order to obtain an optimal test under a particular hazard ratio θ_1—an alternative hypothesis—, it is necessary to compute the partial conditional likelihood for the data under the alternative of having observed O^B given $(\bar{y}^A_{k-1}, \bar{y}^B_{k-1}, \bar{N}_{k-1})$. This conditional likelihood is given by Fisher's noncentral hypergeometric distribution with parameter ω. Unfortunately, ω depends on the baseline hazard function λ, which is assumed to be unknown (see Appendix B.1.3 for details). It is for this reason that we restrict the null hypothesis to $\theta_0 = 1$. Luckily, since the test based on $S^{q}_{\theta_0,t}$ remains valid even if q is only approximately correct, this problem can be skirted. As also noted by Mehrotra and Roth [2001], when the times between observations are short, the parameter ω is well approximated by θ_1, the hazard ratio under the alternative hypothesis—no knowledge of λ is needed for the approximation. With this in mind, we put forward the use of $S^{\theta_1}_{\theta_0,k}$

$$S^{\theta_1}_{1,k} := \prod_{l\leq k} \frac{p_{\theta_1,k}(O^B_k)}{p_{1,k}(O^B_k)}$$

where $S^{\theta_1}_{1,k}$ is a an instance of (3.13) with $q_k(O^B_k) = p_{\theta_1,k}(O^B_k)$ and $p_{\theta_1,k}(O^B_k)$ is Fisher's noncentral hypergeometric distribution with parameter $\omega = \theta_1$, that is,

$$\begin{aligned} p_{\theta_1,k}(o^B) &= p_{\text{FNCH}}(o^B;\ \bar{y}, \bar{y}^B, o, \omega = \theta_1) \\ &= \frac{\binom{\bar{y}^B}{o^B}\binom{\bar{y}-\bar{y}^B}{o-o^B}\theta_1^{o^B}}{\sum_{\max\{0,o^B-\bar{y}^B\}\leq u\leq \min\{\bar{y}^B,o^B\}}\binom{\bar{y}^B}{u}\binom{\bar{y}-\bar{y}^B}{o^B-u}\theta_1^u}. \end{aligned}$$

We remark that despite $p_{(\theta_1),k}$ being only approximately the correct distribution for the observations under the alternative, type-I error guarantees are not compromised (see the discussion on luckiness in Section 3.3.2). In any case, this approximation is accurate when the time between two consecutive observation times is not very long and when the number of tied observations is small. Two reassuring remarks are in order. First, in the special case when only one observation is made in each time interval between two consecutive observation moments, the statistic $S^{\theta_1}_{1,k}$ reduces to the continuously monitored AV logrank test (3.6) at time t_k. Second, the score test associated to $S^{\theta_1}_{1,k}$ coincides with the logrank test as is conventionally computed in the presence of ties.

3.4. A Gaussian approximation to the AV logrank test

In this section we present an approximation to the AV logrank test introduced in the previous section. This is based on a sequential-Gaussian approximation to the logrank statistic. The approximation is of interest for two reasons. First, in practical situations, only the logrank Z-statistic (a standardized form of the classic logrank statistic) and other summary statistics may be available—and not the full risk-set process. This is often the case in medical trials, where the full data sets are confidential. If we also know the number of events $\bar{N}_k$ and the initial number of participants in both groups, m^A and m^B, the Gaussian approximation to the AV logrank statistic can still be used. The second reason, which we address in Section 3.4.1, is related to the fact that α-spending and group-sequential approaches, which we use as benchmarks, are also based on Gaussian approximations to the classic logrank statistic. Consequently, the behavior of the Gaussian approximation gives further insights into how the AV logrank statistic compares to group-sequential and α-spending approaches as well. We henceforth focus on the main case of interest, $\theta_0 = 1$.

Our general strategy is close in spirit to that followed in the construction of the exact AV logrank statistic in Section 3.3. We build likelihood ratios using a classic approximation for the distribution of the original logrank statistic [Schoenfeld, 1981]. If the distribution of this statistic was exactly normal, we could monitor continuously its likelihood ratio. We show through extensive simulation in which regimes this approximation behaves similarly to the AV logrank statistic.

We begin by recalling the definition of the Z-score associated to the classic logrank test. Let $E_i^B = O_i p_i^B$ with $p_i^B = \bar{y}_i^B/(\bar{y}_i^A + \bar{y}_i^B)$ be the expected (under the null) number of events witnessed in the time interval $(t_{i-1}, t_i]$ in group B, and let $V_i^B = O_i\ p_i^B(1-p_i^B)\frac{\bar{y}_i - O_i}{\bar{y}_i - 1}$ be its variance. After k observations the Z-score associated to the classic logrank statistic, Z_k, is given by

$$Z_k = \frac{\sum_{i\leq k}\left\{O_i^B - E_i^B\right\}}{\sqrt{\sum_{i\leq k} V_i^B}}. \tag{3.14}$$

The numerator in Z_k is the classic logrank statistic $H_k = \sum_{i\leq k}\left\{O_i^B - E_i^B\right\}$, which is typically interpreted as the cumulative difference between observed counts O_i^B and the expected counts E_i^B in Group B. The factor $\frac{\bar{y}_i - O_i}{\bar{y}_i - 1}$ found in V_i^B can be interpreted as a multiplicity correction, that is, a correction for ties [Klein and Moeschberger, 2003, p. 207]. When only one event is witnessed between two consecutive observation times, then $O_i = 1$, $E_i^B = p_i^B$, and $V_i^B = p_i^B(1-p_i^B)$. We remark that the above formulation is also found in the work of Cox [1972, (26)].

We put forward the Gaussian approximation S_k^G to the logrank statistic $S_{1,k}^{\theta_1}$—we show its derivation in Appendix B.3—, given by

$$S_k^G := \exp\left(-\frac{1}{2}\bar{N}_k\mu_1^2 + \sqrt{\bar{N}_k}\mu_1 Z_k\right), \tag{3.15}$$

where $\bar{N}_k$ is the total number of observations up until time t_k and

$$\mu_1 = \log(\theta)\sqrt{m^B m^A/(m^A + m^B)^2}.$$

For an arbitrary random observation time $t_K \in \{t_1, t_2, \dots\}$, we refer to the test $\xi_K^G = \mathbf{1}\left\{S_K^G \geq 1/\alpha\right\}$ as the Gaussian AV logrank test for (3.3). Recall that we test $\theta_0 = 1$, which corresponds to the asymptotic mean of the Z-score under the null hypothesis being $\mu_0 = 0$. In Appendix B.3.1 extensive simulations are performed to show in which regimes the Gaussian logrank test retains type-I error guarantees. In Appendix B.3.2, it is shown that, under continuous monitoring, the Gaussian AV logrank test tends to be more conservative—it needs more data than the exact one. The conclusion is the following: S_K^G can be used for designs with balanced allocation, and it approximates $S_{1,K}^{\theta_1}$ well for hazard ratios between 0.5 and 2.

We now compare the rejection regions defined by the Gaussian logrank test to those of continuously monitoring using α-spending and group-sequential approaches.

3.4.1. Rejection region and α-spending

In this section we compare the rejection regions of the Z-scores for which α-spending approaches and the AV logrank test for the null hypothesis of no effect (hazard ratio $\theta_0 = 1$). The two main α-spending approaches discussed here are due to Pocock [1977] and O'Brien and Fleming [1979]. We provide two reasons why the main focus of the comparison, however, will be on the O'Brien-Fleming approach. Firstly, in retrospect, Pocock himself believes that his approach leads to boundaries that are unsuitable [Pocock, 2006]. One main feature of the Pocock procedure is that the rejection regions are the same regardless of whether the (interim) analyses are conducted at the start or the end of the trial. In practice this leads to many stopped trials for benefits based on (too) small sample sizes and with unrealistically large treatments effects [Pocock, 2006]. In contrast, the rejection boundary of the O'Brien-Fleming is more conservative at the start than at the end of the trial. Secondly, the Pocock procedure only allows for a finite number of planned analyses and, therefore, cannot be monitored continuously, whereas this is possible with the O'Brien-Fleming α-spending approach. Hence, the fair comparison is between the two procedures (the AV logrank test and the O'Brien-Fleming α-spending approach) that allow for continuous monitoring.

We begin by specifying the rejection regions for both the Gaussian AV logrank test and that of the O'Brien-Fleming α-spending procedure. For the Gaussian AV logrank we compute the region for the Z-score that rejects the null hypothesis. Indeed, using (3.15), we can compute that whenever $m^A = m^B$, the null hypothesis is rejected as soon as

$$\begin{aligned} Z_n &\geq \frac{\sqrt{n}}{4}\ln(\theta_1) - \frac{2}{\sqrt{n}}\frac{\log(\alpha)}{\log(\theta_1)} \quad \text{if} \quad \theta_1 > 1, \text{ or} \\ Z_n &\leq \frac{\sqrt{n}}{4}\ln(\theta_1) - \frac{2}{\sqrt{n}}\frac{\log(\alpha)}{\log(\theta_1)} \quad \text{if} \quad \theta_1 < 1. \end{aligned}$$

The O'Brien-Fleming procedure is based on a Brownian-motion approximation to the sequentially computed logrank statistic Z-score. Indeed, for large values of $n_{\max}$ and $t \in [0,1]$, the process $t \mapsto \sqrt{\frac{\lfloor tn_{\max}\rfloor}{n_{\max}}} Z_{\lfloor tn_{\max}\rfloor}$ can be approximated by a Brownian motion B_t. We stress the fact that $n_{\max}$ has to be set in advance. If B_t is a Brownian motion, the reflection principle, a well-known but nontrivial application of the symmetry of B_t, implies that

$$\mathbf{P}\{\max_{0 \le t \le 1} B_t \ge c\} = 2\mathbf{P}\{B_1 \ge c\}$$

Since B_1 is Gaussian with mean zero and standard deviation 1, setting $c = q_{1-\alpha/2}$, the $(1-\alpha/2)$-quantile of a standard Gaussian distribution, then

$$\mathbf{P}\{\max_{0 \le t \le 1} B_t \ge q_{1-\alpha/2}\} = \alpha.$$

This implies that

$$\mathbf{P}\{\max_{n \le n_{\max}} \sqrt{n} Z_n \ge \sqrt{n_{\max}} q_{1-\alpha/2}\} \approx \alpha,$$

or, in other words, the procedure that continuously monitors whether the Z-score crosses the boundary $\sqrt{n_{\max}} q_{1-\alpha/2}$ guarantees approximate type-I error α. Given a hazard ratio θ_1 under the alternative hypothesis, $n_{\max}$ can be set to achieve a desired type-II error. The left-handed procedure can be worked out similarly, and we obtain that, for $m^A = m^B$, the continuous-monitoring version of the O'Brien-Fleming procedure rejects as soon as

$$Z_n \ge \sqrt{\frac{n}{n_{\max}}} q_{1-\alpha/2} \quad \text{if} \quad \theta_1 > 1 \text{ (right-sided test), or}$$
$$Z_n \le \sqrt{\frac{n}{n_{\max}}} q_{1-\alpha/2} \quad \text{if} \quad \theta_1 < 1 \text{ (left-sided test).}$$

The two regions of the Z-statistic values share an important feature: they are more conservative to reject the null hypothesis at small sample sizes than at larger ones, requiring more extreme values for the Z-statistic at the start of the trial. This sets them apart from the Pocock spending function that requires equally extreme values for the Z-statistic at small and large sample size. Figure 3.2 shows both the Gaussian AV logrank and the O'Brien-Fleming α-spending rejection regions. Additionally, Figure 3.2 shows the boundary of the Pocock α-spending function for 10 interim analyses. Note that the definition of the AV logrank test rejection region requires a very explicit value for the effect size $\theta_1 = \theta_{\min}$ of minimum clinical relevance, while that value is implicit in the definition of the α-spending rejection region: To specify an maximum sample size $n_{\max}$ to achieve a certain power, an effect size of minimal interest is also assumed. A fixed-sample-size analysis designed to detect a minimum hazard ratio of 0.7 would need 195 events to achieve 80% power if the true hazard ratio is also 0.7. A sequential analysis using α-spending requires a slightly larger maximum number of events: 205 with the O'Brien-Fleming spending function; 245, with the Pocock α-spending function—when we design for 10 interim analyses. We investigate the number of events needed by the Gaussian AV logrank test in Appendix B.3.2. For the α-spending procedures continuing beyond $n_{\max}$ is problematic. This is not the case for the AV logrank test, as it allows for unlimited monitoring, then $n_{\max}$ is only a soft constraint on the study—there is no penalty in type-I error for continuing after $n_{\max}$ events have been witnessed.

The benefit of a sequential approach is that if there is evidence that the hazard ratio is more extreme than it was anticipated under the alternative hypothesis, we can detect that with fewer events than the maximum sample size. The left column of Figure 3.3 illustrates that we benefit because the true hazard ratio could be more extreme than we designed for (e.g. 0.5 instead of 0.7; a larger risk reduction in the treatment group) and the data reflects that. We also benefit from a sequential analysis if the true hazard ratio is 0.7 but by chance the values of our Z-statistics are more extreme than expected. The major difference between α-spending approaches and the AV logrank test is that the AV test does not require to set a maximum sample size. It in fact allows to indefinitely increase the sample size without ever spending all α. An α-spending approach designed to have 80% power will miss out on rejecting the null hypothesis in 20% (the type-II error) of the cases as is illustrate in the bottom middle plot of Figure 3.3 by the sample paths that remain (dark) green. In contrast, the AV logrank test can potentially reject with 100% power by continue sampling. In the sample paths of 500 events in Figure 3.3, all but one sample path of Z-statistics could be rejected at a larger sample size by the AV logrank test. By extending the trial, the AV logrank test can potentially have 100% power if the true hazard ratio is at least as small as the hazard ratio set for minimum clinical relevance in the design of the test. Still, type-I error is controlled. The bottom right plot of Figure 3.3 shows two null sample paths with a true hazard ratio of 1 that are rejected by the O'Brien-Fleming α-spending region, but not by the AV logrank test. Here, the AV logrank test is more conservative.

It is known that α-spending methods behave poorly in case of unbalanced allocation [Wu and Xiong, 2017]. In Appendix B.3.1 we showed that our Gaussian approximation to the logrank test is also not an E-variable in case of unbalanced allocation. Our exact AV logrank test, however, is an E-variable under any allocation since it is defined directly on the risk-set process (3.8). This suggests that if the complete data set is available and allocation is unbalanced, the exact logrank test should be preferred over the Gaussian approximation and the α-spending methods.

3.5. Optional continuation and live meta-analysis

In this section, we address optional continuation and live meta-analysis—the continuous aggregation of evidence from multiple experiments. For instance, data could come from medical trials conducted in different hospitals or in different countries. In such cases, we compare a global null hypothesis $\mathcal{H}_0$ that is addressed in all trials (for instance, $\theta_0 = 1$) to an alternative hypothesis $\mathcal{H}_1$ that allows for different hazard ratios in each experiment. The present approach covers even the case in which the decision to start each experiment might depend on the observations made in experiments that are already in progress. Assume that there are k_E experiments, $E_{(1)}, \dots, E_{(k_E)}$, ordered by their respective starting times $V_{(1)} \leq \cdots \leq V_{(k_E)}$, each performed on different and independent populations. Assume further that the starting time $V_{(k)}$, of experiment

$E_{(k)}$ depends only on the data observed in the ongoing experiments $E_{(1)}, \dots, E_{(k-1)}$. If each experiment $E_{(k)}$ monitors the AV logrank statistic $S^k_{\theta_0,t}$, where $S^k_{\theta_0,t} = 1$ for $t \le V_{(k)}$, then the product statistic $S^{\text{meta}}_{\theta_0,t} = \prod_{i \le k_E} S^i_{\theta_0,t}$ is a test martingale with respect to the filtration generated by all observations. Consequently, the meta-test based on it enjoys anytime validity.

Proposition 3.5.1. *Let τ be any random time. The test $\xi^{\text{meta}}_{\theta_0,\tau} = \mathbf{1}\left\{S^{\text{meta}}_{\theta_0,\tau} \ge 1/\alpha\right\}$, where $S^{\text{meta}}_{\theta_0,t} = \prod_{i \le k_E} S^i_{\theta_0,t}$, has type-I error smaller than α.*

This result follows from a reduction to independent left-truncation—we refer to left-truncation in the specific sense defined by Andersen et al. [1993]. Indeed, even in the presence of dependencies on other studies, the observations made in $E_{(k)}$ can be regarded as a left-truncated sample. Here, the time at which observation in $E_{(k)}$ is started is random and only participants that have not witnessed an event are recruited into the study. One may worry that these dependencies may alter the sequential properties of $S^{\text{meta}}_{\theta_0,t}$, but this is not the case. Since the truncation time for $E_{(k)}$ is based on data that are independent of that of experiment $E_{(k)}$—it is possibly based on the observations made in all other experiments, it follows from results of Andersen et al. [1993] (see Appendix B.1.2) that the sequential-multinomial interpretation of the partial likelihood for the truncated data remains valid. Consequently, so does the sequentially computed AV logrank statistic and the product statistic $S^{\text{meta}}_{\theta_0,t}$. By continuously monitoring $S^{\text{meta}}_{\theta_0,t}$, we effectively perform an *online, cumulative* and possibly *live* meta-analysis that remains valid irrespective of the order in which the events of the different trials are observed. Importantly, unlike in α-spending approaches, the maximum number of trials and the maximum sample size (number of events) per trial do not have to be fixed in advance; we can always decide to start a new trial, or to postpone to end a trial and wait for additional events.

3.6. Anytime-valid confidence sequences

Anytime-valid (AV) confidence sequences corresponds to anytime-valid tests in the same way fixed-sample tests correspond to confidence intervals. Indeed, it is possible to "invert" a fixed-sample test to build a confidence interval: the parameters of the null hypothesis that are not rejected by a the test form a confidence interval. Analogously, test martingales can be used to derive AV confidence sequences [Darling and Robbins, 1967, Lai, 1976, Howard et al., 2018a,b]. In our setting, a $(1-\alpha)$-AV confidence sequence is a sequence of confidence intervals $\{\text{CI}_t\}_{t \ge 0}$, such that

$$\mathbf{P}_\theta\{\theta \notin \text{CI}_t \quad \text{for some} \quad t \ge 0\} \le \alpha. \tag{3.16}$$

A standard way to design $(1-\alpha)$-AV confidence sequences, translated to our logrank setting, is to use a prequential plug-in test martingale $S^{\text{preq}}_{\theta_0,t}$ or the Bayesian version $S^{\mathbf{W}}_{\theta_0,t}$ as in Section 3.3.2. At time t, one reports $\text{CI}_t = [\theta^L_t, \theta^U_t]$ where CI_t is the smallest interval containing the values of θ_0 such that $S^{\text{preq}}_{\theta_0,t} > 1/\alpha$ outside this interval. Ville's inequality readily implies that this is indeed an AV confidence sequence. The same construction can be made for arbitrary instances of $S^q_{\theta_0,t}$ as in (3.10).

3.7. Power and sample size

In this section, we investigate the power properties of the AV logrank test—we will study specific stopping times. We have seen that by observing arbitrarily long sequences of events the logrank test can achieve type-II errors that are as close to zero as desired. However, in practice it is necessary to plan for a maximum number of events $n_{\max}$ so that either the experiment is stopped as soon as the null hypothesis is rejected or when $n_{\max}$ events have been observed. In the latter case, there is no evidence to reject the null hypothesis. We assess via simulation the value of $n_{\max}$ needed to guarantee 20% type-II error (80% power) for the exact and Gaussian AV logrank tests. We compare this to the $n_{\max}$ needed to achieve the same power using the continuous-monitoring O'Brien-Fleming α-spending procedure introduced in the previous section, and the fixed-sample-size classic logrank test. Figure 3.4 show simulation results establishing three types of sample sizes. The leftmost panels ("Maximum") shows the sample size $n_{\max}$ described earlier, which would be required to design the experiment. We stress the fact that using the classic logrank test or α-spending designs events beyond $n_{\max}$ cannot be analyzed. The rightmost panel of Figure 3.4 ("Mean") shows the sample sizes that capture the expected duration of the trial. It expresses the mean number of events, under the alternative hypothesis, that will be observed before the trial can be stopped. Here, for the AV logrank tests, we use the aggressive stopping rule that stops as soon as $S^{\theta_1}_{\theta_0,t} \geq 1/\alpha = 20$ or $n = n_{\max}$. In case of α-spending approaches and the AV logrank test this number of events is always smaller than the maximum needed in the design stage. Lastly, the middle panel ("Conditional Mean") shows an even smaller number for those tests that have a flexible sample size: the expected stopping time *given* that the trial is stopped before the maximum $n_{\max}$ was reached—this only happens if the null is rejected. For comparison purposes, all sample sizes are shown relative to (i.e., divided by) the fixed sample size needed by the classical logrank test to obtain 80% power. Note that for small sample size (for small hazard ratios), both the classic logrank test and O'Brien-Fleming α-spending are not recommended due to lack of type-I error control. They are based on Schoenfeld's Gaussian approximation, which underestimates the number of events required for hazard ratios far away from 1. For example, simulations show that for $\theta_1 = 0.1$, $n = 6$ or 7 events will be necessary—for small sample sizes the classical logrank test is not recommended due to lack of type-I error control. We give further details in Appendix B.1.4 (see also Figure 3.4). In summary, at all hazard ratios at which the Gaussian approximation to the classic logrank test is accurate (say for $\theta_1 \geq 0.3$), the mean number of events needed by the AV logrank tests is about the same or noticeably smaller than that needed when using a fixed-sample-size analysis.

3.8. Discussion, Conclusion and Future Work

We introduced the AV logrank test, a version of the logrank test that retains type-I error guarantees under optional stopping and continuation. Extensive simulations reveal that, if we do engage in optional stopping, it is competitive with the classic logrank test (which neither allows in-trial optional stopping nor optional continuation) and α-spending procedures (which allows forms of optional stopping but not optional continuation). We provided an approximate test for applications in which only summary statistics are available and also showed how the AV logrank test can be used in combination with (informative) priors and prequential learning approaches, when no effect size of minimal clinical relevance can be specified. Two of our extensions invite further research: we introduced anytime-valid confidence sequences for the hazard ratio, and will study their performance in comparison to other approaches in future work. We also introduced an extension to Cox' proportional hazards regression, which guarantees type-I error guarantees even if the alternative model is equipped with arbitrary priors. In future work, we plan to implement this extension—which requires the use of sophisticated methods for estimating mixture models. The GROW AV logrank tests (exact and Gaussian) are already available in our `safestats` R package [Turner et al., 2022]. We end with two final points of discussion: *staggered entries* and *doomed trials*.

3.8.1. Staggered entry

Earlier approaches to sequential time-to-event analysis were also studied under scenarios of staggered entry, where each patient has its own event time (e.g., time to death since surgery), but patients do not enter the follow-up simultaneously (such that the risk set of, say, a two-day-after-surgery event changes when new participants enter and survive two days). Sellke and Siegmund [1983] and Slud [1984] show that, in general, martingale properties cannot be preserved under such staggered entry settings, but that asymptotic results are hopeful [Sellke and Siegmund, 1983] as long as certain scenarios are excluded [Slud, 1984]. When all participants' risk is on the same (calendar) time scale (e.g., infection risk in a pandemic; staggered entry now amounts to left-truncation, which we can deal with), or new patients enter in large groups (allowing us to stratify), staggered entry poses no problem for our methods. But research is still ongoing into those scenarios in which our inference is fully AV for patient time under staggered entry, and those that need extra care.

3.8.2. Your trial is not doomed

In their summary of conditional power approaches in sequential analysis Proschan, Lan, and Wittes [2006] write that low conditional power makes a trial futile. Continuing a trial in such case could only be worth the effort to rule out an effect of clinical relevance, when the effect can be estimated with enough precision. However, if "both conditional and revised unconditional power are low, the trial is doomed because a null result is both likely and uninformative" [Proschan et al., 2006, p. 63]. While this is the case for all existing sequential approaches that set a maximum sample size, this is not the case for AV tests. Any trial can be extended and possibly achieve 100% power or in an anytime-valid confidence sequence show that the effect is too small to be of interest. This is especially useful for time-to-event data when sample size can increase by extending the follow-up time of the trial, without recruiting more participants.

Moreover, new participants can always be enrolled either within the same trial or by spurring new trials that can be combined indefinitely in a cumulative meta-analysis.

Acknowledgements

This work is part of the research program with project number 617.001.651, which is financed by the Dutch Research Council (NWO). We thank Henri van Werkhoven, Richard Gill, Wouter Koolen, Aaditya Ramdas and Rosanne Turner for useful conversations.

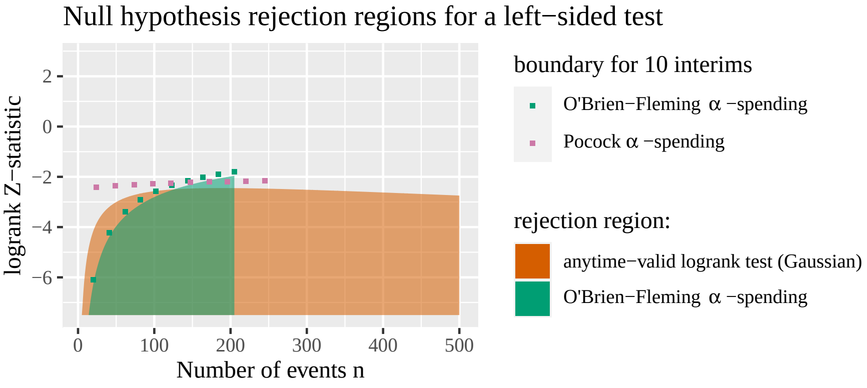

Figure 3.2.: Left-sided rejection regions for continuous-monitoring using O'Brien-Fleming α-spending or the Gaussian AV logrank test. Allocation is balanced ($m^A = m^B$) and $\alpha = 0.05$. Also shown are the O'Brien-Fleming and Pocock α-spending boundaries for 10 interim analyses. The α-spending boundaries are designed to have 80% power when detecting a hazard ratio 0.7. For more details, including the values of $n_{\max}$, see Section 3.4.1.

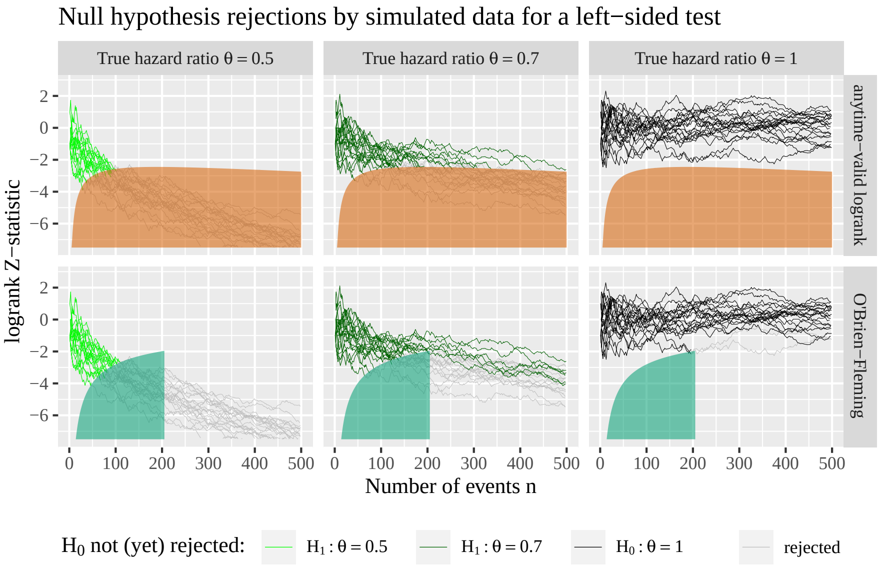

Figure 3.3.: Null hypothesis rejections on simulated data. The rejection regions are the same as shown in Figure 3.2 (designed to detect a hazard ratio of 0.7 with 80% power). Data are simulated under balanced allocation ($m_1 = m_0 = 5000$) and as time-to-event data with possible ties. The logrank Z-statistic does not have a value for all n; it sometimes jumps with several additional events at a time.

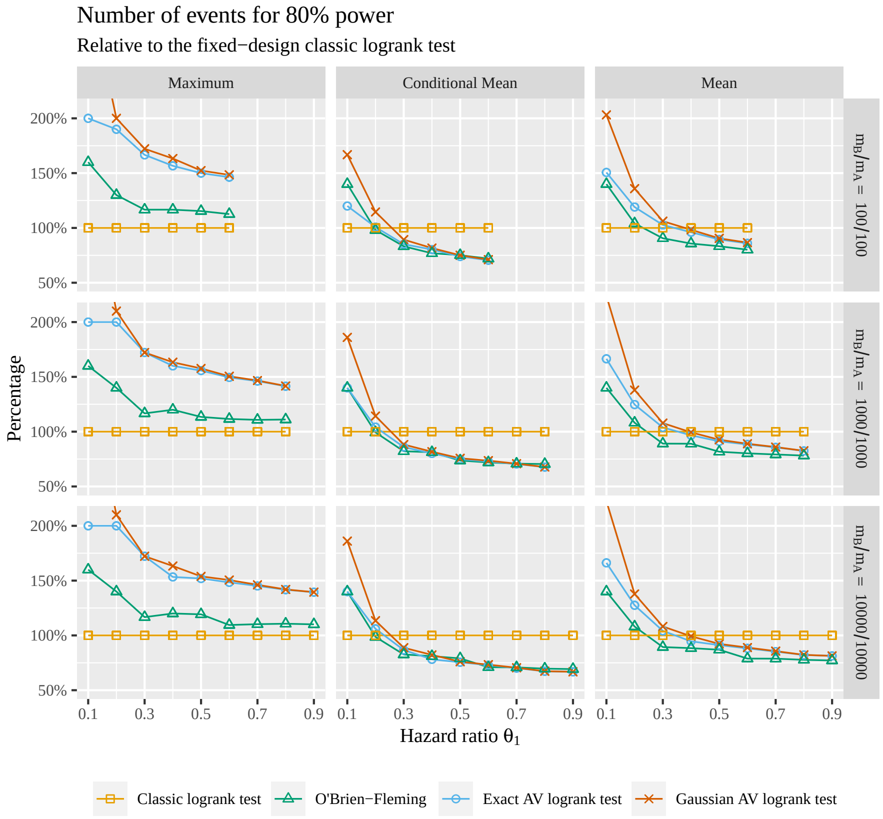

Figure 3.4.: Maximum, expected (Mean) number of events needed to reject the null hypothesis with 80% power. 'Conditional Mean' makes reference to the number of events needed given that the null hypothesis is indeed rejected. The maximum number of events needed using AV logrank statistics is higher than that of a fixed-sample test, but lower in expectation (see Section 3.7). All simulations are performed with $\alpha = 0.05$ and tests are designed to detect the hazard ratio θ_1 shown on the x-axis. Data are generated using that same hazard ratio. The classical logrank test needs the following sample sizes (number of events) $n(\theta_1)$ for an 80%-power design to detect hazard ratio θ_1: $n(0.1) = 5$, $n(0.2) = 10$, $n(0.3) = 18, n(0.4) = 30$, $n(0.5) = 52$, $n(0.6) = 95$, $n(0.7) = 195$, $n(0.8) = 497$ and $n(0.9) = 2228$. These sample sizes represent the 100% line in all plots.

4. Luckiness in Multiscale Online Learning[1]

Algorithms for full-information online learning are classically tuned to minimize their worst-case regret. Modern algorithms additionally provide tighter guarantees outside the adversarial regime, most notably in the form of constant pseudoregret bounds under statistical margin assumptions. We investigate the multiscale extension of the problem where the loss ranges of the experts are vastly different. Here, the regret with respect to each expert needs to scale with its range, instead of the maximum overall range. We develop new multiscale algorithms, tuning schemes and analysis techniques to show that worst-case robustness and adaptation to easy data can be combined at a negligible cost. We further develop an extension with optimism and apply it to solve multiscale two-player zero-sum games. We demonstrate experimentally the superior performance of our scale-adaptive algorithm and discuss the subtle relationship of our results to Freund's 2016 open problem.

4.1. Introduction

The abstract problem of *online prediction with expert advice* [Littlestone and Warmuth, 1994, Freund and Schapire, 1997] is of fundamental importance in computational learning theory. Efficient and optimal algorithms for solving it have a substantial impact on various problems in general online convex optimization [Hazan, 2021], online model selection [Foster et al., 2017], boosting [Freund and Schapire, 1997], and maximal probabilistic inequalities [Rakhlin and Sridharan, 2017], to name a few. Concretely, a decision maker chooses among experts' advices sequentially, and the environment assigns each advice a scalar loss. If all losses have the same numerical range $[-\sigma, \sigma]$, the situation is well understood. Indeed, Freund and Schapire [1997] showed that, for K experts and t rounds, the Hedge algorithm guarantees the minimax regret (defined below) $\sigma\sqrt{2t \ln K}$. Furthermore, modern algorithms additionally guarantee lower or even constant regret when the sequence of losses is more benign [see De Rooij et al., 2014, Koolen and van Erven, 2015, Mourtada and Gaïffas, 2019].

[1] This chapter is based on M. F. Pérez-Ortiz and W. M. Koolen. Luckiness in Multiscale Online Learning. *Advances in Neural Information Processing Systems*, 35: 25160–25170, Dec. 2022. URL `https://papers.nips.cc/paper_files/paper/2022/hash/a0d2345b43e66fa946155c98899dc03b-Abstract-Conference.html`

In the multiscale setting, where the experts' loss ranges may differ by orders of magnitude, it is natural to ask about the existence of algorithms that guarantee an optimal worst-case regret bound that scales with the loss range of the best expert instead of the maximum range. This question has been answered affirmatively [Chen et al., 2021, Bubeck et al., 2019, Cutkosky and Orabona, 2018, Foster et al., 2017]. The algorithms developed in this line of work have had a significant impact in different areas of computational learning theory and practice. Unfortunately, as we will see, the best known algorithms still fail to guarantee lower regret even for the simplest benign statistical cases. Ensuring these goals poses serious technical challenges. In particular, Bernstein's inequality, the engine of classical same-scale luckiness arguments, has no suitable multiscale upgrade. Moreover, intuitive candidate upgrades of same-scale results would contradict recent lower bounds (see Section 4.7). To make things worse, in order to obtain multiscale regret bounds, close attention needs to be paid to terms that are conventionally insignificant but now carry the maximum scale of the problem. This motivates our main question: *can a single algorithm have multiscale worst-case regret guarantees and, in addition, exhibit constant (pseudo)regret in stochastic lucky cases?*

We answer the previous question affirmatively. The key contribution in this chapter is MUSCADA (multiscale adaptive), a computationally efficient algorithm that simultaneously guarantees a worst-case regret that grows with the scale of the best expert, and constant expected pseudoregret under a stochastic margin condition. MUSCADA uses a refined version of Follow the Regularized Leader based on the multiscale entropy of Bubeck et al. [2019]. Its crucial improvement is a second-order variance-like adaptation, the tightest possible for the analysis of this regularizer. This second-order adaptation is close in spirit to, and an improvement of, that of AdaHedge by De Rooij et al. [2014] and those of Chen et al. [2021]. As a result of careful analysis, MUSCADA has the following attractive properties: it does not need knowledge of the length of the game in advance without resorting to any doubling trick, the presence of zero-regret rounds does not change the state of the algorithm or its regret guarantees; it is invariant both under per-round, possibly unknown, translations of each expert's losses, and under a global known scaling common to all losses and ranges.

As an application of MUSCADA and its analysis techniques, we build an optimistic variant of the algorithm and use it to solve two-person zero-sum games that have a multiscale structure. The optimistic variant makes use of a guess of what the losses in the next round will be, and achieves lower regret when the guesses are adequate. This interest originates in the fact that optimistic algorithms converge to the solutions of such games at faster rates than their nonoptimistic counterparts [Syrgkanis et al., 2015]. We find experimentally that MUSCADA outperforms existing single-scale algorithms when the payoff matrix of the game exhibits a multiscale structure.

In the rest of this introduction we lay out formally the multiscale experts problem, review existing work, present a summary of the main contributions (Section 4.1.1), and outline the rest of the chapter.

Full-information online learning. In its simplest form, we must decide sequentially in rounds how to aggregate the predictions made by a fixed number K of *experts.* At each round t, we choose an aggregation strategy, a probability distribution $\boldsymbol{w}_t \in \mathcal{P}(K)$ over experts. After choosing $\boldsymbol{w}_t$, we assess the quality of the experts' predictions with a numerical loss $\ell_t = (\ell_{t,k})_{k \in K}$ and judge the performance of our aggregation strategy by the $\boldsymbol{w}_t$-weighted losses $\langle \boldsymbol{w}_t, \ell_t \rangle = \sum_{k \in K} w_{t,k} \ell_{t,k}$. Our objective is to minimize the cumulative gap between the losses incurred by our aggregation strategy $t \mapsto \boldsymbol{w}_t$ and the best expert in hindsight. This cumulative gap is the *regret* $\mathcal{R}_t = \sum_{s=1}^{t} \langle \boldsymbol{w}_s, \ell_s \rangle - \min_{k \in K} \sum_{s=1}^{t} \ell_{t,k}$. Other than range restrictions on the losses, no assumptions are made about the mechanism that generates them. More precisely, for each expert $k \in K$ and all rounds t, we only assume that $\ell_{k,t} \in [-\sigma_k, \sigma_k]$ for known nonnegative scales $\{\sigma_k\}_{k \in K}$. We call $\boldsymbol{R}_t$ the vector of regrets with respect to each expert, that is, the vector with entries $R_{t,k} = \sum_{s=1}^{t} \{\langle \boldsymbol{w}_s, \ell_s \rangle - \ell_{s,k}\}$.

Existing results. Several algorithms have been proposed that achieve the worst-case regret in the multiscale setting, but none of them achieve constant regret in stochastic lucky cases. Motivated by the problem of online model selection, Foster et al. [2017] used a technique of adaptive relaxations to produce randomized algorithms that guarantee

$$\mathbf{E}_{\mathbf{P}}[R_{t,k}] = O\left(\sigma_k \sqrt{t(\ln t + \ln(1/\pi_k) + \ln(\sigma_k/\sigma_{\min}))}\right) \quad \text{as } t \to \infty,$$

where $\boldsymbol{\pi}$ is a prior distribution on experts that generalizes the uniform $1/K$ of the Hedge algorithm and the expectation is over the algorithm's randomness. Bubeck et al. [2019] first proposed a Follow-the-Regularized-Leader algorithm with a multiscale entropy regularization that guarantees

$$R_{t,k} = O\left(\sigma_k \sqrt{t(\ln K + \ln(\sigma_{\max}/\sigma_{\min}))}\right) \quad \text{as } t \to \infty$$

when the number of rounds t is known in advance. Bubeck et al. [2019, Theorem 20] also construct an instance of the $K = 2$ experts problem in which there exists a time t for which any algorithm must have $R_{t,k'} \gtrsim \sigma_{k'} \sqrt{t(\ln K + \ln(\frac{\sigma_{\max}}{\sigma_{\min}}))}$ for some expert k', shedding some light on the minimax picture. Recently, Chen et al. [2021] designed an optimistic algorithm that uses the same regulatization as Bubeck et al. [2019] with an additional ingredient: at each round, a second-order correction is added to the losses before computing the next round's weights. At every round, their algorithm makes use of a guess vector $\boldsymbol{m}_t$ that can depend on the losses up to time $t-1$. The scale of the guesses $\boldsymbol{m}_t$ are assumed to be the same as that of the losses; $|m_{t,k}| \leq \sigma_k$. For instance, valid choices for the guess $\boldsymbol{m}_t$ are $\mathbf{0}$ and the loss ℓ_{t-1} of the previous round. The algorithm of Chen et al. [2021] achieves

$$R_{t,k} = O\left(\sigma_k \sqrt{\beta_{t,k} \ln t} + \sigma_{\max} \ln t\right) \quad \text{as } t \to \infty,$$

now scaling with the expert-dependent "time" $\beta_{t,k} = \sum_{s=1}^{t} \frac{(\ell_{s,k} - m_{s,k})^2}{\sigma_k^2} \leq 4t$. Furthermore, they show that a different single-scale tuning of their algorithm exhibits stochastic luckiness. Namely, if the losses are sampled from a distribution with a gap $d_{\min} > 0$

between the expected loss of the best expert k^* and that of any other expert, their algorithm guarantees that

$$R_{t,k^*} = O_{\mathbf{P}}\left(\frac{\ln t}{d_{\min}}\right) \quad \text{as } t \to \infty,$$

where $\mathbf{P}$ is the distribution of the losses. Their technique for stochastic luckiness uses the upcoming learner's loss as the guess $m_{t,k} = \langle \boldsymbol{w}_t, \boldsymbol{\ell}_t \rangle$. Unfortunately, this approach cannot be extended to the multiscale case, as these guesses may violate the experts' loss ranges.

4.1.1. Main results

In this section we present succinctly the regret guarantees for Muscada. Firstly, we present multiscale worst-case regret guarantees. Secondly, we present the stochastic luckiness results and Massart's margin condition. We then prove analogs of these results for an optimistic modification of Muscada in Section 4.4. We close this introduction with an outline of the rest of the chapter.

Worst-case bounds. We propose two tunings for Muscada; they cover the cases where there is or is not an expert with loss range equal to zero. Our results imply Theorem 4.1.1 below; it contains the regret guarantees for Muscada, expressed in terms of v_t, an implicitly defined variance-like second-order data-dependent quantity. The quantity v_t, defined by the algorithm, is the tightest allowed by our analysis and enables our luckiness result, Theorem 4.3.1. We interpret v_t through the upper bounds of Theorem 4.1.2, also below, as an internal scale-free measure of time, as $v_t \leq 4t$.

Theorem 4.1.1 (Regret Bounds)**.** *Consider* Muscada, *$t \mapsto v_t$ defined in Figure 4.1, and any initial probability distribution $\boldsymbol{\pi}$.*

- *If $\sigma_{\min} = \min_{k \in K} \sigma_k > 0$, Tuning 1 guarantees, for any loss sequence,*

$$R_{t,k} \leq c\,\sigma_k \sqrt{v_t(\ln(1/\pi_k) + \ln(\sigma_k/\sigma_{\min}))} + O(1) \quad \textit{as } t \to \infty, \tag{4.1}$$

where c is a constant depending only on π. The constant c is well-behaved: if $\max_{k \in K} \pi_k = 1 - \varepsilon$, then $c \leq 4\sqrt{2}(1 + 1/(2\ln(1+\varepsilon)))$.

- *Even if $\min_{k \in K} \sigma_k = 0$, Tuning 2 ensures, for any loss sequence,*

$$R_{t,k} \leq 2\sigma_k \sqrt{2\ v_t(\ln(1/\pi_k) + \ln(1 + v_t))}(1 + o(1)) \quad \textit{as } t \to \infty. \tag{4.2}$$

The following theorem (proven in Appendix C.7) shows that v_t is bounded by a second-order quantity. If $w_{t,k}$ are the weights played by Muscada at round t and $\eta_{t-1,k}$ are its learning rates, v_t is bounded by the variance over experts of the losses w.r.t. a tilted probability distribution $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$. The shape of this quantity may seem surprising, but it is not artificial; our analysis shows that it is the tightest and, consequently, the natural second-order quantity associated to this choice of regularization. In Appendix C.7, we further motivate, via a Taylor approximation, the shape of the resulting upper bound.

Theorem 4.1.2. *Let $\tilde{w}_{t,k}$ be the probability distribution $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$ and let $\Delta v_t = v_t - v_{t-1}$. Then, with either tuning from Figure 4.2, v_t, from Figure 4.1, satisfies*

$$\Delta v_t \leq 4 \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_t}(\ell_t)}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \leq 4, \qquad \textit{where} \qquad \mathrm{Var}_{\tilde{\boldsymbol{w}}_t}(\ell_t) = \langle \tilde{\boldsymbol{w}}_t, (\ell_t - \langle \tilde{\boldsymbol{w}}_t, \ell_t \rangle)^2 \rangle.$$

Stochastic luckiness. We now turn to our results for stochastic easy data. Not all stochastic scenarios are easy (in fact, worst-case regret lower bounds are proved using stochastic data). We use Massart's margin condition, a standard benchmark for easy data.

Definition 4.1.3 (Massart's easiness condition). The losses $\ell_1, \ell_2, \ldots$ satisfy Massart's easiness condition if they are generated i.i.d. from a distribution $\mathbf{P}$ with the following property: there exists a constant c_{M} and an expert $k^* \in K$ such that

$$\mathbf{E}_{\mathbf{P}}[(\ell_{t,k} - \ell_{t,k^*})^2] \leq c_{\mathrm{M}} \mathbf{E}_{\mathbf{P}}[\ell_{t,k} - \ell_{t,k^*}]$$

for all $k \in K$ and $t \geq 1$. In that case, $k^* = \arg\min_{k \in K} \mathbf{E}_{\mathbf{P}}[\ell_{t,k}]$ for all t.

Massart's condition is implied by a more interpretable gap condition [Koolen et al., 2016, Lemma 3]. If there exist a gap $d_{\min} > 0$ in expectation between the loss of any expert and that of the best one k^*, that is, if, for every $k \neq k^*$, $\mathbf{E}_{\mathbf{P}}[\ell_{1,k}] \geq d_{\min} + \mathbf{E}_{\mathbf{P}}[\ell_{1,k^*}]$, Massart's condition is satisfied with $c_{\mathrm{M}} = 1/d_{\min}$. We show the following theorem.

Theorem 4.1.4 (Constant regret under Massart's condition)**.** *Under Massart's condition (Definition 4.1.3),* Muscada *with either Tuning 1 or 2 has constant expected pseudoregret over time, that is,*

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \lesssim 1.$$

Outline. The rest of this chapter is organized as follows. In Section 4.2, we introduce and analyze Muscada. In Section 4.3, we state the main results on stochastic luckiness for Muscada. In Section 4.4, we introduce an optimistic variant of Muscada, give remarks about its numerical implementation in Section 4.5, and apply it to accelerating the solution of multiscale games in Section 4.6. We end this chapter with a discussion of our results in Section 4.7.

4.2. The Muscada Multiscale Online Learning Algorithm

In this section, we describe our algorithm and motivate its design. We present two useful tunings and prove the corresponding worst-case regret guarantees. For the sake of intuition, we specialize the algorithm to the case of same-scale experts with uniform prior and compare its resulting closed form to AdaHedge [De Rooij et al., 2014]. Stochastic luckiness results are found in Section 4.3. We begin by introducing some notation.

Notation. We use boldface type for vectors in $\mathbb{R}^K$ ($\boldsymbol{R}_t, \boldsymbol{L}_t, \boldsymbol{\mu}_t, \boldsymbol{\eta}_t, \boldsymbol{\sigma}, \boldsymbol{u}$) and distributions on K experts $(\boldsymbol{p}, \boldsymbol{w}, \boldsymbol{\pi})$. We number rounds so that all quantities indexed by t depend on the information witnessed by the learner in the first t rounds. Exceptionally, we use weights $\boldsymbol{w}_t$ at round t. For two functions f and g we write "$f = O(g)$ as $t \to \infty$" if there exists $c > 0$ such that $\lim_{t\to\infty} f(t)/g(t) \leq c$. Similarly, we write "$f(t) \sim g(t)$ as $t \to \infty$" if $\lim_{t\to\infty} f(t)/g(t) = 1$, and $f \lesssim g$ if there is $c > 0$ so that $f \leq cg$. We denote the simplex of probability distributions on K experts by $\mathcal{P}(K)$ and use K interchangeably for a number $K \in \mathbb{N}$ and the set $\{1, \dots, K\}$.

We define Muscada in Figure 4.1 and give its two main tunings in Figure 4.2. At round t, after observing cumulative corrected losses $\boldsymbol{L}_{t-1} + \boldsymbol{\mu}_{t-1}$, Muscada plays weights

$$w_{t,k} = u_k \mathrm{e}^{-\eta_{t-1,k}(L_{t-1,k}+\mu_{t-1,k}+a^*_{t-1})},$$

where $u_k > 0$ is a tuning parameter related to the prior weights, $\boldsymbol{\eta}_{t-1}$ are learning rates that decrease over time, $\boldsymbol{\mu}_t$ are corrections incrementally computed at every round, and the scalar a^*_{t-1} ensures normalization (see Lemma C.6.7). The weights $\boldsymbol{w}_t$ are reminiscent of those played by the Hedge algorithm, but the normalization a^*_t cannot be computed explicitly in general. The weights $\boldsymbol{w}_t$ are the result of a Follow-the-Regularized-Leader update on a vector of corrected losses $\boldsymbol{L}_{t-1} + \boldsymbol{\mu}_{t-1}$. The regularizer employed is the multiscale entropy: for a fixed $\boldsymbol{u} > 0$, its Bregman divergence is

$$\boldsymbol{w} \mapsto D_{\boldsymbol{\eta}}(\boldsymbol{w}, \boldsymbol{u}) = \sum_{k\in K} w_k \frac{\ln(w_k/u_k) - (1 - u_k/w_k)}{\eta_k}, \quad \boldsymbol{w} \in \mathcal{P}(K) \tag{4.3}$$

[see Bubeck et al., 2019, Chen et al., 2021]. The goal substracting the data-dependent second-order corrections $\boldsymbol{\mu}_t$ from the experts' regrets is to keep a scalar potential function Φ_t negative. Here, the potential $t \mapsto \Phi_t$ is defined by convex conjugacy with respect to the multiscale entropy as

$$\Phi_t := \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_t) = \max_{\boldsymbol{w}\in\mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{R}_t - \boldsymbol{\mu}_t \rangle - D_{\boldsymbol{\eta}_t}(\boldsymbol{w}, \boldsymbol{u}), \tag{4.4}$$

for which $\boldsymbol{w}_{t+1}$ is the maximizer. The corrections $\boldsymbol{\mu}_t$ and the consequent negativity of the potential Φ_t are the main ingredients in the regret analysis of Muscada. We next motivate these choices.

The shape of the corrections $\boldsymbol{\mu}_t$. We designed Muscada to favor experts with low corrected regret $\boldsymbol{R}_t - \boldsymbol{\mu}_t$. For the sake of informal discussion, our goal is to obtain $\mu_{t,k} \approx \sigma_k\sqrt{v_t \ln(1/\pi_k)}$. The algorithm achieves this by additively correcting the regrets in each round. Indeed, from the analysis of entropy-regularized algorithms, one would expect learning rates of the shape $\eta_{t,k} \approx \frac{1}{\sigma_k}\sqrt{\frac{\ln(1/\pi_k)}{v_t}}$ to be optimal. With this learning rates in mind, the desired correction $\boldsymbol{\mu}_t$ can be approximated using a Riemann-sum approximation of $\sqrt{v_t} = \int_0^{v_t} \frac{1}{2\sqrt{v}}\mathrm{d}v$. Indeed, for the conjectured learning rates, our target $\mu_{t,k}$ satisfies $\mu_{t,k} \approx \sigma_k^2 \sum_{s\leq t} \eta_{s-1,k}\Delta v_s$, where $\Delta v_t = v_t - v_{t-1}$. This implies that the choice $\Delta\mu_{t,k} = \sigma_k^2 \eta_{t-1,k}\Delta v_t$ as our per-round additive correction is helpful for achieving our goal. We discuss our precise choice of learning rates after the formal statement of Proposition 4.2.2 below.

Parameters: A vector $u_k > 0$ of initial weights, initial strictly positive learning rates $\eta_{0,k} \le 1/(2\sigma_k)$, and real, continuous nonincreasing functions $H_k : \mathbb{R}^+ \mapsto \mathbb{R}$ with $H_k(0) = 1$. **Initialization:** Let $\mu_{0,k} = 0$, $v_0 = 0$, $R_{0,k} = 0$ and $L_{0,k} = 0$. For each round $t = 1, 2, 3, \ldots$

1. Play (follow the multiscale-entropy regularized leader of the corrected losses)

$$\boldsymbol{w}_t = \operatorname*{arg\,min}_{\boldsymbol{w} \in \mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{L}_{t-1} + \boldsymbol{\mu}_{t-1} \rangle + D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}, \boldsymbol{u}), \tag{4.5}$$

   where $D_{\boldsymbol{\eta}}$ is the multiscale relative entropy given in (4.3).

2. Observe loss ℓ_t. Update $R_{t,k} = R_{t-1,k} + \langle \boldsymbol{w}_t, \ell_t \rangle - \ell_{t,k}$ and $L_{t,k} = L_{t-1,k} + \ell_{t,k}$.

3. Compute Δv_t, the value $\Delta v \ge 0$ such that

$$\Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_{t-1} - \boldsymbol{\sigma}^2 \boldsymbol{\eta}_{t-1} \Delta v, \boldsymbol{\eta}_{t-1}) = \Phi(\boldsymbol{R}_{t-1} - \boldsymbol{\mu}_{t-1}, \boldsymbol{\eta}_{t-1}), \tag{4.6}$$

   where Φ is the potential function defined in (4.4).

4. Compute $\Delta\mu_{t,k} = \sigma_k^2 \eta_{t-1,k} \Delta v_t$. Update $\mu_{t,k} = \mu_{t-1,k} + \Delta\mu_{t,k}$ and $v_t = v_{t-1} + \Delta v_t$.

5. Set the new learning rate $\eta_{t,k} = \eta_{0,k} H_k(v_t)$.

Figure 4.1.: MUSCADA

Negativity of Φ. Our regret bounds are a direct consequence of the negativity of the potential $t \mapsto \Phi_t$. Indeed, by its definition, $\Phi_0 \le 0$, and, because of our choice of nonincresing learning rates and corrections, the change in potential $\Delta\Phi_t = \Phi_t - \Phi_{t-1}$ can be bounded by

$$\Delta\Phi_t \le \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_{t-1}) - \Phi(\boldsymbol{R}_{t-1} - \boldsymbol{\mu}_{t-1}, \boldsymbol{\eta}_{t-1}) = 0,$$

where the last equality follows from (4.6), the choice of corrections $\Delta\boldsymbol{\mu}_t$. This implies the following lemma, of which we give a more general proof in Section C.3.1.

Lemma 4.2.1. The potential $t \mapsto \Phi_t$ starts at $\Phi_0 \le 0$ and is decreasing for $t \ge 0$.

Once we prove that the potential Φ_t is negative, we are ready to derive regret guarantees for MUSCADA. The maximal nature of the potential $t \mapsto \Phi_t$ and its nonpositivity together imply that, *simultaneously* for all distributions $\boldsymbol{p} \in \mathcal{P}(K)$,

$$\langle \boldsymbol{p}, \boldsymbol{R}_t - \boldsymbol{\mu}_t \rangle \le D_{\boldsymbol{\eta}_t}(\boldsymbol{p}, \boldsymbol{u}). \tag{4.7}$$

We choose $\boldsymbol{p}$ concentrated on each expert $k \in K$ to deduce the next proposition (proof in Section C.3.1).

Let $\boldsymbol{\pi} \in \mathcal{P}(K)$ be a probability distribution on K experts.

Tuning 1 Requires $\sigma_{\min} > 0$. Set $u_k = \pi_k \frac{\sigma_{\min}}{\sigma_k}$, $\eta_{0,k} = \frac{1}{2\sigma_{\max}}$, $\gamma_k = 8\frac{\sigma_{\max}^2}{\sigma_k^2}\ln(1/u_k)$ and

$$H_{1,k}(v) = \frac{\mathrm{d}}{\mathrm{d}v}\left[\frac{v}{\sqrt{1+v/\gamma_k}}\right] = \frac{v/\gamma_k+2}{2(1+v/\gamma_k)^{3/2}}.$$

Tuning 2 Set $u_k = \pi_k, \eta_{0,k} = \frac{1}{2\sigma_{\max}}$, $\alpha_k = 32\frac{\sigma_{\max}^2}{\sigma_k^2}$, $\gamma_k = \alpha_k \ln(1/u_k)$ and

$$H_{2,k}(v) = \frac{\mathrm{d}}{\mathrm{d}v}\left[\sqrt{\alpha_k^2\left\{(1+v/\alpha_k)\ln(1+v/\alpha_k) - v/\alpha_k\right\} + \frac{v^2}{2(1+v/(2\gamma_k))}}\right]$$

$$= \frac{\alpha_k \ln(1+v/\alpha_k) + \frac{1}{2}\frac{2v+v^2/(2\gamma_k)}{(1+v/(2\gamma_k))^2}}{2\sqrt{\alpha_k^2\left\{(1+v/\alpha_k)\ln(1+v/\alpha_k) - v/\alpha_k\right\} + \frac{v^2}{2(1+v/(2\gamma_k))}}}.$$

If, for some k, $\sigma_k = 0$, define $H_{2,k}$ to be the limit value $\lim_{\sigma\downarrow 0} H_{2,k}(v_t) = 1$.

Figure 4.2.: Tunings

Proposition 4.2.2. *Assume that the learning rates $t \mapsto \boldsymbol{\eta}_t$ are decreasing.* Muscada *guarantees that, for any $t = 1, 2, 3, \ldots$ and all $k \in K$,*

$$R_{t,k} \le \mu_{t,k} + \frac{\ln(1/u_k)}{\eta_{t,k}} + \sum_{j\in K}\frac{u_j}{\eta_{t,j}} - \frac{1}{\eta_{t,k}}, \tag{4.8}$$

where $\mu_{t,k} = \sigma_k^2 \sum_{s\le t}\eta_{s-1,k}\Delta v_s$. Furthermore, for $\eta_{t,k} = \eta_0 H_k(v_t)$ as in Figure 4.1, $\boldsymbol{\mu}_t$ satisfies

$$\mu_{t,k} \le \sigma_k^2\eta_{0,k}\int_0^{v_t} H_k(x)\mathrm{d}x + \sigma_k^2(\eta_{0,k} - \eta_{t,k})\max_{s\le t}\Delta v_s. \tag{4.9}$$

Choice of learning rates. Proposition 4.2.2 guides us in choosing the learning rates presented in Figure 4.2. The starting value of the learning rates influences our ability to control v_t in terms of the variance of the losses of the algorithm while their behavior for large v_t determines the long-term growth of the regret bounds. The learning rates presented in Figure 4.2 interpolate smoothly between these two regimes by taking the form $\eta_{t,k}^{(1)} = \eta_{0,k}H_{1,k}(v_t)$ and $\eta_{t,k}^{(2)} = \eta_{0,k}H_{2,k}(v_t)$. Here, the starting learning rates are set to $\eta_{0,k} = 1/(2\sigma_{\max})$. The functions $H_{1,k}, H_{2,k} \le 1$ decrease monotonically from their initial values $H_{1,k}(0) = H_{2,k}(0) = 1$ in such a way that, as $v_t \to \infty$,

$$\eta_{t,k}^{(1)} \sim \frac{\sqrt{2}}{\sigma_k}\sqrt{\frac{\ln(1/\pi_k)}{v_t}} \quad \text{and} \quad \eta_{t,k}^{(2)} \sim \frac{\sqrt{2}}{\sigma_k}\sqrt{\frac{\ln(1/\pi_k)+\ln v_t}{v_t}}.$$

The asymptotic expresion for $\eta_{t,k}^{(1)}$ is reminiscent of the optimal learning rates for the Hedge algorithm with the number of rounds t replaced by the refined v_t and the uniform $\ln K$ replaced by $\ln(1/\pi_k)$. Finally, with the Riemann sum bound (4.9) from Proposition 4.2.2 in mind, the learning rates were chosen as the derivatives of functions that will become the dominant term in the regret guarantees.

Tuned regret bounds. The learning rates from Figure 4.2 can be readily used in Proposition 4.2.2 to derive regret guarantees for MUSCADA. However, to facilitate interpretation, we bound the learning rates and their reciprocals in order to obtain the regret bounds contained in the following proposition (proof in Appendix C.3.2). After its statement, we prove Theorem 4.1.1 from the introduction.

Proposition 4.2.3. *Let $\boldsymbol{\pi}$ be a probability distribution on K.*

- MUSCADA *run with Tuning 1 depicted in Figure 4.2 guarantees that, for any $t = 1, 2, \ldots$,*

$$R_{t,k} \le 2\sigma_k\sqrt{2v_t\ln(1/u_k)} + c_{\boldsymbol{\sigma},\boldsymbol{\pi}}\sigma_{\min}\sqrt{2v_t} + 8\sigma_{\max}\ln(1/u_k) + 4\sigma_{\max} + \frac{\sigma_k}{2}\max_{s\le t}\Delta v_s, \tag{4.10}$$

where the constant $c_{\boldsymbol{\sigma},\boldsymbol{\pi}} = \sum_{k\in K}\pi_k(1/\sqrt{\ln(1/u_k)})$ and $u_k = \pi_k\frac{\sigma_{\min}}{\sigma_k}$.

- MUSCADA *run with Tuning 2 depicted in Figure 4.2 guarantees that, for any $t = 1, 2, \ldots$,*

$$R_{t,k} \le 2\sigma_k\sqrt{2v_t\left(\ln\left(1 + \tfrac{\sigma_k^2 v_t}{32\sigma_{\max}^2}\right) + \ln(1/\pi_k)\right)} + \sigma_k\ln(1/\pi_k)Z_k + \sum_{j\in K}\pi_j\sigma_j Z_j + \tfrac{\sigma_k}{2}\max_{s\le t}\Delta v_t, \tag{4.11}$$

where $Z_k = \sqrt{\frac{v_t}{2\ln\left(1+\frac{\sigma_k^2 v_t}{32\sigma_{\max}^2}\right)}}\left(1 + \sqrt{\frac{\min\{\ln(1/\pi_k), \frac{\sigma_k^2 v_t}{16\sigma_{\max}^2}\}}{\ln\left(1+\frac{\sigma_k^2 v_t}{32\sigma_{\max}^2}\right)}}\right) = O\left(\sqrt{\frac{v_t}{\ln v_t}}\right)$ *as $v_t \to \infty$.*

Proof of Main Theorem 4.1.1. With Proposition 4.2.3 at hand, we can prove the claims made in Section 4.1.1. Use the fact that $\sigma_{\min} \le \sigma_k$ to conclude from (4.10) that, as $t \to \infty$,

$$R_{t,k} \le 2\sigma_k\sqrt{2v_t\ln(1/u_k)} + 2c_{\boldsymbol{\sigma},\boldsymbol{\pi}}\sigma_k\sqrt{2v_t} + O(1).$$

We can bound $c_{\boldsymbol{\sigma},\boldsymbol{\pi}}/\sqrt{\ln(1/u_k)} \le 1/\ln(1/\pi_{\max})$, where $\pi_{\max} = \max_{k\in K}\pi_k$. Consequently,

$$R_{t,k} \le 2\sigma_k\left\{1 + 1/(2\ln(1+\varepsilon))\right\}\sqrt{2v_t\ln(1/u_k)} + O(1)$$

as $t \to \infty$ any time that $\pi_{\max} = 1 - \varepsilon$. This coincides with (4.1). Similarly, (4.11) implies (4.2). □

4.2.1. Closed-form solutions in the single-scale uniform-prior case

To help in the interpretation and to illustrate the challenges of the multiscale problem, we instantiate Muscada to a situation where all calculations can be carried out in closed form: when all scales are the same and equal to σ, and the initial weights $\boldsymbol{\pi}_{\text{Unif}}$ are uniform; $\pi_{\text{Unif},k} = 1/K$. This is the setting in which AdaHedge by De Rooij et al. [2014] operates. In this case, the learning rates and corrections of Muscada are the same for all experts; $\eta_{t,k} = \eta_t$ and $\Delta\mu_{t,k} = \Delta\mu_t$. The potential Φ_t and the corrections $\Delta\mu_t$ take the familiar form

$$\Phi_t = \frac{1}{\eta_t}\ln\left(\frac{1}{K}\sum_{k\in K} \mathrm{e}^{\eta_t(R_{t,k}-\mu_{t,k})}\right), \qquad \text{and} \qquad \Delta\mu_t = \frac{1}{\eta_{t-1}}\ln\sum_{k\in K} w_{t,k}\mathrm{e}^{\eta_{t-1}(\langle \boldsymbol{w}_t,\boldsymbol{\ell}_t\rangle-\ell_t)}.$$

These two quantities play a central role in the analysis of AdaHedge, where De Rooij et al. [2014] called $\Delta\mu_t$ the *mixability gap*, the difference between the average $\langle \boldsymbol{w}_t, \boldsymbol{\ell}_t\rangle$ and the *mixed average* $-\frac{1}{\eta_{t-1}}\ln\sum_{k\in K} w_{t,k}\mathrm{e}^{-\eta_{t-1}\ell_{t,k}}$. The main quantity in our analysis, Δv_t, becomes

$$\Delta v_t = \frac{1}{\eta_{t-1}^2\sigma^2}\ln\sum_{k\in K} w_{t,k}\mathrm{e}^{\eta_{t-1}(\langle \boldsymbol{w}_t,\boldsymbol{\ell}_t\rangle-\ell_{t,k})}.$$

Using well-known estimates for cumulant generating functions, Δv_t can be bounded by the ratio $\operatorname{Var}_{\boldsymbol{w}_t}(\boldsymbol{\ell}_t)/\sigma^2$. Indeed, Hoeffding's inequality implies the worst-case bound $\Delta v_t \le \frac{1}{2}$; Bernstein's, the second-order $\Delta v_t \lesssim \operatorname{Var}_{\boldsymbol{w}_t}(\boldsymbol{\ell}_t)/\sigma^2$. Since it is v_t that appears in the regret bounds in Proposition 4.2.3, they are a refinement over those of AdaHegde[2]. Additionally, the present analysis yields improvements that are apparent in lower-order terms. Indeed, the last two terms in the regret bound (4.8) in Proposition 4.2.2 vanish, and the analysis used in the proof of Proposition 4.2.3 with $\eta_0 = \sqrt{2}/\sigma$ and the instantiation of H_1 from Figure 4.2, $H_1(x) = \frac{x/\ln(K)+2}{2(1+x/\ln(K))^{3/2}}$, give the regret bound

$$\mathcal{R}_t \le \begin{cases} c_1\sigma v_t + c_2\sigma\ln K + \sigma/2 & \text{if } v_t \le \ln K,\\ 2\sigma\sqrt{2v_t\ln K} + \sigma/2 & \text{if } v_t > \ln K \end{cases}$$

with $c_1 = 3/\sqrt{2}$ and $c_2 = 1/\sqrt{2}$. Unfortunately, multiscale analogs of Bernstein and Hoeffding's inequalities on Δv_t are not available; considerably more technical work needs to be carried out to prove Theorem 4.1.2. A multiscale analog of Bernstein's estimate for Δv_t is only available when all the learning rates are smaller than $1/(2\sigma_{\max})$ (see the proof of Theorem 4.1.2 in Appendix C.7).

4.3. Multiscale Stochastic Luckiness

In this section we show, under easiness conditions, that the expected pseudoregret of Muscada is constant. Assume that the loss vectors $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \ldots$ are i.i.d. and are generated according to a distribution $\mathbf{P}$ that satisfies Massart's easiness condition (see

[2] Our algorithm with learning rate tuning function $H(v) = \sqrt{\frac{\ln K}{4v}}$ comes closest to AdaHedge.

Definition 4.1.3). For Tuning 1, assume that the minimum scale among experts $\sigma_{\min}$ is strictly positive. The analysis technique in this case is similar to that of Koolen et al. [2016] with an extra step. A use of Theorem 4.1.2 shows that Δv_t can be estimated in terms of $\mathrm{Var}_{\boldsymbol{w}_t}(\ell_t)$. This estimate possibly incurs in a multiplicative factor that can be as high as $1/\sigma^2_{\min}$. There are examples for which this constant is necessary (not shown). After this, standard arguments show that the expected pseudoregret is constant. See Appendix C.5 for proofs.

Theorem 4.3.1. *Under Massart's condition and using Tuning 1 from Figure 4.2, the expected pseudoregret of* Muscada *is bounded by a constant in the number of rounds. Specifically, for any $t \geq 0$,*

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \lesssim a^2 c_{\mathrm{M}} + b$$

with $a = \sqrt{2 \max_{i,j \in K} \left\{ \frac{1}{\sigma_i \sigma_j} \frac{\ln\left(\frac{1}{\pi_i}\right) + \ln\left(\frac{\sigma_i}{\sigma_{\min}}\right)}{\ln\left(\frac{1}{\pi_j}\right) + \ln\left(\frac{\sigma_j}{\sigma_{\min}}\right)} \right\}} \left(4\sigma_{k^*} \sqrt{2 \ln\left(\frac{1}{u_{k^*}}\right)} + 2\sqrt{2} c_{\boldsymbol{\sigma},\boldsymbol{\pi}} \sigma_{\min} \right)$ *and* $b = 8\sigma_{\max} \ln(\frac{1}{u_{k^*}}) + 4\sigma_{\max} + 2\sigma_{k^*}$.

For Tuning 2, where we do not assume that $\sigma_{\min} > 0$, still $\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \lesssim 1$ using a different proof technique. Using the expression for the weights of the algorithm, we show that they concentrate on the best expert k^*. The analysis here is similar to that of Mourtada and Gaïffas [2019], but the lack of an expression for the normalizing a^*_t presents with an additional technical difficulty. The result is the following theorem.

Theorem 4.3.2. *Let $d_k = \mathbf{E}_{\mathbf{P}}[\ell_{t,k} - \ell_{t,k^*}]$ and assume that $\min_{k \neq k^*} d_k > 0$. Using Tuning 2 in Figure 4.2,* Muscada *guarantees constant expected pseudoregret. Specifically,*

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \leq \sum_{k \in K} f(d_k), \qquad \text{where} \qquad f(d) = O\left(\frac{\sigma^2_{\max}}{d} \ln\left(\frac{\sigma^2_{\max}}{d^2} \right) \right) \text{ as } d \to 0.$$

Standard modifications of the arguments presented may be used to prove that the pseudoregret is constant with $\mathbf{P}$-high probability (not shown).

4.4. Optimism

In this section we show an optimistic variant of Muscada. Suppose that, before round t, we count on guesses $\boldsymbol{m}_t$ for what ℓ_t will be. Assume that $\boldsymbol{m}_t$ is of the same scale as ℓ_t, that is, $|m_{t,k}| \leq \sigma_k$. In particular, this entails that $|\ell_{t,k} - m_{t,k}| \leq 2\sigma_k$. A modification of Muscada, presented in Figure 4.1, puts these guesses to good use. These modifications allow for regret guarantees similar to those contained in Proposition 4.2.3, but in this case $\Delta v^\circ_t \lesssim \mathrm{Var}_{\tilde{\boldsymbol{w}}^\circ_t}(\ell_t - \boldsymbol{m}_t)/\langle \tilde{\boldsymbol{w}}^\circ_t, \boldsymbol{\sigma}^2 \rangle$, where the superscript $\circ$ signals the optimistic analogs of the quantities from Muscada. These modifications are shown in Figure 4.3 and the regret bounds in the following proposition (proofs in Appendix C.4).

Proposition 4.4.1. *If $t \mapsto v^\circ_t$ is the variance process defined by Optimistic* Muscada *in Figure 4.3, the same regret bounds presented Proposition 4.2.3 hold with two modifications: v°_t instead of v_t and all scales doubled, that is, $2\boldsymbol{\sigma}$ instead of $\boldsymbol{\sigma}$. Furthermore, for each $t = 1, 2, \ldots$, $\Delta v^\circ_t \leq 4\,\mathrm{Var}_{\tilde{\boldsymbol{w}}^\circ_t}(\ell_t - \boldsymbol{m}_t)/\langle \tilde{\boldsymbol{w}}^\circ_t, \boldsymbol{\sigma}^2 \rangle \leq 4t$, where $\tilde{w}^\circ_{t,k} \propto w^\circ_{t,k} \eta_{t-1,k}$.*

1’ Compute the guess $\boldsymbol{m}_t$ and play

$$\boldsymbol{w}_t^\circ = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{L}_{t-1} + \boldsymbol{m}_t + \boldsymbol{\mu}_{t-1}\rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}, \boldsymbol{u}).$$

3’ Let Δv_t° be the value $\Delta v^\circ \geq 0$ such that

$$\Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_{t-1} - \boldsymbol{\eta}_{t-1}\boldsymbol{\sigma}^2 \Delta v^\circ, \boldsymbol{\eta}_{t-1}) = \Phi(\boldsymbol{R}_{t-1} + \langle \boldsymbol{w}_t^\circ, \boldsymbol{m}_t\rangle - \boldsymbol{m}_t - \boldsymbol{\mu}_{t-1}, \boldsymbol{\eta}_{t-1}). \quad (4.12)$$

Tuning 1’ and Tuning 2’. As in Figure 4.2 but with halved starting learning rate $\eta_{0,k} = \frac{1}{4\sigma_{\max}}$.

Figure 4.3.: Optimistic Muscada, given as update w.r.t. Figure 4.1.

4.5. Computation

At each round, Muscada requires two computations. We now argue that both can be executed to machine precision in $O(K)$ time. First, computing the weights (4.5) given the losses $\boldsymbol{L}_{t-1}$ and correction terms $\boldsymbol{\mu}_{t-1}$ can be reduced, by Lemma C.6.6, to a single scalar convex minimization problem. Cancelling the derivative of the objective amounts to searching for the normalizing offset a_t. To that end, binary search to machine precision takes $O(K)$ time per round. Notice that this also allows us to compute the potential value. Second, for computing the variance contribution (4.6), we observe that the right hand side of (4.6) is decreasing in Δv_t. Since the potential can be computed in $O(K)$ time, we can use an outer binary search to compute Δv_t to machine precision in $O(K)$ time as well. Alternatively, Newton’s method may be employed; both of the previous problems require finding a root of a convex function. When deferring to a convex optimization library, a convenient expression is the jointly convex minimization form (see Lemma C.6.6)

$$\Delta v_t \;=\; \inf_{a,\Delta v} \Delta v \quad \text{subject to} \quad a + \sum_{k\in K} w_{t,k} \frac{\mathrm{e}^{\eta_{t-1,k}(\langle \boldsymbol{w}_t, \boldsymbol{\ell}_t\rangle - \ell_{t,k} - a) - \eta_{t-1,k}^2 \sigma_k^2 \Delta v} - 1}{\eta_{t-1,k}} \;\leq\; 0.$$

4.6. Experiments on Synthetic Data

We investigate the performance of our multiscale method on two experiments: one for illustrating the performance of Muscada under Massart’s condition, another for solving multiscale two-player zero-sum games.

The aim of the first experiment is to compare the performance of Muscada in easy and hard stochastic data sequences. To this end, we compared a sequence of hard stochastic data with no gap vs. easy data sampled i.i.d. from a distribution

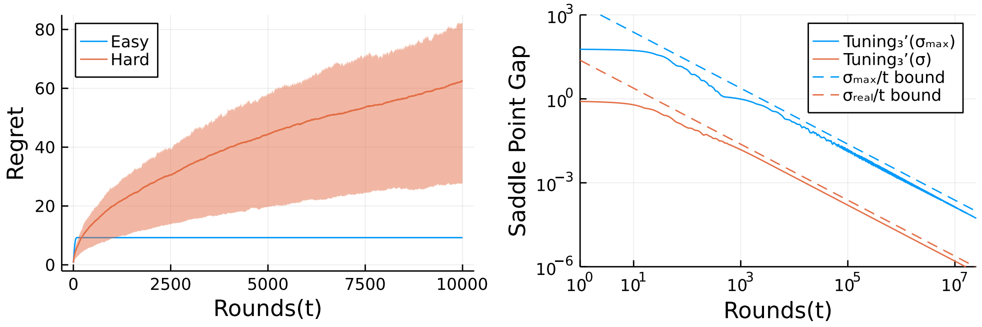

Figure 4.4.: Left: empirical mean and quartiles of 2000 realizations of the regret $t \mapsto R_{t,k^*}$ of Muscada. For easy i.i.d. Massart distribution, the regret is constant; for a hard distribution without a gap, $\Omega(\sqrt{t})$. Right: optimistic Muscada (solid red) achieves an iterate-average saddle-point gap of σ_{real}/t where $\sigma_{\text{real}} = \sigma_{\max}/100$ is the relevant scale of the Nash equilibrium. Other methods scale as $\sigma_{\max}/t$.

satisfying Massart's condition. We witnessed constant regret for the easy data, as shown in Figure 4.4 (Left). We take $K = 50$ experts and set $\sigma_k = 1/k$ for each $k \in K$. To generate our data, we fix some mean $\lambda_k \in [-\sigma_k, \sigma_k]$ and generate binary expert losses $\ell_{t,k} \in \{-\sigma_k, +\sigma_k\}$ independently between rounds and experts, with probability $\mathbf{P}\{\ell_{t,k} = \sigma_k\} = \frac{\sigma_k+\lambda_k}{2\sigma_k}$. For the hard case, we set $\lambda_k = 0$ for all k. For the lucky case, we set $\lambda_2 = -1/5$ instead. Generating this figure with the code in the supplementary material takes 3 seconds on an Intel i7-7700 processor.

The aim of the second experiment is to show the performance of Muscada for solving multiscale zero-sum games. Here, the payoff matrix is unknown, but row and column scales are available and vastly different. As detailed in Appendix C.1, we run two instances of appropriately tuned Optimistic Muscada against each other. As shown in Figure 4.4 (Right), the pair of time-average iterates converges to the saddle point with a suboptimality gap of order σ_{real}/t instead of the worst-case $\sigma_{\max}/t$, where σ_{real} is the maximum range within the support of the saddle point. In Appendix C.1, we conjecture that this rate holds for any such game and prove a weaker result: without optimism, the slower but scale-adaptive rate $\sigma_{\text{real}}/\sqrt{t}$ is achieved.

4.7. Discussion

We developed a new algorithm for multiscale online learning that is both worst-case safe and achieves constant pseudoregret in stochastic lucky cases. Our method is a refinement of the Follow-the-Regularized-Leader template with a weighted entropy. The main innovation is in the correction terms added to the losses, which are the tightest the technique admits. This suggests that these variance-like terms are in fact intrinsic to the problem of obtaining scale-dependent regret bounds. Lastly, we relate this newfound variance to the variance asked for by Freund [2016], we comment on the advantage of second-order guarantees over zeroth-order ones, and we state an open problem.

Quantile bounds and solving Freund's problem. Freund [2016] asked whether quantile adaptivity and variance adaptivity are compatible, that is, whether one can have $\langle \boldsymbol{p}, \boldsymbol{R}_t \rangle \le \sqrt{\mathrm{KL}(\boldsymbol{p}, \boldsymbol{u}) \sum_{s \le t} \mathrm{Var}_{\boldsymbol{w}_s}(\ell_s)}$ for all comparator distributions $\boldsymbol{p} \in \mathcal{P}(K)$ simultaneously. Even though our tuning of $\boldsymbol{\eta}_t$ does not yield quantile bounds, these can, however, be added employing a now-standard method [Koolen and van Erven, 2015]. Namely, instead of only including every expert with a private learning rate tuned to its prior complexity level (the typical $\ln K$ or $\ln(1/\pi_k)$ term), we include multiple copies of each expert, each with a learning rate tuned to a smaller complexity level. We then start from (4.7) with comparator distribution $\boldsymbol{p}$ concentrated on the ε-quantile of interest and carry out all future steps (from Proposition 4.2.2 on), ending up with the quantile regret bound $\langle \boldsymbol{p}, \boldsymbol{R}_t \rangle \le \max_{k : p_k > 0} \sigma_k \sqrt{v_t(\ln C + D_{\boldsymbol{\eta}_0}(\boldsymbol{p}, \boldsymbol{u}))}$, where C is the number of learning rates thus created. As these learning rates can be exponentially spaced in an interval of width $\ln K$, C is of order $\ln \ln K$. Does this procedure answer Freund's question? For our notion of variance, v_t, which our results suggest is a rather useful notion, the answer is yes. However, to relate Δv_t to $\mathrm{Var}_{\boldsymbol{w}_t}(\ell_t)$, we incur a multiplicative ratio $\eta_{t,\max}/\eta_{t,\min}$, which, for the quantile case, is of order $\sqrt{\ln K}$, turning the prior-in-the-square-root bound into a prior-outside-the-square-root bound. The latter was already achievable by not tuning $\boldsymbol{\eta}$ to the prior complexities at all. This problem does not arise in the same-scale uniform-prior case; there, Δv_t is bounded by a small multiple of $\mathrm{Var}_{\boldsymbol{w}_t}(\ell_t)$ [De Rooij et al., 2014]. Note that this problem is present even when K is fixed while t grows, which is narrowly outside the scope of the impossibility results of Marinov and Zimmert [2021]. This discussion sheds light from another angle on why Freund's problem is hard; we present a desirable multiscale alternative.

Luckiness, gap, and Massart's condition. We now address the advantage of MUSCADA's refined second-order measure of time v_t over the zeroth-order number of rounds t. Multiscale zeroth-order regret bounds (growing with t) can be guaranteed either by tuning MUSCADA crudely to a constant multiple of t or by building an any-time improvement of the algorithm of Bubeck et al. [2019], also tuned to t. Both t-tuned and v_t-tuned algorithms have constant expected pseudoregret in stochastic lucky cases, but the constant can be widely different. Indeed, the constant for t-tuned algorithms scales with the inverse $1/d_{\min}$ of the gap $d_{\min} = \min_{k \ne k^*} \mathbf{E}[\ell_{t,k} - \ell_{t,k^*}]$, while the constant for v_t-tuned algorithms scales with the constant c_{M} from Massart's condition (see Definition 4.1.3). The difference stems from the fact that c_{M} is at most $1/d_{\min}$, but it can be arbitrarily smaller. This separation appears to be fundamental. In the single-scale uniform-prior case, the above t-tuned algorithms are closely related to Decreasing Hedge [Mourtada and Gaïffas, 2019], just as MUSCADA is related to AdaHedge (see Section 4.2.1). Mourtada and Gaïffas [2019] show that, in the single-scale case, even under Massart's condition with $c_{\mathrm{M}} = 1$, Decreasing Hedge and, consequently, Bubeck et al.'s algorithm with decreasing learning rates, has expected pseudoregret

$\mathbf{E}[R^B_{t,k}] \gtrsim 1/d_{\min}$. If the smallest scale $\sigma_{\min} > 0$, by taking $d_{\min}$ small, this lower bound can be made arbitrarily worse than the guarantee of Muscada, $\mathbf{E}[R^{\textsc{Muscada}}_{t,k^*}] \lesssim c_{\mathrm{M}} + 1$, from Theorem 4.3.1.

Open problem. Our ability to incorporate an arbitrary prior suggests that the results should extend to countably many experts. However, the current techniques do break down. When $\max_{k\in\mathbb{N}} \sigma_k < \infty$ Muscada with Tuning 1 (if $\inf_{k\in\mathbb{N}} \sigma_k > 0$) or Tuning 2 would still deliver the worst-case bound. Yet our luckiness result currently requires $\max_{k,l,t} \frac{\eta_{t,k}}{\eta_{t,l}\sigma_l^2} < \infty$. Even with a common scale σ, this is never the case due to the dependence of $\boldsymbol{\eta}_t$ on the prior $\boldsymbol{\pi}$, which is necessarily decreasing. Is luckiness actually possible, for example, in the online learning analog of the elegant challenge example presented by Talagrand [2014, Chapter 2]?

5. Exponential Stochastic Inequality[1]

We develop the concept of *exponential stochastic inequality* (ESI), a novel notation that simultaneously captures high-probability and in-expectation statements. It is especially well suited to succinctly state, prove, and reason about excess-risk and generalization bounds in statistical learning; specifically, but not restricted to, the PAC-Bayesian type. We show that the ESI satisfies transitivity and other properties which allow us to use it like standard, nonstochastic inequalities. We extend to a large degree the original definition from 2016 and show that general ESIs satisfy a host of useful additional properties, including a novel Markov-like inequality. We show how ESIs relate to, and clarify, PAC-Bayesian bounds, subcentered subgamma random variables and *fast-rate conditions* such as the central and Bernstein conditions. We also show how the ideas can be extended to random scaling factors (learning rates).

5.1. Introduction

Let X, Y be two random variables. For fixed $\eta > 0$, we define

$$X \trianglelefteq_\eta Y \text{ if and only if } \mathbf{E}[\mathrm{e}^{\eta(X-Y)}] \leq 1. \tag{5.1}$$

If $X \trianglelefteq_\eta Y$ we say that X *is stochastically exponentially smaller than* Y, and we call a statement of the form $X \trianglelefteq_\eta Y$ an *Exponential Stochastic Inequality* or *ESI* (pronounce as "easy").

The ESI is a useful tool to express certain nonasymptotic probabilistic concentration inequalities, and generalization and excess risk bounds in statistical learning, especially but not exclusively of the *PAC-Bayesian* kind—it allows theorems to be stated more succinctly and their proofs to be simultaneously streamlined, clarified and shortened. This is enabled by the ESI's two main characteristics: first, the ESI simultaneously expresses that random variables are ordered both in expectation and with high probability—consequences of Jensen's and Markov's inequality, respectively. Indeed, if $X \trianglelefteq_\eta Y$ then both

$$\text{(a) } \mathbf{E}[X] \leq \mathbf{E}[Y] \text{ and (b), with probability at least } 1-\delta, X \leq Y + \frac{\ln(1/\delta)}{\eta}, \tag{5.2}$$

[1] This chapter is based on P. D. Grünwald, M. F. Pérez-Ortiz, and Z. Mhammedi. Exponential Stochastic Inequality, Apr. 2023. URL `http://arxiv.org/abs/2304.14217`. arXiv: 2304.14217 [math, stat]

for all $0 < \delta \leq 1$—this is formalized more generally in Proposition 5.2.3. These simultaneous inequalities are in contrast with considering either ordering in probability or in expectation separately: it is easy to construct random variables that are ordered in expectation but not with high probability and vice versa. The second main characteristic of the ESI is that it satisfies a useful transitivity-like property. As shown in Section 5.2.4 below, if separately and with high probability $X \leq Y$ and $Y \leq Z$, the common technique of applying the union bound to obtain a high-probability statement for $X \leq Z$ would lead to slightly worse bounds than using ESI transitivity. ESI notation was originally introduced by Koolen et al. [2016] and Grünwald and Mehta [2020] (the first arXiv version of which came out in 2016) to improve precisely such chained bounds and to avoid stating essentially the same statement twice, once in probability and once in expectation—both statements were highly relevant in the context of the latter article. A third reason was that the bounds from Grünwald and Mehta [2020] often involved annealed expectations (normalized log-moment generating functions, see the next section), and writing them out explicitly would require unwieldy nested statements like $\mathbf{E}[\exp(\mathbf{E}(\exp(\eta(...))))] \leq 1$, as can be found in for instance the pioneering work of Zhang [2006a]. ESI notation makes such expressions much more readable by expressing the outer expectation as an ESI, and the inner one as an annealed expectation (as defined in the next section). The ESI was later used in several follow-up articles [Mhammedi et al., 2019, Grünwald and Mehta, 2019, Grünwald et al., 2021], but its properties were never spelled out fully and in much detail.

This chapter gives a detailed development of the ESI. We extend its definition and notation to cover many more cases, making a novel distinction between "weak" and "strong" ESI. We provide a list of useful properties—a calculus as it were—that can be used for manipulating ESIs. Our purpose is twofold: first, we want to showcase the ease and advantages of working with the ESI; second, we derive some new technical results—that are very conveniently expressed using the ESI—that provide a characterization of classical *subcentered random variables that are subgamma on the right* (which have been well studied before, e.g. Boucheron et al. [2013]) and of the main *fast-rate conditions* in statistical learning theory, the *Bernstein* and *central* conditions, extending results of Van Erven et al. [2015] to unbounded random variables. We find that such conditions only require exponential-moment control on one tail; only minimal control—of the first and second moments—for the other tail.

In the remainder of this introduction, we give a brief overview of what is to come, starting with the generalized definition of ESI. As a running example, we use the determination of stochastic bounds on averages of i.i.d. random variables. We say that $u : \mathbb{R}^+ \to \mathbb{R}^+$ is an *ESI function* if it is continuous, nondecreasing, and strictly positive.

Definition 5.1.1 (ESI). Let u be an ESI function u—continuous, nondecreasing and strictly positive. We define

$$X \trianglelefteq_u Y \text{ if and only if for all } \epsilon > 0, \ \mathbf{E}[\mathrm{e}^{u(\epsilon)\cdot(X-Y)}] \leq \mathrm{e}^{u(\epsilon)\cdot\epsilon}. \tag{5.3}$$

This definition entails that, using the original ESI notation (5.1), for all $\epsilon > 0$, if $\eta = u(\epsilon)$, then $X \trianglelefteq_\eta Y + \epsilon$. Henceforth, we shall refer to the original type of ESI

(5.2) as *strong ESI* and to the new form (5.3) simply as *ESI.* The strong ESI is a special instance of the ESI, as can be seen by taking the constant function $u(\epsilon) \equiv \eta$ in (5.3). In the special case that $\lim_{\epsilon \downarrow 0} u(\epsilon) = 0$, we shall refer to $X \trianglelefteq_u 0$ as a *weak* ESI. The main reason for introducing a general ESI is that it allows us to extend all major useful properties of the strong ESI to the weak ESI, which provides a weaker exponential right-tail control than the strong ESI and thus hold more often in practice. We will consistently use Greek letters (usually η) to refer to constants, i.e. strong ESIs, and Latin letters (usually u) to refer to functions, i.e. general ESIs. The most basic properties of the general ESI as well as fully precise definitions of all notations are given in Section 5.2.

Transitivity, summation and averaging. As we mentioned earlier, a key property of the strong ESI is its transitivity-like property, which leads to sharper bounds than those obtained through the union bound. This property is a consequence of the fact that strong ESIs are preserved under summation, and general ESIs under averaging (Section 5.2.4, Proposition 5.2.6, Corollary 5.2.7). To demonstrate the latter property, let $\{X_f : f \in \mathcal{F}\}$ be a family of random variables and let $X_{f,1}, \ldots, X_{f,n}$ be i.i.d. copies of each X_f. Suppose we are given the ESIs

$$X_{f,i} \trianglelefteq_u 0 \text{ for all } f \in \mathcal{F} \text{ and } i \in [n]. \tag{5.4}$$

Then, we can conclude via Corollary 5.2.7, for all $f \in \mathcal{F}$, that

$$\frac{1}{n} \sum_{i=1}^{n} X_{f,i} \trianglelefteq_{n \cdot u} 0. \tag{5.5}$$

This does not only imply that $\mathbf{E}[\sum X_{f,i}] \leq 0$, but also the high-probability statement that for all $0 < \delta \leq 1$,

$$\frac{1}{n} \sum_{i=1}^{n} X_{f,i} \leq \inf_{\epsilon > 0} \left(\epsilon + \frac{\ln(1/\delta)}{n \cdot u(\epsilon)} \right). \tag{5.6}$$

Additionally, the ESI (5.4) implies that all the moments of the right tail of each $X_{f,i}$ are finite. Under the quite weak condition that the $X_{f,i}$ also have uniformly bounded second moment on the left tail, we can infer via Proposition 5.3.1 in Section 5.3.1—under the assumption that (5.4) holds for some common ESI function u—that they also satisfy a (weak) ESI for a function $u(\epsilon) = C^* \epsilon \wedge \eta^*$ for some $C^*, \eta^* > 0$. Thus, without loss of generality, we can take a u that is linear near the origin. We can then see that for large enough n, the minimum in (5.6) is achieved at an ϵ with $u(\epsilon) = C^* \epsilon < \eta^*$. In that case, the infimum can be computed through differentiation and (5.6) becomes

$$\frac{1}{n} \sum_{i=1}^{n} X_{f,i} \leq c \cdot \left(\frac{\ln(1/\delta)}{n} \right)^{\alpha} \tag{5.7}$$

for some $c > 0$ and $\alpha = 1/2$, a standard bound in statistical learning theory [Vapnik, 1998, Shalev-Shwartz and Ben-David, 2014]. In Section 5.3.1 (Proposition 5.3.1), we give a number of equivalent characterizations of the general ESI in terms of *subcentered, subgamma random variables* of which the result that "u can be taken linear near the origin" is just one instance.

From weak to strong ESI: excess risk bounds. The transitivity property also allows us to prove fast rates of convergence of empirical averages to their expected value. This is of particular interest, as we will recall, for proving excess risk bounds of Machine Learning algorithms. Now we consider $\{X_f : f \in \mathcal{F}\}$ that all satisfy the ESI $X_f \trianglelefteq_u 0$ for a common ESI function u of the form $u(\epsilon) = C^* \epsilon^\gamma \wedge \eta^*$ with $0 \leq \gamma \leq 1$ and C^*, η^* positive constants. Again, for large enough n, the minimum in (5.6) is achieved at an ϵ with $u(\epsilon) < \eta^*$, and differentiation gives that (5.7) now holds with $\alpha = 1/(1+\gamma)$. If $\gamma < 1$, we say that the average satisfies a *fast-rate* statement. To see why, we briefly need to explain one of the most important applications of the ESI, namely, providing *excess-risk bounds* in statistical learning theory [Zhang, 2006b,a, Grünwald and Mehta, 2020]. Here, we assume that there is an underlying sequence of i.i.d. *data* $Z_1, \ldots, Z_n$, each Z_i having the same distribution as Z. Each $f \in \mathcal{F}$ represents a *predictor*, and there is a loss function $\ell_f(Z) \in \mathbb{R}$ quantifying the loss that the predictor f makes on Z. Often, Z is of the form $Z = (U, Y)$, and f represents a function mapping covariates—or features—U to $Y \subset \mathbb{R}$. An example of this setup is regression with the squared error loss $\ell_f((U,Y)) = \frac{1}{2}(Y - f(U))^2$. One can fit other prediction and inference problems such as classification and density estimation into this framework as well [Van Erven et al., 2015, Grünwald and Mehta, 2020]. We now define the *excess loss* that the predictor f makes on the outcome Z as $L_f = L_f(Z) = \ell_f(Z) - \ell_{f^*}(Z)$ where f^* is the minimizer of $f \mapsto \mathbf{E}[\ell_f(Z)]$ over Z—for simplicity, we assume f^* to exist and be unique. Thus, L_f measures how much better or worse f performs compared to the theoretically optimal f^* on a particular Z. Based on a sample $Z^n = (Z_1, \ldots, Z_n)$, learning algorithms output an “estimate” or “learned predictor” $\hat{f} \coloneqq \hat{f}|Z^n$, the latter notation indicating the dependence of $\hat{f}$ on Z^n. Sometimes, e.g. in Bayesian and PAC-Bayesian inference (see below), they output, more generally, a learned distribution $\hat{\Pi} = \hat{\Pi}|Z^n$ on $f \in \mathcal{F}$. The goal is to design an algorithm whose *excess risk* $\mathbf{E}_{Z\sim\mathbf{P}}[L_{\hat{f}|Z^n}(Z)]$ (or $\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\mathbf{E}_{Z\sim\mathbf{P}}[L_{\bar{f}|Z^n}(Z)]$ if the algorithm outputs a distribution) converges to zero fast, with high probability and/or in expectation. To this end, it is crucial to control how fast the *empirical excess risk* $n^{-1}\sum_{i=1}^n L_{f,i}$ (where $L_{f,i} = \ell_f(Z_i) - \ell_{f^*}(Z_i)$) of each fixed $f \in \mathcal{F}$ converges to its expectation $\mathbf{E}[L_f]$. In practice, in simple cases (e.g. bounded losses) the collection of negative excess risks $\{X_f : f \in \mathcal{F}\}$ with $X_f = -L_f$ satisfies a weak ESI, so that (5.7) holds with $\alpha = 1/2$—in line with what one might expect from the central limit theorem. However, in many interesting cases (e.g. bounded squared error loss), something better (larger α) can be attained, because (5.6) holds, for all $f \in \mathcal{F}$, with $u(\epsilon) = C^*\epsilon^\gamma \wedge \eta^*$ for a $\gamma < 1$ (in the specific case of bounded squared error loss it even holds with $\gamma = 0$). Then (5.7) implies that, for each individual f, $n^{-1}\sum_{i=1}^n L_{f,i} = O(n^{-\alpha})$ with $\alpha = 1/(1+\gamma)$, and this usually translates into learning algorithms that also converge at this fast (i.e., faster than $1/\sqrt{n}$, since $\gamma > 0$) rate; an example for empirical risk minimization (ERM) is given below.

Using different terminology and notation (not ESI), Van Erven et al. [2015] already identified that collections $\{L_f : f \in \mathcal{F}\}$ such that all $X_f = -L_f$ satisfy $X_f \trianglelefteq_u 0$ for $u(\epsilon) = C^*\epsilon^\gamma \wedge \eta^*$—as above—allow for fast rates; in their terminology, such a family satisfies the *u-central fast-rate condition.* They showed that, for bounded loss functions (and hence uniformly bounded L_f), satisfying this property for some γ is equivalent to $\mathcal{F}$

satisfying the celebrated *β-Bernstein* condition, with $\beta = 1-\gamma$. The Bernstein condition [Audibert, 2004, Bartlett and Mendelson, 2006, Audibert, 2009] is a more standard, well-known condition for fast rates. Van Erven et al. [2015] left open the nagging question whether the Bernstein and central fast-rate conditions remain equivalent for unbounded loss functions. As one of the main results in this chapter, we show in Theorem 5.3.11 (Section 5.3.2) that this is indeed the case as long as the left tail of the excess risk is exponentially small, and the right tail satisfies a mild condition on its second moment.

PAC-Bayesian bounds. The ESI is particularly well suited to PAC-Bayesian analysis. To demonstrate this, we continue to assume that there are i.i.d. random variables $Z_1, Z_2 \ldots, Z_n$ such that, for all $f \in \mathcal{F}$, $X_{f,i} = g_f(Z_i)$, that is, $X_{f,i}$ can be written as a function of Z_i for some function g_f which may, but does need to be a negative excess loss (in fact, in many applications it will be an expected loss minus an absolute, non-excess empirical loss; see e.g. Grünwald et al. [2021]). We can easily combine the ESIs as (5.4) into a statement that simultaneously involves all $f \in \mathcal{F}$ by using *PAC-Bayesian* bounds [see Catoni, 2007, McAllester, 1998, Van Erven, 2014, Guedj, 2019, Alquier, 2023]. As we show in Section 5.4, in ESI notation such bounds take a simple form, and become easy to manipulate and combine. By Proposition 5.4.1, Part 2, from (5.5) we immediately get the ESI

$$\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\left[\frac{1}{n}\sum_{i=1}^{n} X_{\bar{f},i}\right] \trianglelefteq_{nu} \frac{\mathrm{KL}(\hat{\Pi},\Pi_0)}{nu} \tag{5.8}$$

—the notation is explained in more detail in the next section. Here, KL is the Kullback-Leibler divergence; Π_0 is a distribution on $\mathcal{F}$ called a “prior” in analogy to prior distributions in Bayesian statistics; and $\hat{\Pi}$ is allowed to be any distribution on $\mathcal{F}$ that may depend on data Z^n and that represents the learning algorithm of interest. If we write $\mathbf{E}$ without subscript, we refer to the expectation of Z and hence to that of X_f; with subscript $\bar{f} \sim \hat{\Pi}$, the expectation is taken over $\hat{\Pi}$. In simple cases, $\hat{\Pi}$ will be a degenerate distribution with mass one on an estimator (learning algorithm) $\hat{f} = \hat{f}_{|Z^n}$, as above, and Π_0 will have a probability mass function π_0 on a countable subset of $\mathcal{F}$, and then $\mathrm{KL}(\hat{\Pi},\Pi_0) = -\ln \pi_0(\hat{f})$. Now, Lemma 5.3.7 in Section 5.3.2, adapted from Grünwald and Mehta [2020] but receiving a very different interpretation in the present ESI context, shows that, if the ESI (5.4) holds with $u(\epsilon) = C^*\epsilon^\gamma \wedge \eta^*$ (providing right-tail control of the X_f), then under a weak additional condition on the left tail, the so-called *witness condition*, there exists a constant $c > 0$ such that, for all $f \in \mathcal{F}$, $i \in [n]$,

$$X_{f,i} - c\mathbf{E}[X_{f,i}] \trianglelefteq_{u/2} 0 \tag{5.9}$$

((5.9) is not a trivial consequence of (5.4) because we have $\mathbf{E}[X_{f,i}] \le 0$). Using again Corollary 5.2.7 about ESI averages and PAC-Bayes Proposition 5.4.1, Part 2, from (5.9) we immediately get the ESI

$$\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\left[\frac{1}{n}\sum_{i=1}^{n} X_{\bar{f},i} - c\mathbf{E}[X_{\bar{f}}]\right] \trianglelefteq_{nu/2} \frac{2\mathrm{KL}(\hat{\Pi},\Pi_0)}{nu},$$

which, barring suboptimal constant factors, coincides with the main excess risk bound of Grünwald and Mehta [2020]. Indeed, in the case with $L_f = -X_f$, an excess loss, the above can be rewritten as

$$c\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\mathbf{E}_{Z\sim\mathbf{P}}[L_{\bar{f}}] \trianglelefteq_{nu/2} \mathbf{E}_{\bar{f}\sim\hat{\Pi}}\left[\frac{1}{n}\sum_{i=1}^{n} L_{\bar{f},i}\right] + \frac{2\mathrm{KL}(\hat{\Pi},\Pi_0)}{nu},$$

which provides an excess risk bound for the learning algorithm embodied by $\hat{\Pi}$. It says that the expected performance on future data—if we use the randomized predictor obtained by sampling from $\hat{\Pi}$—is in expectation as good as it performed on the sample Z^n itself, up to a KL/n complexity term. If $\hat{\Pi}$ implements empirical risk minimization, placing mass 1 on the $\hat{f} \in \mathcal{F}$ that minimizes the loss on Z^n, then the empirical excess loss $\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\left[\frac{1}{n}\sum_{i=1}^{n} L_{\bar{f},i}\right] = \frac{1}{n}\sum_{i=1}^{n} L_{\hat{f},i}$ must be ≤ 0; if further $\mathcal{F}$ is finite and Π_0 is uniform on $\mathcal{F}$, this implies, following a minimization analogous to (5.6) but now with $\ln(1/\delta)+2\mathrm{KL}(\hat{\Pi},\Pi_0)$ in the numerator, that depending on γ, a rate of $O((\ln|\mathcal{F}|)^{1/(1+\gamma)})$ is achieved both in expectation and in probability. Grünwald and Mehta [2020] show variations of this bound (with discretized infinite $\mathcal{F}$) to be minimax optimal in some situations.

Further developments: partial order, ESI Markov, Random η, non-i.i.d. Besides the properties needed for the above-illustrated applications to fast-rate, PAC-Bayesian, excess-risk bounds, we provide some further properties of the ESI that are of general interest. We start in Section 5.2 with basic properties of the ESI, including an extensive treatment of transitivity. We show that the strong ESI formally defines a *partial order* relation. We also provide answers to natural questions such as "does the ESI characterization (5.3) admit a converse?" and we show that ESIs imply some other curious stochastic inequalities. In particular, we show an *ESI Markov inequality*, which we find intriguing—whether it will prove useful in applications remains to be seen, though.

Section 5.3 gives detailed characterization of strong and general ESI, and contains, besides new notation, also some truly novel results. Section 5.4 revamps existing results to provide the connection to PAC-Bayes; its main result, Proposition 5.4.1, was already illustrated above. While strictly speaking not containing anything new, it reorganizes and disentangles existing PAC-Bayesian proof techniques, showing that there really are at least three inherently different basic PAC-Bayesian results that are used as building blocks in other works. Section 5.5 contains some new results again, concerning the situation that the η in strong ESIs itself is not fixed but itself a random, i.e. data-dependent, variable. The chapter ends with Section 5.6 that extends ESIs to the non-i.i.d. case, connecting them to random processes, showing that ESIs defined on a sequence of random variables remain valid under optional stopping. Example 5.6.4 in that section lays out an intriguing connection between Zhang's PAC-Bayesian inequality and the Wald identity, a classic result in sequential analysis. All longer proofs are deferred to appendices.

ESI, Annealed expectation and log-moment generating function Of course, it has always been common to abbreviate notations for moment and cumulant moment generating functions in order to get more compact representations and proofs of concentration inequalities. For example, the classic work of Boucheron et al. [2013] uses $\psi_X(\eta) = \ln \mathbf{E}[e^{\eta X}]$ for the cumulant moment generating function. Instead of this, we use ESI and, as will become useful later, the annealed expectation (5.10) $\mathbf{A}^{\eta}[X] = \eta^{-1} \ln \mathbf{E}[e^{\eta X}]$, i.e. $\psi_X(\eta) = \eta\, \mathbf{A}^{\eta}(X)$. We stress that we do not claim that our notations are inherently better or more useful. Rather, we think that in some contexts uses of unnormalized $\psi_X(\eta)$ together with high-probability statements may be preferable; in other—especially related to excess- and generalization risk bounds—the normalized version $\mathbf{A}^{\eta}[X]$ and the ESIs are more convenient. These new notations are meant to complement, not replace, the existing.

5.2. Basic ESI Properties

In this section, we show the properties of the ESI that were anticipated in the introduction. We start with Section 5.2.1, where we lay down the notation that will be used in the rest of the chapter; in particular, for the annealed expectation. In Section 5.2.2, we show basic properties of the ESI. There, we show the main implications of a random variable satisfying an ESI, and layout useful properties that will be used in the next sections. In Section 5.2.3 we show a partial converse to definition of the ESI: if a random variable has a subexponential right tail, it satisfies an ESI—we show a more definitive converse in Section 5.3. In Section 5.2.4 we show the main properties of the ESI in relation to its transitivity and its use to bounding sums of independent random variables. In Section 5.2.5, we show that the ESI defines a partial order on random variables. We end with Section 5.2.6 with a curiosity, a Markov-like inequality that replaces the requirement of positivity in Markov's inequality with the weaker $0 \trianglelefteq_{\eta} X$.

5.2.1. Preliminaries: additional definitions and notation

Throughout the chapter, we fix some probability space (Ω, Σ, P). Whenever we speak of random variables or a class of random variables without indicating their distribution, we assume that they are all defined relative to this triple, and that their expectation is well-defined. To be more precise, we call a function $X : \Omega \to \mathbb{R}$ a random variable if it is measurable; we may have $\mathbf{E}[X_+] = \infty$ (then $\mathbf{E}[X] = \infty$) or $\mathbf{E}[X_-] = \infty$ (then $\mathbf{E}[X] = -\infty$), but not both. Here and in the sequel, $\mathbf{E}$ denotes expectation under P and $X_+ = 0 \vee X; X_- = 0 \vee (-X)$.

Definition 5.2.1 (Subcentered and regular). We call a random variable X *subcentered* if $\mathbf{E}[X] \leq 0$ and *regular* if $\mathbf{E}[X^2] < \infty$. We call a family of random variables $\{X_f : f \in \mathcal{F}\}$ regular if $\sup_{f \in \mathcal{F}} \mathbf{E}[X_f^2] < \infty$.

The reason for reserving the grand word "regular" for this simple property is that, as we will see in Section 5.3, as long as it holds everything works out nicely; in particular, we obtain an equivalence between random variables satisfying an ESI being*subcentered, uniformly subgamma random variables.*

Definition 5.2.2 (Annealed expectation). Let $\eta > 0$ and let X be a random variable. We define the *annealed expectation* as

$$\mathbf{A}^{\eta}[X] = \frac{1}{\eta} \ln \mathbf{E}[\mathrm{e}^{\eta X}]. \tag{5.10}$$

The annealed expectation is a rescaling of the cumulant generating function, "a well-known provider of nonasymptotic bounds" [Catoni, 2007]; we remark that in some other works, "annealed expectation of X" refers to what is $-\mathbf{A}^{\eta}[-X]$ in our notation. Of course, the definition of the ESI could have been written using the annealed expectation as

$$X \trianglelefteq_{\eta} Y \quad \text{if and only if} \quad \mathbf{A}^{\eta}[X - Y] \leq 0. \tag{5.11}$$

We need one more, final extension of the ESI notation. Let u be an ESI function—a continuous, positive, increasing function. For any random variables X and Y and function $f : \mathbb{R}^{+} \times \mathbb{R} \to \mathbb{R}$, we write

$$X \trianglelefteq_{u} f(u, Y) \text{ as shorthand for: for all } \epsilon > 0, \text{ with } \eta = u(\epsilon),\ \mathbf{E}[\mathrm{e}^{\eta(X - f(\eta, Y))}] \leq e^{\eta\epsilon}. \tag{5.12}$$

Notice that we already used this notation implicitly in (5.8).

5.2.2. Basic Properties of the ESI

In the following proposition, we state the main consequences of two random variables X, Y satisfying an ESI $X \trianglelefteq_{\eta} Y$; namely, that they are ordered both in expectation and with high probability. In the next section we give a partial converse to this definition: if two random variables X, Y are ordered with high probability, they satisfy an ESI with modified constants. A more definitive characterization is the subject of Section 5.3.

Proposition 5.2.3 (ESI characterization). *Let X, Y be two random variables such that $X \trianglelefteq_{u} Y$ for some ESI function u. Then*

1. $\mathbf{E}[X] \leq \mathbf{E}[Y]$. *If $u \equiv \eta$ is constant (strong ESI), then the inequality is strict unless $X = Y$ a.s.*

2. *X and Y are ordered with high probability, that is, for all $\epsilon > 0$, $\mathbf{P}\{X \geq Y + \epsilon + K\} \leq \mathrm{e}^{-u(\epsilon)K}$, or equivalently, for any $\delta \in [0, 1]$*

$$X \leq Y + \inf_{\epsilon > 0} \left(\frac{1}{u(\epsilon)} \ln \frac{1}{\delta} + \epsilon \right), \tag{5.13}$$

with probability higher than $1 - \delta$. In the special case of $u \equiv \eta$ constant, i.e. a strong ESI, $\mathbf{P}\{X \geq Y + K\} \leq \mathrm{e}^{-\eta K}$ or, for any $0 < \delta \leq 1$,

$$X \leq Y + \frac{1}{\eta} \ln \frac{1}{\delta}, \tag{5.14}$$

with probability higher than $1 - \delta$.

Proof. Jensen's inequality and the fact that the function $x \mapsto \mathrm{e}^{-\eta x}$ is strictly convex yields Part 1 (including strictness for the strong ESI case). For Part 2, apply Markov's inequality to $e^{u(\epsilon)(X-Y-\epsilon)}$ to give $\mathbf{P}\{X \geq Y + \epsilon + (\ln(1/\delta)/u(\epsilon))\} \leq \delta$. Since this holds simultaneously for all $\delta > 0$, the result follows. □

For simplicity, we did not spell out the consequences of an ESI of the form $X \trianglelefteq_u f(u, Y)$ asdefined above in (5.12); the extension of Proposition 5.2.3 to this case is entirely straightforward.

Remark If the ESI $X \trianglelefteq_u Y$ is not strong, then it is possible that the inequality in Part 1 of the proposition is not strict, i.e. that $\mathbf{E}[X] = \mathbf{E}[Y]$. An example is given by $\mathbf{P}\{X = 1\} = \mathbf{P}\{X = -1\} = 1/2$, $\mathbf{P}\{Y = 0\} = 1$. By the cosh inequality we have $X \trianglelefteq_u Y$ for $u(\epsilon) = \epsilon/2$, yet obviously $\mathbf{E}[X] = \mathbf{E}[Y]$.

We now introduce some very basic useful properties of ESIs that we will freely use in the remainder of the chapter.

Proposition 5.2.4 (Useful Properties). *Let X, Y, Z be three random variables and let u and u^* be ESI functions. The following hold:*

1. *If $X \trianglelefteq_u Y$ and $Y \leq Z$ almost surely then $X \trianglelefteq_u Z$.*
2. *$X \leq Y$ almost surely if and only if $X \trianglelefteq_\eta Y$ (strong ESI) for every $\eta > 0$.*
3. *If $X \trianglelefteq_{u^*} Y$, then $X \trianglelefteq_{u^\circ} Y$ for each ESI function u° with $u^\circ \leq u^*$ (by which we mean: for all $\epsilon > 0, u^\circ(\epsilon) \leq u^*(\epsilon)$).*
4. *Suppose that $Z \trianglelefteq_u 0$. Then $Z_+ - \mathbf{E}[Z_+] \leq Z_+ \trianglelefteq_u (\ln 2)/u$ and similarly, for every $c > 0$, we have $Z\mathbf{1}\{Z \geq c\} \leq Z_+ \trianglelefteq_u (\ln 2)/u$.*
5. *For $\eta > 0$, it holds that*
$$X - \mathbf{A}^\eta[X] \trianglelefteq_\eta 0. \tag{5.15}$$
and hence
$$\mathbf{E}[X] \leq \mathbf{A}^\eta[X]. \tag{5.16}$$

Proof. We only give the proofs for strong ESIs with constant u; the generalizations to general ESI functions $u = \eta$ are immediate. For 1, notice that if $Y \leq Z$, then $X - Y \geq X - Z$. This in turn implies $0 \geq \mathbf{A}^\eta[X - Y] \geq \mathbf{A}^\eta[X - Z]$ so that $X \trianglelefteq_\eta Z$. For 2 it is clear that if $X - Y \leq 0$, then $\mathbf{A}^\eta[X - Y] \leq 0$ for each η. For the converse recall that if the p–norm $\|X\|_p = (\mathbf{E}[|X|^p])^{1/p}$ of a random variable X is finite for all $p > 0$, then, as $p \to \infty$, $\|X\|_p \to \operatorname{ess\,sup}|X|$, the essential supremum[2] of X. Note that by assumption $\mathbf{A}^\eta[X - Y] = \ln \left\|\mathrm{e}^{X-Y}\right\|_\eta \leq 0$ for all $\eta > 0$, and thus taking $\eta \to \infty$ we can conclude that $\ln(\operatorname{ess\,sup} \mathrm{e}^{X-Y}) \leq 0$, that is, $X - Y \leq 0$ almost surely. 3 follows from the convexity of the function $x \mapsto \mathrm{e}^{\eta x}$. 4 follows since

$$\mathbf{E}[\mathrm{e}^{\eta Z}] = \mathbf{E}[\mathrm{e}^{\eta Z_+}] + \mathbf{E}[\mathrm{e}^{-\eta Z_-}] - 1, \tag{5.17}$$

[2] The essential supremum of a random variable X is the smallest constant c such that $X \leq c$ almost surely.

so that

$$\mathbf{E}\left[\mathrm{e}^{\eta(Z_+ + (\ln 2)/\eta)}\right] = \frac{1}{2}\mathbf{E}\left[\mathrm{e}^{\eta Z_+}\right] \leq \frac{1}{2}\left(\mathbf{E}\left[\mathrm{e}^{\eta Z}\right] + 1\right) \leq 1,$$

where the final inequality follows by assumption. 5 follows from Jensen's inequality and (5.15) is just definition chasing. □

5.2.3. A partial converse to the basic ESI characterization

Proposition 5.2.5. *Let Z be a random variable. If there exist $a, b > 0$ such that*

$$\mathbf{P}\{Z \geq \epsilon\} \leq a\mathrm{e}^{-b\epsilon} \tag{5.18}$$

for each $\epsilon > 0$, then, for each $0 < \eta' < b$, there is a constant $c > 0$ such that $Z \trianglelefteq_{\eta'} c$, where

$$c = \frac{1}{\eta'} \ln\left(1 + \frac{a\eta'}{b - \eta'}\right). \tag{5.19}$$

In particular, if for some η the precise statement (5.14) holds for all $0 < \delta \leq 1$ with probability at least $1 - \delta$, then by taking $a = 1$, $b = \eta$, $\eta' = \eta/2$, $Z = X - Y$, we find that $X \trianglelefteq_{\eta/2} Y + (2/\eta)\ln 2$.

This proposition shows that if we have an exponentially small right-tail probability for Z, then an ESI statement with a $C^* > 0$ on the right must already hold; in particular, if we weaken an ESI to its high-probability implication and then convert back to an ESI, we loose both a factor of 2 in the scale factor η and an additive constant. If we can additionally assume that $\mathbf{E}[Z] \leq 0$, then both main ESI implications from Proposition 5.2.3 hold and indeed, if additionally Z is regular—if its second moment is bounded—, we get a more complete converse of Proposition 5.2.3 (allowing ESI functions u rather than just fixed η); this is done in Proposition 5.3.1 later on.

5.2.4. Sums of random variables and transitivity

In this subsection we show how ESIs are useful when proving probabilistic bounds for sums $\sum_{i=1}^{n} X_i$ of random variables—not necessarily independent—, and how this leads to a transitivity-like property. All our results are stated, and valid for, strong ESIs; in Corollary 5.2.7 we look at averages rather than sums and, as stated there, the results become valid for general ESIs.

Thus, consider the sum $S_n = \sum_{i=1}^{n} X_n$. In the case that strong ESI bounds are available for each of them individually, that is, when $X_i \trianglelefteq_{\eta_i} 0$ for some $\eta_i > 0$ and $i = 1, \ldots, n$, then we seek to obtain a similar statement for S_n—in analogy to the sum of negative numbers remaining negative. In order for S_n to remain negative with large probability, independence or, more generally, association assumptions need to be made. We discuss this fact after the statement of the bounds. A set of random variables $X_1, \ldots, X_n$ is said to be negatively associated [cf. Joag-Dev and Proschan, 1983, Dubhashi and Ranjan, 1998] if for any two disjoint index sets $I, J \subset \{1, \dots, n\}$ it holds that $\mathrm{Cov}(f(X_i, i \in I), g(X_j, j \in J)) \leq 0$, or more succinctly, if

$$\mathbf{E}[f(X_i, i \in I) g(X_j, j \in J)] \leq \mathbf{E}[f(X_i, i \in I)] \mathbf{E}[g(X_j, j \in J)]$$

for any choice of monotone increasing functions[3] f and g. Examples of negatively associated random variables include independent random variables, but also include negatively correlated jointly Gaussian random variables, and permutation distributions. The following proposition can be obtained.

Proposition 5.2.6. *Let $X_1, \dots, X_n$ be random variables such that $X_i \trianglelefteq_{\eta_i} 0$ for some $\eta_1, \dots, \eta_n > 0$. Then*

1. *Under no additional assumptions, $S_n \trianglelefteq_{\eta} 0$ with $\eta = \left(\sum_{i=1}^n \frac{1}{\eta_i} \right)^{-1}$.*

2. *If $X_1, \dots, X_n$ are negatively associated random variables—in particular, if they are independent—, then $S_n \trianglelefteq_{\eta} 0$ with $\eta = \min_i \eta_i$.*

Proof. We prove the case $n = 2$; its generalization is straightforward. Note that $\mathbf{A}^{\eta}[X] = \ln \|e^X\|_{\eta}$, where $\|\cdot\|_{\eta}$ denotes the p-norm at $p = \eta$ given by $\|Y\|_{\eta} = (\mathbf{E}|Y|^{\eta})^{1/\eta}$. Using Hölder's inequality we get

$$\mathbf{A}^{\eta}[X_1 + X_2] \leq \mathbf{A}^{\eta p}[X_1] + \mathbf{A}^{\eta q}[X_2], \tag{5.20}$$

where $p, q \geq 1$ are Hölder conjugates related by $p^{-1} + q^{-1} = 1$. Replacing $p = 1 + \frac{\eta_1}{\eta_2}$ and η as in 1, the result follows. For Part 2, note that for independent or negatively associated random variables it holds that $\mathbf{A}^{\eta}[S_n] \leq \sum_{i=1}^n \mathbf{A}^{\eta}[X_i] \leq 0$ with $\eta = \min_i \eta_i$, from which the result follows. □

With an eye towards the PAC-Bayesian bounds anticipated in the introduction, we now present a corollary of the previous proposition which holds for averages instead of sums. Its proof is omitted as it is a direct application of the previous proposition. Under this modification, the results hold for arbitrary ESI functions u instead of constants η; thus, it is this corollary that allows for the ESI treatment of PAC-Bayesian bounds. As above, consider random variables $X_1, \dots, X_n$ and let $\bar{X} = n^{-1} S_n$ be their average. We obtain:

Corollary 5.2.7. *Suppose that $X_i \trianglelefteq_{u_i}$ for ESI functions $u_1, \dots, u_n$. Then*

1. *Under no additional assumptions, $\bar{X} \trianglelefteq_{nu} 0$ with $u = \left(\sum_{i=1}^n \frac{1}{u_i} \right)^{-1}$.*

2. *If $X_1, \dots, X_n$ are i.i.d. and $u = u_1 = u_2 = \dots = u_n$, then $\bar{X} \trianglelefteq_{nu} 0$.*

The results obtained in Part 1 and 2 of the Proposition 5.2.6 above have very different quantitative consequences because of the difference in their association assumptions. In the case that for some fixed $\eta > 0$ it holds that $X_i \trianglelefteq_{\eta} 0$ for $i = 1, \dots, n$, then Proposition

[3] We mean that the functions are increasing in each argument when the others are held fixed.

5.2.6 implies that $S_n \trianglelefteq_{\eta/n} 0$. Through Proposition 5.2.3 this in turn implies that with probability higher than $1 - \delta$ it holds that

$$S_n \leq \frac{n}{\eta} \ln \frac{1}{\delta}.$$

This does not rule out the possibility that, even if all of the X_i are with large probability negative, their sum might still grow linearly with the number of terms n—for instance under complete dependency, when all $X_i = X_1$. On the other hand, when $X_i, \ldots, X_n$ are independent or negatively associated, this cannot be the case. Indeed, Proposition 5.2.6 implies $S_n \trianglelefteq_\eta 0$ which after using again Proposition 5.2.3, implies that with probability higher than $1 - \delta$

$$S_n \leq \frac{1}{\eta} \ln \frac{1}{\delta}.$$

As a corollary, the anticipated property that is reminiscent of transitivity holds for $\trianglelefteq_\eta$.

Corollary 5.2.8 (Transitivity). *If $X \trianglelefteq_{\eta_1} Y$ and $Y \trianglelefteq_{\eta_2} Z$, then*

1. *$X \trianglelefteq_\eta Z$ with $\eta = (1/\eta_1 + 1/\eta_2)^{-1}$.*
2. *If X, Y and Z are negatively associated, then $X \trianglelefteq_\eta Z$ with $\eta = \min\{\eta_1, \eta_2\}$.*

Proof. Use that $X - Z = (X - Y) + (Y - Z)$ and Proposition 5.2.6. □

We close this subsection with an observation about the common practice of using probabilistic union bounds. Even though in general the union bound is tight, in the presence of ESIs it is loose.

Remark 5.2.9 (Chaining ESI bounds improves on union bound). Suppose X,Y,Z are random variables such that $X \trianglelefteq_\eta Y$, and $Y \trianglelefteq_\eta Z$. For each $a > 0$, Proposition 5.2.3 implies both that $\mathbf{P}\{X \geq Y + a\} \leq \mathrm{e}^{-\eta a}$ and that $\mathbf{P}\{Y \geq Z + a\} \leq \mathrm{e}^{-\eta a}$. Using directly the union bound on these two events, one would obtain that $\mathbf{P}\{X \geq Z + 2a\} \leq 2\mathrm{e}^{-\eta a}$, or equivalently that with probability higher than $1 - \delta$

$$X \leq Z + \frac{2}{\eta} \ln \frac{2}{\delta} \tag{5.21}$$

while using Proposition 5.2.8 one obtains that $X \trianglelefteq_{\eta/2} Z$, which, again using Proposition 5.2.3 implies that with probability higher than $1 - \delta$

$$X \leq Z + \frac{2}{\eta} \ln \frac{1}{\delta}. \tag{5.22}$$

This is better than the previous bound because of the factor appearing in the logarithm. This seems like a minor difference, but the effect adds up when chaining n inequalities of this type. Indeed, in that case one obtains (by using ESI) in-probability bounds that tighter than the union bound by a $\ln n$ factor.

5.2.5. ESI as a stochastic ordering

ESIs are different from standard ordering relations in that they depend on the parameter u. We may view them as such standard ordering relations simply by adding existential quantifiers. Thus we may set

$X \trianglelefteq_{\text{GENERAL}} Y$ if and only if there exists an ESI function u s.t. $X \trianglelefteq_u Y$

$X \trianglelefteq_{\text{STRONG}} Y$ if and only if there exists $\eta^* \in \mathbb{R}^+$ s.t. $X \trianglelefteq_{\eta^*} Y$

Proposition 5.2.10. *Let $\{X_f : f \in \mathcal{F}\}$ be a set of random variables. Then $\trianglelefteq_{\text{STRONG}}$ defines a partial order on this set.*

We note that $\trianglelefteq_{\text{GENERAL}}$ does not define a partial order. Indeed, if $\mathbf{P}\{X = 1\} = \mathbf{P}\{X = -1\} = 1/2$ and $\mathbf{P}\{Y = 0\} = 1$ we have, as a consequence of a small computation, both $X \trianglelefteq_{\text{GENERAL}} Y$ and $Y \trianglelefteq_{\text{GENERAL}} X$. However, $X \neq Y$ a.s.

Proposition 5.2.10. We need to check whether the order is reflexive, transitive and antisymmetric. Reflexivity is immediate, transitivity follows from Corollary 5.2.8 above, and antisymmetry from Proposition 5.2.3, Part 1. □

In light of this proposition, it might be of interest to compare this partial order to the usual order of stochastic dominance, and its generalization, kth order stochastic dominance.

5.2.6. ESI-positive random variables: a curious Markov-like inequality

In this section we deal with random variables X that are positive in the strong ESI sense, that is, $0 \trianglelefteq_\eta X$ for some $\eta > 0$. Notice that by Proposition 5.2.3, we know that for each $a > 0$, we can bound the probability that X is smaller than $-a$—a left-tail bound—by $\mathbf{P}\{X \leq -a\} \leq \mathrm{e}^{-a}$. Additionally, we can obtain a Markov-style inequality for the probability that X is large—a right-tail bound.

Proposition 5.2.11. *Let X be a random variable such that $0 \trianglelefteq_\eta X$. Then, for any $a > 0$,*

$$\mathbf{P}\{X \geq a\} \leq \frac{\mathbf{E}[X]}{a} + \frac{p\ln(1/p)}{\eta a} \leq \frac{\mathbf{E}[X]}{a} + \frac{1}{\mathrm{e}\eta a},$$

where $p = \mathbf{P}\{X < 0\}$

Remark 5.2.12. Notice that the first inequality reduces to Markov's inequality in the case that $p = \mathbf{P}\{X < 0\} = 0$, that is, when X is a nonnegative random variable, the requirement for the standard Markov's inequality to hold. Thus, the intuition behind the proposition is that, since $0 \trianglelefteq_\eta X$ expresses X is "highly likely almost positive", it allows us to get something close to Markov after all.

Notice that for any increasing real-valued function f it holds that

$$\mathbf{P}\{X \geq a\} = \mathbf{P}\{f(X) \geq f(a)\}$$

and consequently if $f(X)$ is positive in the ESI sense, that is, $0 \trianglelefteq_\eta f(X)$ for some $\eta > 0$, our version of Markov's inequality can be used in the same spirit in which Chebyshev's inequality follows from Markov's inequality.

Corollary 5.2.13. *If f is increasing and X is a random variable such that $0 \trianglelefteq_\eta f(X)$, then*

$$\mathbf{P}\{X \geq a\} \leq \frac{\mathbf{E}[f(X)]}{f(a)} + \frac{p\ln(1/p)}{\eta f(a)} \leq \frac{\mathbf{E}[f(X)]}{f(a)} + \frac{1}{\mathrm{e}\eta f(a)} \tag{5.23}$$

where $p = \mathbf{P}\{f(X) < 0\}$.

5.3. When does a family of RVs satisfy an ESI?

In this section, we show a converse to the definition of the ESI. A special role will be payed by regular, subgamma, subcentered random variabes. As we will see, subgamma makes reference to random variables whose (right tail) is lighter than that of a gamma distribution. Recall from Section 5.2 that we call a family of random variables regular if its second moment is uniformly bounded; subcentered, if their expectation is negative.

5.3.1. General ESIs and subcentered subgamma random variables

We say that a random variable X has a (c, v)-*subgamma right tail* if it satisfies

$$X - \mathbf{E}[X] \trianglelefteq_\eta \frac{1}{2}\frac{v\eta}{1 - c\eta} \tag{5.24}$$

for some $c, v > 0$ and all η with $0 \leq c\eta \leq 1$ [see Boucheron et al., 2013, Section 2.4]. This name is in relation to the fact that random variables that are gamma distributed satisfy it. Subgamma random variables are well-studied: Van de Geer and Lederer [2013] studied empirical processes of random variables that satisfy a tail condition implied by (5.24). Sufficient conditions for (possibly unbounded) random variables to satisfy a subgamma bound have been known for a long time [cf. Uspensky, 1937, p. 202-204]. This topic has been also treated by Van der Vaart and Wellner [1996, Section 2.2.2] and by Boucheron et al. [2013, Section 2.8].

The following proposition shows that for a regular family, that is, a family satisfying $\sup_{f\in\mathcal{F}} \mathbf{E}[X_f^2] < \infty$, ESI families—families that satisfy $X_f \trianglelefteq_u 0$ for all f and some u—can be equivalently characterized in a number of ways. Its most important implications are that a regular family of random variables satisfies an ESI, i.e. for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$,

(a) if and only if its elements are all subcentered and uniformly subgamma on the right, and

(b) if and only if it satisfies an ESI for a function h that is linear near 0.

We also note that the first converse that we presented to the main ESI implications, Proposition 5.2.5, was still relatively weak, in the sense that if we have an ESI of the form $Z \trianglelefteq_u 0$, we apply the central Proposition 5.2.3 to calculate that for all $\epsilon > 0$, (a) $\mathbf{P}\{Z \geq K + \epsilon\} \leq e^{-u(\epsilon)K}$ and (b) $\mathbf{E}[Z] \leq 0$, and we "back-transform" (a) to an ESI via the converse in Proposition 5.2.5 (which only uses (a)), we obtain $Z \trianglelefteq_{u'} c$ for some ESI function u' and some $c > 0$, i.e. we loose a additive constant term. With the help of the proposition below, we can use (a) jointly with (b) to conclude (using 6. below) that $Z \trianglelefteq_{u'} 0$ for an ESI function u', i.e. we can "back-transform" without loosing any additive terms in the ESI.

Proposition 5.3.1. *Let $\{X_f\}_{f \in \mathcal{F}}$ be a regular family, i.e. $\sup_{f \in \mathcal{F}} \mathbf{E}[X_f^2] < \infty$. Then, the following statements are equivalent:*

1. *There is an ESI function u such that for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$.*
2. *There is a constant $C^* > 0$ and a constant $\eta^* > 0$ such that, uniformly over all $f \in \mathcal{F}$, $X_f \leq X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C^*$.*
3. *There exist $c, v > 0$ such that, for all $f \in \mathcal{F}$, the X_f are subcentered and have a (c, v)-subgamma right tail.*
4. *There is an ESI function h such that, for all $f \in \mathcal{F}$, we have $X_f \leq X_f - \mathbf{E}[X_f] \trianglelefteq_h 0$ where h is of the form $h(\epsilon) = C\epsilon \wedge \eta^*$.*
5. *There exists $c, v > 0$ such that, for all $f \in \mathcal{F}$, the X_f are subcentered and, for each $f \in \mathcal{F}$ and $0 < \delta \leq 1$, with probability at least $1 - \delta$,*
$$X_f \leq \sqrt{2v \ln(1/\delta)} + c \ln(1/\delta). \qquad (5.25)$$
6. *There exists $a > 0$ and a differentiable function $h : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ with $h(\epsilon) > 0$, $h'(\epsilon) \geq 0$ for $\epsilon > 0$, such that for all $f \in \mathcal{F}$, the X_f are subcentered and $\mathbf{P}\{X \geq \epsilon\} \leq a \exp(-h(\epsilon))$ (in particular h may be a positive constant or a linear function of ϵ).*

In Appendix D.2, we state and prove an extended version of this result, Proposition D.2.1, which shows that if (3) holds for some pair (c, v), then (5) holds for the same (c, v) and (4) holds for $\eta^* = 1/(2c)$ and $C = 1/(2v)$. It also shows that regularity is only required for some of the implications between the four statements above. In particular, it is not needed for (3) ⇒ (4), (3) ⇒ (5) and (4) ⇒ (1), and for (1) ⇒ (2); a strictly weaker condition—control of the first rather then second moment of the X_f—is sufficient. However, Example 5.3.3 below shows that, in general, some sort of minimal control of the supremum of the second moment, and hence of the left tails of the X_f, is needed (note though that higher moments of $|X_f|$ need not exist) to get (2) ⇒ (3) and hence the full range of equivalences. Indeed, the only difficult part in the proposition above is the implication (2) ⇒ (3). It is a direct consequence of Theorem 5.3.2 below (again proved in Appendix D.2), which shows that we can actually directly relate the constants (c, v) in "right subgamma" to the constants C^* and η^*. The proof extends an argument from [Boucheron et al., 2013, Theorem 2.10].

Theorem 5.3.2. *Let U be a random variable such that $U - \mathbf{E}[U] \trianglelefteq_{\eta^*} C$ for some fixed constants C and $\eta^* > 0$. Then for $0 < \eta \leq \eta^*$, we have $U - \mathbf{E}[U] \trianglelefteq_{\eta} \frac{1}{2}\frac{v\eta}{1-c\eta}$ for $v = \mathrm{Var}(U) + 2\exp(\eta^* C)$ and $c = 1/\eta^*$.*

Example 5.3.3. Let U be a random variable with, for $U \leq -1$, density $p(u) = 1/|u|^{\nu}$ for some ν with $5/2 < \nu < 3$. Then $\mathbf{P}\{U \leq -1\} = \int_{-\infty}^{-1} p(u) = 1/(\nu - 1)$. We set $x_\nu = (\nu-1)/(\nu-2)^2$ and $\mathbf{P}\{U = x_\nu\} = 1 - \mathbf{P}\{U \leq -1\}$ so that $\mathbf{P}\{-1 < U < x_\nu\} = \mathbf{P}\{U > x_\nu\} = 0$. Then $\mathbf{E}[U \cdot \mathbf{1}\{U \leq -1\}] = -1/(\nu-2)$ and hence $\mathbf{E}[U] = 0$, and an easy calculation shows that $U = U - \mathbf{E}[U] \trianglelefteq_1 C^*$ with $C^* = \ln(1 + \exp(x_\nu))$. Hence the premise inside (3) of Proposition 5.3.1 is satisfied for family $\{U\}$, but $\mathrm{Var}(U) = \infty$ so that $\{U\}$ is not regular so that the general precondition of Proposition 5.3.1 does not hold. And indeed (proof in Appendix D.2) we find that $(\mathbf{E}[\exp(\eta U)] - 1)/\eta^2 \to \infty$ as $\eta \downarrow 0$, showing that the right-subgamma property is not satisfied.

5.3.2. Interpolating between weak and strong ESIs

We may think of a weak and a strong ESI as two extremes in a hierarchy of possible tail bounds—the strong ESI given the lightest tails; the weak, the heaviest. We now define ESI families and γ-strong ESI family, where $\gamma \in [0,1]$ is the interpolating factor.

Definition 5.3.4. We say that a family of random variables $\{X_f : f \in \mathcal{F}\}$ is an *ESI family* if there exists an ESI function u such that for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$. For $0 \leq \gamma \leq 1$, we say that the family is a γ-strong ESI family if there exist $C^* > 0, \eta^* > 0$ and a function $u(\epsilon) = C^*\epsilon^\gamma \wedge \eta^*$ such that for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$. For an interval $I \subseteq [0,1]$, we say that the family is an I-strong ESI family if for all $\gamma \in I$, it is a γ-strong ESI family.

Note that if for some $\eta > 0$, all X_f satisfy the strong ESI $X_f \trianglelefteq_\eta 0$, then in this terminology they form a 0-strong ESI family.

Proposition 5.3.5. *Fix $\gamma \in [0,1]$. A regular family $\{X_f : f \in \mathcal{F}\}$ is a γ-strong ESI family if and only if there exists $C^\circ > 0, 0 < \eta^\circ < 1$ such that for all $f \in \mathcal{F}$,*

$$\textit{for all } 0 < \eta \leq \eta^\circ\textit{: } X_f \trianglelefteq_\eta C^\circ \eta^{\frac{1}{\gamma}} \tag{5.26}$$

where we set $\eta^{1/0} := \lim_{\gamma\downarrow 0} \eta^{1/\gamma} = 0$.

Proof. Let $u(\epsilon) = C^*\epsilon^\gamma \wedge \eta^*$ as in the definition of γ-strong. Set $\epsilon^* > 0$ to be such that $C^* e^{*\gamma} = \eta^*$, i.e. the value of ϵ at which $u(\epsilon)$ starts to become a horizontal line. By definition, we have

$$(5.26) \Leftrightarrow \forall \eta \in (0, \eta^\circ]\text{: } \mathbf{E}[\mathrm{e}^{\eta X_f}] \leq e^{\eta \cdot C^\circ \eta^{1/\gamma}} \text{ and } X_f \trianglelefteq_u 0 \Leftrightarrow$$
$$\forall \epsilon \in (0, \epsilon^*]\text{: } \mathbf{E}[\mathrm{e}^{C^*\epsilon^\gamma X_f}] \leq e^{C^*\epsilon^\gamma \cdot \epsilon}$$

If we set $C^\circ = 1/C^{*\gamma}$ and for each $\epsilon \in (0, \epsilon^*]$, we set $\eta = C^*\epsilon^\gamma$ then both expressions coincide for each such ϵ and for each $\eta \in (0, \eta^*]$; the result follows. □

The importance and motivation of γ-strong ESI families comes from their application in fast-rate results as already indicated in the introduction. As there, let $\{L_f : f \in \mathcal{F}\}$ be a collection of excess-loss random variables, L_f being the excess loss of predictor f, and let $X_f = -L_f$ be the negative excess loss. Then $\{X_f : f \in \mathcal{F}\}$ being a γ-strong ESI family coincides with, under the definitions of Van Erven et al. [2015], $\mathcal{F}$ satisfying the *u-central fast rate condition* for $u(\epsilon) = C^* \epsilon^\gamma \wedge \eta^*$. They showed that, for bounded loss functions (implying that the L_f are uniformly bounded), under the u-central fast-rate condition with u as above, and with a suitable notion of complexity $\textsc{comp}$, one can get an excess risk rate of order $O((\textsc{comp}/n)^{1/(1+\gamma)}$, as was illustrated for the special case of ERM with finite $\mathcal{F}$ in the introduction. Grünwald and Mehta [2020] (GM from now on) extended their result to the case that the L_f are unbounded, and only have minimal tail control on the right tail, the tail satisfying a condition they called the *witness-of-badness* or just *witness* condition. They showed that both this condition and a u-central fast-rate condition hold in many practically interesting learning situations. We state the witness-of-badness condition here in terms of $X_f = -L_f$ rather than L_f, since it can then also be used for collections $\{X_f\}_{f \in \mathcal{F}}$ that simply satisfy an ESI and have no excess-loss interpretation.

Definition 5.3.6 (Witness-of-Badness Condition). There exists $0 < c < 1$ and $C > 0$ such that for all $f \in \mathcal{F}$,

$$\mathbf{E}[(-X_f)\mathbf{1}\{-X_f \geq C\}] \leq c\mathbf{E}[-X_f]. \tag{5.27}$$

Note that this condition only makes sense for random variables with $\mathbf{E}[X_f] \leq 0$ (which automatically holds if $X_f \trianglelefteq_u 0$). It then automatically holds whenever the X_f have uniformly bounded left tail; GM show that it holds in many other cases as well, with the caveat that the constant C often scales linearly in the (suitably defined) dimension, making the resulting bounds not always optimal in terms of this dimension.

GM's Lemma 21, translated into ESI notation, now says the following:

Lemma 5.3.7 (GM's Lemma 21, rephrased as ESI). Suppose that $\{X_f : f \in \mathcal{F}\}$ is an ESI family, i.e. $X_f \trianglelefteq_u 0$, such that $\sup_{\epsilon > 0} u(\epsilon) < \infty$ (in particular, any ESI family can be expressed as such if it is regular) and suppose that the witness-of-badness condition as above holds. Then, there is a $c^* > 0$ such that, for all $f \in \mathcal{F}$,

$$X_f - c^* \mathbf{E}[X_f] \trianglelefteq_{u/2} 0. \tag{5.28}$$

GM go out of their way to optimize for the constant c^*; our interest being in the big picture here, we will not provide details about the constant. It can be seen from the proof that their result does not rely on the L_f having an interpretation as excess risks: it holds for general families $\{X_f : f \in \mathcal{F}\}$.

It was already discussed in the introduction how GM's Lemma (Lemma 5.3.7) can lead to fast rates. Essentially, to design learning algorithms that attain fast rates within this framework one needs that $\{X_f\}_{f \in \mathcal{F}}$ is a γ-strong ESI family for $\gamma < 1$, which gives exponential control of the X_f's right tail, ensuring that empirical losses converge to their mean faster than $1/\sqrt{n}$, and on top of that one needs witness-of-badness, which gives a very different kind of control of the potentially heavy-tailed left tail, to ensure that this convengence also holds for empirical losses with a constant times their expectation, the empirical risk, subtracted; finally one uses a PAC-Bayesian combination of all $f \in \mathcal{F}$ to get the desired excess risk bound.

5.3.3. The Bernstein conditions and the γ-strong ESI

In their general treatment of fast rate conditions, Van Erven et al. [2015] showed how the u-central condition for $\{L_f : f \in \mathcal{F}\}$ with $u(\epsilon) = C^* \epsilon^\gamma \wedge \eta^*$ is equivalent to the β-Bernstein condition, with $\beta = 1 - \gamma$. The β-Bernstein condition is a better known condition for obtaining fast rates in excess risk bounds [see Bartlett and Mendelson, 2006, Van Erven et al., 2015, Audibert, 2009]. Their equivalence result only holds for uniformly bounded L_f; extending it to general—unbounded—excess risks remained a nagging open question. In Theorem 5.3.11 below, we fully resolve this issue for abstract families of random variables that do not require an excess risk interpretation. As a by-product, the theorem implies an analogue to Lemma 5.3.7 that relates γ-strong ESIs to strengthenings thereof with $c\mathbf{E}[X]$ subtracted as in (5.28).

We first recall the standard definition of the Bernstein condition:

Definition 5.3.8 (β-Bernstein Condition). Let $\beta \in [0,1]$. We say that a family of random variables $\{L_f\}_{f \in \mathcal{F}}$ satisfies the β-Bernstein condition if, for all $f \in \mathcal{F}$, $\mathbf{E}[L_f] \geq 0$ and there is some $B > 0$ such that

$$\text{for all } f \in \mathcal{F},\ \mathbf{E}[L_f^2] \leq B(\mathbf{E}[L_f])^\beta. \tag{5.29}$$

The 1-Bernstein condition is also known as the *strong Bernstein condition.*

Suppose that $\{L_f\}_{f \in \mathcal{F}}$ is a regular family. Then, it is straightforward to show that the family satisfies β-Bernstein if and only if satisfies β'-Bernstein for all $\beta' \in [0, \beta]$. Motivated by this equivalence, we may start considering half-open intervals $[0, \beta)$ — it turns out that this gives a version of Bernstein that is much better suited for comparing with ESI families for unbounded random variables. Formally:

Definition 5.3.9. Let $I \subseteq [0,1]$ be an interval. We say that a family of random variables $\{L_f\}_{f \in \mathcal{F}}$ satisfies the I-Bernstein condition if for all $f \in \mathcal{F}$, $\mathbf{E}[L_f] \geq 0$ and for all $\beta' \in I$, there is some $B > 0$ such that

$$\text{for all } f \in \mathcal{F},\ \mathbf{E}[L_f^2] \leq B(\mathbf{E}[L_f])^{\beta'}.$$

It is immediately verified that every family that satisfies the I-Bernstein for nonempty I automatically has uniformly bounded second moment, i.e. it is regular.

The following theorem shows that for regular families of random variables, the notions of $(b, 1]$-strong ESI and $[0, b)$-strong Bernstein coincide (with for the Bernstein condition, X_f replaced by $-X_f$), under a "squared version" of the witness condition defined below; this condition has not been proposed before in the literature, as far as we know.

In Appendix D.3 we state and prove an extended version of the theorem, in which the various conditions on $\{X_f\}_{f \in \mathcal{F}}$ needed for the various implications in the theorem below are spelled out; these conditions are all implied by regularity but are in some cases weaker.

Definition 5.3.10 (Squared-Witness Condition). We consider the following condition for a family of random variables $\{U_f : f \in \mathcal{F}\}$: there exists $0 < c < 1$ and $C > 0$ such that for all $f \in \mathcal{F}$,

$$\mathbf{E}[U_f^2 \mathbf{1}\left\{U_f^2 \geq C\right\}] \leq c\mathbf{E}[U_f^2]. \tag{5.30}$$

The original witness-of-badness condition (Definition 5.3.6) is just (5.30) with $-X_f$ in the role of U_f^2, where $-X_f$ represents, as in this section, the excess risk. Below we use the equation with $U_f^2 = X_f^2 = (-X_f)^2$ and also with $U_f^2 = ((X_f)_-)^2$.

Special cases of parts of the following theorem for uniformly bounded X_f, for which regularity and squared witness automatically hold, were stated and proven by Koolen et al. [2016], and earlier by Gaillard et al. [2014].

Theorem 5.3.11. *Let $\{X_f : f \in \mathcal{F}\}$ be a regular family of random variables that satisfies the squared-witness condition above for $U_f = X_f$ or for $U_f = (X_f)_-$. Then the following statements are equivalent:*

1. *$\{-X_f : f \in \mathcal{F}\}$ satisfies the $[0,b)$-Bernstein condition for some $0 < b < 1$ and $\{X_f : f \in \mathcal{F}\}$ is an ESI family.*

2. *For all $\beta \in [0,b)$, for all $c \geq 0$, all $0 \leq c^* < 1$, there exists $\eta^\circ > 0$ and $C^\circ > 0$ such that for all $f \in \mathcal{F}$, all $0 < \eta \leq \eta^\circ$,*

$$X_f + c \cdot \eta \cdot X_f^2 - c^* \cdot \mathbf{E}[X_f] \trianglelefteq_\eta C^\circ \cdot \eta^{\frac{1}{1-\beta}}, \tag{5.31}$$

   or equivalently, by Proposition 5.3.5, there exists $\eta^*, C^* > 0$ such that

$$X_f + c \cdot \eta \cdot X_f^2 - c^* \cdot \mathbf{E}[X_f] \trianglelefteq_u 0,$$

   where $u(\epsilon) = C^* \epsilon^{1-\beta} \wedge \eta^*$.

3. *For all $\beta \in [0,b)$, there exists $\eta^\circ > 0$ and $C^\circ > 0$ such that for all $f \in \mathcal{F}$, all $0 < \eta \leq \eta^\circ$, we have*

$$X_f \trianglelefteq_\eta C^\circ \eta^{\frac{1}{1-\beta}},$$

   i.e. $\{X_f : f \in \mathcal{F}\}$ is a $(b,1]$-strong ESI family. Equivalently, by Proposition 5.3.5, there exists $\eta^*, C^* > 0$ such that for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$, where $u(\epsilon) = C^* \epsilon^{1-\beta} \wedge \eta^*$.

Note that, if $\{-X_f : f \in \mathcal{F}\}$ satisfies $[0,1)$-Bernstein, then $\mathbf{E}[X_f] \leq 0$ for all $f \in \mathcal{F}$; also, $X_f^2 \geq 0$. Therefore, the implication (2) ⇒ (3) is trivial. The proof of Theorem 5.3.11 is based on Theorem D.3.1 and Lemma D.3.2 in the appendix, which, taken together, are a bit stronger than Theorem 5.3.11, which comes at the price of a more complicated statement. In a nutshell, on one hand, the implication (1) ⇒ (2) still holds even if the witness-type condition does not hold. On the other hand, the implication (2) ⇒ (3) ⇒ (1) still holds if the right-hand side of (5.31) is replaced by 0 (strong ESI family) and then the conclusion in (3) also becomes that (a) $X_f \trianglelefteq_{\eta^*} 0$ and, in (1), that (b) $\{-X_f : f \in \mathcal{F}\}$ satisfies $[0,1]$-Bernstein. Thus, the implication (3) ⇒ (2) can be seen as a second-order analogue of Lemma 5.3.7, allowing not just

$c^*\mathbf{E}[X_f]$ but also $c\eta X_f^2$ to be added to X_f, at the price of requiring the witness-squared rather than the standard witness condition.

Having an ESI with X^2 outside of the expectation is not needed for the excess-risk bound discussed in the introduction, but it is crucial for several other PAC-Bayesian generalization bounds that also achieve faster rates if the data are sampled from a distribution such that a Bernstein condition holds [see Mhammedi et al., 2019].

5.4. PAC-Bayes

In this section we prove and write the PAC-Bayesian bounds [see McAllester, 1998, Van Erven, 2014, Catoni, 2007, Guedj, 2019, Alquier, 2023] in ESI notation, under which they take a pleasant look. Importantly, we find that in the existing literature, "applying the PAC-Bayesian" or "Donsker-Varadhan change-of-measure" technique can really mean at least three different things. Using the annealed expectation notation together with ESI can disentangle these different uses, appearing as the three different parts in Proposition 5.4.1 below. We let $\{X_f\}_{f\in\mathcal{F}}$ again be a family of random variables. Let Π and $\hat{\Pi}$ be two equivalent probability measures (two probability measures with the same null sets) on $\mathcal{F}$ such that their mutual Radon-Nykodim derivatives exist. Define the Kulback-Leibler divergence $\mathrm{KL}(\hat{\Pi},\Pi)$ as

$$\mathrm{KL}(\hat{\Pi},\Pi) = \mathbf{E}_{\hat{\Pi}}\left[\ln\frac{\mathrm{d}\hat{\Pi}}{\mathrm{d}\Pi}\right].$$

PAC-Bayesian theorems are based on the relation of convex duality that exists between the Kullback-Leibler divergence and the cumulant generating function—the logarithmic moment generating function. We state them here as strong ESIs but, since the following results hold for all $\eta > 0$, it also follows that they also hold with η replaced by any ESI function u.

We continue to assume that there are i.i.d. $Z, Z_1, \ldots, Z_n$ such that, for all $f \in \mathcal{F}$, $X_f = g_f(Z)$ and $X_{f,i} = g_f(Z_i)$ can be written as a function of Z and Z_i respectively for some function g_f. Hence the distribution of Z determines the distribution of X_f and $X_{f,i}$ for all $i \in [n], f \in \mathcal{F}$. In this section we need to pay special attention to the notation; P is not the only measure that plays a role. We will write the relevant measure in subscript of $\mathbf{E}$ and $\mathbf{A}$. Thus, $\mathbf{E}_{(Z,f)\sim P\otimes\Pi}[X_f] = \iint g_f(Z)\mathrm{d}\Pi(f)\mathrm{d}\mathbf{P}(Z)$—notice that Π might depend on Z. With this in mind, we state the following proposition.

Proposition 5.4.1. *Let $\{X_f\}_{f\in\mathcal{F}}$ be a family of random variables and let $\eta > 0$. Then for any two equivalent distributions Π_0 and $\hat{\Pi}$ on $\mathcal{F}$ we have:*

1. *The following ESI holds:*

$$\mathbf{E}_{\bar{f}\sim\hat{\Pi}}[X_{\bar{f}}] - \mathbf{A}^{\eta}_{(Z,\bar{f})\sim\mathbf{P}\otimes\Pi_0}[X_{\bar{f}}] \trianglelefteq_\eta \frac{\mathrm{KL}(\hat{\Pi},\Pi_0)}{\eta}. \tag{5.32}$$

2. Suppose further that for each $f \in \mathcal{F}$, $X_f \trianglelefteq_\eta 0$. Then we have:

$$\mathbf{E}_{\bar{f}\sim\hat{\Pi}}[X_{\bar{f}}] \trianglelefteq_\eta \frac{\mathrm{KL}(\hat{\Pi}, \Pi_0)}{\eta}. \tag{5.33}$$

3. Now let again $\{X_f\}_{f\in\mathcal{F}}$ be an arbitrary family. We have:

$$\mathbf{E}_{\bar{f}\sim\hat{\Pi}}[X_{\bar{f}} - \mathbf{A}^{\eta}_{Z\sim\mathbf{P}}[X_{\bar{f}}]] \trianglelefteq_\eta \frac{\mathrm{KL}(\hat{\Pi}, \Pi_0)}{\eta}. \tag{5.34}$$

In case the family $\{X_f\}_{f\in\mathcal{F}}$ satisfies $X_f \trianglelefteq_\eta 0$ for all $f \in \mathcal{F}$, then $\mathbf{A}^{\eta}[X_f]$ is negative and the third result is a “boosted” version of the second, and therefore should usually give stronger consequences.

Proof. For Part 1: the variational formula for the KL divergence (which appeared already in the work of Gibbs [1902]) states that

$$\ln \mathbf{E}_{\bar{f}\sim\Pi}[\mathrm{e}^{\eta X_{\bar{f}}}] \geq \eta \mathbf{E}_{\bar{f}\sim\hat{\Pi}}[X_{\bar{f}}] - \mathrm{KL}(\hat{\Pi}, \Pi)$$

Taking exponentials, P-expected value on both sides, and using Fubini's theorem,

$$\mathbf{E}_{\bar{f}\sim\Pi}\mathbf{E}_{Z\sim\mathbf{P}}[\mathrm{e}^{\eta X_f}] \geq \mathbf{E}_{Z\sim\mathbf{P}}\left[\exp\left(\eta\mathbf{E}_{\bar{f}\sim\hat{\Pi}}[X_{\bar{f}}] - \mathrm{KL}(\hat{\Pi}, \Pi)\right)\right],$$

which is a rewriting of the result. Part 2 follows from Part 1 by noting that in this case we further have $1 \geq \mathbf{E}_{\bar{f}\sim\Pi}\mathbf{E}_{Z\sim\mathbf{P}}[\mathrm{e}^{\eta X_{\bar{f}}}]$. Part 3 follows from using Part 2 with $X_{\bar{f}}$ replaced by $X_{\bar{f}} - \mathbf{A}^{\eta}_{Z\sim\mathbf{P}}[X_f]$; for this new random variable, ESI is guaranteed by the simple observation (5.15) in Proposition 5.2.4 so that Part 3 follows. □

The second result is the most straightforward one and has been used to derive many PAC-Bayesian results, e.g. Seldin et al. [2012], Tolstikhin and Seldin [2013], Wu and Seldin [2022], Mhammedi et al. [2019]. The third result, illustrated in Example 5.4.2 below, has been (implicitly) used to get PAC-Bayesian *excess-risk* bounds such as those by Zhang [2006a,b] and Grünwald and Mehta [2020]. The first result, illustrated in Example 5.4.3, can be used to derive a whole class of PAC-Bayesian bounds that include one of the strongest and best-known early bounds, the Langford-Seeger-Maurer bound [Seeger, 2002, Langford and Shawe-Taylor, 2002, Maurer, 2004, Alquier, 2023]. It would be interesting to see how recent articles establishing bounds based on conditional mutual information (which can be thought of as an in-expectation version of a specific PAC-Bayesian bound) fit in. For example, Grünwald et al. [2021] uses the second result, but this is not so clear for recent bounds such as those by Hellström and Durisi [2022].

Example 5.4.2 (Zhang's Inequality). Zhang's inequality [Zhang, 2006b,a] provides one of the strongest PAC-Bayesian-type excess-risk bounds in the literature; more precisely, it gives a “proto-bound” which can then be further specialized to a wide variety of settings. For $i = 1, \ldots, n$ and each $f \in \mathcal{F}$, let $X_{f,i}$ be i.i.d. copies of X_f. By (5.15) in Proposition 5.2.4 combined with Proposition 5.2.6 we automatically have that, for all $\eta > 0$, for all $f \in \mathcal{F}$, $\sum_{i=1}^{n} X_{f,i} - n\,\mathbf{A}^{\eta}[X_f] \trianglelefteq_{\eta} 0$ for every ESI function η. Zhang's bound, which using ESI notation we can give simultaneously in its expectation and in-probability version, is quite simply the result of applying the PAC-Bayes bound of Part 3 in Proposition 5.4.1 to these ESIs, and then dividing everything by n:

$$\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\left[\frac{1}{n}\sum_{i=1}^{n} X_{\bar{f},i} - \mathbf{A}^{\eta}_{Z\sim\mathbf{P}}[X_{\bar{f}}]\right] \trianglelefteq_{n\eta} \frac{\mathrm{KL}(\hat{\Pi},\Pi_0)}{n\eta}, \tag{5.35}$$

where we note that, as defined in the introduction, in Zhang's work the X_f represent minus excess risks, $X_f = L_f$ with $L_f = L_f(Z) = \ell_f(Z) - \ell_{f^*}(Z)$. The basic bound can then further be refined by bounding $\mathbf{A}^{\eta}$. In the setting of well-specified density estimation (in which f^* is the density of the underlying P) the f's represent densities and $\ell_f(z) = -\ln f(z)$ is the log-score. For fixed $\eta = 1/2$, the annealed expectation $\mathbf{A}^{1/2}$ is the Rényi divergence of order $1/2$ [Van Erven and Harremoes, 2014], which is an upper bound on the Hellinger distance. In that case, Zhang's bound becomes a risk bound for density estimation. For other loss functions we proceed as follows: since the bound holds under no further conditions at all, for every $\eta > 0$, it still holds if we replace η by an arbitrary ESI u. $\mathbf{A}^{u}_{Z\sim\mathbf{P}}[X_f]$ can then be bounded in terms of $\mathbf{E}_{Z\sim\mathbf{P}}[X_f]$ for appropriate γ-strong ESI function u. This is what was done in Grünwald and Mehta [2020, Lemma 21] which we restated here in ESI language as Proposition 5.3.7—we essentially followed their reasoning in the introduction while avoiding the explicit use of $\mathbf{A}^{u}$ there.

Example 5.4.3 (Bégin et al.'s unified derivation). Bégin et al. [2016] implicitly used the first result (Part 1 of the proposition above) to unify several PAC-Bayesian *generalization* bounds. They work in the same statistical learning setup as in the introduction, so $\ell_f(Z)$ represents the loss predictor f makes on outcome Z, and the aim is to bound, with high probability, the expected loss $\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\mathbf{E}_{Z\sim\mathbf{P}}[\ell_{\bar{f}}(Z)]$ of the learned distribution on classifiers $\hat{\Pi}$, when applied by drawing a $\bar{f}$ randomly from $\hat{\Pi}$, in terms of the behaviour of $\hat{\Pi}$ on the training sample, $\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_{\bar{f}}(Z_i)\right]$. In our (significantly compressed) language, they reason as follows: let $a \in \mathbb{R}^{+} \cup \{\infty\}$ and suppose we have a jointly convex divergence $\Delta : [0,a] \times [0,a] \to \mathbb{R}_0^{+}$, where by "divergence" we mean that $\Delta(c,c') \geq 0$ for all $c, c' \in [0,a]^2$ and $\Delta(c,c') = 0$ iff $c = c'$. Upon defining $X_f = \Delta(n^{-1}\sum_{i=1}^{n}\ell_f(Z_i), \mathbf{E}_P[\ell_f])$, we get, using Jensen's inequality,

$$\begin{aligned}
&\Delta\left(\mathbf{E}_{f\sim\hat{\Pi}}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_f(Z_i)\right], \mathbf{E}_{f\sim\hat{\Pi}}\mathbf{E}_{Z\sim\mathbf{P}}[\ell_f(Z)]\right)\\
\leq &\mathbf{E}_{f\sim\hat{\Pi}}\left[\Delta\left(\frac{1}{n}\sum_{i=1}^{n}\ell_f(Z_i), \mathbf{E}_{Z\sim\mathbf{P}}[\ell_f(Z)]\right)\right]\\
= &\frac{1}{n}\mathbf{E}_{f\sim\hat{\Pi}}\left[n\cdot\Delta\left(\frac{1}{n}\sum_{i=1}^{n}\ell_f(Z_i), \mathbf{E}_{Z\sim\mathbf{P}}[\ell_f(Z)]\right)\right]\\
\trianglelefteq_{\eta} &\frac{1}{n}\left(\mathbf{A}^{\eta}_{(Z_i,f)\sim\mathbf{P}\otimes\Pi_0}\left[n\Delta\left(\frac{1}{n}\sum_{i=1}^{n}\ell_f(Z_i), \mathbf{E}_{Z\sim\mathbf{P}}[\ell_f(Z)]\right)\right] + \frac{\mathrm{KL}(\hat{\Pi},\Pi_0)}{\eta}\right),
\end{aligned}$$

where $\mathbf{A}^{\eta}_{(Z_i,f)\sim\mathbf{P}\otimes\Pi_0}[\Delta(n^{-1}\sum_{i=1}^{n}\ell_f(Z_i),\mathbf{E}_P[\ell_f])]$ can be further bounded to get some well-known existing PAC-Bayes bounds such as the *Langford-Seeger-Maurer bound* [Alquier, 2023]. The latter is obtained by taking Δ as the KL divergence and letting η depend on $n^{-1}\sum\ell_f(Z_i)$ in a clever way.

5.5. ESI with random η

In some applications, we will want η to be estimated itself in terms of underlying data, i.e. it becomes a random variable $\hat{\eta}$; trying to learn η from the data is a recurring theme in one of the author's work, starting with his first learning theory article [Grünwald, 1999], and shown to be possible in some situations using the *safe-Bayesian algorithm* [Grünwald, 2012] while leading to gross problems in others [Grünwald and van Ommen, 2017]. Also, the fine-tuning of parameters in several PAC-Bayes bounds (e.g. Catoni's [2007] or the one in Mhammedi et al. [2019]) can be reinterpreted in terms of an η determined by the data. The goal of the present section is to extend the ESI definition to this case, allowing us to get a more general idea of what is possible with random η than in the specific cases treated in the aforementioned articles. In this section we only consider strong ESIs, i.e. $X \trianglelefteq_{\eta} Y$ rather than $X \trianglelefteq_{u} Y$.

Interestingly, many properties still go through for ESI with random η, but the in-expectation implication gets weakened—and its proof is not trivial any more.

Definition 5.5.1 (ESI with random η). Let $\hat{\eta}$ be a random variable with range $H \subset \mathbb{R}^+$ such that $\inf H > 0$. Let $\{X_\eta : \eta \in H\}$ and $\{Y_\eta : \eta \in H\}$ be two collections of random variables. We will write

$$X_{\hat{\eta}} \trianglelefteq_{\hat{\eta}} \quad Y_{\hat{\eta}} \quad \text{as shorthand for} \quad \hat{\eta}(X_{\hat{\eta}} - Y_{\hat{\eta}}) \trianglelefteq_1 0 \tag{5.36}$$

We can still get an in-expectation result from random-η-ESI with a small correction. It is trivial to give bounds for the expectation with $1/\eta_{\min}$—with $\eta_{\min}$ the largest lower bound of H— as a leading constant. However, since we want to work with $\hat{\eta}$ that are very small in "unlucky" cases but large in lucky cases, and we want to exploit lucky cases, this is not good enough. The following result, which extends Proposition 5.2.3 and 5.2.5 to the random η case and, in contrast to those propositions, is far from trivial, shows that we can instead get a dependence of the form $1/\hat{\eta}$, which is of the same order as what we lose anyway, even for fixed η, if we want our results to hold with high probability.

We let $\{X_\eta : \eta \in \mathcal{G}\}$ and $\{Y_\eta : \eta \in \mathcal{G}\}$ be any two collections of random variables.

Theorem 5.5.2. *Let $\mathcal{G}$, X_η and Y_η be as above, with H finite. We have:*

1. *If $X_{\hat{\eta}} \trianglelefteq_{\hat{\eta}} \ Y_{\hat{\eta}}$, then for any $\delta \in]0,1[$,*

$$\mathbf{P}\left\{X_{\hat{\eta}} \le Y_{\hat{\eta}} + \frac{\ln\frac{1}{\delta}}{\hat{\eta}}\right\} \ge 1-\delta, \tag{5.37}$$

$$\textit{and} \quad \mathbf{E}\left[X_{\hat{\eta}}\right] \le \mathbf{E}\left[Y_{\hat{\eta}} + \frac{1}{\hat{\eta}}\right]. \tag{5.38}$$

2. As a partial converse, if (5.37) holds, then we have

$$X_{\hat{\eta}} \trianglelefteq_{\frac{\hat{\eta}}{2}} Y_{\hat{\eta}} + \frac{2 \ln 2}{\hat{\eta}}. \tag{5.39}$$

Remark 5.5.3. The following simple example shows that even though $\mathbf{E}[\mathrm{e}^{\hat{\eta} W_{\hat{\eta}}}] \le 1$ it can happen that $\mathbf{E}[W_{\hat{\eta}}]$ is unbounded, showing that in general one cannot get rid of the additive $1/\hat{\eta}$ on the right-hand side of (5.38): Let $H = \{\eta_1, \eta_2\}$ and $W_{\eta_1} \equiv C_1 < 0$ and $W_{\eta_2} \equiv C_2 > 0$. We then set $\eta_1 = \frac{1}{-C_1}$ and $\eta_2 = \frac{1}{C_2}$; note that $\eta_1, \eta_2 > 0$ as required. The term $\mathbf{E}[\mathrm{e}^{\hat{\eta} W_{\hat{\eta}}}]$ does then not depend on C_1 and C_2 and computes to

$$\mathbf{E}[\mathrm{e}^{\hat{\eta} W_{\hat{\eta}}}] = p(\eta_1) e^{-1} + p(\eta_2) e^{1}$$

This term is smaller than 1 if we set for example $p(\eta_1) = \frac{3}{4}$. But for $C_2 \to \infty$ we observe that $\mathbf{E}[W_{\hat{\eta}}] \to \infty$.

5.5.1. Additional properties for random ESI: transitivity, PAC-Bayes on $\hat{\eta}$

Having established that the basic interpretation of an ESI as simultaneously expressing inequality in expectation and in probability still holds for the random η case, we may next ask whether the additional properties we showed for strong ESIs still hold in the random case, or even with random variables X_η indexed by η rather than f. We do this for the summation and transitivity properties of Section 5.2.4 and the PAC-Bayesian results of Section 5.4.

Random η ESI Sums and Transitivity In what follows, given random variables $Z_1, \ldots, Z_n$, $n \in \mathbb{N}$, in some set $\mathcal{Z}$, we denote

$$Z^{n \setminus i} \coloneqq (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n) \in \mathcal{Z}^{n-1}.$$

The following is a result analogous to Proposition 5.2.6, Part 2, with "negative correlation" replaced by "ESI holding conditionally given all variables except 1". Of course, it would be interesting to extend both results to make them more similar; whether this can be done will be left for future work.

Proposition 5.5.4 (ESI for sums and transitivity with random η)**.** *Let $\mathcal{G}$ be a finite subset of $\mathbb{R}^+$, and let $Z_1, \ldots, Z_n$ be $\mathcal{Z}$-valued i.i.d random variables distributed according to $\mathbf{P}$. For every $\eta \in \mathcal{G}$ and $i \in [n]$, let $X_{i,\eta} : \mathcal{Z}^n \to \mathbb{R}$ be a measurable function such that*

$$\text{for all } i \in [n] \text{ and } z^{n \setminus i} \in \mathcal{Z}^{n-1}, \quad X_{i,\eta}(z_1, \ldots, z_{i-1}, Z, z_{i+1}, \ldots, z_n) \trianglelefteq_{\eta} 0. \tag{5.40}$$

Then, for any random $\hat{\eta} \in \mathcal{G}$, we have

$$\mathbf{E}\left[\sum_{i=1}^{n} X_{i,\hat{\eta}}(Z^n)\right] \le \mathbf{E}\left[\frac{\ln |\mathcal{G}| + 1}{\hat{\eta}}\right]. \tag{5.41}$$

Random η and PAC-Bayes We now investigate whether strong ESIs for individual fixed η's are as easily combined into an ESI involving all η's, the particular η chosen in a data-dependent manner, as they are for individual X_f's. There we used general PAC-Bayesian combinations with arbitrary 'posterior' (data-dependent) $\hat{\Pi}$ on $f \in \mathcal{F}$. Here we consider the analogue with a data-dependent distribution $\hat{\Pi}$ on $\hat{\eta}$. We find that the resulting bound is slightly different, involving the likelihood ratio between posterior and prior for the chosen $\hat{\eta} \sim \hat{\Pi}$ rather than in expectation over $\hat{\Pi}$ (which would be the direct analogue of the PAC-Bayesian result Proposition 5.4.1) Still, if we focus on the special but important case with $\hat{\Pi}$ a degenerate distribution, almost surely putting all its mass on a single estimator $\hat{\eta}$, then we get a precise analogy to the PAC-Bayes result.

Proposition 5.5.5 (PAC-Bayes on Random η)**.** *Let Π_0 be any prior distribution on $\mathcal{G}$, and $\hat{\Pi} : \Omega \to \Delta(\mathcal{G})$ be any random estimator such that $\hat{\Pi}(\omega)$ is absolutely continuous with respect to Π_0, for all $\omega \in \Omega$. If $X_\eta \trianglelefteq_\eta 0$, for all $\eta \in \mathcal{G}$, then for $\hat{\eta} \sim \hat{\Pi}$, we have:*

$$X_{\hat{\eta}} \trianglelefteq_{\hat{\eta}} \frac{\ln \left.\frac{\mathrm{d}\hat{\Pi}}{\mathrm{d}\Pi_0}\right|_{\hat{\eta}}}{\hat{\eta}}. \tag{5.42}$$

Proof. For $\eta \in \mathcal{G}$, let W_η be the random variable defined by $W_\eta \coloneqq X_\eta - \frac{1}{\eta} \ln \left.\frac{\mathrm{d}\hat{\Pi}}{\mathrm{d}\Pi_0}\right|_\eta$. We have

$$\mathbf{E}\left[e^{\hat{\eta}W_{\hat{\eta}}}\right] = \mathbf{E}_{Z\sim\mathbf{P}}\mathbf{E}_{\hat{\eta}\sim\hat{\Pi}}\left[e^{\hat{\eta}W_{\hat{\eta}}}\right] = \\ \mathbf{E}_{Z\sim\mathbf{P}}\mathbf{E}_{\hat{\eta}\sim\hat{\Pi}}\left[\mathrm{e}^{\eta X_{\hat{\eta}} - \ln \left.\frac{\mathrm{d}\hat{\Pi}}{\mathrm{d}\Pi_0}\right|_\eta}\right] = \mathbf{E}_{Z\sim\mathbf{P}}\mathbf{E}_{\eta\sim\Pi_0}\left[\mathrm{e}^{\eta X_\eta}\right] \leq 1, \tag{5.43}$$

where the last step follows from the fact that $X_\eta \trianglelefteq_\eta 0$, for all $\eta \in \mathcal{G}$. This completes the proof. □

Corollary 5.5.6. *Let Π_0 be any prior distribution on $\mathcal{G}$, and $\hat{\Pi} : \Omega \to \Delta(\mathcal{G})$ be any random estimator such that $\hat{\Pi}(\omega)$ is absolutely continuous with respect to Π_0, for all $\omega \in \Omega$. If $X_\eta \trianglelefteq_\eta 0$, for all $\eta \in \mathcal{G}$, then for any $0 < \delta 1$ and $\hat{\eta} \sim \hat{\Pi}$:*

$$\mathbf{P}\left\{X_{\hat{\eta}} \leq \frac{\ln \left.\frac{\mathrm{d}\hat{\Pi}}{\mathrm{d}\Pi_0}\right|_{\hat{\eta}} + \ln\frac{1}{\delta}}{\hat{\eta}}\right\} \geq 1 - \delta, \tag{5.44}$$

$$\textit{and} \quad \mathbf{E}_{Z\sim\mathbf{P}}\left[X_{\hat{\eta}} - \frac{\ln \left.\frac{\mathrm{d}\hat{\Pi}}{\mathrm{d}\Pi_0}\right|_{\hat{\eta}} + 1}{\hat{\eta}}\right] \leq 0. \tag{5.45}$$

In particular, if $\hat{\Pi}$ a.s. puts mass 1 on a particular $\hat{\eta}$, where $\hat{\eta}$ is a random variable taking values in $\mathcal{G}$, and $\mathcal{G}$ is a countable set, Π_0 having probability mass function π_0, then the $\ln \left.\frac{\mathrm{d}\hat{\Pi}}{\mathrm{d}\Pi_0}\right|_{\hat{\eta}}$ term is equal to $-\ln \pi_0(\hat{\eta})$.

Proof. The result follows by applying Propositions 5.5.5 and Theorem 5.5.2 to the random variable $W_\eta \coloneqq X_\eta - \frac{1}{\eta} \ln \left.\frac{\mathrm{d}\hat{\Pi}}{\mathrm{d}\Pi_0}\right|_\eta$, $\eta \in \mathcal{G}$. □

5.6. Non-iid Sequences

Here we extend our previous results to sequences of random variables $X_1, X_2, \ldots$ that might not be independent and identically distributed. We find that, if an ESI hold for each X_i conditionally on the past, ESI statements about the sums of the X_i's remain valid under optional stopping, thereby connecting ESIs to the recent surge of work in *anytime-valid confidence sequences*, *e-values, e-variables* and *e-processes* [Grünwald et al., 2020, Ramdas et al., 2022a]. As a consequence, we reprove Wald's identity, a well-known result in sequential analysis dating back to the 1950s, and show that it is related to Zhang's inequality treated before, and implies that Zhang's inequality remains valid under optional stopping. Relatedly, it has recently been noted that PAC-Bayesian inequalities are closely related to e-processes as well [Jang et al., 2023, Chugg et al., 2023]. Let us clarify the straightforward connection between e-variables as defined in the above references and strong ESIs. Formally, e-variables S are defined relative to some random variable Y and a *null hypothesis* $\mathcal{H}_0$, a set of distributions on Y. We call nonnegative random variable S an e-variable relative to Y and $\mathcal{H}_0$ if it can be written as a function $S = S(Y)$ of Y and, for all $P \in \mathcal{H}_0$, $\mathbf{E_P}[S(Y)] \leq 1$. To clarify the connection to ESIs, let $\{X_f : f \in \mathcal{F}\}$ be a family of random variables with $\mathbf{P}_f$ the marginal distribution of X_f as induced by $\mathbf{P}$, and suppose that $\{X_f : f \in \mathcal{F}\}$ all satisfy $X_f \trianglelefteq_{\eta^*} 0$. Suppose that we observe random variable Y. Under the null hypothesis, $Y = X_f$ for some $f \in \mathcal{F}$ (they take on the same values). Equivalently, under the null hypothesis, $Y \sim \mathbf{P}_f$ for some $f \in \mathcal{F}$, i.e. $\mathcal{H}_0 = \{\mathbf{P}_f : f \in \mathcal{F}\}$. Then, clearly, $S(Y) := \exp(\eta^* Y)$ is an e-variable. We will not further exploit or dwell on this fact below, but rather concentrate on the development of ESI for random processes.

Definition 5.6.1 (Conditional ESI). Let let X and Y be two random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbf{P})$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra. Define

$$X \trianglelefteq_{\eta,\mathcal{G}} Y \text{ if and only if } \mathbf{A}^\eta[X - Y|\mathcal{G}] \leq 0 \text{ almost surely,}$$

where we call $\mathbf{A}^\eta[X - Y|\mathcal{G}] = \frac{1}{\eta} \ln \mathbf{E}[\mathrm{e}^{\eta(X-Y)}|\mathcal{G}]$ the conditional annealed expectation of $X - Y$ given $\mathcal{G}$.

The following properties can be checked; they follow from the standard properties of the conditional expectation—"pulling out known factors", and the tower property.

Proposition 5.6.2. *Let let X be an $\mathcal{F}$-measurable random variable and let $\mathcal{H} \subseteq \mathcal{G} \subseteq \mathcal{F}$ be σ-algebras. The following hold:*

1. *If $X \trianglelefteq_{\eta,\mathcal{G}} 0$ and X is $\mathcal{G}$-measurable, then $X \leq 0$ almost surely.*
2. *If $X \trianglelefteq_{\eta,\mathcal{G}} 0$, then $X \trianglelefteq_{\eta} 0$.*
3. *If $X \trianglelefteq_{\eta,\mathcal{G}} 0$, then $X \trianglelefteq_{\eta,\mathcal{H}} 0$.*

Proof. 1 follows from the fact that if X is $\mathcal{G}$-measurable, then $\mathbf{A}^\eta[X|\mathcal{G}] = X$. 2 follows from the fact that $\mathbf{A}^\eta[X|\mathcal{G}] \leq 0$ implies that $\mathbf{A}^\eta[X] = \mathbf{A}^\eta[\mathbf{A}^\eta[X|\mathcal{G}]] \leq 0$. 3 follows from the tower property of conditional expectations because $\mathbf{A}^\eta[X|\mathcal{H}] = \mathbf{A}^\eta[\mathbf{A}^\eta[X|\mathcal{G}]|\mathcal{H}] \leq 0$. □

Let $(\Omega, \mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{N}}, P)$ be a filtered probability space. Let $(X_t)_{t\in\mathbb{N}}$ be a sequence of random variables adapted to $\mathbb{F}$ and assume that $X_t \trianglelefteq_{\eta,t-1} 0$ (where we write $X_t \trianglelefteq_{\eta,t-1} 0$ instead of $X_t \trianglelefteq_{\eta,\mathcal{F}_{t-1}} 0$ to avoid double subindexes). This statement expresses the fact that $(\prod_{s\le t} \mathrm{e}^{\eta X_s})_{t\in\mathbb{N}}$ is a supermartingale.

Proposition 5.6.3. *Let $(X_t)_{t\in\mathbb{N}}$ be an adapted sequence such that $X_t \trianglelefteq_{\eta,t-1} 0$ for each t and for some $\eta > 0$. Let τ be an almost surely bounded stopping time with respect to $(X_t)_{t\in\mathbb{N}}$. Then, if $S_t = \sum_{s\le t} X_s$,*

$$S_\tau \trianglelefteq_\eta 0.$$

Proof. The result is an application of the Optional Stopping Theorem. □

We now present two applications of this result, Example 5.6.4 and Proposition 5.6.5.

Example 5.6.4. [Zhang meets Wald] Let $X_1, X_2, \ldots$ be i.i.d. copies of some random variable X, fix arbitrary $\eta > 0$ and let $Z_i = X_i - \mathbf{A}^\eta[X]$. Then the Z_i are also i.i.d., hence $(Z_t)_{t\in\mathbb{N}}$ is adapted, and by (5.15 in Proposition 5.2.4, they satisfy $Z_t \trianglelefteq_{\eta,t-1} 0$. Therefore we can use Proposition 5.6.3 to infer that for any a.s. bounded stopping time τ, with $S_t := \sum_{i=1}^t Z_i$ that $S_\tau \trianglelefteq_\eta 0$, i.e.

$$\sum_{i=1}^{\tau} X_i - \tau \cdot \mathbf{A}^\eta[X] \trianglelefteq_\eta 0, \tag{5.46}$$

which must hold for all $\eta > 0$ and thus also if η is replaced by any ESI function u. But (5.46) is just the celebrated *Wald identity* [Skorokhod, 2012] as expressed in ESI notation, which we have thus reproved. (the Wald identity is not to be confused with the more well-known basic Wald's equation, which says that $\mathbf{E}[S_\tau] = \mathbf{E}[\tau] \cdot \mathbf{E}[X]$). We may now, just as in Example 5.4.2, combine this with a PAC-Bayes bound and then divide everything by τ to get, for a family of random variables $\{X_f : f \in \mathcal{F}\}$ with $X_{f,1}, X_{f,2}, \ldots$ i.i.d. copies of X_f as in the introduction,

$$\mathbf{E}_{\bar{f}\sim\hat{\Pi}}\left[\frac{1}{\tau}\sum_{i=1}^{\tau} X_{\bar{f},i} - \mathbf{A}^\eta[X_{\bar{f}}]\right] \trianglelefteq_{\tau\eta} \frac{\mathrm{KL}(\hat{\Pi}, \Pi_0)}{\tau\eta}.$$

We see that this is identical to *Zhang's inequality* (5.35), which we have therefore shown to be "anytime valid" (it holds for any stopping time τ), something that, it seems, has not been noted before. Since the scaling $\tau\eta$ is now data-dependent, we have to use Theorem 5.5.2 rather than Proposition 5.2.3 if we want to turn this in an in-probability or in-expectation bound though.

Finally, we note that using Ville's maximal inequality[4] [Ville, 1939, p.35] we can obtain the following proposition.

Proposition 5.6.5. *Let $(X_t)_{t\in\mathbb{N}}$ be a sequence of random variables such that $X_t \trianglelefteq_{\eta^*,t-1} 0$ for each t and some $\eta^* > 0$. Let $0 < \eta < \eta^*$. Then, there is a fixed constant c such that*

$$\sup_{t\in\mathbb{N}} X_t \trianglelefteq_\eta c.$$

[4] This inequality is also commonly attributed to J.L. Doob.

Proof. By Ville's maximal inequality,

$$\mathbf{P}\left\{\sup_{t\in\mathbb{N}} X_t \geq x\right\} = \mathbf{P}\left\{\sup_{t\in\mathbb{N}} \mathrm{e}^{\eta^* X_t} \geq \mathrm{e}^{\eta^* x}\right\} \leq \mathbf{E}[\mathrm{e}^{\eta^* X_1}]\mathrm{e}^{-\eta^* x} \leq \mathrm{e}^{-\eta^* x}.$$

The result follows from Proposition 5.2.5. □

5.7. Discusion

It is sometimes the case that half the way to solving a problem is finding the correct notation to state it. We have emphasized that many results in probability theory and statistical learning theory—especially PAC-Bayesian bounds—are obtained through bounds for cumulant generating functions. In this chapter we have have introduced a notational device with the goal of systematizing such bounds. The result is the Exponential Stochastic Inequality (ESI), which the authors have found helpful—we do not claim its absolute superiority, though. The strong ESI $X \trianglelefteq_\eta 0$ can be thought of as an interpolation between positivity in expectation (the case that $\eta \downarrow 0$) and almost-sure positivity ($\eta \to \infty$). Its main properties, shown in Section 5.2, allow for the derivation of high-probability and in-expectation bounds, and its transitivity-like property allows for chaining such bounds in a way that is superior to a straightforward union bound.

Inventing new notation is, however, a contentious affair. We have found the community to be rather conservative about notational changes. Like many things, this has two sides. On the positive side, it allows for easy understanding of a wide variety of articles at a low overhead. Standard notation serves as a *lingua franca* for conveying mathematical ideas. On the other side, sometimes good ideas are obscured for the sole reason that they are awkward to write in standard notation. We believe that in these cases—such as, we argue, PAC-Bayesian bounds—, new notation can help clarify and systematize the key techniques of the field. This is not a new idea; for instance, mathematicians (sometimes) and physicists (more often) have been inventing new notation for the better part of last century—think of Feynman diagrams or Einstein's summation convention. Hopefully, as it has already happened in other areas, new notation will help easier communication, provide a deeper understanding of the present techniques, and help the rise of new ones.

Having said that, we note once more that the ESI does have limitations. Of course, not all tails are exponential, and not all bounds are obtained through the analysis of cumulant generating functions. The arguments that we have presented using the ESI are consequences of the use of a particular convex duality relation—the one that exists between the cumulant generating function and the Kulback-Leibler divergence. A particular and interesting extension of the ESI might come from applying the same reasoning to other convex duality relationships. For example, Lugosi and Neu [2022] provide PAC-Bayesian-like bounds based on other convex dualities; their work might be considered a first step in this direction. This enterprise is worthwhile because convex duality arguments are the bread and butter of statistical learning theory, online learning and optimization.

5.8. Acknowledgements

This chapter benefited enormously from several conversations with Wouter Koolen, Tim van Erven, Nishant Mehta and Odalric-Ambrym Maillard; in particular, some results in Maillard's (2019) habilitation, although not directly used, were inspirational to the developments in Section 5.5. This research was supported by the Dutch Research Council (NWO) via research project 617.001.651, *Safe Bayesian Learning*.

6. Discussion

> As a word of warning, almost any set of summary statistics can have a story woven about them—we are good at making up stories.
>
> Persi Diaconis

In this dissertation we have shown optimal AV tests for group-invariant problems; an AV test that replaces the logrank test for two-group survival-time data; we formulated strategies for online prediction under multiscale range constraints; and proposed a new notation device to reason about probabilistic inequalities that hold in expectation and with high probability. In this discussion we review the main results of this dissertation, and point at open problems and future lines of research.

6.1. Anytime-valid methods

In this dissertation we have considered E-statistics that are growth rate optimal in the worst case (GROW), as proposed by [Grünwald et al., 2020]. One might wonder about the necessity of using this specific optimality criterion; at first sight it might seem rather arbitrary. The main goal that is achieved by using GROW E-statistics is that the evidence of consecutive experiments, as measured using E-statistics, can be combined through multiplying their respective E-values—the observed value of the E-statistic after the experiment is conducted. GROW E-statistics maximize their worst-case expected logarithmic value under the alternative hypothesis. This criterion has nothing to do with the value of evidence—as could be measured with other functions rather than with the logarithm—, but with the fact that the logarithm is additive over repetitions of the experiment and the law of large numbers applies to it. This optimality criterion can be understood using the metaphor of a repeated gambling game where the player is allowed to reinvest their revenue. Under this interpretation, the E-value is the payout of the game after betting one monetary unit; for example €1. An observed E-value of 20 would correspond to a payout of €20 after having bet €1. GROW E-statistics maximize the worst-case growth rate of the capital of a gambler in this imaginary game when the odds of the outcomes are given by an unknown distribution in the alternative hypothesis. In turn, this capital growth rate is the reinterpretation of Kelly Jr. [1956] of the rate information transmission in communication channels when they can be used repeatedly with the goal of error correction [Shannon, 1948]. Shannon used the logarithm because (1) it is practically useful, (2) it is intuitive, and (3) it is mathematically more suitable than other functions. Furthermore, they provide a plausibility argument based on an axiomatic derivation of the logarithmic criterion [see Shannon, 1948, Theorem 2]. Even though GROW criterion has the notion of "repeated experimentation" built in its definition, using it is not a mathematical necessity—but neither is using power maximization for fixed-sample size experiments.

In the examples studied in this dissertation—group-invariant problems in Chapter 2 and time-to-event data in Chapter 3—, this criterion yielded optimal tests statistics that coincided with either a conventional likelihood ratio or a partial likelihood ratio. There are, however, problems for which this GROW criterion yields nonstandard E-statistics. Although being nonstandard is not a problem in itself—being GROW is, we believe, a good property to satisfy—, their computation can be challenging. This is the case of the extension to the full proportional hazards ratio model of the results from Chapter 3.

In Chapter 3, we proposed the AV logrank test, an anytime-valid upgrade of the classic logrank test, when the comparison is between the event times of two groups of subjects. The AV logrank test is anytime-valid in the sense that given an initial design where a fixed number of subjects is followed, the events can be monitored continuously in time. We devised a test that guarantees a type-I error guarantee under any stopping decision in the duration of the study. Furthermore, the resulting test statistic can be multiplied with that of other studies that are evaluating the same null hypothesis concurrently. The resulting meta-analytic E-statistic can also be monitored continuously in time and decisions can be made based on it, going beyond the realm of conventional meta-analysis. Given its importance in statistical practice (see Chapter 3), designing an anytime-valid treatment of the full proportional hazards model of Cox [1972] is an interesting line of research. This is a hard problem because there exists no closed-form GROW E-statistic and computing one numerically is a formidable computational problem. Either finding an efficient KL minimization algorithm for this task or finding near-optimal E-statistic is a worthwhile project. The results of this chapter, as we will see, can be reinterpreted in terms of the results of Chapter 2.

In Chapter 2, we investigated group-invariant problems. The problems covered by these results include dominated models that are invariant under the action of fairly general groups—many groups of interest for parametric estimation are included—with certain geometric properties. The main result of that chapter is that the overall GROW E-statistic for group-invariant problems resides within the family of group-invariant E-statistics. When the invariance-reduced problem is a simple-vs.-simple test, an anytime-valid test can be constructed using the GROW E-statistic. This result draws a parallel with the theory of most powerful tests for group-invariant problems, where the main result is that of Hunt and Stein [Lehmann and Romano, 2005]. An abridged and informal version of main assumptions for the main results of Chapter 2—in addition to invariance of the model—are the following:

1. The models under consideration are both dominated by a common measure.
2. The group under which the problem is invariant is, roughly speaking "not too big"—it is a locally compact Hausdorff amenable group.

In relation to the second item above, we saw in Chapter 2 that there exist nonamenable groups for which a group-invariant GROW E-statistic exists. On the other hand, the first item above excludes the application of our results to nonparametric infinite-dimensional models where no densities with respect a common measure are available. As we will see, relaxing these assumptions would open the door to proving the conjecture that the AV logrank test statistic—Cox' partial likelihood ratio—from Chapter 3 is GROW for the proportional hazards alternative.

The results of Chapter 3 about the AV logrank test can be understood in terms of a group-invariant structure. Indeed, Kalbfleisch and Prentice [1973] showed that, in the fixed-sample case, Cox' partial likelihood used for the AV logrank test can also be interpreted as a partial likelihood ratio of the rank vector of the survival times. This follows from a more general fact about rank statistics shown by Savage [1956]. In this case, the partial likelihood ratio is between the null hypothesis under which the distribution of the survival times of all subjects is the same against the proportional hazards alternative [see Kalbfleisch and Prentice, 1973]. In this case, the rank statistic can be seen as an invariant statistic under the action of the group of all increasing functions of the survival times—all time changes that preserve the order of the events—with composition as the group operation. Needless to say, this group is not even locally compact—no Haar measure exists on it, and our techinques do not apply. A natural question is whether our group-invariance results also extend to Cox' model, that is, whether the partial likelihood employed in the AV logrank test is a GROW E-statistic for this problem.

6.2. Individual-sequence prediction

In Chapter 4, we introduced a multiscale adaptive algorithm, Muscada, for the game of prediction with expert advice [Cesa-Bianchi and Lugosi, 2006]. This algorithm is computationally efficient and provides a regret guarantee with the following two desirable properties:

1. If the ranges of the losses of the experts vary by orders of magnitude, the regret of the algorithm scales with the range of the best expert, not with the largest one—which would result from naively using a single-scale experts algorithm.

2. If the losses are samples from an easy distribution—a distribution according to which the best expert is better than any other bay a constant margin in expectation—, then the regret incurred by the algorithm assured to be constant.

The first property refers to the multiscale adaptiveness; the second one, to luckiness. Existing algorithms achieved either of these objectives, but not both. The proof ideas and techniques that were developed in designing and analyzing Muscada are broadly applicable to the family of Follow-the-Leader algorithms, and have the potential to motivate a general theory of second-order parameter-free procedures. For example, an instantiation of these techniques to gradient descent yield guarantees such as those of AdaGrad [Duchi et al., 2011] (not shown). The crux of the algorithm is a subtle use of convex duality, where a potential function of the experts' corrected regrets—the dual of the regularizer—is kept negative.

At least two problems are left open in this line of research by this dissertation. The first one is related to the application of MUSCADA to the solution of two-player zero-sum games. We have shown in Section C.1.1 convinvincing numerical evidence that if two instances of an optimistic version of MUSCADA play against each other, their strategies will converge to the saddle point of the game. An open problem is to provide a proof that this in fact is the case. The second one, already mentioned in Section 4.7, is the extension of the algortithm to infinitely many experts. This would open the possibility of formulating improved online learning algorithms for nonparametric online regression [Cesa-Bianchi et al., 2017, Gaillard and Gerchinovitz, 2015, Kuzborskij and Cesa-Bianchi, 2020].

6.3. Concentration Inequalities

In Chapter 5, we studied the exponential stochastic inequality, a new notational device designed to reason about random variables that are ordered both in expectation and with high probability. This notation is specially well suited to the study of PAC-Bayesian bounds, and has been useful to reason about excess risk bounds for machine learning algorithms. This chapter also serves as a nonextensive survey on PAC-Bayesian bounds. These bounds can be understood through the convex duality that exists between the cumulant generating function and the Kulback-Leibler divergence. The notational device is a tool simplify and hopefuly derive new results in this area.

Appendices

A. Appendix to Chapter 2

A.1. Computations

Proposition A.1.1. *Let $X \sim N(\gamma, I)$, and let $mS \sim W(m, I)$ be independent random variables. Let $LL' = S$ be the Cholesky decomposition of S, and let $M = \frac{1}{\sqrt{m}} L^{-1} X$. If $\mathbf{P}_{0,n}$ is the probability distribution under which $X \sim N(0, I)$, then, the likelihood $p^M_{\gamma,m}/p^M_{0,m}$ ratio is given by*

$$\frac{p^M_{\gamma,m}(M)}{p^M_{0,m}(M)} = \mathrm{e}^{-\frac{1}{2}\|\gamma\|^2} \int \mathrm{e}^{\langle \gamma, TA^{-1}M\rangle} \mathrm{d}\mathbf{P}_{m+1,I}(T)$$

where $A \in \mathcal{L}^+$ is the Cholesky factor $AA' = I + MM'$, and $\mathbf{P}^T_{m+1,I}$ is the probability distribution on $\mathcal{L}^+$ such that $TT' \sim W(m+1, I)$.

Proof. Let $\Sigma = \Lambda\Lambda'$ be the Cholesky decomposition of Σ. The density $p^X_{\gamma,\Lambda}$ of X with respect to the Lebesgue measure on $\mathbb{R}^d$ is

$$p^X_{\gamma,\Lambda}(X) = \frac{1}{(2\pi)^{d/2}\det(\Lambda)} \mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}X - \gamma)(\Lambda^{-1}X - \gamma)'\right),$$

where, for a square matrix A, we define $\mathrm{etr}(A)$ to be the exponential of the trace of A. Let $W = mS$. Then, the density $p^W_{\gamma,\Lambda}$ of W with respect to the Lebesgue measure on $\mathbb{R}^{d(d-1)/2}$ is

$$p^W_{\gamma,\Lambda}(W) = \frac{1}{2^{md/2}\Gamma_d(n/2)\det(\Lambda)^m} \det(S)^{(m-d-1)/2} \mathrm{etr}\left(-\frac{1}{2}(\Lambda\Lambda')^{-1}W\right).$$

Now, let $W = TT'$ be the Cholesky decomposition of W. We seek to compute the distribution of the random lower lower triangular matrix T. To this end, the change of variables $W \mapsto T$ is one-to-one, and has Jacobian determinant equal to $2^d \prod_{i=1}^d t_{ii}^{d-i+1}$. Consequently, the density $p^T_{\gamma,\Lambda}(T)$ of T with respect to the Lebesgue measure is

$$p^T_{\gamma,\Lambda}(T) = \frac{2^d}{2^{md/2}\Gamma_d(m/2)} \det(\Lambda^{-1}T)^m \mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)'\right) \prod_{i=1}^d t_{ii}^{-i}.$$

We recognize $\mathrm{d}\nu(T) = \prod_{i=1}^d t_{ii}^{-i} \mathrm{d}T$ to be a left Haar measure on $\mathcal{L}_+$, and consequently

$$\tilde{p}^T_{\gamma,\Lambda}(T) = \frac{2^d}{2^{md/2}\Gamma_d(m/2)} \det(\Lambda^{-1}T)^m \mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)'\right)$$

is the density of T with respect to $\mathrm{d}\nu(T)$. After this, the density $\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T)$ of the pair (X,T) with respect to $\mathrm{d}X \times \mathrm{d}\nu(T)$ is given by

$$\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T) = \frac{2^d}{K}\frac{\det(\Lambda^{-1}T)^m}{\det(\Lambda)}\operatorname{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}X-\gamma)(\Lambda^{-1}X-\gamma)'\right)$$

with $K = (2\pi)^{d/2}2^{md/2}\Gamma_d(n/2)$. The change of variables $(X,T) \mapsto (T^{-1}X,T)$ has Jacobian determinant equal to $\det(T)$. If $M = T^{-1}X$, then, the density $\tilde{p}^{M,T}_{\gamma,\Lambda}$ of (M,T) with respect to $\mathrm{d}M \times \mathrm{d}\nu(T)$ is given by

$$\tilde{p}^{M,T}_{\gamma,\Lambda}(M,T) = \frac{\det(\Lambda^{-1}T)^{m+1}}{K''}\operatorname{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}TM-\gamma)(\Lambda^{-1}TM-\gamma)'\right).$$

We now marginalize T to obtain the distribution of the maximal invariant M. Since the integral is with respect to the left Haar measure $\mathrm{d}\nu(T)$, we have that

$$\int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,\Lambda}(M,T)\mathrm{d}\nu(T) = \int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,I}(M,\Lambda^{-1}T)\mathrm{d}\nu(T) = \int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,I}(M,T)\mathrm{d}\nu(T),$$

and consequently,

$$\begin{aligned} p^M_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\operatorname{etr}\left(-\frac{1}{2}TT' - \frac{1}{2}(TM-\gamma)(TM-\gamma)'\right)\mathrm{d}\nu(T) \\ &= \frac{2^d}{K}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\operatorname{etr}\left(-\frac{1}{2}T(I+MM')T' + \gamma(TM)'\right)\mathrm{d}\nu(T). \end{aligned}$$

The matrix $I+MM'$ is positive definite and symmetric. It is then possible to perform its Cholesky decomposition $(I+MM') = AA'$. With this at hand, the previous display can be written as

$$p^M_{\gamma,\Lambda}(M) = \frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}}{K}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\operatorname{etr}\left(-\frac{1}{2}(TA)(TA)' + \gamma(TM)'\right)\mathrm{d}\nu(T).$$

We now perform the change of variable $T \mapsto TA^{-1}$. To this end, notice that $\mathrm{d}\nu(A^{-1}) = \mathrm{d}\nu(T)\prod_{i=1}^d a_{ii}^{-(d-2i+1)}$, and consequently

$$\begin{aligned} p^M_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\operatorname{etr}\left(-\frac{1}{2}TT' + \gamma(TA^{-1}M)'\right)\mathrm{d}\nu(T) \\ &= \frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)}\frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\mathbf{P}^T_{m+1}\left[\mathrm{e}^{\langle\gamma,TA^{-1}M\rangle}\right], \end{aligned}$$

so that that at $\gamma = 0$ the density $p^M_{0,\Lambda}(M)$ takes the form

$$p^M_{0,\Lambda}(M) = \frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)} \frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}},$$

and consequently the likelihood ratio is

$$\frac{p^M_{\gamma,\Lambda}(M)}{p^M_{0,\Lambda}(M)} = \mathrm{e}^{-\frac{1}{2}\|\gamma\|^2} \int \mathrm{e}^{\langle \gamma, TA^{-1}M\rangle} \mathrm{d}\mathbf{P}_{m+1}(T).$$

□

Remark A.1.2 (Numerical computation). Computing the optimal E-value is feasible numerically. We are interested in computing

$$\int \mathrm{e}^{\langle x, Ty\rangle} \mathrm{d}\mathbf{P}_{m+1}(T),$$

where T is a $\mathcal{L}^+$-valued random lower triangular matrix such that $TT' \sim W(m+1, I)$, and $x, y \in \mathbb{R}^d$. Define, for $i \geq j$, the numbers $a_{ij} = x_i y_j$. Then $\langle x, Ty\rangle = \sum_{i\geq j} a_{ij}T_{ij}$. By Bartlett's decomposition, the entries of the matrix T are independent and $T_{ii}^2 \sim \chi^2((m+1) - i + 1)$, and $T_{ij} \sim N(0,1)$ for $i > j$. Hence, our target quantity satisfies

$$\int [\mathrm{e}^{\langle x, Ty\rangle}]\mathbf{P}_{m+1}(T) = \int \mathrm{e}^{\sum_{i\geq j} a_{ij}T_{ij}} \mathrm{d}\mathbf{P}_{m+1}(T) = \int \prod_{i\geq j} \mathrm{e}^{a_{ij}T_{ij}} \mathrm{d}\mathbf{P}_{m+1}(T).$$

On the one hand, for the off-diagonal elements satisfy, using the expression for the moment generating function of a standard normal random variable,

$$\mathbf{E}^{\mathbf{P}}_{m+1}[\mathrm{e}^{a_{ij}T_{ij}}] = \exp\left(\frac{1}{2}a_{ij}^2\right).$$

For the diagonal elements the situation is not as simple, but a numerical solution is possible. Indeed, for $a_{ii} \geq 0$, and $k_i = (m+1) - i + 1$

$$\begin{aligned}\mathbf{E}^{\mathbf{P}}_{m}[\mathrm{e}^{a_{ii}T_{ii}}] &= \frac{1}{2^{\frac{k_i}{2}}\Gamma\left(\frac{k_i}{2}\right)} \int_0^\infty x^{\frac{k_i}{2}-1} \exp\left(-\frac{1}{2}x + a_{ii}\sqrt{x}\right) \mathrm{d}x \\ &= {}_1F_1\left(\frac{k_i}{2}, \frac{1}{2}, \frac{a_{ii}^2}{2}\right) + \frac{\sqrt{2}a_{ii}\Gamma\left(\frac{k_i+1}{2}\right)}{\Gamma\left(\frac{k_i}{2}\right)} {}_1F_1\left(\frac{k_i+1}{2}, \frac{3}{2}, \frac{a_{ii}^2}{2}\right),\end{aligned}$$

where ${}_1F_1(a, b, z)$ is the Kummer confluent hypergeometric function. For $a_{ii} < 0$,

$$\frac{1}{2^{k_i/2}\Gamma\left(\frac{k_i}{2}\right)} \int_0^\infty x^{k_i/2-1} \exp\left(-\frac{1}{2}x + a_{ii}\sqrt{x}\right) \mathrm{d}x = \frac{\Gamma(k_i)}{2^{k_i-1}\Gamma\left(\frac{k_i}{2}\right)} U\left(\frac{k_i}{2}, \frac{1}{2}, \frac{a_{ii}^2}{2}\right),$$

and U is Kummer's U function.

A.2. Importance of the filtration for randomly stopped E-Statistics

Consider the the t-test as in Example 2.1.1. Fix some $0 < a < b$, and define the stopping time $\tau^* := 1$ if $|X_1| \notin [a,b]$. $\tau^* = 2$ otherwise. Then clearly τ^* is not adapted to (hence not a stopping time relative to) $(M_n)_n$ as defined in that example, since $M_1 \in \{-1,1\}$ coarsens out all information in X_1 except its sign. Now let $\delta_0 := 0$ (so that $\mathcal{H}_0$ represents the normal distributions with mean $\mu = 0$ and arbitrary variance). Let $T_n^{*,\delta_1}(X^n)$ be equal to the GROW E-statistic $T_n^*(X^n)$ as in (2.13); here we make explicit its dependence on δ_1. For $\mathcal{H}_1$, to simplify computations, we put a prior $\tilde{\mathbf{\Pi}}_1^\delta$ on $\Delta_1 := \mathbb{R}$. We take $\tilde{\mathbf{\Pi}}_1^\delta$ to be a normal distribution with mean 0 and variance κ. We can now apply Corollary 2.8.3 (with prior $\tilde{\mathbf{\Pi}}_0^\delta$ putting mass 1 on $\delta = \delta_0 = 0$), which gives that $\tilde{T}_n(X^n)$ is an E-statistic, where

$$\tilde{T}_n(x^n) = \int \frac{1}{\sqrt{2\pi\kappa^2}} \exp\left(-\frac{\delta_1^2}{2\kappa^2}\right) \cdot T_n^{*,\delta_1}(x^n) \mathrm{d}\delta_1$$

coincides with a standard type of Bayes factor used in Bayesian statistics. By exchanging the integrals in the numerator, this expression can be calculated analytically. The Bayes factor $\tilde{T}_1(x_1)$ for $x^1 = x_1$ is found to be equal to 1 for all $x_1 \neq 0$, and the Bayes factor for (x_1, x_2) is given by:

$$\tilde{T}_2(x_1,x_2) = \frac{\sqrt{2\kappa^2+1}\cdot(x_1^2+x_2^2)}{\kappa^2(x_1-x_2)^2+(x_1^2+x_2^2)}.$$

Now we consider the function

$$f(x) := \mathbf{E}_{X_2 \sim N(0,1)}[\tilde{T}_2(x, X_2)].$$

$f(x)$ is continuous and even. We want to show that, with τ^* as above, $\tilde{T}_{\tau^*}(X^{\tau^*})$ is not an E-variable for some specific choices of a, b and κ. Since, for any $\sigma > 0$, the null contains the distribution under which the X_i are i.i.d. $N(0,\sigma)$, the data may, under the null, in particular be sampled from $N(0,1)$. It thus suffices to show that

$$\mathbf{E}_{X_1,X_2 \sim N(0,1)}[\tilde{T}_{\tau^*}(X^{\tau^*})] = \\ \mathbf{P}_{X_1\sim N(0,1)}\{|X_1| \notin [a,b]\} + \mathbf{E}_{X_1\sim N(0,1)}[\mathbf{1}\{|X_1| \in [a,b]\} f(X_1)] > 1.$$

But from numerical integration we find that $f(x) > 1$ on $[a,b]$ and $[-b,-a]$ if we take $\kappa = 200$, $a \approx 0.44$ and $b \approx 1.70$. Using again numerical integration, we find that the above expectation is then approximately equal to 1.19, which shows that, even though $\tilde{T}_n$ is an E-statistic at each n by Corollary 2.8.3 (it is even a GROW one), $\tilde{T}_{\tau^*}$ is not an E-statistic (its expectation is 0.19 too large), providing the desired counterexample.

B. Appendix to Chapter 3

B.1. Omitted Proofs and Details

In this section we provide proofs and remarks omitted from previous sections. In Appendix B.1.1 we relate growth-rate optimality to the minimum expected stopping time. In Appendix B.1.2, we show that the AV logrank statistic is a continuous-time martingale, and show that this is also true for general patterns of incomplete observation, such as left truncation and filtering as a consequence of the results of Andersen et al. [1993]. In Appendix B.1.3, we proof the claims made in Section 3.3.3 about the martingale structure of the AV logrank test under the presence of ties. Lastly, in Appendix B.1.4, we give further details on the simulations used to compute the planned maximum sample sizes for a given targeted power. Under the alternative and optional stopping, the observed sample size is in many cases lower.

B.1.1. Expected Stopping Time, GROW and Wald's Identity

Here we motivate the GROW criterion by showing that it minimizes, in a worst-case sense, the expected number of events needed before there is sufficient evidence to stop. Let $\mathbf{P}_0$ represent our null model, and let, as before, the alternative hypothesis be $\mathcal{H}_1 : \theta \leq \theta_1$ for some $\theta_1 < \theta_0$. Suppose we perform a level-α test based on a test martingale $S^q_{\theta_0,t}$ using the stopping rule τ that stops as soon as $S^q_{\theta_0,t}$ exceeds the threshold $1/\alpha$, that is, $\tau^q = \inf_t\{t : S^q_{\theta_0,t} \geq 1/\alpha\}$. In the main text we elaborated on how $S^{\theta_1}_{\theta_0,t}$ is optimal with respect to the GROW criterion. We now show that the problem of minimizing the worst-case, the expected number of events $\mathbf{E}_\theta[\bar{N}_{\tau^q}]$ over q is approximately equivalent to finding the GROW test martingale. To do so, we make simplifying assumptions that reduce the problem to an i.i.d. experiment. This allows us to employ a standard argument based on an identity of Wald [1947], originally due to Breiman [1961]. For this we assume that the initial risk sets (i.e., $\bar{y}^A_0$ and $\bar{y}^B_0$) are large enough so that, for all sample sizes we will ever encounter, $\bar{y}^A_t/\bar{y}^B_t \approx \bar{y}^A_0/\bar{y}^B_0$. This allows us to treat the likelihood of the participant(s) $I_{(k)}$ having witnessed the event at time $T^{(k)}$ to be independent of t, that is, as an i.i.d. experiment.

The argument of Breiman [1961] relates the expected number of events to the expected value of our stopped AV logrank statistic. Suppose first that we happen to know that the data come from a specific θ in the alternative hypothesis. Then $S^q_{\theta_0,\tau}$ is the product of $\bar{N}_\tau$ factors of ratios $R^q_{\theta_0,(i)} = q_{(i)}(I_{(i)})/p_{\theta_0,(i)}(I_{(i)})$ at the ith event.

Wald's identity applied to its logarithm implies

$$\mathbf{E}_\theta[\bar{N}_\tau] = \frac{\mathbf{E}_\theta[\ln S^q_{\theta_0,\tau^q}]}{\mathbf{E}_\theta[\ln R^q_{\theta_0,(1)}]}. \tag{B.1}$$

For simplicity we will further assume that the number of participants at risk is large enough so that the probability that we run out of data before we can reject is negligible. Because of the choice of the stopping rule τ^q, the right-hand side of the last display can then be further rewritten as

$$\frac{\mathbf{E}_\theta[\ln S^q_{\theta_0,\tau^q}]}{\mathbf{E}_\theta[\ln R^q_{\theta_0,(1)}]} = \frac{\ln(1/\alpha) + \text{VERY SMALL}}{\mathbf{E}_\theta\left[\ln\left(q_{(1)}(I_{(1)})/p_{(1),\theta_0}(I_{(1)})\right)\right]},$$

where VERY SMALL between 0 and $\log|\theta_1/\theta_0|$. The equality follows because we reject as soon as $S^q_{\theta_0,t} \geq 1/\alpha$, so $S^q_{\theta_0,\tau}$ cannot be smaller than $1/\alpha$, and it cannot be larger by more than a factor equal to the maximum likelihood ratio at a single outcome (if we would not ignore the probability of stopping because we run out of data, there would be an additional small term in the numerator).

With (B.1) at hand, we can relate our choice of q to the expected number of events witnessed before stopping. If, for a fixed θ, we try find the q that minimizes the expected number of events $\mathbf{E}_\theta[\bar{N}_{\tau^q}]$, and, as is customary in sequential analysis, we approximate the minimum by ignoring the VERY SMALL part, we see that the expression is minimized by maximizing the numerator $\mathbf{E}_\theta\left[\ln\left(Q_{(1)}/P_{\theta_0,(1)}\right)\right]$ over q. The maximum is achieved by $Q_{(1)} = P_{\theta,(1)}$; the expression in the denominator then becomes the Kulback-Leibler divergence between two Bernoulli distributions. It follows that, under θ, the expected number of outcomes until rejection is minimized by $Q_{(1)} = P_\theta$. Thus, in this case, we use the GROW $S^\theta_{\theta_0,t}$ as test statistic. However, we still need to consider the fact that the real $\mathcal{H}_1$ is composite: as statisticians, we do not know the actual θ; we only know $0 < \theta \leq \theta_1$. A worst-case approach uses the q achieving

$$\max_q \min_{\theta \leq \theta_1} \mathbf{E}_\theta\left[\ln\left(p_{(1)}(I_{(1)})/q_{(1),\theta_0}(I_{(1)})\right)\right]$$

since, repeating the reasoning leading to (B.1), this q should be close to achieving the min-max number of events until rejection, given by

$$\min_q \max_{\theta \leq \theta_1} \mathbf{E}_\theta[\bar{N}_{\tau^q}]$$

But this just tells us to use the GROW E-variable relative to $\mathcal{H}_1$, which is what we were arguing for.

B.1.2. Continuous time and anytime validity

In this section, we show the anytime validity of the AV logrank test. This is done via Ville's inequality for which it suffices to show that $S^q_{\theta_0} = (S^q_{\theta_0,t})_{t \geq 0}$ is a nonnegative (super) martingale. To do so, we use the counting process formalism. A few definitions are in order. Only in this section, we assume knowledge of counting process theory [see Andersen et al., 1993, Fleming and Harrington, 2011]. Denote, for $i = 1, \dots, m$, $\tilde{N}_t^i = \mathbf{1}\left\{t \le T^i\right\}$ the counting processes associated to each participant, and let y_t^i be the at-risk process. For each participant, the censored process N_t^i, which is observed, is given by $\mathrm{d}N_t^i = y_t^i \mathrm{d}\tilde{N}_t^i$—we use this convention to signify that $N_t^i = \int_0^t y_s^i \mathrm{d}\tilde{N}_s^i$. We define the sigma-algebra $\mathcal{F}_t := \sigma(N_s^j : 0 \le s \le t, j = 1, \dots, n)$, which, as usual, can be interpreted as the information in the study up to time t.

One of the results of the counting process theory is that the processes $\mathrm{d}N_t^i - y_t^i \mathrm{d}\lambda_t^i$ are martingales, where, recall, $y_t^i = \mathbf{1}\left\{X^i \ge t\right\}$ is the at-risk process, and λ_t^i is the hazard function associated to T^i. In that case, $y_t^i \mathrm{d}\lambda_t^i$ is called the compensator of N_t^i. The result that the AV logrank test is a martingale hinges specifically on this structure. Thus, any pattern that preserves this martingale structure also preserves the martingale property for the AV logrank test, and consequently its type-I error guarantees. Andersen et al. [1993, III.4] show exactly this under general patterns of incomplete observation provided that the mechanisms are independent of the observations. With this in mind, in the following, we only assume that the counting processes N_t^i have compensators A_t^i given by $\mathrm{d}A_t^i = y_t^i \mathrm{d}\lambda_t^i$.

The filtration $\mathcal{F} = (\mathcal{F}_s)_{s \ge 0}$ is right-continuous and we can safely identify predictable processes with left-continuous process. For some θ_0, denote by $\mathbf{P}_0$ the distribution under which, for each $i = 1, \dots, m$, the hazard function for T^i is $\lambda_t^i = \theta_0^{z^i} \lambda_t^A$, where $g^i = 0$ if $i \in A$ and $g^i = 1$ if $i \in B$. Recall from Section 3.2, if participant i belongs to Group B, $\lambda_t^i = \theta_0 \lambda_t^B = \theta_0 \lambda_t^A$; otherwise, $\lambda_t^i = \lambda_t^A$. Let $q_t^1, \dots, q_t^m$ be predictable processes such that $\sum_{i \le m} q_t^i y_t^i = 1$ a.s. for all t, that is, $\{q_t^i\}_{i \in \mathcal{R}_t}$ at each t is a probability distribution over the participants at risk at time t. Define r_t^i to be each of the ratios $r_t^i = q_t^i / p_{\theta_0,t}^i$. Define the predictable process $S^q_{\theta_0,t^-} = \lim_{s \uparrow t} S^q_{\theta_0,t^-}$. As such, at each t, the change $\mathrm{d}S^q_{\theta_0,t} = S^q_{\theta_0,t} - S^q_{\theta_0,t^-}$ of the AV logrank statistic $S^q_{\theta_1}$ at time t, given in (3.10), can be computed as

$$\mathrm{d}S^q_{\theta_0,t} = \sum_{i \le m} S^q_{\theta_0,t^-}(r_t^i - 1)\mathrm{d}N_t^i,$$

because no two events happen simultaneously with positive probability. Since $S^q_{\theta_0,t^-}$ is predictable, it is enough to prove that the process M_t defined by $\mathrm{d}M_t = \sum_{i \le m}(1 - r_t^i)\mathrm{d}N_t^i$ is a martingale [see Fleming and Harrington, 2011, Theorem 1.5.1]. Recall that $\bar{y}_t^A = \sum_{i \in A} y_t^i$ and $\bar{y}_t^B = \sum_{i \in B} y_t^i$. Then both $\bar{y}^A$ and $\bar{y}^B$ are left-continuous processes.

Lemma B.1.1. Let $\{q_t^i\}_{i \le m}$ be a collection of nonnegative left-continuous processes $q^i = (q_t^i)_{t \ge 0}$ such that $\sum_{i \le m} y_t^i q_t^i = 1$ for all t. Let $\{p_{\theta_0,t}^i\}_{i \le m}$ be the collection of processes given by

$$p_{\theta_0,t}^i = \frac{\theta_0^{g^i} y_t^i}{\bar{y}_t^A + \theta_0 \bar{y}_t^B}.$$

The process $M = (M_t)_{t \ge 0}$ given by $\mathrm{d}M_t = \sum_{i \le m}(1 - r_t^i)\mathrm{d}N_t^i$ is a martingale under $\mathbf{P}_0$ with respect to the filtration $\mathcal{F} = (\mathcal{F}_t)_{t \ge 0}$.

Proof. It suffices to show that the compensator A_t of M_t, given by $\mathrm{d}A_t = \sum_{i \le m} \sum_{i \le m}(r_t^i - 1) y_t^i \lambda_t^i \mathrm{d}t$ is zero. Define $\bar{q}_t^A = \sum_{i \in A} y_t^i q_t^i$ and $\bar{q}_t^B = \sum_{i \in B} y_t^i q_t^i$. Notice that by assumption

$\bar{q}_t^A + \bar{q}_t^B = 1.$, and recall that, under the null $\lambda_t^B = \theta_0 \lambda_t^A$. We can compute

$$
\begin{aligned}
\sum_{i\leq m} (r_t^i - 1) y_t^i \lambda_t^i &= \sum_{i\in A} y_t^i \lambda_t^A (r_t^i - 1) + \sum_{i \in B} y_t^i \lambda_t^B (r_t^i - 1) \\
&= \lambda_t^A [(\bar{y}_t^A + \theta_0 \bar{y}_t^B) \bar{q}_t^A - \bar{y}_t^A + \\
&\qquad (\bar{y}_t^A + \theta_0 \bar{y}_t^B) \bar{q}_t^B - \theta_0 \bar{y}_t^B] \\
&= \lambda_t^A [(\bar{y}_t^A + \theta_0 \bar{y}_t^B) \overbrace{(\bar{q}_t^A + \bar{q}_t^B)}^{=1} - \\
&\qquad (\bar{y}_t^A + \theta \bar{y}_t^B)] \\
&= 0,
\end{aligned}
$$

where we used the assumption that $\sum_{i\leq m} y_t^i q_t^i = \bar{y}_t^A q_t^A + \bar{y}_t^B q_t^B = 1$. As the compensator A_t of M_t is zero at each t, we conclude that M_t is a martingale, as was to be shown. □

Our previous discussion and the preceding lemma have the following corollary as a consequence.

Corollary B.1.2. $S_{\theta_0}^q = (S_{\theta_0,t}^q)_{t\geq 0}$ *is a nonnegative martingale with expected value equal to one.*

Hence, Ville's inequality holds for $S_{\theta_0}^q$, which implies that

$$\mathbf{P}_0\{S_{\theta_0,t}^q \geq 1/\alpha \text{ for some } t \geq 0\} \leq \alpha.$$

This implies the anytime validity of the test $\xi_{\theta_0}^q = (\xi_{\theta_0,t}^q)_{t\geq 0}$ given by the AV logrank test $\xi_{\theta_0,t}^q = \mathbf{1}\left\{S_{\theta_0,t}^q \geq 1/\alpha\right\}$.

B.1.3. Ties

The purpose of this section is twofold. Firstly, we prove Lemma 3.3.2. Secondly, we show that the conditional likelihood given in Section 3.3.3 indeed approximates the true conditional partial likelihood ratio under any distribution such that the hazard ratio is θ_1.

Our general strategy in this case is similar to the one undertaken in the continuous-monitoring case: we build a test martingale with respect to a filtration $\mathcal{G}^\star$, and use Ville's inequality to derive anytime-valid type-I error guarantees. Define, for each $k = 1, 2, \ldots$, the sigma-algebra $\mathcal{G}_k$ generated by all observations made in times $t_1, \ldots, t_k$, that is, $\mathcal{G}_k = \sigma(N_{t_l}^i, \tilde{N}_{t_l}^i \ : \ i = 1, \ldots, m; \ l = 1, \ldots, k)$, and the corresponding filtration $\mathcal{G} = (\mathcal{G}_k)_{k=1,2,\ldots}$. Under Cox's proportional hazard model, conditionally on $\mathcal{G}_{k-1}$, our observations $\Delta\bar{N}_k^A$ and $\Delta\bar{N}_k^B$ are binomially distributed with parameters depending on the hazard function (see Lemma B.1.3 below). By conditioning both on $\mathcal{G}_{k-1}$ and on the total number of events $\Delta\bar{N}_k = \Delta\bar{N}_k^A + \Delta\bar{N}_k^B$, we use the likelihood of having observed $\Delta\bar{N}_k^B$, which follows Fisher's noncentral hypergeometric distribution, as detailed in Corollary B.1.4. We gather these observations in the following two lemmas.

Lemma B.1.3. Conditionally on $\mathcal{G}_{k-1}$, the following hold:

1. The number of events $\Delta\bar{N}_k^A$ has a binomial distribution with parameters $\bar{y}_k^A$ and p_k^A where $p_k^A = 1 - \exp\left(-\int_{t_{k-1}}^{t_k} \lambda_s^A \mathrm{d}s\right)$.

2. The number of events $\Delta\bar{N}_k^B$ has a binomial distribution with parameters $\bar{y}_k^B$ and p_k^B where $p_k^B = 1 - \exp\left(-\theta\int_{t_{k-1}}^{t_k} \lambda_s^A \mathrm{d}s\right)$ and θ is the hazard ratio.

Proof. The result is standard, and it follows from explicitly solving for λ in (3.1) and computing the conditional probability in (3.2) for each group. □

Next, we use a standard result: given two binomially distributed random variables X and Y, the distribution of X conditionally on $X+Y$ is Fisher's noncentral hypergeometric distribution. We apply this to $\Delta\bar{N}_k^A$ and $\Delta\bar{N}_k^B$ from the previous lemma and spell out the corresponding parameters in the following corollary.

Corollary B.1.4. *Let $\mathcal{G}_{k-1}^\star = \mathcal{G}_{k-1} \vee \sigma(\Delta\bar{N}_k)$, and let p_k^A and p_k^B be as in Lemma B.1.3. Define the odd ratios $\omega_k^A = p_k^A/(1-p_k^A)$, $\omega_k^B = p_k^B/(1-p_k^B)$ and the ratio $\omega_k = \omega_k^B/\omega_k^A$. Then, conditionally on $\mathcal{G}_{k-1}^\star$, the likelihood of having observed $\Delta\bar{N}_k^B$ events in group B is given by Fisher's noncentral hypergeometric distribution with probability mass function $p_{\mathrm{FNCH}}(\Delta\bar{N}_k^B\ ;\ \bar{y}_{k-1}^B, \bar{y}_{k-1}^A, \Delta\bar{N}_k, \omega_k)$ given by*

$$p_{\mathrm{FNCH}}(n^B;\ \bar{y}^B, \bar{y}^A, n, \omega) = \frac{\binom{\bar{y}^B}{n^B}\binom{\bar{y}^A}{n-n^B}\omega^{n^B}}{\sum_{\max\{0, n^B-\bar{y}^B\} \le u \le \min\{\bar{y}^B, n^B\}} \binom{\bar{y}^B}{u}\binom{\bar{y}^A}{n^B-u}\omega^u}.$$

Naively, one could use a partial likelihood ratio just as in the absence of ties to derive a sequential test. This, however, is not satisfactory, because, in general, the parameter ω_k depends heavily on the unknown baseline hazard function λ^A. Contrary to the general case, when the hazard ratio θ is one, the parameter $\omega_k = 1$, and Fisher's noncentral hypergeometric distribution reduces to the conventional hypergeometric distribution. With this observation at hand, if $\{q_k\}_{k=1,2,\ldots}$ is a sequence of conditional distributions $q_k(\ \cdot\ )$ on the possible values of $\Delta\bar{N}_k^B$, we can build a sequential tests for (3.3), with its corresponding type-I error guarantee. We give the details in the following corollary, and subsequently point at a useful choice for q that approximates the real likelihood.

The choice of q for our statistic presented in Section 3.3.3 follows from an approximation of the parameter ω for small $\Delta t_k = t_k - t_{k-1}$. As noted by Mehrotra and Roth [2001], if $\int_{t_{k-1}}^{t_k} \lambda_1(s)\mathrm{d}s$ is small, then $p_k^A \approx \lambda_{t_{k-1}}\Delta t_k$ and $p_k^A \approx \theta p_k^A$. With these two approximations, $\omega_k \approx \theta$. This means that the choice $q_k(\Delta\bar{N}_k^B) = p_{\theta_1,k}(\Delta\bar{N}_k^B) := p_{\mathrm{FNCH}}(\Delta\bar{N}_k^B;\ \bar{y}_k^B,\ \bar{y}_k^A,\ \Delta\bar{N}_k,\ \omega = \theta_1)$ approximates the real conditional likelihood under any alternative for which the true hazard ratio is θ_1. Hence, the sequentially computed statistic

$$S_k^{\theta_1} = \prod_{l\le k} \frac{p_{\theta_1,k}(\Delta\bar{N}_k^B)}{p_{1,k}(\Delta\bar{N}_k^B)}$$

approximates the true partial likelihood ratio of the data observed up to time t_k in the presence of ties, and we recommend its use.

B.1.4. Details of sample size comparison simulations

In this section we lay out the procedure that we used to estimate the expected and maximum number of events required to achieve a predefined power as shown in Figure 3.4 and Figure 3.1 in Section 3.7. First we describe how we sampled the survival processes under a specific hazard ratio. We then describe how we estimated the maximum and expected sample size required to achieve a predefined power (80% in our case) for any of the test martingales that we considered (that of the exact AV logrank, its Gaussian approximation, and the prequential plugin variant). Finally, we explain how the same quantitiees for the classical logrank test and the O'Brien-Fleming procedure were obtained.

In order to simulate the order in which the events in a survival processes happens, we used the sequential-multinomial risk-set process from Section 3.3. As before, we consider the general testing problem with $\theta_0 = 1$ and a minimal clinically relevant effect size $\theta_1 < 1$, and we denote the true data generating parameter by θ, typically, $\theta \leq \theta_1$. Under θ, the odds of the next event at the i^{th} event time happening in Group B are $\theta \bar{y}^B_{(i)} : \bar{y}^A_{(i)}$—the odds change at each time step. Thus, simulating in which group the next event happens only takes a biased coin flip. For the problem of testing (3.11) with θ_0 we fix the tolerate a type-I error to $\alpha = 0.05$ and the type-II error to $\beta = 0.2$. For each test martingale $S^q_{\theta_0}$ of interest we first consider the stopping rule $\tau^q = \inf\{k : S^q_{\theta_0,(k)} \geq 1/\alpha\}$, that is, we stop as soon as $S^q_{\theta_0,(i)}$ crosses the threshold $1/\alpha$. Recall that in the worst case, $\theta = \theta_1$ the expected stopping time τ^q is lowest when we use $S^{\theta_1}_{\theta_0,(k)}$, see Appendix B.1.1.

To estimate the maximum number of events needed to achieve a predefined power with a given test martingale, we turned our attention to a modified stopping rule $\tilde{\tau}^q$. Under $\tilde{\tau}^q$ we stop at the first of two moments: either when our test martingale $S^q_{\theta_0,(k)}$ crosses the threshold $1/\alpha$ (i.e., at τ) or once we have witnessed a predefined maximum number of events $n_{\max}$. More compactly, this means using the stopping rule $\tilde{\tau}^q$ given by $\tilde{\tau}^q = \min(\tau^q, n_{\max})$. In those cases in which the test based on the stopping rule τ^q achieves a power higher than $1 - \beta$, a maximum number of events $n_{\max}$ smaller than the initial size of the combined risk groups can be selected to achieve approximate power $1 - \beta$ using the rule $\tilde{\tau}^q$.

A quick computation shows that $n_{\max}$ has the following property: it is the smallest number of events n such that stopping after n events has probability smaller than $1-\beta$ under the alternative hypothesis, that is,

$$\mathbf{P}_\theta\{\tau^q \geq n\} \leq 1 - \beta.$$

More succinctly, $n_{\max}$ is the (approximate) $(1 - \beta)$-quantile of the stopping time τ^q, which can be estimated experimentally in a straightforward manner.

To estimate $n_{\max}$ for an initial risk set sizes m_1, m_0, we sampled 10^4 realizations of the survival process (under θ) using the method described at the beginning of this section. This allowed us to obtain the same number of realizations of the stopping time τ^q. We then computed the $(1-\beta)$-quantile of the simulated first passage time distribution of τ^q, and reported it as an estimate of the number of events $n_{\max}$ in the 'maximum' column in Figure 3.4.

We assessed the uncertainty in the estimation $n_{\max}$ using the bootstrap. We performed 1000 bootstrap rounds on the sampled empirical distribution of τ^q, and found that the number of realizations that we sampled (10^4) was high enough so that plotting the uncertainty estimates was not meaningful relative to the scale of our plots. For this reason we omitted the error bars in Figure 3.4 and Figure 3.1.

In the "mean" column of Figure 3.4 and Figure 3.1 we plot an estimate of the expected number of events $\tilde{\tau}^q = \min(\tau^q, n_{\max})$. For this, we used the empirical mean of the stopping times that were smaller than $n_{\max}$ on the sample that we obtained by simulation, with 20% of the stopping times being $n_{\max}$ itself. In the "conditional mean" column, we plot an estimate of $\tilde{\tau}^q \mid \tilde{\tau}^q < n_{\max}$, i.e., the stopping time given that we stop early (and hence reject the null).

For comparison, we also show the number of events that one would need under the Gaussian non-sequential approximation of Schoenfeld [1981], and under the continuous monitoring version of the O'Brien-Fleming procedure. In order to judge Schoenfeld's approximation, we report the number of events required to achieve 80% power. This is equivalent to treating the logrank statistic as if it were normally distributed, and rejecting the null hypothesis using a z-test for a fixed number of events. The power analysis of this procedure is classic, and the number of events required is $n^S_{\max} = 4(z_\alpha + z_\beta)^2/\log^2\theta_1$, where z_α, and z_β are the α, and β-quantiles of the standard normal distribution. In the case of the continuous monitoring version of O'Brien-Fleming's procedure, we estimated the number of events $n^{OF}_{\max}$ needed to achieve 80% as follows. For each experimental setting (m^A, m^B, θ), we generated 10^4 realizations of the survival process under θ and computed the corresponding trajectories of the logrank statistic. For each possible value n of $n^{OF}_{\max}$, we computed the fraction of trajectories for which the O'Brien-Fleming procedure correctly stopped when used with the maximum number of events set to n. We report as an estimate of the true $n^{OF}_{\max}$ the first value of n for which this fraction is higher than 80%, our predefined power.

B.2. Covariates: the full Cox Proportional Hazards E-Variable

We extend the AV logrank test to the situation when time-dependent covariates are present, as done in Section 3.3 with the same notation used there. Assume now the presence of d covariates and let, for each participant i, $\mathbf{z}^i_t = (z^i_{t,1}, \ldots, z^i_{t,d})$ be the covariate vector consisting left-continuous time-dependent covariates $z^i_{t,1}, \ldots, z^i_{t,d}$. Denote by $\mathbf{z}^i_{(k)}$ the value of the covariates of participant i at the time $T_{(k)}$ when the kth event is witnessed. We let random variable $I_{(k)}$ denote the index of the patient to which the kth event happens, and consider the extended process $I_{(1)}, I_{(2)}, \ldots$ where the information that is available at time $T_{(k)}$ is, $I_{(1)}, I_{(2)}, \ldots, I_{(k)}$, and $z_{(1)}, \ldots, z_{(k)}$. The conditional partial likelihood underlying the process is now denoted $\mathbf{P}_{\beta,\theta}$ with $\theta > 0$, $\boldsymbol{\beta} \in \mathbb{R}^d$, and $\beta_\theta = \ln \theta \in \mathbb{R}$, defined as follows:

$$
\begin{aligned}
&\mathbf{P}_{\boldsymbol{\beta},\theta}\{I_{(k)} = i \mid \mathbf{z}^j_{(l)}, y^j_{(l)};\ j = 1, \ldots, n;\ l = 1, \ldots, k\} := \\
&\qquad\qquad P_{\beta,\theta}\{I_{(k)} \mid \mathbf{z}^i_{(k)}, y^i_{(k)};\ i = 1, \ldots, n\}, \text{ and} \\
&\mathbf{P}_{\boldsymbol{\beta},\theta}\{I_{(k)} = i \mid \mathbf{z}^i_{(k)}, y^i_{(k)};\ i = 1, \ldots, n\} := \\
&p_{\boldsymbol{\beta},\theta,(k)}(\ i\ ) := \frac{y^i_{(k)} \exp\left(\langle \boldsymbol{\beta}, \mathbf{z}^i_{(k)} \rangle + g^i \beta_\theta\right)}{\sum_{j \in \mathcal{R}_{(k)}} \exp\left(\langle \boldsymbol{\beta}, \mathbf{z}^j_{(k)} \rangle + g^i \beta_\theta\right)},
\end{aligned}
$$

This is consistent with Cox' (1972) proportional hazards regression model: the probability that the ith participant witnesses an event, assuming he/she is still at risk, is proportional to the exponentiated weighted covariates, with group membership being one of the covariates. In case $\boldsymbol{\beta} = 0$, this is easily seen to coincide with the definition of $\mathbf{P}_\theta$ via (3.5) with $\theta = \mathrm{e}^{\beta_\theta}$.

B.2.1. E-Variables and Martingales

Let $\mathbf{W}$ be a prior distribution on $\beta \in \mathbb{R}^d$ for some $d > 0$. ($\mathbf{W}$ may be degenerate, i.e., put mass one on a specific parameter vector β_1). For each such $\mathbf{W}$, we let $q_{\mathbf{W},\theta,(k)}$ be the probability distribution on $\mathcal{R}_{(k)}$ defined by

$$
q_{\mathbf{W},\theta,(k)}(\ i\ ) := \int p_{\boldsymbol{\beta},\theta,(k)}(\ i\ ) \mathrm{d}\mathbf{W}(\boldsymbol{\beta}).
$$

Consider a measure ρ on $\mathbb{R}^d$ (e.g., Lebesgue or some counting measure) and we let $\mathcal{W}$ be the set of all distributions on $\mathbb{R}^d$ which have a density relative to ρ, and $\mathcal{W}^\circ \subset \mathcal{W}$ be any convex subset of $\mathcal{W}$ (we may take $\mathcal{W}^\circ = \mathcal{W}$, for example). We define $\tilde{q}_{\leftarrow\mathbf{W},\theta_0}$ to be the *reverse information projection* [Li, 1999] (RIPr) of $q_{\mathbf{W},\theta,(k)}$ on $\{q_{\mathbf{W},\theta_0,(k)} : \mathbf{W} \in \mathcal{W}^\circ\}$, defined as the probability distribution on $\mathcal{R}_{(k)}$ such that

$$
\mathrm{KL}(q_{\mathbf{W},\theta_1,(k)} \| \tilde{q}_{\leftarrow\mathbf{W},\theta_0,(k)}) = \inf_{\mathbf{W}^\circ \in \mathcal{W}^\circ} \mathrm{KL}(q_{\mathbf{W},\theta_1,(k)} \| q_{\mathbf{W}^\circ,\theta_0,(k)}).
$$

We know from Li [1999] and Grünwald et al. [2020] that $\tilde{q}_{\leftarrow\mathbf{W},\theta_0,(k)}$ exists for each k. Grünwald et al. [2020] show, in the context of E-variables for 2×2 contingency tables, that the infimum in the previous display is in fact achieved by some distribution $\mathbf{W}^\star$ with finite support on $\mathbb{R}^d$ if the random variables $y^1_{(k)}, \ldots, y^m_{(k)}$ constituting our random process have a finite range. For given hazard ratios $\theta_0, \theta_1 > 0$, let

$$
R^{\theta_1}_{\mathbf{W},\theta_0,(k)} = \frac{q_{\mathbf{W},\theta_1,(k)}(I_{(k)})}{q_{\leftarrow\mathbf{W},\theta_0,(k)}(I_{(k)})} \tag{B.2}
$$

be our analogue of (3.8).

Theorem B.2.1 (Corollary of Theorem 1 from Grünwald et al. [2020]). *For every prior* $\mathbf{W}$ *on* $\mathbb{R}^d$*, for all* $\boldsymbol{\beta} \in \mathbb{R}^d$*,*

$$\mathbf{E}_{\boldsymbol{\beta},\theta_0}[R^{\theta_1}_{\mathbf{W},\theta_0,(k)} \mid \mathbf{z}^i_{(l)}, y^i_{(l)};\ i = 1,\dots,m;\ l = 1,\dots,k] = \\ \sum_{i \in \mathcal{R}_{(k)}} q_{\boldsymbol{\beta},\theta_0,(k)}(i) \frac{q_{\mathbf{W},\theta_1,(k)}(i)}{q_{\leftarrow \mathbf{W},\theta_0,(k)}(i)} \leq 1$$

so that $R^{\theta_1}_{\mathbf{W},\theta_0,(k)}$ *is an E-variable conditionally on* $\mathbf{z}^i_{(l)}, y^i_{(l)}$ *with* $i = 1,\dots,m;\ l = 1,\dots,k$*.*

Note that the result does not require the prior $\mathbf{W}$ to be well specified in any way: under any $(\boldsymbol{\beta}, \theta_0)$ in the null distribution, even if $\boldsymbol{\beta}$ is completely disconnected to $\mathbf{W}$, $R^{\theta_1}_{\mathbf{W},\theta_0,(k)}$ is an E-variable conditional on past data.

In particular, since the result holds for arbitrary priors, it holds, at the kth event time, for the Bayesian posterior $\mathbf{W}_{k+1} = \mathbf{W}_1 \mid \mathbf{z}^i_{(l)}, y^i_{(l)};\ i = 1,\dots,m;\ l = 1,\dots,k$, based on arbitrary prior $\mathbf{W}_1$ with density w_1, i.e., the density of $\mathbf{W}_{k+1}$ is given by

$$w_{k+1}(\boldsymbol{\beta}) \propto \prod_{l \leq k} q_{\boldsymbol{\beta},\theta,(l)}(I_{(l)}) w_1(\boldsymbol{\beta}).$$

In parallel to the discussion in Section 3.3.1, we can therefore, for each prior $\mathbf{W}_1$, construct a test martingale $S_k := \prod_{l \leq k} R^{\theta_1}_{\mathbf{W}_l,\theta_0,(l)}$ that "learns" $\boldsymbol{\beta}$ from the data, analogously to (3.12), and computes a new RIPr at each event time k.

B.2.2. Finding the RIPr

While it is not clear how to calculate the RIPr $q_{\leftarrow \mathbf{W},\theta_0,(k)}$ in general, it can be well approximated with the efficient algorithm design by Li [1999] and Li and Barron [1999]. Their algorithm is computationally feasible as long as we restrict $\mathbf{W}^\circ_\delta$ to be the set of all priors $\mathbf{W}$ for which $\min_{i \in \mathcal{R}_{(k)}} q_{\mathbf{W},\theta_0,(k)}(i) \geq \delta$, for some $\delta > 0$. In that case, when run for M steps, the algorithm achieves an approximation error of $O(\ln(1/\delta)/M)$, where each step is linear in the dimension d. Since the approximation error is logarithmic in $1/\delta$, we can take a very small value of δ, which makes the requirement less restrictive. Exploring whether the Li-Barron algorithm really allows us to compute the RIPr for the Cox model, and hence $R^{\theta_1}_{\mathbf{W}_k,\theta_0,(k)}$ in practice, is a major goal for future work.

B.2.3. Ties

Without covariates, our E-variables allow for ties correspond to a likelihood ratio of Fisher's noncentral hypergeometric distributions (see Section 3.3.3), the situation is not so simple in the presence of covariates. Although deriving the appropriate extension of the noncentral hypergeometric partial likelihood is possible, one ends up with a hard-to-calculate formula [Peto, 1972]. Various approximations have been proposed in the literature [Cox, 1972, Efron, 1977]. In case these preserve the E-variable and martingale properties, they would retain type-I error probabilities under optional stopping and we could use them without problems. We do not know whether this is the case however; for the time being, we recommend handling ties by putting the events in a worst-case order, leading to the smallest values of the E-variable of interest, as this is bound to preserve the type-I error guarantees.

B.3. Gaussian AV logrank test

In this section we derive the Gaussian AV logrank test of Section 3.4, and investigate the validity of the Gaussian approximation. In Appendix B.3.1, we show that this approximation is only valid when the allocation of participants to each group under investigation is balanced, that is, when $m^A = m^B$. In Appendix B.3.2 we investigate numerically the sample size needed to reject the null hypothesis under both the exact AV logrank test and its Gaussian approximation.

We start with the derivation of (3.15). For this we use (local) asymptotic normality of the Z-score (3.14). Under the null distribution, Z_k from (3.14) has an asymptotic standard Gaussian distribution. Under any alternative distribution under which the hazard ratio is θ, Schoenfeld [1981] showed that, in the absence of ties, the Z-statistic also follows a Gaussian distribution with unit variance, but this time with mean $\mu_1^\star$ given by

$$\mu_1^\star = \frac{\sum_{i\leq k} E_i^B(1-E_i^B)}{\sqrt{\sum_{i\leq k} E_i^B(1-E_i^B)}} \log(\theta).$$

Note that $\mu_1^\star$ depends on more than the summary statistic Z_k. In the case that the number of observed events is much smaller than the initial risk set sizes, the mean $\mu_1^\star$ under the alternative can be further approximated by

$$\mu_1^\star \approx \sqrt{\bar{N}_k}\mu_1 = \sqrt{\bar{N}_k}\sqrt{\frac{m^B m^A}{(m^B+m^A)^2}} \log(\theta), \tag{B.3}$$

where $\bar{N}_k$ is the total number of observations up until time t_k, and the resulting approximation only depends on summary statistics. It is exactly this value μ_1 that we use in the Gaussian AV logrank test. The asymptotic result of Schoenfeld relies on two conditions: (1) that the hazard ratio θ_1 under the alternative is close enough to one so that a first-order Taylor approximation around $\theta_0 = 1$ is adequate; (2) that the expected number of events E_k^B stays approximately constant over time, that is, close to the initial allocation proportion $E_1^B = m^B/(m^B + m^A)$. This indicates that the asymptotic approximation is reasonable for values of θ_1 close to 1 and the initial risk sets are both large in comparison to the number of events witnessed. Notice that in this regime of large risk sets the multiplicity correction in V_k is also negligible.

This raises the question whether a sequential Gaussian approximation is sensible for the logrank statistic— a priori it is not at all clear whether Schoenfeld's asymptotic fixed-sample result has a nonasymptotic counterpart. Define the the logrank statistic per observation time

$$Z_i = \frac{O_i^B - E_i^B}{\sqrt{V_i^B}}.$$

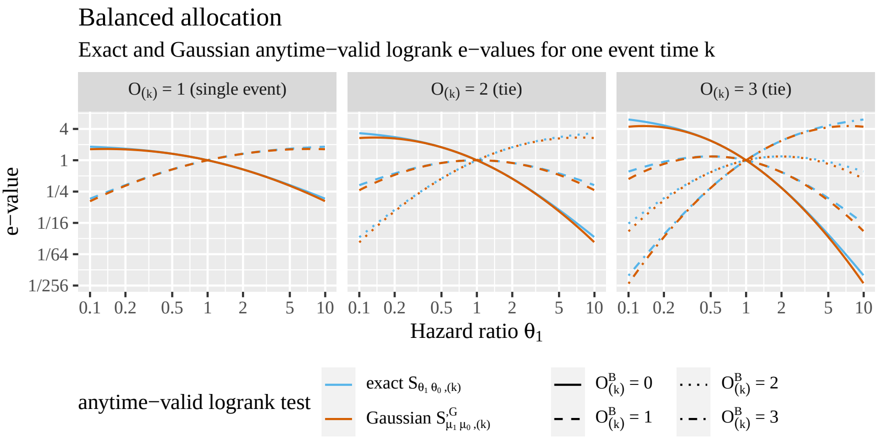

Figure B.1.: For balanced allocation ($m^A = m^B$) $S_1'^G$ is very similar to $S_{(1)}$ when $0.5 \leq \theta_1 \leq 2$. Here $\theta_0 = 1$, $\mu_0 = 0$, and $\mu_1 = \mu_1(\theta_1)$ as in in (B.3). Note that both axis are logarithmic.

We investigate whether the exact AV logrank statistic behaves similarly to the Gaussian likelihood ratio

$$S_k'^G = \prod_{i \leq k} \frac{\phi_{\mu_1 \sqrt{O_i}}(Z_i)}{\phi_{\mu_0}(Z_i)} = \exp\left(-\frac{1}{2}\sum_{i \leq k}\left\{O_i \mu_1^2 - 2\mu_1 \sqrt{O_i Z_i}\right\}\right)$$

for $\theta_0 = 1$ we have $\mu_0 = 0$, $\mu_1 = \log(\theta)\sqrt{m^B m^A/(m^A + m^B)^2}$, and ϕ_μ is the Gaussian density with unit variance and mean μ. Note that the statistic still depends on elements of the full data set; more approximations are needed. Write the Gaussian densities, and use that in the limit of large risk sets $p_i^B \approx m^B/(m^A + m^B)$ and that consequently $V_i \approx \sqrt{O_i \frac{m^A m^B}{(m^A+m^B)^2}}$. This approximations valid under Schoenfeld's second assumption. With these approximations at hand, the Z-statistic is approximated by

$$Z_k \approx \frac{\sum_{i\leq k}\left\{O_i^B - E_i^B\right\}}{\sqrt{O_i \frac{m^A m^B}{(m^A+m^B)^2}}}$$

and consequently

$$S'^G_k \approx S_k^G \approx S_k^G = \exp\left(-\frac{1}{2}\bar{N}_k \mu_1^2 + \sqrt{\bar{N}_k}\mu_1 Z_k\right),$$

where S_k^G is as in (3.15). In Figure B.1 we show, in case of balanced allocation, that the Gaussian approximation S_k^G a single event time from the Gaussian approximation are very similar to the exact $S_{\theta_0,(k)}^{\theta_1}$ for alternative hazard ratios θ_1 between 0.5 and 2.

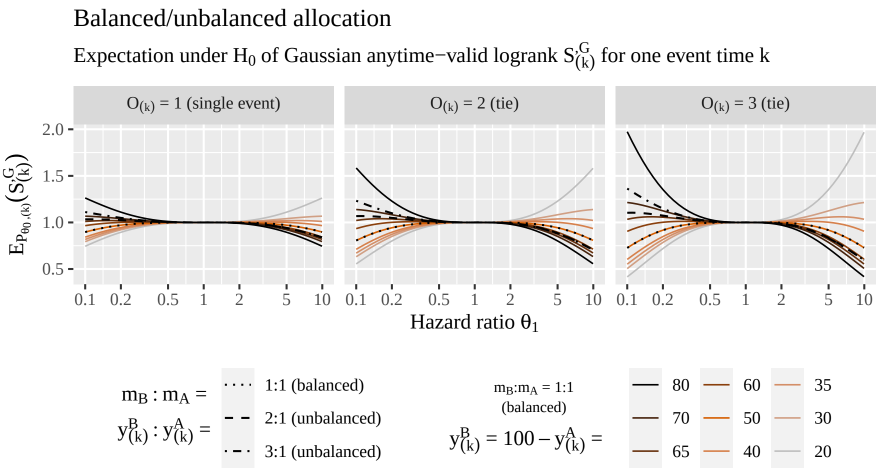

Figure B.2.: Expected value of the increments of the Gaussian AV logrank statistic as a function of the hazard ratio θ_1. For balanced allocation R_i^G is an E-variable, but it is not for unbalanced allocation. The risk set can also start out balanced but become unbalanced; this is unlikely under the null hypothesis (see Appendix B.3.1). Note that the x-axis is logarithmic.

B.3.1. Safety only for balanced allocation

In order to assess whether the Gaussian AV logrank test is indeed AV, that is, whether the type-I error guarantees holds, we inspect whether the expected value of each of its multiplicative increments is bellow 1. In relation to our discussion in Section 3.3.1, this would imply that all multiplicative increments are conditional E-variables and that the resulting test is, at least approximately, a test martingale. Figure B.2 shows the expectation of these increments as a function of the hazard ratio for several initial allocation ratios. In case of balanced 1:1 allocation S_k^G is an E-variable, since its expectation is 1 or smaller. However, in case of unbalanced 2:1 or 3:1 allocation and designs with hazard ratio $\theta_1 < 1$, S_k^G is not an E-variable. Of course, even if the initial allocation is balanced, it can become unbalanced. Figure B.2 shows that in case of designs outside the range $0.5 \leq \theta_1 \leq 2$ the deviations from expectation 1 can be problematic. Hence we do not recommend to use the Gaussian approximation on the logrank statistic for unbalanced designs and designs for $\theta_1 < 0.5$ or $\theta_1 > 2$. For balanced designs with $0.5 \leq \theta_1 \leq 2$, we found that in practice they are safe to use, the reason being that scenarios in which the allocation becomes highly unbalanced after some time (e.g. $y_i^B = 80, y_i^A = 20$) are extremely unlikely to occur under the null.

B.3.2. Sample size

In this section we compare the stopping time distribution $\tau^G := \inf\{k : \xi_k^G = 1\}$ of the Gaussian approximation to that of $\tau = \inf\{k : \xi_k = 1\}$. We use tests with tolerable type I error $\alpha = 0.05$, thus, the threshold $1/\alpha = 20$ for both tests. In the previous section we showed that the Gaussian approximation to the AV logrank statistic is valid when the initial allocation is 1:1 and for values $0.5 \leq \theta_1 \leq 2$, where θ_1 is the hazard ratio under the alternative. In these scenarios, we simulate a survival process from a distribution according to which the true data generating hazard ratio is $\theta = \theta_1$ and sampled realizations τ^G and τ for the same data set. The results of the simulation are shown in Figure B.3, where we plot the realizations of τ^G against those of τ. We see that in most cases both tests reject at the same time $\tau^G = \tau$, and that the approximation becomes better as θ_1 moves closer to $\theta_0 = 1$ (Schoenfeld's assumption 1). When both tests do not reject at the same time, the Gaussian approximation errs on the conservative side. The deviations from the constant large and balanced risk set do not seem to occur often for this range of hazard ratios. After all, the risk set needs to be large to observe the number of events to detect hazard ratios in the range $0.5 \leq \theta_1 \leq 2$.

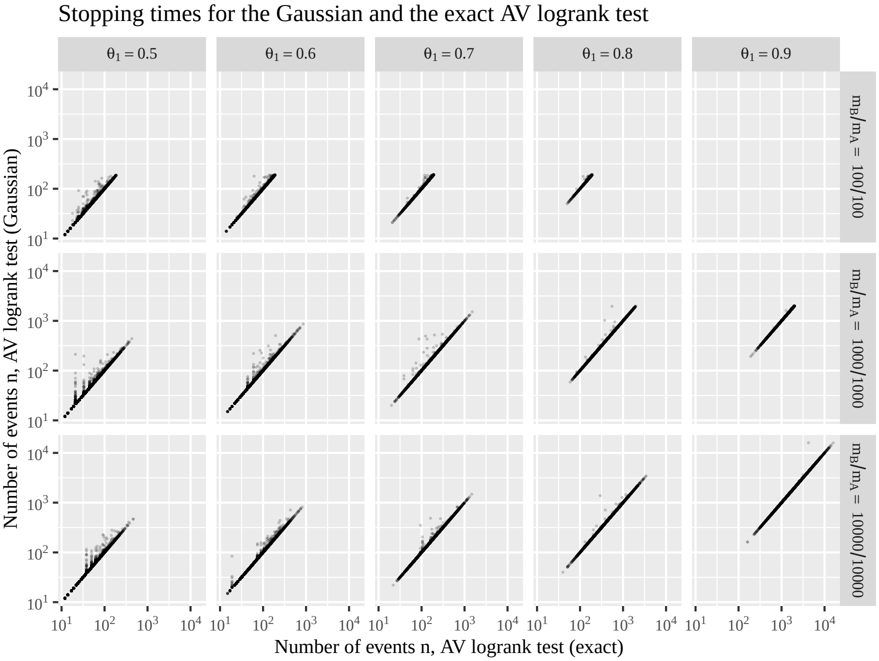

Figure B.3.: Stopping times for the Gaussian and exact AV logrank tests under continuous monitoring (no ties) with threshold $1/\alpha = 20$. The stopping times under the Gaussian approximation often coincide with the exact ones, and are often more conservative (see Appendix B.3.2). Note that both axes are logarithmic.

.

C. Appendix to Chapter 4

C.1. Saddle-Point Computation in Multiscale Games

One application of online learning is computing approximate mixed-strategy Nash equilibria in finite two-player zero-sum games (and more generally, to approximate saddle points of convex-concave functions). Here, we investigate a multiscale version of that problem. Our main focus is to find methods whose performance does not depend on the *maximum scale*, but on the *relevant scale* to the problem instance at hand. In this case, this means the scale of the payoffs in the subset of rows and columns in the support of the Nash equilibrium. In Section C.1.1 we lay out the setup of two-player zero-sum finite games. In Section C.1.2 we define the suboptimality gap, the main measure of performance in judging the solution to these games. In Section C.1.3 we define the payoff matrices used in the experiments that produced Figure 4.4. We conjecture that MUSCADA achieves fast scale-dependent convergence in Section C.1.5 and provide the additional details of the experiments that produced Figure 4.4(right) in Section C.1.6.

C.1.1. Two-player zero-sum finite games

Given a payoff matrix $A \in \mathbb{R}^{K \times M}$ (specifying losses for the row player and gains for the column player) we are looking for the mixed-strategy saddle point $(\boldsymbol{p}_*, \boldsymbol{q}_*) \in \mathcal{P}(K) \times \mathcal{P}(M)$ such that

$$\min_i \boldsymbol{e}_i^\intercal A \boldsymbol{q}_* \;\geq\; \max_j \boldsymbol{p}_*^\intercal A \boldsymbol{e}_j.$$

Our approach will be based on oracle access to the matrix-vector products $\boldsymbol{q} \mapsto A\boldsymbol{q}$ and $\boldsymbol{p} \mapsto A^\intercal \boldsymbol{p}$. We will use the scheme of running two online learners against each other, with loss vectors $\boldsymbol{\ell}_t^{\text{row}} = A\boldsymbol{q}_t$ and $\boldsymbol{\ell}_t^{\text{col}} = -A^\intercal \boldsymbol{p}_t$ and optimistic estimates given by the past loss vector $\boldsymbol{m}_t^{\text{row/col}} = \ell_{t-1}^{\text{row/col}}$. For the same-scale case, Rakhlin and Sridharan [2013] show that uncoupled adaptive schemes benefit from convergence of the gap of the pair of iterate averages at rate $O(\sigma_{\max} \frac{\ln K + \ln M}{T})$, while recently Hsieh et al. [2021] showed last iterate convergence as well. Here we investigate the advantage of using adaptive multiscale learners to improve the dependence in $\sigma_{\max}$.

C.1.2. The metric of success: suboptimality gap

We are looking for the equilibrium in mixed strategies, i.e. $\min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\intercal A \boldsymbol{q}$. The social exploitability of a candidate saddle point pair $\boldsymbol{p}, \boldsymbol{q}$ is defined as the gap

$$\operatorname{gap}(\boldsymbol{p}, \boldsymbol{q}) \;=\; \max_j \boldsymbol{p}^\intercal A \boldsymbol{e}_j - \min_i \boldsymbol{e}_i^\intercal A \boldsymbol{q}.$$

We use the common technique of employing online learning with linear loss functions $\boldsymbol{p} \mapsto A\boldsymbol{q}_t$ and $\boldsymbol{q} \mapsto -A^\intercal \boldsymbol{p}_t$. A standard analysis [Freund and Schapire, 1997] bounds the gap of the iterate averages $\bar{\boldsymbol{p}}_t = \frac{1}{t}\sum_{s\le t} \boldsymbol{p}_s$ and $\bar{\boldsymbol{q}}_t = \frac{1}{t}\sum_{s\le t} \boldsymbol{q}_s$ from above by the social (sum-of) regret

$$\begin{aligned} \operatorname{gap}(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{q}}_t) &= \max_j \bar{\boldsymbol{p}}_t^\intercal A \boldsymbol{e}_j - \min_i \boldsymbol{e}_i^\intercal A \bar{\boldsymbol{q}}_t = \frac{1}{t}\left(\max_j \sum_{s\le t} \boldsymbol{p}_s^\intercal A \boldsymbol{e}_j - \min_i \sum_{s\le t} \boldsymbol{e}_i^\intercal A \boldsymbol{q}_s \right) \\ &= \frac{1}{t} \max_{i,j} \Big(\underbrace{\sum_{s\le t} \boldsymbol{p}_s^\intercal A \boldsymbol{e}_j - \sum_{s\le t} \boldsymbol{p}^\intercal A \boldsymbol{p}_s}_{R_t^{\boldsymbol{q}}(j)} + \underbrace{\sum_{s\le t} \boldsymbol{p}^\intercal A \boldsymbol{p}_s - \sum_{s\le t} \boldsymbol{e}_i^\intercal A \boldsymbol{q}_s}_{R_t^{\boldsymbol{p}}(i)} \Big). \end{aligned}$$

Having multiscale regret bounds at our disposal, it is natural to look at multiscale payoff matrices.

C.1.3. Multiscale structure

We will assume that our payoff matrix is multiscale in the sense that we are given row and column range vectors $\boldsymbol{\sigma}^{\text{row}}$ and $\boldsymbol{\sigma}^{\text{col}}$ such that $|A_{ij}| \le \min\{\sigma_i^{\text{row}}, \sigma_j^{\text{col}}\}$. The main point is to learn the saddle point faster if the maximum range is much larger than the range in the support of the saddle point, i.e. $\sigma_{\max}^{\text{row}} \gg \sigma_{\text{real}}^{\text{row}} := \max\{\sigma_i^{\text{row}} | \boldsymbol{e}_i^\intercal \boldsymbol{p}_* > 0\}$ and/or $\sigma_{\max}^{\text{col}} \gg \sigma_{\text{real}}^{\text{col}} := \max\{\sigma_j^{\text{col}} | \boldsymbol{e}_j^\intercal \boldsymbol{q}_* > 0\}$. We will denote that largest relevant scale by $\sigma_{\text{real}} = \max\left\{\sigma_{\text{real}}^{\text{row}}, \sigma_{\text{real}}^{\text{col}}\right\}$. Our aim is to get gap bounds that scale with σ_{real}, not $\sigma_{\max}$.

Example C.1.1 (Simple multiscale Game). For the purpose of our experiment, we will construct our multiscale payoff matrices following the template

$$A = \begin{bmatrix} B & -\mathbf{1}\mathbf{1}^\intercal \\ \mathbf{1}\mathbf{1}^\intercal & C \end{bmatrix}$$

where B_{ij} are i.i.d. Rademacher $\{\pm 1\}$ and C_{ij} are i.i.d. Rademacher $\{\pm\sigma_{\max}\}$ for some pre-specified $\sigma_{\max} \gg 1$. By construction, any saddle point for the submatrix B is (upon padding with zeros) also a saddle point for the full matrix A. Moreover, it is a strict saddle point for A if it is a strict saddle point for B with value $\min_{\boldsymbol{p}} \max_{\boldsymbol{q}} \boldsymbol{p}^\intercal B \boldsymbol{q} \in (\pm 1)$. We will assume throughout that we are in this latter strict case. Here $\sigma_{\text{real}} = 1$ regardless of $\sigma_{\max}$.

C.1.4. What can one hope to achieve?

Throughout the remainder we assume for simplicity that the saddle point $\boldsymbol{p}_*, \boldsymbol{q}_*$ of the payoff matrix A is unique (a common situation). We define the *optimality gap* of row i by $\delta^{\text{row}}(i) = (\boldsymbol{e}_i - \boldsymbol{p}_*)^\intercal A \boldsymbol{q}_* \ge 0$ and of column j by $\delta^{\text{col}}(j) = \boldsymbol{p}_*^\intercal A(\boldsymbol{q}_* - \boldsymbol{e}_j) \ge 0$. We are interested in scenarios where at least one player has strictly positive optimality gap on the action(s) of largest scale. We will show that multiscale regret bounds allow the learning to accelerate. Moreover, the learner does not need to know about this structure and will adapt automatically.

Let us assume without loss of generality that $\delta^{\text{row}}(k) > 0$ while $\sigma_k^{\text{row}} = \max_i \sigma_i^{\text{row}}$ where $\sigma_i^{\text{row}} = \max_j |A_{i,j}|$. The general idea now is to use that $\bar{\boldsymbol{p}}_T \to \boldsymbol{p}_*$. This means that from some point t on,

$$\max_j \bar{\boldsymbol{p}}_t^\intercal A \boldsymbol{e}_j = \max_{j:\boldsymbol{q}_*(j)>0} \bar{\boldsymbol{p}}_t^\intercal A \boldsymbol{e}_j = \\ \frac{1}{t} \max_{j:\boldsymbol{q}_*(j)>0} \sum_{s\le t} \boldsymbol{p}_s^\intercal A \boldsymbol{e}_j \le \frac{1}{t} \sum_{s\le t} \boldsymbol{p}_s^\intercal A \boldsymbol{q}_s + \max_{j:\boldsymbol{q}_*(j)>0} \frac{1}{t} R_t^{\text{col}}(j)$$

A similar argument for the row player then allows us to conclude

$$\begin{aligned} \operatorname{gap}(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{q}}_t) &\le \frac{1}{t}\left(\sum_{s\le t} \boldsymbol{p}_s^\intercal A \boldsymbol{q}_s + \max_{j:\boldsymbol{q}_*(j)>0} R_t^{\text{col}}(j) - \sum_{s\le t} \boldsymbol{p}_s^\intercal A \boldsymbol{q}_s + \max_{i:\boldsymbol{p}_*(i)>0} R_t^{\text{row}}(i) \right) \\ &= \frac{1}{t}\left(\max_{j:\boldsymbol{q}_*(j)>0} R_t^{\text{col}}(j) + \max_{i:\boldsymbol{p}_*(i)>0} R_t^{\text{row}}(i) \right). \end{aligned}$$

The main point is that this bound scales with $\max_{i:\boldsymbol{p}_*(i)>0} \sigma_i^{\text{row}} + \max_{j:\boldsymbol{q}_*(j)>0} \sigma_j^{\text{col}}$ and not with the respective unconstrained maxima.

Proposition C.1.2. *Any pair of multiscale online learning algorithms with bounds of order $R_t^i \le O(\sigma_i \sqrt{T})$ ensures iterate average gap*

$$\operatorname{gap}(\bar{p}_t, \bar{q}_t) = O(\sigma_{\text{real}}/\sqrt{t})$$

as $t \to \infty$. In particular, this holds for Muscada *with Tuning 3 (see Lemma C.2.1).*

Note that single-scale algorithms would only deliver $\operatorname{gap}(\bar{p}_t, \bar{q}_t) = O(\sigma_{\max}/\sqrt{t})$; a weaker guarantee.

C.1.5. Why our approach may achieve the hope optimistically

Rakhlin and Sridharan [2013] show that using optimism in saddle point interactions can improve the rate to $O(\sigma_{\max}/t)$. We first show that this is true for Muscada as well, after which we will investigate achieving $O(\sigma_{\text{real}}/t)$. The mechanism for this proof is to show that the social regret is constant. Technically, one would explicitly keep track of the slack in (C.5) and (C.6), and use these harvested slacks to cancel the $\sqrt{t}$ term of the regret bound. Only the constant-order term measuring the entropy of the initial weights remains. For this to be a constant, we further need that the learning rate stops decreasing once the regret stabilizes. Following exactly the steps of Rakhlin and Sridharan [2013], we can prove the following proposition.

Proposition C.1.3. *For same-scale games, the optimistic version (see Figure 4.3) of* Muscada *with Tuning 3 and uniform prior (see Lemma C.2.1) achieves average iterate gap gap$(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{q}}_t) = O(\sigma_{\max}/t)$ as $t \to \infty$.*

The same-scale assumption makes all $\boldsymbol{\sigma}$ equal, while the uniform-prior assumption in addition makes all $\boldsymbol{\eta}$ equal. This makes the standard argument from the literature apply.

We further forward the natural conjecture that we state next.

Conjecture C.1.4. For the multiscale case, the optimistic version (see Figure 4.3) of Muscada with Tuning 3 (see Lemma C.2.1) and any nondegenerate prior achieves average iterate gap bounded by $gap(\bar{\boldsymbol{p}}_t, \bar{\boldsymbol{q}}_t) = O(\sigma_{\text{real}}/t)$.

The reason that our Tuning 3 has any chance here is that *no* terms (not even the additive constant) in the regret bound scale with $\sigma_{\max}$. This in contrast to the algorithms of Foster et al. [2017], Cutkosky and Orabona [2018], Bubeck et al. [2019], Chen et al. [2021], whose existing multiscale analyses all result in a lower-order term scaling with $\sigma_{\max}$.[1] We next provide empirical support for our conjecture.

C.1.6. Numerical results

We investigate three algorithms: Hedge with classic time-decreasing learning rate $\eta_t = \sqrt{\frac{\ln(K)}{\sigma_{\max}^2 t}}$, Muscada with all scales set to $\sigma_{\max}$ and Muscada with actual knowledge of the multiscale vectors. All algorithms are run in optimistic mode with guesses $\boldsymbol{m}_t = \ell_{t-1}$, the loss vector of the previous round (and $\boldsymbol{m}_{1,k} = 0$). We choose a matrix of structure given in Example C.1.1, with B and C of size 10×10, and pick $\sigma_{\max} = 100$. We give all algorithms the uniform prior $\pi_k = 1/20$. The results are displayed in Figure C.1, where we show the saddle point gap for the average iterate, the last iterate and the theoretical regret bounds that we obtain from the analysis. In the main text, Figure 4.4(right) shows only the saddle point gap for the average iterate of optimistic Muscada with the optimistic modification of Tuning 3 from Figure C.2. Generating this figure with the code from the supplementary material takes 30 minutes on an Intel i7-7700 processor. Memory usage is negligible.

We see in Figure C.1 that the gap of optimistic Hedge decays at the slow rate $O(\sigma_{\max}/\sqrt{t})$. This means that optimism alone is insufficient to obtain a faster $O(\sigma_{\max}/t)$ convergence rate; it is also necessary that the learning rates stop decreasing when the regret plateaus. It is also apparent that Muscada tuned to $\sigma_{\max}$ has the fast $O(1/t)$ rate, but at the $\sigma_{\max}$ scale. Finally, the numerical experiments show evidence that our multiscale algorithm does exploit the small scale of the actions in the support of the saddle point, exhibiting the desired $O(\sigma_{\text{real}}/t)$ regret conjectured above. The plot also includes the quality of the last iterate. Hsieh et al. [2021] prove convergence of the last iterate for the common scale, common prior case. In our experiment the iterate average can be seen to converge quickly in the multiscale case, but convergence is terribly slow in the same-scale case. This is not inconsistent; no rates are currently known for the last iterate.

C.2. Tuning 3

In this section we describe a third tuning, defined in Figure C.2. In contrast to Tunings 1 and 2, the learning rates in Tuning 3 start *higher*, namely at $1/(2\sigma_k)$ instead of $1/(2\sigma_{\max})$. The downside of this aggressive tuning is that the variance bound is not available (though the weaker, uncentered second-moment analog is). The upside

[1] Which is hard to spot in some of the literature because of a global $\sigma_{\max} = 1$ convention.

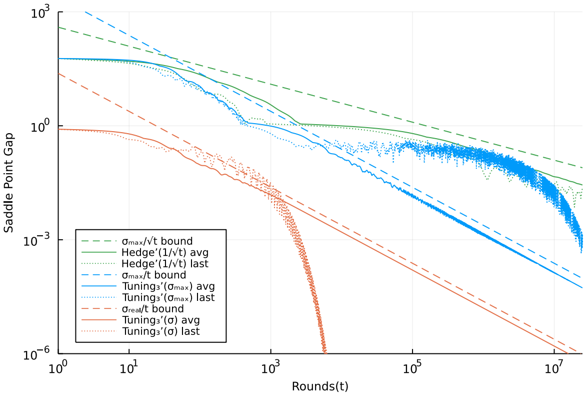

Figure C.1.: Quality of average iterate (solid) and last iterate (dotted) for three optimistic algorithms, compared to their relevant bounds (dashed). The multiscale-aware algorithm (red) outperforms the non-scale-aware competitors by the factor $\sigma_{\max}/\sigma_{\text{real}} = 100$. See Section C.1.6 for further discussion.

Tuning 3 $u = \pi\frac{\sigma_{\min}}{\sigma_k}$, $\eta_{0,k} = \frac{1}{2\sigma_k}$, $\gamma = 8\ln(1/u_k)$, and

$$H_{1,k}(v_t) = \frac{\mathrm{d}}{\mathrm{d}v_t}\left[\frac{v_t}{\sqrt{1+v_t/\gamma_k}}\right] = \frac{v_t/\gamma_k+2}{2(1+v_t/\gamma_k)^{3/2}}.$$

Figure C.2.: Tuning 3 for Muscada

is that the resulting regret bound compared to expert k features only σ_k and has *no* occurrence of $\sigma_{\max}$ whatsoever, not even in the additive constants.

Lemma C.2.1. Let $\boldsymbol{\pi}$ be a probability distribution on K experts. Muscada run with Tuning 3 depicted in Figure C.2 guarantees that, for any $t = 1, 2, \ldots,$

$$R_{t,k} \le 2\sigma_k\sqrt{2v_t\ln(1/u_k)} + c_{\boldsymbol{\sigma},\boldsymbol{\pi}}\sigma_{\min}\sqrt{2v_t} + 8\sigma_k\ln(1/u_k) + 4\sigma_{\min} + \frac{\sigma_k}{2}\max_{s\le t}\Delta v_s, \quad \text{(C.1)}$$

where $c_{\boldsymbol{\sigma},\boldsymbol{\pi}} = \sum_{k\in K}\pi_k(1/\sqrt{\ln(1/u_k)})$ and $u_k = \pi_k\frac{\sigma_{\min}}{\sigma_k}$. Additionally, we have that $v_t \le 4\sum_{s\le t}\frac{\langle\tilde{\boldsymbol{w}}_s,\ell_s^2\rangle}{\langle\tilde{\boldsymbol{w}}_s,\boldsymbol{\sigma}^2\rangle} \le 4t$, where, for each $t = 1, 2, \ldots,$ the weights are $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$.

Proof. Follow the same steps as in the proof of the regret bound for Tuning 1 in

Lemma 4.2.3. Obtain that

$$\mu_{t,k} \le \sigma_k \sqrt{2v_t \ln(1/u_k)} + 4\sigma_k \ln(1/u_k) \tag{C.2}$$

$$\frac{\ln(1/u_k)}{\eta_{t,k}} \le \sigma_k \sqrt{2v_t \ln(1/u_k)} + 4\sigma_k \ln(1/u_k), \text{ and} \tag{C.3}$$

$$\sum_{k\in K} \frac{u_k}{\eta_k} \le c_{\boldsymbol{\sigma},\boldsymbol{\pi}} \sigma_{\min} \sqrt{2v_t} + 4\sigma_{\min} \tag{C.4}$$

with $c_{\boldsymbol{\sigma},\boldsymbol{\pi}} = \sum_{k\in K} \pi_k \left(\frac{1}{\sqrt{\ln(1/u_k)}}\right)$. Use Proposition 4.2.3 to conclude the first claim. For the additional claim, use Lemma C.7.2 with $\lambda = 0$. □

C.3. Algorithm Analysis

The only step in the algorithm that may be problematic is the definition of Δv_t at every round, which one might think can take infinite values. We show in Proposition C.7.1 that this is not the case and that consequently $t \mapsto v_t$ is well defined.

C.3.1. Untuned regret bound, proof of Proposition 4.2.2

We prove that the potential $t \mapsto \Phi_t$ is decreasing for optimistic Muscada. The result for the nonoptimistic version follows by setting the guesses $\boldsymbol{m}_t$ to $\boldsymbol{0}$. Recall from (4.4) in Section 4.2 that the potential Φ_t is defined by

$$\Phi_t \;=\; \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_t) \;=\; \max_{\boldsymbol{w}\in\mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{R}_t - \boldsymbol{\mu}_t\rangle - D_{\boldsymbol{\eta}_t}(\boldsymbol{w}, \boldsymbol{u}).$$

Proof of Lemma 4.2.1. We prove the result in the optimistic case. The nonoptimistic case is recovered for $\boldsymbol{m}_t = \boldsymbol{0}$ and replacing $4\sigma_k^2$, which is a bound on $|m_{t,k} - \ell_{t,k}|^2$, by σ_k^2, which bounds $|\ell_{t,k}|^2$. The result is a consequence of the following inequalities:

$$\begin{aligned}
\Phi_t &\le \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_{t-1}) && \boldsymbol{\eta} \mapsto D_{\boldsymbol{\eta}} \text{ decr.} \quad \text{(C.5)}\\
&= \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_{t-1} - 4\boldsymbol{\eta}_{t-1}\boldsymbol{\sigma}^2 \Delta v_t, \boldsymbol{\eta}_{t-1}) && \text{by def. of } \boldsymbol{\mu}_t\\
&= \Phi(\boldsymbol{R}_{t-1} + \langle \boldsymbol{w}_t, \boldsymbol{\mu}_t\rangle - \boldsymbol{m}_t - \boldsymbol{\mu}_{t-1}, \boldsymbol{\eta}_{t-1}) && \text{by def. of } \Delta v_t\\
&= \max_{\boldsymbol{w}\in\mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{R}_{t-1} + \langle \boldsymbol{w}_t, \boldsymbol{m}_t\rangle - \boldsymbol{m}_t - \boldsymbol{\mu}_{t-1}\rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}, \boldsymbol{u}) && \text{by def. of } \Phi\\
&= \langle \boldsymbol{w}_t, \boldsymbol{R}_{t-1} + \langle \boldsymbol{w}_t, \boldsymbol{m}_t\rangle - \boldsymbol{m}_t - \boldsymbol{\mu}_{t-1}\rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}_t, \boldsymbol{u}) && \text{by def. of } \boldsymbol{w}_t\\
&= \langle \boldsymbol{w}_t, \boldsymbol{R}_{t-1} - \boldsymbol{\mu}_{t-1}\rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}_t, \boldsymbol{u}) && \langle \boldsymbol{w}_t, \boldsymbol{m}_t\rangle \text{ cancels}\\
&\le \max_{\boldsymbol{w}\in\mathcal{P}(K)} \langle \boldsymbol{w}, \boldsymbol{R}_{t-1} - \boldsymbol{\mu}_{t-1}\rangle - D_{\boldsymbol{\eta}_{t-1}}(\boldsymbol{w}, \boldsymbol{u}) && \text{since } \boldsymbol{w}_t \in \mathcal{P}(K)\\
& && \text{(C.6)}\\
&= \Phi(\boldsymbol{R}_{t-1} - \boldsymbol{\mu}_{t-1}, \boldsymbol{\eta}_{t-1}) \;=\; \Phi_{t-1} && \text{by def. of } \Phi,\ \Phi_t.
\end{aligned}$$

Hence, $\Phi_t \le \Phi_{t-1}$, as we were to show. □

Proof of Proposition 4.2.2. Lemma 4.2.1 shows that the potential $t \mapsto \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_t)$ is decreasing in t and that consequently $\Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_t) \le \Phi(\boldsymbol{R}_0 - \boldsymbol{\mu}_0, \boldsymbol{\eta}_0) = -D_{\boldsymbol{\eta}_0}(\boldsymbol{w}_1, \boldsymbol{u})$. The maximal nature of the definition of Φ implies that, for any probability distribution $\boldsymbol{p} \in \mathcal{P}(K)$,

$$\langle \boldsymbol{p}, \boldsymbol{R}_t \rangle \le \langle \boldsymbol{p}, \boldsymbol{\mu}_t \rangle + D_{\boldsymbol{\eta}_t}(\boldsymbol{p}, \boldsymbol{u}) - D_{\boldsymbol{\eta}_0}(\boldsymbol{w}_1, \boldsymbol{u}). \tag{C.7}$$

The second claim contained in (4.8) follows from the special case where $\boldsymbol{p} = \boldsymbol{\delta}_k$, the probability distribution that puts all of its mass on expert k, and by bounding the last term in (C.7) by zero. The last statement contained in (4.9) is proven in Lemma C.6.3. This is all that we had set ourselves to prove. □

C.3.2. Tuning, proof of Proposition 4.2.3

Proof of Proposition 4.2.3. The main tool that is employed here to derive the regret bounds is Proposition 4.2.2. The fact that the learning rates at hand are decreasing is a consequence of Lemma C.6.4; we give more details in the following. A slightly stronger result than what we claim could be obtained by replacing directly the learning rates in Proposition 4.2.2. However, the result is not amenable to an easy interpretation, and we use upper bounds on the learning rates and their reciprocals. Recall that $\gamma_k = 8\frac{\sigma_{\max}^2}{\sigma_k^2}\ln(1/u_k)$. The learning rate is of the form $\eta_{t,k} = \eta_{0,k}H_{1,k}(v_t) = \eta_{0,k}h(v_t/\gamma_k)$ with $h(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[\sqrt{\frac{x^2}{1+x}}\right] = \frac{x+2}{2(1+x)^{3/2}}$ and $\eta_{0,k} = 1/(2\sigma_{\max})$. That this choice of learning rate is indeed nondecreasing can be proven using Lemma C.6.4. We use the following two elementary inequalities in relation to this specific choice of function h.

Lemma C.3.1. Let $x \ge 0$. The function $h(x) = \frac{x+2}{2(1+x)^{3/2}}$ satisfies

$$\int_0^x h(x')\mathrm{d}x' \le \min\left\{x, \sqrt{x}\right\} \le \max\left\{1, \sqrt{x}\right\}, \quad \text{and} \tag{C.8}$$

$$\frac{1}{h(x)} \le \begin{Bmatrix} 1+x & \text{if } x \le 1 \\ 2\sqrt{x} & \text{if } x > 1 \end{Bmatrix} \le 2\max\left\{1, \sqrt{x}\right\}, \tag{C.9}$$

where the first minimum is equalized at $x = 1$.

Using these upper bounds and the choice $u_k = \pi_k \frac{\sigma_{\min}}{\sigma_k}$ in Proposition 4.2.2 gives the claimed result. Indeed, recall that Proposition 4.2.2 implies that

$$R_{t,k} \le \sigma_k^2 \eta_{0,k} \int_0^{v_t} h(x/\gamma_k)\mathrm{d}x + \frac{\ln(1/u_k)}{\eta_{t,k}} + \sum_{j \in K} \frac{u_j}{\eta_{t,j}} + \sigma_k^2 \eta_{0,k} \max_{s \le t} \Delta v_s. \tag{C.10}$$

We now focus on bounding each term. First,

$$\int_0^{v_t} h(x/\gamma_k)\mathrm{d}x = \gamma_k \int_0^{v_t/\gamma_k} h(x')\mathrm{d}x' \le \max\left\{\gamma_k,\ \sqrt{v_t \gamma_k}\right\}.$$

Consequently,

$$\sigma_k^2 \eta_{0,k} \int_0^{v_t} h(x/\gamma_k)\mathrm{d}x \le \sigma_k \sqrt{2 v_t \ln(1/u_k)} + 4\sigma_{\max}\ln(1/u_k). \tag{C.11}$$

Next,

$$\frac{1}{\eta_k} = \frac{2\sigma_{\max}}{h(v/\gamma_k)} \le 4\sigma_{\max}\max\left\{1, \sqrt{\frac{v_t}{\gamma_k}}\right\} \le 4\sigma_{\max} + \sigma_k\sqrt{\frac{2v_t}{\ln(1/u_k)}}.$$

With this at hand, the second and third term on the right hand side of (C.10) can be bounded by

$$\frac{\ln(1/u_k)}{\eta_{t,k}} \le \sigma_k\sqrt{2v_t\ln(1/u_k)} + 4\sigma_{\max}\ln(1/u_k), \text{ and} \tag{C.12}$$

$$\sum_{j\in K}\frac{u_j}{\eta_j} \le c_{\boldsymbol{\sigma},\boldsymbol{\pi}}\sigma_{\min}\sqrt{2v_t} + 4\sigma_{\max} \tag{C.13}$$

with $c_{\boldsymbol{\sigma},\boldsymbol{\pi}} = \sum_{k\in K}\pi_k\left(\frac{1}{\sqrt{\ln(1/u_k)}}\right)$. Replace (C.11), (C.12), and (C.13) in the the regret bound (C.10) to obtain the result. In order to prove the second claim we follow a similar path; we use Proposition 4.2.2 as our main tool. Recall that in this case the learning rate is of the form $\eta_{t,k} = \eta_{0,k}H_{2,k}(v_t)$ with $\eta_{0,k} = 1/(2\sigma_{\max})$ and

$$H_{2,k}(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[\sqrt{\alpha_k^2\left\{\left(1+\frac{x}{\alpha_k}\right)\ln\left(1+\frac{x}{\alpha_k}\right) - \frac{x}{\alpha_k}\right\} + \frac{x^2}{2(1+x/(2\gamma_k))}}\right]$$

with $\alpha_k = 32\frac{\sigma_{\max}^2}{\sigma_k^2}$ and $\gamma_k = \alpha_k\ln(1/\pi_k)$. The fact that $k \mapsto H_{2,k}(x)$ is decreasing follows from Lemma C.6.4 after performing the change of variable $x' = x/\alpha_k$. We use the inequalities for $H_{2,k}$ that are proven in the following lemma.

Lemma C.3.2. Let $\beta_k = \ln(1/\pi_k)$. The function $H_{2,k}$ satisfies

$$\int_0^x H_{2,k}(x')\mathrm{d}x' \le \sqrt{\alpha_k x(\ln(1+x/\alpha_k)+\beta_k)}, \text{ and} \tag{C.14}$$

$$\frac{1}{H_{2,k}(x)} \le 2\sqrt{\frac{x/\alpha_k}{\ln(1+x/\alpha_k)}}\sqrt{1+\frac{\min\{\beta_k, \frac{1}{2}\frac{x}{\alpha_k}\}}{\ln(1+x/\alpha_k)}}. \tag{C.15}$$

We can now compute the analogs of (C.11), (C.12), and (C.13) to obtain that

$$\sigma_k^2\eta_{0,k}\int_0^{v_t}H_{2,k}(x)\mathrm{d}x \le 2\sigma_k\sqrt{2v_t\left(\ln\left(1+\frac{\sigma_k^2}{32\sigma_{\max}^2}v_t\right)+\ln(1/\pi_k)\right)},$$

$$\frac{\ln(1/\pi_k)}{\eta_{t,k}} \le \sigma_k\ln(1/\pi_k)\sqrt{\frac{v_t}{2\ln\left(1+\frac{\sigma_k^2}{32\sigma_{\max}^2}v_t\right)}}\left(1+\sqrt{\frac{\min\{\ln\left(\frac{1}{\pi_k}\right), \frac{\sigma_k^2}{16\sigma_{\max}^2}v_t\}}{\ln\left(1+\frac{\sigma_k^2}{32\sigma_{\max}^2}v_t\right)}}\right),$$

$$\sum_{j\in K}\frac{u_j}{\eta_j} \le \sum_{j\in K}\pi_j\left(\sigma_j\sqrt{\frac{v_t}{2\ln\left(1+\frac{\sigma_j^2}{32\sigma_{\max}^2}v_t\right)}}\left(1+\frac{\sqrt{\min\{\ln\left(\frac{1}{\pi_j}\right), \frac{\sigma_j^2}{16\sigma_{\max}^2}v_t\}}}{\sqrt{\ln\left(1+\frac{\sigma_j^2}{32\sigma_{\max}^2}v_t\right)}}\right)\right),$$

and employ them in Proposition 4.2.2 to obtain the result. □

Proof of Lemma C.3.1. The relations are clear for $x = 0$. Let $x > 0$. Recall that $\int_0^x h(x')\mathrm{d}x' = \frac{x}{\sqrt{1+x}}$. We start by proving (C.8). The fact that $\frac{x}{\sqrt{1+x}} \leq x$ is clear. The inequality $\frac{x}{\sqrt{1+x}} \leq \sqrt{x}$ follows from dividing both sides of the inequality $x \leq \sqrt{x^2+x}$ by $\sqrt{1+x}$. Thus, the first inequality in (C.8) follows, and the second is direct after observing that $x \leq \sqrt{x} \leq 1$ for $x \leq 1$. We now turn to proving (C.9). Recall that $1/h(x) = \frac{2(1+x)^{3/2}}{2+x}$. We start by showing that $1/h(x) \leq 1+x$ for all $x > 0$. Note that $\frac{2(1+x)^{3/2}}{2+x} = (1+x)\frac{2\sqrt{1+x}}{2+x}$. Thus, the claim holds if and only if $2\sqrt{1+x} \leq 2+x$, which is easily checked to be the case. Now let $x > 1$. Observe that the second claim in the first inequality holds if and only if $2(1+x)^{3/2} \leq 2\sqrt{x}(2+x)$. Square both members and rearrange to conclude that the sought relation holds if and only if $0 \leq 4x^2 + 4x - 4$, which is the case as $x > 1$. The second inequality in (C.9) is clear. □

Proof of Lemma C.3.2. The inequalities contained in (C.14) and (C.15) are a consequence of the fact that

$$\int_0^x H_2(x')\mathrm{d}x' = \sqrt{\alpha^2\left\{\left(1+\frac{x}{\alpha}\right)\ln\left(1+\frac{x}{\alpha}\right) - \frac{x}{\alpha}\right\} + \frac{x^2}{2(1+x/(2\gamma))}}$$

and the inequalities

$$(1+x')\ln(1+x') - x' \leq x'\ln(1+x') \quad \text{and} \quad \frac{a^2x'^2}{2(1+x'/(2b))} \leq \min\{bx', \tfrac{1}{2}a^2x'^2\},$$

that hold for $x', a, b \geq 0$. From this, (C.14) is immediate once we use the substitutions $x' = x/\alpha$, $a = \alpha$, and $b = \beta$. To prove (C.15), use the same substitution and estimate

$$\begin{aligned}\frac{1}{H(x')} &= 2\frac{\sqrt{(1+x')\ln(1+x') - x' + \frac{x'^2}{2(1+x'/(2b))}}}{\ln(1+x') + \frac{1}{2}\frac{2x'+x'^2/(2b)}{(1+x'/(2b))^2}}\\ &\leq 2\frac{\sqrt{x'\ln(1+x') + \min\{bx', \frac{1}{2}x'^2\}}}{\ln(1+x')}\\ &= 2\sqrt{\frac{x'}{\ln(1+x')}}\sqrt{1+\frac{\min\{b, \frac{1}{2}x'\}}{\ln(1+x')}}\\ &\leq 2\sqrt{\frac{x'}{\ln(1+x')}}\left(1+\sqrt{\frac{\min\{b, \frac{1}{2}x'\}}{\ln(1+x')}}\right).\end{aligned}$$

This is all we set ourselves to prove. □

C.4. Optimism, proof of Proposition 4.4.1

Proof of Proposition 4.4.1. In Lemma 4.2.1 we show that the potential $t \mapsto \Phi_t$ is decreasesing. The rest of the proof is identical to that of Proposition 4.2.3 after multiplying all scales by 2. The "furthermore" claim follows from a direct modification of Proposition C.7.1. □

C.5. Luckiness

This appendix contains the proofs of the luckiness results in Section 4.3.

C.5.1. Proof of Theorem 4.3.1

Proof of Theorem 4.3.1. Let $s_t = \sum_{s\le t} \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_s}(\ell_t)}{\langle \tilde{\boldsymbol{w}}_s, \boldsymbol{\sigma}^2\rangle}$. It is shown in Proposition C.7.1 that v_t can be bounded in terms of s_t. Indeed, in any case $v_t \le 4$, and because the learning rates are low enough at the start of the protocol, namely $\eta_{t,k} \le 1/(2\sigma_{\max})$, the upper bound $v_t \le 4s_t$ also holds. A verification of the regret bound obtained in Proposition 4.2.2 shows that it is increasing in v_t, and consequently the same regret bound holds once we replace v_t with the larger quantity $4s_t$, and the proof of Proposition 4.2.3 can be repeated with no problems. Consequently the regret bounds in Proposition 4.2.3 are available with $4s_t$ occupying the place of v_t. The next step that we follow is to show that $\mathbf{E}_{\mathbf{P}}[s_t] \lesssim \mathbf{E}_{\mathbf{P}}[R_{t,k^*}]$, which is done in the following lemma.

Lemma C.5.1. Under Massart's condition (see Definition 4.1.3),

$$\mathbf{E}_{\mathbf{P}}[s_t] \le k_{\mathrm{M}} \mathbf{E}_{\mathbf{P}}[R_{t,k^*}],$$

where $k_{\mathrm{M}} = c_{\mathrm{M}} \max_{i,j\in K} \sup_{v\ge 0}\left\{\frac{\eta_{0,i}H_i(v)}{\eta_{0,j}H_j(v)\sigma_j^2}\right\}$ satisfies

$$k_{\mathrm{M}} \le 2c_{\mathrm{M}} \max_{i,j\in K}\left\{\frac{1}{\sigma_i\sigma_j}\frac{\ln(1/\pi_i)+\ln(\sigma_i/\sigma_{\min})}{\ln(1/\pi_j)+\ln(\sigma_j/\sigma_{\min})}\right\}.$$

From the previous discussion, a small modification of Proposition 4.2.3 shows that this tuning guarantees a regret bound of the form

$$R_{t,k^*} \le a'\sqrt{s_t} + b \tag{C.16}$$

with $a' = 4\sigma_{k^*}\sqrt{2\ln(1/u_{k^*})} + 2\sqrt{2}c_{\boldsymbol{\sigma},\boldsymbol{\pi}}\sigma_{\min}$ and $b = 8\sigma_{\max}\ln(1/u_{k^*}) + 4\sigma_{\max} + 2\sigma_{k^*}$. Take $\mathbf{P}$-expectations in the last display, use the concavity of $x \mapsto \sqrt{x}$ to invoke Jensen's inequality, and use Lemma C.5.1 to obtain that

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \le a'\sqrt{k_{\mathrm{M}}\mathbf{E}_{\mathbf{P}}[R_{t,k^*}]} + b. \tag{C.17}$$

This implies that the expected regret satisfies $\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \lesssim 1$, that is, it is constant. Indeed, using Lemma C.5.2 yields that

$$\mathbf{E}_{\mathbf{P}}[R_{t,k^*}] \le {a'}^2 k_{\mathrm{M}} + b. \tag{C.18}$$

The upper bound for k_{M} is contained in Lemma C.5.3. Given the definition of a in the claim, this is what we set ourselves to prove. □

Proof of Lemma C.5.1. Recall that $s_t = \sum_{s\leq t}\Delta s_s = \sum_{s\leq t}\frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_s}(\ell_s)}{\langle\tilde{\boldsymbol{w}}_s,\boldsymbol{\sigma}^2\rangle}$ with the weights $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$. Define $\ell^*_s = \ell_{s,k^*}$ to be the loss of the best expert k^*, and use that the variance $\mathrm{Var}_{\tilde{\boldsymbol{w}}_s}(\ell_s)$ satisfies $\mathrm{Var}_{\tilde{\boldsymbol{w}}_s}(\ell_s) \leq \langle\tilde{\boldsymbol{w}}_s,(\ell_s-\ell^*_s)^2\rangle$ to obtain the estimate

$$\Delta s_s \leq \frac{\langle\tilde{\boldsymbol{w}}_s,(\ell_s-\ell^*_s)^2\rangle}{\langle\tilde{\boldsymbol{w}}_s,\boldsymbol{\sigma}^2\rangle}.$$

Recall that, under $\mathbf{P}$, the loss vector ℓ_s is assumed to be independent of ℓ_{s-1}. This implies that

$$\begin{aligned}\mathbf{E_P}[\Delta s_s] &\leq \sum_{k\in K}\left(\mathbf{E_P}\left[\frac{\tilde{w}_{s,k}}{\langle\tilde{\boldsymbol{w}}_s,\boldsymbol{\sigma}^2\rangle}\right]\mathbf{E_P}\left[(\ell_{s,k}-\ell^*_s)^2\right]\right)\\ &\leq c_{\mathrm{M}}\sum_{k\in K}\left(\mathbf{E_P}\left[\frac{\tilde{w}_{s,k}}{\langle\tilde{\boldsymbol{w}}_s,\boldsymbol{\sigma}^2\rangle}\right]\mathbf{E_P}\left[\ell_{s,k}-\ell^*_s\right]\right).\end{aligned}$$

Sum the last display over rounds, and use the fact that the weights $\tilde{w}_{t,k}\propto w_{t,k}\eta_{t-1,k}$ to deduce that

$$\mathbf{E_P}\left[s_t\right] \leq c_{\mathrm{M}}\left\|\max_{s\leq t}\left\{\frac{\max_{k\in K}\eta_{s-1,k}}{\min_{k\in K}\eta_{s-1,k}\sigma^2_k}\right\}\right\|_\infty \mathbf{E_P}\left[R_{t,k^*}\right],$$

where $\|\cdot\|_\infty$ is the infinity norm w.r.t. $\mathbf{P}$ (recall that $\eta_{t-1,k}$ depend on the random losses ℓ_{t-1}). Since, for any $s=1,\ldots,$ and $k\in K$, the learning rate $\eta_{s-1,k}=\eta_{0,k}H_k(v)$, we can deduce that $c_{\mathrm{M}}\left\|\max_{s\leq t}\left\{\frac{\max_{k\in K}\eta_{s-1,k}}{\min_{k\in K}\eta_{s-1,k}\sigma^2_k}\right\}\right\|_\infty \leq k_{\mathrm{M}}$, where k_{M} is as defined in the claim of the proposition. This implies what we set ourselves to prove. □

Lemma C.5.2. Let $y,a,b\geq 0$. If $y^2\leq ay+b$ then $y\leq b+\sqrt{a}$.

Proof. The quadratic polynomial y^2-ay-b has a zero at $y^*=\frac{b+\sqrt{b^2+4a}}{2}\leq b+\sqrt{a}$. Hence, if $y^2\leq ay+b$, then $y\leq y^*$, and the result follows. □

C.5.2. Proof of Theorem 4.3.2

Proof of Theorem 4.3.2. Call $\Delta_{t,k}=L_{s,k}-L_{s,k^*}$, and $d_k=\mathbf{E_P}[\Delta_{t,k}]$. Since ℓ_s and ℓ_{s-1} are independent, the expected value of the increment of the regret R_{t,k^*} is

$$\begin{aligned}\mathbf{E_P}[\Delta R_{t,k^*}] &= \sum_{k\neq k^*}\mathbf{E_P}[w_{t,k}]\mathbf{E_P}[\ell_{t,k}-\ell_{t,k^*}] &\text{(C.19)}\\ &= \sum_{k\neq k^*}\mathbf{E_P}[w_{t,k}]d_k. &\text{(C.20)}\end{aligned}$$

We seek to prove that for $k\neq k^*$, in an event $\Omega_{t,k}$ which we define next, the weight $w_{t,k}$ is small. Define, for each $k\neq k^*$ and $t\geq 1$, the event $\Omega_{t,k}$ by

$$\Omega_{t,k}=\left\{L_{t,k^*}-L_{t,k}\leq \mu_{t,k}-\mu_{t,k^*}-\frac{1}{\eta_{t,k^*}}\ln\left(1/\pi_{k^*}\right)-\frac{1}{\eta_{t,k}}\ln\left(\frac{1}{\pi_k\varepsilon_t}\right)\right\},$$

for deterministic constants $\varepsilon_t = 1/t^2$. Recall that the weights have the form $w_{t,k} = \pi_k \mathrm{e}^{-\eta_{t,k}(L_{t,k}+\mu_{t,k}+a_t^*)}$, where a_t^* is such that $\sum_k w_{t,k} = 1$. Next, we show that, in each event $\Omega_{t,k}$, for carefully chosen $\tilde{a}_t = -\frac{1}{\eta_{t,k^*}} \ln(1/\pi_{k^*}) - L_{t,k^*} - \mu_{t,k^*}$, it holds that $a_t^* \geq \tilde{a}_t$. Indeed, this follows because, by design $\pi_{k^*}\mathrm{e}^{-\eta_{t,k}(L_{t,k}+\mu_{t,k}+\tilde{a}_t)} = 1$, and consequently,

$$\sum_{k\in K} \pi_k\big(\mathrm{e}^{-\eta_{t,k}(L_t+\mu_{t,k}+\tilde{a}_{t,k})}\big) \geq 1 = \sum_{k\in K} \pi_k\big(\mathrm{e}^{-\eta_{t,k}\left(L_t+\mu_{t,k}+a^*_{t,k}\right)}\big),$$

which implies $a_t^* \geq \tilde{a}_t$. We use this in the weight $w_{t,k}$ of expert k to conclude that

$$\mathbf{E_P}[w_{t,k}\mathbf{1}\{\Omega_{t,k}\}] \leq \pi_k\big(\mathrm{e}^{-\eta_{t,k}(L_t+\mu_{t,k}+\tilde{a}_{t,k})}\big) = \pi_k\varepsilon_t,$$

hence

$$\mathbf{E_P}[w_{t,k}] \leq \pi_k\varepsilon_t + \mathbf{P}\left\{\Omega^c_{t,k}\right\}.$$

Consequently, using (C.20),

$$\begin{aligned}\mathbf{E_P}[R_{t,k^*}] &= \sum_{s\leq t}\sum_{k\neq k^*} d_k\mathbf{E_P}[w_{t,k}]d_k &\text{(C.21)}\\ &\leq \sum_{s\leq t}\sum_{k\neq k^*}\{\pi_k d_k\varepsilon_s + d_k\mathbf{P}\{\Omega_{s,k}\}\} &\text{(C.22)}\\ &\leq 2\sum_{k\in K}\pi_k d_k + \sum_{s\leq t}\sum_{k\neq k^*} d_k\mathbf{P}\{\Omega^c_{s,k}\}, &\text{(C.23)}\end{aligned}$$

where we used that $\sum_s \varepsilon_s \leq \pi^2/6 \leq 2$. We now focus on bounding the probabilities $\mathbf{P}\{\Omega^c_{s,k}\}$. We use that $\Delta\mu_{t,k} \geq 0$ to deduce that

$$\begin{aligned}\mathbf{P}\left\{\Omega^c_{t,k}\right\} &= \mathbf{P}\left\{L_{t,k^*} - L_{t,k} > \mu_{t,k} - \mu_{t,k^*} - \frac{1}{\eta_{t,k^*}}\ln(1/\pi_{k^*}) - \frac{1}{\eta_{t,k}}\ln(1/\varepsilon_t)\right\}\\ &\leq \mathbf{P}\left\{L_{t,k^*} - L_{t,k} > -\mu_{t,k^*} - \frac{1}{\eta_{t,k^*}}\ln(1/\pi_{k^*}) - \frac{1}{\eta_{t,k}}\ln(1/\varepsilon_t)\right\}.\end{aligned}$$

In order to continue, we derive an upper bound on μ_{t,k^*}, and a lower bound on $\eta_{t,k}$, and η_{t,k^*} consisting of deterministic functions of time. Recall from Lemma C.7.1 that $v_t \leq 4t$, and that Lemma C.3.2 can be used to bound μ_{t,k^*} in terms of the integral of the function $x \mapsto H_{2,k^*}(x)$ (see proof of Proposition 4.2.3) to obtain that

$$\begin{aligned}\mu_{t,k^*} &\leq \sigma^2_{k^*}\eta_{0,k^*}\int_0^{4t} H_{2,k^*}(v)\mathrm{d}v + 4\sigma^2_{k^*}\eta_{0,k^*}\\ &\leq 4\sigma_{k^*}\sqrt{2t(\ln(1+t/8)+\ln(1/\pi_*))} + 4\sigma_{k^*}\end{aligned}$$

Now fix $k \in K$, and use again that $v_t \leq 4t$ and that $x \mapsto H_{2,k}(x)$ is decreasing (see Lemma C.6.4) to deduce that $\eta_{t,k} = \eta_{0,k}H_k(v_t) \geq \eta_{0,k}H_k(4t)$. From these observations, $\mathbf{P}\left\{\Omega^c_{t,k}\right\}$ can be further bounded by

$$\mathbf{P}\left\{\Omega^c_{t,k}\right\} \leq \mathbf{P}\left\{L_{t,k^*} - L_{t,k} > -F_k(t)\right\},$$

where $F_k(t) = 4\sigma_{k^*}\sqrt{2t(\ln(1+t/8)+\ln(1/\pi_{k^*}))} + 4\sigma_{k^*} + \frac{\ln(1/\pi_{k^*})}{\eta_{0,k^*}H_{2,k^*}(4t)} + \frac{\ln(1/\varepsilon_t)}{\eta_{0,k}H_{2,k}(4t)}$ is a deterministic function of time. Recall that the gap $d_{\min}$ was defined as $d_{\min} = \min_{k\neq k^*} d_k$ and that is assumed to be strictly positive. Recall that $\Delta_{t,k} = L_{t,k} - L_{t,k^*}$ is the gap in losses between expert k and the best expert k^*. Hoeffding's inequality implies that

$$\begin{aligned}\mathbf{P}\left\{L_{t,k^*} - L_{t,k} > F_k(t)\right\} &= \mathbf{P}\left\{td_k - \Delta_{t,k} > td_k - F_k(t)\right\}\\ &\leq \exp\left(-\frac{t}{2\sigma^2_{\max}}((d_k - F_k(t)/t)_+)^2\right)\\ &= \exp\left(-\frac{td_k^2}{2\sigma^2_{\max}}((1 - F_k(t)/(d_kt))_+)^2\right),\end{aligned}$$

where $x \mapsto (x)_+ = \max\{0,x\}$. We now seek a bound on the point $t_k^\star$ at which $F_k(t_k^\star)/t_k^\star = d_k/2$. For these values $t_k^\star$, we have, using (C.23), that

$$\mathbf{E_P}[R_{t,k^*}] \leq 2\sum_{k\in K}\pi_k d_k + \sum_{k\neq k^*}(t_k^\star d_k + \sum_{s\geq t^\star} d_k\mathbf{P}\{\Omega^c_{t,k}\}). \tag{C.24}$$

We now concentrate on bounding $t_k^\star$ and the probability of the event $\Omega^c_{t,k}$ for each k. In the limit that $d_k \to 0$, the time $t_k^\star \to \infty$. A quick computation shows that, as $t \to \infty$, $H_{2,k}(t) \sim \sqrt{\frac{2\ln t}{t}}$, and, in the same limit, $4\sigma_{k^*}\sqrt{2t(\ln(1+t/8)+\ln(1/\pi_*))} + 4\sigma_{k^*} \sim 4\sigma_{k^*}\sqrt{2t\ln t}$. Hence, as $t \to \infty$, the function F_k satisfies $F_k(t) \sim (4\sigma_{k^*} + 2\sigma_{\max})\sqrt{2t\ln t}$. We now give a bound on the solution $x_k^\star$ to the equation $xd_k/2 = (4\sigma_{k^*} + 2\sigma_{\max})\sqrt{2x\ln x}$ that holds asymptotically as $d_k \to 0$. Call $c = d_k/(2\sqrt{2}(4\sigma_{k^*} + 2\sigma_{\max}))$. Our equation of interest can be rewritten as $xc^2 = \ln x$. Linearize $x\ln x$ around $x = 2/c^2$, and use its concavity to obtain that $\ln(x) \leq \ln(2/c^2) + (c^2/2)(x - 2/c^2)$. With this estimate at hand, the solution to the simpler, linear equation $xc^2 = \ln(2/c^2) + (c^2/2)(x - 2/c^2)$ is an upper bound on $x_k^\star$. From this discussion it follows that the point $t_k^\star$ of interest satisfies $t_k^\star \leq 2\frac{\ln(1/c^2)}{c^2} - \frac{2}{c^2}$. Hence, as $d_k \to 0$,

$$d_kt_k^\star \leq \frac{2(4\sigma_{k^*} + 2\sigma_{\max})^2}{d_k}\left\{\ln\left(\frac{8(4\sigma_{k^*} + 2\sigma_{\max})^2}{d_k^2}\right) - 1\right\} = O\left(\frac{\sigma^2_{\max}}{d_k}\ln\left(\frac{\sigma^2_{\max}}{d_k^2}\right)\right). \tag{C.25}$$

We deduce that, as $d_k \to 0$, for $t \geq t_k^\star$ and any $k \neq k^*$, the probability $\mathbf{P}\{\Omega^c_{t,k}\} \leq \exp\left(-\frac{t}{8\sigma^2_{\max}}d_k^2\right)$. We sum $\mathbf{P}\{\Omega^c_{t,k}\}$ over rounds to conclude that

$$\mathbf{E_P}[R_{t,k^*}] \leq 2\sum_{k\in K}\pi_k d_k + \sum_{k\neq k^*}(t_k^\star d_k + \sum_{s\geq t^\star} d_k\mathbf{P}\{\Omega^c_{t,k}\}). \tag{C.26}$$

We now concentrate on bounding $t_k^\star$ and the probability of the event $\Omega^c_{t,k}$ for each k. In the limit that $d_k \to 0$, the time $t_k \to \infty$. A quick computation shows that, as $t \to \infty$, $H_{2,k}(t) \sim \sqrt{\frac{2\ln t}{t}}$, and, in the same limit, $4\sigma_{k^*}\sqrt{2t(\ln(1+t/8)+\ln(1/\pi_*))} + 4\sigma_{k^*} \sim 4\sigma_{k^*}\sqrt{2t\ln t}$. Hence, as $t \to \infty$, the function F_k satisfies $F_k(t) \sim (4\sigma_{k^*} + 2\sigma_{\max})\sqrt{2t\ln t}$. We know give a bound on the solution $x_k^\star$ to the equation $xd_k/2 =$

$(4\sigma_{k^*} + 2\sigma_{\max})\sqrt{2x\ln x}$ that holds asymptotically as $d_k \to 0$. Call $c = d_k/(2\sqrt{2}(4\sigma_{k^*} + 2\sigma_{\max}))$. Our equation of interest can be rewritten as $xc^2 = \ln x$. Linearize $x \ln x$ around $x = 2/c^2$, and use its concavity to obtain that $\ln(x) \le \ln(2/c^2) + (c^2/2)(x - 2/c^2)$. With this estimate at hand, the solution to the simpler, linear equation $xc^2 = \ln(2/c^2) + (c^2/2)(x-2/c^2)$ is an upper bound on $x_k^\star$. From this discussion it follows that the point $t_k^\star$ of interest satisfies

$$t_k^\star \le 2\frac{\ln(1/c^2)}{c^2} - \frac{2}{c^2} = O\left(\frac{\sigma_{\max}}{d_k^2}\ln\frac{\sigma_{\max}^2}{d_k^2}\right), \tag{C.27}$$

as $d_k \to 0$. Hence, again, as $d_k \to 0$,

$$\sum_{t \ge t^\star} d_k \mathbf{P}\left\{\Omega_{t,k}^c\right\} \le \sum_{t \ge t^\star} d_k \mathrm{e}^{-td_k^2/(8\sigma_{\max}^2)} \le \frac{d_k}{1 - \mathrm{e}^{-d_k^2/(8\sigma_{\max}^2)}}. \tag{C.28}$$

We use (C.27) and (C.28) in (C.26), and the fact that $d/(1 + \mathrm{e}^{d^2/\sigma^2}) = O(\sigma^2/d)$ as $d \to 0$ to conclude the proof. □

C.5.3. In Lemma C.5.1, k_{M} is bounded

Lemma C.5.3. In Lemma C.5.1, the constant k_{M} is bounded for Tuning 1, shown in Figure 4.2. More precisely,

$$k_{\mathrm{M}} \le 2 \max_{i,j \in K}\left\{\frac{1}{\sigma_i\sigma_j}\frac{\ln(1/\pi_i) + \ln(\sigma_i/\sigma_{\min})}{\ln(1/\pi_j) + \ln(\sigma_j/\sigma_{\min})}\right\}.$$

Proof. Recall that in both tunings of the algorithm we use the starting learning rate $\eta_{0,k} = 1/(2\sigma_{\max})$, a constant over the experts. As long as this is the case, the constant of interest k_{M} can be bounded by

$$k_{\mathrm{M}} \le \max_{i,j \in K}\sup_v \frac{H_i(v)}{\sigma_j^2 H_j(v)}. \tag{C.29}$$

Recall from Figure 4.2 that $H_{1,k}(v)$ is defined as $H_{1,k}(v) = \frac{v/\gamma_k + 2}{2(1+v/\gamma_k)^{3/2}}$ with $\gamma_k = 8\frac{\sigma_{\max}^2}{\sigma_k^2}(\ln(1/\pi_k) + \ln(\sigma_k/\sigma_{\min}))$. We can estimate the ratio

$$\begin{aligned}\frac{H_i(v)}{H_j(v)} &= \frac{v/\gamma_i + 2}{(1+v/\gamma_i)^{3/2}}\frac{(1+v/\gamma_j)^{3/2}}{v/\gamma_j + 2}\\ &\le \frac{2v/\gamma_i + 2}{(1+v/\gamma_i)^{3/2}}\frac{(1+v/\gamma_j)^{3/2}}{v/\gamma_j + 1}\\ &= 2\sqrt{\frac{1+v/\gamma_j}{1+v/\gamma_i}}\\ &\le 2\max\left\{1, \sqrt{\frac{\gamma_i}{\gamma_j}}\right\}.\end{aligned}$$

Hence

$$k_{\mathrm{M}} \le 2 \max_{i,j\in K} \left\{ \frac{1}{\sigma_i \sigma_j} \frac{\ln(1/\pi_i) + \ln(\sigma_i/\sigma_{\min})}{\ln(1/\pi_j) + \ln(\sigma_j/\sigma_{\min})} \right\},$$

as it was to be shown. □

C.6. Technical Lemmas

In this appendix we gather technical results used in previous sections.

C.6.1. For showing that the potential decreases

Lemma C.6.1. For fixed $\boldsymbol{X}$, the function $\boldsymbol{\eta} \mapsto \Phi(\boldsymbol{X}, \boldsymbol{\eta})$ is increasing, that is, if $\eta_k \le \eta'_k$, then, for fixed X, it holds that $\Phi(\boldsymbol{X}, \boldsymbol{\eta}) \le \Phi(\boldsymbol{X}, \boldsymbol{\eta}')$.

Proof. It follows from the definition of Φ and the fact that, for all $x \ge 0$, the function $x \mapsto -\ln(x) - 1 + x$ is nonnegative. Indeed, for any $w \in \mathcal{P}(K)$, it holds that

$$\begin{aligned} D_{\boldsymbol{\eta}}(\boldsymbol{w}, \boldsymbol{u}) &= \sum_{k\in K} w_k \left(\frac{\ln(w_k/u_k) - (1 - u_k/w_k)}{\eta_k} \right) \\ &\ge \sum_{k\in K} w_k \left(\frac{\ln(w_k/u_k) - (1 - u_k/w_k)}{\eta'_k} \right) \\ &= D_{\boldsymbol{\eta}'}(\boldsymbol{w}, \boldsymbol{u}). \end{aligned}$$

The result follows from the definition of Φ contained in (4.4). □

Lemma C.6.2. Fix vectors $\boldsymbol{X}, \boldsymbol{m} \in \mathbb{R}^K$ and $\boldsymbol{u}, \boldsymbol{\eta} \in \mathbb{R}^K_+$. Let $\boldsymbol{w}$ be the optimum value $\boldsymbol{w} = \arg\max_{\boldsymbol{p}\in\mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X} + \boldsymbol{m} \rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u})$. Then,

$$\Phi(\boldsymbol{X} + \boldsymbol{m} - \langle \boldsymbol{w}, \boldsymbol{m} \rangle, \boldsymbol{\eta}) \le \Phi(\boldsymbol{X}, \boldsymbol{\eta})$$

Proof. The result follows from the chain of inequalities

$$\begin{aligned} \Phi(\boldsymbol{X} + \boldsymbol{m} - \langle \boldsymbol{w}, \boldsymbol{m} \rangle, \boldsymbol{u}) &= \langle \boldsymbol{w}, \boldsymbol{X} + \boldsymbol{m} - \langle \boldsymbol{w}, \boldsymbol{m} \rangle \rangle - D_{\boldsymbol{\eta}}(\boldsymbol{w}, \boldsymbol{u}) \\ &= \langle \boldsymbol{w}, \boldsymbol{X} \rangle - D_{\boldsymbol{\eta}}(\boldsymbol{w}, \boldsymbol{u}) \\ &\le \Phi(\boldsymbol{X}, \boldsymbol{\eta}). \end{aligned}$$

□

C.6.2. For bounding $\boldsymbol{\mu}$ with v

The following is the consequence of a standard result in the theory of Riemann integration.

Lemma C.6.3. Let $x \mapsto H(x)$ be a decreasing, positive, real, and continuous function such that $H(x) < \infty$ on $0 \le x < \infty$. If $\Delta v_s \ge 0$ for $s = 1, 2, \dots, t$ then

$$\sum_{s \le t} H(v_{s-1}) \Delta v_s \le \int_0^{v_t} H(x) \mathrm{d}x + (H(0) - H(v_t)) \max_{s \le t} \Delta v_t,$$

where $v_t = \sum_{s \le t} \Delta v_s$.

Proof. Because H is decreasing and $t \mapsto v_t = \sum_{s \le t} \Delta v_t$ is nondecreasing,

$$\int_0^{v_t} H(x) \mathrm{d}x \ge \sum_{s \le t} H(v_s) \Delta v_s.$$

Use this observation to deduce that

$$\begin{aligned} \sum_{s=1} H(v_s) \Delta v_s - \int_0^{v_t} H(x) \mathrm{d}x &\le \sum_{s \le t} (H(v_{s-1}) - H(v_s)) \Delta v_s \\ &\le (H(0) - H(v_s)) \max_{s \le t} \Delta v_s, \end{aligned}$$

which is what we set ourselves to prove. □

C.6.3. The learning rates decrease

Lemma C.6.4. The functions $f(x) = \frac{x+2}{2(1+x)^{3/2}}$ and $g(x) = \frac{\ln(1+x) + \frac{2x + x^2/a}{(1+x/a)^2}}{\sqrt{(1+x)\ln(1+x) - x + \frac{x^2}{2(1+x/a)}}}$ are decreasing in $x \ge 0$ for any fixed $a > 0$.

Proof. The function f is differentiable in $x \ge 0$, and its derivative is $f'(x) = -\frac{x+4}{(1+x)^{5/2}}$, a negative function. Thus, f is decreasing. We turn our attention to the function g. Let $h_1(x) = \ln(1+x)$, $h_2(x) = \frac{2x + x^2/a}{(1+x/a)^2}$, and let $H_1(x) = \int_0^x h_1(s) \mathrm{d}s = (1+x)\ln(1+x) - x$, and $H_2(x) = \int_0^x h_2(s) \mathrm{d}s = \frac{x^2}{2(1+x/a)}$. Then, the function g is of the form $h/(2\sqrt{H})$ with $h = h_1 + h_2$, and $H = H_1 + H_2$. Since $g(x)$ is differentiable in $x \ge 0$, it is enough to prove that $g' \le 0$. We compute the derivative $g' = \frac{h'(x)\sqrt{H(x)} - h^2(x)/(2\sqrt{H(x)})}{H(x)}$ and conclude that $g' \le 0$ if and only if

$$h'(x) H(x) \le \frac{1}{2} h^2(x). \tag{C.30}$$

Since $h_1/\sqrt{H_1} = \frac{\sqrt{2}}{2} f(x/a)$, the analog of the last display holds for the pair h_1, H_1. We will show that the same holds true for the pair h_2, H_2 at the end of the proof. For now, use that (C.30) holds for both pairs, replace the definition of h and H, and conclude that it is enough to show that

$$h_1' H_2 + h_2' H_1 \le h_1 h_2.$$

We now focus on showing that $\delta^\star = h_1 h_2 - h_1' H_2 - h_2' H_1$ is nonnegative. Define $\delta(x) = (1+x/a)^3(x+1)2a^3\delta^\star(x)$. It is clear that it is sufficient to our purposes to show that $\delta(x) \geq 0$ for $x \geq 0$. Computation shows that

$$\delta(x) = a^3x^2 - 2\,a^2x^3 - ax^4 + 2\,a^3x + \\ \left((4\,a+1)x^4 + x^5 - 2\,a^3x + 5\,a^2x^2 + \left(5\,a^2 + 4\,a\right)x^3 - 2\,a^3\right)\ln(x+1)\,.$$

Since $\delta(0) = 0$, it is enough to show that its derivative is positive; that $\delta'(x) \geq 0$ for $x \geq 0$. Computation shows that

$$\delta'(x) = 2\,a^3x - a^2x^2 + x^4 + \\ \left(4\,(4\,a+1)x^3 + 5\,x^4 - 2\,a^3 + 10\,a^2x + 3\left(5\,a^2 + 4\,a\right)x^2\right)\ln(x+1)\,.$$

We now pay attention to the first three summands of the previous display. We use that $2a^3x - a^2x^2 + x^4 = x(2a^3 - a^2x + x^3) \geq \ln(1+x)(2a^3 - a^2x + x^3)$, which follows from the fact that last factor of the last equation is a depressed cubic that is nonnegative for $x, a \geq 0$. This fact, the previous display, and a short computation together imply that

$$\frac{\delta'(x)}{\ln(1+x)} \geq (16\,a+5)x^3 + 5\,x^4 + 9\,a^2x + 3\left(5\,a^2 + 4\,a\right)x^2,$$

which shows that $\delta'(x) \geq 0$ for $x \geq 0$. This in turn shows that the function δ is positive, that consequently the relation (C.30) holds, and finally, that the original function of interest g is decreasing. □

C.6.4. For bounding Δv in terms of Δs

Lemma C.6.5. Let $y, x, b \in \mathbb{R}$ be such that $b \geq 0$, $x \leq b$, and $y > 0$. Let $\varphi = \frac{\mathrm{e}^b - 1 - b}{\frac{1}{2}b^2} \geq 1$. Then the following statements hold.

1. For $g(y) = \frac{\varphi - 1 - \sqrt{(\varphi-1)^2 + 2\varphi y}}{\varphi} - \ln\left(\varphi - \sqrt{(\varphi-1)^2 + 2\varphi y}\right)$, we have

$$\mathrm{e}^{x - g(y)} - 1 - x \leq \frac{1}{2}\varphi x^2 - y$$

   any time that $y \leq \frac{2\varphi - 1}{\varphi}$.

2. Let $c = \varphi/(\varphi - 1)$. For any $0 < s < 1/c$ it holds that

$$\mathrm{e}^{x - s - h(cs)} - 1 - x \leq \frac{1}{2}\varphi x^2 - s,$$

   where

$$h(u) = -u - \ln(1-u) \leq \frac{1}{2}\frac{u^2}{1-u}$$

   for $0 < u < 1$.

Proof. Proving our claim is equivalent to proving that

$$g(z) \geq x - \ln\left(1 - z + x + \frac{1}{2}\varphi x^2\right).$$

The condition that $z < \frac{2\varphi-1}{2\varphi}$ ensures that the logarithm is well defined. The first claim follows because g was chosen as the maximizer over $x \leq b$ of the right hand side of the previous display. Indeed, the maximizer is $-x^\star(z)$ with $x^\star(z) = -\frac{\varphi-1-\sqrt{(\varphi-1)^2+2\varphi z}}{\varphi} \geq 0$. Now we turn to proving the second claim, which will follow from a series of rewritings of the first claim. The previous display can be rewritten as

$$g(z) = -x^\star(z) - \ln(1 - \varphi x^\star(z)).$$

Let $s' = x^\star(z)$ so that $z = \frac{1}{2}\varphi {s'}^2 + (\varphi - 1)s'$. If we let $h(u) = -u - \ln(1-u)$, the previous display can be rewritten as

$$g(z) = (\varphi - 1)s' + h(\varphi s').$$

In these terms, the first claim that we already proved takes the shape

$$\mathrm{e}^{x-(\varphi-1)s'-h(\varphi s')} - 1 - x \leq \frac{1}{2}\varphi x^2 - (\varphi - 1)s' - \frac{1}{2}\varphi {s'}^2$$

any time that $s' \leq 1/\varphi$. Define $s = (\varphi - 1)s'$. Replace this in the last display and bound the last, negative term by 0 to obtain that, as long as $s \leq \frac{\varphi-1}{\varphi}$,

$$\mathrm{e}^{x-s-h(cs)} - 1 - x \leq \frac{1}{2}\varphi x^2 - s.$$

This is our claim. The additional bound on h is well known and can be proven with a term-wise bound on the Taylor expansion of $u \mapsto -u - \ln(1-u)$. □

C.6.5. Dual formulation of $\Delta\Phi$

Recall from the definitions in Section 4.2 that the Bregman divergence $D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u})$ between $\boldsymbol{p}$ and $\boldsymbol{u}$, two vectors in $\mathbb{R}_+^K$, was defined in (4.3) as

$$D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}) = \sum_{k \in K} p_k \left(\frac{\ln(p_k/u_k) - (p_k - u_k)}{\eta_k}\right);$$

and the corresponding potential Φ, in (4.4) as

$$\Phi(\boldsymbol{X}, \boldsymbol{\eta}) = \sup_{\boldsymbol{p} \in \mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X} \rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}).$$

In the implementation of the algorithm, we rely on the dual formulation of the potential Φ and its change $\Delta\Phi$ between rounds. We compute these in the following two lemmas.

Lemma C.6.6 (Potential difference in dual form). Let $\boldsymbol{X}, \Delta\boldsymbol{X} \in \mathbb{R}^K$ and $\boldsymbol{u}, \boldsymbol{\eta} \in \mathbb{R}_+^K$, $\Delta\Phi = \Phi(\boldsymbol{X} + \Delta\boldsymbol{X}, \boldsymbol{\eta}) - \Phi(\boldsymbol{X}, \boldsymbol{\eta})$, and $\boldsymbol{w} = \arg\max_{\boldsymbol{p}\in\mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X}\rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u})$. Then

$$\Delta\Phi = \inf_{\Delta a\in\mathbb{R}} \sum_{k\in K} w_k \left(\frac{\mathrm{e}^{\eta_k(\Delta X_k - \Delta a)} + \eta_k \Delta a - 1}{\eta_k} \right).$$

Proof. From Lemma C.6.7 we know that

$$w_k = u_k \mathrm{e}^{\eta_k(X_k - a^*)},$$

where a^* is such that $\sum_{k\in K} w_k = 1$, and that

$$\Phi(\boldsymbol{X}, \boldsymbol{\eta}) = a^* + \sum_{k\in K} u_k \left(\frac{\mathrm{e}^{\eta_k(X_k - a^*)} - 1}{\eta_k} \right).$$

Use the same lemma and the change of variable $a = a^* + \Delta a$ to obtain that

$$\Phi(\boldsymbol{X} + \Delta\boldsymbol{X}, \boldsymbol{\eta}) = \inf_{\Delta a\in\mathbb{R}} \left\{ a^* + \Delta a + \sum_{k\in K} u_k \left(\frac{\mathrm{e}^{\eta_k(X_k - a^* + \Delta X_k - \Delta a)} - 1}{\eta_k} \right) \right\}.$$

Substract these two displays and use the explicit expresion for w. In this way, we obtain the result. □

Lemma C.6.7 (Potential Dual). Let $\boldsymbol{X} \in \mathbb{R}^K$ be a vector, and let $\boldsymbol{u}, \boldsymbol{\eta} \in \mathbb{R}_+^K$ be positive vectors. Then

1. The potential Φ satisfies

$$\Phi(\boldsymbol{X}, \boldsymbol{\eta}) = \langle \boldsymbol{p}^*, \boldsymbol{X}\rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}^*, \boldsymbol{u}),$$

where $p_k^* = u_k \mathrm{e}^{\eta_k(X_k - a^*)}$, and a^* is such that $\sum_{k\in K} p_k^* = 1$.

2. The potential Φ satisfies the identity

$$\Phi(\boldsymbol{X}, \boldsymbol{\eta}) = \inf_{a\in\mathbb{R}} \left\{ a + \sum_{k\in K} u_k \left(\frac{\mathrm{e}^{\eta(X_k - a)} - 1}{\eta_k} \right) \right\}.$$

Proof. Consider the optimization problem

$$\sup_{p\in\mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X}\rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}).$$

Its Lagrangian function is

$$\mathcal{L}(a, \boldsymbol{p}) = \langle \boldsymbol{p}, \boldsymbol{X}\rangle - D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}) - a\left(\sum_{k\in K} p_k - 1 \right).$$

The strong duality relation

$$\sup_{\boldsymbol{p}\in\mathcal{P}(K)} \langle \boldsymbol{p}, \boldsymbol{X}\rangle + D_{\boldsymbol{\eta}}(\boldsymbol{p}, \boldsymbol{u}) = \inf_{a\in\mathbb{R}} \sup_{\boldsymbol{p}\in\mathbb{R}^K} \mathcal{L}(a, \boldsymbol{p}) \tag{C.31}$$

holds, and the maximum on the right hand side can be computed by diferentiation. The gradient with respect to $\boldsymbol{p}$ is

$$\nabla_{\boldsymbol{p}}\mathcal{L}_k = X_k - a - \frac{\ln(p_k/u_k)}{\eta_k},$$

which is zero at

$$p_k^* = u_k \mathrm{e}^{\eta_k(X_k-a)}.$$

Replace $\boldsymbol{p}^*$ in the Lagrangian $\mathcal{L}$ to conclude that

$$\mathcal{L}(a,\boldsymbol{p}^*) = a + \sum_{k\in K} u_k \left(\frac{\mathrm{e}^{\eta_k(X_k-a)} - 1}{\eta_k}\right).$$

Replace this in (C.31) to obtain the second claim. For the first claim, differentiate $\inf_{a\in\mathbb{R}} \mathcal{L}(a,p^*)$ with respect to a and equate to 0. □

C.7. Proof of Theorem 4.1.2

Recall that Δv_t is implicitly specified in the definition of MUSCADA, in Figure 4.1. The main intuition driving the result contained in Theorem 4.1.2 stems from a Taylor approximation of the increment of the potential function at round t for small learning rates. The duality computation for the potential increment $\Delta\Phi$ of Lemma C.6.6 implies that, at round t, Δv_t is the value of Δv that satisfies

$$\inf_{\lambda\in\mathbb{R}} \sum_{k\in K} w_{t,k} \left(\frac{\mathrm{e}^{-\eta_{t-1,k}(\ell_{t,k}-\lambda)-\eta_{t-1,k}^2\sigma_k^2\Delta v} + \eta_{t-1,k}(\ell_{t,k}-\lambda) - 1}{\eta_{t-1,k}}\right) = 0, \tag{C.32}$$

where, in the notation of Lemma C.6.6, we used $\Delta\boldsymbol{X} = \Delta\boldsymbol{R}_t$ and reparametrized by $\lambda = \langle \boldsymbol{w}_t, \ell_t\rangle - \Delta a$. For small values of η, the Taylor approximation $\mathrm{e}^{\eta x - \eta^2 b} = 1 + \eta x + \frac{1}{2}\eta^2(x^2 - 2b) + O(\eta^3)$ gives that, if all the learning rates are small, the quantity being minimized in the previous display can be approximated as

$$\begin{aligned}\sum_{k\in K} w_{t,k} \left(\frac{\mathrm{e}^{-\eta_{t-1,k}(\ell_{t,k}-\lambda)-\eta_{t-1,k}^2\sigma_k^2\Delta v} + \eta_{t-1,k}(\ell_{t,k}-\lambda) - 1}{\eta_{t-1,k}}\right) \approx \\ \frac{1}{2}\sum_{k\in K} w_{t,k}\eta_{t-1,k}(\ell_{t,k}-\lambda)^2 - \Delta v \sum_{k\in K} w_{t,k}\eta_{t-1,k}\sigma_k^2.\end{aligned} \tag{C.33}$$

If this approximate expression could be plugged into (C.32), we could solve the infimum and obtain that

$$\Delta v_t \approx \frac{1}{2}\frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}}(\ell_t)}{\langle \tilde{\boldsymbol{w}}, \boldsymbol{\sigma}^2\rangle}$$

with $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$. However, this approximation is only valid under range restrictions in the values of λ. This is the subject of Lemma C.7.2, whose main technical ingredient is the inequality obtained in Lemma C.6.5, which contains an estimate that makes (C.33) precise. We gather theses results in the following proposition. Used with $b = 1$, it implies Theorem 4.1.2 because the learning rates from Figure 4.2 are all smaller than $1/(2\sigma_{\max})$.

Proposition C.7.1. *Fix $t \geq 1$. Let $\tilde{w}_{t,k} \propto w_{t,k}\eta_{t-1,k}$, where $\boldsymbol{w}_t$ are the weights played by* MUSCADA *at round t, and $\boldsymbol{\eta}_{t-1}$ its learning rates. The following statements hold.*

1. *If $\max_k 2\eta_{t-1,k}\sigma_k \leq b$ and $b \leq 1$, then*

$$\Delta v_t \leq c_0 \frac{\langle \tilde{\boldsymbol{w}}_t, \ell_t^2 \rangle}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \leq c_0, \tag{C.34}$$

where the constant c_0 satisfies $c_0 \leq 3.1$ and depends only on b.

2. *If $\max_k 2\eta_{t-1,k}\sigma_{\max} \leq b$ for some $b \leq 1$, and*

$$\Delta s_t = \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_t}(\ell_t)}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle},$$

then

$$\Delta v_t \leq c_1 \Delta s_t + c_2 \Delta s_t^2, \tag{C.35}$$

and consequently

$$v_t \leq c_3 s_t,$$

where $c_1 \leq 0.72$, $c_2 \leq 2.4$, and $c_3 = c_1 + c_2 \leq 3.1$ depend on b only.

Proof of Proposition C.7.1. First, we prove 1. Assume that $\max_k 2\eta_{t-1,k}\sigma_k \leq b'$ and that $b' \leq 1$. Our objective is to use Lemma C.7.2 with $\lambda = 0$. To this end, let $\varphi' = \frac{e^{b'} - b' - 1}{\frac{1}{2}b'^2} \geq 1$, $c_1' = \frac{b'^2\varphi'^2}{8(\varphi'-1)}$, and $c_2' = \frac{\varphi'^4 b'^2}{8(\varphi'-1)^2} - \frac{\varphi'^3 b'^2}{8(\varphi'-1)}$ be as in Lemma C.7.2. Since we assumed that $b \leq 1$, we have that $c_1' \leq 1/2$, and we can conclude that

$$\Delta v_t \leq \frac{\varphi'}{2}\Delta s_{t,0} + \frac{1}{2}\frac{c_2' \Delta s_{t,0}^2}{1 - c_1' \Delta s_{t,0}}$$

with $\Delta s_{t,0} = \frac{\langle \tilde{\boldsymbol{w}}_t, \ell_t^2 \rangle}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \leq 1$. Use this to conclude that

$$\Delta v_t \leq \frac{\varphi'}{2} + \frac{1}{2}\frac{c_2'}{1 - c_1'}.$$

This last display is exactly our first claim once we set $c_0 = \frac{\varphi'}{2} + \frac{1}{2}\frac{c_2'}{1-c_1'}$. The value of c_0' depends monotonically on that of b'. Compute the value of c_0' for $b' = 1$ to confirm that $c_0' \leq 3.1$.

We now turn our attention to the second claim. We proceed in a similar fashion as before. Assume that $\max_k 2\eta_{t-1,k}\sigma_{\max} \leq b$ for some $b \geq 1$. Let φ, c_1, c_2 be defined as before but now in terms of b. Use Lemma C.7.2 to obtain that

$$\Delta v_t \leq \frac{\varphi}{2}\Delta s_t + \frac{1}{2}\frac{c_2 \Delta s_t^2}{1 - c_1 \Delta s_t}$$

with $\Delta s_t = \frac{\mathrm{Var}_{\tilde{\boldsymbol{w}}_t}(\ell_t)}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \le 1$. Use this to conclude that

$$\Delta v_t \le \frac{\varphi}{2} \Delta s_t + \frac{1}{2} \frac{c_2}{1 - c_1} \Delta s_t^2.$$

This is exactly the second claim up to a redefinition of constants. The "consequenlty" part of the claim follows from the observation that $\Delta s_t^2 \le \Delta s_t$ and a summation over time. The computation of the upper bound on the constants is similar as before. □

Lemma C.7.2. Let $t \ge 1$, $\lambda \in \mathbb{R}$, and let

$$\Delta s_t = \Delta s_t(\lambda) = \frac{\langle \tilde{\boldsymbol{w}}_t, (\ell_t - \lambda)^2 \rangle}{\langle \tilde{\boldsymbol{w}}_t, \boldsymbol{\sigma}^2 \rangle} \tag{C.36}$$

with $\tilde{w}_{t,k} \propto w_{t,k} \eta_{t-1,k}$. Then, if $\max_k \eta_{t-1,k}(\ell_{k,t} - \lambda) \le b$ and $\max_k (2\eta_{t-1,k}\sigma_k) \le b$ for some $b \ge 0$, we have that

$$\Delta v_t \le \frac{\varphi}{2} \Delta s_t + c_1 \Delta v_t \Delta s_t + \frac{1}{2} c_2 \Delta s_t^2, \tag{C.37}$$

where $\varphi = \frac{\mathrm{e}^b - b - 1}{\frac{1}{2} b^2} \ge 1$, $c_1 = \frac{b^2 \varphi^2}{8(\varphi - 1)}$, and $c_2 = \frac{\varphi^4 b^2}{8(\varphi-1)^2} - \frac{\varphi^3 b^2}{8(\varphi - 1)}$. If additionally $c_1 \Delta v_t < 1$, then

$$\Delta v_t \le \frac{\varphi}{2} \Delta s_t + \frac{1}{2} \frac{c_2 \Delta s^2}{1 - c_1 \Delta s_t}. \tag{C.38}$$

Proof. Let $t \ge 1$. First note that if $c_1 \Delta s_t \ge 1$, our claim becomes trivial. We can safely assume that that $c_1 \Delta s_t < 1$. We proceed in the following steps. Use Lemma C.6.6 to express the increase in the potential function $\Delta \Phi_t(\Delta v) = \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_{t-1} - \boldsymbol{\eta} \boldsymbol{\sigma}^2 \Delta v, \boldsymbol{\eta}_{t-1}) - \Phi(\boldsymbol{R}_t - \boldsymbol{\mu}_t, \boldsymbol{\eta}_{t-1})$ in dual form as

$$\Delta \Phi_t(\Delta v) = \inf_{\lambda \in \mathbb{R}} \sum_{k \in K} w_{t,k} \left(\frac{\mathrm{e}^{-\eta_{t-1,k}(\ell_{t,k} - \lambda) - \eta_{t-1,k}^2 \sigma_k^2 \Delta v} + \eta_{t-1,k}(\ell_{t,k} - \lambda) - 1}{\eta_{t-1,k}} \right).$$

From now and until the end of the proof, omit the time indexes for readability.

Because of our assumption that $\eta_k |\ell_k - \lambda| \le b$, Lemma C.6.5 can be used to obtain that

$$\Delta \Phi(\Delta v) \le \frac{1}{2} \varphi \sum_{k \in K} w_k [\eta_k (\ell_k - \lambda)^2] - \sum_{k \in K} w_k \left(\frac{g^{-1}(\eta_k^2 \sigma_k^2 \Delta v)}{\eta_k} \right)$$

where $g(x) = x + h(cx)$, $h(u) = \frac{1}{2} \frac{u^2}{1-u}$ and $c = \varphi / (\varphi - 1)$. Use the concavity of $x \mapsto g^{-1}(\Delta v x)/x$ and Jensen's inequality to deduce that

$$\begin{aligned} \sum_{k \in K} w_k \left(\frac{g^{-1}(\eta_k^2 \sigma_k^2 \Delta v)}{\eta_k} \right) &= \sum_{k \in K} w_k \left(\eta_k \sigma_k^2 \frac{g^{-1}(\eta_k^2 \sigma_k^2 \Delta v)}{\eta_k^2 \sigma_k^2} \right) \\ &\ge \langle \boldsymbol{w}, \boldsymbol{\eta} \boldsymbol{\sigma}^2 \rangle \frac{g^{-1}\left(\Delta v \langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle \right)}{\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle}, \end{aligned}$$

where we defined $\hat{w}_k \propto w_k \eta_k \sigma_k^2$. This is useful for obtaining the bound

$$\Delta\Phi(\Delta v) \leq \frac{1}{2}\varphi \sum_{k \in K} w_k \left(\eta_k (\ell_k - \lambda)^2\right) - \langle \boldsymbol{w}, \boldsymbol{\eta}\boldsymbol{\sigma}^2 \rangle \frac{g^{-1}\left(\Delta v \langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle\right)}{\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle}.$$

Consequenlty, $\Delta\Phi(\Delta v^\star) \leq 0$ for

$$\Delta v^\star = \frac{1}{\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle} g\left(\frac{1}{2}\varphi \langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle \frac{\langle \tilde{\boldsymbol{w}}, (\ell - \lambda)^2 \rangle}{\langle \tilde{\boldsymbol{w}}, \boldsymbol{\sigma}^2 \rangle}\right),$$

where $\tilde{w}_k \propto w_k \eta_k$. Use the definition of Δv and the continuity of $\Delta\Phi$ to conclude that $\Delta v \leq \Delta v^\star$. Unpack the definition of g to obtain that

$$\Delta v \leq \frac{1}{2}\varphi \Delta s + \frac{1}{2} \frac{\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle (c' \Delta s)^2}{1 - c' \langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle \Delta s}$$

with $c' = \frac{1}{2}\frac{\varphi^2}{\varphi - 1}$. Next, we will use use that $\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle \leq \frac{1}{4} b^2$ to bound further Δv. Use this observation and the definition of c_1 to deduce the inequality $\langle \hat{\boldsymbol{w}}, \boldsymbol{\eta}^2 \boldsymbol{\sigma}^2 \rangle c' \Delta s \leq c_1 \Delta s < 1$. Plug this in the previous display and rearrange to obtain the result:

$$\Delta v \leq \frac{1}{2}\varphi \Delta s + \frac{1}{2}\Delta s^2 \left(\frac{1}{4} c'^2 b^2 - \frac{1}{4}\varphi c' b^2\right) + \frac{1}{4} c' b^2 \Delta v \Delta s,$$

exactly what we claimed. □

D. Appendix to Chapter 5

D.1. Proofs for Section 5.2

of Proposition 5.2.5. Since $\exp(\eta Z) \le \exp(\eta Z_+)$, we have for each $0 < \eta < b$, using also Fubini's theorem and the tail condition on Z

$$\mathbf{E}[\mathrm{e}^{\eta Z} - 1] \le \mathbf{E}[\mathrm{e}^{\eta Z_+} - 1] = \mathbf{E}[\int_0^{Z_+} \eta \mathrm{e}^{\eta Z} \mathrm{d}z] = \eta \int_0^{\infty} \mathbf{P}\{Z \ge z\} \mathrm{e}^{\eta z} \mathrm{d}z \le$$
$$a\eta \int_0^{\infty} \mathrm{e}^{-(b-\eta)z} \mathrm{d}z = \frac{a\eta}{b-\eta},$$

which means that

$$Z \trianglelefteq_\eta \frac{1}{\eta} \ln\left(1 + \frac{a\eta}{b-\eta}\right) \tag{D.1}$$

For the first claim, pick $0 \le \eta^* < b$ and call c the right hand side of (D.1) when evaluated at $\eta = \eta^*$. The result follows by item 3 in Proposition 5.2.4, ahead. The converse follows from

$$\mathbf{P}\{X \ge Y + \epsilon\} \le \mathrm{e}^{\eta(\mathbf{A}^\eta[X-Y]-\epsilon)} \le \mathrm{e}^{\eta(c-\epsilon)}$$

with $a = \mathrm{e}^{\eta c}$, and $b = \eta$. □

of Proposition 5.2.11. We can write

$$\mathbf{E}[X] = \mathbf{E}[X \mid X \ge 0]\mathbf{P}\{X \ge 0\} + \mathbf{E}[X \mid X < 0]\mathbf{P}\{X < 0\}.$$

We will bound both terms on the right hand side from bellow. For the first one, use Markov's inequality and the definition of conditional expectation to obtain that

$$\mathbf{E}[X \mid X \ge 0]\mathbf{P}\{X \ge 0\} = \mathbf{E}[[X]_+] \ge a\mathbf{P}\{X \ge a\}. \tag{D.2}$$

For the second term, use the conditional version of Jensen's inequality to obtain that

$$\begin{aligned} \mathbf{E}[X \mid X < 0] &\ge -\frac{1}{\eta} \ln \mathbf{E}[e^{-\eta X} \mid X < 0], \\ &\ge -\frac{1}{\eta} \ln \frac{1}{\mathbf{P}\{X < 0\}}. \end{aligned} \tag{D.3}$$

where the last inequality holds because by the assumption that $0 \trianglelefteq_\eta X$, which implies

$$\begin{aligned} 1 \ge \mathbf{E}[\mathrm{e}^{-\eta X}] &= \mathbf{E}[\mathrm{e}^{-\eta X} \mid X \ge 0]\mathbf{P}\{X \ge 0\} + \mathbf{E}[\mathrm{e}^{-\eta X} \mid X < 0]\mathbf{P}\{X < 0\}, \\ &\ge \mathbf{E}[\mathrm{e}^{-\eta X} \mid X < 0]\mathbf{P}\{X < 0\}. \end{aligned}$$

Gathering (D.2) and (D.3) together implies

$$\mathbf{E}[X] \geq a\mathbf{P}\{X \geq a\} - \frac{1}{\eta}\mathbf{P}\{X < 0\} \ln \frac{1}{\mathbf{P}\{X < 0\}},$$

which after rearrangement implies the first inequality. The second inequality follows from maximizing the function $x \mapsto x \ln(1/x)$, which is a concave function that attains its maximum value 1/e at $x^* = 1/\mathrm{e}$. □

D.2. Proofs for Section 5.3.1

Proposition D.2.1 below strictly strengthens Proposition 5.3.1. To see how, note that (1) ⇒ (2) in Proposition 5.3.1 is implied by 1. below, noting that in Proposition 5.3.1 we assume that $\{X_f\}_{f\in\mathcal{F}}$ is regular, which implies the condition of (1). (2) ⇒ (3) in Proposition 5.3.1 is implied by 2. below, again since in Proposition 5.3.1 we assume that $\{X_f\}_{f\in\mathcal{F}}$ is regular, together with the fact that (2) in Proposition 5.3.1 already implies that for all $f \in \mathcal{F}$, the X_f are subcentered. (3) ⇒ (4) in Proposition 5.3.1 is directly implied by 3. below and again the fact that (3) in Proposition 5.3.1 already implies that for all $f \in \mathcal{F}$, the X_f are subcentered, so that $X_f \leq X - f - \mathbf{E}[X_f]$ for all $f \in \mathcal{F}$. (4) ⇒ (1) in Proposition 5.3.1 is directly implied by 4. below. (3) ⇒ (5) is implied by 5. below and (5) ⇒ (3) is implied by 6. below.

Proposition D.2.1. *1. Let $\{X_f\}_{f\in\mathcal{F}}$ be a family of random variables such that $\inf_{f\in\mathcal{F}} \mathbf{E}[X_f] > -\infty$. Suppose there is an ESI function u such that for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$. Then there are constants $C^* > 0$ and $\eta^* > 0$ such that uniformly for all $f \in \mathcal{F}$, $X_f \leq X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C^*$ (in particular, for all $f \in \mathcal{F}$, X_f is subcentered, i.e. $\mathbf{E}[X_f] \leq 0$).*

2. *Suppose $\sup_{f\in\mathcal{F}} \mathrm{Var}(X_f) < \infty$. Suppose there is a constant $C^* > 0$ and a constant $\eta^* > 0$ such that uniformly for all $f \in \mathcal{F}$, $X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C^*$. Then there exists $c, v > 0$ such that for all $f \in \mathcal{F}$, the X_f are (c, v)-subgamma on the right, i.e. they satisfy (5.24).*

3. *Suppose there exists $c, v > 0$ such that for all $f \in \mathcal{F}$, the X_f are (c, v)-subgamma on the right. Then there is an ESI function h such that for all $f \in \mathcal{F}$, we have $X_f - \mathbf{E}[X_f] \trianglelefteq_h 0$ where h is of the form $h(\epsilon) = C\epsilon \wedge \eta^*$ for some constants $C > 0, \eta^* > 0$.*

4. *Suppose that for all $f \in \mathcal{F}$, we have $X_f \leq X_f - \mathbf{E}[X_f] \trianglelefteq_h 0$ where $h(\epsilon) = C\epsilon \wedge \eta^*$ for some $C, \eta^* > 0$. Then there is an ESI function u such that for all $f \in \mathcal{F}$, $X_f \trianglelefteq_u 0$.*

5. *Suppose there exists $c, v > 0$ such that for all $f \in \mathcal{F}$, the X_f are (c, v)-subgamma on the right. Then for all $0 < \delta \leq 1$, all $f \in \mathcal{F}$, (5.25) holds with probability at least $1 - \delta$.*

6. *Suppose there exists $c, v > 0$ such that for all $f \in \mathcal{F}$, the X_f are subcentered and, for each $f \in \mathcal{F}$, for each $0 < \delta \leq 1$, with probability at least $1 - \delta$, (5.25) holds. Then:*
   there exists $a > 0$ and a differentiable function $h : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ with $h(\epsilon) > 0$ and $h'(\epsilon) \geq 0$ for $\epsilon > 0$, such that for all $f \in \mathcal{F}$, the X_f are subcentered and $\mathbf{P}\{X \geq \epsilon\} \leq a \exp(-h(\epsilon))$.

7. *Suppose the condition above holds. Then there is $C^* > 0, \eta^* > 0$ such that uniformly for all $f \in \mathcal{F}$, $X_f \leq X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C^*$.*

Proof. Part 1. Let $C' := -\inf_{f \in \mathcal{F}} \mathbf{E}[X_f] < \infty$. Take $C'' > 0$ such that $\eta^* := u(C'') > 0$. Then $X_f \trianglelefteq_{\eta^*} C''$ and $X_f - \mathbf{E}[X_f] \trianglelefteq_{\eta^*} C'' + C' := C^*$. Also $\mathbf{E}[X_f] \leq \epsilon$ for all $\epsilon > 0$ and hence $\mathbf{E}[X_f] \leq 0$, which implies subcenteredness.

Part 2. see Theorem 5.3.2 and its proof below.

Part 3. Let $U_f = X_f - \mathbf{E}[X_f]$. The assumption of right subgammaness implies that for all $0 < \eta \leq \eta^*$ with $\eta^* = 1/(2c)$ and $C' = 2v$, we have

$$\mathbf{E}[\mathrm{e}^{\eta U_f}] \leq \exp\left(\eta^2 \cdot \frac{v}{1 - c\eta}\right) \leq \exp\left(\eta^2 \cdot \frac{v}{1 - (1/2)}\right) = \exp\left(\eta^2 \cdot C'\right).$$

Now take $h(\epsilon) = \epsilon / C'$ if $\epsilon \leq C'/2c$ and $h(\epsilon) = 1/2c$ for $\epsilon > C'/2c$. Then the above display implies that $\mathbf{E}[U_f] \trianglelefteq_h 0$, and h is seen to be equal to the h in the proposition statement.

Part 4. The premise implies that $\mathbf{E}[X_f] \leq 0$ for all $f \in \mathcal{F}$ and $X_f \trianglelefteq_h 0$, so we can (trivially) take $u = h$.

Part 5. (5.25) be rewritten as: for every $t > 0$,

$$\mathbf{P}\{X_f > \sqrt{2vt} + ct\} \leq \exp(-t), \tag{D.4}$$

which is shown to be implied by (c, v)-subgammaness on the right in [Boucheron et al., 2013, Section 2.4].

Part 6. [Boucheron et al., 2013, Section 2.4]. shows that (D.4) is equivalent to $\mathbf{P}\{X_f > t\} \leq \exp(-h(t))$ where $h(t) = (v/c^2) h_1(ct/v))$, with $h_1(u) = 1 + u - \sqrt{1 + 2u}$; this is of the required form.

Part 7. If $h(0) > 0$, then we can simply apply Proposition 5.2.5 to get the desired result; if $h(0) = 0$, apply the proposition with a set in the proposition to $2a$. □

of Theorem 5.3.2. Suppose that $U - \mathbf{E}[U] \trianglelefteq_{\eta^*} C$ holds and suppose without loss of generality that U is centered. Let $M = \mathrm{e}^{\eta^* \mathbf{A}^{\eta^*}[U]} = \mathrm{e}^{\eta^* C}$, which is finite by assumption.

We bound the moments of the right part of U.

$$\begin{aligned}\mathbf{E}[(U_f)_+^n] &= \mathbf{E}[\int_0^{(U_f)_+} nu^{n-1}\mathrm{d}u] \\ &= \mathbf{E}[\int_0^\infty \mathbf{P}\{U_f \geq u\}nu^{n-1}\mathrm{d}u] \\ &\leq \mathrm{e}^{\eta^* \mathbf{A}^{\eta^*}[U_f]}n\int_0^\infty \mathrm{e}^{-\eta^* u}u^{n-1}\mathrm{d}u \\ &= n!\frac{M}{\eta^{*n}}\end{aligned}$$

Now let $v = \mathrm{Var}(U) + \frac{2M}{\eta^{*2}}$ and $c = \frac{1}{\eta^*}$. This means that $\mathbf{E}[(U)_+^n] \leq n!\frac{v}{2}c^{n-2}$ for $n \geq 3$ and $\mathbf{E}[U^2] \leq v$ uniformly over $\mathcal{F}$. The rest of the proof follows that of Boucheron et al. [2013, Theorem 2.10]: note that $\mathrm{e}^x \leq 1 + x + \frac{1}{2}x^2$ for $x \leq 0$ so that

$$\mathrm{e}^x \leq 1 + x + \frac{1}{2}x^2 + \sum_{n\geq 3}\frac{x_+^n}{n!}$$

for all x. With this in mind, we obtain a bound on the moment generating function of U:

$$\begin{aligned}\mathbf{E}[\mathrm{e}^{\eta U}] &\leq 1 + \mathbf{E}[U] + \frac{1}{2}\eta^2\,\mathrm{Var}(U) + \sum_{n\geq 3}\frac{\eta^n\mathbf{E}[(U)_+^n]}{n!} \\ &\leq 1 + \frac{1}{2}v\eta^2\sum_{n\geq 2}(c\eta)^{n-2} \\ &= 1 + \frac{1}{2}\frac{v\eta^2}{1-c\eta}\end{aligned}$$

for $0 < c\eta < 1$. Taking logarithms on both sides, using that $\ln(1+x) \leq x$ for $x \geq 0$ and rewriting leads to

$$\mathbf{A}^\eta[U] \leq \frac{1}{2}\frac{v\eta}{1-c\eta},$$

which is exactly what we were after. □

Proof of the Claim in Example 5.3.3 Let $\eta < 1$. Using $\mathbf{P}\{-1/\eta \leq U < -1\} = \int_{-1/\eta}^{-1} p(u)du = (1-\eta^{v-1})/(v-1)$ and $\mathbf{E}[U^2 \cdot \mathbf{1}\{-1/\eta < U < 0\}] = \int_{-1/\eta}^{-1} u^2 p(u)du =$

$(1-\eta^{\nu-3})/(\nu-3)$ we find:

$$\begin{aligned}
\mathbf{E}[\exp(\eta U)] &\geq \mathbf{E}[\exp(\eta U)\cdot \mathbf{1}\{-1/\eta < U < 0\}] + \mathbf{E}[\exp(\eta U)\cdot \mathbf{1}\{U \geq 0\}] \\
&\geq \mathbf{E}[(1+\eta U + U^2\eta^2/4)\cdot \mathbf{1}\{-1/\eta < U < 0\}] + \mathbf{E}[(1+\eta U)\cdot \mathbf{1}\{U\geq 0\}] \\
&\geq \mathbf{P}\{U > -1/\eta\} + \eta\mathbf{E}[U] + (\eta^2\cdot\exp(-1)\cdot\mathbf{E}[U^2\cdot \mathbf{1}\{-1/\eta < U < 0\}] \\
&= \mathbf{P}\{-1/\eta \leq U < -1\} + (1-\mathbf{P}\{U\leq -1\}) + \\
&\qquad \eta\mathbf{E}[U] + \eta^2\cdot\exp(-1)\cdot\mathbf{E}[U^2\cdot\mathbf{1}\{-1/\eta < U < 0\}] \\
&= \frac{1-\eta^{\nu-1}}{\nu-1} + (1-\frac{1}{\nu-1}) + \eta\mathbf{E}[U] + \eta^2\cdot\exp(-1)\cdot\frac{\eta^{\nu-3}-1}{3-\nu} \\
&= 1 - \frac{1}{\nu-1}\eta^{\nu-1} + \frac{\exp(-1)}{(3-\nu)}\cdot(\eta^{\nu-1}-\eta^2).
\end{aligned}$$

Since for $5/2 < \nu < 3$, we have $\exp(-1)/(3-\nu) > 1/(\nu-1)$, we find that, as $\eta \downarrow 0$, $(\mathbf{E}[\exp(\eta U)]-1)/\eta^2 \to \infty$, showing that right subgamma-ness is violated.

D.3. Proofs for Section 5.3.2

Below we first state Theorem D.3.1 and then Lemma D.3.2. We then show how, taken together, these two results imply the result in the main text, Theorem 5.3.11, as an almost direct corollary. After that, we provide first the proof of Lemma D.3.2 and then the proof of Theorem D.3.1 itself, followed by the statement and proof of Lemma D.3.3, a slight extension of the standard second-order Taylor approximation of the moment generating function that is crucial for proving Lemma D.3.2.

Theorem D.3.1. *1. Suppose $\{-X_f : f \in \mathcal{F}\}$ satisfies the $[0,b)$-Bernstein condition for some $0 < b \leq 1$ and the conclusion* ***C1*** *of Lemma D.3.2, Part 1 holds. Then for all $\beta \in [0,b)$, all $c \geq 0$, all $0 < c^* < 1$, there exists $\eta^\circ > 0$ and $C^\circ > 0$ such that for all $f \in \mathcal{F}$, all $0 < \eta \leq \eta^*$,*

$$X_f + c\eta X_f^2 - c^*\mathbf{E}[X_f] \trianglelefteq_\eta C^\circ\eta^{1/(1-\beta)}.$$

2. Suppose there exists $\eta^\circ > 0, C^\circ > 0$ such that, for some $0 < \beta \leq 1$, for all $f \in \mathcal{F}$, $X_f \trianglelefteq_{\eta^\circ} C^\circ\eta^{1/(1-\beta)}$ and the conclusion ***C2*** *of Lemma D.3.2, Part 2 holds. Then, (a) $\{-X_f : f \in \mathcal{F}\}$ satisfies the β-Bernstein condition. If furthermore $\sup_{f\in\mathcal{F}}\mathbf{E}[-X_f] < \infty$ then (b), $\{-X_f : f \in \mathcal{F}\}$ also satisfies the $[0,\beta]$-Bernstein condition and also $\sup_{f\in\mathcal{F}}\mathbf{E}[X_f^2]$ is bounded, so that the family is regular.*

Lemma D.3.2. Let $\{X_f : f \in \mathcal{F}\}$ be an ESI family.

1. Suppose that the family is regular. Then **C1** holds, with:
   C1: for each $k \geq 0$, for all $0 < \delta \leq 1$, there exist $C^*, C^\circ > 0$ and $\eta^\circ > 0$ (that may depend on δ) such that for all $0 < \eta \leq \eta^\circ$, all $f \in \mathcal{F}$, $X_f \trianglelefteq_{\eta^\circ} C^*$ and

$$\mathbf{E}[|X_f|^{2+k}\exp(\eta X_f)] \leq C^\circ(\mathbf{E}[X_f^2])^{1-\delta} \tag{D.5}$$

2. Suppose that the witness-type condition (5.30) holds for the family $\{X_f : f \in \mathcal{F}\}$ or for the family $\{(X_f)_- : f \in \mathcal{F}\}$. Then **C2** holds, with:
C2: there exist $C > 0$ and an $\eta^\circ > 0$ such that for all $0 < \eta \le \eta^\circ$, all $f \in \mathcal{F}$,
$$\mathbf{E}[X_f^2] \le C \cdot \mathbf{E}[X_f^2 \exp(\eta X_f)]. \tag{D.6}$$

To see how the above two results together imply Theorem 5.3.11 in the main text, note that the implication (1) ⇒ (2) in that theorem is a direct consequence of the fact that **C1** in Lemma D.3.2 holds for regular ESI families for $\delta = 1 - \beta$, any $0 \le \beta < 1$, as expressed by Part 1 of that lemma, combined with Theorem D.3.1, Part 1.

Implication (2) ⇒ (3) of Theorem 5.3.11 is trivial. Implication (3) ⇒ (1) follows by the fact that **C2** in Lemma D.3.2 holds for families satisfying the witness-type condition, as expressed by Part 2 of that lemma, combined with Theorem D.3.1, Part 2.

of Lemma D.3.2. Part 1. Let $s = \sup_{f \in \mathcal{F}} \mathbf{E}[X_f^2]$ and set $X = X_f$ for arbitrary $f \in \mathcal{F}$. Since we assume the family is regular, we can use Proposition 5.3.1 to infer that there exists $\eta^* > 0$, $C^* > 0$ such that $X \trianglelefteq_\eta C^*$ for all $0 < \eta \le \eta^*$.

For all $\eta' > 0, \eta > 0$, all $0 < \gamma < 2$ there must be constants $C, C' > 0$, such that for all $p > 0, q > 0$ with $1/p + 1/q = 1$, it holds that :

$$\begin{aligned}
&\mathbf{E}[|X|^{2+k} \exp(\eta X)] \\
&= \mathbf{E}[\mathbf{1}\{X \le -1\} |X|^2 \cdot \left(|X|^k \exp(\eta X)\right)] + \\
&\qquad \mathbf{E}[\mathbf{1}\{-1 < X < 1\} |X|^{2+k} \exp(\eta X)] + \mathbf{E}[\mathbf{1}\{X \ge 1\} |X|^{2+k} \exp(\eta X)] \\
&\le C' \mathbf{E}[\mathbf{1}\{X \le -1\} X^2] + \exp(\eta) \mathbf{E}[\mathbf{1}\{-1 < X < 1\} X^2] + \\
&\qquad\qquad \mathbf{E}[\mathbf{1}\{X \ge 1\} |X|^{2-\gamma} \cdot \left(|X|^{k+\gamma} \exp(\eta X)\right)] \\
&\le (C' + \exp(\eta)) s (\mathbf{E}[X^2]/s)^{1-\delta} + C \mathbf{E}[\mathbf{1}\{X \ge 1\} X^{2-\gamma} \exp((\eta' + \eta) X)] \\
&\le (C' + \exp(\eta) s^\delta \cdot (\mathbf{E}[X^2])^{1-\delta} + \\
&\qquad C \left(\mathbf{E}[(\mathbf{1}\{X \ge 1\} X^{2-\gamma})^p]\right)^{1/p} (\mathbf{E}[\exp(q(\eta' + \eta) X)])^{1/q},
\end{aligned}$$

where we in the first inequality we used that $|X|^k \exp(\eta X)$ is bounded on $X \le -1$ and in the second we used that $|X|^{k+\gamma} \exp(\eta X)$ is bounded by a constant times $\exp((\eta' + \eta) X)$ on $X \ge 1$. Then we used Hölder's inequality and the fact that $\mathbf{E}[X^2]/s \le 1$. We now take $0 < \delta \le 1$ as in the theorem statement and bound the second term further setting $1/p = 1 - \delta$, $1/q = \delta$ and $\gamma = 2\delta$ (so that $1/p + 1/q = 1$ as required and $(2 - \gamma)p = 2$):

$$\begin{aligned}
&\left(\mathbf{E}[(\mathbf{1}\{X \ge 1\} X^{2-\gamma})^p]\right)^{1/p} (\mathbf{E}[\exp(q(\eta' + \eta) X)])^{1/q} \\
&= \left(\mathbf{E}[(\mathbf{1}\{X \ge 1\} X^{2-\gamma})^p]\right)^{1-\delta} \left(\mathbf{E}[\exp(\delta^{-1}(\eta' + \eta) X)]\right)^\delta \\
&\le \left(\mathbf{E}[\mathbf{1}\{X \ge 1\} X^2]\right)^{1-\delta} \left(\mathbf{E}[\exp(\delta^{-1}(\eta' + \eta) X)]\right)^\delta \le \left(\mathbf{E}[X^2]\right)^{1-\delta} C^*
\end{aligned}$$

where the final equation follows for the specific choice $\eta' = \eta^*/(2\delta)$ and any η with $0 < \eta \le \eta^\circ := \eta^*/(2\delta)$. Combining the two equations we find that for any such η, using that the constants C and C^* do not depend on the f with $X = X_f$, the result (D.5) follows.

Part 2. Let $X = X_f$ for arbitrary $f \in \mathcal{F}$. First assume (5.30) for the family $\{X_f : f \in \mathcal{F}\}$. We have:

$$\mathbf{E}[X^2] = \mathbf{E}[X^2 \mathbf{1}\{X \geq 0\}] + \mathbf{E}[X^2 \mathbf{1}\{X < 0; X^2 \leq C\}] + \mathbf{E}[X^2 \mathbf{1}\{X < 0; X^2 > C\}] \tag{D.7}$$

$$\leq \mathbf{E}[X^2 \exp(\eta X) \mathbf{1}\{X \geq 0\}] + \mathbf{E}[X^2 \mathbf{1}\{X < 0; X^2 \leq C\}] + \mathbf{E}[X^2 \mathbf{1}\{X^2 > C\}] \tag{D.8}$$

$$\leq \mathbf{E}[X^2 \exp(\eta X) \mathbf{1}\{X \geq 0\}] + \exp(\eta^\circ \sqrt{C}) \mathbf{E}[X^2 \mathbf{1}\{X < 0; X^2 \leq C\} e^{\eta X}] + c\mathbf{E}[X^2]$$

for some $0 < c < 1$. Moving the rightmost term to the left-hand side and dividing both sides by $1 - c$, we find that

$$\mathbf{E}[X^2] \leq \frac{1}{1-c} \left(\exp(\eta^\circ \sqrt{C}) \mathbf{E}[X^2 \exp(\eta X)] \right),$$

which is what we had to prove. In case that (5.30) holds the family $\{(X_f)_- : f \in \mathcal{F}\}$, the result follows by a minor variation of the above argument: the rightmost term in (D.7) can then be bounded by $c\mathbf{E}[X_-^2]$, so the rightmost term in (D.8) can be replaced by $c\mathbf{E}[X_-^2]$, and then the final inequality still holds. □

of Theorem D.3.1. Part 1. A short calculation using Jensen's inequality shows that for all $c, c^*, \eta > 0$, we have

$$(X_f - c^* \cdot \mathbf{E}[X_f] + \eta c X_f^2)^2 \leq 3X_f^2 + 3c^{*2}(\mathbf{E}[X_f])^2 + 3\eta^2 c^2 X_f^4. \tag{D.9}$$

Take $0 < c^* < 1$ and $\beta \in [0, b)$ as in the theorem statement. Set $\delta := 1 - \beta$ and choose η° for this δ as in Lemma D.3.2, Part 1 (which we can use because $0 < \delta \leq 1$). Without loss of generality let $\eta^\circ \leq 1$. We find for all $0 < \eta < \eta^\circ$, that for some η' with $0 < \eta' < \eta^\circ$, and for some constants $C_0, C_1, C_2, C_3, C_4, C^\circ > 0$, for $\delta' = 2\delta - \delta^2$ that

$$\begin{aligned}
&\mathbf{E}[\exp(\eta(X_f - c^* \cdot \mathbf{E}[X_f] + \eta c X_f^2)] \\
&\leq 1 + \eta \mathbf{E}[X_f - c^* \cdot \mathbf{E}[X_f] + \eta c X_f^2] + \frac{1}{2}\eta^2 \mathbf{E}[(X_f - c^* \cdot \mathbf{E}[X_f] + \eta c X_f^2)^2 e^{\eta' X_f}] \\
&\leq 1 + \eta(1 - c^*) \cdot \mathbf{E}[X_f] + \eta^2 c \mathbf{E}[X_f^2] + \\
&\qquad \frac{3}{2}\eta^2 \left(c^* 2 \mathbf{E}[X_f^2] \mathbf{E}[\mathrm{e}^{\eta' X_f}] + \mathbf{E}[X_f^2 e^{\eta' X_f}] + 3c^2 \mathbf{E}[X_f^4 \mathrm{e}^{\eta' X_f}] \right) \\
&\leq 1 + \eta(1 - c^*) \cdot \mathbf{E}[X_f] + \eta^2 c \mathbf{E}[X_f^2] + \frac{3}{2} c^{*2} \eta^2 C^* \mathbf{E}[X_f^2] + \frac{3}{2}(1 + c^2)\eta^2 C^\circ (\mathbf{E}[X_f^2])^{1-\delta} \\
&\leq 1 + \eta(1 - c^*) \cdot \mathbf{E}[X_f] + \eta^2 C_1 B \mathbf{E}[-X_f]^{1-\delta} + \eta^2 C_2 B^{1-\delta} (\mathbf{E}[-X_f])^{(1-\delta)^2} \\
&\leq 1 + \eta(1 - c^*) \cdot \mathbf{E}[X_f] + \eta^2 C_3 \mathbf{E}[-X_f]^{1-\delta'} + \eta^2 C_4 (\mathbf{E}[-X_f])^{(1-\delta')} \\
&\leq 1 + \eta \left((1 - c^*) \cdot \mathbf{E}[X_f] + \eta \frac{C_3 + C_4}{(1 - c^*)^{1-\delta'}} \left((1 - c^*) \mathbf{E}[-X_f] \right)^{1-\delta'} \right) \\
&\leq 1 + \eta \left((1 - c^*) \cdot \mathbf{E}[X_f] + C^\circ \eta^{1/\delta} + (1 - c^*) \mathbf{E}[-X_f] \right) \leq 1 + \eta \cdot C^\circ \eta^{1/\delta} \leq e^{\eta C^\circ \eta^{1/\delta}}.
\end{aligned}$$

Here the first inequality is our extended Taylor approximation from Lemma D.3.3, stated and proved further below. For the second we used (D.9). The third follows by Lemma D.3.2, Part 1, applied with $k = 0$ (for the first and second term within the rightmost brackets, and for determining C^*) and $k = 2$, for the third term within those brackets, and with constant C_0 taken to be the maximum of the two corresponding constants C° in that lemma obtained with $k = 0$ and $k = 2$. The fourth follows from the β-Bernstein condition (applied to the two final terms) for $\beta = 1 - \delta$, which holds by assumption. The fifth follows because by Cauchy-Schwarz,

$$\mathbf{E}[-X_f]^{1-\delta} = \mathbf{E}[-X_f]^{1-\delta'} \cdot \mathbf{E}[X_f]^{\delta'-\delta} \leq \mathbf{E}[-X_f]^{1-\delta'} \cdot \mathbf{E}[X_f^2]^{(\delta'-\delta)/2}$$

and the latter factor is bounded since we assume the $\{X_f\}$ to be regular. The sixth inequality above is just rearranging, and the seventh inequality is a 'linearization' step. To see how it follows, note first that, for $p, q > 0$ with $1/p + 1/q = 1$, Young's inequality, $xy \leq |x|^p/p + |y|^q/q$, implies that for $0 < \beta < 1$,

$$ab^\beta \leq \frac{1-\beta}{\beta}(\beta a)^{\frac{1}{1-\beta}} + b \tag{D.10}$$

which follows for $a, b > 0$ by taking $\beta = 1/p$, $a = x^p$, and $b = y$. We apply this with $\beta = 1 - \delta'$, $b = (1 - c^*)\mathbf{E}[-X_f]$, $a = \eta(C_3 + C_4)/(1 - c^*)^{1-\delta'}$. The final inequality then gives the desired result.

Part 2 is similar to 1 but much easier; we omit the details. □

We end the section with the statement and proof of the validity of the second order Taylor approximation of the moment generating function, as used in the proof of Lemma D.3.2.

Lemma D.3.3 ("Extended Taylor"). Suppose that $\mathbf{E}[X^2] < \infty$ and also $\mathbf{E}[\mathrm{e}^{\eta X}] < \infty$ for all $\eta \in [0, \eta_{\max}]$ and let η^* be any number with $0 < \eta^* < \eta_{\max}$. Then for all $0 < \eta < \eta^*$ we have:

$$\mathbf{E}[\mathrm{e}^{\eta X}] = 1 + \eta\mathbf{E}[X] + \frac{1}{2}\eta^2\mathbf{E}[X^2\mathrm{e}^{\eta' X}] \tag{D.11}$$

for some η' with $\eta \leq \eta' < \eta^*$.

This is just the standard Taylor approximation of the moment generating function for random variable X at $\eta = 0$. However, in this chapter we need this approximation also for the case that $\mathbf{E}[\mathrm{e}^{\eta X}] = \infty$ for all $\eta < 0$ (e.g. if X has polynomial left tail), in which the standard Taylor's theorem does not apply any more, since the standard (two-sided) derivative at $\eta = 0$ is undefined. The lemma shows that nevertheless, everything still works as one would expect.

Proof. Fix $\eta_0 \in (0, \eta^*)$. Then all derivatives of $\mathbf{E}[\exp(\eta X)]$ exist at η with $\eta_0 \leq \eta \leq \eta^*$, so that:

$$\mathbf{E}[\mathrm{e}^{\eta X}] = \mathbf{E}[\mathrm{e}^{\eta_0 X}] + (\eta - \eta_0)\mathbf{E}[X\mathrm{e}^{\eta_0 X}] + \frac{1}{2}(\eta - \eta_0)^2\mathbf{E}[X^2\mathrm{e}^{s(\eta_0,\eta)X}] \tag{D.12}$$

with s some function $s : [0, \eta^*]^2 \to [\eta_0, \eta]$. Since

$$|e^{\eta X} X| \leq |e^{\eta X_+} X| \leq |e^{\eta^* X_+} X| \leq |X_-| + |\mathbf{1}\{X \geq 0\} e^{\eta^* X_+} X| \leq |X_-| + |e^{\eta^* X} X|$$

and we know $\mathbf{E}[|e^{\eta^* X} X|] < \infty$, we can use the dominated convergence theorem to conclude that $\lim_{\eta_0 \downarrow 0} \mathbf{E}[X e^{\eta_0 X}] = \mathbf{E}[X]$. Analogously one shows that $\lim_{\eta_0 \downarrow 0} \mathbf{E}[X^2 e^{s(\eta_0, \eta) X}] = \mathbf{E}[X^2 e^{\eta' X}]$ for an $\eta' \in [\eta_0, \eta]$. The result now follows by using these two limiting results in taking the limit for $\eta_0 \downarrow 0$ in (D.12). □

D.4. Proofs for Section 5.5

of Theorem 5.5.2. Inequality (5.37) follows from Proposition 5.2.3, applied with $X = \hat{\eta} X_{\hat{\eta}}, Y = \hat{\eta} Y_{\hat{\eta}}, \eta = 1$. (5.39) follows from Proposition 5.2.5, with X, Y and η set in the same way (see the remark at the end of the proposition statement).

Now we prove (5.38). Define the random variable $W_{\hat{\eta}} := X_{\hat{\eta}} - Z_{\hat{\eta}}$, and for $(a, k) \in (0, \infty) \times \mathbb{N}$ define the event $\mathcal{E}_{a,k} := \{a \cdot (k-1) \leq \hat{\eta} W_{\hat{\eta}} \leq ak\}$. With $Z = \hat{\eta} W_{\hat{\eta}}$, we will first show that

$$\limsup_{a \downarrow 0} \sum_{k=1}^{\infty} ak \cdot \mathbf{P}\{\mathcal{E}_{a,k}\} \leq \int_0^{\infty} \mathbf{P}\{Z \geq z\} \mathrm{d}z. \tag{D.13}$$

For $a, b > 0$ and $k_{a,b} := \lfloor b/a \rfloor$, we have

$$\begin{aligned}
\sum_{k=1}^{k_{a,b}} ak \cdot \mathbf{P}\{\mathcal{E}_{a,k}\} &= \sum_{k=1}^{k_{a,b}} ak \cdot (\mathbf{P}\{Z > a \cdot (k-1)\} - \mathbf{P}\{Z \geq ak\}), \\
&\leq \sum_{k=1}^{k_{a,b}} ak \cdot (\mathbf{P}\{Z \geq a \cdot (k-1)\} - \mathbf{P}\{Z \geq ak\}), \\
&\leq \sum_{k=0}^{k_{a,b}-1} a \cdot \mathbf{P}\{Z \geq ak\}, \\
&\leq a\mathbf{P}\{Z \geq 0\} + \int_0^{k_{a,b}} a\mathbf{P}\{Z \geq at\} \mathrm{d}t, \quad (t \mapsto a \cdot \mathbf{P}\{Z \geq at\} \text{ nonincr.}) \\
&= a\mathbf{P}\{Z \geq 0\} + \int_0^{ak_{a,b}} \mathbf{P}\{Z \geq z\} \mathrm{d}z, \quad (\text{change of variable } y = at) \\
&\leq a + \int_0^{b} \mathbf{P}\{Z \geq z\} \mathrm{d}z.
\end{aligned} \tag{D.14}$$

Since (D.14) holds for all $a, b > 0$, we have

$$\limsup_{a \downarrow 0} \sum_{k=1}^{\infty} ak \cdot \mathbf{P}\{\mathcal{E}_{a,k}\} = \limsup_{a \downarrow 0} \left\{ \sup_{b > 0} \sum_{k=1}^{k_{a,b}} ak \cdot \mathbf{P}\{\mathcal{E}_{a,k}\} \right\} \overset{(D.14)}{\leq} \int_0^{\infty} \mathbf{P}\{Z \geq z\} \mathrm{d}z, \tag{D.15}$$

and thus, the desired inequality (D.13) follows. We now have, for all $a > 0$,

$$\begin{aligned}\mathbf{E}\left[W_{\hat{\eta}}\right] &\le \sum_{k=1}^{\infty} \mathbf{P}\left\{\mathcal{E}_{a,k}\right\} \cdot \mathbf{E}\left[W_{\hat{\eta}} \middle| \mathcal{E}_{a,k}\right], \\ &\le \sum_{k=1}^{\infty} \mathbf{P}\left\{\mathcal{E}_{a,k}\right\} \cdot \mathbf{E}\left[\frac{ak}{\hat{\eta}}\right] \quad \text{(by the definition of } \mathcal{E}_{a,k}\text{)} \\ &= \left(\sum_{k=1}^{\infty} \mathbf{P}\left\{\mathcal{E}_{a,k}\right\} \cdot ak\right) \cdot \mathbf{E}\left[\frac{1}{\hat{\eta}}\right].\end{aligned}$$

Since this inequality holds for all $a > 0$, using (D.13) implies that

$$\begin{aligned}\mathbf{E}\left[W_{\hat{\eta}}\right] &\le \mathbf{E}\left[\frac{1}{\hat{\eta}}\right] \cdot \int_0^{\infty} \mathbf{P}\{\hat{\eta} W_{\hat{\eta}} \ge t\} \mathrm{d}t \\ &\le \mathbf{E}\left[\frac{1}{\hat{\eta}}\right] \cdot \mathbf{E}\left[\mathrm{e}^{\hat{\eta} W_{\hat{\eta}}}\right] \int_0^{\infty} \mathrm{e}^{-t} \mathrm{d}t, \text{(Markov's Inequality)} \\ &\le \mathbf{E}\left[\frac{1}{\hat{\eta}}\right], \end{aligned} \tag{D.16}$$

where the last step follows from the fact that $W_{\hat{\eta}} \trianglelefteq_{\hat{\eta}} 0$. □

of Proposition 5.5.4. Let's denote $X_{i,\eta}(Z; z^{n \smallsetminus i}) \coloneqq X_{i,\eta}(z_1, \dots, z_{i-1}, Z, z_{i+1}, \dots, z_n)$, for all $\eta \in \mathcal{G}$, $i \in [n]$, $z^{n \smallsetminus i} \in \mathcal{Z}^{n-1}$, and $Z \in \mathcal{Z}$. In this way, (5.40) can be written as

$$X_{i,\eta}(Z_i; z^{n \smallsetminus i}) \trianglelefteq_{\eta} 0, \quad \text{for all } i \in [n] \text{ and } z^{n \smallsetminus i} \in \mathcal{Z}^{n-1}. \tag{D.17}$$

In particular, since this holds for all $z^{n \smallsetminus i} \in \mathcal{Z}^{n-1}$, we can also write

$$X_{i,\eta}(Z_i; Z^{n \smallsetminus i}) \trianglelefteq_{\eta} 0. \tag{D.18}$$

Now let $Z_1^n, \dots, Z_n^n \in \mathcal{Z}^n$ be n i.i.d copies of Z^n. From (D.18), we have, for each $i \in [n]$,

$$Y_{i,\eta} \coloneqq X_{i,\eta}(Z_{i,i}; Z_i^{n \smallsetminus i}) \trianglelefteq_{\eta} 0. \tag{D.19}$$

Since $(Y_{i,\eta})_{i \in [n]}$ are i.i.d, we can chain (D.19), for $i = 1, \dots, n$, using Proposition 5.2.6 to get

$$\sum_{i=1}^{n} Y_{i,\eta} \trianglelefteq_{\eta} 0. \tag{D.20}$$

By applying Proposition 5.5.5, we have, for any random $\hat{\eta}$ in $\mathcal{G}$,

$$\begin{aligned}\mathbf{E}\left[\frac{\ln|\mathcal{G}| + 1}{\hat{\eta}}\right] &\ge \mathbf{E}\left[\sum_{i=1}^{n} Y_{i,\hat{\eta}}\right], \\ &= \mathbf{E}\left[\sum_{i=1}^{n} X_{i,\hat{\eta}}(Z_{i,i}, Z_i^{n \smallsetminus i})\right], \\ &= \mathbf{E}\left[\sum_{i=1}^{n} X_{i,\hat{\eta}}(Z^n)\right], \end{aligned} \tag{D.21}$$

the last equality follows from the fact that $(Z_i^n)_{i \in [n]}$ are i.i.d copies of Z^n. □

Samenvatting

Deze dissertatie gaat, in grote lijnen, over statistische hypothesetoetsen. Dit onderwerp is belangrijk omdat wetenschappelijke vraagstukken vaak worden uitgedrukt in termen van statistische hypothesetoetsen. Bij het bestuderen van het effect van een medische behandeling vergelijken onderzoekers bijvoorbeeld de klinische resultaten van een groep mensen die de behandeling hebben ontvangen met die van een controlegroep die deze niet heeft ontvangen. Als de behandeling effectief is, worden betere klinische resultaten verwacht bij de behandelingsgroep dan bij de controlegroep. Daarom kan de wetenschappelijke hypothese over de werkzaamheid van de behandeling worden bestudeerd door middel van een statistische vergelijking van de resultaten van de twee groepen patiënten.

Deze dissertatie is gericht op flexibele methoden voor statistische monitoring. Om een idee te krijgen van het soort flexibiliteit waar we naar verwijzen, vergelijken we de methoden die het onderwerp van onze studie zijn met klassieke steekproefmethoden. De meeste klassieke methoden die worden onderwezen in inleidende statistiekvakken vereisen dat onderzoekers de experimenten van tevoren plannen. In het bijzonder vereisen klassieke methoden dat onderzoekers een vaste hoeveelheid data verzamelen voordat ze enige analyse uitvoeren. Deze vereisten zijn nodig om fouten die kunnen optreden als gevolg van toeval te beperken. In het medische voorbeeld betekent dit dat onderzoekers de klinische resultaten van een vast aantal patiënten - een vaste steekproefgrootte - moeten waarnemen voordat ze hun bevindingen analyseren. De aanname van een vaste steekproefgrootte kan echter beperkend zijn in bepaalde situaties. Onderzoekers kunnen bijvoorbeeld geïnteresseerd zijn in het analyseren van de data terwijl deze worden verzameld om het experiment eerder te stoppen indien nodig. Dit is bijvoorbeeld het geval bij menselijke overlevingstijd-experimenten. Sterker nog, vroegtijdig stoppen kan zelfs een ethische plicht zijn wanneer een behandeling levensreddend blijkt te zijn. Er zijn krachtige methoden ontworpen om deze beperking te overwinnen. De methoden die de meeste flexibiliteit bieden worden altijd-geldig genoemd ("anytime-valid"). Deze methoden stellen onderzoekers in staat om een lopend experiment voort te zetten of te stoppen, of zelfs op elk moment een nieuw experiment te starten zonder de statistische geldigheid van hun analyse te beïnvloeden. Deze flexibele, altijd-geldige methoden zijn het hoofdonderwerp van deze dissertatie.

In dit werk worden verschillende wiskundige resultaten getoond met betrekking tot de theorie van altijd-geldige toetsen en voorspellingen. Zoals vaak het geval is bij wiskundige werken, kunnen de resultaten in deze dissertatie worden toegepast in meerdere contexten. De inhoud van deze dissertatie heeft betrekking op het ontwerp van - in een specifieke zin - optimale, altijd-geldige toetsen in abstracte settings die relevante toepassingen omvatten.

Hoofdstuk 3 toont een altijd-geldige toets voor tijd-tot-gebeurtenisgegevens. Dit is nuttig bij het monitoren van experimenten waarbij de focus ligt op de verstreken tijd tot een uitkomst van belang wordt waargenomen. Bijvoorbeeld de tijd die het kost voordat iemand ziek wordt na een behandeling, de bedrijfstijd van een mechanisch apparaat nadat een onderdeel is vervangen, of de tijd die een persoon besteedt aan het kijken naar een video. In het bijzonder ontwikkelen we een altijd-geldige tegenhanger van de logrank toets, wellicht de belangrijkste toets die wordt gebruikt voor de vergelijking van de overlevingstijden van twee groepen patiënten in medische onderzoeken. Deze toets is met succes gebruikt door enkele van de medeauteurs van het hoofdstuk bij het beoordelen van de werkzaamheid van het Bacillus Calmette-Guérin (BCG)-vaccin, een vaccin tegen tuberculose, bij de behandeling van de Covid-19-ziekte in de vroege dagen van de pandemie van 2019.

Hoofdstuk 2 toont optimale toetsen voor hypothesen die bepaalde vormen van symmetrie vertonen, dat wil zeggen problemen die onveranderd blijven onder bepaalde transformaties. Dit is met name belangrijk omdat probabilistische modellen voor fysieke systemen deze symmetrieën vertonen. Fysische theorieën blijven bijvoorbeeld onveranderd onder veranderingen in de eenheden van meting, het referentiekader ten opzichte waarvan metingen worden verricht, of de volgorde waarin ze worden uitgevoerd. Het belangrijkste resultaat van dit hoofdstuk kan worden samengevat in de stelling dat als een probleem een dergelijke invariantie vertoont, de optimale test ook dezelfde invariantie moet respecteren. Dit is in principe niet triviaal, omdat de verzameling van alle mogelijke altijd-geldige toetsen veel groter is dan de verzameling van invariante toetsen.

Hoofdstuk 4 biedt een oplossing voor het probleem van voorspellingen met behulp van deskundig advies. In dit klassieke probleem wordt een spel gespeeld in rondes en moet de speler kiezen uit een vast aantal acties. Na elke beslissing krijgt de speler de kwaliteit van alle acties te zien in de vorm van een getal. Als de speler van tevoren wist welke actie de beste is, zou hij die vanaf het begin kiezen. De uitdaging van dit probleem ligt in het ontbreken van deze kennis. We kunnen bijvoorbeeld elke dag vertrouwen op meerdere weersvoorspellingen, en elke dag kunnen we evalueren hoe goed elke voorspeller was door het verschil te meten tussen hun voorspellingen en het waargenomen weer. Opmerkelijk genoeg is het zelfs zonder enige aannames over hoe de kwaliteit van de acties wordt toegekend - ze zouden kunnen worden gegeven door een kwaadwillende tegenstander - mogelijk om strategieën te ontwerpen die niet veel slechter presteren dan wanneer de beste actie vanaf het begin was gekozen. De bijdrage van dit hoofdstuk is een algoritme voor dit voorspellingsprobleem wanneer de grootte van de kwaliteit van de acties ordes van grootte kan verschillen.

Hoofdstuk 5 introduceert een stuk wiskundige notatie dat is ontworpen om bepaalde soorten probabilistische argumenten te bestuderen. Deze probabilistische argumenten worden gebruikt om te beoordelen hoe ver de empirische prestaties van statistische en voorspellingssystemen afwijken van hun theoretische waarde.

Summary

From a broad perspective, the main topic of this dissertation is statistical hypothesis testing. This topic is important because scientific hypothesis are often translated to statistical hypothesis testing problems. For instance, when studying the effect of a medical treatment, researchers compare the clinical outcomes of a group of people that received the treatment to those of a control group that did not receive it. If the treatment is effective, the clinical outcomes of the treatment group are expected to be better than those of the control group. Hence, the scientific hypothesis about the efficacy of the treatment can be studied through the statistical comparison of the outcomes of the two groups of patients.

This dissertation is concerned with flexible methods for statistical monitoring. To gain some appreciation of the type of flexibility that is alluded, we compare the methods that are the subject of our study to classic fixed-sample methods. Most classic methods taught in introductory statistics courses demand that researchers plan the experiments in advance. In particular, classic methods demand that researchers gather a fixed amount of data before performing any analysis. This requirements are necessary to control errors that may occur due to random chance. In the medical example, this entails that researchers must witness the clinical outcomes of a fixed number—a fixed sample size—of patients before analyzing their findings. Unfortunately, the fixed-sample-size assumption can be restricting in certain situations. For instance, researchers may be interested in analyzing the data as it is gathered in order to stop the experiment earlier if needed. This is the case, for instance, in human survival-time experiments. Indeed, early stopping may even be an ethical imperative when a treatment shows to be life-saving. Powerful methods have been designed to overcome this restriction. In particular, the methods that provide the most flexibility, known as anytime valid, allow researchers to either continue or stop a running experiment, or even to start a new one at any moment without altering the statistical validity of their analysis. These flexible anytime-valid methods are the main topic of this dissertation.

In this work, a number of mathematical results on the theory of anytime-valid testing and prediction are shown. As it is often the case with mathematical works, the results in this dissertation can be applied in multiple contexts. The contents of this dissertation pertain the design of optimal—in a specific sense—anytime-valid tests in abstract settings that aim at capturing relevant applications.

Chapter 3 shows an anytime-valid test for the analysis of time-to-event data. This is useful when monitoring experiments whose focus is the time elapsed until an outcome of interest is observed. For instance, the time that it takes for a person to become ill after a treatment, the up time of a mechanical device after a piece is replaced or the time that a person spends watching a video. In particular, we develop an anytime-valid counterpart of the logrank test, arguably the most important test used for the com-

parison of the survival times of two groups of patients in medical trials. This test has been successfully used by some of the coauthors of the chapter in assessing the efficacy of the Bacillus Calmette-Guérin (BCG) vaccine—a vaccine against tuberculosis—in treating Covid-19 disease in the early days of the 2019 pandemic.

Chapter 2 shows optimal tests for hypothesis problems that present certain types of symmetry, that is, problems that remain unchanged under certain transformations. This is specially important because probabilistic models for physical systems show these symmetries. For instance, physical theories remain unchanged under changes in the units of measurement, the frame of reference with respect to which measurements are made or the order in which they are executed. The main result of this chapter can be summarized in the statement that if a problem shows such an invariance, the optimal test should also respect the same invariance. This is, in principle, nontrivial, because the set of all possible anytime-valid tests is much larger than the set of invariant tests.

Chapter 4 provides a solution to the problem of prediction with expert advice. In this classic problem, a game is played in rounds and the player must decide among a fixed number of actions. After each decision, the player is shown the quality of all actions in the form of a number. If the player knew in advance which action is best, they would choose it from the onset. The challenge of this problem lies in the absence of this knowledge. For example, we may count on several weather forecasts each day, and each day we can evaluate how well each forecaster was by measuring the difference between their predictions and the observed weather. Remarkably, even without any assumptions on how the quality of the actions are assigned—they could be given by an evil adversary—, it is possible to design strategies that perform not much worse than having chosen the best action from the beginning. The contribution of this chapter is an algorithm to this problem of prediction when the magnitude of the quality of the actions may vary by orders of magnitude.

Chapter 5 introduces a piece of mathematical notation designed to study certain types of probabilistic arguments. These probabilistic arguments are used to judge how far the empirical performance of statistical and prediction systems is from their theoretical value.

Resumen

A grandes rasgos, el objeto de esta disertación es la prueba estadística de hipótesis. Este es un tema importante porque, con frecuencia, las hipótesis científicas pueden probarse a través de pruebas estadísticas de hipótesis. Por ejemplo, durante el estudio de un tratamiento médico, un grupo de investigadores puede comparar los resultados clínicos de dos grupo de personas; uno que recibió el tratamiento y los de un grupo de control que no lo recibió. Si el tratamiento es efectivo, se espera que los resultados clínicos del grupo que recibió el tratamiento sean mejores que los del grupo de control. Por lo tanto, la hipótesis científica sobre la eficacia del tratamiento puede ser estudiada a través de la comparación estadística entre los resultados clínicos de los dos grupos de pacientes. Para este fin, es crucial contar con métodos estadísticos confiables.

Esta disertación se ocupa de diseñar métodos flexibles para el monitoreo estadístico de experimentos científicos. Para entender de qué tipo de flexibilidad se habla, es útil comparar los métodos de los que se ocupa esta tésis con métodos clásicos que requieren un tamaño de muestra predeterminada. La mayoría de métodos clásicos que se enseñan en cursos introductorios a la estadística exigen que los investigadores planeen sus experimentos con antelación. En particular, los métodos clásicos prescriben que los investigadores recojan una muestra de un tamaño fijo antes de hacer cualquier análisis. Este requerimiento es necesario para controlar la probabilidad de cometer errores que son producto del muestreo aleatorio. En el ejemplo anterior, este requerimiento hace que los investigadores deban esperar a presenciar los resultados clínicos de un número fijo de pacientes antes de analizar sus hallazgos. Desafortunadamente, esto puede ser una limitación en algunas aplicaciones. Por ejemplo, en estudios clínicos sobre el tiempo de supervivencia de ciertos pacientes, detener un experimento se puede convertir en un imperativo ético si el tratamiento demuestra salvar vidas humanas. Existen métodos diseñados para superar esta restricción. En particular, los métodos que proveen una mayor flexibilidad son aquellos que son siempre válidos (*anytime valid*). Estos métodos son el objeto principal de esta disertación.

En este trabajo se presenta una colección de resultados matemáticos sobre la teoría de predicción y de pruebas estadísticas siempre válidas. Dada la naturaleza matemática de los resultados aquí mostrados, estos pueden ser aplicados en múltiples contextos. Esta disertación trata sobre el diseño de procedimientos siempre válidos en marcos abstractos con el propósito de capturar aplicaciones relevantes.

En el capítulo 3 se muestra una prueba siempre válida para el análisis de supervivencia. Este tipo de análisis es útil para monitorear experimentos cuyo objetivo es estudiar el tiempo que toma presenciar un evento de interés. Por ejemplo, se puede tratar del tiempo que toma antes de que una persona presente una enfermedad tras un tratamiento, del tiempo que le toma a una máquina descomponerse, o del tiempo que una persona dura viendo un video. El capítulo muestra una versión siempre válida de la prueba de Mantel-Cox, que es utilizada para comparar los tiempos de supervivencia de dos grupos de pacientes en ensayos clínicos. Típicamente, se utiliza para comparar un tratamiento y un procedimiento de control. La prueba aquí descrita ha sido utilizada con éxito por algunos de los coautores del capítulo para evaluar la eficacia de la vacuna BCG (bacilo de Calmette y Guérin) –una vacuna contra la tuberculosis– para tratar la enfermedad por coronavirus (COVID-19) durante la etapa temprana de la pandemia de 2019.

El capítulo 2 muestra pruebas siempre válidas óptimas para problemas que presentan ciertos tipos de simetrías, esto es, problemas que permanecen invariantes bajo ciertas transformaciones. La importancia de este tipo de problemas radica en que los modelos probabilísticos para sistemas físicos presentan este tipo de invarianzas. Por ejemplo, las predicciones de las teorías físicas no cambian si se alteran las unidades de medida, el marco de referencia con respecto al cual se realizan mediciones o el orden en que estas se realizan. El resultado principal de este capítulo se puede resumir en pocas palabras: si los modelos considerados presentan una invarianza, una prueba siempre válida óptima debe presentar la misma invarianza. En principio, eso no es obvio, porque el conjunto de todos las pruebas siempre válidas en mucho más grande que el conjunto de las que adicionalmente son invariantes.

El capítulo 4 presenta una solución a una modificación del problema de predicción con consejo experto. Se trata de un problema clásico. En él, un jugador debe decidir en cada ronda entre un número predeterminado de acciones. Después de elegir una acción, el jugador puede ver cuál habría sido su desempeño si hubiera elegido cualquier otra acción. Por ejemplo, si cada día contamos con múltiples predicciones sobre cuánto va a llover y debemos decidir cuál creer, cada día podemos evaluar nuestras decisiones comparando las predicciones con la realidad. En este caso y en general, el desempeño de cada acción es juzgado con una medida numérica; el error de cada predicción, por ejemplo. Si el jugador supiera cuál es la mejor acción, siempre la elegiría, pero ésta no es conocida. Sorprendentemente, incluso si no se asume nada sobre el mecanismo que genera las observaciones, es posible diseñar estrategias que no son mucho peores que haber elegir la mejor acción desde el principio. La contribución de este capítulo es un algoritmo para resolver este problema cuando la magnitud numérica de la calidad de las acciones puede variar en órdenes de magnitud.

El capítulo 5 introduce notación matemática para estudiar cierto tipo de argumentos probabilísticos. Estos argumentos son utilizados para juzgar la diferencia entre el desempeño observado y el desempeño teórico de sistemas estadísticos y de predicción.

Curriculum Vitae

Muriel was born in Bogotá in 1993. He spent his high school years between El Colegio Filipense Sagrado Corazón de Jesús in Alcalá de Henares, Spain (2006-2008), and El Colegio Tolimense, in Ibagué, Colombia (2008-2009). He went on to study a BSc in Physics at the University of the Andes (2010-2015) in Bogotá and a BSc in Mathematics (2011-2016) at the same university. After a year working at the Astronomy Group of the University of the Andes, Muriel went to Amsterdam, where he completed a MSc degree (cum laude) in Mathematics at the University of Amsterdam (2016-2018). With the guidance of Peter Grünwald and Wouter Koolen, at the Machine Learning group of *Centrum Wiskunde & Informatica* (CWI), he completed is PhD in 2023.

List of Publications

The contents of this dissertation are based on the following publications.

- Chapter 2 is based on

  M. F. Pérez-Ortiz, T. Lardy, R. de Heide, and P. Grünwald. E-Statistics, Group Invariance and Anytime Valid Testing, Aug. 2022. URL http://arxiv.org/abs/2208.07610. arXiv:2208.07610 [math, stat], under submission.

- Chapter 3 is based on

  J. ter Schure, M. F. Pérez-Ortiz, A. Ly, and P. Grünwald. The Safe Logrank Test: Error Control under Continuous Monitoring with Unlimited Horizon, July 2021. URL http://arxiv.org/abs/2011.06931. arXiv:2011.06931 [math, stat], under submission.

- Chapter 4 is based on

  M. F. Pérez-Ortiz and W. M. Koolen. Luckiness in Multiscale Online Learning. *Advances in Neural Information Processing Systems*, 35:25160–25170, Dec. 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/a0d2345b43e66fa946155c98899dc03b-Abstract-Conference.html.

- Chapter 5 is based on

  P. D. Grünwald, M. F. Pérez-Ortiz, and Z. Mhammedi. Exponential Stochastic Inequality, Apr. 2023. URL http://arxiv.org/abs/2304.14217. arXiv: 2304.14217 [math, stat].

Bibliography

P. Alquier. User-friendly introduction to PAC-Bayes bounds, Mar. 2023. URL http://arxiv.org/abs/2110.11216. arXiv:2110.11216 [cs, math, stat].

V. Amrhein, D. Trafimow, and S. Greenland. Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician*, 73(sup1):262–270, Mar. 2019. ISSN 0003-1305. doi: 10.1080/00031305.2018.1543137. URL https://doi.org/10.1080/00031305.2018.1543137. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2018.1543137.

P. K. Andersen, O. Borgan, R. D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Science & Business Media, 1993. ISBN 978-1-4612-4348-9. Google-Books-ID: fh3SBwAAQBAJ.

S. Andersson. Distributions of Maximal Invariants Using Quotient Measures. *Annals of Statistics*, 10(3):955–961, Sept. 1982. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345885. URL https://projecteuclid.org/euclid.aos/1176345885. Publisher: Institute of Mathematical Statistics.

F. J. Anscombe. Fixed-Sample-Size Analysis of Sequential Observations. *Biometrics*, 10(1):89–100, 1954. ISSN 0006-341X. doi: 10.2307/3001665. URL https://www.jstor.org/stable/3001665. Publisher: [Wiley, International Biometric Society].

P. Armitage. The Search for Optimality in Clinical Trials. *International Statistical Review / Revue Internationale de Statistique*, 53(1):15–24, 1985. ISSN 0306-7734. doi: 10.2307/1402871. URL https://www.jstor.org/stable/1402871. Publisher: [Wiley, International Statistical Institute (ISI)].

J.-Y. Audibert. PAC-Bayesian statistical learning theory. *These de doctorat de l'Université Paris*, 6:29, 2004.

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, Aug. 2009. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS623. URL https://projecteuclid.org/euclid.aos/1245332827.

P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, July 2006. ISSN 0178-8051, 1432-2064. doi: 10.1007/s00440-005-0462-3. URL https://link.springer.com/article/10.1007/s00440-005-0462-3.

D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson. Redefine statistical significance. *Nature Human Behaviour*, 2(1):6–10, Jan. 2018. ISSN 2397-3374. doi: 10.1038/s41562-017-0189-z. URL https://www.nature.com/articles/s41562-017-0189-z. Number: 1 Publisher: Nature Publishing Group.

J. O. Berger and D. Sun. Objective priors for the bivariate normal model. *Annals of Statistics*, 36(2):963–982, Apr. 2008. ISSN 0090-5364, 2168-8966. doi: 10.1214/07-AOS501. URL https://projecteuclid.org/euclid.aos/1205420525. Publisher: Institute of Mathematical Statistics.

J. O. Berger and R. L. Wolpert. *The Likelihood Principle*. Institute of Mathematical Statistics, 1984. ISBN 978-0-940600-06-5. Google-Books-ID: a6NUBG4WZPMC.

J. O. Berger, L. R. Pericchi, and J. A. Varshavsky. Bayes Factors and Marginal Distributions in Invariant Situations. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 60(3):307–321, 1998. ISSN 0581-572X. URL https://www.jstor.org/stable/25051210. Publisher: Springer.

R. H. Berk. A Note on Sufficiency and Invariance. *The Annals of Mathematical Statistics*, 43(2):647–650, Apr. 1972. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177692645. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-43/issue-2/A-Note-on-Sufficiency-and-Invariance/10.1214/aoms/1177692645.full. Publisher: Institute of Mathematical Statistics.

J. L. Bhowmik and M. L. King. Maximal invariant likelihood based testing of semi-linear models. *Statistical Papers*, 48(3):357–383, Sept. 2007. ISSN 1613-9798. doi: 10.1007/s00362-006-0342-7. URL https://doi.org/10.1007/s00362-006-0342-7.

J. V. Bondar. Borel Cross-Sections and Maximal Invariants. *Annals of Statistics*, 4 (5):866–877, Sept. 1976. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176343585. URL https://projecteuclid.org/euclid.aos/1176343585. Publisher: Institute of Mathematical Statistics.

J. V. Bondar and P. Milnes. Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(1):103–128, 1981. Publisher: Springer.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, Feb. 2013. ISBN 978-0-19-953525-5. Google-Books-ID: koNqWRluhP0C.

N. Bourbaki. *Integration II: Chapters 7–9*. Elements of Mathematics. Springer-Verlag, Berlin Heidelberg, 2004. ISBN 978-3-540-20585-2. doi: 10.1007/978-3-662-07931-7. URL `https://www.springer.com/gp/book/9783540205852`.

L. Breiman. Optimal Gambling Systems for Favorable Games. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 4.1:65–79, Jan. 1961. URL `https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fourth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Optimal-Gambling-Systems-for-Favorable-Games/bsmsp/1200512159`. Publisher: University of California Press.

S. Bubeck, N. R. Devanur, Z. Huang, and R. Niazadeh. Multi-scale Online Learning: Theory and Applications to Online Auctions and Pricing. *Journal of Machine Learning Research*, 20(62):1–37, 2019. ISSN 1533-7928. URL `http://jmlr.org/papers/v20/17-498.html`.

L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. PAC-Bayesian Bounds based on the Rényi Divergence. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 435–444. PMLR, May 2016. URL `https://proceedings.mlr.press/v51/begin16.html`. ISSN: 1938-7228.

O. Catoni. Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. *IMS Lecture Notes Monograph Series*, 56:1–163, 2007. ISSN 0749-2170. doi: 10.1214/074921707000000391. URL `http://arxiv.org/abs/0712.0248`. arXiv: 0712.0248.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

N. Cesa-Bianchi, P. Gaillard, C. Gentile, and S. Gerchinovitz. Algorithmic Chaining and the Role of Partial Feedback in Online Nonparametric Learning. In *Conference on Learning Theory*, pages 465–481, June 2017. URL `http://proceedings.mlr.press/v65/cesa-bianchi17a.html`.

J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51 (3):287–317, 1997. ISSN 1467-9574. doi: 10.1111/1467-9574.00056. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9574.00056`.

L. Chen, H. Luo, and C.-Y. Wei. Impossible Tuning Made Possible: A New Expert Algorithm and Its Applications. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 1216–1259. PMLR, July 2021. URL `https://proceedings.mlr.press/v134/chen21f.html`. ISSN: 2640-3498.

Y. J. Choe and A. Ramdas. Comparing Sequential Forecasters, Dec. 2022. URL `http://arxiv.org/abs/2110.00115`. arXiv:2110.00115 [cs, math, stat].

B. Chugg, H. Wang, and A. Ramdas. A unified recipe for deriving (time-uniform) PAC-Bayes bounds, Feb. 2023. URL `http://arxiv.org/abs/2302.03421`. arXiv:2302.03421 [cs, math, stat].

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience, New York, NY, USA, 2006. ISBN 978-0-471-24195-9.

D. R. Cox. Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2):290–299, Apr. 1952. ISSN 1469-8064, 0305-0041. doi: 10.1017/S030500410002764X. URL `https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/abs/sequential-tests-for-composite-hypotheses/3424D357C438E368282290BE6D28A99A`. Publisher: Cambridge University Press.

D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 0035-9246. URL `https://www.jstor.org/stable/2985181`. Publisher: [Royal Statistical Society, Wiley].

D. R. Cox. Partial Likelihood. *Biometrika*, 62(2):269–276, 1975. ISSN 0006-3444. doi: 10.2307/2335362. URL `https://www.jstor.org/stable/2335362`. Publisher: [Oxford University Press, Biometrika Trust].

A. Cutkosky and F. Orabona. Black-Box Reductions for Parameter-free Online Learning in Banach Spaces. In *Conference On Learning Theory*, pages 1493–1529, July 2018. URL `http://proceedings.mlr.press/v75/cutkosky18a.html`.

D. A. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, July 1967. doi: 10.1073/pnas.58.1.66. URL `https://www.pnas.org/doi/abs/10.1073/pnas.58.1.66`. Publisher: Proceedings of the National Academy of Sciences.

D. A. Darling and H. Robbins. Some further remarks on inequalities for sample sums*. *Proceedings of the National Academy of Sciences*, 60(4):1175–1182, Aug. 1968a. doi: 10.1073/pnas.60.4.1175. URL `https://www.pnas.org/doi/abs/10.1073/pnas.60.4.1175`. Publisher: Proceedings of the National Academy of Sciences.

D. A. Darling and H. Robbins. Some nonparametric sequential tests with power one*. *Proceedings of the National Academy of Sciences*, 61(3):804–809, Nov. 1968b. doi: 10.1073/pnas.61.3.804. URL `https://www.pnas.org/doi/abs/10.1073/pnas.61.3.804`. Publisher: Proceedings of the National Academy of Sciences.

A. P. Dawid. Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292, 1984. ISSN 0035-9238. doi: 10.2307/2981683. URL `https://www.jstor.org/stable/2981683`. Publisher: [Royal Statistical Society, Wiley].

A. P. Dawid, M. Stone, and J. V. Zidek. Marginalization Paradoxes in Bayesian and Structural Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(2):189–233, 1973. ISSN 0035-9246. URL `https://www.jstor.org/stable/2984907`. Publisher: [Royal Statistical Society, Wiley].

R. de Heide and P. D. Grünwald. Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review*, 28(3):795–812, June 2021. ISSN 1531-5320. doi: 10.3758/s13423-020-01803-x. URL `https://doi.org/10.3758/s13423-020-01803-x`.

S. de Rooij, T. van Erven, P. D. Grünwald, and W. M. Koolen. Follow the Leader If You Can, Hedge If You Must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014. ISSN 1533-7928. URL `http://jmlr.org/papers/v15/rooij14a.html`.

D. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, 13(2):99–124, 1998.

J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12 (61):2121–2159, 2011. ISSN 1533-7928. URL `http://jmlr.org/papers/v12/duchi11a.html`.

R. Durrett. *Probability: Theory and Examples*. Number 49 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, Apr. 2019. ISBN 978-1-108-47368-2. Google-Books-ID: b22MDwAAQBAJ.

M. L. Eaton. Group Invariance Applications in Statistics. *Regional Conference Series in Probability and Statistics*, 1:i–133, 1989. ISSN 1935-5912. URL `https://www.jstor.org/stable/4153172`.

M. L. Eaton and W. D. Sudderth. Consistency and Strong Inconsistency of Group-Invariant Predictive Inferences. *Bernoulli*, 5(5):833–854, 1999. ISSN 1350-7265. doi: 10.2307/3318446. URL `https://www.jstor.org/stable/3318446`. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability.

M. L. Eaton and W. D. Sudderth. Group invariant inference and right Haar measure. *Journal of Statistical Planning and Inference*, 103(1):87–99, Apr. 2002. ISSN 0378-3758. doi: 10.1016/S0378-3758(01)00199-9. URL http://www.sciencedirect.com/science/article/pii/S0378375801001999.

B. Efron. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, 72(359):557–565, Sept. 1977. ISSN 0162-1459. doi: 10.1080/01621459.1977.10480613. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1977.10480613. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1977.10480613.

R. Fisher. Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(1):69–78, 1955. ISSN 0035-9246. URL https://www.jstor.org/stable/2983785. Publisher: [Royal Statistical Society, Wiley].

T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, Sept. 2011. ISBN 978-1-118-15066-5.

D. J. Foster, S. Kale, M. Mohri, and K. Sridharan. Parameter-Free Online Learning via Model Selection. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6020–6030. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7183-parameter-free-online-learning-via-model-selection.pdf.

Y. Freund. Open Problem: Second order regret bounds based on scaling time. In *Conference on Learning Theory*, pages 1651–1654, 2016.

Y. Freund and R. E. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55 (1):119–139, Aug. 1997. ISSN 0022-0000. doi: 10.1006/jcss.1997.1504. URL https://www.sciencedirect.com/science/article/pii/S002200009791504X.

P. Gaillard and S. Gerchinovitz. A Chaining Algorithm for Online Nonparametric Regression. In *Conference on Learning Theory*, pages 764–796, June 2015. URL http://proceedings.mlr.press/v40/Gaillard15.html.

P. Gaillard, G. Stoltz, and T. van Erven. A second-order bound with excess losses. In *Proceedings of The 27th Conference on Learning Theory*, pages 176–196. PMLR, May 2014. URL https://proceedings.mlr.press/v35/gaillard14.html. ISSN: 1938-7228.

J. W. Gibbs. *Elementary Principles in Statistical Mechanics*. Charles Scribner's Sons, New York, 1902.

N. Giri, J. Kiefer, and C. Stein. Minimax Character of Hotelling's T^2 Test in the Simplest Case. *The Annals of Mathematical Statistics*, 34(4):1524–1535, Dec. 1963. ISSN 0003-4851, 2168-8990. doi:

10.1214/aoms/1177703884. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-34/issue-4/Minimax-Character-of-Hotellings-T2-Test-in-the-Simplest-Case/10.1214/aoms/1177703884.full. Publisher: Institute of Mathematical Statistics.

S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N. Goodman, and D. G. Altman. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31:337–350, 2016. ISSN 0393-2990. doi: 10.1007/s10654-016-0149-3. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4877414/.

P. Grünwald. The Safe Bayesian. In N. H. Bshouty, G. Stoltz, N. Vayatis, and T. Zeugmann, editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 169–183. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-34106-9.

P. Grünwald. Beyond Neyman-Pearson, Feb. 2023. URL http://arxiv.org/abs/2205.00901. arXiv:2205.00901 [stat].

P. Grünwald, R. de Heide, and W. Koolen. Safe Testing. *arXiv:1906.07801 [cs, math, stat]*, June 2020. URL http://arxiv.org/abs/1906.07801. Accepted for publication as a read paper in the Journal of the Royal Statistical Society, Series B. 2023.

P. D. Grünwald. Viewing all models as "probabilistic". In *Proceedings of the twelfth annual conference on Computational learning theory*, COLT '99, pages 171–182, New York, NY, USA, July 1999. Association for Computing Machinery. ISBN 978-1-58113-167-3. doi: 10.1145/307400.307436. URL https://dl.acm.org/doi/10.1145/307400.307436.

P. D. Grünwald. *The Minimum Description Length Principle*. Adaptive Computation and Machine Learning series. Mar. 2007. ISBN 978-0-262-52963-1.

P. D. Grünwald and N. A. Mehta. A tight excess risk bound via a unified PAC-Bayesian–Rademacher–Shtarkov–MDL complexity. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, pages 433–465. PMLR, Mar. 2019. URL https://proceedings.mlr.press/v98/grunwald19a.html. ISSN: 2640-3498.

P. D. Grünwald and N. A. Mehta. Fast Rates for General Unbounded Loss Functions: From ERM to Generalized Bayes. *Journal of Machine Learning Research*, 21(56):1–80, 2020. ISSN 1533-7928. URL http://jmlr.org/papers/v21/18-488.html.

P. D. Grünwald and T. Roos. Minimum description length revisited. *International Journal of Mathematics for Industry*, 11(01):1930001, Dec. 2019. ISSN 2661-3352. doi: 10.1142/S2661335219300018. URL https://www.worldscientific.com/doi/10.1142/S2661335219300018. Publisher: World Scientific Publishing Co.

P. D. Grünwald and T. van Ommen. Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, 12(4): 1069–1103, Dec. 2017. ISSN 1936-0975, 1931-6690. doi: 10.1214/17-BA1085. URL https://projecteuclid.org/euclid.ba/1510974325.

P. D. Grünwald, T. Steinke, and L. Zakynthinou. PAC-Bayes, MAC-Bayes and Conditional Mutual Information: Fast rate bounds that handle general VC classes. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 2217–2247. PMLR, July 2021. URL https://proceedings.mlr.press/v134/grunwald21a.html. ISSN: 2640-3498.

P. D. Grünwald, M. F. Pérez-Ortiz, and Z. Mhammedi. Exponential Stochastic Inequality, Apr. 2023. URL http://arxiv.org/abs/2304.14217. arXiv: 2304.14217 [math, stat].

B. Guedj. A Primer on PAC-Bayesian Learning, May 2019. URL http://arxiv.org/abs/1901.05353. arXiv:1901.05353 [cs, stat].

W. J. Hall, R. A. Wijsman, and J. K. Ghosh. The Relationship Between Sufficiency and Invariance with Applications in Sequential Analysis. *The Annals of Mathematical Statistics*, 36(2):575–614, 1965. ISSN 0003-4851. URL https://www.jstor.org/stable/2238164. Publisher: Institute of Mathematical Statistics.

W. J. Hall, R. A. Wijsman, and J. K. Ghosh. Correction: The Relationship Between Sufficiency and Invariance with Applications in Sequential Analysis. *The Annals of Statistics*, 23(2):705–705, 1995. ISSN 0090-5364. URL https://www.jstor.org/stable/2242359. Publisher: Institute of Mathematical Statistics.

E. Hazan. Introduction to Online Convex Optimization, Dec. 2021. URL http://arxiv.org/abs/1909.05207. arXiv:1909.05207 [cs, math, stat].

F. Hellström and G. Durisi. A New Family of Generalization Bounds Using Samplewise Evaluated CMI. *Advances in Neural Information Processing Systems*, 35:10108–10121, Dec. 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/41b6674c28a9b93ec8d22a53ca25bc3b-Abstract-Conference.html.

A. Henzi and J. F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, page asab047, Sept. 2021. ISSN 0006-3444. doi: 10.1093/biomet/asab047. URL https://doi.org/10.1093/biomet/asab047.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Exponential line-crossing inequalities. *arXiv:1808.03204 [math]*, Aug. 2018a. URL http://arxiv.org/abs/1808.03204. arXiv: 1808.03204.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Uniform, nonparametric, non-asymptotic confidence sequences. *arXiv:1810.08240 [math, stat]*, Oct. 2018b. URL http://arxiv.org/abs/1810.08240. arXiv: 1810.08240.

Y.-G. Hsieh, K. Antonakopoulos, and P. Mertikopoulos. Adaptive Learning in Continuous Games: Optimal Regret Bounds and Convergence to Nash Equilibrium. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pages 2388–2422. PMLR, July 2021. URL `https://proceedings.mlr.press/v134/hsieh21a.html`. ISSN: 2640-3498.

K. Jang, K.-S. Jun, I. Kuzborskij, and F. Orabona. Tighter PAC-Bayes Bounds Through Coin-Betting, Feb. 2023. URL `http://arxiv.org/abs/2302.05829`. arXiv:2302.05829 [cs, stat].

S. H. Jeffreys and S. H. Jeffreys. *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. Oxford University Press, Oxford, New York, third edition, third edition edition, Aug. 1998. ISBN 978-0-19-850368-2.

K. V. Jespersen, V. Pando-Naude, J. Koenig, P. Jennum, and P. Vuust. Listening to music for insomnia in adults. *The Cochrane Database of Systematic Reviews*, 8(8): CD010459, Aug. 2022. ISSN 1469-493X. doi: 10.1002/14651858.CD010459.pub3.

K. Joag-Dev and F. Proschan. Negative Association of Random Variables with Applications. *The Annals of Statistics*, 11(1):286–295, 1983. ISSN 0090-5364. URL `https://www.jstor.org/stable/2240482`.

J. D. Kalbfleisch and R. L. Prentice. Marginal Likelihoods Based on Cox's Regression and Life Model. *Biometrika*, 60(2):267–278, 1973. ISSN 0006-3444. doi: 10.2307/2334538. URL `https://www.jstor.org/stable/2334538`. Publisher: [Oxford University Press, Biometrika Trust].

T. Kariya. Locally Robust Tests for Serial Correlation in Least Squares Regression. *The Annals of Statistics*, 8(5):1065–1070, Sept. 1980. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345143. URL `https://projecteuclid.org/journals/annals-of-statistics/volume-8/issue-5/Locally-Robust-Tests-for-Serial-Correlation-in-Least-Squares-Regression/10.1214/aos/1176345143.full`. Publisher: Institute of Mathematical Statistics.

J. L. Kelly Jr. A New Interpretation of Information Rate. *Bell System Technical Journal*, 35(4):917–926, 1956. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1956.tb03809.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1956.tb03809.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1956.tb03809.x.

K. Kim and D. L. DeMets. Confidence Intervals Following Group Sequential Tests in Clinical Trials. *Biometrics*, 43(4):857–864, 1987. ISSN 0006-341X. doi: 10.2307/2531539. URL `https://www.jstor.org/stable/2531539`. Publisher: [Wiley, International Biometric Society].

J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer, New York, NY,

2003. ISBN 978-0-387-95399-1 978-0-387-21645-4. doi: 10.1007/b97377. URL http://link.springer.com/10.1007/b97377.

W. M. Koolen and T. van Erven. Second-order Quantile Methods for Experts and Combinatorial Games. In *Proceedings of The 28th Conference on Learning Theory*, pages 1155–1175. PMLR, June 2015. URL https://proceedings.mlr.press/v40/Koolen15a.html. ISSN: 1938-7228.

W. M. Koolen, P. D. Grünwald, and T. van Erven. Combining Adversarial Guarantees and Stochastic Fast Rates in Online Learning. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/db116b39f7a3ac5366079b1d9fe249a5-Abstract.html.

I. Kuzborskij and N. Cesa-Bianchi. Locally-Adaptive Nonparametric Online Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1679–1689. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/12780ea688a71dabc284b064add459a4-Abstract.html.

T. L. Lai. On Confidence Sequences. *The Annals of Statistics*, 4(2):265–280, Mar. 1976. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176343406. URL https://projecteuclid.org/journals/annals-of-statistics/volume-4/issue-2/On-Confidence-Sequences/10.1214/aos/1176343406.full. Publisher: Institute of Mathematical Statistics.

T. L. Lai. Sequential Analysis: Some Classical Problems and New Challenges. *Statistica Sinica*, 11(2):303–351, 2001. ISSN 1017-0405. URL https://www.jstor.org/stable/24306854. Publisher: Institute of Statistical Science, Academia Sinica.

J. Langford and J. Shawe-Taylor. PAC-Bayes & Margins. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/hash/68d3098125488874 00e375eaa036d2f1-Abstract.html.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer-Verlag, New York, 3 edition, 2005. ISBN 978-0-387-98864-1. doi: 10.1007/0-387-27605-X. URL https://www.springer.com/gp/book/9780387988641.

J. Li and A. Barron. Mixture Density Estimation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://papers.nips.cc/paper/1999/hash/a0f3601dc682036423013a5d965db9aa-Abstract.html.

Q. J. Li. *Estimation of Mixture Models*. PhD Thesis, Yale University, New Haven, CT, USA, 1999.

F. Liang and A. Barron. Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50(11):2708–2726, Nov. 2004. ISSN 1557-9654. doi: 10.1109/TIT.2004.836922. Conference Name: IEEE Transactions on Information Theory.

M. Lindon and A. Malek. Anytime-Valid Inference for Multinomial Count Data, May 2022. URL `http://arxiv.org/abs/2011.03567`. arXiv:2011.03567 [math, stat].

N. Littlestone and M. K. Warmuth. The Weighted Majority Algorithm. *Information and Computation*, 108(2):212–261, Feb. 1994. ISSN 0890-5401. doi: 10.1006/inco.1994.1009. URL `https://www.sciencedirect.com/science/article/pii/S0890540184710091`.

G. Lugosi and G. Neu. Generalization Bounds via Convex Analysis. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 3524–3546. PMLR, June 2022. URL `https://proceedings.mlr.press/v178/lugosi22a.html`. ISSN: 2640-3498.

O.-A. Maillard. *Mathematics of Statistical Sequential Decision Making*. Habilitation à diriger des recherches, Université de Lille, Sciences et Technologies, Feb. 2019. URL `https://hal.archives-ouvertes.fr/tel-02162189`.

N. Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170, Mar. 1966. ISSN 0069-0112.

T. V. Marinov and J. Zimmert. The Pareto Frontier of model selection for general Contextual Bandits. In *Advances in Neural Information Processing Systems*, volume 34, pages 17956–17967. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/hash/9570efef719d705326f0ff817ef084e6-Abstract.html`.

A. Maurer. A Note on the PAC Bayesian Theorem, Nov. 2004. URL `http://arxiv.org/abs/cs/0411099`. arXiv:cs/0411099.

D. A. McAllester. Some PAC-Bayesian Theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 230–234, New York, NY, USA, 1998. ACM. ISBN 978-1-58113-057-7. doi: 10.1145/279943.279989. URL `http://doi.acm.org/10.1145/279943.279989`.

D. V. Mehrotra and A. J. Roth. Relative risk estimation and inference using a generalized logrank statistic. *Statistics in Medicine*, 20 (14):2099–2113, 2001. ISSN 1097-0258. doi: 10.1002/sim.854. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.854`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.854.

Z. Mhammedi, P. D. Grunwald, and B. Guedj. PAC-Bayes Un-Expected Bernstein Inequality. *arXiv:1905.13367 [cs, stat]*, May 2019. URL http://arxiv.org/abs/1905.13367. arXiv: 1905.13367.

J. Mourtada and S. Gaïffas. On the optimality of the Hedge algorithm in the stochastic regime. *Journal of Machine Learning Research*, 20(83):1–28, 2019. ISSN 1533-7928. URL http://jmlr.org/papers/v20/18-869.html.

A. G. Nogales and J. A. Oyola. Some Remarks on Sufficiency, Invariance and Conditional Independence. *The Annals of Statistics*, 24(2):906–909, 1996. ISSN 0090-5364. URL https://www.jstor.org/stable/2242682. Publisher: Institute of Mathematical Statistics.

P. C. O'Brien and T. R. Fleming. A Multiple Testing Procedure for Clinical Trials. *Biometrics*, 35(3):549–556, 1979. ISSN 0006-341X. doi: 10.2307/2530245. URL https://www.jstor.org/stable/2530245. Publisher: [Wiley, International Biometric Society].

Office of Scientific Research and Development. Summary Technical Report of the Applied Mathematics Panel, NDRC. Volume 3. Probability and Statistical Studies in Warfare Analysis. Part 1: Bombing Studies. Part 2: Miscellaneous Studies. Technical report, Washington DC, 1946. URL https://apps.dtic.mil/sti/citations/ADB809137. Section: Technical Reports.

A. L. Paterson. *Amenability*. Number 29. American Mathematical Soc., 2000.

R. Peto. Discussion of: Regression models and life tables, by DR Cox. *Journal of the Royal Statistical Society, Series B*, 26:205–207, 1972.

R. Peto and J. Peto. Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2):185–207, 1972. ISSN 0035-9238. doi: 10.2307/2344317. URL https://www.jstor.org/stable/2344317. Publisher: [Royal Statistical Society, Wiley].

S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, Aug. 1977. ISSN 0006-3444. doi: 10.1093/biomet/64.2.191. URL https://doi.org/10.1093/biomet/64.2.191.

S. J. Pocock. Current controversies in data monitoring for clinical trials. *Clinical Trials*, 3(6):513–521, Dec. 2006. ISSN 1740-7745. doi: 10.1177/1740774506073467. URL https://doi.org/10.1177/1740774506073467. Publisher: SAGE Publications.

M. A. Proschan, K. K. G. Lan, and J. T. Wittes. *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer Science & Business Media, Dec. 2006. ISBN 978-0-387-44970-8. Google-Books-ID: BCu8c8NwXxcC.

M. F. Pérez-Ortiz and W. M. Koolen. Luckiness in Multiscale Online Learning. *Advances in Neural Information Processing Systems*, 35:25160–25170, Dec. 2022. URL `https://papers.nips.cc/paper_files/paper/2022/hash/a0d2345b43e66fa946155c98899dc03b-Abstract-Conference.html`.

M. F. Pérez-Ortiz, T. Lardy, R. de Heide, and P. Grünwald. E-Statistics, Group Invariance and Anytime Valid Testing, Aug. 2022. URL `http://arxiv.org/abs/2208.07610`. arXiv:2208.07610 [math, stat], under submission.

A. Rakhlin and K. Sridharan. On Equivalence of Martingale Tail Bounds and Deterministic Regret Inequalities. In *Proceedings of the 2017 Conference on Learning Theory*, pages 1704–1722. PMLR, June 2017. URL `https://proceedings.mlr.press/v65/rakhlin17a.html`. ISSN: 2640-3498.

S. Rakhlin and K. Sridharan. Optimization, Learning, and Games with Predictable Sequences. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL `https://proceedings.neurips.cc/paper_files/paper/2013/hash/f0dd4a99fba6075a9494772b58f95280-Abstract.html`.

A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167 [math, stat]*, Sept. 2020. URL `http://arxiv.org/abs/2009.03167`. arXiv: 2009.03167.

A. Ramdas, P. D. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference, Oct. 2022a. URL `http://arxiv.org/abs/2210.01948`. arXiv:2210.01948 [cs, math, stat].

A. Ramdas, J. Ruf, M. Larsson, and W. M. Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, Feb. 2022b. ISSN 0888-613X. doi: 10.1016/j.ijar.2021.06.017. URL `https://www.sciencedirect.com/science/article/pii/S0888613X21000980`.

H. Reiter and J. D. Stegeman. *Classical Harmonic Analysis and Locally Compact Groups*. London Mathematical Society Monographs. Oxford University Press, Oxford, New York, second edition edition, Nov. 2000. ISBN 978-0-19-851189-2.

Z. Ren and R. F. Barber. Derandomized knockoffs: leveraging e-values for false discovery rate control, Feb. 2023. URL `http://arxiv.org/abs/2205.15461`. arXiv:2205.15461 [stat].

H. Robbins and D. Siegmund. Boundary Crossing Probabilities for the Wiener Process and Sample Sums. *The Annals of Mathematical Statistics*, 41(5):1410–1429, Oct. 1970. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177696787. URL `https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-41/issue-5/Boundary-Crossing-Probabilities-for-the-Wiener-Process-and-`

Sample-Sums/10.1214/aoms/1177696787.full. Publisher: Institute of Mathematical Statistics.

J. N. Rouder, P. L. Speckman, D. Sun, R. D. Morey, and G. Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237, Apr. 2009. ISSN 1531-5320. doi: 10.3758/PBR.16.2.225. URL https://doi.org/10.3758/PBR.16.2.225.

S. N. Roy and R. E. Bargmann. Tests of Multiple Independence and the Associated Confidence Bounds. *The Annals of Mathematical Statistics*, 29(2):491–503, June 1958. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177706624. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-29/issue-2/Tests-of-Multiple-Independence-and-the-Associated-Confidence-Bounds/10.1214/aoms/1177706624.full. Publisher: Institute of Mathematical Statistics.

R. Royall. *Statistical Evidence: A Likelihood Paradigm*, volume 71. Chapman and Hall/CRC, New York, 1 edition, 1997. ISBN 978-1-03-247800-5. URL https://www.routledge.com/Statistical-Evidence-A-Likelihood-Paradigm/Royall/p/book/9781032478005.

S. Rushton. On a Two-Sided Sequential t-Test. *Biometrika*, 39(3/4):302–308, 1952. ISSN 0006-3444. doi: 10.2307/2334026. URL https://www.jstor.org/stable/2334026. Publisher: [Oxford University Press, Biometrika Trust].

I. R. Savage. Contributions to the Theory of Rank Order Statistics-the Two-Sample Case. *The Annals of Mathematical Statistics*, 27(3): 590–615, Sept. 1956. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177728170. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-27/issue-3/Contributions-to-the-Theory-of-Rank-Order-Statistics-the-Two/10.1214/aoms/1177728170.full. Publisher: Institute of Mathematical Statistics.

D. Schoenfeld. The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions. *Biometrika*, 68(1):316–319, 1981. ISSN 0006-3444. doi: 10.2307/2335833. URL https://www.jstor.org/stable/2335833. Publisher: [Oxford University Press, Biometrika Trust].

M. Seeger. PAC-Bayesian Generalisation Error Bounds for Gaussian Process Classification. *Journal of Machine Learning Research*, 3(Oct):233–269, 2002. ISSN ISSN 1533-7928. URL https://www.jmlr.org/papers/v3/seeger02a.html.

Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer. PAC-Bayesian Inequalities for Martingales. *IEEE Transactions on Information Theory*, 58(12): 7086–7093, Dec. 2012. ISSN 1557-9654. doi: 10.1109/TIT.2012.2211334. Conference Name: IEEE Transactions on Information Theory.

T. Sellke and D. Siegmund. Sequential analysis of the proportional hazards model. *Biometrika*, 70(2):315–326, 1983. Publisher: Oxford University Press.

G. Shafer. The Language of Betting as a Strategy for Statistical and Scientific Communication. *arXiv:1903.06991 [math, stat]*, Oct. 2019. URL `http://arxiv.org/abs/1903.06991`. arXiv: 1903.06991.

G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A*, 184(2):407–431, 2021. Publisher: Royal Statistical Society.

G. Shafer and V. Vovk. *Probability and Finance: It's Only a Game!* New York, 2001. ISBN 978-0-471-40226-8.

G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test Martingales, Bayes Factors and p-Values. *Statistical Science*, 26(1):84–101, Feb. 2011. ISSN 0883-4237, 2168-8745. doi: 10.1214/10-STS347. URL `https://projecteuclid.org/euclid.ss/1307626567`. Publisher: Institute of Mathematical Statistics.

O. V. Shalaevskii. Minimax Character of Hotelling's T2 Test. I. In V. M. Kalinin and O. V. Shalaevskii, editors, *Investigations in Classical Problems of Probability Theory and Mathematical Statistics: Part I*, pages 74–101. Springer US, Boston, MA, 1971. ISBN 978-1-4684-8211-9. doi: 10.1007/978-1-4684-8211-9_2. URL `https://doi.org/10.1007/978-1-4684-8211-9_2`.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, Cambridge, 2014. ISBN 978-1-107-05713-5. doi: 10.1017/CBO9781107298019. URL `https://www.cambridge.org/core/books/understanding-machine-learning/3059695661405D25673058E43C8BE2A6`.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948. ISSN 1538-7305. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x.

D. Siegmund. *Sequential Analysis.* Springer Series in Statistics. Springer, New York, NY, 1985. ISBN 978-1-4419-3075-0 978-1-4757-1862-1. doi: 10.1007/978-1-4757-1862-1. URL `http://link.springer.com/10.1007/978-1-4757-1862-1`.

A. V. Skorokhod. *Random Processes with Independent Increments*, volume 47 of *Mathematics and its Applications.* Springer Dordrecht, 1 edition, 2012. ISBN 978-94-010-5650-2. URL `https://link.springer.com/book/9780792303404`. Originally published in Russian.

E. V. Slud. Sequential Linear Rank Tests for Two-Sample Censored Survival Data. *Annals of Statistics*, 12(2):551–571, June 1984. ISSN 0090-5364, 2168-8966. doi:

10.1214/aos/1176346505. URL https://projecteuclid.org/euclid.aos/1176346505. Publisher: Institute of Mathematical Statistics.

E. V. Slud. Partial Likelihood for Continuous-Time Stochastic Processes. *Scandinavian Journal of Statistics*, 19(2):97–109, 1992. ISSN 0303-6898. URL https://www.jstor.org/stable/4616231. Publisher: [Board of the Foundation of the Scandinavian Journal of Statistics, Wiley].

P. Subbaiah and G. S. Mudholkar. A Comparison of Two Tests for the Significance of a Mean Vector. *Journal of the American Statistical Association*, 73(362):414–418, June 1978. ISSN 0162-1459. doi: 10.1080/01621459.1978.10481592. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1978.10481592. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1978.10481592.

D. Sun and J. O. Berger. Objective Bayesian analysis for the multivariate normal model. *Bayesian Statistics*, 8:525–562, 2007.

V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire. Fast Convergence of Regularized Learning in Games. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html.

M. Talagrand. *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems.* Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics. Springer-Verlag, Berlin Heidelberg, 2014. ISBN 978-3-642-54074-5. URL //www.springer.com/la/book/9783642540745.

A. Tartakovsky, I. Nikiforov, and M. Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection.* Chapman and Hall/CRC, New York, Aug. 2014. ISBN 978-0-429-15234-4. doi: 10.1201/b17279.

J. B. Taylor. An Interview with Milton Friedman. *Macroeconomic Dynamics*, 5(1):101–131, Feb. 2001. ISSN 1469-8056, 1365-1005. doi: 10.1017/S1365100501018053. URL https://www.cambridge.org/core/journals/macroeconomic-dynamics/article/abs/an-interview-with-milton-friedman/4F9B7590919E146F06B80BFAD022080B. Publisher: Cambridge University Press.

J. ter Schure and P. Grünwald. ALL-IN meta-analysis: breathing life into living systematic reviews. Technical Report 11:549, F1000Research, May 2022. URL https://f1000research.com/articles/11-549. Type: article.

J. ter Schure and P. D. Grünwald. Accumulation Bias in meta-analysis: the need to consider *time* in error control. Technical Report 8:962, F1000Research, June 2019. URL https://f1000research.com/articles/8-962. Type: article.

J. ter Schure, M. F. Pérez-Ortiz, A. Ly, and P. Grünwald. The Safe Logrank Test: Error Control under Continuous Monitoring with Unlimited Horizon, July 2021. URL http://arxiv.org/abs/2011.06931. arXiv:2011.06931 [math, stat], under submission.

I. O. Tolstikhin and Y. Seldin. PAC-Bayes-Empirical-Bernstein Inequality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/a97da629b098b75c294dffdc3e463904-Abstract.html.

A. A. Tsiatis. A Large Sample Study of Cox's Regression Model. *The Annals of Statistics*, 9(1):93–108, Jan. 1981. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345335. URL https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-1/A-Large-Sample-Study-of-Coxs-Regression-Model/10.1214/aos/1176345335.full. Publisher: Institute of Mathematical Statistics.

A. A. Tsiatis. *Group sequential methods for survival analysis with staggered entry.* Institute of Mathematical Statistics, 1982. ISBN 978-0-940600-02-7. doi: 10.1214/lnms/1215464854. URL https://projecteuclid.org/euclid.lnms/1215464854. Pages: 257-268 Publication Title: Survival Analysis.

J. W. Tukey. Review of Sequential Analysis of Statistical Data: Applications. *The Annals of Mathematical Statistics*, 18(1):142–144, 1947. ISSN 0003-4851. URL https://www.jstor.org/stable/2236115. Publisher: Institute of Mathematical Statistics.

R. Turner, A. Ly, and P. Grünwald. Two-Sample Tests that are Safe under Optional Stopping, with an Application to Contingency Tables. *arXiv:2106.02693 [cs, math, stat]*, Oct. 2021. URL http://arxiv.org/abs/2106.02693. arXiv: 2106.02693.

R. Turner, A. Ly, M. F. Perez-Ortiz, J. ter Schure, and P. D. Grunwald. safestats: Safe Anytime-Valid Inference, Nov. 2022. URL https://CRAN.R-project.org/package=safestats.

S. Urban, R. Sreenivasan, and V. Kannan. It's All A/Bout Testing, Nov. 2021. URL https://netflixtechblog.com/its-all-a-bout-testing-the-netflix-experimentation-platform-4e1ca458c15.

J. V. Uspensky. Introduction to mathematical probability. 1937.

S. van de Geer and J. Lederer. The Bernstein–Orlicz norm and deviation inequalities. *Probability Theory and Related Fields*, 157(1):225–250, Oct. 2013. ISSN 1432-2064. doi: 10.1007/s00440-012-0455-y. URL https://doi.org/10.1007/s00440-012-0455-y.

A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 978-0-387-94640-5. URL //www.springer.com/la/book/9780387946405.

T. van Erven. PAC-Bayes Mini-tutorial: A Continuous Union Bound. *arXiv:1405.1580 [stat]*, May 2014. URL http://arxiv.org/abs/1405.1580. arXiv: 1405.1580.

T. van Erven and P. Harremoes. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014. ISSN 0018-9448. doi: 10.1109/TIT.2014.2320500.

T. Van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16: 1793–1861, 2015.

V. Vapnik. *Statistical learning theory. 1998*, volume 3. Wiley, New York, 1998.

J. Ville. *Étude Critique de la Notion de Collectif.* Number 218 in Thèses de l'entre-deux-guerres. Paris, 1939. URL http://archive.numdam.org/item/THESE_1939__218__1_0/.

V. Vovk. A Game of Prediction with Expert Advice. *Journal of Computer and System Sciences*, 56(2):153–173, Apr. 1998. ISSN 0022-0000. doi: 10.1006/jcss.1997.1556. URL http://www.sciencedirect.com/science/article/pii/S0022000097915567.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, June 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS2020. URL https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-3/E-values-Calibration-combination-and-applications/10.1214/20-AOS2020.full. Publisher: Institute of Mathematical Statistics.

V. G. Vovk. Aggregating Strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory*, COLT '90, pages 371–386, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-146-8. URL http://dl.acm.org/citation.cfm?id=92571.92672.

E.-J. Wagenmakers. A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, 14(5):779–804, Oct. 2007. ISSN 1531-5320. doi: 10.3758/BF03194105. URL https://doi.org/10.3758/BF03194105.

A. Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, June 1945. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177731118. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-16/issue-2/Sequential-Tests-of-Statistical-Hypotheses/10.1214/aoms/1177731118.full. Publisher: Institute of Mathematical Statistics.

A. Wald. *Sequential Analysis.* 1947. URL http://archive.org/details/in.ernet.dli.2015.510091.

A. Wald and J. Wolfowitz. Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, 19(3):326–339, Sept. 1948. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177730197. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-19/issue-3/Optimum-Character-of-the-Sequential-Probability-Ratio-Test/10.1214/aoms/1177730197.full. Publisher: Institute of Mathematical Statistics.

W. A. Wallis. The Statistical Research Group, 1942-1945. *Journal of the American Statistical Association*, 75(370):320–330, 1980. ISSN 0162-1459. doi: 10.2307/2287451. URL https://www.jstor.org/stable/2287451. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

Q. Wang, R. Wang, and J. Ziegel. E-backtesting, Sept. 2022. URL https://papers.ssrn.com/abstract=4206997.

R. Wang and A. Ramdas. False discovery rate control with e-values. *arXiv:2009.02824 [math, stat]*, Nov. 2020. URL http://arxiv.org/abs/2009.02824. arXiv: 2009.02824.

R. Wang and A. Ramdas. False Discovery Rate Control with E-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, July 2022. ISSN 1369-7412. doi: 10.1111/rssb.12489. URL https://doi.org/10.1111/rssb.12489.

R. A. Wijsman. Cross-sections of orbits and their application to densities of maximal invariants. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 5.1:389–401, Jan. 1967. URL https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Cross-sections-of-orbits-and-their-application-to-densities-of/bsmsp/1200512999. Publisher: University of California Press.

R. A. Wijsman. Proper Action in Steps, with Application to Density Ratios of Maximal Invariants. *The Annals of Statistics*, 13(1):395–402, Mar. 1985. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176346600. URL https://projecteuclid.org/journals/annals-of-statistics/volume-13/issue-1/Proper-Action-in-Steps-with-Application-to-Density-Ratios-of/10.1214/aos/1176346600.full. Publisher: Institute of Mathematical Statistics.

J. Wolfowitz. Abraham Wald, 1902-1950. *The Annals of Mathematical Statistics*, 23(1):1–13, 1952. ISSN 0003-4851. URL https://www.jstor.org/stable/2236396. Publisher: Institute of Mathematical Statistics.

J. Wu and X. Xiong. Group Sequential Survival Trial Design and Monitoring Using the Log-Rank Test. *Statistics in Biopharmaceutical Research*, 9(1):35–43, Jan. 2017. ISSN null. doi: 10.1080/19466315.2016.1189355. URL `https://doi.org/10.1080/19466315.2016.1189355`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/19466315.2016.1189355.

Y.-S. Wu and Y. Seldin. Split-kl and PAC-Bayes-split-kl Inequalities for Ternary Random Variables. *Advances in Neural Information Processing Systems*, 35:11369–11381, Dec. 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/49ffa271264808cf500ea528ed8ec9b3-Abstract-Conference.html`.

T. Zhang. From Epsilon-Entropy to KL-Entropy: Analysis of Minimum Information Complexity Density Estimation. *The Annals of Statistics*, 34(5):2180–2210, 2006a. ISSN 0090-5364. URL `https://www.jstor.org/stable/25463505`. Publisher: Institute of Mathematical Statistics.

T. Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, Apr. 2006b. ISSN 0018-9448. doi: 10.1109/TIT.2005.864439.