Universiteit Leiden
The Netherlands

# Optimal test statistics for anytime-valid hypothesis tests

Lardy, T.D.

# Optimal Test Statistics for Anytime-Valid Hypothesis Tests

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 woensdag 18 juni 2025
klokke 16:00 uur

door

Tyron Darnell Lardy
geboren te Haarlem, Nederland
in 1996

**Promotor:**

Prof.dr. P.D. Grünwald (Universiteit Leiden, CWI)

**Co-promotor:**

Prof.dr. W.M. Koolen (Universiteit Twente, CWI)

**Promotiecommissie:**

Prof.dr.ir. G.L.A. Derks

Prof.dr. M. Fiocco

Prof.dr. J. Ziegel (ETH Zürich)

Prof.dr. J. Ruf (London School of Economics)

Dr. I. Waudby-Smith (UC Berkeley)

Cover design by Leroy Soesman

*When all else fails, you have to be able to go home again and have people call your name in a way that is familiar to only them.*

Hanif Abdurraqib

*For my family.*

# Preface

This dissertation marks a significant milestone in my academic journey in the field of statistics. When I first embarked on this journey, I was intrigued by the ability of statistical tools to uncover complex patterns and provide clarity in the face of uncertainty. At the same time, I became aware of all the ways in which statistics could be misused, be it intentional or through ignorance, to reach unwarranted conclusions. This matter is becoming more pressing with the increasing reliance on data analytics in modern science and business. One of the most common problems is that researchers often repeatedly analyze data and change or halt data collection based on interim results. Such practices invalidate the conclusions drawn by traditional statistical methods. In response, the field of anytime-valid inference has emerged, which comprises methods that lead to reliable statistical conclusions regardless of whether inference is made at the start, during, or at the end of data collection. This thesis contains a number of results on the construction of such methods.

The intention behind this work is not to portray anytime-valid methods as the ultimate or all-encompassing solution. Instead, the aim is to provide a candid overview of their strengths and weaknesses, along with a comparison to established methods whenever applicable. It is my hope that this dissertation will equally encourage reflection, critique, and new insights; ultimately leading to further advancement of the field of anytime-valid inference.

# Acknowledgments

This dissertation would not have seen the light of day without the guidance, support, and encouragement of many people, to whom I owe my heartfelt gratitude. To my advisor, Peter Grünwald, for sparking my interest in anytime-valid inference and, most importantly, for inspiring me to explore and grow independently. To the exceptional individuals with whom I have had the privilege of working, Peter Harremoës, Alexander Henzi, Wouter Koolen, and Muriel Felipe Pérez-Ortiz, for countless hours of problem solving on the whiteboard. To Yunda Hao, for all the games of table tennis. To Alexander Ly and Udo Böhm, for the many discussions during our reading group. To Marta Fiocco, for showing me the importance of networking and marketing. To my office mates, Daniel Gomon and Marta Spreafico, for their camaraderie and the much-needed coffee breaks. To Leroy Soesman, for the determination to see me pursue a PhD, whether I liked it or not. To Jelle Vogel, for being a great friend. To my sisters, Joslyn and Tabitha, for putting up with me ever since we were little. To my parents, Esther and Neubury, for being the best support system imaginable; words cannot express how grateful I am to you. To Annabel, for lighting up my life, always.

# Sources

This dissertation is based on the following manuscripts, to all of which the author of this dissertation contributed significantly:

**Chapter 3** is based on

Lardy, T., Grünwald, P., and Harremoës, P. (2024). Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631. © 2024 IEEE. Reproduced with permission and with minor modifications to fit the style of this dissertation.

**Chapter 4** is based on

Hao, Y., Grünwald, P., Lardy, T., Long, L., and Adams, R. (2024). E-values for k-sample tests with exponential families. *Sankhya A*, 86(1):596–636. This article is licensed under a Creative Commons Attribution 4.0 International License. For details, visit `https://creativecommons.org/licenses/by/4.0/`. Minor changes were made to the text and formatting to fit the style of this dissertation.

**Chapter 5** is based on

Grünwald, P., Lardy, T., Hao, Y., Bar-Lev, S. K., and De Jong, M. (2024). Optimal e-values for exponential families: the simple case. ArXiv preprint, arXiv:2404.19465. A revised version of this manuscript has been accepted as a contribution to the Springer festschrift "Information Theory, Probability and Statistical Learning: A Festschrift in Honor of Andrew Barron."

**Chapter 6** is based on

Grünwald, P., Henzi, A., and Lardy, T. (2023). Anytime valid tests of conditional independence under model-X. *Journal of the American Statistical Association*, 119(546):1554–1565. Reproduced with permission and with minor modifications to fit the style of this dissertation.

**Chapter 7** is based on

Lardy, T. and Pérez-Ortiz, M. F. (2024). Anytime-valid tests of group invariance through conformal prediction. ArXiv preprint arXiv:2401.15461

**Chapter 8** is based on

Pérez-Ortiz, M. F., Lardy, T., De Heide, R., and Grünwald, P. D. (2024). E-statistics, group invariance and anytime-valid testing. *The Annals of Statistics*, 52(4):1410–1432. This manuscript was originally published by the Institute of Mathematical Statistics. Reproduced with permission and with minor modifications to fit the style of this dissertation.

# Contents

**1 Introduction** **1**
1.1 Hypothesis Testing . . . 2
1.2 Optional Stopping in Academia . . . 4
1.3 Sequential Methods . . . 5
1.4 Anytime-Valid Inference . . . 6
1.5 Contributions of This Thesis . . . 7

**2 Preliminaries** **11**
2.1 The E-Family . . . 12
2.2 Optimality . . . 15
2.3 From Experiment to Meta-Analysis . . . 18

**3 On the Optimality of E-statistics** **19**
3.1 Introduction . . . 20
3.2 Background . . . 22
3.3 The Reverse Information Projection . . . 24
3.4 Optimal E-Statistics . . . 32
3.5 Summary and Future Work . . . 38

**4 k-Sample Tests With Exponential Families** **41**
4.1 Introduction . . . 42
4.2 The Four Types of E-Variables . . . 49
4.3 Growth Rate Comparison of Our E-Variables . . . 53
4.4 Growth Rate Comparison for Specific Exponential Families . . . 55
4.5 Simulations to Approximate the RIPr . . . 58
4.6 Conclusion and Future Work . . . 60

**5 Simple E-Variables for Exponential Families** **63**
5.1 Introduction . . . 64
5.2 Existence of Simple Local E-Variables . . . 71
5.3 Existence of Simple Global E-Variables . . . 73
5.4 Examples . . . 75
5.5 Proof of Theorem 5.3 . . . 88
5.6 Conclusion and Future Work . . . 90

**6 Tests of Conditional Independence Under Model-X** **93**
6.1 Introduction . . . 94
6.2 Background . . . 96
6.3 Conditional Independence Testing With E-Statistics . . . 100
6.4 Simulations . . . 107
6.5 Data Application . . . 113
6.6 Discussion, Related and Future Work . . . 114

**7 Tests of Group Invariance** **117**
7.1 Introduction . . . 118
7.2 Problem Statement . . . 120
7.3 Sequential Group Actions are Online Compression Models . . . 123
7.4 Testing Group Invariance With Conformal Martingales . . . 126
7.5 Applications and Extension . . . 131
7.6 Discussion . . . 136

**8 Testing With Group-Invariant Models** **139**
8.1 Introduction . . . 140
8.2 Preparation for the Main Results . . . 143
8.3 Main Results . . . 151
8.4 Testing Multivariate Normal Distributions . . . 157
8.5 Discussion and Future Work . . . 160
8.6 Proofs . . . 162

**9 Discussion** **171**
9.1 Publish or perish . . . 171
9.2 How Much Freedom Is Too Much? . . . 172

**Summary** **175**

**Samenvatting** **177**

**Bibliography** **179**

**Appendices** **193**

**A Appendix to Chapter 3** **195**

**B Appendix to Chapter 4** **217**

**C Appendix to Chapter 5** **239**

**D Appendix to Chapter 6** **241**

**E Appendix to Chapter 7** **253**

**F Appendix to Chapter 8** **261**

**Curriculum Vitae** **273**

# 1 | Introduction

> Statistical procedure and experimental design are only two different aspects of the same whole....
>
> *R. A. Fisher, The Design of Experiments*

At the core of science is a systematic process of inquiry, where observations lead to hypotheses, experiments provide evidence in favor of or against these hypotheses, and this in turn shapes our understanding of the world around us. To illustrate the complexities of this process, imagine that you frequently play table tennis with a friend. After some time, you get the feeling that you are not equally matched and decide to test this by playing a game against each other. A choice presents itself: one approach is to play until a fixed number of rallies has been played, another is to follow the official rules (first to 11, minimum of two-point difference), and a third option is to play until, for whatever reason, either of you has had enough. Whichever option is chosen, the outcome of the match will offer evidence in favor of or against the hypothesis that you are (not) equally matched. Once again, this requires deliberation. Would you consider winning by a single point to be evidence in favor of your hypothesis? Or should the difference in points be more extreme before you believe that you are not equally matched? In this analogy, the match represents an experiment and the questions about its outcome mirror the subsequent statistical analysis. The first match type listed above corresponds to experiments in which the number of data points to be collected is fixed in advance. The second option relates to experiments where the data collection plan is decided in advance, but the total number of data points collected is variable (you could win 11-0, or a closer game might end 12-10). Finally, the third variant corresponds to experiments without a predetermined sampling plan, that is, where experimentation may be stopped at any point in time. The analysis of data collected from this type of experiment lies at the heart of this thesis.

The remainder of this introduction serves to further outline this problem and contextualize the results that are given in the subsequent chapters. The structure is as follows: First, we introduce the general problem of statistical hypothesis testing, leading up to an explanation showing that experiments with a variable sample size require different statistical methodology than those with a fixed sample size. Next, we discuss one of the repercussions of not taking this into account properly. Following this (in Section 1.3), a review is provided of the first statistical methods that were designed to be used with a fixed data collection plan, but without a fixed sample size. Then, an introduction to the theory of hypothesis testing for experiments without any sampling plan is given. Finally, we give an overview of the content of this thesis.

## 1.1 Hypothesis Testing

A statistical hypothesis test is a data-driven method that is used to decide whether a hypothesis of interest, called the null hypothesis, can be falsified. The way in which hypothesis tests are used is comparable to a proof by contradiction. For example, suppose that you want to verify the claim that you and your friend are not equally good at table tennis. You would then take the null hypothesis to be the assumption that this claim is false, that is, that you are equally good. After collecting data (playing a match), a hypothesis test is used to determine whether the null hypothesis is still plausible in light of the results. If the outcome of the test is that the null hypothesis can be falsified, then the data are so incompatible with the null hypothesis that you should no longer believe it to be correct. This would happen if, for example, one of you wins by a landslide. You can then proceed as if your original claim were true, that is, as if you are not equally good at table tennis. This is commonly referred to as "rejecting" the null hypothesis. The other possibility is that the data do not give sufficient reason to doubt the null hypothesis (for example, if the match is very close), in which case you fail to reject the null hypothesis. This does not mean that the null hypothesis is true: absence of evidence is not evidence of absence. In this case, you can therefore not be sure about your original claim.

Of course, the conclusion drawn by a hypothesis test is useful only if it does not lead to too many mistakes. To quantify how often a hypothesis test is wrong, the null hypothesis is translated to a probabilistic model, which we denote by $\mathcal{H}_0$. This translation allows probabilistic statements about the data to be made. For example, in the table tennis example above, the null hypothesis postulates that you and your friend are equally good. One way of modeling this is by saying that both of you are

equally likely to win each point. If this model were actually true, then the probability that the difference in points is at least four in a match with a total of five points is equal to 0.375, or 37.5%. This follows because there are five ways in which the score could end up being 4-1 and only one way in which the score could be 5-0, and similarly for 1-4 and 0-5. Under $\mathcal{H}_0$, all of these scores have associated probability $\frac{1}{2}^5$, so the total probability of this happening is $12 \times \frac{1}{2}^5 = 0.375$.

Such statements allow us to only consider hypothesis tests such that, if $\mathcal{H}_0$ were actually true, the probability of seeing data for which the test would reject $\mathcal{H}_0$ is not too high. What qualifies as too high should be determined by the experimenters beforehand and is referred to as the significance level $\alpha$; common values include $0.1, 0.05$ and $0.01$. False rejections of the null hypothesis are also called type-I errors, so the above can be summarized as follows: the probability of making a type-I error should be smaller than $\alpha$. If this is true for a certain hypothesis test, we say that the test is valid at the significance level $\alpha$. Returning to the example, if a match with five rallies is played, one might define a hypothesis test by rejecting the null hypothesis whenever the difference in points is four or more. This test has a type-I error probability of 0.375 according to the above reasoning. That is, even if the null hypothesis were true, this test would reject it with a probability of 37.5%. Therefore, it is not a valid hypothesis test at any significance level $\alpha < 0.375$. To obtain a smaller type-I error probability, one can consider a test that only rejects the null hypothesis if the difference in points is more extreme than four. In fact, it is a standard result in statistics that, for a fixed sample size of $n$, rejecting when the difference in points exceeds $1.96 \times \sqrt{n}$ gives a type-I error probability of approximately 0.05. For $n = 5$, this amounts to rejecting the null hypothesis whenever the difference in points is greater than $1.96 \times \sqrt{5} \approx 4.4$, which only happens when the score is 5-0 in favor of either player.

The type-I error probability of this test is calculated under the assumption that the sample size is fixed beforehand and that the data are only analyzed once. Such methods will be referred to as fixed-sample methods. Now, suppose that the null hypothesis is true but that you have a certain incentive to reject it nonetheless. If the sample size $n$ has not been fixed beforehand, you could adopt a technique known as optional stopping: continue to play—increase the sample size—until the difference in points first exceeds $1.96 \times \sqrt{n}$. It follows from a standard result in probability theory that this will surely happen infinitely often. Hence, you are certain to come to a point at which the difference in points exceeds $1.96 \times \sqrt{n}$. By stopping at such a point, the hypothesis test will suggest to reject the null hypothesis, even though it is true. The intuition is that, even if you and your friend are equally matched, one of you will at

some point reach a long enough lucky streak to be able to reject the null hypothesis. This immediately disqualifies the use of fixed-sample methods whenever $n$ is not fixed in advance. The problem is neither that the method for data collection is invalid nor that the hypothesis test is incorrect. They are simply not compatible. The test is designed for a fixed sample size, while the experiment follows a sequential sampling plan. The latter would be valid if the rejection rule were adjusted accordingly. That is, serial data require different statistical methods from data collected in a batch.

## 1.2 Optional Stopping in Academia

One of the first documented cases of optional stopping was when researchers saw it unfold in the study of extrasensory perception, studies to prove that the sixth sense exists (see e.g Greenwood, 1938; Greenwood and Greville, 1939; Feller, 1940). It has since been the subject of continuous debate among statisticians and has, more critically, been referred to as "sampling to a foregone conclusion" (Anscombe, 1954). The problem is that researchers, either aware or unaware, might engage in optional stopping and still report that they have conducted a fixed-sample hypothesis test at some significance level $\alpha$, while the true type-I error probability can be much higher. This results in frequent controversies stemming from statistical investigations in various fields. For example, an article published in 2018 in a *Nature* journal showed that activation of certain neurons in animals suppresses REM sleep (Weber et al., 2018). This conclusion was based on hypothesis tests designed for a fixed sample size, yet the authors state that "we continuously increased the number of animals until statistical significance was reached to support our conclusions." This sparked some debate on social media among statisticians (Barnett, 2018), but it only came to light because the authors were transparent regarding their methodology.

An example in which there was no such transparency was when a team of researchers supposedly showed the benefits of power posing (Carney et al., 2010). They reported that assuming a powerful position helped boost confidence, increase testosterone, and decrease cortisol. Five years later, another research group tried to replicate their findings with a larger sample size, and while they indeed found that power posing increased subjective feelings of power, they found no significant effect on hormonal levels (Ranehill et al., 2015). In light of this and other failed replication attempts (see e.g. Simmons and Simonsohn, 2017), the first author of the original article eventually disassociated herself from the claims in the article and mentioned optional stopping as one of her concerns: "We ran subjects in chunks and checked the effect along the way.

It was something like 25 subjects run, then 10, then 7, then 5" (Carney, 2016, fact 5). Unfortunately, these are not isolated incidents. In 2012, 56% of participants in a survey of more than 2000 psychologists admitted to "deciding whether to collect more data after looking to see whether the results were significant" (John et al., 2012).

This compromises the reliability of results derived from statistical analyses and, in turn, might cause a distrust of statistics in general. Feller (1940) eloquently puts it as follows: "...statistics makes claim to mathematical rigor, and still its practical applications are often disputed or rejected as absurd. I fear it may sometimes produce a feeling that mathematics is, after all, 'one of those rational and scientific paths, which is nothing more than a narrow, short, and dirty dead end, at the end of which one hits their nose ingloriously'" (second part translated from French). Rather than taking such a bleak point of view, statisticians have recently emphasized the need to develop tools that can accommodate optional stopping. That is, to focus on the development of methods that can handle serial data collection without losing type-I error control (see e.g. Ramdas et al., 2023).

## 1.3 Sequential Methods

The first developments of methods that were designed for experiments with a variable sample size were based on a simple observation: There are situations where one can reach the same conclusions as when using fixed-sample methods but with a strictly smaller sample size. For example, for a fixed sample size of $n = 20$, the table tennis hypothesis test described above would reject the null hypothesis if the difference in points is greater than or equal to 9. Now suppose that after 15 points the score is 15-0. At this point, the game might as well be stopped, since no matter how it plays out, the null hypothesis will always be rejected. That is, the sample size can be strictly reduced while making the exact same decisions (a similar realization is already present in early work on lot inspection by Dodge and Romig, 1929, p. 626).

This insight led to the idea that it might be beneficial to directly design hypothesis tests that incorporate serial data collection. The culmination was the development of a widely applicable sequential method: the sequential probability ratio test (SPRT) (Wald, 1947). The SPRT works by updating a single number, the probability ratio, that represents the evidence in favor of or against the null hypothesis after each collected data point. Based on this number, the SPRT prescribes whether to collect another data point or to stop and draw conclusions about the null hypothesis. In this way, the total number of data points that will be collected in an experiment is variable

and depends on the data that are observed. However, if one were to repeatedly follow this procedure, the average sample size that is needed to reach a conclusion is smaller than the average sample size that would be needed to draw similar conclusions using any other method (fixed-sample and sequential alike). That is, the SPRT achieves the optimal expected sample size.

This was a highly desirable feat because, in the intended applications, data collection was extremely costly or even destructive. Indeed, the SPRT was developed in the context of World War II and, as such, its applications included testing whether ordnance (artillery, ammunition, explosives, etc.) was faulty or not (see also Wallis, 1980). The caveat to this cost efficiency is that the probability of falsely rejecting the null hypothesis is only bounded by the prescribed significance level $\alpha$ if the SPRT's sampling plan is followed to a tee. If one deviates from it, then the type-I error probability might still be higher, so the SPRT does not fix the problems with optional stopping that were outlined in the previous section.

## 1.4 Anytime-Valid Inference

The problem of optional stopping was specifically addressed in work by, among others, Robbins (1952, 1970), Darling and Robbins (1968), and Lai (1976). They developed hypothesis testing methods for which the type-I error probability remained below the prespecified level of significance regardless of the data sampling rule that was employed. Their methods are now referred to as anytime-valid. Similarly to the SPRT, anytime-valid methods are based on a test statistic that can be thought of as a measure of evidence against the null hypothesis at a certain time. This test statistic is updated after observing each data point. If the total evidence is large enough at some point, the null hypothesis is rejected. The key difference from the SPRT is that anytime-valid methods do not prescribe whether or not to collect more data. Instead, this choice is left to the experimenters. In particular, they might choose to keep collecting data until the test statistic gives enough evidence to reject the null hypothesis. The probability that this will ever happen if the null hypothesis is actually true, is smaller than $\alpha$. As such, anytime-valid methods are unaffected by optional stopping.

For a long time, anytime-valid methods made up only a minor segment of the area of sequential analysis. A shift occurred when interest in them took off in recent years, as demonstrated by the mere existence of this thesis. The key year for this shift was 2019, when at least four breakthrough articles on the subject were made available online (Wasserman et al. (2020); Shafer (2021); Vovk and Wang (2021); Grünwald

et al. (2024); all of these articles first appeared on ArXiv in 2019). In addition to these breakthroughs, the surge of interest may partly be caused by what is known as the replication crisis, which refers to the widespread phenomenon that many scientific studies cannot be reliably replicated or reproduced. One of the many facets of this problem is that the number of scientific conclusions based on false-positive results is much higher than one might theoretically expect. As argued in Section 1.1, this could be caused by practices such as optional stopping. Therefore, anytime-valid methods are sometimes presented as a partial solution to the replication crisis. Furthermore, anytime-valid methods might be more appealing now than they used to be because technological advances have transformed the way data is collected and stored (Cukier and Mayer-Schoenberger, 2013; Sagiroglu and Sinanc, 2013). Datasets used to be collected manually and stored on paper, film, or other analogue media with careful thought as to what was stored and what was not. Today, the standard in many fields—finance, commerce, and many more—has become storing data digitally, and as much of it as possible. Think of stock prices, the amount of time customers spend at a certain webshop, or atmospheric conditions at various points in time. These are all examples of data that are inherently serial, and digital records allow them to be accessed easily and at all times. Anytime-valid methods enable researchers to analyze these data in real time and draw conclusions accordingly without breaking the type-I error control, that is, without making too many mistakes.

## 1.5 Contributions of This Thesis

In the discussion of hypothesis testing so far, only one type of error has been considered: false rejection of the null hypothesis. If this were the only basis on which hypothesis tests are evaluated, then one should be content with never rejecting the null hypothesis, and the experiment might as well not have been performed. Therefore, to determine whether testing methods are actually useful, it is custom to consider a second type of mistake: failure to reject the null hypothesis when it is actually false. This is commonly referred to as a type-II error. For a fixed significance level $\alpha$, the typical approach for fixed-sample methods is to find the hypothesis test that minimizes the probability of making a type-II error, among all methods with a type-I error probability below $\alpha$. Equivalently, one can consider maximizing one minus the probability of making a type-II error, which is referred to as the power of the test. Which test has the most power will depend on the specific hypotheses under consideration and should be determined on a case-by-case basis.

For anytime-valid tests, the concept of power is ambiguous. Indeed, the advantage of anytime-valid tests lies in the fact that they can be used with an unknown data sampling rule. One might keep collecting data forever, in which case the null hypothesis will almost certainly be rejected if it is false, or one might stop after a single data point, making it highly unlikely to be rejected. Due to this ambiguity, it is customary to consider optimality criteria different from power for anytime-valid methods. To this end, Koolen and Grünwald (2022); Grünwald et al. (2024) propose to consider anytime-valid tests that are based on log-optimal test statistics. A formal definition of this concept, as well as a rigorous introduction to the theory of anytime-valid testing, is given in Chapter 2. As is the case for power, the criterion of log optimality gives an abstract notion of how optimal tests can be found. In practice, actually finding them is not a straightforward exercise. This thesis is concerned with the criterion of log-optimality and finding log-optimal test statistics for a variety of hypotheses. We now give a brief overview of all the chapters.

## Log Optimality

Grünwald et al. (2024) show that, under certain conditions, there exists a log-optimal test statistic that takes the form of a likelihood ratio—an object that is well studied in the classical literature on testing. However, when these conditions are not met, the log-optimality criterion cannot fully differentiate between different test statistics. That is, there are settings for which a wide range of test statistics all seem equally good when judged by this criterion. One can even construct examples where a certain test statistic always provides more evidence against the null hypothesis than another, yet both statistics seem equally good according to the log-optimality criterion. Chapter 3 discusses a way to redefine the log-optimality criterion in such situations to avoid this. Optimal test statistics are shown to still take on the form of likelihood ratios.

## Exponential Families

Many common statistical models, such as the normal, Bernoulli, and Poisson model, are instances of exponential families. The latter are collections of probabilistic models with a specific form that offers great mathematical convenience for a variety of objectives. As such, exponential families are used to model data by researchers in a wide range of scientific fields for many different applications.

An example is for $k$-sample tests, where $k$ groups of data (or samples) are observed and the goal is to test whether they have the same distribution, under the assumption

that this distribution comes from a certain exponential family. In Chapter 4, three different test statistics for this problem are studied. The first is the log-optimal test statistic for a specific alternative, which, as it turns out, cannot always be computed efficiently. The second test statistic is generally suboptimal but can always be easily evaluated. Finally, the third test statistic is designed to be robust against misspecification of the exponential family. That is, it is still valid if the data have the same distribution in all groups but this distribution is not a member of the exponential family. For small effects, that is, if the distributions of the different streams under the alternative are not too different, it is shown that all three statistics give a surprisingly similar amount of evidence.

Although the third statistic mentioned is robust against misspecification of the exponential family, the first two are not. This means that if those statistics are used but the model is wrong, then the type-I error probability is not guaranteed to be below the desired significance level. Similar reliability issues can occur with many statistical tools when the underlying models are misspecified. It is therefore important to test whether the exponential family is well specified, that is, whether the data could indeed have been generated by a specific exponential family. This problem is studied in Chapter 5. Conditions are given under which log-optimal test statistics are easy to compute. We furthermore discuss a variety of exponential families for which these conditions hold, so that anytime-valid tests can be constructed to test whether they are well specified.

## Model-X

An important task in many branches of science is to detect whether there is an association, or dependence, between a response and an explanatory variable. Consider, for example, testing whether a certain medication (explanatory variable) has an impact on a patient's health (response). Often, there are also other variables that could potentially impact the response (e.g. age) and therefore need to be controlled for. These variables are called covariates. To capture this in a hypothesis testing setting, one needs to construct a probabilistic model for the data. The difficulty with that is that there is often little prior information about what the explanatory variable and response will behave like. However, in specific cases, it is known how the explanatory variable should be modeled conditional on knowledge of the covariates. The assumption that this conditional distribution is known, is referred to as the model-X assumption. For example, in many clinical trials, the medication is administered to

patients in a randomized manner, irrespective of the covariates. Therefore, the distribution of the explanatory variable (whether the medication has been administered to a certain patient) conditional on the covariates (that patient's characteristics) is known: It is fully characterized by the randomization. Chapter 6 shows that, under the model-X assumption, it is possible to construct anytime-valid tests of independence without further assumptions on the way in which the data are generated. In particular, log-optimal test statistics for specific alternatives are derived.

### Group Invariance

Transformations of the data often do not have any meaningful effect from a statistical point of view. For example, changing the units of some measurement from kilometers to miles should generally not impact the information that can be extracted from the data. It is therefore an accepted principle that statistical inferences should exhibit certain invariance properties. In particular, if the data are not impacted by a certain transformation, then the conclusions drawn from a hypothesis test should also not be. Tests that have this property are referred to as invariant, and similarly for the corresponding test statistics. Anytime-valid tests that are invariant are the subject of Chapter 8. In particular, conditions are given under which the log-optimal test statistic is invariant. That is, the conclusion that will be reached by the optimal test will not depend on irrelevant aspects of the data.

However, it is not always entirely clear which facets of the data are actually irrelevant, that is, which transformations have a meaningful impact on the data. For example, it is a common assumption that changing the order of the data points does not have a meaningful effect on any important aspects of the data. However, there are also situations where, for example, seasonal effects are lost by reordering the data. Carrying out analyses under the assumption that certain transformations do not impact the data, while in reality they do, might cause researchers to draw wrong conclusions. It is therefore also important to be able to test whether certain transformations have an effect on the data. General methodology to construct anytime-valid tests for this purpose is discussed in Chapter 7. Furthermore, conditions are shown under which the corresponding test statistics are log optimal.

# 2 | Preliminaries

In this chapter, the theory of anytime-valid testing is formally introduced. Specifically, we will define what it means for a test to be anytime valid, discuss test statistics that are commonly used to construct anytime-valid tests, and define appropriate notions of optimality for those statistics. The concepts that are most relevant to the subsequent chapters will be reiterated there; the aim here is solely to sketch the context in which the results should be understood. We will therefore leave the measure-theoretic details implicit for the most part.

Throughout, $X_1, X_2, \dots$ denotes the data that are observed in a certain experiment. The purpose of said experiment is to test the null hypothesis $\mathcal{H}_0$. That is, the random variables $X_1, X_2, \dots$ are assumed to be independent and identically distributed (i.i.d.)[1] following an unknown probability distribution on, and taking values in, the sample space $\mathcal{X}$. The null hypothesis $\mathcal{H}_0$ is a collection of distributions on $\mathcal{X}$ and the objective is to collect evidence against the claim that one of these distributions generated $X_1, X_2, \dots$. The strength of this evidence is defined relative to an alternative hypothesis $\mathcal{H}_1$, which is a collection of plausible distributions on $\mathcal{X}$ for when $\mathcal{H}_0$ is not true. For the sake of brevity, we will use $X^n$ as shorthand for $(X_1, \dots, X_n)$ for all $n \in \mathbb{N}$, where we use the convention $\mathbb{N} = \{1, 2, \dots\}$. Furthermore, a stopping time $\tau$ is defined as a random variable that takes values in $\mathbb{N} \cup \{\infty\}$, such that the event $\{\tau = n\}$ is $\sigma(X^n)$-measurable for all $n \in \mathbb{N}$. The intuition is that $\tau$ denotes the sample size at which the experiment is stopped, and the event $\{\tau = n\}$ being $\sigma(X^n)$-measurable means that the decision to stop at time $n$ may only be based on the data available up to that time.[2] The set of all stopping times will be denoted by $\mathcal{T}$.

[1] The assumption that data are i.i.d. is made purely for ease of exposition. It is not actually required for most of the theory discussed here.

[2] More generally, there could be a filtration $\mathbb{F} = (\mathcal{F}_n)_{n \in \mathbb{N}}$ on $\mathcal{X}$, such that $\mathcal{F}_n$ represents all information available at time $n$. For example, researchers might make the decision to ignore part of the data, or to add e.g. randomization. This would result in $\mathbb{F}$ being a poorer or richer filtration than $\sigma(X^n)$, respectively. In this section, we only consider the case $\mathcal{F}_n = \sigma(X^n)$ for simplicity's sake.

The main goal in anytime-valid testing is to construct a sequence of decision rules with a type-I error probability (the probability of rejecting $\mathcal{H}_0$ if it is actually true) that is uniformly bounded over time.

**Definition 2.1** (Anytime-valid sequential test)**.** A sequential test is a sequence $(\phi_n)_{n\in\mathbb{N}}$ of functions $\phi_n : \mathcal{X}^n \to \{0, 1\}$. A sequential test is said to be anytime valid at significance level $\alpha \in (0, 1)$ if the following holds:

$$Q(\exists n \in \mathbb{N} : \phi_n(X^n) = 1) \leq \alpha \text{ for all } Q \in \mathcal{H}_0. \tag{2.1}$$

In this definition, $\phi_n(X^n) = 1$ indicates that the null hypothesis is rejected after observing the first $n$ data points and $\phi_n(X^n) = 0$ means that there is not enough evidence to reject the null. Inequality (2.1) ensures that the type-I error probability of an anytime-valid test $(\phi_n)_{n\in\mathbb{N}}$ is smaller than $\alpha$, even if the experimenter uses the most aggressive stopping rule, that is, if they continue to collect data until $\phi_n(X^n) = 1$. However, the latter is a choice, not a requirement; the test would keep the type-I error guarantee if the experimenter chooses any other moment to stop collecting data. To emphasize that the type-I error guarantee also holds for other stopping rules, an equivalent definition of anytime-validity (see e.g. Howard et al., 2021, Lemma 3) is given by

$$Q(\phi_\tau(X^\tau) = 1) \leq \alpha \text{ for all } \tau \in \mathcal{T} \text{ and } Q \in \mathcal{H}_0. \tag{2.2}$$

In particular, anytime-valid tests are also valid under sampling rules that do not require the experimenter to reanalyze after each data point. That is, one might choose to first collect a batch of data points of size $n_1 \in \mathbb{N}$, then compute $\phi_{n_1}(X^{n_1})$ and choose whether to collect another batch of some size $n_2 \in \mathbb{N}$ or not. If so, then $\phi_{n_1+n_2}(X^{n_1+n_2})$ is computed, etc.

The output of a decision rule (that is, to reject or not) is generally based on a numerical measure of evidence against $\mathcal{H}_0$. The guarantee in (2.1), or equivalently (2.2), will only hold if this measure is chosen appropriately. Appropriate measures of evidence are the subject of the next section.

## 2.1 The E-Family

One of the central statistics in the construction of anytime-valid tests is the *e*-statistic.

**Definition 2.2** (E-statistic)**.** For fixed $n \in \mathbb{N}$, an *e*-statistic is any nonnegative statistic $E$ such that $E = E(X^n)$ and $\mathbb{E}_Q[E] \leq 1$ for all $Q \in \mathcal{H}_0$.

Here, $\mathbb{E}_Q[\cdot]$ is used to denote the expected value under $Q$. E-statistics are also known as $e$-variables. Both terms will be used interchangeably in the following chapters. Furthermore, the realization of an $e$-statistic will be referred to as an $e$-value.

Intuitively, it should not occur frequently under the null hypothesis that we observe a large $e$-value. Large $e$-values can thus be interpreted as evidence against the null hypothesis. To turn this evidence into a hypothesis test, we can apply Markov's inequality to see $Q(E \geq 1/\alpha) \leq \alpha$. It follows that the test defined by $\mathbf{1}\{E \geq 1/\alpha\}$ has a type-I error probability that is bounded by $\alpha$. However, this test is only defined after observing the first $n$ data points, and thus requires the sample size to be fixed. A first step to turn this into a sequential procedure is to realize that, if $E$ and $E'$ are independent $e$-statistics, then $E \cdot E'$ is again an $e$-statistic because $\mathbb{E}_Q[E \cdot E'] = \mathbb{E}_Q[E] \cdot \mathbb{E}_Q[E'] \leq 1$ for any $Q \in \mathcal{H}_0$. We can thus consider each data point separately, define an $e$-statistic on each data point, and then multiply the $e$-statistics to get a measure of the total evidence. That is, let $E_i = E_i(X_i)$ be an $e$-statistic for each $i$, then the product $\prod_{i=1}^n E_i$ is again an $e$-statistic for any $n \in \mathbb{N}$, because the $X_i$'s are independent. This idea can be generalized to allow dependence between the $e$-statistics by considering past-conditional $e$-statistics.

**Definition 2.3** (Past-conditional $e$-statistic)**.** A past-conditional $e$-statistic at time $n \in \mathbb{N}$ is a nonnegative statistic $E = E(X^n)$ s.t. $\mathbb{E}_Q[E \mid X^{n-1}] \leq 1$ for all $Q \in \mathcal{H}_0$.[3]

Here, the expectation is to be read unconditionally for $n = 1$. The intuition is that the past-conditional $e$-statistic at time $n$ measures the evidence against $\mathcal{H}_0$ in round $n$ conditional on the past data. The total evidence at time $n$ is then measured by the product $S_n = \prod_{i=1}^n E_i$. It follows from the law of total expectation that $S_n$ is again an $e$-statistic. However, a stronger property also holds: the cumulative product $(S_n)_{n\in\mathbb{N}}$ forms a test supermartingale with respect to $\mathcal{H}_0$.

**Definition 2.4** (Test supermartingale[4])**.** A test supermartingale for $\mathcal{H}_0$ is a sequence $(M_n)_{n\in\mathbb{N}}$ of nonnegative statistics $M_n = M_n(X^n)$ such that $\mathbb{E}_Q[M_1] \leq 1$ and $\mathbb{E}_Q[M_n \mid X^{n-1}] \leq M_{n-1}$ for $n > 1$ and all $Q \in \mathcal{H}_0$.

[3] In this definition, the relevant randomness in a past-conditional $e$-statistic originates from $X_n$ only, whereas the randomness in a 'regular' $e$-statistic may come from $n$ data points. However, as we will see in Section 2.3, it might be that $X_n = (Y_1, \ldots, Y_m)$ for some sequence of data $Y_1, \ldots, Y_m$. Then, in round $n$, we could compute a statistic $E = E(Y_1, \ldots, Y_m)$ such that either $\mathbb{E}_Q[E] \leq 1$ or $\mathbb{E}_Q[E \mid X^{n-1}] \leq 1$ for all $Q \in \mathcal{H}_0$, corresponding to a regular or a past-conditional $e$-statistic, respectively. Past-conditional $e$-statistics can thus be seen as $e$-statistics with an additional property.

[4] Conventionally, (test) supermartingales are defined with respect to a single distribution. However, the composite definition given here is particularly useful in the context of anytime-valid testing (see e.g. Ramdas et al. (2023), who call them composite test supermartingales).

The fact that the cumulative product of past-conditional $e$-statistics forms a test supermartingale for $\mathcal{H}_0$ follows from $\mathbb{E}_Q[S_n \mid X^{n-1}] = S_{n-1}\mathbb{E}_Q[E_n \mid X^{n-1}] \leq S_{n-1}$ for all $Q \in \mathcal{H}_0$ and $n > 1$. Here, the last inequality follows from the definition of the past-conditional $e$-statistic at time $n$. Furthermore, we have $\mathbb{E}_Q[S_1] = \mathbb{E}_Q[E_1] \leq 1$ for all $Q \in \mathcal{H}_0$ by definition, so that $(S_n)_{n\in\mathbb{N}}$ indeed forms a test martingale. Conversely, any test martingale $(M_n)_{n\in\mathbb{N}}$ can be decomposed into past-conditional $e$-statistics by considering $E_n = M_n/M_{n-1}$. Depending on the situation, it can be easier to consider test supermartingales as a whole or to think in terms of the decomposition in past-conditional $e$-statistics. The most well-known examples of $e$-statistics and test supermartingales are likelihood ratios. To illustrate this, suppose that $\mathcal{H}_0 = \{Q\}$ and $\mathcal{H}_1 = \{P\}$ for some $P$ and $Q$ under which data are i.i.d. and that have densities $p$ and $q$ with respect to a common background measure. Then the likelihood ratio process between $P$ and $Q$ is defined as $(M_n)_{n\in\mathbb{N}}$ with $M_n = p(X^n)/q(X^n)$. This process is a test martingale for $\mathcal{H}_0$, which can be shown as follows

$$\mathbb{E}_Q[M_n \mid X^{n-1}] = \frac{p(X^{n-1})}{q(X^{n-1})}\mathbb{E}_Q\left[\frac{p(X_n)}{q(X_n)}\right] = M_{n-1},$$

where the density in the denominator is canceled against that of the expectation. The process $(M_n)_{n\in\mathbb{N}}$ can be decomposed into (past-conditional) $e$-statistics by defining, for each $n \in \mathbb{N}$, $E_n = p(X_n)/q(X_n)$.

Test supermartingales are useful for sequential testing because any test supermartingale $(M_n)_{n\in\mathbb{N}}$ for $\mathcal{H}_0$ satisfies

$$Q(\exists n \in \mathbb{N} : M_n \geq 1/\alpha) \leq \alpha \text{ for all } Q \in \mathcal{H}_0, \tag{2.3}$$

which is an immediate implication of Ville's inequality (Ville, 1939). That is, the probability that a test martingale ever grows large is bounded under the null hypothesis. In particular, applying this to the cumulative product of past-conditional $e$-variables, it follows that the sequential test $(\phi_n)_{n\in\mathbb{N}}$ defined by $\phi_n(X^n) = \mathbf{1}\{S_n \geq 1/\alpha\}$ is anytime valid at level $\alpha$. That is, we can monitor the cumulative product over time and reject the null hypothesis as soon as it exceeds $1/\alpha$. However, as alluded to below equation (2.2), it is not necessary to monitor the test martingale after each data point to have a bounded type-I error probability. We might, for example, choose to collect data in blocks and only compute the test martingale on a new block of data rather than after each individual data point. In fact, inequality (2.3) is a consequence of a more general property: any test supermartingale $(M_n)_{n\in\mathbb{N}}$ with respect to $Q \in \mathcal{H}_0$

satisfies $\mathbb{E}_Q[M_\tau] \leq 1$ for any stopping time $\tau \in \mathcal{T}$, which follows from Doob's optional stopping theorem (Williams, 1991, Section 10.10). Combined with Markov's inequality, it follows that $Q(M_\tau \geq 1/\alpha) \leq \alpha$. We can recover inequality (2.3) by applying this to $\tau^* = \inf\{n \in \mathbb{N} : M_n \geq 1/\alpha\}$. This suggests that it is not necessary to have the extra structure of a test supermartingale (decomposability into past-conditional $e$-statistics) to define an anytime-valid test. All we need is a process that, when stopped, has expected value bounded by one. This idea gives rise to $e$-processes.

**Definition 2.5** (E-process)**.** An $e$-process is a sequence $(E_n)_{n\in\mathbb{N}}$ of nonnegative statistics $E_n = E_n(X^n)$ such that $\mathbb{E}_Q[E_\tau] \leq 1$ for all $\tau \in \mathcal{T}$ and all $Q \in \mathcal{H}_0$.

Similar to the discussion above, the definition of an $e$-process ensures that $Q(E_\tau \geq 1/\alpha) \leq \alpha$ for any $Q \in \mathcal{H}_0$, so that an anytime-valid test can be defined by $(\phi_n)_{n\in\mathbb{N}}$ where $\phi_n(X^n) = \mathbf{1}\{E_n \geq 1/\alpha\}$. It should be clear that any test martingale is also an $e$-process, however, the other way around is not necessarily true. Ramdas et al. (2022) show that it is possible to define a nontrivial $e$-process for a certain problem for which no nontrivial test martingales exist.

## 2.2 Optimality

Whenever the null hypothesis is not true, we would actually like to gather evidence against it—that is, to obtain a large $e$-value. One idea is to use the $e$-statistic that is expected to grow as fast as possible when the alternative hypothesis $\mathcal{H}_1$ is true. There are many ways to define what this means exactly; here, the approach of Grünwald et al. (2024) is followed. To this end, assume first that the alternative hypothesis is simple, that is, $\mathcal{H}_1 = \{P\}$.

**Definition 2.6** (Log-optimal $e$-statistic)**.** For a fixed $n$, the log-optimal $e$-statistic is the maximizer of $E \mapsto \mathbb{E}_P[\ln E]$ over all $e$-statistics $E = E(X^n)$.

Here, $\mathbb{E}_P[\ln E]$ can be understood as the expected rate at which evidence is accumulated per batch of $n$ data points, if one uses the $e$-statistic $E$. To explain, suppose that we partition the data into blocks of size $n$ and calculate the same $e$-statistic on each block, that is, $E_1 = E(X_1, \ldots, X_n), E_2 = E(X_{n+1}, \ldots, X_{2n})$, etc. The total evidence after $m$ blocks of data will be measured by $\prod_{i=1}^m E_i$. If the alternative $P$ is true, then by the law of large numbers, it will $P$-a.s. hold that $\prod_{i=1}^m E_i = \exp(m\mathbb{E}_P[\ln E] + o(m))$. That is, the accumulated evidence will grow exponentially fast in the number of blocks at a rate of $\mathbb{E}_P[\ln E]$. The log-optimal $e$-statistic is defined to maximize this rate. For

this reason, it is also known as the growth-rate optimal (GRO) $e$-statistic. When applied to blocks of size $n = 1$, the cumulative product $(\prod_{i=1}^{m} E_i)_{m\in\mathbb{N}}$ defines a test martingale, so that this procedure can be used for anytime-valid testing.

Alternatively, we can directly define the log-optimal $e$-process as the dynamic counterpart of the log-optimal $e$-statistic (Koolen and Grünwald, 2022).

**Definition 2.7** (Log-optimal $e$-process)**.** For a fixed randomized stopping time $\tau$, the log-optimal $e$-process is the maximizer of $(E_n)_{n\in\mathbb{N}} \mapsto \mathbb{E}_{X^n\sim P^\tau}[\ln E_n(X^n)]$ over all $e$-processes.

Here, a randomized stopping time $\tau$ is a process $\tau = (\tau_n)_{n\in\mathbb{N}}$ where $\tau_n = \tau_n(X^n) \in [0, 1]$ gives the conditional probability of stopping after having seen data $X^n$. That is, for randomized stopping times, the choice to stop after $n$ data points need not be determined deterministically by $X^n$, but may be randomized. Furthermore, $P^\tau$ denotes the distribution induced by $\tau$ together with the alternative $P$. Randomized stopping times are needed because, for regular stopping times, the log-optimal $e$-process might remain zero until that specific stopping time is reached. This is undesirable if, for whatever reason, we decide to employ a stopping time different from the one with respect to which was optimized, because we might then be left without any evidence.

The motivation of the log-optimality criterion for $e$-processes is through repeated testing. That is, instead of thinking about using the same statistic on different blocks of data, we can think about using the same $e$-process and stopping rule over independent repetitions of the same experiment. By a similar argument as above, the rate at which evidence is accumulated over repetitions of the experiment is given by $(E_n)_{n\in\mathbb{N}} \mapsto \mathbb{E}_{X^n\sim P^\tau}[\ln E_n(X^n)]$ and the log-optimal $e$-process is defined to maximize this rate.

For the stopping time $\tau \equiv n$, the log-optimal $e$-process coincides with the log-optimal $e$-statistic at sample size $n$. However, which $e$-process is log optimal generally depends on the stopping rule and can be difficult to compute—if it is even known how to do so (see e.g. Koolen and Grünwald, 2022, Section 4.5). To avoid this dependency and complexity, it is common to consider test martingales that are the cumulative product of log-optimal $e$-statistics instead. In general, it is unclear whether this is an effective approach; for example, if no nontrivial test martingales exist, it is doomed to fail. Remarkably, however, there are settings where the resulting test martingale is also the log-optimal $e$-process with respect to any stopping time, randomized or not (Koolen and Grünwald, 2022, Theorem 12). As we will see, this is true in the contexts of Chapters 4–7. Although these chapters are framed from the perspective of finding log-optimal $e$-statistics (with the exception of Chapter 7), the optimality

results they contain therefore also hold with respect to (randomized) stopping times.

So far, all of the optimality criteria were defined for a simple alternative. There are multiple ways in which they can be adjusted to handle composite $\mathcal{H}_1$. One such approach is the method of mixtures (see e.g. Robbins, 1970). That is, one can first consider the log-optimal $e$-statistic or $e$-process against each of the elements of $\mathcal{H}_1$ separately. These can then be combined by taking a convex mixture, which will result in a valid $e$-statistic or $e$-process that can be used as a measure of evidence for the entire alternative. To further illustrate, suppose that $\mathcal{H}_1 = \{P_1, P_2\}$, and that $\hat{E}_1$ and $\hat{E}_2$ are the log-optimal $e$-statistics for $P_1$ and $P_2$ respectively. That is, for $j \in \{1, 2\}$, $\hat{E}_j$ maximizes $\mathbb{E}_{P_j}[\ln E_j]$ over all $e$-statistics. Then for any $w \in [0, 1]$, the mixture $w\hat{E}_1 + (1-w)\hat{E}_2$ is an $e$-statistic. This mixture $e$-statistic serves a measure of evidence against $\mathcal{H}_0$ relative to both $P_1$ and $P_2$ simultaneously. The mixture weight $w$ can either be chosen to reflect some prior belief as to which of the alternatives is more likely, or, if no such information is available, it can be set to $w = 1/2$.

Alternatively, in the case of the multiplication of conditional $e$-statistics, one can make use of prequential plug-in estimates (see e.g. Robbins and Siegmund, 1974). That is, at each time, the (smoothed) maximum likelihood (or any other estimator) under the alternative can be computed on the basis of previous data. Then, the log-optimal $e$-statistic with respect to that estimate can be constructed. For example, let us again consider $\mathcal{H}_1 = \{P_1, P_2\}$ and suppose that $P_1$ and $P_2$ have densities $p_1$ and $p_2$ with respect to some background measure. After $n-1$ rounds, we can use these densities to determine which of the alternatives maximizes the likelihood, that is, $\hat{j}_n = \arg\max_{j \in \{1,2\}} p_j(X^{n-1})$. Then $\hat{E}_{\hat{j}_n}(X_n)$ can be used as test statistic at time $n$, so that the total evidence equals $\prod_{i=1}^{n} \hat{E}_{\hat{j}_i}(X_i)$. This can be seen as an estimate of the evidence we would have accumulated if we had known the true distribution and used the log-optimal $e$-statistic for that distribution.

The method of mixtures and prequential estimation have in common that they can be efficiently implemented as long as the log-optimal $e$-statistic for each alternative can be computed. The problem of finding log-optimal $e$-statistics for a simple alternative is therefore important even when the true alternative one has in mind is composite. This problem is at the heart of Chapters 3–7. Another method of dealing with composite alternatives is by taking a worst-case approach. That is, one can consider the maximizer of $\inf_{P \in \mathcal{H}_1} \mathbb{E}_P[\ln E]$ over all $e$-statistics $E$. This requires a fundamentally different analysis, because it cannot be implemented based on the log-optimal $e$-statistic for fixed alternatives. However, a specific instance where it can be applied efficiently is discussed in Chapter 8.

## 2.3 From Experiment to Meta-Analysis

The setup so far has suggestively been framed as a single experiment. However, there is no restriction on what $X_n$ represents for a certain $n \in \mathbb{N}$. For example, each $X_n$ could correspond to a sequence of data collected in an experiment to test $\mathcal{H}_0$. In this case, considering an anytime-valid test on $X_1, X_2, \ldots$ corresponds to performing a meta-analysis—combining the results—of all the different experiments. Viewed in this light, the above theory reveals a straightforward method of conducting meta-analyses: all experimenters should report an $e$-value and the results should be multiplied. Moreover, there might be multiple layers of anytime-validity here: within each sub-experiment, data might again be serial. That is, the data might be given by $X_n = (Y_{n,1}, \ldots, Y_{n,\tau_n})$, where $Y_{n,1}, Y_{n,2}, \ldots$ denotes the data in the $n$th experiment and $\tau_n$ is the stopping time for that experiment. If, for each experiment $n$, the researchers base their results on a conditional $e$-process $E_{n,m} = E_n(Y_{n,1}, \ldots, Y_{n,m})$, that is, such that $\mathbb{E}_Q[E_{n,\tau_n} \mid X^{n-1}] \leq 1$ for all $Q \in \mathcal{H}_0$, then the stopped $e$-process defines a past-conditional $e$-statistic for time $n$ to use in the meta-analysis. The total evidence can therefore be measured as $\prod_{i=1}^{n} E_{i,\tau_i}$.

To illustrate, suppose that some institute conducts a small-scale trial to test the efficacy of a certain drug or vaccine. Due to the limited scope of the trial, no significant conclusions could be drawn based on the results. However, the results might seem promising enough for the institute itself or another to conduct a second trial. By definition, this introduces a sequential dependence between the trials, making it very difficult to combine the results through standard meta-analysis techniques (Ter Schure and Grünwald, 2019). However, if an $e$-process $E_1 = (E_{1,n})_{n\in\mathbb{N}}$ was used in the first trial and a conditional $e$-process $E_2 = (E_{2,n})_{n\in\mathbb{N}}$ is used in the second trial, then the total evidence can be measured by multiplying the stopped $e$-process from the first trial with the $e$-process from the second trial. The total evidence can then be updated and analyzed (that is, reject the null if it exceeds $1/\alpha$) after each new observation of the second trial, while retaining type-I error guarantees. Note in particular that this guarantee does not require the number of patients in either trial to be specified upfront. Furthermore, this reasoning can be extended to multiple trials, that is, we might add a third trial, and then a fourth, etc. Hence, $e$-processes enable straightforward combination of results from separate studies (see Ter Schure and Grünwald (2022) for more details).

# 3 | On the Optimality of E-statistics

In the previous chapter, we defined the log-optimal, or GRO, $e$-statistic in the context of anytime-valid testing. In this chapter, we shift the focus to the properties of the GRO $e$-statistic, independent of that specific context. In particular, we discuss the form of the GRO $e$-statistic, as well as limitations of Definition 2.6.

To this end, it has recently been shown that, for testing simple alternatives against composite null hypotheses, there is a one-to-one correspondence between the GRO $e$-statistic and the reverse information projection (RIPr). The latter is an object that arises in information theory and is defined as the measure in the null that is closest to the alternative in information divergence. However, the RIPr as well as the GRO $e$-statistic are not uniquely defined when the infimum information divergence between the null and alternative hypothesis is infinite. We show that in such scenarios, under some assumptions, there still exists a measure in the null that is closest to the alternative in a specific sense. Whenever the information divergence is finite, this measure coincides with the usual RIPr. It therefore gives a natural extension of the RIPr to certain cases where the latter was previously not defined. This extended notion of the RIPr is shown to lead to optimal $e$-statistics in a sense that is a novel, but natural, extension of the GRO criterion. We also give conditions under which the (extension of the) RIPr is a strict sub-probability measure, as well as conditions under which an approximation of the RIPr leads to approximate $e$-statistics. For this case we provide tight relations between the corresponding approximation rates.

Throughout this chapter, we assume that the null hypothesis is convex, which is not true for most models used in practice. However, from the perspective of testing with $e$-statistics, it does not matter whether one considers a given model or its convex hull, because taking the convex hull does not change the set of all $e$-statistics.

## 3.1 Introduction

We write $D(\nu\|\lambda)$ for the information divergence (Kullback-Leibler divergence, (Kullback and Leibler, 1951; Csiszár, 1963; Liese and Vajda, 1987)) between two finite measures $\nu$ and $\lambda$ given by

$$D(\nu\|\lambda) = \begin{cases} \int_\Omega \ln\left(\frac{\mathrm{d}\nu}{\mathrm{d}\lambda}\right) \mathrm{d}\nu - (\nu(\Omega) - \lambda(\Omega)), & \text{if } \nu \ll \lambda; \\ \infty, & \text{else.} \end{cases}$$

For probability measures the interpretation of $D(\nu\|\lambda)$ is that it measures how much we gain by coding according to $\nu$ rather than coding according to $\lambda$ if data are distributed according to $\nu$. Many problems in probability theory and statistics, such as conditioning and maximum likelihood estimation, can be cast as minimization in either or both arguments of the information divergence. In particular, this is the case within the recently established and now flourishing theory of hypothesis testing based on $e$-statistics that allows for optional continuation of experiments (see Section 3.2.3)(Grünwald et al., 2024; Ramdas et al., 2023; Vovk and Wang, 2021; Shafer, 2021; Henzi and Ziegel, 2022). That is, a duality has been established between optimal $e$-statistics for testing a simple alternative $P$ against a composite null hypothesis $\mathcal{C}$ and reverse information projections (Grünwald et al., 2024). Here, the reverse information projection (RIPr) of $P$ on $\mathcal{C}$ is — if it exists — a unique measure $\hat{Q}$ such that every sequence $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ with $D(P\|Q_n) \to \inf_{Q\in\mathcal{C}} D(P\|Q)$ converges to $\hat{Q}$ in a particular norm (Li, 1999; Csiszár and Matúš, 2003). Li (1999) showed that whenever $\mathcal{C}$ is convex and $D(P\|\mathcal{C}) := \inf_{Q\in\mathcal{C}} D(P\|Q) < \infty$, the RIPr $\hat{Q}$ exists and the likelihood ratio between $P$ and $\hat{Q}$ is an $e$-statistic (this result is restated as Theorem 3.1 below). Grünwald et al. (2024) showed (restated as Theorem 3.3 below) that it is even the optimal $e$-statistic for testing $P$ against $\mathcal{C}$. However, it is clear that the RIPr cannot be defined in this way if the information divergence between $P$ and $\mathcal{C}$ is infinite, i.e. $D(P\|\mathcal{C}) = \infty$. This leaves a void in the theory of optimality of $e$-statistics. In this chapter we remedy this by realizing that even if all measures in $\mathcal{C}$ are infinitely worse than $P$ at describing data distributed according to $P$ itself, there can still be a measure that performs best relative to the elements of $\mathcal{C}$. To find such a measure, we consider the *description gain* (Topsøe, 2007) given by

$$D(P\|Q \rightsquigarrow Q') = \int_\Omega \ln\left(\frac{\mathrm{d}Q'}{\mathrm{d}Q}\right) \mathrm{d}P - (Q'(\Omega) - Q(\Omega)) \tag{3.1}$$

whenever this integral is well-defined. If the quantities involved are finite then the description gain reduces to

$$D(P\|Q \rightsquigarrow Q') = D(P\|Q) - D(P\|Q'). \tag{3.2}$$

In analogy to the interpretation of information divergence for coding, the description gain measures how much we gain by coding according to $Q'$ rather than $Q$ if data are distributed according to $P$. Furthermore denote

$$D(P\|Q \rightsquigarrow \mathcal{C}) := \sup_{Q' \in \mathcal{C}} D(P\|Q \rightsquigarrow Q'),$$

where undefined values are counted as $-\infty$ when taking the supremum. If there exists at least one $Q^* \in \mathcal{C}$ such that $P \ll Q^*$, then $D(P\|Q \rightsquigarrow \mathcal{C})$ is a well-defined number in $[0, \infty]$ for any $Q \in \mathcal{C}$. This quantity should be seen as the maximum description gain one can get by switching from $Q$ to any other measure in $\mathcal{C}$. Intuitively, if there is a best descriptor in $\mathcal{C}$, nothing can be gained by switching away from it. Indeed, in Proposition 3.6 we show that $\inf_{Q \in \mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C})$ is finite if and only if it is equal to zero.

### 3.1.1 Contents and Overview

Below, in Section 3.2, we start by giving an overview of existing results on both the reverse information projection and $e$-statistics, which we define and briefly motivate, and the growth-rate optimality (GRO) criterion, a natural replacement of statistical power within the context of $e$-value based hypothesis testing. Section 3.3 states Theorem 3.5, our first central result. It shows that — under very mild conditions — there exists a unique measure $\hat{Q}$ such that every sequence $(Q_n)_{n \in \mathbb{N}}$ in $\mathcal{C}$ with

$$D(P\|Q_n \rightsquigarrow \mathcal{C}) \to 0$$

converges to $\hat{Q}$ in a specific metric which we define. Thus, Theorem 3.5 may be viewed as a generalization of Li's result stated below as Theorem 3.1. We refer to $\hat{Q}$ as the RIPr, as it coincides with the original notion of the RIPr whenever the information divergence is finite. The remainder of Section 3.3 provides further discussion of this result, as well as an example showing that our extended notion of the RIPr can be well-defined whereas the RIPr was previously undefined. In the specific case that all initial measures are probability measures, both Li's original result and ours leave open

the possibility that $\hat{Q}$ may be a strict sub-probability measure, integrating to less than 1. In Sub-Section 3.3.1 we give a further example showing that this can indeed be the case, and we provide, via Theorem 3.9, a condition under which $\hat{Q}$ is guaranteed to be a standard (integrating to 1) probability measure. Sub-Section 3.3.2 then extends the greedy algorithm of Li and Barron (1999) and Brinda (2018) for approximating the RIPr in settings where $D(P\|\mathcal{C}) < \infty$ to settings where the information divergence might be infinite.

In Section 3.4 we turn to $e$-statistics. It contains our second central result, Theorem 3.13, which shows that whenever our extended notion of the RIPr $\hat{Q}$ exists, the likelihood ratio of $P$ and $\hat{Q}$ is an optimal $e$-statistic according to the criterion of Definition 3.11, which can be seen as a strict generalization of GRO, the standard optimality criterion for $e$-statistics. As such, this result may be viewed as a generalization of a result of Grünwald et al. (2024) stated below as Theorem 3.3. After illustrating the result by an example, Sub-Section 3.4.2 provides another technical result, Theorem 3.16, which relates approximations in terms of information gain, to approximations in terms of $e$-statisticity: conditions are given under which a sequence $Q_1, Q_2, \ldots$ converging to $\hat{Q}$ in terms of information gain at a certain rate also satisfies that the likelihood ratio between $P$ and $Q_1, Q_2, \ldots$ converges to an $e$-statistic, and tight bounds on the corresponding rates are given. After a discussion of related work, the chapter ends with a summary and ideas for future work in Section 3.5. All proofs are delegated to Appendix A.1. Appendix A.2 provides a general method for constructing RIPrs that are strict sub-probability measures. Finally, Appendix A.3 provides a discussion on the assumption of convexity that we will make throughout.

## 3.2 Background

### 3.2.1 Preliminaries

We work with a measurable space $(\Omega, \mathcal{F})$ and, unless specified otherwise, all measures will be defined on this space. Throughout, $P$ will denote a finite measure and $\mathcal{C}$ a set of finite measures, such that $P$ and all $Q \in \mathcal{C}$ have densities w.r.t. a common $\sigma$-finite measure $\mu$. These densities will be denoted with lowercase, i.e. $p$ and $q$ respectively. We will assume throughout that $\mathcal{C}$ is convex, i.e. closed under finite mixtures. In Section 3.4.1 and in more detail in Appendix A.3 we discuss how our results would be affected if we were to adopt stronger notions of convexity like $\sigma$-convexity (closed under countable mixtures), or Choquet-convexity (closed under arbitrary mixtures).

Furthermore, we assume that there exists at least one $Q^* \in \mathcal{C}$ such that $P \ll Q^*$. This assumption is needed to ensure that $D(P\|Q \rightsquigarrow \mathcal{C})$ is a well-defined number in $[0, \infty]$ for any $Q \in \mathcal{C}$.

### 3.2.2 The Reverse Information Projection

As mentioned briefly above, the reverse information projection is the result of minimizing the information divergence between $P$ and $\mathcal{C}$. If $\mathcal{C}$ is an exponential family, this problem is well understood (Csiszár and Matúš, 2003), but we focus here on the case that $\mathcal{C}$ is a general convex set. In this setting, the following theorem establishes existence and uniqueness of a limiting object for any sequence $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ such that $D(P\|Q_n) \to D(P\|\mathcal{C})$ whenever the latter is finite. This limit (i.e. $\hat{Q}$ in the following) is called the reverse information projection of $P$ on $\mathcal{C}$.

**Theorem 3.1** (Li (1999))**.** *If $P$ and all $Q \in \mathcal{C}$ are probability measures s.t. $D(P\|\mathcal{C}) < \infty$, then there exists a unique (potentially sub-) probability measure $\hat{Q}$ such that:*

1. *We have that $\ln q_n \to \ln \hat{q}$ in $L_1(P)$ for all sequences $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ such that $\lim_{n\to\infty} D(P\|Q_n) = D(P\|\mathcal{C})$.*
2. *$\int_\Omega \ln \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}P = D(P\|\mathcal{C})$,*
3. *$\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}Q \leq 1$ for all $Q \in \mathcal{C}$.*

### 3.2.3 E-Statistics and Growth Rate Optimality

The *e*-value has recently emerged as a popular alternative to the *p*-value for hypothesis testing (Ramdas et al., 2023; Henzi and Ziegel, 2022). Unlike the *p*-value, it is eminently suited for testing under optional continuation — and more generally, when the rule for stopping or continuing to analyze an additional batch of data is not under control of the data analyst, and may even be unknown or unknowable. It can be thought of as a measure of statistical evidence that is intimately linked with numerous ideas, such as likelihood ratios, test martingales (Ville, 1939) and tests of randomness (Levin, 1976). Formally, an *e*-value is defined as the value taken by an *e*-statistic, which is defined as a random variable $E : \Omega \to [0, \infty]$ that satisfies $\int_\Omega E \,\mathrm{d}Q \leq 1$ for all $Q \in \mathcal{C}$ (Vovk and Wang, 2021). The set of all *e*-statistics is denoted as $\mathcal{E}_\mathcal{C}$. Large *e*-values constitute evidence against $\mathcal{C}$ as null hypothesis, so that the null can be rejected when the computed *e*-value exceeds a certain threshold. For example, the test that rejects the null hypothesis when $E \geq 1/\alpha$ has a type-I error guarantee of $\alpha$ by

a simple application of Markov's inequality: $Q(E \geq 1/\alpha) \leq \alpha \int_\Omega E \, \mathrm{d}Q \leq \alpha$. For all further details, as well as an extensive introduction to the concept, and how it relates to optional stopping and continuation, we refer to Grünwald et al. (2024) and the overview paper by Ramdas et al. (2023).

In general, the set $\mathcal{E}_\mathcal{C}$ of $e$-statistics is quite large, and the above does not tell us *which* $e$-statistic to pick. This question was studied by Grünwald et al. (2024) and a log-optimality criterion coined GRO (*Growth-Rate Optimality*) was introduced for the case that the interest is in gaining as much evidence as possible relative to an alternative hypothesis given by a single probability measure $P$. GRO is a natural replacement of statistical power, which cannot meaningfully be used in an optional stopping/continuation context. This criterion can be traced back to the information-theoretic Kelly betting criterion by Kelly (1956) and is further discussed at length by Shafer (2021); Ramdas et al. (2023); Grünwald et al. (2024), to which we refer for more discussion.

**Definition 3.2.** If it exists, an $e$-statistic $\hat{E} \in \mathcal{E}_\mathcal{C}$ is Growth-Rate Optimal (GRO) if it achieves

$$\int_\Omega \ln \hat{E} \, \mathrm{d}P = \sup_{E \in \mathcal{E}_\mathcal{C}} \int_\Omega \ln E \, \mathrm{d}P.$$

The following theorem establishes a duality between GRO $e$-statistics and reverse information projections. For a limited set of testing problems, it states that GRO $e$-statistics exist and are uniquely given by likelihood ratios.

**Theorem 3.3** (Grünwald et al. (2024), Theorem 1)**.** *If $P$ and all $Q \in \mathcal{C}$ are probability measures such that $D(P\|\mathcal{C}) < \infty$, $p(\omega) > 0$ for all $\omega \in \Omega$, and $\hat{Q}$ is the RIPr of $P$ on $\mathcal{C}$, then $\hat{E} = \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}}$ is GRO with rate equal to $D(P\|\mathcal{C})$, i.e.*

$$\sup_{E \in \mathcal{E}_\mathcal{C}} \int_\Omega \ln E \, \mathrm{d}P = \int_\Omega \ln \hat{E} \, \mathrm{d}P = D(P\|\mathcal{C}).$$

*Furthermore, for any GRO e-statistic $\tilde{E}$, we have that $\tilde{E} = \hat{E}$ holds $P$-almost surely.*

## 3.3 The Reverse Information Projection

In this section, we state a result analogous to Theorem 3.1 in a more general setting. Rather than convergence of the logarithm of densities in $L_1(P)$, we consider convergence with respect to a different metric on the set of measurable positive functions,

i.e. $\mathcal{M}(\Omega, \mathbb{R}_{>0}) = \{f : \Omega \to \mathbb{R}_{>0} : f \text{ measurable}\}$. For $f, f' \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$ we define

$$m_P^2(f, f') := \frac{1}{2} \int_\Omega \ln\left(\frac{\overline{f}}{f}\right) + \ln\left(\frac{\overline{f}}{f'}\right) \mathrm{d}P, \tag{3.3}$$

where $\overline{f} := {}^{(f+f')}/_2$. This is a divergence that can be thought of as the averaged Bregman divergence associated with the convex function $\gamma(x) = x - 1 - \ln(x)$. In particular, this means that for $Q, Q' \in \mathcal{C}$ such that $P \ll Q$ and $P \ll Q'$, we have that

$$m_P^2(q, q') = \frac{1}{2} D(P \| Q \rightsquigarrow \bar{Q}) + \frac{1}{2} D(P \| Q' \rightsquigarrow \bar{Q}). \tag{3.4}$$

Chen et al. (2008) study averaged Bregman divergences in detail for general $\gamma$, and they show that the function

$$m_\gamma^2(x, y) = \frac{1}{2}\gamma(x) + \frac{1}{2}\gamma(y) - \gamma\left(\frac{x+y}{2}\right)$$

is the square of a metric if and only if $\ln(\gamma''(x))'' \geq 0$. In our case, $\ln(\gamma''(x))'' = 2x^{-2}$, so this result holds. This can be used together with an application of the Minkowski inequality to show that the triangle inequality holds for the square root of the divergence (3.3), i.e. $m_P$, on $\mathcal{M}(\Omega, \mathbb{R}_{>0})$. It should also be clear that for $f, g \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$ if $f = g$ everywhere, then $m_P(f, g) = 0$. Conversely $m_P(f, g) = 0$ only implies that $P(f \neq g) = 0$. This prevents us from calling $m_P$ a metric on $\mathcal{M}(\Omega, \mathbb{R}_{>0})$, and we therefore define, analogous to $\mathcal{L}^p$ and $L^p$ spaces, $M(\Omega, \mathbb{R}_{>0})$ as the set of equivalence classes of $\mathcal{M}(\Omega, \mathbb{R}_{>0})$ under the relation '$\sim$' given by $f \sim g \Leftrightarrow P(f \neq g) = 0$. By the discussion above, $m_P$ properly defines a metric on $M(\Omega, \mathbb{R}_{>0})$. In the following we will often ignore this technicality and simply act as if $m_P$ defines a metric on $\mathcal{M}(\Omega, \mathbb{R}_{>0})$, since we are not interested in what happens on null sets of $P$.

Considering convergence with respect to $m_P$ will be useful for our analyses in the following. In particular, we will exploit on numerous occasions that $m_P$ can be interpreted as a symmetrized version of the description gain, as described in Equation (3.4). However, other than mathematical convenience, there is no fundamental difference between considering convergence with respect to $m_P$ and convergence of the logarithms in $L_1(P)$, as considered in Theorem 3.1. Indeed, Lemma A.2 in Appendix A.1 shows that the two types of convergence are equivalent. It is also this result from which the following proposition follows.

**Proposition 3.4.** *The metric space* $(M(\Omega, \mathbb{R}_{>0}), m_P)$ *is complete.*

Everything is now in place to state the main result.

**Theorem 3.5.** *If* $\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) < \infty$*, then there exists a measure* $\hat{Q}$ *that satisfies the following for every sequence* $(Q_n)_{n\in\mathbb{N}}$ *in* $\mathcal{C}$ *such that* $D(P\|Q_n \rightsquigarrow \mathcal{C}) \to \inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C})$ *as* $n \to \infty$*:*

1. $q_n \to \hat{q}$ *in* $m_P$*.*
2. *If* $P'$ *is a measure such that* $|\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow P')| < \infty$*, then*

$$\int_\Omega \ln \frac{\mathrm{d}P'}{\mathrm{d}\hat{Q}} \,\mathrm{d}P = \lim_{n\to\infty} \int_\Omega \ln \frac{\mathrm{d}P'}{\mathrm{d}Q_n} \,\mathrm{d}P.$$

3. *For any* $Q \in \mathcal{C}$*,*

$$\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}Q \leq P(\Omega) + Q(\Omega) - \liminf_{n\to\infty} Q_n(\Omega).$$

Theorem 3.1 is a special case of Theorem 3.5 when $P$ and all $Q \in \mathcal{C}$ are probability measures such that $D(P\|\mathcal{C}) < \infty$. This follows because Equation (3.2) implies that minimizing $D(P\|Q \rightsquigarrow Q')$ over $Q$ is equivalent to minimizing $D(P\|Q)$ and because convergence of the densities in $m_P$ is equivalent to convergence of the logarithms in $L_1(P)$ by Lemma A.2 in Appendix A.1.1. We therefore refer to $\hat{Q}$ as the *reverse information projection* of $P$ on $\mathcal{C}$, thereby extending the definition of the latter (we refrain from the term 'generalized RIPr', because it has already been used for the RIPr whenever it is not attained by an element of $\mathcal{C}$ (Csiszár and Matúš, 2003) or when the log score is replaced by another loss function (Grünwald and Mehta, 2020)). However, the density of the measure $\hat{Q}$ is only unique as an element of $M(\Omega, \mathbb{R}_{>0})$, since convergence of the densities holds in $m_P$. This causes no ambiguity here, so that we simply refer to it as 'the' RIPr.

Note that Theorem 3.5 implies that if there exists a $Q \in \mathcal{C}$ with $D(P\|Q \rightsquigarrow \mathcal{C}) = 0$, then $Q$ is the RIPr of $P$ on $\mathcal{C}$. This matches with the intuition that the maximum gain we can get from switching away from the 'best' code in $\mathcal{C}$ should be equal to zero. The following result establishes this more formally.

**Proposition 3.6.** *The following conditions are equivalent:*

1. *There exists a measure* $P'$ *such that* $D(P\|P' \rightsquigarrow \mathcal{C})$ *is finite.*
2. *There exists a measure* $Q$ *in* $\mathcal{C}$ *such that* $D(P\|Q \rightsquigarrow \mathcal{C})$ *is finite.*

3. *There exists a sequence of measures* $Q_n \in \mathcal{C}$ *such that* $D(P\|Q_n \rightsquigarrow \mathcal{C}) \to 0$ *for* $n \to \infty$.

*Consequently, whenever* $\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) < \infty$*, it must actually be equal to zero.*

To show that the reverse information projection exists, it is therefore enough to prove that one of these equivalent conditions holds. Which condition is easiest to check will depend on the specific setting, as exemplified by the following propositions.

**Proposition 3.7.** *Suppose that* $\mathcal{C}$ *is the convex hull of finitely many distributions, that is,* $\mathcal{C} = \operatorname{conv}(\{Q_1, \ldots, Q_n\})$*, then for any probability measure* $P$ *with* $P \ll Q_i$ *for at least one* $i$*, it holds that* $D(P\|\frac{1}{n}\sum Q_i \rightsquigarrow \mathcal{C}) < \infty$.

**Example 3.1.** Let $\mathcal{C}$ be a singleton whose single element $Q$ is given by the standard Gaussian and let $P$ be the standard Cauchy distribution. Since the Cauchy distribution is exponentially heavier-tailed than the Gaussian, we have that $D(P\|\mathcal{C}) = \infty$. However, since both distributions have full support, it follows that

$$D(P\|Q \rightsquigarrow \mathcal{C}) = D(P\|Q \rightsquigarrow Q) = 0.$$

By Theorem 3.5 (1), $Q$ is therefore the reverse information projection of $P$ on $\mathcal{C}$.

This example can be extended to composite $\mathcal{C}$ by considering all mixtures of the Gaussian distributions $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 1)$ with mean $\pm 1$ and variance 1. Proposition 3.7 guarantees the existence of a reverse information projection although the information divergence is still infinite because a Cauchy distribution is more heavy tailed than any finite mixture of Gaussian distributions. Symmetry implies that the reverse information projection must be equal to the uniform mixture of $\mathcal{N}(-1, 1)$ and $\mathcal{N}(1, 1)$, which coincides with the result one would intuitively expect.

**Proposition 3.8.** *Assume that* $\mathcal{C}$ *is a convex set of probability measures that has finite minimax regret and with normalized maximum likelihood distribution* $Q^* \in \mathcal{C}$*. Then for any probability measure* $P$ *that is absolutely continuous with respect to* $Q^*$*, it holds that* $D(P\|Q^* \rightsquigarrow \mathcal{C}) < \infty$.

For an extensive discussion on minimax regret in the present coding context, as well as the normalized maximum likelihood distribution (also known as *Shtarkov* distribution), see e.g. Grünwald and Harremoës (2009); van Erven and Harremoës (2014). In short, the minimax regret is defined as $\inf_{Q\in\mathcal{C}} \sup_{Q'\in\mathcal{C},\omega\in\Omega} \ln {}^{q'(\omega)}/_{q(\omega)}$. This quantity is known to be finite if and only if the normalized maximum likelihood distribution $Q^*$, defined as $q^*(\omega) = {}^{\sup_{Q\in\mathcal{C}} q(\omega)}/\left(\int_\Omega \sup_{Q\in\mathcal{C}} q \,\mathrm{d}\mu\right)$, is well-defined. One-dimensional

exponential families with finite minimax regret have been classified by Grünwald and Harremoës (2009).

### 3.3.1 Strict Sub-Probability Measure

We return now to the familiar setting where $P$ is a probability measure and $\mathcal{C}$ a convex set of probability measures. It is easy to verify that the RIPr $\hat{Q}$ of $P$ on $\mathcal{C}$ is then a sub-probability measure. This follows because we know that there exists a sequence $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ such that $q_n$ converges point-wise $P$-a.s. to $\hat{q}$ and Fatou's Lemma tells us

$$\int_\Omega \hat{q}\,\mathrm{d}\mu = \int_\Omega \liminf_{n\to\infty} q_n\,\mathrm{d}\mu \leq \liminf_{n\to\infty} \int_\Omega q_n\,\mathrm{d}\mu = 1. \tag{3.5}$$

It is not clear a priori whether this can ever be a strict inequality. For example, if the sample space is finite, the set of probability measures is compact, so the limit of any sequence of probability measures (i.e. the reverse information projection) will also be a probability measure. The following example illustrates that this is not always the case for infinite sample spaces, and it can in fact already go wrong for a countable sample space with $D(P\|\mathcal{C}) < \infty$.

**Example 3.2.** Let $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\mathbb{N}}$. Furthermore, let $P$ denote the probability measure $\delta_1$ concentrated in the point $i = 1$ and $\mathcal{C}$ the set of distributions $Q$ satisfying

$$\sum_{i=1}^{\infty} \frac{1}{i} q(i) = \frac{1}{2}.$$

This set is defined by a linear constraint, so that $\mathcal{C}$ is convex, and for any $Q \in \mathcal{C}$, we have

$$q(1) + \sum_{i=2}^{\infty} \frac{1}{i} q(i) = \sum_{i=1}^{\infty} \frac{1}{i} q(i) = \frac{1}{2},$$

implying that $q(1) \leq 1/2$. It follows that $D(P\|Q) = -\ln(q(1)) \geq \ln(2)$. The sequence $Q_n = \frac{n-2}{2n-2}\delta_1 + \frac{n}{2n-2}\delta_n$ satisfies $Q_n \in \mathcal{C}$ and

$$D(P\|Q_n) = \ln \frac{2n-2}{n-2} \to \ln(2).$$

Consequently, it must hold that $D(P\|\mathcal{C}) = \ln(2)$. The sequence $Q_n$ converges to the strict sub-probability measure $(1/2)\delta_1$, which must therefore be the RIPr of $P$ on $\mathcal{C}$.

A more general example, which can be seen as a template to create such situa-

tions, is given in Appendix A.2. The common theme is that $\mathcal{C}$ is defined using only constraints of the form $\sum_i f_1(i)q(i) = c$, where $f_1$ is some positive function such that $\lim_{n\to\infty} f_1(n) = 0$. Since $\mathcal{C}$ only contains probability measures, there is the additional constraint that $\sum_i f_0(i)q(i) = 1$, where $f_0$ denotes the constant function $f_0 \equiv 1$. This function $f_0$ dominates all other constraints $f_1$ in the sense that $\lim_{i\to\infty} {}^{f_1(i)}/_{f_0(i)} = 0$, but is itself not dominated by any of the constraints in the same manner. It turns out that this is the precise condition that dictates whether or not a constraint has to be respected when taking point-wise limits of elements in $\mathcal{C}$. Indeed, as shown in the theorem below, any constraint on $\mathcal{C}$ that is dominated by another constraint in the sense described above cannot be violated by taking point-wise limits. Therefore, if we add a restriction to $\mathcal{C}$ that dominates the constant function 1, i.e. that is defined by some function $f_1$ with $\lim_{n\to\infty} f_1(n) = \infty$, then the RIPr cannot be a strict sub-probability measure.

**Theorem 3.9.** *Take $\Omega = \mathbb{N}, \mathcal{F} = 2^{\mathbb{N}}$, and let $\mathcal{C}$ be a convex set of probability measures. Suppose that for $f_0, f_1 : \mathbb{N} \to \mathbb{R}_{>0}$, we have that $\sum_i f_0(i)q(i) \leq \lambda_0$ and $\sum_i f_1(i)q(i) = \lambda_1$ for all $Q \in \mathcal{C}$. If $Q_n$ denotes a sequence of measures in $\mathcal{C}$ that converges point-wise to some distribution $Q^*$, and $f_0$ dominates $f_1$ in the sense that*

$$\lim_{i\to\infty} \frac{f_1(i)}{f_0(i)} = 0, \tag{3.6}$$

*then*

$$\sum_i f_1(i) \cdot q^*(i) = \lambda_1. \tag{3.7}$$

### 3.3.2 Greedy Approximation

So far, we have discussed the existence and properties of the RIPr of $P$ on $\mathcal{C}$. However, there will be many situations where it is infeasible to compute this exact projection, as it requires solving a complex minimization problem. For example, if $\mathcal{C}$ is given by the convex hull of some parameterized family of distributions, the reverse information projection might be an arbitrary mixture of elements of this family, and the minimization problem need not be convex in the parameters of the family. To this end, Li and Barron (1999) propose an iterative greedy algorithm for the case that $\mathcal{C}$ is given by the $\sigma$-convex hull (all countable mixtures, see Appendix A.3) of a parameterized family of distributions, i.e. $\mathcal{C} = \sigma\text{-conv}(\{Q_\theta : \theta \in \Theta\})$, and $D(P\|\mathcal{C}) < \infty$. The algorithm starts by setting $Q_1 := Q_{\theta_1}$, where $\theta_1$ minimizes $D(P\|Q_{\theta_1})$, and then

iteratively defining $Q_k := (1-\alpha_k)Q_{k-1} + \alpha_k Q_{\theta_k}$, where $\alpha_k = 2/(k+1)$[1] and $\theta_k$ is chosen to minimize $D(P\|Q_k)$. It is shown that, if $\sup_{x,\theta_1,\theta_2} \log q_{\theta_1}(x)/q_{\theta_2}(x)$ is bounded, then $D(P\|Q_k)$ converges to $D(P\|\mathcal{C})$ at rate $1/k$. Later, Brinda (2018) showed that the condition that the likelihood ratio has to be uniformly bounded in $x$ can be relaxed to the condition that (3.8) below is finite. In both of these previous works, it is simply assumed that a minimizer in each step exists, though it need not necessarily be unique. We will do likewise in the following, where we give an adaptation of the algorithm that works when the KL divergence is infinite.

**Algorithm 1** Greedy Approximation of the RIPr

Fix $Q^* \in \mathcal{C}$ s.t. $|\inf_{\theta\in\Theta} \int_\Omega \log q^*/q_\theta \,\mathrm{d}P| < \infty$
Let $Q_1 = Q_{\theta_1}$, where $\theta_1 = \arg\min_{\theta'\in\Theta} D(P\|Q_{\theta'} \rightsquigarrow Q^*)$
**for** $k = 2, 3, \ldots$ **do**
    Choose $\alpha_k = \frac{2}{k+1}$ and $\theta_k = \arg\min_{\theta'\in\Theta} D(P\|(1-\alpha_k)Q_{k-1} + \alpha_k Q_{\theta'} \rightsquigarrow Q^*)$
    Let $Q_k = (1-\alpha_k)Q_{\theta_{k-1}} + \alpha_k Q_{\theta_k}$
**end for**

**Proposition 3.10.** *Suppose that* $\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) < \infty$*, let* $(Q_k)_{k\in\mathbb{N}}$ *be the output of Algorithm 1, and let* $Q$ *be any measure in* $\mathcal{C}$*, so that* $q = \sum_{\theta\in\Theta'} q_\theta \cdot w_Q(\theta)$ *for some probability mass function* $w_Q$ *on a countable* $\Theta' \subset \Theta$*. If* $D(P\|Q' \rightsquigarrow Q'')$ *is finite for all* $Q', Q'' \in \mathcal{C}$*, then it holds that*

$$D(P\|Q_k \rightsquigarrow Q) \leq \frac{b_Q^{(k)}(P)}{k},$$

*where* $b_Q^{(k)}(P)$ *is given by*

$$\int_\Omega \left(1 + \sup_{\theta^*\in\{\theta_i\}_{i=1}^k} \log \frac{\sup_{\theta\in\Theta} q_\theta}{q_{\theta^*}}\right) \frac{\sum_{\theta\in\Theta'} q_\theta^2 \cdot w_Q(\theta)}{q^2} \,\mathrm{d}P \leq \sup_{Q\in\mathcal{C}} \int_\Omega \left(1 + \sup_{\theta^*,\theta\in\Theta} \log \frac{q_\theta}{q_{\theta^*}}\right) \frac{\sum_{\theta\in\Theta'} q_\theta^2 \cdot w_Q(\theta)}{q^2} \,\mathrm{d}P. \tag{3.8}$$

It follows that if $b_Q^{(k)}$ is uniformly bounded over all $Q \in \mathcal{C}$, in particular if (3.8) is finite, then $D(P\|Q_k \rightsquigarrow \mathcal{C})$ converges to zero, i.e. $Q_k$ converges to the RIPr of $P$ on $\mathcal{C}$, at rate $1/k$. The former holds under the strong, but often imposed assumption that the

[1] Li actually proposes to either minimize over $\alpha_k$ or use $\alpha_2 = 2/3$ and $\alpha_k = 2/k$ for $k > 2$; the formulation given here is a slight simplification by Brinda (2018).

likelihood ratios in $\mathcal{C}$ are uniformly bounded; for example when $\mathcal{C}$ is given by the $\sigma$-convex hull of Gaussian densities restricted to a cube (Li, 1999, Example 1). However, (3.8) might also be finite under weaker assumptions. For example, consider the set of Gaussian mixtures as in Example 3.1, that is, $\mathcal{C} = \{w \cdot \mathcal{N}(-1,1) + (1-w) \cdot \mathcal{N}(1,1) : w \in [0,1]\}$. It can be seen that $b_Q(P) < \infty$ for all $Q \in \mathcal{C}$ whenever $P$ has a finite first moment. Moreover, if the latter holds, then $b_Q(P)$ is uniformly bounded over all $Q \in \mathcal{C}'$ where $\mathcal{C}' = \{w \cdot \mathcal{N}(-1,1) + (1-w) \cdot \mathcal{N}(1,1) : w \in [c, 1-c]\}$ for some $c \in (0, 1/2)$.

Whereas Proposition 3.10 is a satisfying theoretical result, we must concede that Algorithm 1 might not be the fastest to implement in practice. This arises from the fact that the objective $D(P\|(1-\alpha_k)Q_{k-1} + \alpha_k Q_{\theta'} \rightsquigarrow Q^*)$ need not be convex in $\theta'$. One might therefore have to resort to an exhaustive search over a discretization of the parameter space. On top of that, there is no guarantee that the information gain is easily computable. As an alternative for the case that $\Theta$ is finite and $D(P\|\mathcal{C}) < \infty$, one might use the iterative algorithm proposed by Csiszár and Tusnády (1984, Theorem 5). A big advantage of the latter is that their recursive update step has an explicit formula, which makes each iteration considerably faster. The downside is that, whereas convergence in terms of KL is guaranteed, it is unclear at what rate this happens in general. Furthermore, proving convergence of their algorithm in the setting where $D(P\|\mathcal{C}) = \infty$ seems far from a straightforward exercise.

### 3.3.3 Discussion

The results in this section might be regarded as a generalization of large parts of Chapters 3 and 4 in Li's Ph.D. thesis (Li, 1999) and in fact the tools in this section were initially developed to clear up some ambiguity around the proof of Theorem 3.1, Part 1 as provided by Li. That is, Li states that for all sequences $(Q_n)_{n\in\mathbb{N}}$ in $\mathcal{C}$ such that $\lim_{n\to\infty} D(P\|Q_n) = D(P\|\mathcal{C})$ it holds that $\ln q_n \to \ln \hat{q}$ in $L_1(P)$. However, the proof thereof refers to his Lemma 4.3, which only shows existence of one such a sequence. Then, in Lemma 4.4, Li also shows that if $\hat{Q}$ is such that $\log q_n \to \log \hat{q}$ in $L_1(P)$ for some sequence $(Q_n)_{n\in\mathbb{N}}$ that achieves $\lim_{n\to\infty} D(P\|Q_n) = D(P\|\mathcal{C})$, then it must hold that $D(P\|\hat{Q}) = D(P\|\mathcal{C})$. However, it is a priori not clear whether every sequence $(Q_n)_{n\in\mathbb{N}}$ that achieves $\lim_{n\to\infty} D(P\|Q_n) = D(P\|\mathcal{C})$ has such a limit. Moreover, it is never shown that, if it exists, this limit must be the same for every such sequence. Note that it is not at all our intention here to criticize Li's fundamental and ground-breaking work. Li's is one of those rare theses that have had a major impact outside of their own research area: being a thesis on information-theory, it served

as the central tool and inspiration for papers on fast convergence rates in machine learning theory (van Erven et al., 2015; Grünwald and Mehta, 2020), and also for Grünwald et al. (2024), which led to a breakthrough in ($e$-based) hypothesis testing. Our aim is merely to indicate that Theorem 3.5 ties up some loose ends in Li's original, pioneering results.

## 3.4 Optimal E-Statistics

In this section, we assume that $P$ and all $Q \in \mathcal{C}$ are probability measures, and we are interested in the hypothesis test with $P$ as alternative and $\mathcal{C}$ as null. To this end, Theorem 3.5 shows that — whenever it exists — the likelihood ratio of $P$ and its RIPr is an $e$-statistic. A natural question is whether the optimality of the RIPr in terms of describing data distributed according to $P$ carries over to some sort of optimality of the $e$-statistic, as is true for the GRO criterion in the case that $D(P\|\mathcal{C}) < \infty$. It turns out that this is true in terms of an intuitive extension of the GRO criterion. Completely analogously to the coding story, we simply have to change from absolute to pairwise comparisons.

**Definition 3.11.** For $e$-statistics $E, E' \in \mathcal{E}_{\mathcal{C}}$, we say that $E$ is *stronger* than $E'$ if the following integral is well-defined and nonnegative, possibly infinite:

$$\int_{\Omega} \ln \left( \frac{E}{E'} \right) \mathrm{d}P, \tag{3.9}$$

where we adhere to the conventions $\ln({}^{0}\!/_{c}) = -\infty$ and $\ln({}^{c}\!/_{0}) = \infty$ for all $c \in \mathbb{R}_{>0}$. Furthermore, an $e$-statistic $E^* \in \mathcal{E}_{\mathcal{C}}$ is a *strongest* $e$-statistic if it is stronger than any other $e$-statistic $E \in \mathcal{E}_{\mathcal{C}}$.

The notion of optimality in Definition 3.11 comes down to the simple idea that if one $e$-statistic $E$ is stronger than another $e$-statistic $E'$, then *repeatedly* testing based on $E$ eventually becomes more powerful than repeatedly testing based on $E'$ in the sense that there is a higher probability of rejecting a false null-hypothesis. Let us explain in more detail what we mean by this. Suppose that we conduct the same experiment $N$ times independently to test the veracity of the hypothesis $\mathcal{C}$, resulting in outcomes $\omega_1, \ldots, \omega_N$. For any given $e$-statistic $E \in \mathcal{E}_{\mathcal{C}}$, we have that $\prod_{i=1}^{N} E(\omega_i)$ is still an e-statistic, not just for fixed $N$ but even if $N$ is a random (i.e. data-dependent) stopping time. So, as indicated before, it can be used to test $\mathcal{C}$ with type-I error guarantees. Yet, for two $e$-statistics $E, E' \in \mathcal{E}_{\mathcal{C}}$, the law of large numbers states that

if $P$ is true, it will almost surely hold that

$$\frac{\prod_{i=1}^{n} E(\omega_i)}{\prod_{i=1}^{n} E'(\omega_i)} = \exp\left(n \int_\Omega \ln\left(\frac{E}{E'}\right) \mathrm{d}P + o(n)\right).$$

It follows that if the integral $\int_\Omega \ln\left({}^{E}/_{E'}\right) \mathrm{d}P$ is positive then with high probability $E$ will, for large enough $n$, give more evidence against $\mathcal{C}$ than $E'$ if the alternative is true, i.e. a test based on $E$ will asymptotically have more power than a test based on $E'$.

Since we assume throughout that there exists a $Q^* \in \mathcal{C}$ such that $P \ll Q^*$, it follows that for any $e$-statistic $E$ we must have $P(E = \infty) = 0$, which simplifies any subsequent analyses greatly.

**Proposition 3.12.** *Assume that $\mathcal{C}$ is a set of probability measures and that $P$ is a probability measure. If there is an $E' \in \mathcal{E}_\mathcal{C}$ such that $\sup_{E \in \mathcal{E}_\mathcal{C}} \int_\Omega \ln\left({}^{E}/_{E'}\right) \mathrm{d}P < \infty$, then a strongest e-statistics exists. Furthermore, if $E_1$ and $E_2$ are both strongest e-statistics then $E_1 = E_2$ holds $P$-a.s.*

The strongest $e$-statistic in Definition 3.11 can be seen as a generalization of the GRO $e$-statistic, because if $\int_\Omega \ln E \, \mathrm{d}P$ and $\int_\Omega \ln E' \, \mathrm{d}P$ are both finite, (3.9) can be written as the difference between the two logarithms, that is, $\int_\Omega \ln\left({}^{E}/_{E'}\right) \mathrm{d}P = \int_\Omega \ln E \, \mathrm{d}P - \int_\Omega \ln E' \, \mathrm{d}P$. In this case, finding the strongest e-statistic therefore corresponds to maximizing $\int_\Omega \ln E \, \mathrm{d}P$ over all $e$-statistics, thus recovering the original GRO criterion. As an extension of that case, we prove that whenever the RIPr exists, it always leads to the strongest $e$-statistic.

**Theorem 3.13.** *Suppose that both $P$ and all $Q \in \mathcal{C}$ are probability measures and that $\inf_{Q \in \mathcal{C}} D(P \| Q \rightsquigarrow \mathcal{C}) < \infty$. If $\hat{Q}$ denotes the RIPr of $P$ on $\mathcal{C}$, then $\hat{E} = {}^{\mathrm{d}P}/_{\mathrm{d}\hat{Q}}$ is the strongest e-statistic.*

The likelihood ratio between $P$ and its RIPr is in fact the only $e$-statistic in the form of a likelihood ratio with $P$ in the numerator, as the following proposition shows. Though the statement is more general, the proof is completely analogous to part of the proof of Lemma 4.1 by Li (1999).

**Proposition 3.14.** *Suppose that $\mathcal{C}$ is a set of probability measures and that $P$ is a probability measure. If there exists a measure $Q^* \in \mathcal{C}$ such that ${}^{\mathrm{d}P}/_{\mathrm{d}Q^*} \in \mathcal{E}_\mathcal{C}$, then $D(P \| Q^* \rightsquigarrow \mathcal{C}) = 0$, i.e. $Q^*$ is the RIPr of $P$ on $\mathcal{C}$.*

We now return to Example 3.1, where the GRO criterion is not able to distinguish between $e$-variables, but we are able to do so with Definition 3.11 and Theorem 3.13.

**Example 3.1** (continued)**.** In the case that $P$ is the standard Cauchy and $\mathcal{C} = \{Q\}$, where $Q$ is the standard Gaussian, it is straightforward to see that the likelihood ratio between $P$ and $Q$ is an $e$-statistic, i.e.

$$\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}Q}\,\mathrm{d}Q = \int_\Omega \mathrm{d}P = 1.$$

However, for the growth rate it holds that

$$\int_\Omega \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\,\mathrm{d}P = D(P\|Q) = \infty.$$

The same argument can be used to show that for any $0 < c \leq 1$, we have an $e$-statistic given by $c\,\mathrm{d}P/\mathrm{d}Q$, which still has infinite growth rate. The GRO criterion in Definition 3.2 is not able to tell which of these $e$-statistics is preferable. However, since $Q$ is the RIPr of $P$ on $\mathcal{C}$, it follows from Theorem 3.13 that $\mathrm{d}P/\mathrm{d}Q$ is the strongest $e$-statistic, and in particular it is stronger than $c\,\mathrm{d}P/\mathrm{d}Q$ for all $0 < c < 1$.

### 3.4.1 Convexity

In the discussion above, the null hypothesis $\mathcal{C}$ is assumed to be convex, which does not hold for many of the null hypotheses commonly employed in statistics, such as the set of all Gaussian distributions with varying mean and/or variance. However, it follows from the Fubini-Tonelli theorem that the set of $e$-statistics on $\mathcal{C}$ equals the set of $e$-statistics on the convex hull of $\mathcal{C}$. The same is true if the convex hull is replaced by the $\sigma$-convex hull where countable mixtures are allowed or by the Choquet-convex hull where arbitrary mixtures are allowed (see Appendix A.3 for precise definitions). Therefore, if there exists a strongest $e$-statistic for testing the alternative $P$ against any of these notions of the convex hull, then that is also the strongest $e$-statistic for testing $P$ against the original null hypothesis, regardless of whether that was convex. It follows from Theorem 3.13 that, to find the strongest $e$-statistic, it suffices to find the RIPr of $P$ on any of the notions of the convex hull. In particular, if RIPrs exists on more than one of these, they must coincide; on the other hand, none of the three RIPrs may exist, and our results also do not rule out the possibility that the RIPr exists on just one or two of the three convex hulls. To witness, in Appendix A.3 we give an example (Example A.1) in which the RIPr of $P$ on the $\sigma$-convex hull exists, whereas the RIPr of $P$ on the convex hull does not. At the same time, there are constraints: Theorem A.8 in Appendix A.3 implies that if the RIPr on the convex hull of $\mathcal{C}$ exists, then the RIPr on the $\sigma$-convex hull of $\mathcal{C}$ also exists (and then they must be equal).

Things become much more clear-cut if the RIPr $\hat{Q}$ of $P$ on a certain notion of the convex hull exists *and is an element of that set*. In that case, $\hat{Q}$ is also the RIPr of $P$ on any stronger notion of the convex hull. Indeed, the different levels of convex hulls are nested, and their corresponding sets of $e$-statistics coincide, so this follows directly from Proposition 3.14:

**Corollary 3.15.** *Let $\mathcal{C}$ denote a set of probability measures (not necessarily convex) and let $P$ denote a probability measure. If the RIPr of $P$ on the convex hull of $\mathcal{C}$ exists and is given by $\hat{Q} \in \mathrm{conv}(\mathcal{C})$, then $\hat{Q}$ is also, (a) the RIPr of $P$ on the $\sigma$-convex and, (b), on the Choquet-convex hull of $\mathcal{C}$. Similarly, if $\hat{Q} \in \sigma\text{-conv}(\mathcal{C})$ is the RIPr of $P$ on $\sigma\text{-conv}(\mathcal{C})$, then (c) $\hat{Q}$ is also the RIPr of $P$ on the Choquet-convex hull of $\mathcal{C}$.*

Further details regarding convexity are presented in Appendix A.3. In particular, Theorem A.8 in the latter gives an analogous result to Corollary 3.15, Part (a), for the case that $P$ and $\mathcal{C}$ are not restricted to be probability measures, and the RIPr is not assumed to be attained in the set.

### 3.4.2 Approximation

In Section 3.3.2, we discussed an algorithm that provides an approximation of the RIPr for scenarios where it is not possible to explicitly compute the latter. However, the convergence guarantee given by Proposition 3.10 is in terms of the information gain. That is, if $Q_k$ is the approximation of the projection after $k$ iterations, then under suitable conditions it holds that $D(P\|Q_k \rightsquigarrow \mathcal{C}) \to 0$. This is not enough if we want to use such an approximation for hypothesis testing: we need that $p/q_k$ gets closer and closer to being an $e$-statistic. The following theorem gives a condition under which this is true.

**Theorem 3.16.** *Assume $\inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) < \infty$, fix $Q, Q' \in \mathcal{C}$, set $\delta := D(P\|Q \rightsquigarrow \mathcal{C})$ and suppose that there exists $\beta \in (0, \infty]$ such that $\|q'/q\|_{1+\beta} < \infty$. If $\beta \leq 1$ or $D(P\|Q' \rightsquigarrow \mathcal{C}) \leq K\delta$, then it holds that*

$$\int_\Omega \frac{p}{q}\,\mathrm{d}Q' = \int_\Omega \frac{q'}{q}\,\mathrm{d}P = 1 + O\left(C_\beta \cdot \delta^{\frac{\beta}{1+\beta}}\right) \text{ as } \delta \to 0, \tag{3.10}$$

*where $C_\beta = \|q'/q\|_{1+\beta}$ if $\beta \leq 1$ and $C_\beta = K^{\frac{\beta-1}{2(1+\beta)}} \|q'/q\|_{1+\beta}$ otherwise.*

Here, we use $\|f\|_p$ for $p \in (0, \infty]$ to denote the $\mathcal{L}^p(\Omega, P)$ norm of a function $f \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$, i.e. $\left(\int_\Omega (f)^p \,\mathrm{d}P\right)^{1/p}$. Explicit values for the constants in (3.10) can be

found in the proof in the appendix. In particular, Theorem 3.16 implies the following: if there are $C, \delta_0 > 0$ such that $\|q'/q\|_2 \leq C$ for all $Q' \in \mathcal{C}$ and all $Q \in \mathcal{C}$ with $D(P\|Q \rightsquigarrow \mathcal{C}) \leq \delta_0$, then any sequence $Q_1, Q_2, \ldots$ with $D(P\|Q_k \rightsquigarrow \mathcal{C}) \to 0$ will have $\sup_{q' \in \mathcal{C}} \int_\Omega q'/q_k \, \mathrm{d}P = 1 + O(\delta_k^{1/2})$, where $\delta_k = D(P\|Q_k \rightsquigarrow \mathcal{C})$. This gives an easy to check condition for the convergence of $p/q_k$ to an $e$-statistic. This square-root rate of convergence cannot be improved in general without an extra assumption, even if all likelihood ratios are bounded, i.e. $\|q'/q\|_\infty < \infty$. This can be seen by taking $P$ and $Q$ to be Bernoulli distributions with parameter $1/2$ and $1/2 + \epsilon$ respectively, $\mathcal{C}$ the set of Bernoulli distributions with parameters in $[1/4, 3/4]$ and $Q'$ Bernoulli $1/4$. Then $\delta = D(P\|Q \rightsquigarrow \mathcal{C}) = 2\epsilon^2(1 + o(1))$ yet $\int_\Omega q'/q \, \mathrm{d}P = 1 + 4\epsilon(1 + o(1))$. But if likelihood ratios are bounded and we additionally consider $Q'$ in a 'neighborhood' of $Q$ (i.e. $D(P\|Q' \rightsquigarrow \mathcal{C}) \leq K\delta$), then a linear rate is possible as shown in Theorem 3.16 by letting $\beta$ tend to infinity; the rate then interpolates between $\delta^{1/2}$ and $\delta$ depending on the largest $\beta$ for which the $(1 + \beta)$-th moment exists. Furthermore the following example shows that in general bounds on the integrated likelihood ratios are necessary for the convergence to hold at all.

**Example 3.3.** Let $\mathcal{Q}$ represent the family of geometric distributions on $\Omega = \mathbb{N}_0$ and let $\mathcal{C} = \mathrm{conv}(\mathcal{Q})$. The elements of $\mathcal{Q}$ are denoted by $Q_\theta$ with density $q_\theta(n) = \theta^n(1-\theta)$, where $\theta \in [0, 1)$ denotes the probability of failure. For simplicity, assume that $P \in \mathcal{Q}$ so that the reverse information projection of $P$ on $\mathcal{C}$ is equal to $P$. Take for example $P = Q_{1/2}$, then for any $\theta, \theta' \in [0, 1)$

$$\begin{aligned} \int_\Omega \frac{q_{\theta'}}{q_\theta} \, \mathrm{d}P &= \sum_{n=0}^{\infty} \left(\frac{1}{2}\frac{\theta'}{\theta}\right)^n \frac{1}{2}\frac{1-\theta'}{1-\theta} \\ &= \begin{cases} \frac{1}{1-\frac{\theta'}{2\theta}} \cdot \frac{1}{2}\frac{1-\theta'}{1-\theta}, & \text{if } \theta' < 2\theta; \\ \infty, & \text{otherwise;} \end{cases} \end{aligned} \tag{3.11}$$

whereas

$$\begin{aligned} D(P\|Q_\theta) &= \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^{n+1} (-n\log(2\theta) - \log 2(1-\theta)) \\ &= \log\frac{1/2}{1-\theta} + \log\frac{1/2}{\theta}, \end{aligned}$$

Now consider a sequence $1/3 < \theta_1 < \theta_2 < \theta_3 \ldots$ that converges to $1/2$. Then by the

above,

$$D(P\|Q_{\theta_i}) \to 0 = D(P\|\mathcal{C}).$$

We also see that for all $i$ and all $\theta' \in [2\theta_i, 1)$, we have

$$\int_\Omega \frac{q_{\theta'}}{q_{\theta_i}} \mathrm{d}P = \infty,$$

i.e. for all $i$ we have $\sup_{\theta' \in [0,1)} \int_\Omega {}^{q_{\theta'}}/_{q_{\theta_i}} \mathrm{d}P = \infty$.

### 3.4.3 Related Work

The results on the existence of optimal $e$-statistics displayed in this section bear similarities with work concurrently done by Zhang et al. (2024). In particular, they show that if $\mathcal{C}$ is a convex polytope, then there exists an $e$-statistic in the form of a likelihood ratio between two unspecified measures. Since a convex polytope contains the uniform mixture of its vertices, which can be shown to have finite information gain, this also follows from our Proposition 3.7. However, the techniques used to prove their results appear to be of a completely different nature than the ones used in this chapter, as they rely mostly on classical results in convex geometry together with results on optimal transport (and with these techniques, they provide various other results incomparable to ours).

In the case of compact alternative they furthermore discuss a property which they refer to as nontrivial $e$-power. That is, if the alternative is a convex polytope $\mathcal{A}$, then at least one of their $e$-statistics in the form of a likelihood ratio satisfies $\inf_{P\in\mathcal{A}} \int_\Omega \ln E \, \mathrm{d}P > 0$. We now show that the existence of such an $e$-statistic also follows from our results. In fact, if $\mathcal{A}$ is any convex set (not just a polytope) such that $\inf_{P\in\mathcal{A}} D(P\|\mathcal{C}) < \infty$, then (as Zhang et al. (2024) point out) a similar result is already implied by Grünwald et al. (2024) as long as the infimum is achieved on the left. Indeed, they show that the likelihood ratio of the distribution that achieves the infimum and its RIPr is an $e$-statistic that has nontrivial $e$-power. This leaves the case that $\inf_{P\in\mathcal{A}} D(P\|\mathcal{C}) = \infty$. Indeed, the current work implies that also in this case, an $e$-statistic with nontrivial (in fact, infinite) $e$-power exists, as long as $\mathcal{A}$ is a convex polytope. That is, if we use $P^*$ to denote the uniform mixture of the vertices of $\mathcal{A}$, then for any vertex $P \in \mathcal{C}$, we have that

$$\int_\Omega \ln \frac{\mathrm{d}P^*}{\mathrm{d}\hat{Q}^*} \, \mathrm{d}P \geq \int_\Omega \ln \frac{\frac{1}{n}\mathrm{d}P}{\mathrm{d}\hat{Q}^*} \, \mathrm{d}P \geq D(P\|\mathcal{C}) - \ln(n),$$

where $\hat{Q}^*$ denotes the RIPr of $P^*$. It follows that $\inf_{P \in \mathcal{A}} \int_\Omega \ln \frac{\mathrm{d}P^*}{\mathrm{d}\hat{Q}^*} \, \mathrm{d}P = \infty$, so that the $e$-statistic given by the likelihood ratio of $P^*$ to its RIPr has "nontrivial $e$-power". However, more work is needed to determine whether such constructions are in any way optimal and whether the restriction that $\mathcal{A}$ is a convex polytope can be relaxed.

Second, after the first version of this manuscript was made available online, a follow-up paper appeared by Larsson et al. (2024). They show that, under no conditions on $P$ and $\mathcal{C}$ whatsoever, there exists an $e$-statistic $E^*$ that is the strongest $e$-statistic in the sense of Definition 3.11. This $e$-statistic, which they call 'the numeraire', gives rise to a measure $Q^*$ such that ${}^{\mathrm{d}Q^*}/_{\mathrm{d}P} = {}^{1}/_{E^*}$. Whenever the conditions of Theorem 3.5 hold, $Q^*$ coincides with the reverse information projection of $P$ on $\mathcal{C}$, so that it provides (in their words) "[...] a natural definition of the RIPr in the absence of any assumptions on $\mathcal{C}$ or $P$." We refer to their work (Larsson et al., 2024) for all further details.

## 3.5 Summary and Future Work

We have shown that, under very mild conditions, there exists a measure that achieves the minimax description gain over a convex set of measures $\mathcal{C}$ relative to a measure $P$. Whenever the information divergence between $P$ and $\mathcal{C}$ is finite, this measure coincides with the reverse information projection of $P$ on $\mathcal{C}$. As such, it provides a natural extension of the reverse information projection to cases where the the minimax description gain is finite, while the information divergence is infinite. In the context of hypothesis testing, this extended notion of the RIPr can be used to define an $e$-statistic for testing the simple alternative $P$ against the composite null $\mathcal{C}$. This $e$-statistic is optimal in a sense that is a natural, but novel extension of the previously known GRO optimality criterion for $e$-statistics. We have shown an example where GRO is unable to differentiate between $e$-statistics, whereas our novel criterion can, so that it is a strict extension. Additionally, we discussed an algorithm that can be used to approximate the reverse information projection in scenarios where it is not explicitly computable and show under what circumstances this also leads to an approximation of the optimal $e$-statistic.

The results presented thus far suggest various avenues for further research of which we discuss two. First, Theorem 3.5 is formulated for general measures so one may ask for an interpretation of the RIPr in the case that $P$ and $\mathcal{C}$ are not probability measures. If $\Omega$ is finite and $\lambda$ is a measure on $\Omega$, then we may define a probability measure $Po(\lambda)$ as the product measure $Po(\lambda) = \bigotimes_{\omega \in \Omega} Po\left(\lambda(\omega)\right)$, where $Po\left(\lambda(\omega)\right)$

denotes the Poisson distribution with mean $\lambda(\omega)$. With this definition we get

$$D(P\|Q \rightsquigarrow Q') = D(Po(P)\|Po(Q) \rightsquigarrow Po(Q')).$$

Furthermore, it can be shown that if the RIPr $\hat{Q}$ of $P$ on $\mathcal{C}$ exists and is an element of $\mathcal{C}$, then $Po(\hat{Q})$ is also the RIPr of $Po(P)$ on the convex hull of $\mathcal{C}' := \{Po(Q)|Q \in \mathcal{C}\}$. Consequently, ${}^{Po(P)}/_{Po(\hat{Q})}$ can be thought of as an $e$-statistic for $\mathcal{C}'$. More work is needed to determine whether this interpretation has any applications and if it can be generalized to arbitrary $\Omega$.

Second, even if $D(P\|\mathcal{C}) = \infty$, the Rényi divergence $D_\alpha(P\|Q)$ (see e.g. van Erven and Harremoës (2014)) may be a well-defined nonnegative real number for $\alpha \in (0, 1)$ and $Q \in \mathcal{C}$. These Rényi divergences are jointly convex in $P$ and $Q$ (van Erven and Harremoës, 2014) and for each $0 < \alpha < 1$ one may define a reverse Rényi projection $\hat{Q}_\alpha$ of $P$ on $\mathcal{C}$ (Kumar and Sason, 2016). Larsson et al. (2024) show that one may use this projection to define an e-statistic that is optimal for a polynomial rather than a logarithmic utility function — the theory is developed in completely analogous fashion to the logarithmic/standard Kullback-Leibler information case. We conjecture that the projections $\hat{Q}_\alpha$ will converge to the RIPr for $\alpha$ tending to 1, which might lead to further applications.

# 4 | k-Sample Tests With Exponential Families

The discussion of the correspondence between log-optimal $e$-variables and the reverse information projection in the previous chapter leads us to consider the problem of finding the reverse information projection in concrete settings. This is complicated because the reverse information projection is generally an element of the convex hull of the null hypothesis, so it could be any mixture of elements of the null. However, in this chapter we show that there are certain k-sample tests for which the reverse information projection lies in the null hypothesis itself, that is, is not a mixture.

In particular, we consider the problem of testing whether $k$ samples of data are drawn from the same element of an exponential family, the alternative being that they come from different elements of the same exponential family. We show that for some exponential families, there exists an $e$-variable that is a likelihood ratio between the alternative and an element of the null. The denominator of this likelihood ratio must be the reverse information projection by Proposition 3.14, so this $e$-variable is GRO. We also propose two other $e$-variables for when this is not the case, thus considering three $e$-variables in total: the GRO $e$-variables for (1) the null itself, and (2) a larger nonparametric null, as well as (3) an $e$-variable arrived at by conditioning on the sum of the sufficient statistics. (2) and (3) are always efficiently computable, and extend ideas from Turner et al. (2024) and Wald (1947) respectively from Bernoulli to general exponential families. We provide theoretical and simulation-based comparisons of these $e$-variables in terms of their logarithmic growth rate, and find that for small effects all four $e$-variables behave surprisingly similarly; for the Gaussian location and Poisson families, e-variables (1) and (3) coincide; for Bernoulli, (1) and (2) coincide; but in general, whether (2) or (3) grows faster under the alternative is family-dependent. Finally, we discuss algorithms for numerically approximating (1).

## 4.1 Introduction

E-variables (and the value they take, the *e-value*) provide an alternative to p-values that is inherently more suitable for testing under optional stopping and continuation, and that lies at the basis of *anytime-valid* confidence intervals that can be monitored continuously (Grünwald et al., 2024; Vovk and Wang, 2021; Shafer, 2021; Ramdas et al., 2023; Henzi and Ziegel, 2022; Grünwald, 2023). While they have their roots in the work on anytime-valid testing by H. Robbins and students (e.g. (Darling and Robbins, 1967)), they have begun to be investigated in detail for composite null hypotheses only very recently. E-variables can be associated with a natural notion of optimality, called GRO (growth-rate optimality), introduced and studied in detail by Grünwald et al. (2024). GRO may be viewed as an analogue of the uniformly most powerful test in an optional stopping context. In this chapter, we develop GRO and near-GRO $e$-variables for a classical statistical problem: parametric $k$-sample tests. Pioneering work in this direction appears already in Wald (1947): as we explain in Example 4.1, his SPRT for a sequential test of two proportions can be re-interpreted in terms of $e$-values for Bernoulli streams. Wald's $e$-values are not optimal in the GRO sense — GRO versions were derived only very recently by Turner et al. (2024); Turner and Grünwald (2023), but again only for Bernoulli streams. Here we develop $e$-variables for the case that the alternative is associated with an arbitrary but fixed exponential family, $\mathcal{M}$, with data in each of the $k$ groups sequentially sampled from a different distribution in that family. We mostly consider tests against the null hypothesis, denoted by $\mathcal{H}_0(\mathcal{M})$ that states that outcomes in all groups are i.i.d. by a single member of $\mathcal{M}$. We develop the GRO $e$-variable $S_{\text{GRO}(\mathcal{M})}$ for this null hypothesis, but it is not efficiently computable in general. Therefore, we introduce two more tractable $e$-variables: $S_{\text{GRO}(\text{IID})}$ and $S_{\text{COND}}$. The former is defined as the GRO $e$-variable, for the much larger null hypothesis that the $k$ groups are i.i.d. from an arbitrary distribution, denoted by $\mathcal{H}_0(\text{IID})$: since an e-variable relative to a null hypothesis $\mathcal{H}_0$ is automatically an e-variable relative to any null that is a subset of $\mathcal{H}_0$, $S_{\text{GRO}(\text{IID})}$ is automatically also an e-variable relative to $\mathcal{H}_0(\mathcal{M})$. Whenever below we refer to 'the null', we mean the smaller $\mathcal{H}_0(\mathcal{M})$. The use of $S_{\text{GRO}(\text{IID})}$ rather than $S_{\text{GRO}(\mathcal{M})}$ for this null, for which it is not GRO, is justifiable by ease of computation and robustness against misspecification of the model $\mathcal{M}$. However, exactly this robustness might also cause it to be too conservative when $\mathcal{M}$ is well-specified. The third $e$-variable we consider, $S_{\text{COND}}$, does not have any GRO status, but is specifically tailored to $\mathcal{H}_0(\mathcal{M})$, so that it might still be better than $S_{\text{GRO}(\text{IID})}$ in practice. Finally, we introduce a pseudo-$e$-variable

$S_{\text{PSEUDO}(\mathcal{M})}$, which coincides with $S_{\text{GRO}(\mathcal{M})}$ whenever the latter is easy to compute; in other cases it is not a real $e$-variable, but it is still highly useful for our theoretical analysis.

**Results** Besides defining $S_{\text{GRO}(\mathcal{M})}$, $S_{\text{GRO}(\text{IID})}$ and $S_{\text{COND}}$ and proving that they achieve what they purport to, we analyze their behavior both theoretically and by simulations. Our main theoretical results, Theorem 4.12 and 4.13 reveal some surprising facts: for any exponential family, the four types of (pseudo-) $e$-variables achieve almost the same growth rate under the alternative, hence are almost equally good, whenever the 'distance' between null and alternative is sufficiently small. That is, suppose that the (shortest) $\ell_2$-distance between the $k$ dimensional parameter of the alternative and the parameter space of the null is given by $\delta$. Then for any two of the aforementioned $e$-variables $S, S'$, we have $\mathbb{E}[\log S - \log S'] = O(\delta^4)$, where the expectation is taken under the alternative. Here, $\mathbb{E}[\log S]$ can be interpreted as the growth rate of $S$, as explained in Section 4.1.1.

While $S_{\text{GRO}(\text{IID})}$ and $S_{\text{COND}}$ are efficiently computable for the families we consider, this is generally not the case for $S_{\text{GRO}(\mathcal{M})}$, since to compute it we need to have access to the *reverse information projection* (RIPr; (Li, 1999; Grünwald et al., 2024)) of a fixed simple alternative to the set $\mathcal{H}_0(\mathcal{M})$. In general, this is a convex combination of elements of $\mathcal{H}_0(\mathcal{M})$, which can only be found by numerical means. Interestingly, we find that for three families, Gaussian with fixed variance, Bernoulli and Poisson, the RIPr is attained at a single point (i.e. a mixture putting all its mass on that point) that can be efficiently computed. Furthermore, in these cases $S_{\text{GRO}(\mathcal{M})}$ coincides with one of the other $e$-variables ($S_{\text{GRO}(\text{IID})}$ for Bernoulli, $S_{\text{COND}}$ for Gaussian and Poisson). For other exponential families, for $k = 2$, we approximate the RIPr and hence $S_{\text{GRO}(\mathcal{M})}$ using both an algorithm proposed by Li (1999) and a brute-force approach. We find that we can already get an extremely good approximation of the RIPr with a mixture of just *two* components. This leads us to conjecture that perhaps the deviation from the RIPr is just due to numerical imprecision and that the actual RIPr really can be expressed with just two components. The theoretical interest of such a development notwithstanding, we advise to use $S_{\text{COND}}$ or $S_{\text{GRO}(\text{IID})}$ rather than $S_{\text{GRO}(\mathcal{M})}$ for practical purposes whenever more than one component is needed for the RIPr, as their growth rates are not much worse, and they are much easier to compute. If furthermore robustness against misspecification of the null is required, then $S_{\text{GRO}(\text{IID})}$ is the most sensible choice.

**Method: Restriction to Single Blocks and Simple Alternatives** The main interest of $e$-variables is in analyzing sequential, anytime-valid settings: the data arrives in $k$ streams corresponding to $k$ groups, and we may want to stop or continue sampling at will (optional stopping); for example, we only stop when the data looks sufficiently good; or we stop unexpectedly, because we run out of money to collect new data. Nevertheless, in this chapter we focus on what happens in a single *block*, i.e. a vector $X^k = (X_1, \ldots, X_k)$, where each $X_j$ denotes a single outcome in the $j$-th stream. By now, there are a variety of papers (see e.g. Grünwald et al. (2024); Ramdas et al. (2023); Turner et al. (2024)) that explain how $e$-variables defined for such a single block can be combined by multiplication to yield e-processes (in our context, coinciding with *nonnegative supermartingales*) that can be used for testing the null with optional stopping if blocks arrive sequentially — that is, one observes one outcome of each sample at a time. Briefly, one multiplies the $e$-variables and at any time one intends to stop, one rejects the null if the product of $e$-values observed so-far exceeds $1/\alpha$ for pre-specified significance level $\alpha$. This gives an *anytime-valid* test at level $\alpha$: irrespective of the stopping rule employed, the type-I error is guaranteed to be below $\alpha$. Similarly, one can extend the method to design *anytime-valid confidence intervals* by inverting such tests, as described in detail by Ramdas et al. (2023). This is done for the 2-sample test with Bernoulli data by Turner and Grünwald (2023); their inversion methods are extendable to the general exponential family case we discuss here. Thus, we refer to the aforementioned papers for further details and restrict ourselves here to the 1-block case. Also, Turner et al. (2024); Turner and Grünwald (2022) describe how one can adapt an e-process for data arriving in blocks to general streams in which the $k$ streams do not produce data points at the same rate; we briefly extend their explanation to the present setting in Appendix B.1. Finally, we mainly restrict to the case of a simple alternative, i.e. a single member of the exponential family under consideration. While this may seem like a huge restriction, extension from simple to composite alternatives (e.g. the full family under consideration) is straightforward using the *method of mixtures* (i.e. Bayesian learning of the alternative over time) and/or the plug-in method. We again refer to Grünwald et al. (2024); Ramdas et al. (2023) for detailed explanations, and Turner et al. (2024) for an explanation in the 2-sample Bernoulli case, and restrict here to the simple alternative case: all the 'real' difficulty lies in dealing with composite null hypotheses, and that, we do explicitly and exhaustively in this chapter.

**Related Work and Practical Relevance** As indicated, this chapter is a direct (but far-reaching) extension of the papers Turner et al. (2024); Turner and Grünwald (2023) on 2-sample testing for Bernoulli streams as well as Wald's (1947) sequential two-sample test for proportions to streams coming from an exponential family. There are also *nonparametric* sequential (Lhéritier and Cazals, 2018) and anytime-valid 2-sample tests (Balsubramani and Ramdas, 2016; Pandeva et al., 2022) that tackle a somewhat different problem. They work under much weaker assumptions on the alternative (in some versions the samples could be arbitrary high-dimensional objects such as pictures and the like). The price to pay is that they will need a much larger sample size before a difference can be detected. Indeed, while our main interest is theoretical (how do different $e$-variables compare? in what sense are they optimal?), in settings where data are expensive, such as randomized clinical trials, the methods we describe here can be practically very useful: they are exact (existing methods are often based on chi-squared tests, which do not give exact type-I error guarantees at small sample size), they allow for optional stopping, and they need small amounts of data due to the strong parametric assumptions for the alternative. As a simple illustration of the practical importance of these properties, we refer to the recent SWEPIS study (Wennerholm et al., 2019) which was stopped early for harm. As demonstrated by Turner et al. (2024), if an anytime-valid two-sample test had been used in that study, substantially stronger conclusions could have been drawn.

We also mention that $k$-sample tests can be viewed as independence tests (is the outcome independent of the group it belongs to?) and as such this chapter is also related to recent papers on $e$-values and anytime-valid tests for conditional independence testing (Shaer et al., 2023; Duan et al., 2022); see also Chapter 6. Yet, the setting studied in those papers is quite different in that they assume the covariates (i.e. indicator of which of the $k$ groups the data belongs to) to be i.i.d.

**Contents** In the remainder of this introduction, we fix the general framework and notation and we briefly recall how $e$-variables are used in an anytime-valid/optional stopping setting. In Section 4.2 we describe our four (pseudo-) $e$-variables in detail, and we provide preliminary results that characterize their behavior in terms of growth rate. In Section 4.3 we provide our main theoretical results which show that, for all regular exponential families, the expected growth of the four types of $e$-variables is of surprisingly small order $\delta^4$ if the parameters of the alternative are at $\ell_2$-distance $\delta$ to the parameter space of the null. In Section 4.4 we give more detailed comparisons for a large number of standard exponential families (Gaussian, Bernoulli, Poisson,

exponential, geometric, beta), including simulations that show what happens if $\delta$ gets larger. Section 4.5 provides some additional simulations about the RIPr. All proofs, and some additional simulations, are in the appendix.

### 4.1.1 Formal Setting

Consider a regular one-dimensional exponential family $\mathcal{M} = \{P_\mu : \mu \in \mathtt{M}\}$ given in its mean-value parameterization (see e.g. (Barndorff-Nielsen, 1978) for more on definitions and for all the proofs of all standard results about exponential families that are to follow). Each member of the family is a distribution for some random variable $U$, taking values in some set $\mathcal{U}$, with density $p_{\mu;[U]}$ relative to some underlying measure $\rho_{[U]}$ which, without loss of generality, can be taken to be a probability measure. For regular exponential families, $\mathtt{M}$ is an open interval in $\mathbb{R}$ and $p_{\mu;[U]}$ can be written as:

$$p_{\mu;[U]}(U) = \exp\left(\lambda(\mu) \cdot t(U) - A(\lambda(\mu))\right), \tag{4.1}$$

where $\lambda(\mu)$ maps mean-value $\mu$ to canonical parameter $\beta$, $t(U)$ is a measurable function of $U$ and $A(\beta)$ is the log-normalizing factor. We furthermore have $\mu = \mathbb{E}_{P_\mu}[t(U)]$. The measure $\rho_{[U]}$ induces a corresponding (marginal) measure $\rho := \rho_{[X]}$ on the *sufficient statistic* $X := t(U)$, and similarly the density (4.1) induces a corresponding density $p_\mu := p_{\mu;[X]}$ on $X$, i.e. we have

$$p_\mu(X) := p_{\mu;[X]}(X) = \exp\left(\lambda(\mu) \cdot X - A(\lambda(\mu))\right). \tag{4.2}$$

All $e$-variables that we will define can be written in terms of the induced measure and density of the sufficient statistic of $X$; in other words, we can without loss of generality act as if our family is *natural*. Therefore, from now on we simply assume that we observe data in terms of their sufficient statistics $X$ rather than the potentially more fine-grained $U$, and will be silent about $U$; for simplicity we thus abbreviate $p_{\mu;[X]}$ to $p_\mu$ and $\rho_{[X]}$ to $\rho$. Note that exponential families are more usually defined with a carrier function $h(X)$ and $\rho$ set to Lebesgue or counting measure; we cover this case by absorbing $h$ into $\rho$, which we do not require to be Lebesgue or counting.

The data comes in as a block $X^k = (X_1, \ldots, X_k) \in \mathcal{X}^k$, where $\mathcal{X}$ is the support of $\rho$. To calculate our $e$-values we only need to know $X^k \in \mathcal{X}^k$, and under the alternative hypothesis, all $X_j$, $j = 1 \ldots k$ are distributed according to some element $P_{\mu_j}$ of $\mathcal{M}$. In our main results we take the alternative hypothesis to be *simple*, i.e. we assume that

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \mathtt{M}^k$ is fixed in advance. The alternative hypothesis is thus given by

$$\text{simple } \mathcal{H}_1 : X_1 \sim P_{\mu_1}, X_2 \sim P_{\mu_2}, \ldots, X_k \sim P_{\mu_k} \text{ independent.}$$

Note that we will keep $\boldsymbol{\mu}$ fixed throughout the rest of this section and Section 4.2. This is without loss of generality as $\boldsymbol{\mu}$ is defined as an arbitrary element of $\mathtt{M}^k$, so that all results stated for $\boldsymbol{\mu}$ hold for any element of $\mathtt{M}^k$. The extension to composite alternatives by means of the method of mixtures or the plug-in method is straightforward, and done in a manner that has become standard for $e$-value based testing (Ramdas et al., 2023).

Our null hypothesis is directly taken to be composite, for as regards the null, the composite case is inherently very different from the simple case (Ramdas et al., 2023; Grünwald et al., 2024). It expresses that the $X^k$ are identically distributed. We shall consider various variants of this null hypothesis, all composite: let $\mathcal{P}$ be a set of distributions on $\mathcal{X}$, then the null hypothesis *relative to* $\mathcal{P}$, denoted $\mathcal{H}_0(\mathcal{P})$, is defined as

$$\text{composite } \mathcal{H}_0(\mathcal{P}) : X_1 \sim P, X_2 \sim P, \ldots, X_k \sim P \text{ i.i.d. for some } P \in \mathcal{P}.$$

Our most important instantiation for the null hypothesis will be $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{M})$ for the same exponential family $\mathcal{M}$ from which the alternative was taken; then $\mathcal{H}_0(\mathcal{M})$ is a one-dimensional parametric family expressing that the $X_i$ are i.i.d. from $P_{\mu_0}$ for $\mu_0 \in \mathtt{M}$. Still, we will also consider $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{P})$ where $\mathcal{P}$ is the much larger set of *all* distributions on $\mathcal{X}$. Then the null simply expresses that the $X^k$ are i.i.d.; we shall abbreviate this null to $\mathcal{H}_0(\text{IID})$. Finally we sometimes consider $\mathcal{H}_0 = \mathcal{H}_0(\mathcal{M}')$ where $\mathcal{M}' \subset \mathcal{M}$ is a subset of $P_\mu \in \mathcal{M}$ with $\mu \in \mathtt{M}'$ for some sub-interval $\mathtt{M}' \subset \mathtt{M}$. The statistics that we use to gain evidence against these null hypotheses are $e$-variables.

**Definition 4.1.** We call any nonnegative random variable $S$ on a sample space $\Omega$ (which in this chapter will always be $\Omega = \mathcal{X}^k$) an *e-variable relative to* $\mathcal{H}_0$ if it satisfies

$$\text{for all } P \in \mathcal{H}_0 : \qquad \mathbb{E}_P[S] \leq 1. \tag{4.3}$$

### 4.1.2 The GRO E-Variable for General $\mathcal{H}_0$

In general, there exist many $e$-variables for testing any of the null hypotheses introduced above. Each $e$-variable $S$ can in turn be associated with a growth rate, defined by $\mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S]$. Roughly, this can be interpreted as the (asymptotic) exponential growth rate one would achieve by using $S$ in consecutive independent experiments and multiplying the outcomes if the (simple) alternative was true (see e.g. (Grünwald et al.,

2024, Section 2.1) or (Kelly, 1956)). The Growth Rate Optimal (GRO) $e$-variable is then the $e$-variable with the greatest growth rate among all $e$-variables. The central result (Theorem 1) of Grünwald et al. (2024) states that, under very weak conditions, GRO $e$-variables take the form of likelihood ratios between the alternative and the *reverse information projection* (Li, 1999) of the alternative onto the null. We instantiate their Theorem 1 to our setting by providing Lemma 4.2 and 4.4, both special cases of their Theorem 1. Before stating these, we need to introduce some more notation and definitions. For $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ we use the following notation:

$$p_{\boldsymbol{\mu}}(X^k) := \prod_{i=1}^{k} p_{\mu_i}(X_i).$$

Whenever in this text we refer to KL divergence $D(Q\|R)$, we refer to measures $Q$ and $R$ on $\mathcal{X}^k$. Here $Q$ is required to be a probability measure, while $R$ is allowed to be a sub-probability measure, as in (Grünwald et al., 2024). A *sub-* probability measure $R$ on $\mathcal{X}^k$ is a measure that integrates to 1 or less, i.e $\int_{x\in\mathcal{X}} dR(x) \leq 1$.

The following lemma follows as a very special case of Theorem 1 (simplest version) of Grünwald et al. (2024), when instantiated to our $k$-sample testing set-up:

**Lemma 4.2.** *Let $\mathcal{P}$ be a set of probability distributions on $\mathcal{X}^k$ and let* $\mathrm{conv}(\mathcal{P})$ *be its convex hull. Then there exists a sub-probability measure $P_0^*$ with density $p_0^*$ such that*

$$D(P_{\boldsymbol{\mu}}\|P_0^*) = \inf_{P\in\mathrm{conv}(\mathcal{P})} D(P_{\boldsymbol{\mu}}\|P). \tag{4.4}$$

*$P_0^*$ is called the* reverse information projection (RIPr) *of $P_{\boldsymbol{\mu}}$ onto* $\mathrm{conv}(\mathcal{P})$.

Clearly, if $P_0^* \in \mathrm{conv}(\mathcal{P})$ (the minimum is achieved) then $P_0^*$ is a probability measure, i.e. integrates to exactly one. We show that this happens for certain specific exponential families in Section 4.4. However, in general we can neither expect the minimum to be achieved, nor the RIPr to integrate to one. Lemma 4.4 below, again a special case of (Grünwald et al., 2024, Theorem 1), shows that the RIPr characterizes the GRO $e$-variable, and explains the use of the term GRO in the definition below.

**Definition 4.3.** $S_{\mathrm{GRO}(\mathcal{P})}$ is defined as

$$S_{\mathrm{GRO}(\mathcal{P})} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_0^*(X^k)} \tag{4.5}$$

where $p_0^*$ is the density of the RIPr of $P_{\boldsymbol{\mu}}$ onto $\mathrm{conv}(\mathcal{P})$.

**Lemma 4.4.** *For every set of distributions $\mathcal{P}$ on $\mathcal{X}$, $S_{GRO(\mathcal{P})}$ is an e-variable for $\mathcal{H}_0(\mathcal{P})$. Moreover, it is the GRO* (Growth-Rate-Optimal) *e-variable for $\mathcal{H}_0(\mathcal{P})$, i.e. it essentially uniquely achieves*

$$\sup_S \mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S]$$

*where the supremum ranges over all e-variables for $\mathcal{H}_0(\mathcal{P})$.*

Here, essential uniqueness means that any other GRO $e$-variable must be equal to $S_{\text{GRO}(\mathcal{P})}$ with probability 1 under $P_{\boldsymbol{\mu}}$. This in turn implies that the measure $P_0^*$ is in fact unique, as members of regular exponential families must have full support. Thus, once we have fixed our alternative and defined our null as $\mathcal{H}_0(\mathcal{P})$ for some set of distributions $\mathcal{P}$ on $\mathcal{X}$, the optimal (in the GRO sense) $e$-variable to use is the $S_{\text{GRO}(\mathcal{P})}$ $e$-variable as defined above.

## 4.2 The Four Types of E-Variables

In this section, we define our four types of $e$-variables; the definitions can be instantiated to any underlying 1-parameter exponential family. More precisely, we define three 'real' $e$-variables $S_{\text{GRO}(\mathcal{M})}, S_{\text{COND}}, S_{\text{GRO(IID)}}$ and one 'pseudo' $e$-variable $S_{\text{PSEUDO}(\mathcal{M})}$, a variation of $S_{\text{GRO}(\mathcal{M})}$ which for some exponential families is an $e$-variable, and for others is not.

### 4.2.1 The GRO E-Variable for $\mathcal{H}_0(\mathcal{M})$ and the pseudo *e*-variable

We now consider the GRO $e$-variable for our main null of interest, $\mathcal{H}_0(\mathcal{M})$. In practice, for some exponential families $\mathcal{M}$, the infimum over $\text{conv}(\mathcal{M})$ in (4.4) is actually achieved for some $P_{\mu_0^*} \in \mathcal{M}$. In this *easy* case we can determine $S_{\text{GRO}(\mathcal{M})}$ analytically (this happens if $S_{\text{GRO}(\mathcal{M})} = S_{\text{PSEUDO}(\mathcal{M})}$, see below). For all other $\mathcal{M}$, i.e. whenever the infimum is not achieved at all or is in $\text{conv}(\mathcal{M}) \setminus \mathcal{M}$, we do not know if $S_{\text{GRO}(\mathcal{M})}$ can be determined analytically. In this *hard* case will numerically approximate it by $S'_{\text{GRO}(\mathcal{M})}$ as defined below. First, for a fixed parameter $\mu_0 \in \mathtt{M}$ we define the vector $\langle \mu_0 \rangle$ as the vector indicating the distribution on $\mathcal{X}^k$ with all parameters equal to $\mu_0$:

$$\langle \mu_0 \rangle := (\mu_0, \ldots, \mu_0) \in \mathtt{M}^k \tag{4.6}$$

Next, with $W$ a distribution on $\mathtt{M}$, we define

$$p_W := \int p_{\langle \mu_0 \rangle}(X^k) dW(\mu_0) \tag{4.7}$$

to be the Bayesian marginal density obtained by marginalizing over distributions in $\mathcal{H}_0(\mathcal{M})$ according to $W$. Clearly, if $W$ has finite support then the corresponding distribution $P_W$ has $P_W \in \mathrm{conv}(\mathcal{M})$. We now set

$$S'_{\mathrm{GRO}(\mathcal{M})} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{W'_0}(X^k)}$$

where $W'_0$ is chosen so that $p_{W'_0}$ is within a small $\epsilon$ of achieving the minimum in (4.4), i.e. $D(P_{\mu_1,\ldots,\mu_k} \| P'_{W_0}) = \inf_{P \in \mathrm{conv}(\mathcal{M})} D(P_{\mu_1,\ldots,\mu_k} \| P) + \epsilon'$ for some $0 \leq \epsilon' < \epsilon$. Then, by Corollary 2 of Grünwald et al. (2024), $S'_{\mathrm{GRO}(\mathcal{M})}$ will *not* be an $e$-variable unless $\epsilon' = 0$, but in each case (i.e. for each choice of $\mathcal{M}$) we verify numerically that $\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\mu_0,\ldots,\mu_0}}[S] = 1 + \delta$ for negligibly small $\delta$, i.e. $\delta$ goes to 0 quickly as $\epsilon'$ goes to 0. We return to the details of the calculations in Section 4.5.

We now consider the 'easy' case in which $P_0^* = P_{\langle \mu_0^* \rangle}$ for some $\mu_0^* \in \mathtt{M}$. Clearly, we must have $\mu_0^* := \arg\min_{\mu_0 \in \mathtt{M}} D(P_{\boldsymbol{\mu}} \| P_{\langle \mu_0 \rangle})$. An easy calculation shows that then

$$\mu_0^* = \frac{1}{k} \sum_{i=1}^{k} \mu_i. \tag{4.8}$$

**Definition 4.5.** $S_{\mathrm{PSEUDO}(\mathcal{M})}$ is defined as

$$S_{\mathrm{PSEUDO}(\mathcal{M})} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0^* \rangle}(X^k)}.$$

$S_{\mathrm{PSEUDO}(\mathcal{M})}$ is not always a real $e$-variable relative to $\mathcal{H}_0(\mathcal{M})$, which explains the name 'pseudo'. Still, it will be very useful as an auxiliary tool in Theorem 4.12 and derivations. Note that, if it is an $e$-variable then we know that it is equal to $S_{\mathrm{GRO}(\mathcal{M})}$:

**Proposition 4.6.** *$S_{\mathrm{PSEUDO}(\mathcal{M})}$ is an $e$-variable for $\mathcal{M}$ iff $S_{\mathrm{PSEUDO}(\mathcal{M})} = S_{\mathrm{GRO}(\mathcal{M})}$.*

The proposition above does not give any easily verifiable condition to check if $S_{\mathrm{PSEUDO}(\mathcal{M})}$ is an $e$-variable or not. The following proposition does provide a condition which is sometimes easy to check (and which we will heavily employ below). With $\mu_0^*$

as in (4.8), define

$$f(\mu_0) := \sum_{i=1}^{k} \mathrm{VAR}_{P_{\mu_i+\mu_0-\mu_0^*}}[X] - k\mathrm{VAR}_{P_{\mu_0}}[X].$$

**Proposition 4.7.** *If $f(\mu_0^*) > 0$, then $S_{PSEUDO(\mathcal{M})}$ is not an e-variable. If $f(\mu_0^*) < 0$, then there exists an interval $\mathtt{M}' \subset \mathtt{M}$ with $\mu_0^*$ in the interior of $\mathtt{M}'$ so that $S_{PSEUDO(\mathcal{M})}$ is an e-variable for $\mathcal{H}_0(\mathcal{M}')$, where $\mathcal{M}' = \{P_\mu : \mu \in \mathtt{M}'\}$.*

### 4.2.2 The GRO E-Variable for $\mathcal{H}_0(\mathrm{IID})$

Recall that we defined $\mathcal{H}_0(\mathrm{IID})$ as the set of distributions under which $X_j$, $j = 1, \ldots k$, are i.i.d. from some arbitrary distribution on $\mathcal{X}$. By the defining property of *e*-variables, i.e. expected value bounded by one under the null (4.3), it should be clear that any *e*-variable for $\mathcal{H}_0(\mathrm{IID})$ is also an *e*-variable for $\mathcal{H}_0(\mathcal{M})$, since $\mathcal{H}_0(\mathcal{M}) \subset \mathcal{H}_0(\mathrm{IID})$. In particular, we can also use the GRO *e*-variable for $\mathcal{H}_0(\mathrm{IID})$ in our setting with exponential families. It turns out that this *e*-variable, which we will denote as $S_{\mathrm{GRO(IID)}}$, has a simple form that is generically easy to compute. We now show this:

**Theorem 4.8.** *The minimum KL divergence $\inf_{P \in \mathrm{conv}(\mathcal{H}_0(\mathit{IID}))} D(P_{\boldsymbol{\mu}} \| P)$ as in Lemma 4.2 is achieved by the distribution $P_0^*$ on $\mathcal{X}^k$ with density*

$$p_0^*(x^k) = \prod_{j=1}^{k} \frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(x_j).$$

*Hence, $S_{\mathit{GRO(IID)}}$, as defined below, is the GRO e-variable for $\mathcal{H}_0(\mathit{IID})$.*

**Definition 4.9.** $S_{\mathrm{GRO(IID)}}$ is defined as

$$S_{\mathrm{GRO(IID)}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} \left( \frac{1}{k} \sum_{i=1}^{k} p_{\mu_i}(X_j) \right)}.$$

The proof of Theorem 4.8 extends an argument of Turner et al. (2024) for the 2-sample Bernoulli case to the general $k$-sample case. The argument used in the proof does not actually require the alternative to equal the product distribution of $k$ independent elements of an exponential family — it could be given by the product of $k$ arbitrary distributions. However, we state the result only for the former case, as that is the setting we are interested in here.

### 4.2.3 The Conditional E-Variable

So far, we have defined $e$-variables as likelihood ratios between $P_{\boldsymbol{\mu}}$ and cleverly chosen elements of either $\mathcal{H}_0(\mathcal{M})$ or $\mathcal{H}_0(\text{IID})$. We now do things differently by not considering the full original data $X_1, \ldots X_k$, but instead conditioning on the sum of the sufficient statistics, i.e. $Z = \sum_{i=1}^k X_i$. It turns out that doing so actually collapses $\mathcal{H}_0(\mathcal{M})$ to a single distribution, so that the null becomes simple. That is, the distribution of $X^k \mid Z$ is the same under all elements of $\mathcal{H}_0(\mathcal{M})$, as we will prove in Proposition 4.11. This means that instead of using a likelihood ratio of the original data, we can use a likelihood ratio conditional on $Z$, which 'automatically' gives an $e$-variable.

**Definition 4.10.** Setting $Z$ to be the random variable $Z := \sum_{i=1}^k X_i$, $S_{\text{COND}}$ is defined as

$$S_{\text{COND}} := \frac{p_{\boldsymbol{\mu}}\left(X^{k-1} \mid Z\right)}{p_{\langle \mu_0 \rangle}\left(X^{k-1} \mid Z\right)},$$

with $\mu_0 \in \mathtt{M}$ and $(X)$ the sufficient statistic as in (4.2).

**Proposition 4.11.** *For all $\boldsymbol{\mu}' = (\mu'_1, \ldots, \mu'_k) \in \mathtt{M}^k$ , we have that $p_{\boldsymbol{\mu}'}(x^{k-1} \mid Z = z)$ depends on $\boldsymbol{\mu}'$ only through $\lambda_j := \lambda(\mu'_j) - \lambda(\mu'_k)$, $j = 1, \ldots k-1$, i.e. it can be written as a function of $(\lambda_1, \ldots, \lambda_{k-1})$. As a special case, for all $\mu_0, \mu'_0 \in \mathtt{M}$, it holds that $p_{\langle \mu_0 \rangle}(x^k \mid Z) = p_{\langle \mu'_0 \rangle}(x^k \mid Z)$. As a direct consequence, $S_{COND}$ is an e-variable for $\mathcal{H}_0(\mathcal{M})$,*

**Example 4.1. [The Bernoulli Model]** If $\mathcal{M}$ is the Bernoulli model and $k = 2$, then the conditional $e$-variable reduces to a ratio between the conditional probability of $(X_1, X_2) \in \{0,1\}^2$ given their sum $Z \in \{0,1,2\}$. Clearly, for all $\mu'_1, \mu'_2 \in \mathtt{M} = (0,1)$, we have $p_{\mu'_1,\mu'_2}((0,0) \mid Z = 0) = p_{\mu'_1,\mu'_2}((1,1) \mid Z = 2) = 1$, so $S_{\text{COND}} = 1$ whenever $Z = 0$ or $Z = 2$, irrespective of the alternative: data with the same outcome in both groups is effectively ignored. A nonsequential version of $S_{\text{COND}}$ for the Bernoulli model was analyzed earlier in great detail by Adams (2020).

Furthermore, for any $c \in \mathbb{R}$, we have that $\mathtt{M}_c := \{(\mu'_1, \mu'_2) : \lambda(\mu_1) - \lambda(\mu_2) = c\}$ is the line of distributions within $\mathtt{M}^2$ with the same odds ratio $\log(\mu_1(1-\mu_2)/((1-\mu_1)\mu_2)) = c$. The sequential probability ratio test of two proportions from Wald (1947) was based on fixing a $c$ for the alternative (viewing it as a notion of 'effect size') and analyzing sequences of paired data $X_{(1)}, X_{(2)}, \ldots$ with $X_{(i)} = (X_{i,1}, X_{i,2})$ by the product of conditional probabilities

$$\frac{p_c(X_{(i)} \mid Z_{(i)})}{p_0(X_{(i)} \mid Z_{(i)})} = S_{\text{COND}}(X_i),$$

thus effectively using $S_{\text{COND}}$ (here, we abuse notation slightly, writing $p_c(x \mid z)$ when we mean $p_{\mu_1',\mu_2'}(x \mid z)$ for any $\mu_1', \mu_2' \in \mathtt{M}_c$). It is, however, important to note that this product was not used for an anytime-valid test but rather for a classical sequential test with a fixed stopping rule especially designed to optimize power.

## 4.3 Growth Rate Comparison of Our E-Variables

Above we provided several recipes for constructing $e$-variables $S = S^{\boldsymbol{\mu}}$ whose definition implicitly depended on the chosen alternative $\boldsymbol{\mu}$. To compare these, we define, for any nonnegative random variables $S_1^{\boldsymbol{\mu}}$ and $S_2^{\boldsymbol{\mu}}$, $S_1^{\boldsymbol{\mu}} \succeq S_2^{\boldsymbol{\mu}}$ to mean that for all $\boldsymbol{\mu} \in \mathtt{M}^k$, it holds that $\mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S_1^{\boldsymbol{\mu}}] \geq \mathbb{E}_{P_{\boldsymbol{\mu}}}[\log S_2^{\boldsymbol{\mu}}]$. We write $S_1^{\boldsymbol{\mu}} \succ S_2^{\boldsymbol{\mu}}$ if $S_1^{\boldsymbol{\mu}} \succeq S_2$ and there exists $\boldsymbol{\mu} \in \mathtt{M}^k$ for which equality does not hold. From now on we suppress the dependency on $\boldsymbol{\mu}$ again, i.e. we write $S$ instead of $S^{\boldsymbol{\mu}}$. We trivially have, for every underlying exponential family $\mathcal{M}$,

$$S_{\text{PSEUDO}(\mathcal{M})} \succeq S_{\text{GRO}(\mathcal{M})} \succeq S_{\text{GRO(IID)}} \text{ and } S_{\text{GRO}(\mathcal{M})} \succeq S_{\text{COND}}. \tag{4.9}$$

We proceed with Theorem 4.12 and 4.13 below (proofs in the Appendix). These results go beyond the qualitative assessment above, by numerically bounding the difference in growth rate between $S_{\text{PSEUDO}(\mathcal{M})}$ and $S_{\text{GRO(IID)}}$ (and, because $S_{\text{GRO}(\mathcal{M})}$ must lie in between them, also between these two and $S_{\text{GRO}(\mathcal{M})}$) and $S_{\text{PSEUDO}(\mathcal{M})}$ and $S_{\text{COND}}$ respectively. Theorem 4.12 and 4.13 are asymptotic (in terms of difference between mean-value parameters) in nature. To give more precise statements rather than asymptotics we need to distinguish between individual exponential families; this is done in the next section.

To state the theorems, we need a notion of effect size, or discrepancy between the null and the alternative. So far, we have taken the alternative to be fixed and given by $\boldsymbol{\mu}$, but effect sizes are usually defined with the null hypothesis as starting point. To this end, note that each $P_{\langle\mu_0\rangle} \in \mathcal{H}_0(\mathcal{M})$ corresponds to a whole set of alternatives for which $P_{\langle\mu_0\rangle}$ is the closest point in KL within the null. This set of alternatives is parameterized by $\mathtt{M}^{(k)}(\mu_0) = \{\mu_1', \ldots, \mu_k' \in \mathtt{M} : \frac{1}{k}\sum_{i=1}^k \mu_i' = \mu_0\}$, as in (4.8). We can re-parameterize this set as follows, using the special notation $\langle\mu_0\rangle$ as given by (4.6). Let $\mathtt{A}$ be the set of unit vectors in $\mathbb{R}^k$ whose entries sum to 0, i.e. $\boldsymbol{\alpha} \in \mathtt{A}$ iff $\sqrt{\sum_{j=1}^k \alpha_j^2} = 1$ and $\sum_{j=1}^k \alpha_j = 0$. Clearly $\boldsymbol{\mu} \in \mathtt{M}^{(k)}(\mu_0)$ if and only if $\mu_1, \ldots, \mu_k \in \mathtt{M}$ and $\boldsymbol{\mu} = \langle\mu_0\rangle + \delta\boldsymbol{\alpha}$ for some scalar $\delta \geq 0$ and $\boldsymbol{\alpha} \in \mathtt{A}$. We can think of $\delta$ as expressing the magnitude of an effect and $\boldsymbol{\alpha}$ as its direction. Note that, if $k = 2$, then there

are only two directions, $\mathtt{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_{-1}\}$ with $\boldsymbol{a}_1 = (1/\sqrt{2}, -1/\sqrt{2})$ and $\boldsymbol{a}_{-1} = -\boldsymbol{a}_1$, corresponding to positive and negative effects: we have $\mu_1 - \mu_2 = \sqrt{2} \cdot \delta$ if $\boldsymbol{\alpha} = \boldsymbol{a}_1$ and $\mu_1 - \mu_2 = -\sqrt{2} \cdot \delta$ if $\boldsymbol{\alpha} = \boldsymbol{a}_{-1}$, as illustrated later on in Figure 4.1. Also note that, for general $k$, in the theorem below, we can simply interpret $\delta$ as the Euclidean distance between $\boldsymbol{\mu}$ and $\langle \mu_0 \rangle$.

**Theorem 4.12.** *Fix some $\mu_0 \in \mathtt{M}$, some $\boldsymbol{\alpha} \in \mathtt{A}$ and let $\boldsymbol{\mu} = \langle \mu_0 \rangle + \delta \boldsymbol{\alpha}$ for $\delta \geq 0$ such that $\boldsymbol{\mu} \in \mathtt{M}^{(k)}(\mu_0)$. The difference in growth rate between $S_{PSEUDO(\mathcal{M})}$ and $S_{GRO(IID)}$ is given by*

$$\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{PSEUDO(\mathcal{M})} - \log S_{GRO(IID)}\right] = \frac{1}{8} \int_x \frac{\left(f_x''(0)\right)^2}{f_x(0)} d\rho(x) \cdot \delta^4 + o\left(\delta^4\right) = O\left(\delta^4\right), \tag{4.10}$$

*where $f_x(\delta) = \sum_{i=1}^k p_{\mu_0 + \delta\alpha_i}(x) = \sum_{i=1}^k p_{\mu_i}(x)$ and $f_x''$ is the second derivative of $f_x$, so that $f_x(0) = k p_{\mu_0}(x)$ and (with some calculation) $f_x''(0) = \frac{d^2}{d\mu^2} p_\mu(x) \mid_{\mu=\mu_0}$.*

As is implicit in the $O(\cdot)$-notation, the expectation on the left is well-defined and finite and the integral in the middle equation is finite as well. The theorem implies that for general exponential families, $S_{\text{GRO(IID)}}$ is surprisingly close ($O(\delta^4)$) to the optimal $S_{\text{GRO}(\mathcal{M})}$ in the GRO sense, whenever the distance $\delta$ between $\mathcal{H}_1$ and $\mathcal{H}_0(\mathcal{M})$ is small. This means that, whenever $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{PSEUDO}(\mathcal{M})}$ (so $S_{\text{GRO}(\mathcal{M})}$ is hard to compute and $S_{\text{PSEUDO}(\mathcal{M})}$ is not an $e$-variable), we might consider using $S_{\text{GRO(IID)}}$ instead: it will be more robust (since it is an $e$-variable for the much larger hypothesis $\mathcal{H}_0(\text{IID})$) and it will only be slightly worse in terms of growth rate.

Theorem 4.12 is remarkably similar to the next theorem, which involves $S_{\text{COND}}$ rather than $S_{\text{GRO(IID)}}$. To state it, we first set $X_k(x^{k-1}, z) := z - \sum_{i=1}^{k-1} x_i$, and we denote the marginal distribution of $Z = \sum_{i=1}^k X_i$ under $P_{\boldsymbol{\mu}}$ as $P_{\boldsymbol{\mu};[Z]}$, noting that its density $p_{\boldsymbol{\mu};[Z]}$ is given by

$$p_{\boldsymbol{\mu};[Z]}(z) = \int_{\mathcal{C}(z)} p_{\boldsymbol{\mu}}\left(x^{k-1}, x_k\right) d\rho(x^{k-1}), \tag{4.11}$$

where $\rho$ is extended to the product measure of $\rho$ on $\mathcal{X}^{k-1}$ and

$$\mathcal{C}(z) := \left\{x^{k-1} \in \mathcal{X}^{k-1} : X_i(x^{k-1}, z) \in \mathcal{X}\right\}. \tag{4.12}$$

**Theorem 4.13.** *Fix some $\mu_0 \in \mathtt{M}$, $\boldsymbol{\alpha} \in \mathtt{A}$ and let $\boldsymbol{\mu} = \langle \mu_0 \rangle + \delta \boldsymbol{\alpha}$ for $\delta \geq 0$ such that*

*$\boldsymbol{\mu} \in \mathrm{M}^{(k)}(\mu_0)$. The difference in growth rate between $S_{PSEUDO(\mathcal{M})}$ and $S_{COND}$ is given by*

$$\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{PSEUDO(\mathcal{M})} - \log S_{COND}\right] = \frac{1}{8}\int_z \frac{\left(g_z''(0)\right)^2}{g_z(0)} d\rho_{[Z]}(z) \cdot \delta^4 + o\left(\delta^4\right) = O(\delta^4), \tag{4.13}$$

*where $g_z(\delta) := p_{\langle\mu_0\rangle + \boldsymbol{\alpha}\delta;[Z]}(z)$ and $\rho_{[Z]}$ denotes the measure on $Z$ induced by the product measure of $\rho$ on $\mathcal{X}^k$; an explicit expression for $g_z''(0)$ is*

$$\int_{\mathcal{C}(z)} p_{\langle\mu_0\rangle}\left(x^k\right) \sum_{j=1}^{k} \left[I'(\mu_0)(x_j - \mu_0) - I(\mu_0)\right] d\rho(x^{k-1}),$$

*where $I(\mu)$ denotes the Fisher information for $\mu$ and $I'(\mu)$ is its first derivative.*

Again, the expectation on the left is well-defined and finite and the integral on the right is finite. Comparing Theorem 4.13 to Theorem 4.12, we see that $f_x(0)$, the sum of $k$ identical densities evaluated at $x$, is replaced by $g_z(0)$, the density of the sum of $k$ i.i.d. random variables evaluated at $z$.

**Corollary 4.14.** *With the definitions as in the two theorems above, the growth-rate difference $\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{COND} - \log S_{GRO(IID)}\right]$ can be written as*

$$\frac{1}{8}\left(\int_x \frac{\left(f_x''(0)\right)^2}{f_x(0)} d\rho(x) - \int_z \frac{\left(g_z''(0)\right)^2}{g_z(0)} d\rho_{[Z]}(z)\right) \cdot \delta^4 + o\left(\delta^4\right) = O\left(\delta^4\right). \tag{4.14}$$

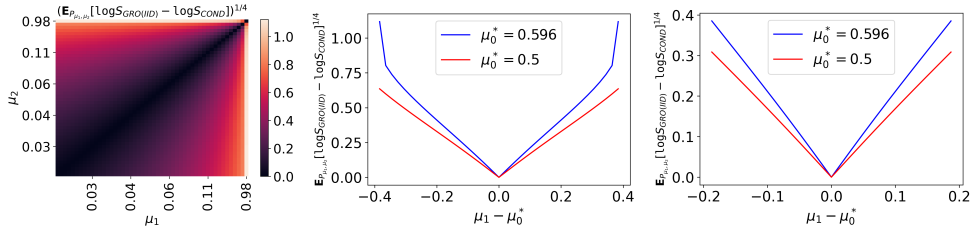
## 4.4 Growth Rate Comparison for Specific Exponential Families

We will now establish more precise relations between the four (pseudo-) $e$-variables in $k$-sample tests for several standard exponential families, namely those listed in Table 4.1 and a few related ones, as listed at the end of this section. For each family $\mathcal{M}$ under consideration, we give proofs for which different $e$-variables are the same, i.e. $S = S'$, where $S, S' \in \{S_{\mathrm{GRO}(\mathcal{M})}, S_{\mathrm{COND}}, S_{\mathrm{GRO(IID)}}, S_{\mathrm{PSEUDO}(\mathcal{M})}\}$. Whenever we can prove that $S_{\mathrm{GRO}(\mathcal{M})} \neq S$ for another $e$-variable $S \in \{S_{\mathrm{COND}}, S_{\mathrm{GRO(IID)}}\}$, we can infer that $S_{\mathrm{GRO}(\mathcal{M})} \succ S$ because $S_{\mathrm{GRO}(\mathcal{M})}$ is the GRO $e$-variable for $\mathcal{H}_0(\mathcal{M})$. Whenever both $S_{\mathrm{COND}}$ and $S_{\mathrm{GRO(IID)}}$ are not equal to $S_{\mathrm{GRO}(\mathcal{M})}$, we will investigate via simulation whether $S_{\mathrm{GRO(IID)}} \succ S_{\mathrm{COND}}$ or vice versa — our theoretical results do not extend to this case. All simulations are carried out for the case $k = 2$ in the chapter. Theorem 4.12 and

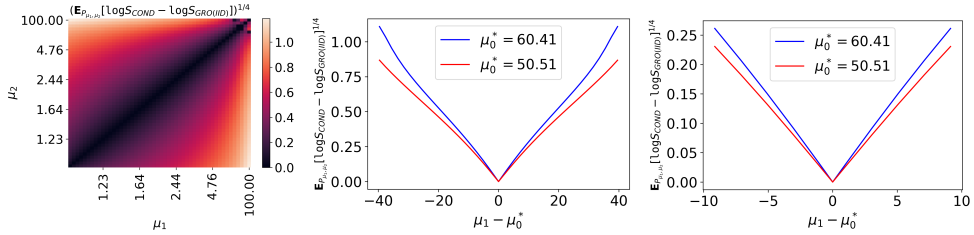
Theorem 4.13 show that in the neighborhood of $\delta = 0$ ($\mu_1, \ldots, \mu_k$ all close together), the difference $\mathbb{E}_{P_\mu}[\log S - \log S']$ is of order $\delta^4$ when $S, S' \in \{S_{\text{GRO}(\mathcal{M})}, S_{\text{PSEUDO}(\mathcal{M})}, S_{\text{GRO(IID)}}, S_{\text{COND}}\}$. Hence in the figures we will show $(\mathbb{E}_{P_\mu}[\log S - \log S'])^{1/4}$, since then we expect the distances to increase linearly as we move away from the diagonal, making the figures more informative.

Our findings, proofs as well as simulations, are summarised in Table 4.1. For each exponential family, we list the rank of the (pseudo-)$e$-variables when compared with the order '$\succ$'. The ranks that are written in black are proven in Appendix B.4, while the ranks in blue are merely conjectures based on our simulations as stated above. The results of the simulations on which these conjectures are based are given in Figure 4.1. Furthermore, the rank of $S_{\text{PSEUDO}(\mathcal{M})}$ is colored red whenever it is not an $e$-variable for that model, as shown in the Appendix. Note that whenever any of the $e$-variables have the same rank, they must be equal $\rho$-almost everywhere, by strict concavity of the logarithm together with full support of the distributions in the exponential family. For example, the results in the table reflect that for the Bernoulli family, we have shown that $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})} = S_{\text{GRO(IID)}}$ and that $S_{\text{PSEUDO}(\mathcal{M})} \succ S_{\text{COND}}$. Also, for the geometric family and beta with free $\beta$ and fixed $\alpha$, we have proved that $S_{\text{PSEUDO}(\mathcal{M})}$ is not an $e$-variable, that $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{GRO(IID)}}$ and that $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{COND}}$, so that it follows from (4.9) that $S_{\text{PSEUDO}(\mathcal{M})} \succ S_{\text{GRO}(\mathcal{M})}$, $S_{\text{GRO}(\mathcal{M})} \succ S_{\text{GRO(IID)}}$ and $S_{\text{GRO}(\mathcal{M})} \succ S_{\text{COND}}$. Then the findings of the simulations shown in Figure 4.1a suggest that $S_{\text{GRO(IID)}} \succ S_{\text{COND}}$ for beta with free $\beta$ and fixed $\alpha$ and in Figure 4.1b suggest that $S_{\text{COND}} \succ S_{\text{GRO(IID)}}$ for geometric family, but these are not proven. Figure 4.1c shows that $S_{\text{GRO(IID)}} \succ S_{\text{COND}}$ for Gaussians with free variance and fixed mean. Finally, Figure 4.1d shows that for the exponential, there is no clear relation between $S_{\text{GRO(IID)}}$ and $S_{\text{COND}}$. That is, $S_{\text{GRO(IID)}}$ grows faster than $S_{\text{COND}}$ for some $\mu_1, \ldots, \mu_k \in \mathtt{M}$, and slower for others, which is indicated by rank (3) – (4) in the table.

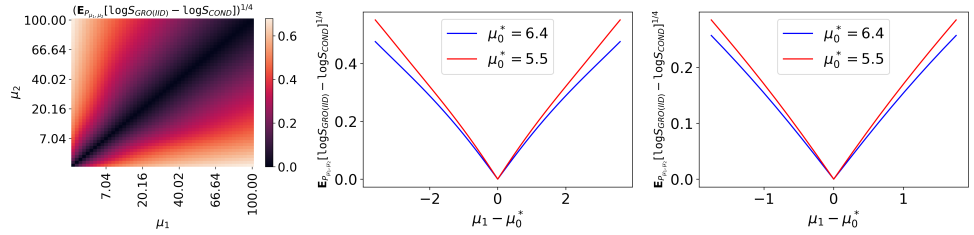
Finally, we note that for each family listed in the table, the results must extend to any other family that becomes identical to it if we reduce it to the natural form (4.2). For example, the family of Pareto distributions with fixed minimum parameter $v$ can be reduced to that of the exponential distributions: if $U \sim \text{Pareto}(v, \alpha)$, then we can do a transformation $X = t(U)$ with $t(U) = \log(U/v)$, and then $X \sim \text{Exp}(\alpha)$. Thus, the $k$-sample problem for $U$ with the Pareto$(v, \alpha)$ distributions, with $\alpha$ as free parameter, is equivalent to the $k$-sample problem for $X$ with the exponential distributions; the $e$-value $S_{\text{GRO}(\mathcal{M})}$ obtained with a particular alternative in the Pareto setting for observation $U$ coincides with $S_{\text{GRO}(\mathcal{M})}$ for the corresponding alternative in the exponential setting for observation $X = t(U)$, and the same holds for $S_{\text{GRO(IID)}}$ and $S_{\text{COND}}$.



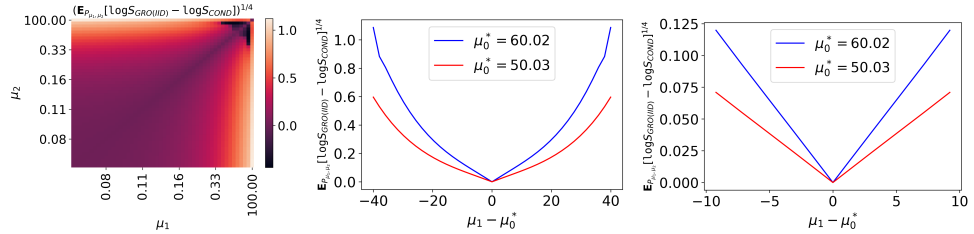
**(a)** beta with free $\beta$ and fixed $\alpha$



**(b)** geometric



**(c)** Gaussian with free variance and fixed mean



**(d)** Exponential

**Figure 4.1:** A comparison of $S_{\mathrm{GRO(IID)}}$ and $S_{\mathrm{COND}}$ for four exponential families. We evaluated the expected growth difference on a grid of $50 \times 50$ alternatives $(\mu_1, \mu_2)$, equally spaced in the standard parameterization (explaining the nonlinear scaling on the depicted mean-value parameterization). On the left are the corresponding heatmaps. On the right are diagonal 'slices' of these heatmaps: the red curve corresponds to the main diagonal (top left - bottom right), the blue curve corresponds to the diagonal starting from the second tick mark (10th discretization point) top left until the second tick mark bottom right. These slices are symmetric around 0, their value only depending on $\delta = \mid \mu_1 - \mu_2 \mid / \sqrt{2} = \mid \mu_1 - \mu_0^* \mid \cdot \sqrt{2}$, where $\mu_0^* = (\mu_1 + \mu_2)/2$ and $\delta$ is as in Theorem 4.12

| Exponential Family | $S_{\text{PSEUDO}(\mathcal{M})}$ | $S_{\text{GRO}(\mathcal{M})}$ | $S_{\text{GRO(IID)}}$ | $S_{\text{COND}}$ |
|---|---|---|---|---|
| Bernoulli | (1) | (1) | (1) | (2) |
| Gaussian with free mean and fixed variance | (1) | (1) | (2) | (1) |
| Poisson | (1) | (1) | (2) | (1) |
| beta with free $\beta$ and fixed $\alpha$ | (1) | (2) | (3) | (4) |
| geometric | (1) | (2) | (4) | (3) |
| Gaussian with free variance and fixed mean | (1) | (2) | (3) | (4) |
| Exponential | (1) | (2) | (3)-(4) | (3)-(4) |

**Table 4.1:** The ranks of the four different $e$-variables when compared with the relation '$\succ$'. The ranks in black are proved in Appendix B.4, while the ranks in blue are conjectures based on the simulations in Figure 4.1. The rank of $S_{\text{PSEUDO}(\mathcal{M})}$ is denoted in red whenever it is not an $e$-variable, as shown in Appendix B.4

Therefore, the ordering for Pareto must be the same as the ordering for exponential in Table 4.1. Similarly, the $e$-variables for the log-normal distributions (with free mean or variance) can be reduced to the two corresponding normal distribution $e$-variables.

## 4.5 Simulations to Approximate the RIPr

Because of its growth optimality property, we may sometimes still want to use the GRO $e$-variable $S_{\text{GRO}(\mathcal{M})}$, even in cases where it is not equal to the easily calculable $S_{\text{PSEUDO}(\mathcal{M})}$. To this end we need to approximate it numerically. The goal of this section is twofold: first, we want to illustrate that this is feasible in principle; second, we show that this raises interesting additional questions for future work. Thus, below we consider in more detail simulations to approximate $S_{\text{GRO}(\mathcal{M})}$ for the exponential families with $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{PSEUDO}(\mathcal{M})}$ that we considered before, i.e. beta, geometric, exponential and Gaussian with free variance; for simplicity we only consider the case $k = 2$. In Appendix B.5 we provide some graphs illustrating the RIPr probability densities for particular choices of $\mu_1, \mu_2$; here, we focus on how to approximate them, taking our findings for $k = 2$ as suggestive for what happens with larger $k$.

### 4.5.1 Approximating the RIPr via Li's Algorithm

Li (1999) provides an algorithm for approximating the RIPr of distribution $Q$ with density $q$ onto the convex hull $\text{conv}(\mathcal{P})$ of a set of distributions $\mathcal{P}$ (where each $P \in \mathcal{P}$ has density $p$) arbitrarily well in terms of KL divergence. At the $m$-th step, this algorithm outputs a finite mixture $P_{(m)} \in \text{conv}(\mathcal{P})$ of at most $m$ elements of $\mathcal{P}$. For $m > 1$, these mixtures are determined by iteratively setting $P_{(m)} := \alpha P_{(m-1)} + (1 -$

$\alpha)P'$, where $\alpha \in [0,1]$ and $P' \in \mathcal{P}$ are chosen so as to minimize KL divergence $D(Q\|\alpha P_{(m-1)} + (1-\alpha)P')$. Here, $P_{(1)}$ is defined as the single element of $\mathcal{P}$ that minimizes $D(Q\|P_{(1)})$. It is thus a greedy algorithm, but Li shows that, under some regularity conditions on $\mathcal{P}$, it holds that $D(Q\|P_{(m)}) \to \inf_{P\in\text{conv}(\mathcal{P})} D(Q\|P)$. That is, $P_{(m)}$ approximates the RIPr in terms of KL divergence. This suggests, but is not in itself sufficient to prove, that $\sup_{P\in\mathcal{P}} \mathbb{E}_P[q(X)/p_{(m)}(X)] \to 1$, i.e. that the likelihood ratio actually tends to an $e$-variable.

We numerically investigated whether this holds for our familiar setting with $k = 2$, $Q$ is equal to $P_{\boldsymbol{\mu}}$ for some $\boldsymbol{\mu} = (\mu_1, \mu_2) \in \mathtt{M}^2$, and $\mathcal{P} = \mathcal{H}_0(\mathcal{M})$. To this end, we applied Li's algorithm to a wide variety of values $(\mu_1, \mu_2)$ for the beta, exponential, geometric and Gaussian with free variance. In all these cases, after at most $m = 15$ iterations, we found that $\sup_{\mu_0\in\mathtt{M}} \mathbb{E}_{P_{\mu_0,\mu_0}}[p_{\mu_1,\mu_2}(X_1,X_2)/q_{(m)}(X_1,X_2)]$ was bounded by 1.005: Li's algorithm convergences quite fast; see Appendix B.5 for a graphical depiction of the convergence and design choices in the simulation.

(note that, since we have proved that $S_{\text{GRO}(\mathcal{M})} = S_{\text{PSEUDO}(\mathcal{M})}$ for Bernoulli, Poisson and Gaussian with free mean, there is no need to approximate $S_{\text{GRO}(\mathcal{M})}$ for those families).

### 4.5.2 Approximating the RIPr via Brute Force

While Li's algorithm converges quite fast, it is still highly suboptimal at iteration $m = 2$, due to its being greedy. This motivated us to investigate how 'close' we can get to an e-variable by using a mixture of just two components. Thus, we set $p_A(x^k) := \alpha p_{\langle\mu_{01}\rangle}(x^k) + (1-\alpha)p_{\langle\mu_{02}\rangle}(x^k)$ and, for various choices of $\boldsymbol{\mu} = (\mu_1, \mu_2)$, considered

$$S_{\text{APPR}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{p_A(X^k)} \tag{4.15}$$

as an approximate e-variable, for the specific values of $\alpha \in [0,1]$ and $\mu_{01}, \mu_{02}$ that minimize

$$\sup_{\mu_0\in\mathtt{M}} \mathbb{E}_{P_{\langle\mu_0\rangle}}[S_{\text{APPR}}].$$

(in practice, we maximize $\mu_0$ over a discretization of $\mathtt{M}$ with 1000 equally spaced grid points and minimize $\alpha, \mu_{01}, \mu_{02}$ over a grid with 100 equally sized grid points, with left- and right- end points of the grids over $\mathtt{M}$ determined by trial and error).

The simulation results, for $k = 2$ and particular values of $\mu_1, \mu_2$ and the exponential families for which approximation makes sense (i.e. $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{PSEUDO}(\mathcal{M})}$) are presented in Table 4.2. We tried, and obtained similar results, for many more

| Distributions | $(\mu_1, \mu_2)$ | $\alpha$ | $(\mu_{01}, \mu_{02})$ | $\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{X_1, X_2 \sim P_{\mu_0, \mu_0}}[S_{\text{APPR}}]$ |
|---|---|---|---|---|
| beta | $(0.5, 0.25)$ | 0.22 | (0.24, 0.81) | 1.0052 |
| Exponential | $(0.5, 0.25)$ | 0.56 | (0.35, 0.51) | 1.0000 |
| Gaussian with free variance and fixed mean | $(0.5, 0.25)$ | 0.37 | (0.5, 0.5) | 1.0000 |
| Exponential | $(\frac{10}{3}, \frac{5}{4})$ | 0.51 | (0.62, 0.31) | 1.0047 |
| geometric | $(\frac{10}{3}, \frac{5}{4})$ | 0.47 | (1.84, 2.97) | 1.0008 |
| Gaussian with free variance and fixed mean | $(\frac{10}{3}, \frac{5}{4})$ | 0.08 | (3.64, 2.73) | 1.0002 |

**Table 4.2:** For given values of $\boldsymbol{\mu} = (\mu_1, \mu_2)$, we show $\alpha, \mu_{01}$ and $\mu_{02}$ for the corresponding two-component mixture $\alpha p_{\mu_{01}}(X_1) p_{\mu_{01}}(X_2) + (1-\alpha) p_{\mu_{02}}(X_1) p_{\mu_{02}}(X_2)$ arrived at by brute-force minimization of the KL divergence as in Section 4.5.2, and we show how close the corresponding likelihood ratio $S_{\text{APPR}}$ is to being an e-variable

parameter values; one more parameter pair for each family is given in Table B.1 in Appendix B.5. The term $\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\langle \mu_0 \rangle}}[S_{\text{APPR}}]$ is remarkably close to 1 for all of these families. Corollary 2 of Grünwald et al. (2024) implies that if the supremum *is* exactly 1, i.e. $S_{\text{APPR}}$ is an e-variable, then $S_{\text{APPR}}$ must also be the GRO e-variable relative to $P_{\boldsymbol{\mu}}$. This leads us to speculate that perhaps all the exceedance beyond 1 is due to discretization and numerical error, and the following might (or might not — we found no way of either proving or disproving the claim) be the case:

**Conjecture** For $k = 2$, the RIPr, i.e. the distribution achieving

$$\min_{Q \in \text{conv}(\mathcal{H}_0(\mathcal{M}))} D(P_{\mu_1, \mu_2} \| Q)$$

can be written as a mixture of just two elements of $\mathcal{H}_0(\mathcal{M})$.

## 4.6 Conclusion and Future Work

In this chapter, we introduced and analyzed four types of $e$-variables for testing whether $k$ groups of data are distributed according to the same element of an exponential family. These four $e$-variables include: the GRO $e$-variable ($S_{\text{GRO}(\mathcal{M})}$), a conditional $e$-variable ($S_{\text{COND}}$), a mixture $e$-variable ($S_{\text{GRO}(\text{IID})}$), and a pseudo-$e$-variable ($S_{\text{PSEUDO}(\mathcal{M})}$). We compared the growth rate of the $e$-variables under a simple alternative where each of the $k$ groups has a different, but fixed, distribution in the same exponential family. We have shown that for any two of the $e$-variables

$S, S' \in \{S_{\text{GRO}(\mathcal{M})}, S_{\text{COND}}, S_{\text{GRO(IID)}}, S_{\text{PSEUDO}(\mathcal{M})}\}$, we have $\mathbb{E}[\log S - \log S'] = O(\delta^4)$ if the $\ell_2$ distance between the parameters of this alternative distribution and the parameter space of the null is given by $\delta$. This shows that when the effect size is small, all the $e$-variables behave surprisingly similar. For more general effect sizes, we know that $S_{\text{GRO}(\mathcal{M})}$ has the highest growth rate by definition. Calculating $S_{\text{GRO}(\mathcal{M})}$ involves computing the reverse information projection of the alternative on the null, which is generally a hard problem. However, we proved that there are exponential families for which one of the following holds $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})}$, $S_{\text{COND}} = S_{\text{GRO}(\mathcal{M})}$ or $S_{\text{GRO(IID)}} = S_{\text{GRO}(\mathcal{M})}$, which considerably simplifies the problem. If one is interested in testing an exponential family for which is not the case, there are algorithms to estimate the reverse information projection. We have numerically verified that approximations of the reverse information projection also lead to approximations of $S_{\text{GRO}(\mathcal{M})}$. However, the use of $S_{\text{COND}}$ or $S_{\text{GRO(IID)}}$ might still be preferred over $S_{\text{GRO}(\mathcal{M})}$ due to the computational advantage. Our simulations show that depends on the specific exponential family which of them is preferable over the other, and that sometimes there is even no clear order.

# 5 | Simple E-Variables for Exponential Families

In the previous chapter, we observed that in certain k-sample tests, there exists an $e$-variable in the form of a simple-vs.-simple likelihood ratio—a likelihood ratio between the alternative and a single element of the null. Equivalently, the reverse information projection is sometimes a distinct element of the null hypothesis rather than a convex combination. In such cases, the GRO $e$-variable is easy to compute. Moreover, while it was not mentioned explicitly in the previous chapter, Koolen and Grünwald (2022, Theorem 12) show that the cumulative product of such 'simple' $e$-variables is also the log-optimal $e$-process, further justifying their use for applications where the sample size is not fixed beforehand.

In this chapter, we provide a general condition under which simple $e$-variables exist when the null hypothesis is a composite, multivariate exponential family. Simple e-variables were previously only known to exist in quite specific settings, but we offer a unifying theorem on their existence for testing exponential families. We start with a simple alternative $Q$ and a regular exponential family null. Together these induce a second exponential family $\mathcal{Q}$ containing $Q$, with the same sufficient statistic as the null. Our theorem shows that simple e-variables exist whenever the covariance matrices of $\mathcal{Q}$ and the null are in a certain relation. A prime example in which this relation holds is testing whether a parameter in a linear regression is 0. Other examples include some $k$-sample tests, Gaussian location- and scale tests, and tests for more general classes of natural exponential families. While in all these examples, the implicit composite alternative is also an exponential family, this is not required in general.

## 5.1 Introduction

Exponential families play a central role in statistical modeling, as they include the Bernoulli-, Gaussian-, Poisson-, and many more models. An important task is to test whether these models are well-specified, that is, whether observed data are indeed distributed by an element of an exponential family; or more specifically whether a specific parameter in an exponential family is 0 or not — the latter including linear regression testing as a special case. Many classic tests are well-suited for this purpose (Anderson and Darling, 1954; Lilliefors, 1967; Stephens, 1974). However, the vast majority of these methods are based on p-values, and thus designed for fixed sample size experiments. Here, we are instead interested in hypothesis tests that are based on e-values (Grünwald et al., 2024), which is the value taken by an e-variable. The most straightforward example of e-variables are likelihood ratios between simple alternatives and simple null hypotheses. E-variables for composite hypotheses, and in particular 'good' e-variables, are generally more complicated. However, e-variables in the form of a likelihood ratio with a single, special element of the null representing the full, composite null sometimes still exist. We refer to such e-variables as 'simple' e-variables. As we shall see below, their existence is intimately tied to properties of the reverse information projection (RIPr).

Simple e-variables, if they exist, can easily be computed, and are known to be optimal in an expected-log-optimality sense (Koolen and Grünwald, 2022; Grünwald et al., 2024). That is, if we combine evidence from a repeated experiment where data is collected using a fixed stopping rule, then using the simple e-variable will asymptotically result in the most evidence against the null, among all e-variables; details can be found in Section 5.1.4. As such, it is desirable to find out whether or not simple e-variables exist in specific settings. The main result of this chapter, Theorem 5.3, provides a set of equivalent conditions under which simple e-variables exist for exponential family nulls.

### 5.1.1 Main Result and Overview

Here we briefly describe Theorem 5.3, assuming prior knowledge on e-variables and exponential families, and we provide an overview of the rest of the chapter — all relevant definitions and explanations are given in Section 5.1.2–5.1.4. We fix a regular multivariate exponential family null $\mathcal{P}$ for data $U$ with some sufficient statistic vector $X = t(U)$ and a distribution $Q$ for $U$, outside of $\mathcal{P}$, and with density $q$. As our

most important regularity condition, we assume that $Q$ has a moment generating function and that there exists $P_{\boldsymbol{\mu}^*} \in \mathcal{P}$ with the same mean of $X$, say $\boldsymbol{\mu}^*$, as $Q$. It is known that $P_{\boldsymbol{\mu}^*}$ is the *Reverse Information Projection (RIPr)* of $Q$ onto $\mathcal{P}$ (Li, 1999), that is, it achieves $\min_{P \in \mathcal{P}} D(Q\|P)$. Denoting the density of $P_{\boldsymbol{\mu}^*}$ by $p_{\boldsymbol{\mu}^*}$, it follows by Theorem 1 of Grünwald et al. (2024) that $q(U)/p_{\boldsymbol{\mu}^*}(U)$ would be an e-variable in case $\inf_{P \in \mathrm{conv}(\mathcal{P})} D(Q\|P) = \min_{P \in \mathcal{P}} D(Q\|P)$. Our theorem establishes a sufficient condition for when this is actually the case. It is based on constructing a second exponential family $\mathcal{Q}$ with densities proportional to $\exp(\boldsymbol{\beta}^T t(U))q(U)$ for varying $\boldsymbol{\beta}$: $\mathcal{Q}$ contains $Q$ and has the same sufficient statistic as $\mathcal{P}$. In some cases, but not all, $\mathcal{Q}$ may be thought of as the composite alternative we are interested in. Letting $\Sigma_p(\boldsymbol{\mu})$ and $\Sigma_q(\boldsymbol{\mu})$ denote the covariance matrices of the $P_{\boldsymbol{\mu}} \in \mathcal{P}$ and $Q_{\boldsymbol{\mu}} \in \mathcal{Q}$ with mean $\boldsymbol{\mu}$, Theorem 5.3 below implies the following: under a further regularity condition on the parameter spaces of $\mathcal{P}$ and $\mathcal{Q}$, simple e-variables exist whenever $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu}$ in the mean-value parameter space of $\mathcal{Q}$ (additionally, three equivalent conditions will be given). If this happens, then we may further conclude that for *every* element $Q_{\boldsymbol{\mu}'}$ of the constructed $\mathcal{Q}$, the likelihood ratio $q_{\boldsymbol{\mu}'}(U)/p_{\boldsymbol{\mu}'}(U)$ is an e-variable, where $P_{\boldsymbol{\mu}}$ is the element of $\mathcal{P}$ to which $Q_{\boldsymbol{\mu}}$ is projected. An example pair $(Q, \mathcal{P})$ to which the theorem applies is when, under $Q$, $U \sim N(m, s^2)$ for fixed $m, s^2$ and $\mathcal{P} = \{N(0, \sigma^2) : \sigma^2 > 0\}$ is the univariate (scale) family of normal distributions. This situation is illustrated in Figure 5.1 and is treated in detail in Section 5.4.3, and extended to linear regression testing – arguably our most important application – in Section 5.4.4.

We stress that, while our approach starts with a simple alternative $Q$, the results are still applicable if one is interested in a composite alternative $\mathcal{H}_1$. To this end, take any $Q \in \mathcal{H}_1$ and use our main result to determine whether a simple e-variable with respect to $Q$ exists. If one exists for every $Q$, an e-variable for the full alternative can easily be constructed either by the method of mixtures or the prequential (sequential plug-in learning) method (Ramdas et al., 2023).

Things conceptually simplify in this composite alternative case if $\mathcal{H}_1$ can be parameterized as $\mathcal{H}_1 = \{\mathcal{Q}^{(\theta)} : \theta \in \Theta\}$ in such a way that for each $Q \in \mathcal{H}_1$, the associated family $\mathcal{Q}$ constructed from $\mathcal{P}$ and $Q$ is equal to $\mathcal{Q}^{(\theta)}$ for some $\theta$. As is suggested by Figure 5.1, this happens, for example, in the Gaussian scale example of Section 5.4.3, if we consider as alternative $\mathcal{H}_1$ the full Gaussian family. We can start with any $Q = N(m, s^2)$ and generate $\mathcal{Q}$ which then coincides with some $\mathcal{Q}^{(\theta)}$, corresponding to a specific sloped line in the figure. Together, all these sloped lines span $\mathcal{H}_1$. In fact, it turns out that a natural choice of $\mathcal{H}_1$ that partitions into $\mathcal{Q}^{(\theta)}$ is possible in *all*

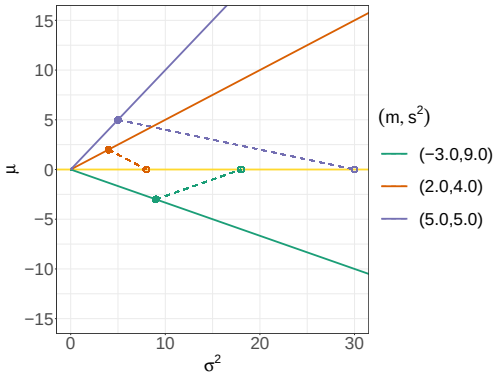



**Figure 5.1:** The family $\mathcal{Q}$ for various $(m, s^2)$. The coordinate grid represents the parameters of the full Gaussian family, the horizontal line shows the parameter space of $\mathcal{P}$, the sloped lines show the parameters of the distributions in $\mathcal{Q}$, and the dashed lines show the projection of $(m, s^2)$ onto the parameter space of $\mathcal{P}$. For example, we may start out with $Q$ expressing $U \sim N(m, s^2)$ with $m = -3.0, s^2 = 9.0$, represented as the green dot on the green line. Its RIPr onto $\mathcal{P}$ is the green point on the yellow line. The corresponding family $\mathcal{Q}$, constructed in terms of $Q$ and $\mathcal{P}$, is depicted by the green solid line. The theorem implies that the likelihood ratio between any point on the green line and its RIPr onto the yellow line is an e-variable; similarly for the red and blue lines.

our examples, and that this $\mathcal{H}_1$ is itself an exponential family in all these examples. Nevertheless, we stress that in general our method does not in any way require $\mathcal{H}_1$ to be an exponential family — only $\mathcal{P}$ is required to be so.

A specific interpretation of the result is obtained when restricting to the 1-dimensional case. The best squared-error predictor of $X$ sampled according to $Q$ has $Q$-expected squared error prediction equal to $\text{VAR}_Q[X] = \mathbf{E}_Q(X - \boldsymbol{\mu}^*)^2$. If $X$ is really sampled from $Q$ but we think it comes from $P_{\boldsymbol{\mu}^*}$ and want to make the best $P_{\boldsymbol{\mu}^*}$-expected squared error predictions, we would predict with the mean, which is still $\boldsymbol{\mu}^*$, but we assess our squared error as $\text{VAR}_{P_{\boldsymbol{\mu}^*}}[X]$ whereas the real expected squared error is still $\text{VAR}_Q[X]$. Thus, in the 1-dimensional case, in the situation that our result does *not* hold, there is a mismatch between $Q$ and its projection $P_{\boldsymbol{\mu}^*}$ in the sense that the closest approximation we can provide to $Q$ promises a better squared-error prediction than can be obtained with $Q$ itself. Our result says that if the mismatch does not occur, then we cannot get closer to $Q$ by convexifying $\mathcal{P}$.

The proof of Theorem 5.3 is based on convex duality properties of exponential families. In the remainder of this introductory section, we fix notation and definitions of exponential families and e-variables. In Section 5.2 we show how, based on the constructed family $\mathcal{Q}$, one can often easily construct *local* e-variables, i.e. e-variables

with the null restricted to a subset of $\mathcal{P}$. Then, in Section 5.3 we present our main theorem, extending the insight to global e-variables. Section 5.4 provides several examples. This includes cases for which simple e-variables were already established, such as certain k-sample tests (Turner et al., 2024) (see also Chapter 4) or — in an unpublished master's thesis — the linear regression model (De Jong, 2021), as well as cases for which it was previously unknown whether simple e-variables exist, such as for a broad class of natural exponential families. Theorem 5.3 can thus be seen as a unification and generalization of known results on the existence of simple e-variables, leading to deeper understanding of why they sometimes exist. Section 5.5 provides the proof for Theorem 5.3. Finally, Section 5.6 provides a concluding discussion and points out potential future directions.

### 5.1.2 Formal Setting

We study general hypothesis testing problems in which the null hypothesis $\mathcal{P}$ is a regular (and hence full) $d$-dimensional exponential family. Here and in the sequel, we will freely use standard properties of exponential families without explicitly referring to their definitions and proofs, for which we refer to e.g. (Barndorff-Nielsen, 1978; Brown, 1986; Efron, 2022). Each member of $\mathcal{P}$ is a distribution for a random element $U$, that takes values in some set $\mathcal{U}$, with a density relative to some given underlying measure $\nu$ on $\mathcal{U}$. The sufficient statistic vector is denoted by $X = (X_1, \ldots, X_d)$ with $X_j = t_j(U)$ for given measurable functions $t_1, \ldots, t_d$. We furthermore define $\mathtt{M}_p$ to be the mean-value parameter space of $\mathcal{P}$, i.e. the set of all $\boldsymbol{\mu}$ such that $\mathbb{E}_P[X] = \boldsymbol{\mu}$ for some $P \in \mathcal{P}$. For any $\boldsymbol{\mu} \in \mathtt{M}_p$, we denote by $P_{\boldsymbol{\mu}}$ the unique element of $\mathcal{P}$ with $\mathbb{E}_{P_{\boldsymbol{\mu}}}[X] = \boldsymbol{\mu}$, so that $\mathcal{P} = \{P_{\boldsymbol{\mu}} : \boldsymbol{\mu} \in \mathtt{M}_p\}$. As usual, this parameterization of $\mathcal{P}$ is referred to as its mean-value parameterization. Furthermore, we use $\Sigma_p$ to denote the variance function of $\mathcal{P}$. That is, for all $\boldsymbol{\mu} \in \mathtt{M}_p$, $\Sigma_p(\boldsymbol{\mu})$ is the covariance matrix corresponding to $P_{\boldsymbol{\mu}}$.

Since $\mathcal{P}$ is an exponential family, the density of any member of $\mathcal{P}$ can be written, for each fixed $\boldsymbol{\mu}^* \in \mathtt{M}_p$, as

$$p_{\boldsymbol{\beta};\boldsymbol{\mu}^*}(u) = \frac{1}{Z_p(\boldsymbol{\beta};\boldsymbol{\mu}^*)} \exp\left(\sum_{j=1}^{d} \beta_j t_j(u)\right) \cdot p_{\boldsymbol{\mu}^*}(u), \tag{5.1}$$

where $Z(\boldsymbol{\beta};\boldsymbol{\mu}^*) = \int \exp(\sum \beta_j t_j(u)) p_{\boldsymbol{\mu}^*}(u) d\nu$, and $\boldsymbol{\beta} \in \mathbb{R}^d$ such that $Z_p(\boldsymbol{\beta};\boldsymbol{\mu}^*) < \infty$. Therefore, $\mathcal{P}$ can also be parameterized as $\mathcal{P} = \{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*} : \boldsymbol{\beta} \in \mathtt{B}_{p,\boldsymbol{\mu}^*}\}$, where $\mathtt{B}_{p;\boldsymbol{\mu}^*} \subset \mathbb{R}^d$

denotes the canonical parameter space with respect to $\boldsymbol{\mu}^*$, i.e. the set of all $\boldsymbol{\beta}$ for which $Z_p(\boldsymbol{\beta};\boldsymbol{\mu}^*) < \infty$. We use $\boldsymbol{\beta}_p(\boldsymbol{\mu}';\boldsymbol{\mu}^*)$ to denote the $\boldsymbol{\beta} \in \mathtt{B}_{p,\boldsymbol{\mu}^*}$ such that $\mathbf{E}_{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}[X] = \boldsymbol{\mu}'$ and set $\boldsymbol{\mu}_p(\cdot;\boldsymbol{\mu}^*) = \boldsymbol{\beta}_p^{-1}(\cdot;\boldsymbol{\mu}^*)$ to be its inverse. $\boldsymbol{\beta}_p(\cdot;\boldsymbol{\mu}^*)$ maps mean-value parameters to corresponding canonical parameters and $\boldsymbol{\mu}_p(\cdot;\boldsymbol{\mu}^*)$ vice versa. Note that $p_{\boldsymbol{\mu}^*} = p_{\mathbf{0},\boldsymbol{\mu}^*}$, and that we can see from the notation (one versus two subscripts) whether a density is given in the mean- or canonical representation, respectively.

The reason for explicitly denoting the mean $\boldsymbol{\mu}^*$ of the carrier density, which is unconventional, is that it will be convenient to simultaneously work with different canonical parameterizations, i.e. with respect to a different element of $\mathtt{M}_p$, below. These are all linearly related to one another in the sense that for each $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathtt{M}_p$, there is a fixed vector $\boldsymbol{\gamma}$ such that for all $\boldsymbol{\beta} \in \mathtt{B}_{p,\boldsymbol{\mu}_1}$ it holds that $p_{\boldsymbol{\beta};\boldsymbol{\mu}_1} = p_{\boldsymbol{\beta}+\boldsymbol{\gamma};\boldsymbol{\mu}_2}$. This can be seen by taking $\boldsymbol{\gamma} = -\boldsymbol{\beta}_p(\boldsymbol{\mu}_2;\boldsymbol{\mu}_1)$, since one then has

$$\begin{aligned} p_{\boldsymbol{\beta};\boldsymbol{\mu}_1}(u) &= \frac{1}{Z_p(\boldsymbol{\beta};\boldsymbol{\mu}_1)} \exp\left(\sum_{j=1}^{d}(\beta_j+\gamma_j)t_j(u)\right)\exp\left(\sum_{j=1}^{d}-\gamma_j t_j(u)\right)p_{\boldsymbol{\mu}_1}(u) \\ &= \frac{Z_p(-\boldsymbol{\gamma};\boldsymbol{\mu}_1)}{Z_p(\boldsymbol{\beta};\boldsymbol{\mu}_1)} \exp\left(\sum_{j=1}^{d}(\beta_j+\gamma_j)t_j(u)\right)p_{-\boldsymbol{\gamma};\boldsymbol{\mu}_1}(u) \\ &= \frac{1}{Z_p(\boldsymbol{\beta}+\boldsymbol{\gamma};\boldsymbol{\mu}_2)} \exp\left(\sum_{j=1}^{d}(\beta_j+\gamma_j)t_j(u)\right)p_{\boldsymbol{\mu}_2}(u) = p_{\boldsymbol{\beta}+\boldsymbol{\gamma};\boldsymbol{\mu}_2}(u). \end{aligned} \tag{5.2}$$

### 5.1.3 The Composite Alternative Generated by A Simple One

We are mostly concerned with testing the null hypothesis $\mathcal{P}$ against simple alternative hypotheses of the form $\{Q\}$ for some distribution $Q$ on $\mathcal{U}$. In particular, we will consider distributions $Q$ that admit a moment generating function and that have a density $q$ relative to the underlying measure $\nu$. While the former is a strong condition, it holds in many cases of interest. For our analysis, it will be beneficial to define a second exponential family $\mathcal{Q}$ for $U$ with distributions $Q_{\boldsymbol{\beta};\boldsymbol{\mu}^*}$ and corresponding densities

$$q_{\boldsymbol{\beta};\boldsymbol{\mu}^*}(u) = \frac{1}{Z_q(\boldsymbol{\beta};\boldsymbol{\mu}^*)} \cdot \exp\left(\sum_{j=1}^{d}\beta_j t_j(u)\right)\cdot q(u), \tag{5.3}$$

where $\boldsymbol{\mu}^*$ is the mean of $X$ under $Q$, and $Z_q(\boldsymbol{\beta};\boldsymbol{\mu}^*)$ is the normalizing constant. The notational conventions that we use for $\mathcal{Q}$ will be completely analogous to that for $\mathcal{P}$, e.g. $\boldsymbol{\beta}_q(\cdot,\mu^*)$, $\boldsymbol{\mu}_q(\cdot,\mu^*)$, $\Sigma_q$, etc. Since $Q$ is assumed to have a moment generating

function, the canonical domain $\mathtt{B}_{q,\boldsymbol{\mu}^*}$ is nonempty and contains a neighborhood of $\mathbf{0}$. Similarly, the mean-value space $\mathtt{M}_q$ is also nonempty and contains a neighborhood of $\boldsymbol{\mu}^*$. We further have the following: if we take any other $Q' \in \mathcal{Q}$, say $Q' = Q_{\boldsymbol{\mu}'}$ for $\boldsymbol{\mu}' \in \mathtt{M}_q$, then the 'constructed' family around $Q'$, i.e. $\{q_{\boldsymbol{\beta};\boldsymbol{\mu}'} : \boldsymbol{\beta} \in \mathtt{B}_{q;\boldsymbol{\mu}'}\}$ coincides with $\mathcal{Q}$ (as was the case for $\mathcal{P}$, by (5.2)).

We may think of the null $\mathcal{P}$ and the generated family $\mathcal{Q}$ as two different exponential families that share the same sufficient statistic. Moreover, as we shall see below, there are many examples where their mean-value spaces are equal, that is, $\mathtt{M}_q = \mathtt{M}_p$. In this case $\mathcal{P}$ and $\mathcal{Q}$ are "matching" pairs: they share the same sufficient statistic as well as the same set of means for this statistic.

### 5.1.4 E-variables

We use e-variables to gather evidence against the null hypothesis $\mathcal{P}$. An e-variable is a non-negative statistic with expected value bounded by one under the null, i.e. a non-negative statistic $S(U)$ such that $\mathbb{E}_P[S(U)] \leq 1$ for all $P \in \mathcal{P}$. We give only a brief introduction to e-variables here and refer to e.g. (Grünwald et al., 2024; Ramdas et al., 2023) for detailed discussions. The realization of an e-variable on observed data will be referred to as an e-value, though the two terms are often used interchangeably. Large e-values give evidence against the null hypothesis, since by Markov's inequality we have that $Q(S(U) \geq \frac{1}{\alpha}) \leq \alpha$ for any e-variable $S(U)$ and $Q \in \mathcal{P}$. The focus here is on a static setting, where e-variables are computed for a single block of data (i.e. one observation of $U$). However, the main application of e-variables is in anytime-valid settings, where data arrives sequentially and one wants a type-I error guarantee uniformly over time. Indeed, it is well-known that the product of sequentially computed e-variables again gives an e-variable, even if the definition of each subsequent e-variable depends on past e-values, which leads to an easy extension of the methods described here to such anytime-valid settings (Ramdas et al., 2023; Grünwald et al., 2024).

Since large e-values give evidence against the null, we look for e-variables that are, on average, 'as large as possible' under the alternative hypothesis. In particular, we study growth-rate optimal (GRO) e-variables, an optimality criterion embraced implicitly or explicitly by most of the e-community (Ramdas et al., 2023). Grünwald et al. (2024) define the GRO e-variable for single outcome $U$, relative to a simple alternative $\{Q\}$, to be the e-variable $S$ that, among all e-variables, maximizes the growth-rate $\mathbb{E}_{U\sim Q}[\log S(U)]$ (also known as e-power (Wang et al., 2024; Zhang et al.,

2024)). In a celebrated result, Grünwald et al. (see also Chapter 3 and (Larsson et al., 2024)) show that the GRO e-variable is given by:

$$\frac{q(U)}{p_{\leftsquigarrow q}(U)}, \tag{5.4}$$

where $p_{\leftsquigarrow q}$ denotes the reverse information projection of $Q$ on the convex hull of the null $\mathcal{P}$. The reverse information projection of $Q$ on $\mathrm{conv}(\mathcal{P})$ is defined as the distribution that uniquely achieves $\inf_{P\in\mathrm{conv}(\mathcal{P})} D(Q\|P)$, which is known to exist whenever the latter is finite (see Chapter 3 for further details). Here, $D(Q\|P)$ denotes the Kullback-Leibler (KL) divergence between $Q$ and $P$, both defined as distributions for $U$. In this chapter, all reverse information projection will be on $\mathrm{conv}(\mathcal{P})$, so we will not explicitly mention the domain of projection everywhere (i.e. referring to it simply as 'the reverse information projection of $Q$'). The growth rate achieved by the GRO e-variable is given by

$$\mathbb{E}_Q\left[\log \frac{q(U)}{p_{\leftsquigarrow q}(U)}\right] = D(Q\|P_{\leftsquigarrow q}) = \inf_{P\in\mathrm{conv}(\mathcal{P})} D(Q\|P). \tag{5.5}$$

However, due to the fact that, with the exception of the Bernoulli and multinomial models, exponential families are not convex sets of distributions, finding the reverse information projection can be quite challenging (Lardy, 2021). In this chapter we provide a simple and easily verifiable condition under which

$$\inf_{P\in\mathrm{conv}(\mathcal{P})} D(Q\|P) = \min_{P\in\mathcal{P}} D(Q\|P), \tag{5.6}$$

that is, the infimum is achieved by an element of $\mathcal{P}$, so that the problem greatly simplifies.

In that case, the GRO e-variable simply takes on the form of a likelihood ratio between $Q$ and a particular member of $\mathcal{P}$, i.e.

$$\frac{q(U)}{p(U)}, \tag{5.7}$$

which we will refer to as a simple e-variable relative to $Q$. We will frequently use the fact (following from Corollary 1 of (Grünwald et al., 2024, Theorem 1)) that there can be at most one simple e-variable with respect to any fixed alternative, i.e. of the form (5.7). This is captured by the following proposition.

**Proposition 5.1.** *Fix a probability measure $Q$ on $U$. If there exists a simple e-variable relative to $Q$, then it must be the GRO e-variable for testing $\mathcal{P}$ against alternative $\{Q\}$.*

A big advantage of simple e-variables—besides their simplicity—is that their optimality extends beyond the static setting. That is, suppose we were to observe independent copies $U_1, U_2, \ldots$ of the data and assume that a simple e-variable of the form (5.7) exists. As alluded to before, we can measure the total evidence as $\prod_{i=1}^{n} q(U_i)/p(U_i)$, which defines an e-variable for any fixed $n \in \mathbb{N}$. Instead of thinking of this as multiplication of individual e-variables, one can think of it as a likelihood ratio of $U_1, \ldots, U_n$. Proposition 5.1 then implies that $\prod_{i=1}^{n} q(U_i)/p(U_i)$ is the GRO e-variable for testing $\mathcal{P}$ against $\{Q\}$ based on $n$ data points. This statement shows that for any fixed sample size $n$, the best e-variable (in the GRO sense of 5.5) is the simple likelihood ratio. Moreover, for applications where the sample size is not fixed beforehand, Koolen and Grünwald (2022, Theorem 12) show that a more flexible statement is also true: if $\tau$ is any stopping time that is adapted to the data filtration, then $q(U^\tau)/p(U^\tau)$ is also a maximizer of $\mathbb{E}[\ln S_\tau]$ over all processes $S = (S_n)_{n\in\mathbb{N}}$ with $\mathbb{E}[S_\tau] \leq 1$. While we will not explicitly consider this type of sequential optimality in the following, it is one of the main motivating factors behind this work.

We assume throughout this chapter that, for any considered alternative $Q$, there exists a $\boldsymbol{\mu}^* \in \mathtt{M}_p$ such that $\mathbb{E}_{X\sim Q}[X] = \boldsymbol{\mu}^*$. By a standard property of exponential families, the KL divergence from $Q$ to $\mathcal{P}$ is then minimized by the element of $\mathcal{P}$ with the same mean as $Q$. If (5.6) holds, then $P_{\boldsymbol{\mu}^*}$ must therefore be the reverse information projection of $Q$. It follows that, if a simple e-variable with respect to $Q$ exists, then it is given by $q(U)/p_{\boldsymbol{\mu}^*}(U)$.

## 5.2 Existence of Simple Local E-Variables

Here we will show how the family $\mathcal{Q}$ is related to the question of whether the likelihood ratio $q(U)/p_{\boldsymbol{\mu}^*}(U)$ is a *local* GRO e-variable around $\boldsymbol{\mu}^*$. We say that a nonnegative statistic $S(U)$ is a local e-variable around $\boldsymbol{\mu}^*$ if there exists a connected open subset $\mathtt{B}'_{\boldsymbol{\mu}^*}$ of $\mathtt{B}_{p;\boldsymbol{\mu}^*} \cap \mathtt{B}_{q;\boldsymbol{\mu}^*}$ containing $\mathbf{0}$ such that $S$ is an e-variable relative to $\mathcal{P}' = \{P_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathtt{B}'_{\boldsymbol{\mu}^*}\}$, i.e. $\sup_{\boldsymbol{\beta}\in\mathtt{B}'_{\boldsymbol{\mu}^*}} \mathbb{E}_{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}[S] \leq 1$. If $S$ also maximizes $\mathbb{E}_Q[\ln S(U)]$ among all e-variables relative to $\mathcal{P}'$, then we say that $S$ is a local GRO e-variable with respect to $Q$. A local (GRO) e-variable may not be an e-variable relative to the full null hypothesis $\mathcal{P}$, but it is a an e-variable relative to some smaller null hypothesis, restricted to all distributions in the null with mean in a neighborhood of $\boldsymbol{\mu}^*$. Investigating when local e-variables exist provides the basic insight on top of which the subsequent, much stronger Theorem 5.3 about 'global' e-variables is built. As stated in Section 5.1.3, we may view $\mathcal{P}$ and $\mathcal{Q}$ as two families with the same sufficient statistic, only differing in

their carrier, which for $\mathcal{P}$ is $p_{\boldsymbol{\mu}^*} = p_{\mathbf{0};\boldsymbol{\mu}^*}$ and for $\mathcal{Q}$ is $q_{\mathbf{0};\boldsymbol{\mu}^*} = q = q_{\boldsymbol{\mu}^*}$: we can and will denote the original $Q$ also by $Q_{\boldsymbol{\mu}^*}$.

Define the function $f(\cdot;\boldsymbol{\mu}^*) : \mathtt{B}_{p;\boldsymbol{\mu}^*} \cap \mathtt{B}_{q;\boldsymbol{\mu}^*} \to \mathbb{R}$ as

$$f(\boldsymbol{\beta};\boldsymbol{\mu}^*) := \log \mathbb{E}_{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}\left[\frac{q_{\boldsymbol{\mu}^*}(U)}{p_{\boldsymbol{\mu}^*}(U)}\right] = \log Z_q(\boldsymbol{\beta};\boldsymbol{\mu}^*) - \log Z_p(\boldsymbol{\beta};\boldsymbol{\mu}^*), \tag{5.8}$$

where the equality comes from the fact that we can rewrite the density in the numerator as $q_{\boldsymbol{\mu}^*}(U) = Z_q(\boldsymbol{\beta};\mu^*)\exp(\sum_{j=1}^d \beta_j t_j(u))^{-1} q_{\boldsymbol{\beta};\boldsymbol{\mu}^*}(U)$ and similar for the density in the denominator. It should be clear that the function $f(\cdot;\boldsymbol{\mu}^*)$ is highly related to the question we are interested in. Indeed, $q_{\boldsymbol{\mu}^*}(U)/p_{\boldsymbol{\mu}^*}(U)$ is a local e-variable relative to $\mathcal{P}' = \{P_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathtt{B}'_{\boldsymbol{\mu}^*}\}$ if and only if $\sup_{\boldsymbol{\beta}\in\mathtt{B}'_{\boldsymbol{\mu}^*}} f(\boldsymbol{\beta};\boldsymbol{\mu}^*) \leq 0$. Equivalently, since $f(\mathbf{0};\boldsymbol{\mu}^*) = \mathbf{0}$, we have that $q_{\boldsymbol{\mu}^*}/p_{\boldsymbol{\mu}^*}$ is a local e-variable around $\boldsymbol{\mu}^*$ if and only if there is a local maximum at $\mathbf{0}$. To investigate when this happens, a standard result on exponential families gives the following:

$$\nabla f(\boldsymbol{\beta};\boldsymbol{\mu}^*) = \mathbb{E}_{Q_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}[X] - \mathbb{E}_{P_{\boldsymbol{\beta};\boldsymbol{\mu}^*}}[X] \tag{5.9}$$

In particular, it follows that $\nabla f(\mathbf{0};\boldsymbol{\mu}^*) = \boldsymbol{\mu}^* - \boldsymbol{\mu}^* = \mathbf{0}$. Thus, $q_{\boldsymbol{\mu}^*}/p_{\boldsymbol{\mu}^*}$ is a local e-variable around $\boldsymbol{\mu}^*$ if and only if the $d \times d$ Hessian matrix of second partial derivatives of $f(\cdot;\boldsymbol{\mu}^*)$, is negative semidefinite in $\mathbf{0}$. By (5.8)-(5.9) and using a convex duality property of exponential families, this is equivalent to

$$I_p(\mathbf{0};\boldsymbol{\mu}^*) - I_q(\mathbf{0};\boldsymbol{\mu}^*) = \Sigma_p(\boldsymbol{\mu}^*) - \Sigma_q(\boldsymbol{\mu}^*) \text{ is positive semidefinite,}$$

where $I_p$ and $I_q$ denote the Fisher information matrix in terms of the canonical parameter spaces of $\mathcal{P}$ and $\mathcal{Q}$, respectively. We have thus proven our first result:

**Proposition 5.2.** *$q_{\boldsymbol{\mu}^*}(U)/p_{\boldsymbol{\mu}^*}(U)$ is a local e-variable around $\boldsymbol{\mu}^*$ (and therefore, by Proposition 5.1, a GRO local e-variable) if and only if $\Sigma_p(\boldsymbol{\mu}^*) - \Sigma_q(\boldsymbol{\mu}^*)$ is positive semidefinite.*

The surprising result that follows below essentially adds to this that, if for every $\boldsymbol{\mu}^* \in \mathtt{M}_q$, $q_{\boldsymbol{\mu}^*}/p_{\boldsymbol{\mu}^*}$ is a local e-variable, then also for every $\boldsymbol{\mu}^*$, we have that $q_{\boldsymbol{\mu}^*}/p_{\boldsymbol{\mu}^*}$ is a full, global e-variable!

## 5.3 Existence of Simple Global E-Variables

The theorem below gives eight equivalent characterizations of when a global GRO e-variable exists. Not all characterizations are equally intuitive and informative: the simplest ones are Part 1 and 3. To appreciate the more complicated characterizations as well, it is useful to first recall some convex duality properties concerning derivatives of KL divergences with regular exponential families (see e.g. Grünwald, 2007, Section 18.4.3):

$$\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}^*) = \nabla_{\boldsymbol{\mu}} D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}^*}), \tag{5.10}$$

$$\left(\Sigma_p^{-1}(\boldsymbol{\mu})\right)_{ij} = \frac{d^2}{d\boldsymbol{\mu}_i d\boldsymbol{\mu}_j} D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}^*}), \tag{5.11}$$

and analogous for $\mathcal{Q}$. That is, the gradient of the KL divergence in its first argument at $\boldsymbol{\mu}$ is given by the canonical parameter vector corresponding to $\boldsymbol{\mu}$, and the Hessian is given by the Fisher information, i.e. the inverse covariance matrix.

**Theorem 5.3.** *Let $\mathcal{P}$ be a regular exponential family with mean-value parameter space $\mathtt{M}_p$. Fix a distribution $Q$ for $U$ with $\mathbb{E}_Q[X] = \boldsymbol{\mu}^*$ for some $\boldsymbol{\mu}^* \in \mathtt{M}_p \subseteq \mathbb{R}^d$ and consider the corresponding $\mathcal{Q}$ as defined above. Suppose that $\mathtt{M}_q$ is convex, $\mathtt{M}_q \subseteq \mathtt{M}_p$, and $\mathtt{B}_{p;\boldsymbol{\mu}} \subseteq \mathtt{B}_{q;\boldsymbol{\mu}}$ for all $\boldsymbol{\mu} \in \mathtt{M}_q$. Then the following statements are equivalent:*

1. *$\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*
2. *$(\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}; \boldsymbol{\mu}'))^T \cdot (\boldsymbol{\mu} - \boldsymbol{\mu}') \leq 0$ for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$.*
3. *$D(Q_{\boldsymbol{\mu}} \| Q_{\boldsymbol{\mu}'}) \geq D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}'})$ for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$.*
4. *$\log Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}) \geq \log Z_q(\boldsymbol{\beta}; \boldsymbol{\mu})$ for all $\boldsymbol{\mu} \in \mathtt{M}_q, \boldsymbol{\beta} \in \mathtt{B}_{p;\boldsymbol{\mu}}$.*
5. *$q_{\boldsymbol{\mu}}(U)/p_{\boldsymbol{\mu}}(U)$ is a global e-variable for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*
6. *$q_{\boldsymbol{\mu}}(U)/p_{\boldsymbol{\mu}}(U)$ is the global GRO e-variable w.r.t. $Q_{\boldsymbol{\mu}}$ for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*
7. *$q_{\boldsymbol{\mu}}(U)/p_{\boldsymbol{\mu}}(U)$ is a local e-variable for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*
8. *$q_{\boldsymbol{\mu}}(U)/p_{\boldsymbol{\mu}}(U)$ is a local GRO e-variable w.r.t. $Q_{\boldsymbol{\mu}}$ for all $\boldsymbol{\mu} \in \mathtt{M}_q$.*

Note that the canonical parameter space of a full exponential family is always convex, but the mean-value space need not be (Efron, 2022). Still, in all examples we consider below, the constructed family $\mathcal{Q}$ will in fact be a *regular* exponential family, and then the convexity requirement must hold.

In the one-dimensional case, the first statement simplifies to $\sigma_p^2(\mu) \geq \sigma_q^2(\mu)$ for all $\mu \in \mathtt{M}_q$. Similarly, the second statement reduces to $\boldsymbol{\beta}_q(\mu;\mu') \geq \boldsymbol{\beta}_p(\mu;\mu')$ for all $\mu \in \mathtt{M}_q$ such that $\mu > \mu'$ and $\boldsymbol{\beta}_q(\mu;\mu') \leq \boldsymbol{\beta}_p(\mu;\mu')$ for all $\mu \in \mathtt{M}_q$ such that $\mu < \mu'$ for all $\mu' \in \mathtt{M}_q$.

Using standard properties of Loewner ordering, it can be established that $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite if and only if $\Sigma_q^{-1}(\boldsymbol{\mu}) - \Sigma_p^{-1}(\boldsymbol{\mu})$ is (see e.g. Agrawal, 2018). Therefore, recalling (5.10) and (5.11), statement 1 in Theorem 5.3 can be thought of as a condition on the second derivative of $D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}^*}) - D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}^*})$, whereas statement 2 refers to its first derivative, and statement 3 to the difference in KL divergence itself. It is somewhat surprising that signs of differences between the second derivatives and separately signs of differences between the first derivatives are sufficient to determine signs of difference between a function itself.

### 5.3.1 Simplifying Situations

In some special situations, the conditions needed to apply Theorem 5.3 may be significantly simplified. We now identify two such situations, embodied by Proposition 5.4 and Corollary 5.5, that will be useful for our examples below.

First, we note that it is sometimes easy to check that either $\mathtt{M}_p = \mathtt{M}_q$ or $\mathtt{B}_{p;\boldsymbol{\mu}^*} = \mathtt{B}_{q;\boldsymbol{\mu}^*}$. The following proposition shows that, in the 1-dimensional setting, this is already sufficient to apply the theorem (we do not know whether an analogous result holds in higher dimensions):

**Proposition 5.4.** *Let $\mathcal{P}$ be a 1-dimensional regular exponential family with mean-value parameter space $\mathtt{M}_p \subseteq \mathbb{R}$. Fix a distribution $Q$ for $U$ with $\mathbb{E}_Q[X] = \mu^*$ for some $\mu^* \in \mathtt{M}_p$ and consider the corresponding $\mathcal{Q}$ as defined above. Suppose that for all $\mu \in \mathtt{M}_q$, we have $\sigma_p^2(\mu) \geq \sigma_q^2(\mu)$, i.e. the first condition of Theorem 5.3 holds. Then:*

1. *If $\mathtt{M}_q = \mathtt{M}_p$ then for all $\mu' \in \mathtt{M}_q$, $\mathtt{B}_{p;\mu'} \subseteq \mathtt{B}_{q;\mu'}$, i.e., Theorem 5.3 is applicable.*
2. *If for some $\mu \in \mathtt{M}_q$, we have that $\mathtt{B}_{p;\mu} = \mathtt{B}_{q;\mu}$ then $\mathtt{M}_q \subseteq \mathtt{M}_p$. Hence if for all $\mu \in \mathtt{M}_q$, we have that $\mathtt{B}_{p;\mu} = \mathtt{B}_{q;\mu}$, then Theorem 5.3 is applicable.*

The proof is simple and we only sketch it here: for part 1, draw the graphs of $\beta_p(\mu;\mu')$ and $\beta_q(\mu;\mu')$ as functions of $\mu \in \mathtt{M}_q$, noting that both functions must take the value 0 at the point $\mu = \mu'$. Using that $1/\sigma_p^2(\mu)$ is the derivative of $\beta_p(\mu;\mu')$ and similarly for $\sigma_q^2(\mu)$, the function $\beta_q(\mu;\mu')$ must lie above $\beta_p(\mu;\mu')$ for $\mu > \mu'$, and below for $\mu < \mu'$. Therefore the co-domain of $\beta_q$ must include that of $\beta_p$. The second

part goes similarly, essentially by flipping the just-mentioned graph of two functions by 90 degrees.

Second, we note that in practice we often have a composite alternative $\mathcal{H}_1$ in mind such that the union of the set of families $\mathcal{Q}$ that can be constructed from $\mathcal{P}$ and any $Q \in \mathcal{H}_1$ in fact coincides with $\mathcal{H}_1$. This is the case in the examples of Section 5.4.1, 5.4.3, 5.4.3 and 5.4.4. The following immediate corollary of Theorem 5.3 simplifies the analysis in such cases (although we will only explicitly need to invoke it in Section 5.4.4). While in that section, $\mathcal{H}_1$ will itself be an exponential family, we stress that in general, this need not be the case: to apply the corollary it is sufficient for $\mathcal{H}_1$ to be a *union* of exponential families.

**Corollary 5.5.** *Let $\mathcal{P}$ be a $d$-dimensional regular exponential family as before with mean-value parameter space $\mathtt{M}_p$, and let $\mathcal{H}_1 = \bigcup_{\theta\in\Theta} \mathcal{Q}^{(\theta)}$ where each $\mathcal{Q}^{(\theta)}$ is a $d$-dimensional regular exponential family with the same sufficient statistic as $\mathcal{P}$ and with mean-value parameter space $\mathtt{M}_q^{(\theta)}$ and canonical parameter spaces $\mathtt{B}_{q;\boldsymbol{\mu}}^{(\theta)}$ for $\boldsymbol{\mu} \in \mathtt{M}_q^{(\theta)}$. Suppose that, for each $\theta \in \Theta$, for each $Q \in \mathcal{Q}^{(\theta)}$, the corresponding set $\mathcal{Q}$ as constructed above in terms of $\mathcal{P}$ and $Q$, happens to be equal to $\mathcal{Q}^{(\theta)}$ and satisfies the pre-condition of Theorem 5.3, i.e. $\mathtt{M}_q^{(\theta)}$ is convex, $\mathtt{M}_q^{(\theta)} \subseteq \mathtt{M}_p$, and $\mathtt{B}_{p;\boldsymbol{\mu}} \subseteq \mathtt{B}_{q;\boldsymbol{\mu}}^{(\theta)}$ for all $\boldsymbol{\mu} \in \mathtt{M}_q^{(\theta)}$. Then we have, with $Q_{\boldsymbol{\mu}}^{(\theta)}$ (density $q_{\boldsymbol{\mu}}^{(\theta)}$) denoting the element of $\mathcal{Q}^{(\theta)}$ with mean $\boldsymbol{\mu}$, for all $\theta \in \Theta$:*

*For all $\boldsymbol{\mu} \in \mathtt{M}_q^{(\theta)}$: $\frac{q_{\boldsymbol{\mu}}^{(\theta)}(U)}{p_{\boldsymbol{\mu}}(U)}$ is the global GRO e-variable w.r.t. $Q_{\boldsymbol{\mu}}^{(\theta)}$ $\Leftrightarrow$*
*For all $\boldsymbol{\mu} \in \mathtt{M}_q^{(\theta)}$: $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q^{(\theta)}(\boldsymbol{\mu})$ is positive semidefinite.*

*Here $\Sigma_q^{(\theta)}(\boldsymbol{\mu})$ denotes the $d \times d$ covariance matrix of the element of $\mathcal{Q}^{(\theta)}$ with mean-value parameter vector $\boldsymbol{\mu}$.*

## 5.4 Examples

In this section we discuss a variety of settings to which Theorem 5.3 can be applied. In some cases, this gives new insights into whether simple e-variables exist, and in others it simply gives a reinterpretation of existing results. The examples are broadly divided in terms of the curvature of the function $f(\cdot; \boldsymbol{\mu}^*)$, as defined in (5.8). Instances where $f(\cdot; \boldsymbol{\mu}^*)$ is constant will be referred to as having 'zero curvature', those with a constant second derivative as having 'constant curvature', and 'nonconstant curvature' otherwise.

### 5.4.1 Zero Curvature: Gaussian and Poisson k-sample tests

In Chapter 4, we studied GRO e-values for $k$-sample tests with regular exponential families. In that setting, data arrived in $k \in \mathbb{N}$ groups, or samples, and the null hypothesis was that all of the data points are distributed according to the same element of some exponential family. That is, let $U = (Y_1, \ldots, Y_k)$ for $Y_i \in \mathcal{Y}$, so that $\mathcal{U} = \mathcal{Y}^k$ for some measurable space $\mathcal{Y}$. Furthermore, fix a one-dimensional regular exponential family on $\mathcal{Y}$, given in its mean-value parameterization as $\mathcal{P}_{\text{START}} = \{P_\mu : \mu \in \mathtt{M}_{\text{START}}\}$ with sufficient statistic $t_{\text{START}}(Y)$. The composite null hypothesis $\mathcal{P}$ considered in the k-sample test expresses that $Y_1, \ldots, Y_k \overset{\text{i.i.d.}}{\sim} P_\mu$ for some $\mu \in \mathtt{M}_{\text{START}}$. On the other hand, the simple alternative $Q$ that was considered in Chapter 4 is characterized by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \mathtt{M}^k_{\text{START}}$, and expresses that the $Y_1, \ldots, Y_k$ are independent with $Y_i \sim P_{\mu_i}$ for $i = 1 \ldots k$. It was shown that, for the case that $\mathcal{P}_{\text{START}}$ is either the Gaussian location family or the Poisson family,

$$S(U) := \prod_{i=1}^{k} \frac{p_{\mu_i}(Y_i)}{p_{\bar{\mu}}(Y_i)}, \text{ with } \bar{\mu} = \frac{1}{k} \sum_{i=1}^{k} \mu_i,$$

is a simple e-value relative to $Q$, and that its expectation is constant as the null varies. That is, for any $\mu' \in \mathtt{M}_{\text{START}}$, it holds that

$$\mathbb{E}_{U \sim P_{\mu'} \times \cdots \times P_{\mu'}} [S(U)] = 1. \tag{5.12}$$

This finding can now be re-interpreted as an instance of Theorem 5.3, as we will show in detail for the Poisson family; the analysis for the Gaussian location family is completely analogous. In the Poisson case, $t_{\text{START}}(Y) = Y$, so that $\mathcal{P}$ defines an exponential family on $\mathcal{U}$ with sufficient statistic $X = \sum_{i=1}^{k} Y_i$ and mean-value space $\mathtt{M}_p = \mathbb{R}^+$. The latter follows because the sum of Poisson data is itself Poisson distributed with mean equal to the sum of means of the original data. Under the alternative, the mean of the sufficient statistic is given by $\mu^* := \mathbb{E}_Q[\sum_{i=1}^{k} Y_i] = \sum_{i=1}^{k} \mu_i$, so that the elements of the auxiliary exponential family $\mathcal{Q}$ as in (5.3) can be written as

$$q_{\beta;\mu^*}(Y_1, \ldots, Y_k) = \frac{1}{Z_q(\beta; \mu^*)} \cdot \exp\left(\beta \sum_{i=1}^{k} Y_i\right) \cdot q(Y_1, \ldots, Y_k). \tag{5.13}$$

Note in particular that $\mathcal{Q}$ is, by construction, a one-dimensional exponential family with sufficient statistic $\sum_{i=1}^{k} Y_i$, which does not equal (yet may be viewed as a subset of) the full $k$-dimensional exponential family from which $Q$ was originally chosen. The

normalizing constant $Z_q(\beta;\mu^*)$ is equal to the moment generating function of $X$ under $Q$, which is given by

$$Z_q(\beta;\mu^*) = \mathbb{E}_Q\left[\exp\left(\beta\sum_{i=1}^k Y_i\right)\right] = \exp\left(\mu^*(e^\beta - 1)\right).$$

It follows that

$$\mathbb{E}_{Q_{\beta;\mu^*}}\left[\sum_{i=1}^k Y_i\right] = \frac{\mathrm{d}}{\mathrm{d}\beta}\log Z_q(\beta;\mu^*) = \mu^* e^\beta,$$

which shows that mean-value space of the alternative is again given by $\mathtt{M}_q = \mathbb{R}^+$. Therefore, via Proposition 5.4, the assumptions of Theorem 5.3 are satisfied. The element of $\mathcal{P}$ with mean $\mu^*$ is given by $P_{\bar{\mu}} \times \cdots \times P_{\bar{\mu}}$, so that

$$\frac{q_{\mu^*}(U)}{p_{\mu^*}(U)} = \prod_{i=1}^k \frac{p_{\mu_i}(Y_i)}{p_{\bar{\mu}}(Y_i)}.$$

Under $P_{\mu^*}$, the sufficient statistic $\sum_{i=1}^k Y_i$ has the same distribution as under $Q_{\mu^*}$, so that $Z_p(\beta;\mu^*) = Z_q(\beta;\mu^*)$. Consequently, $f(\cdot;\boldsymbol{\mu}^*)$ as in (5.8) is zero, so that its second derivative is zero, and condition 1 of Theorem 5.3 is verified. It follows that, $q_{\mu^*}(U)/p_{\mu^*}(U)$ is the global GRO e-variable with respect to $Q_{\mu^*}$.

### 5.4.2 Constant Curvature: Multivariate Gaussian Location

Suppose that $\mathcal{P}$ is the multivariate Gaussian location family with some given nondegenerate covariance matrix $\Sigma_p$ and let $Q$ be any Gaussian distribution with nondegenerate covariance matrix $\Sigma_q$. Note that in this case we have that $X = U$, i.e. the sufficient statistic is simply given by the original data. The family $\mathcal{Q}$, generated from $Q$ and $\mathcal{P}$ as in (5.3), is the full Gaussian location family with fixed covariance matrix $\Sigma_q$. For both $\mathcal{P}$ and $\mathcal{Q}$, the mean-value and canonical spaces are all equal to $\mathbb{R}^d$, so that Theorem 5.3 applies to the pair $\mathcal{P}$ and $\mathcal{Q}$. Furthermore, the covariance functions are constant, since $\Sigma_p(\boldsymbol{\mu}) = \Sigma_p$ and $\Sigma_q(\boldsymbol{\mu}) = \Sigma_q$ for all $\boldsymbol{\mu} \in \mathbb{R}^d$. It follows that, if $\Sigma_p - \Sigma_q$ is positive semidefinite, then $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathbb{R}^d$. In that case, Theorem 5.3 shows that the simple likelihood ratio $q_{\boldsymbol{\mu}}/p_{\boldsymbol{\mu}}$ is the GRO e-value w.r.t. $Q_{\boldsymbol{\mu}}$ for every $\boldsymbol{\mu} \in \mathbb{R}^d$. The growth rate is given by

$$\mathbb{E}_Q\left[\log\frac{q_{\boldsymbol{\mu}}(U)}{p_{\boldsymbol{\mu}}(U)}\right] = D_{\text{GAUSS}}(\Sigma_q\Sigma_p^{-1}),$$

where $D_{\text{GAUSS}}(B) := \frac{1}{2}\left(-\log\det(B) - (d - \text{tr}(B))\right)$, i.e. the standard formula for the KL divergence between two multivariate Gaussians with the same mean.

In the case that $\Sigma_p - \Sigma_q$ is negative semidefinite, the simple likelihood ratio does not give an e-value; the GRO e-value for this case can also be derived however and will be reported on in future work.

### 5.4.3 Nonconstant Curvature: Univariate Examples

We now discuss three examples with nonconstant curvature. In the first two, Theorem 5.3 can be used to show the existence of simple e-variables. All three are univariate in nature; in the separate Section 5.4.4 we provide the example of linear regression, which has nonconstant curvature but is multivariate.

#### More k-Sample Tests

Besides the Gaussian and Poisson case, Chapter 4 identified one more model that gives rise to a $k$-sample test with a simple e-value: the case that $\mathcal{P}_{\text{START}}$ is the Bernoulli model. The difference with the Gaussian location- and Poisson family is that the involved e-value does not have constant expectation 1 here. Nevertheless, this result for the Bernoulli model can also be cast in terms of Theorem 5.3 using a different argument.

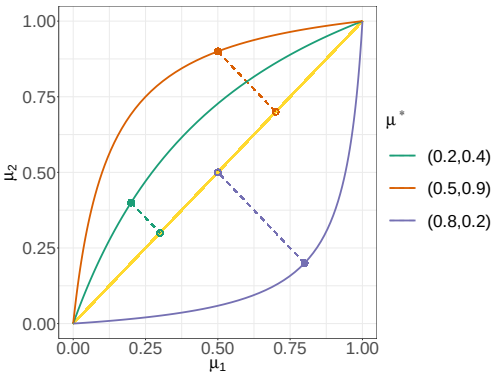



**Figure 5.2:** The family $\mathcal{Q}$ for various $\boldsymbol{\mu}^*$. The coordinate grid represents the parameters of the full 2-sample Bernoulli family, the straight line shows the parameter space of $\mathcal{P}$, the curved lines show the parameters of the distributions in $\mathcal{Q}$, and the dashed lines show the projection of $\boldsymbol{\mu}^*$ onto the parameter space of $\mathcal{P}$.

Again, $\mathcal{P}$ is an exponential family on $\mathcal{U}$ that states that the $k$ samples are i.i.d. Bernoulli, which has sufficient statistic $X = \sum_{i=1}^{k} Y_i$. Its mean-value space is given

by $\mathtt{M}_p = (0, k)$, since the sum of $k$ i.i.d. bernoulli random variables with parameter $\mu$ has a binomial distribution with parameters $(k, \mu)$. Under the alternative $Q$, the $k$ samples are independently Bernoulli distributed with means given by $\boldsymbol{\mu} \in (0,1)^k$, in which case the sum has mean $\mu^* = \sum_{i=1}^k \mu_i$. When constructing the family $\mathcal{Q}$ as in (5.3), it can be verified that $Q_{\beta,\mu^*}$ is the product of Bernoulli distributions with means

$$\left( \frac{e^\beta \mu_1}{1 - \mu_1 + e^\beta \mu_1}, \ldots, \frac{e^\beta \mu_k}{1 - \mu_k + e^\beta \mu_k} \right). \tag{5.14}$$

This family of distributions is illustrated in Figure 5.2 for different choices of $\boldsymbol{\mu}^*$. Seen as a function of $\beta$, all entries in (5.14) behave as sigmoid functions, so that the sum takes values in $(0, k)$. It follows that the mean-value space of $\mathcal{Q}$ is given by $\mathtt{M}_q = (0, k)$, which equals $\mathtt{M}_p$ — ana also, the canonical spaces are all equal to $\mathbb{R}$. Furthermore, the normalizing constant $Z_q(\beta; \mu^*)$ of $\mathcal{Q}$ must be given by

$$Z_q(\beta; \mu^*) = \prod_{i=1}^k (1 - \mu_i + \mu_i e^\beta).$$

We will now verify that item 4 of Theorem 5.3 is satisfied by doing a similar construction for arbitrary $\mu \in (0, k)$. The element in $\mathcal{P}$ with mean $\mu$ corresponds to Bernoulli parameter $\mu/k$, so that we have

$$Z_p(\beta; \mu) = \mathbb{E}_{P_{\mu^*}} \left[ \exp\left( \beta \sum_{i=1}^k Y_i \right) \right] = \left( 1 - \frac{\mu}{k} + \frac{\mu}{k} e^\beta \right)^k.$$

Furthermore, there is a corresponding $\boldsymbol{\mu}' \in (0,1)^k$ such that $\sum_{i=1}^k \mu_i' = \mu$ and $\boldsymbol{\mu}'$ can be written as (5.14) for a specific $\beta$. Repeating the reasoning above gives

$$Z_q(\beta; \mu) = \prod_{i=1}^k (1 - \mu_i' + \mu_i' e^\beta).$$

By concavity of the logarithm, it holds that

$$\log Z_p(\beta; \mu) = k \log \left( 1 - \frac{\mu}{k} + \frac{\mu}{k} e^\beta \right) \geq \sum_{i=1}^k \log(1 - \mu_i' + \mu_i' e^\beta) = \log Z_q(\beta; \mu).$$

We can therefore conclude that $q(U)/p_{\mu^*}(U)$ is the GRO e-variable with respect to $Q$.

Several other exponential families for $k$-sample testing, such as exponential distributions, Gaussian scale, and beta, were investigated in Chapter 4, but none of these

gave rise to a simple e-value. Parts 1-4 of Theorem 5.3 provide some insight into what separates these families from the Gaussian location, Poisson, and Bernoulli.

### Gaussian Scale Family

Another setting in which Theorem 5.3 applies is where $\mathcal{P}$ equals the Gaussian scale family with fixed mean, which we take to be 0 without loss of generality. That is, $\mathcal{P} = \{P_{\sigma^2} : \sigma^2 \in \mathtt{M}_p\}$ where $P_{\sigma^2}$ is the normal with mean 0 and variance $\sigma^2$, i.e.

$$p_{\sigma^2}(U) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{1}{2\sigma^2}U^2}. \tag{5.15}$$

We will substantially extend this null hypothesis, and hence this example, in Section 5.4.4. For now, note that $\mathcal{P}$ is an exponential family with sufficient statistic $X = U^2$, mean-value parameter $\sigma^2$ and mean-value space given by $\mathtt{M}_p = \mathbb{R}^+$. The canonical parameterization of the null relative to any mean-value $\sigma^2 \in \mathtt{M}_p$ is given by

$$p_{\beta;\sigma^2}(U) = \frac{1}{Z_p(\beta;\sigma^2)} \cdot e^{\beta U^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-U^2/(2\sigma^2)} \tag{5.16}$$

with canonical parameter space $\mathtt{B}_{p;\sigma^2} = (-\infty, 1/(2\sigma^2))$.

As alternative, we take $Q$ to be a Gaussian distribution with some fixed mean $m \neq 0$ and variance $s^2$. We use $m$ and $s^2$ instead of $\mu$ and $\sigma^2$ here to avoid confusion with the mean-value parameters of $\mathcal{P}$. The expected value of $X$ under $Q$ is given by $\sigma^{*2} := \mathbb{E}_Q[X] = s^2 + m^2$. The family $\mathcal{Q} = \{Q_\beta : \beta \in \mathtt{B}_{q;\sigma^{*2}}\}$ as defined by (5.3) therefore becomes:

$$q_{\beta;\sigma^{*2}}(U) = \frac{1}{Z_q(\beta;\sigma^{*2})} \cdot e^{\beta U^2} \cdot \frac{1}{\sqrt{2\pi}s} \cdot e^{-c(U-m)^2}, \tag{5.17}$$

where $c = 1/(2s^2)$, with $\mathtt{B}_{q;\sigma^{*2}} = (-\infty, c)$. Comparing (5.16) and the above confirms that $\mathcal{Q}$ is an exponential family that has the same sufficient statistic, namely $U^2$, as $\mathcal{P}$, but different carrier.

The normalizing constant $Z_q$ can be computed using (for example) the moment generating function of the noncentral chi-squared.

$$Z_q(\beta;\sigma^{*2}) = \mathbb{E}_Q\left[e^{\beta U^2}\right] = \mathbb{E}_Q\left[e^{\beta s^2 (\frac{U}{s})^2}\right] = (1-2\beta s^2)^{-1/2} \exp\left(\frac{m^2\beta}{1-2\beta s^2}\right),$$

where we use that $(U/s)^2$ has noncentral chi-squared distribution with one degree of freedom and noncentrality parameter $m^2/s^2$. Plugging this back in (5.17) shows

that $q_{\beta,\sigma^{*2}}$ is a normal density with mean $cm/(c-\beta)$ and variance $1/(2(c-\beta)) = s^2/(1-2\beta s^2)$. This gives

$$\mathbb{E}_{Q_{\beta;\sigma^{*2}}}[U^2] = \frac{2c^2m^2 - (\beta - c)}{2(\beta - c)^2} \tag{5.18}$$

The mean-value parameter space of $\mathcal{Q}$ is thus given by $\mathtt{M}_q = \{\mathbb{E}_{Q_{\beta;\sigma^{*2}}}[U^2], \beta < c\} = \mathbb{R}^+$ which is equal to $\mathtt{M}_p$. Thus, this constructed family does not equal the natural choice of composite alternative that $Q$ was also chosen from, i.e. the (two-dimensional) set of all Gaussians with arbitrary variance mean unequal to zero. However, it does correspond to a specific one-dimensional subset thereof, as was illustrated in Figure 5.1 in the introduction.

Since $\mathtt{M}_q = \mathtt{M}_p$, we get, via Proposition 5.4 that a simple e-variable w.r.t. $Q$ exists if, for all $\sigma^2 > 0$, we have that $\textsc{var}_{P_{\sigma^2}}[U^2] \geq \textsc{var}_{Q_{\sigma^2}}[U^2]$. We now show this to be the case. We have

$$\textsc{var}_{P_{\sigma^2}}[U^2] = 2\sigma^4 = 2(\mathbf{E}_{P_{\sigma^2}}[U^2])^2 = 2(\mathbf{E}_{Q_{\sigma^2}}[U^2])^2.$$

It is therefore sufficient to check whether, for all $\sigma^2 > 0$, it holds that $\textsc{var}_{Q_{\sigma^2}}[U^2] \leq 2(\mathbb{E}_{Q_{\sigma^2}}[U^2])^2$. We can either verify this using existing results by noting that, no matter how $m$ and $s^2$ were chosen, $U^2$ has a noncentral $\chi^2$-distribution under each $Q_{\sigma^2}$, for which it is known that the inequality holds. We can also easily verify it explicitly now that we have already found an expression for $Z_q(\beta;\sigma^{*2})$: since there is no more mention of the null hypothesis, it is equivalent to check whether for each $\beta \in \mathtt{B}_{q;\sigma^{*2}}$ we have

$$\textsc{var}_{Q_{\beta;\sigma^{*2}}}[U^2] \leq 2\left(\mathbb{E}_{Q_{\beta,\sigma^{*2}}}[U^2]\right)^2.$$

To this end, the variance function in terms of $\beta$ can be computed as

$$\textsc{var}_{Q_{\beta;\sigma^{*2}}}[U^2] = \frac{d^2}{d\beta^2}\log Z_q(\beta;\sigma^{*2}) = -\frac{4c^2m^2 - (\beta - c)}{2(\beta - c)^3}. \tag{5.19}$$

Comparing this to (5.18) shows that the condition above indeed holds, from which we can conclude that $q(U)/p_{\sigma^{*2}}(U)$ is an e-value.

Finally, note that even though the mean-value parameter spaces of $\mathcal{P}$ and $\mathcal{Q}$ are equal, the canonical spaces are not: $\mathtt{B}_{p;\sigma^{*2}}$ is a proper subset of $\mathtt{B}_{q;\sigma^{*2}}$. More generally, for any $\sigma'^2 > 0$ different from the $\sigma^{*2}$ we started with, the canonical spaces $\mathtt{B}_{p;\sigma'^2}$ and $\mathtt{B}_{q;\sigma'^2}$ both change but remain unequal. Still, Proposition 5.4 ensures that we will have $\mathtt{B}_{p;\sigma'^2} \subset \mathtt{B}_{q;\sigma'^2}$.

### NEFS and their Variance Functions

In this section, we consider the setting where $\mathcal{P}$ is a one-dimensional natural exponential family (NEF) and $Q$ is also an element of an NEF. This setting is particularly suited for the analysis above, because the constructed family $\mathcal{Q}$ can be seen to equal the NEF that $Q$ was chosen from. We therefore do not differentiate between the simple or composite alternative in this section. Furthermore, NEFs are fully characterized by the pair $(\sigma^2(\mu), \mathtt{M})$, where $\mathtt{M}$ is the mean-value parameter space and $\sigma^2(\mu)$ is the variance function as defined before. A wide variety of NEFS and their corresponding variance functions have been studied in the literature (see e.g. Morris, 1982; Jørgensen, 1997; Bar-Lev et al., 2024) and this can be used in conjunction with Theorem 5.3 to quickly check on a case-by-case basis whether any given pair of NEFs provides a simple e-variable.

For example, let $\mathcal{P} = \{P_{\lambda,r} : \lambda \in \mathbb{R}^+\}$ be the set of Gamma distributions for $U$ with varying scale parameter $\lambda$ and fixed shape parameter $r > 0$. The sufficient statistic is given by $X = U$ and its mean under $P_{\lambda,r}$ equals $r\lambda$, so the mean-value parameter space is $\mathtt{M}_p = \mathbb{R}^+$. The variance function is given by $\sigma_p^2(\mu) = \mu^2/r$. If we set $Q$ to $P_{\lambda^*,r'}$ for specific $\lambda^*, r' \in \mathbb{R}^+$, then $\mathcal{Q}$ is the set of Gamma distributions with fixed shape parameter $r'$.

Similarly, let $\mathcal{P}$ be the set of negative binomial distributions with fixed number of successes $n \in \mathbb{N}$ and let $Q$ be any Poisson distribution, so that $\mathcal{Q}$ equals the Poisson family. The variance functions are given by $\sigma_p^2(\mu) = \mu^2/n + \mu$ and $\sigma_q^2(\mu) = \mu$, respectively. It is trivially true that $\sigma_p^2(\mu) \geq \sigma_q^2(\mu)$ for all $\mu$, so Theorem 5.3 reveals that a simple e-variables exists with respect to any element of the Poisson family. More generally, we may look at the Awad-Bar-Lev-Makov (ABM) class of NEFs (Bar-Lev and Ridder, 2021; Awad et al., 2022; Bar-Lev and Ridder, 2023) that are characterized by mean-value parameter space $\mathtt{M} = \mathbb{R}^+$ and variance function

$$\sigma_s^2(\mu) = \mu\left(1 + \frac{\mu}{s}\right)^r, \ s > 0, \ r = 0, 1, 2, ...$$

This class was proposed as part of a general framework for alternatives to the Poisson model (which would arise for $r = 0$) that are zero-inflated and over-dispersed. The case $r = 1$ recovers the negative binomial distribution and $r = 2$ is called the generalized Poisson or Abel distribution. As was the case for the negative binomial distribution, it follows from Theorem 5.3 that simple e-variables exist for testing any of the ABM NEFs against the Poisson model.

Much more generally, consider the Tweedie-Bar-Lev-Enis class (Bar-Lev, 2020) of

NEFs that have mean-value space $\mathtt{M} = \mathbb{R}^+$ and power variance functions

$$\sigma^2(\mu) = a\mu^\gamma,\ a > 0,\ \mu > 0,\ \gamma \geq 1.$$

We require $\gamma \geq 1$ because there are no families of this form with $\gamma \in (0, 1)$ and while there are families in this class with $\gamma < 0$, they are not regular and therefore beyond the scope of this chapter. The cases $\gamma = 1$ (Poisson) and $\gamma = 2$ (Gamma families, with $a$ depending on the shape parameter) were already encountered above. If we test between two of such families, say $\mathcal{P}$ with $\sigma_p^2(\mu) = a_p\mu^{\gamma_p}$ and $\mathcal{Q}$ with $\sigma_q^2(\mu) = a_q\mu^{\gamma_q}$ that share the same underlying sample space, there do not exist simple e-variables in general. Indeed, we have that $\sigma_p^2(\mu) \geq \sigma_q^2(\mu)$ if and only if $\mu^{\gamma_p - \gamma_q} \geq a_q/a_p$, which, for certain combinations of parameters, does not hold for all $\mu \in \mathtt{M}$. Since this condition might hold for some $\mu$ but not for others, this suggests that there may be cases where we find local e-variables that are not global.

Let us investigate this for $(a_p, \gamma_p) = (1, 2)$ and $(a_q, \gamma_q) = (1/2, 3)$, which corresponds to the family of exponential distributions and the family of inverse Gaussian distributions with shape parameter $\lambda := a_q^{-1} = 2$ respectively. In this case, it holds that $\sigma_p^2(\mu) \geq \sigma_q^2(\mu) \Leftrightarrow \mu \leq a_q^{-1}$. It follows from the analysis in Section 5.2 that $q_\mu(U)/p_\mu(U)$ is a local e-variable for $\mu \leq a_q^{-1}$. However, since the condition does not hold for all $\mu$ we cannot use Proposition 5.4 (or, equivalently, because, as we will see, the preconditions for Theorem 5.3 do not hold), this need not necessarily also be a global e-variable. In fact, the expected value under $\mu' \in \mathtt{M}$ is given by

$$\mathbb{E}_{U\sim P_{\mu'}}\left[\frac{q_\mu(U)}{p_\mu(U)}\right] = \int_0^\infty \frac{1}{\mu'}\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x} + \frac{x}{\mu} - \frac{x}{\mu'}\right)dx, \tag{5.20}$$

which diverges for $\mu' \geq (1/\mu - \lambda/(2\mu^2))^{-1}$. The latter is vacuous for $\mu \leq \lambda/2$, which means that for such $\mu$ we might still get a global e-variable. For $\mu \in (\lambda/2, \lambda)$, this shows that we will get a local e-variable that is not a global e-variable. These different regimes are illustrated in Figure 5.3. For $\mu > 1$, the lines stop when the integral in (5.20) starts diverging. To see how the potential divergence (for large enough $\mu'$, in the regime $1 < \mu < 2$) plays out in terms of the function $f$ in (5.8), consider for example $\mu = 3/2$. Then, as is immediate from the definition of exponential distributions and the inverse Gaussian density with $\lambda = 2$ we have $q_{\beta;\mu}(x) \propto \exp((\beta - 4/9)x)h(x)$ with $h$ the probability density on $\mathbb{R}^+$ given by $h(x) = \sqrt{1/(\pi x^3)}\exp(-1/x)$, whereas $p_{\beta;\mu} \propto \exp((\beta - 2/3)x)$. We see that $\mathtt{B}_{p;\mu} = (-\infty, 6/9)$ and $\mathtt{B}_{q;\mu} = (-\infty, 4/9)$. Thus, as $\beta \uparrow 4/9$, we get that $\log Z_p(\beta)$ converges to a finite constant whereas $\log Z_q(\beta) \uparrow \infty$,

so that $f(\beta, \mu) \to \infty$, with $f$ the function in (5.8), as it should.

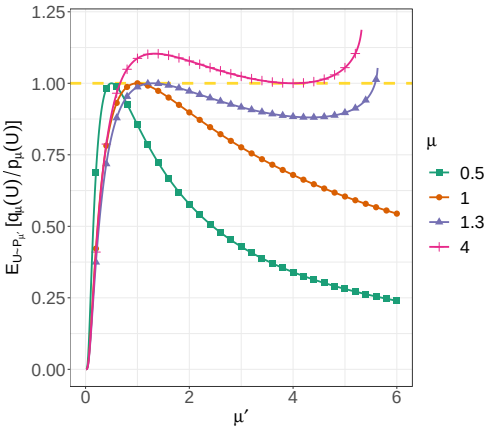



**Figure 5.3:** The expected value of $q_\mu(U)/p_\mu(U)$ under the null $P_{\mu'}$ for varying $\mu'$.

### 5.4.4 The Linear Model

We now show that Theorem 5.3 allows us to conclude that simple e-variables exist for the linear model, i.e. standard linear regression with Gaussian noise, where the null hypothesis $\mathcal{P}$ is a subset of the alternative $\mathcal{H}_1$ obtained by setting the regression parameter of a control random variable to 0, as soon as we allow the variance in $\mathcal{P}$ to be a free parameter. This was shown directly, without associating a specific family $\mathcal{Q}$ to $\mathcal{P}$, in an unpublished master thesis (De Jong, 2021). De Jong's treatment involved a lot of hard-to-interpret calculus, much of it discovered by trial-and-error. The advantage of the present treatment is that Theorem 5.3 clearly guides the reasoning and suggests what formulas to verify. The setting is really a vast extension of that of Section 5.4.3, which is (essentially) retrieved if below we set $d = 0$. Interestingly, e-variables for linear models were already derived by Pérez-Ortiz et al. (2024) and Lindon et al. (2024), based on right-Haar priors. The current approach provides a different type of e-variable which has the advantage that it does not require the variance under the alternative to be equipped with a right-Haar prior: while for convenience we give the treatment below for $\mathcal{H}_1$ with the variance $\sigma^2$ being left a free parameter, we can freely apply the results to any $\mathcal{H}_1' \subset \mathcal{H}_1$, in particular with $\mathcal{H}_1'$ restricted to densities with a fixed variance. The price to pay is that the e-variables derived below, while growth-optimal for the fixed $Q \in \mathcal{H}_1$ relative to which they are defined, will in general not be GROW (*worst-case* growth optimal, see (Grünwald et al., 2024)) in the worst-case

over all distributions in $\mathcal{H}_1$ when $\sigma^2$ varies within $\mathcal{H}_1$.

Assume then that data arrives as a block of outcomes together with given covariate vectors, i.e. $U = ((Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n))$ with $Y_i \in \mathbb{R}$ and $\boldsymbol{x}_i = (x_{i,0}, x_{i,1}, \ldots, x_{i,d})^T \in \mathbb{R}^{d+1}$. Define the conditional normal distributions $G_{\sigma,\boldsymbol{\gamma}}$ with corresponding densities

$$g_{\sigma,\boldsymbol{\gamma}}(Y^n) := g_{\sigma,\boldsymbol{\gamma}}(Y^n \mid \boldsymbol{x}^n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum (Y_i - \nu_i)^2} \tag{5.21}$$

with $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_d)^T \in \mathbb{R}^{d+1}$ and

$$\nu_i := \boldsymbol{\gamma}^T \boldsymbol{x}_i. \tag{5.22}$$

Here and in the sequel, sums without explicitly denoted ranges are invariably taken to be over $i = 1..n$ and we omit the conditional $\boldsymbol{x}^n$ from the notation, since they are fixed throughout the following analysis.

We focus on the most common case in which one of the covariates, $x_{i,0}$, has a special status and we want to test whether the corresponding coefficient $\gamma_0$ is equal to 0. We thus want to design an e-variable for testing any simple alternative $Q$ taken from the full alternative hypothesis $\mathcal{H}_1$ vs. the null $\mathcal{P}$, where $\mathcal{H}_1$ and $\mathcal{P}$ are respectively given by:

$$\mathcal{H}_1 = \{G_{\sigma,\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \mathbb{R}^{d+1}, \gamma_0 \neq 0, \sigma > 0\} \quad ; \quad \mathcal{P} = \{G_{\sigma,\boldsymbol{\gamma}} : \boldsymbol{\gamma} \in \mathbb{R}^{d+1}, \gamma_0 = 0, \sigma > 0\}. \tag{5.23}$$

We make the standard assumption that $n \geq d+1$ and that the matrix $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ has maximal (i.e. $d+1$) rank. Now define the transformed parameters $\lambda := -1/(2\sigma^2)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ with, for $j = 1..d$, $\beta_j := \gamma_j/\sigma^2$ and $\theta := \gamma_0/\sigma^2$ and set $t_j(Y^n) = \sum Y_i x_{i,j}$. Rewriting the likelihood (5.21) in terms of this new parameterization and the $t_j$, denoting the resulting densities by $f^{(\theta)}_{\lambda,\boldsymbol{\beta}}$, we see that

$$f^{(\theta)}_{\lambda,\boldsymbol{\beta}}(y^n) = g_{\sigma,\boldsymbol{\gamma}}(y^n) = \exp\left(\lambda \sum y_i^2 + \theta t_0(y^n) + \sum_{j=1}^{d} \beta_j t_j(y^n)\right) \cdot h_1(y^n) h_2(\sigma, \boldsymbol{\gamma}) \tag{5.24}$$

for some function $h_1$ not depending on the parameters and $h_2$ not depending on the data $y^n$. Let, for $\theta \in \mathbb{R}$, $\mathcal{Q}^{(\theta)}$ be the set of distributions $F^{(\theta)}_{\lambda,\boldsymbol{\beta}}$ with densities $f^{(\theta)}_{\lambda,\boldsymbol{\beta}}$. We see that for each $\theta \in \mathbb{R}$, $\mathcal{Q}^{(\theta)}$ is a $(d+1)$-dimensional exponential family with sufficient

statistic vector

$$\left(\sum Y_i^2, t_1(Y^n), \ldots, t_d(Y^n)\right). \tag{5.25}$$

and mean-value parameter space $\mathtt{M}_q^{(\theta)} = (0,\infty) \times \mathbb{R}^d$. The original parameter vector corresponding to $(\lambda, \boldsymbol{\beta})$ is $(\sigma^2, \boldsymbol{\gamma})$ with $\sigma^2 := -1/(2\lambda)$ and $\boldsymbol{\gamma} = (\sigma^2\theta, \sigma^2\beta_1, \ldots, \sigma^2\beta_d)$ and the corresponding mean-value parameter vector is

$$\boldsymbol{\mu} := \left(n\sigma^2 + \sum \nu_i^2, \sum x_{i,1}\nu_i, \ldots, \sum x_{i,d}\nu_i\right)^T. \tag{5.26}$$

with $\nu_i$ as in (5.22). Observe that $\mathcal{H}_1 = \bigcup_{\theta\in\mathbb{R}\setminus\{0\}} \mathcal{Q}^{(\theta)}$ and $\mathcal{P} = \mathcal{Q}^{(0)}$. For expository convenience, we slightly deviated from our previous notation here by having a canonical parameter space vector of the form $(\lambda, \boldsymbol{\beta})$ rather than $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$; thus $\boldsymbol{\beta}$ is $d$-dimensional but $\boldsymbol{\mu}$ still represents a full $(d+1)$-dimensional mean-value parameter.

Having established that $\mathcal{Q}^{(\theta)}$ and $\mathcal{P}$ are, indeed, exponential families, we will now show that Theorem 5.3 in the form of Corollary 5.5 is applicable to them. Thus, fix arbitrary $Q^\circ \in \mathcal{H}_1$. We must have that $Q^\circ \in \mathcal{Q}^{(\theta^\circ)}$ for some $\theta^\circ$ and the density of $Q^\circ$ can be written as $f^{(\theta^\circ)}_{\lambda^\circ,\boldsymbol{\beta}^\circ}$ or equivalently as $g_{\sigma^\circ,\boldsymbol{\gamma}^\circ}$ with $\sigma^\circ$, $\boldsymbol{\gamma}^\circ$ and $\nu^\circ$ related to $\theta^\circ, \lambda^\circ$ and $\boldsymbol{\beta}^\circ$ in the same way as before, in particular $\nu_i^\circ = \boldsymbol{\gamma}^{\circ T}\boldsymbol{x}_i$ (we can now see how this example extends Section 5.4.3: using the notation from that example, i.e. $m$ the mean of $U$ and $s^2$ its variance under $Q$, we set $n = 1$, $d = 0$, $\boldsymbol{x}_1 = 1$, $\nu_1^\circ = \gamma_0^\circ = m$ and $\sigma^{\circ 2} = s^2$).

Simple differentiation gives that the element in $\mathcal{P}$ that minimizes KL divergence, i.e. achieves $\min_{P\in\mathcal{P}} D(Q\|P)$, is given by $P = G_{\boldsymbol{\gamma}^*,\sigma^*}$ with parameters $\sigma^{*2}$ and $\boldsymbol{\gamma}^* = (0, \gamma_1^*, \ldots, \gamma_d^*)$ where $\boldsymbol{\gamma}^*$ is a Euclidean projection and $\sigma^{*2}$ is related to this projection via

$$\sigma^{*2} = \min_{(\gamma_1,\ldots,\gamma_d)\in\mathbb{R}^d} \frac{1}{n}\mathbf{E}_Q\left[\sum (Y_i - \sum_{j=1}^d \gamma_j \boldsymbol{x}_{i,j})^2\right]$$

This link to Euclidean projection implies, upon setting $\nu_i^* := \boldsymbol{\gamma}^{*T}\boldsymbol{x}_i$ the following easily derivable consequences:

$$\begin{aligned} &\text{for all } j \in \{1,\ldots,d\}\text{: } \textstyle\sum \nu_i^\circ x_{i,j} = \sum \nu_i^* x_{i,j} \\ &\sigma^{*2} = \sigma^{\circ 2} + \frac{1}{n}\sum(\nu_i^* - \nu_i^\circ)^2 = \sigma^{\circ 2} + \frac{1}{n}\left(\sum \nu_i^{*2} - \sum \nu_i^{\circ 2}\right), \end{aligned} \tag{5.27}$$

where we note that (5.27) may be seen as versions of the standard *normal equations* in linear regression analysis. Again we define $\lambda^*, \boldsymbol{\beta}^*, \boldsymbol{\mu}^*$ correspondingly as above, in

particular $\boldsymbol{\mu}^*$ is given in terms of $\sigma^*$ and $\nu^*$ via (5.26) .

We now simply follow the steps needed to apply Theorem 5.3 in the form of Corollary 5.5. First, we reparameterize $\mathcal{P}$ in terms of the specific canonical parameterization in which $(\lambda, \boldsymbol{\beta}) = 0$ must correspond to $G_{\sigma^*,\boldsymbol{\gamma}^*}$. We obtain:

$$p_{\lambda,\boldsymbol{\beta};\boldsymbol{\mu}^*}(y^n) = \frac{1}{Z_p(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)} \cdot \exp\left(\lambda \sum y_i^2 + \sum_{j=1}^{d} \beta_j t_j(y^n)\right) f^{(0)}_{\lambda^*,\boldsymbol{\beta}^*}(y^n), \tag{5.28}$$

with $Z_p(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$ the normalizing constant, defined for all $(\lambda, \boldsymbol{\beta}) \in \mathtt{B}_{p;\boldsymbol{\mu}^*}$ where

$$\mathtt{B}_{p;\boldsymbol{\mu}^*} = \{(\lambda, \boldsymbol{\beta}) : Z_p(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*) < \infty\} = (-\infty, -\lambda^*) \times \mathbb{R}^d.$$

We see that this family coincides with $\mathcal{P}$. Similarly, relative to our fixed $(\theta^\circ, \lambda^\circ, \boldsymbol{\beta}^\circ)$ we define the family with densities

$$q^{(\theta^\circ)}_{\lambda,\boldsymbol{\beta}}(y^n; \boldsymbol{\mu}^*) = \frac{1}{Z_q^{(\theta^\circ)}(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)} \cdot \exp\left(\lambda \sum y_i^2 + \sum_{j=1}^{d} \beta_j t_j(y^n)\right) f^{(\theta^\circ)}_{\lambda^\circ,\boldsymbol{\beta}^\circ}(y^n), \tag{5.29}$$

with normalizing constant $Z_q^{(\theta^\circ)}(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$. We see that this family coincides with $\mathcal{Q}^{(\theta^\circ)}$ and has canonical parameter space $\mathtt{B}^{(\theta^\circ)}_{q;\boldsymbol{\mu}^*} = (-\infty, -\lambda^\circ) \times \mathbb{R}^d$.

To apply Corollary 5.5, we need to verify that for each choice of $\theta^\circ$, we have (i) $\mathtt{M}_q^{(\theta^\circ)} \subseteq \mathtt{M}_p$, and (ii) for each $\boldsymbol{\mu} \in \mathtt{M}_q^{(\theta^\circ)}$, we have that $\mathtt{B}_{p;\boldsymbol{\mu}} \subseteq \mathtt{B}^{(\theta^\circ)}_{q;\boldsymbol{\mu}}$. We already verified $\mathtt{M}_q^{(\theta^\circ)} = \mathtt{M}_p$, implying (i), further above. As to (ii), note that the inclusion holds for $\boldsymbol{\mu} = \boldsymbol{\mu}^*$ since, using (5.27), $-\lambda^* = (1/2\sigma^{*2}) \leq (1/2\sigma^{\circ 2}) = -\lambda^\circ$. We next note that for each $\theta^\circ \in \mathbb{R}$, there is a 1-to-1 correspondence between the choice $(\lambda^\circ, \boldsymbol{\beta}^\circ) \in (-\infty, 0) \times \mathbb{R}^d$ used to determine $f^{(\theta^\circ)}_{\lambda^\circ,\boldsymbol{\beta}^\circ}$ and the resulting $\boldsymbol{\mu}^* \in \mathtt{M}_p$. We thus see that we can obtain the desired inclusion for arbitrarily chosen $\boldsymbol{\mu} \in \mathtt{M}_p$ by picking $(\lambda^\circ, \boldsymbol{\beta}^\circ)$ such that $\boldsymbol{\mu}^*$ becomes equal to this $\boldsymbol{\mu}$. This shows that (ii) holds for all $\boldsymbol{\mu} \in \mathtt{M}_p = \mathtt{M}_q^{(\theta^\circ)}$. Corollary 5.5 now gives the following: for all $\gamma^\circ \in \mathbb{R}^d$ with $\gamma_0^\circ \neq 0$, all $\sigma^\circ > 0$, we have that $g_{\sigma^\circ,\gamma^\circ}(Y^n)/g_{\sigma^*,\gamma^*}(Y^n) = f^{(\theta^\circ)}_{\lambda^\circ,\boldsymbol{\beta}^\circ}(Y^n)/f^{(0)}_{\lambda^*,\boldsymbol{\beta}^*}(Y^n)$ is the GRO e-variable relative to $G_{\sigma^\circ,\gamma^\circ}$ if $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathbb{R}^d$. But this condition is readily established to hold: we do so in Appendix C.1.

## 5.5 Proof of Theorem 5.3

To get some intuition first, we note that the distributions $P_{\boldsymbol{\beta}}$ and $Q_{\boldsymbol{\beta}}$ indexed by the $\boldsymbol{\beta}$ in the definition of $f(\boldsymbol{\beta}; \boldsymbol{\mu}^*)$, i.e. (5.8), are difficult to compare in the sense that they do not necessarily have any properties in common. In particular, $P_{\boldsymbol{\beta}}$ generally does not achieve $\min_{P \in \mathcal{P}} D(Q_{\boldsymbol{\beta}} \| P)$, so that $P_{\boldsymbol{\beta}}$ and $Q_{\boldsymbol{\beta}}$ do not have the same mean. This suggests to replace $f(\boldsymbol{\beta}; \boldsymbol{\mu}^*)$ by a function $g(\boldsymbol{\mu}; \boldsymbol{\mu}^*)$ on the mean-value parameter space and also to re-express $f(\boldsymbol{\beta}; \boldsymbol{\mu}^*) \leq 0$, the condition for being an e-variable, by a condition on $g$ — and this is what we do in the proof of Theorem 5.3: inside the proof below we establish, using well-known convex duality properties of exponential families, that this can be done with function and condition, respectively, given by:

$$g(\boldsymbol{\mu}; \boldsymbol{\mu}^*) = D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}^*}) - D(Q_{\boldsymbol{\mu}} \| Q_{\boldsymbol{\mu}^*}), \tag{5.30}$$

$$g(\boldsymbol{\mu}; \boldsymbol{\mu}^*) \leq 0. \tag{5.31}$$

This condition on $g$ corresponds to item 3 in Theorem 5.3. The key insight for showing the suitability of $g$ is the following well-known convex-duality fact about exponential families: for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_p$, all $\boldsymbol{\beta} \in \mathtt{B}_{p;\boldsymbol{\mu}^*}$, we have:

$$-\log Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}') = D(P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')} \| P_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T \boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') \leq D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T \boldsymbol{\mu}. \tag{5.32}$$

This can be derived as follows:

$$\begin{aligned}
D(P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')} \| P_{\boldsymbol{\mu}'}) - D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}'}) &= \log \frac{Z_p(\beta_p(\boldsymbol{\mu}; \boldsymbol{\mu}'))}{Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}')} + \boldsymbol{\beta}^T \boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') - \boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}')\boldsymbol{\mu} \\
&= \log \frac{Z_p(\beta_p(\boldsymbol{\mu}; \boldsymbol{\mu}'))}{Z_p(\boldsymbol{\beta}; \boldsymbol{\mu}')} + \boldsymbol{\beta}^T (\boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') - \boldsymbol{\mu}) - (\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}') - \boldsymbol{\beta})^T \boldsymbol{\mu} \\
&= \boldsymbol{\beta}^T (\boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') - \boldsymbol{\mu}) - D(P_{\boldsymbol{\mu}} \| P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')}) \\
&\leq \boldsymbol{\beta}^T (\boldsymbol{\mu}_p(\boldsymbol{\beta}; \boldsymbol{\mu}') - \boldsymbol{\mu}).
\end{aligned}$$

We now prove the chain of implications in the theorem.

(1) $\Rightarrow$ (2) Let $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$ and denote $\boldsymbol{\mu}(\alpha) := (1-\alpha)\boldsymbol{\mu}' + \alpha\boldsymbol{\mu}$. By assumption of convexity, we have that $\boldsymbol{\mu}(\alpha) \in \mathtt{M}_q$ for all $\alpha \in [0, 1]$. Furthermore, define $h(\alpha) = (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha); \boldsymbol{\mu}'))^T (\boldsymbol{\mu}(\alpha) - \boldsymbol{\mu}')$, so that $h(0) = 0$ and $h(1) = (\boldsymbol{\beta}_p(\boldsymbol{\mu}; \boldsymbol{\mu}') -$

$\boldsymbol{\beta}_q(\boldsymbol{\mu};\boldsymbol{\mu}'))^T(\boldsymbol{\mu}-\boldsymbol{\mu}')$. The derivative of $h$ is given by

$$\frac{d}{d\alpha}h(\alpha) = \left(\frac{d}{d\alpha}\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}')\right)^T(\boldsymbol{\mu}(\alpha)-\boldsymbol{\mu}') + (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}'))^T\frac{d}{d\alpha}\boldsymbol{\mu}(\alpha).$$

The chain rule gives

$$\frac{d}{d\alpha}\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}') = \Sigma_p^{-1}(\boldsymbol{\mu}(\alpha))^T(\boldsymbol{\mu}-\boldsymbol{\mu}'),$$

where we use (5.10) and (5.11) together with the fact that the Jacobian of the gradient of a function equals the transpose of its Hessian. The derivative of $\boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}')$ can be found with the same argument, so we see

$$\begin{aligned}\frac{d}{d\alpha}h(\alpha) =& \left((\Sigma_p^{-1}(\boldsymbol{\mu}(\alpha)) - \Sigma_q^{-1}(\boldsymbol{\mu}(\alpha)))^T(\boldsymbol{\mu}-\boldsymbol{\mu}')\right)^T(\boldsymbol{\mu}(\alpha)-\boldsymbol{\mu}') \\ &+ (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}'))^T(\boldsymbol{\mu}-\boldsymbol{\mu}') \\ =& \frac{1}{\alpha}(\boldsymbol{\mu}(\alpha)-\boldsymbol{\mu}')^T(\Sigma_p^{-1}(\boldsymbol{\mu}(\alpha)) - \Sigma_q^{-1}(\boldsymbol{\mu}(\alpha)))(\boldsymbol{\mu}(\alpha)-\boldsymbol{\mu}') \\ &+ (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}'))^T(\boldsymbol{\mu}-\boldsymbol{\mu}') \\ =& \frac{1}{\alpha}(\boldsymbol{\mu}(\alpha)-\boldsymbol{\mu}')^T(\Sigma_p^{-1}(\boldsymbol{\mu}(\alpha)) - \Sigma_q^{-1}(\boldsymbol{\mu}(\alpha)))(\boldsymbol{\mu}(\alpha)-\boldsymbol{\mu}') + \frac{1}{\alpha}h(\alpha). \qquad (5.33)\end{aligned}$$

If $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu}$, then $\Sigma_p^{-1}(\boldsymbol{\mu}) - \Sigma_q^{-1}(\boldsymbol{\mu})$ is negative semidefinite (as discussed below the statement of Theorem 5.3). In this case, the first term in (5.33) is negative and, since $h(0) = 0$, the second term is also negative on $[0,1]$. It follows that $h$ is decreasing when $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite, so that $(\boldsymbol{\beta}_p(\boldsymbol{\mu};\boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu};\boldsymbol{\mu}'))^T(\boldsymbol{\mu}-\boldsymbol{\mu}')) \leq 0$, as was to be shown.

(2) ⇒ (3) We use a similar argument as was used to prove the previous implication, so let $\boldsymbol{\mu},\boldsymbol{\mu}' \in \mathtt{M}_q$ and denote $\boldsymbol{\mu}(\alpha) = (1-\alpha)\boldsymbol{\mu}'+\alpha\boldsymbol{\mu}$ as before. Define $h(\alpha) := g(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}')$. Using the chain rule of differentiation together with (5.10), we see that the derivative of $h$ is given by

$$\begin{aligned}\frac{d}{d\alpha}h(\alpha) &= (\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}'))^T(\boldsymbol{\mu}-\boldsymbol{\mu}') \\ &= \frac{1}{\alpha}(\boldsymbol{\beta}_p(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}') - \boldsymbol{\beta}_q(\boldsymbol{\mu}(\alpha);\boldsymbol{\mu}'))^T(\boldsymbol{\mu}(\alpha)-\boldsymbol{\mu}').\end{aligned}$$

If item (2) holds, then we have that $\frac{d}{d\alpha}h(\alpha) \leq 0$. Furthermore, since $h(0) = 0$ and $h(1) = D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}'}) - D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}'})$, we see that item (2) implies that

$$D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}'}) - D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}'}) \leq 0,$$

as was to be shown.

(3) $\Rightarrow$ (4) Assume that $D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}'}) - D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}'}) \leq 0$ for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$. Together with (5.32) this gives, for all $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathtt{M}_q$, all $\boldsymbol{\beta} \in \mathtt{B}_{p;\boldsymbol{\mu}'}$:

$$D(P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')}\|P_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}') \leq D(P_{\boldsymbol{\mu}}\|P_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T\boldsymbol{\mu} \leq D(Q_{\boldsymbol{\mu}}\|Q_{\boldsymbol{\mu}'}) - \boldsymbol{\beta}^T\boldsymbol{\mu}. \tag{5.34}$$

Applying this with $\boldsymbol{\mu} = \boldsymbol{\mu}_q(\boldsymbol{\beta};\boldsymbol{\mu}')$ and re-arranging gives

$$-D(P_{\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}')}\|P_{\boldsymbol{\mu}'}) + \boldsymbol{\beta}^T\boldsymbol{\mu}_p(\boldsymbol{\beta};\boldsymbol{\mu}') \geq -D(Q_{\boldsymbol{\mu}_q(\boldsymbol{\beta};\boldsymbol{\mu}')}\|Q_{\boldsymbol{\mu}'}) + \boldsymbol{\beta}^T\boldsymbol{\mu}_q(\boldsymbol{\beta};\boldsymbol{\mu}'), \tag{5.35}$$

which, by the equality in key fact (5.32) is equivalent to $\log Z_p(\boldsymbol{\beta};\boldsymbol{\mu}') \geq \log Z_q(\boldsymbol{\beta};\boldsymbol{\mu}')$, which is what we had to prove.

**Remaining Implications** (4) $\Rightarrow$ (5) now follows by the equality in (5.8) and the definition of an e-variable. (5) $\Rightarrow$ (6) follows from proposition 5.1, (6) $\Rightarrow$ (7) follows because a global e-variable is automatically also a local one, and (7) $\Rightarrow$ (8) again follows from Proposition 5.1. Finally, (8) $\Rightarrow$ (1) has already been established as Proposition 5.2. □

## 5.6 Conclusion and Future Work

We have provided a theorem that, under regularity pre-conditions, provides a general sufficient condition under which there exists a simple e-variable for testing a simple alternative versus a composite regular exponential family null. The characterization was given in terms of several equivalent conditions, the most direct being perhaps the condition '$\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathtt{M}_q$'. A direct follow-up question is: can we construct GRO or close-to-GRO e-variables, in case either the regularity pre-conditions or the positive definiteness condition do *not* hold? The example of Section 5.4.3, and in particular Figure 5.3, indicated that in that case, many things can happen: under some $\boldsymbol{\mu} \in \mathtt{M}_q$ (green curve), $q_{\boldsymbol{\mu}}/p_{\boldsymbol{\mu}}$ still gives a global

simple e-variable; for other $\boldsymbol{\mu}$ (blue), it gives a local but not global e-variable; for yet other $\boldsymbol{\mu}$ (pink), it does not give an e-variable at all.

Nevertheless, it turns out that if the pre-regularity conditions hold and the 'opposite' of the positive semidefinite condition holds, i.e. if $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q(\boldsymbol{\mu})$ is *negative* semidefinite for all $\boldsymbol{\mu} \in \mathtt{M}_q$, then there is again sufficient structure to analyze the problem. The GRO e-variable will now be based on a mixture of elements of the null, but the specific mixture will depend on the sample size: we now need to look at i.i.d. repetitions of $U$ rather than a single outcome $U$. We will provide such an analysis in future work.

Another interesting avenue for future work is to extend the analysis to *curved* exponential families (Efron, 2022). While we do not have any general results in this direction yet, the analysis by Liang (2023) suggests that this may be possible. Liang considers a variation of the Cochran-Mantel- Haenszel test, in which the null hypothesis expresses that the population-weighted *average* effect size over a given set of strata is equal to, or bounded by, some $\delta$. This can be rephrased in terms of a curved exponential family null, for which Liang (2023) shows that a local e-variable exists by considering the second derivative of the function $f(\boldsymbol{\beta}; \boldsymbol{\mu}^*)$ as in (5.8), just like in the present chapter but with $\boldsymbol{\beta}$ representing a particular suitable parameterization rather than the canonical parameterization of an exponential family. The local e-variable is then shown to be a global e-variable by a technique different from the construction of $\mathcal{Q}$ we use here. Still, the overall derivation is sufficiently similar to suggest that it can be unified with the reasoning underlying Theorem 5.3. Finally, the analysis of the linear model in Section 5.4.4 suggests that the results may perhaps be extended to say something about existence of *generalized* linear models without assuming a *model-X* condition (see Chapter 6) — a situation about which currently next to nothing is known.

# 6 | Tests of Conditional Independence Under Model-X

In Chapters 4 and 5, we discovered that there exist simple-vs.-simple likelihood ratios that are $e$-statistics (hence GRO) for certain parametric hypothesis tests involving exponential families. This considerably simplifies the problem of finding the GRO $e$-statistic in such cases. Furthermore, the discussion of the GRO $e$-statistic for the large nonparametric null in Section 4.2.2—called $S_{\mathrm{GRO(IID)}}$ there—reveals that this can also occur for nonparametric null hypotheses. In this chapter, we continue the investigation of nonparametric problems by studying tests of conditional independence of a response $Y$ and a predictor $X$ given a random vector $Z$. Although we do not make any assumptions on the rest of the distribution, our test depends on the availability of the conditional distribution of $X$ given $Z$, or at least a sufficiently sharp approximation thereof. This is known as the model-X setting. Within this setting, we derive a general method for constructing $e$-statistics to test conditional independence. This method leads to GRO $e$-statistics for simple alternatives, because it gives an $e$-statistic in the form of a simple-vs.-simple likelihood ratio. Furthermore, we prove that our method yields tests with asymptotic power one in the special case of a logistic regression model. A simulation study is done to demonstrate that the approach is competitive in terms of power when compared to established sequential and nonsequential testing methods, and robust with respect to violations of the model-X assumption.

## 6.1 Introduction

A fundamental task in many areas of research, such as economics, genetics, and pharmacology, is to find out whether there is an association between a response $Y$ and an explanatory variable $X$, given a vector of covariates $Z$. Mathematically, absence of such an association is defined as conditional independence (CI) between $X$ and $Y$ given $Z$, denoted by $X \perp\!\!\!\perp Y \mid Z$ (Dawid, 1979). A standard way to tackle these problems is to assume a (semi)parametric model on $Y$ given $X$ and $Z$, encoding the dependence between $Y$ and $X$ in a model parameter. For example, in the logistic model the probability that a binary random variable $Y$ equals one is regressed on $(X, Z)$, and the regression coefficient corresponding to $X$ is zero if and only if $X \perp\!\!\!\perp Y \mid Z$. Within these parametric models, there are well established methods to test conditional independence, such as the likelihood ratio test for generalized linear models (McCullagh and Nelder, 1989). However, all of these tests have in common that they fail to uphold a type-I error guarantee when the model assumptions are not satisfied. If $Z$ is continuously distributed such error inflation is in fact unavoidable: unless further assumptions on the distribution of $(X, Y, Z)$ are imposed, there exist no nontrivial tests of CI which maintain type-I error guarantees (Shah and Peters, 2020).

However, Candès et al. (2018) show that given additional knowledge, i.e. the distribution of $X$ conditional on $Z$, nontrivial tests of conditional independence can be designed without further assumptions on the distribution of $(X, Y, Z)$. This has been dubbed the Model-X (MX) setting, and the condition that the distribution of $X$ conditional on $Z$ is known is called the Model-X (MX) assumption, though there are settings where the MX 'assumption' is *known* to be true. Perhaps the most prominent example is a randomized clinical trial, where the distribution of treatment/control is imposed by the researchers. Another example is conjoint analysis, which is a survey-based experiment where respondents are asked to express a preference between multiple hypothetical products with different attributes. These attributes are randomized according to a distribution chosen by the researchers, with the aim to find out whether one or more of the attributes have an influence on consumer preference (see e.g. Ham et al. (2024)). Furthermore, Candès et al. (2018) show that there are ample scenarios where at least an accurate estimate of the conditional distribution of $X$ given $Z$ is known, while also acknowledging that the MX assumption might not be appropriate if this is not the case.

Under the MX assumption, Candès et al. derive the conditional randomization test (CRT), which has nontrivial power to detect CI. Recently, much effort has gone

into relaxing the MX assumption and improving the robustness of the CRT under misspecification of the conditional distribution of $X \mid Z$ (Katsevich and Ramdas, 2022; Li and Liu, 2023; Niu et al., 2024). The CRT and most of its extensions have in common that they are based on p-values computed on batches of data, and therefore designed for fixed sample size experiments. In this chapter, we focus on anytime-valid tests for CI under MX. Anytime-valid tests allow for more flexibility when testing compared to the nonsequential CRT and also allow covariate adaptive designs, where the distribution of $X$ does not only depend on $Z$, but also on past data, such as in response-adaptive sampling schemes. Previous work on anytime-valid tests of CI under MX has been done by Duan et al. (2022), who propose tests for the case that $X$ is binary. Their approach is based on pragmatic game-theoretic principles, whereas in this chapter, we aim to describe and theoretically analyze anytime-valid tests for CI for general $X$. Shaer et al. (2023) have worked on an extension of the work by Duan et al. (2022) concurrently, and we discuss their work and the connections to this chapter in Section 6.6.

Our hypothesis tests are based on the concept of $e$-statistics[1] (Grünwald et al., 2024; Ramdas et al., 2022; Vovk and Wang, 2021). $E$-statistics have been introduced as an alternative to p-values that is inherently more suitable for testing under optional stopping and continuation (Grünwald et al., 2024; Vovk and Wang, 2021; Ramdas et al., 2020). While they have their roots in the work on anytime-valid testing by H. Robbins and students (e.g. (Darling and Robbins, 1967)), interest in them has exploded in recent years; see, for example, also Shafer (2021); Grünwald (2024). $E$-statistics can be associated with a natural notion of optimality, called GRO (growth-rate optimality) by Grünwald et al. (2024), which may be viewed as an analogue of statistical power in an optional stopping context. We derive a general method for constructing $e$-statistics for conditional independence testing under model-X and show that the method that we propose is optimal in this GRO sense for testing conditional independence against point alternatives under MX. This result should be seen as an anytime-valid analogue to the result by Katsevich and Ramdas (2022), who use the Neyman-Pearson lemma to derive test statistics for which the conditional randomization test is the most powerful conditionally valid MX CI test. Furthermore, we show that under misspecification of the distribution of $X$ given $Z$, our method retains type-I error sequentially just as well as the CRT does for blocks of data (Berrett et al., 2020). Finally, we discuss in detail an application to the setting where $Y$ is binary, where we use logistic regression to construct an anytime-valid test of conditional independence.

[1] E-statistics are commonly known as $e$-variables; we use the former to stress data dependence.

Under the MX assumption, this test is valid even if the logistic model assumptions are not satisfied, and when they are, it is guaranteed to have asymptotic power one.

The rest of this chapter is structured as follows. In Section 6.2, we discuss the necessary theoretical background. This includes an introduction to the conditional randomization test in Section 6.2.1 and to $e$-statistics in Section 6.2.2. In Section 6.3 we discuss in detail our anytime-valid method of testing conditional independence under MX. This includes an analysis of optimality in Section 6.3.1, a bound on the worst-case rejection rate in Section 6.3.2, and an application to binary response variable in Section 6.3.3. We compare our method to established tests of conditional independence in a simulation study in Section 6.4. This includes a comparison in terms of type-I error (Sections 6.4.1 and Appendix D.1) and in terms of power (Section 6.4.2). Our method is further highlighted by an application to a real-world data set in Section 6.5. This chapter is concluded with a discussion of our results in Section 6.6. All proofs are deferred to Appendix D.2.

## 6.2 Background

In this section, we give a brief introduction to the conditional randomization test, as well as to $e$-statistics. Throughout this section, as well as the rest of the chapter, we assume that the data consists of independent and identically distributed (i.i.d.) tuples $D_i = (X_i, Y_i, Z_i) \in \mathcal{D} := \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. In the section on the CRT, we assume that data comes in as a single block $D^n = (D_1, \dots, D_n)$ of $n \in \mathbb{N}$ data points, so that $i$ above ranges from 1 to $n$. In contrast, in the section on $e$-statistics, we assume that the data comes in sequentially as a stream $(D_n)_{n\in\mathbb{N}}$. We use the notation $f_{Y|X,Z}(y \mid x, z)$ to denote the conditional density of $Y$ given $X$ and $Z$ evaluated at $(x, y, z)$ when the joint density of $(X, Y, Z)$ is $f$, and analogous notation for other conditional densities, e.g. $f_{X|Z}(x \mid z)$ for the density of $X$ given $Z$.

### 6.2.1 The Conditional Randomization Test

The conditional randomization test by Candès et al. (2018) works under the MX assumption, i.e. that the distribution of $X \mid Z$ is known. This holds, for example, when $X$ corresponds to a randomized treatment in a clinical trial, since the researchers specify the randomization mechanism themselves. Candès et al. (2018) give other examples where the this distribution is at least known to a much higher precision than the joint distribution of $(X, Y, Z)$ because much more samples of pairs $(X, Z)$ are

available than data including $Y$. We let $Q_z$ denote the distribution of $X$ given that $Z = z$. We are interested in testing $X \perp\!\!\!\perp Y \mid Z$, which in the MX setting corresponds to the null hypothesis

$$\mathcal{H}_0 = \left\{ P \in \mathcal{P}(\mathcal{D}) : X \overset{P}{\perp\!\!\!\perp} Y \mid Z,\ P_{X|Z} = Q_Z \right\}, \tag{6.1}$$

where $\mathcal{P}(\mathcal{D})$ denotes the set of all probability distributions on $\mathcal{D}$, and $P_{X|Z}$ is a shorthand for the conditional law of $X$ given $Z$. The starting point of the CRT is to choose any test statistic $T : \mathcal{D}^n \mapsto \mathbb{R}$ that measures the dependence of $X$ and $Y$ given $Z$, so that larger values of $T$ indicate stronger dependence. Since the conditional distribution of $X$ given $Z$ is known, one can simulate independent new realizations $\tilde{X}_{j,i} \sim Q_{Z_i}$, $j = 1, \ldots, M$, $i = 1, \ldots, n$, so that under the null hypothesis $\tilde{X}_j^n = (X_{j,1}, \ldots, X_{j,n})$ and $X^n$ have the same distribution conditional on $Y^n$ and $Z^n$. Hence the triplets $(\tilde{X}_j^n, Y^n, Z^n)$, $j = 1, \ldots, M$, and $(X^n, Y^n, Z^n)$ are exchangeable, and

$$p_M(X^n, Y^n, Z^n) = \frac{1 + \sum_{j=1}^M 1\{T(\tilde{X}_j^n, Y^n, Z^n) \geq T(X^n, Y^n, Z^n)\}}{1 + M} \tag{6.2}$$

is a "Monte Carlo" p-value for the null hypothesis (6.1). More precisely, for any distribution $P \in \mathcal{H}_0$,

$$P(p_M(X^n, Y^n, Z^n) \leq \alpha) \leq \alpha.$$

In the case that the distributions $Q_z$ are not known exactly, but only some estimate $\hat{Q}_z$, an obvious question is how robust a test based on $\hat{Q}_z$ may be. Berrett et al. (2020, Theorem 4) show that

$$P(p_M(X^n, Y^n, Z^n) \leq \alpha \mid Y^n, Z^n) \leq \alpha + d_{\mathrm{TV}}(\hat{Q}_{Z^n}^n, Q_{Z^n}^n), \tag{6.3}$$

where $Q_{Z^n}^n$ is the product of the distributions $Q_{Z_i}$, $i = 1, \ldots, n$, and $d_{\mathrm{TV}}$ the total variation distance. Precise estimation of conditional distributions in total variation distance is admittedly a challenging problem, but also not completely unrealistic, as discussed among others by Berrett et al. (2020, Section 5.1). We furthermore give a brief discussion of available literature on estimation in terms of KL divergence at the end of Section 6.3.1, and a bound on KL gives a bound on total variation.

There remains the question how to choose the statistic $T$. Katsevich and Ramdas (2022) show that for a point alternative with density $f$ for $(X, Y, Z)$, using the conditional density $f_{Y|X,Z}$ of $Y$ given $(X, Z)$ as statistic $T$ leads to the most powerful

conditionally valid test against $\mathcal{H}_0$. This result suggests that a reasonable method is to use an estimator $\hat{f}_{Y|X,Z}$ of the true conditional density as statistic. Importantly, any estimation error does not lead to a loss of type-I error guarantee, but only to a decrease in power, as the p-value in (6.2) is valid for *any* statistic. Alternatively, one could use any measure of conditional independence, e.g. the absolute value of the fitted coefficient in a lasso regression model, as originally proposed by Candès et al. (2018). A downside to this method is that it is computationally expensive, as it requires a lasso model to be fit for the original data as well as for all the simulated data points. Liu et al. (2021) propose a "leave-one-covariate-out" variant of these lasso statistics, which is significantly less computationally demanding, while leading to similar power.

### 6.2.2 E-Statistics, Test Martingales and Anytime-Valid Tests

An $e$-statistic is any function of the data $S_n : \mathcal{D}^n \to [0, \infty)$ such that $\mathbb{E}_P[S_n(D^n)] \leq 1$ for all $P \in \mathcal{H}_0$. An $e$-statistic evaluated on a realization of the data will be referred to as an $e$-value. Previously, $e$-statistics and $e$-values have appeared in the literature as likelihood ratios (although the concept is vastly more general than such ratios), particular Bayes factors and betting scores (Shafer, 2021). Large $e$-values constitute evidence against the null hypothesis, since $P(S_n(D^n) \geq 1/\alpha) \leq \alpha$ by Markov's inequality, so that the type-I error of the test $1\{S_n(D^n) \geq 1/\alpha\}$ is bounded by $\alpha$. However, such a test is defined for a block of data $D^n$. In a sequential setting, one instead observes a stream of data $(D_n)_{n\in\mathbb{N}}$. We therefore define, more generally, a sequence of conditional $e$-statistics as a sequence of statistics $(E_n(D^n))_{n\in\mathbb{N}}$, such that $\mathbb{E}_P[E_n(D^n) \mid D^{n-1}] \leq 1$ for all $n \in \mathbb{N}$ and $P \in \mathcal{H}_0$. For $n = 1$, the expectation is supposed to be read unconditionally. Intuitively, the conditional $e$-statistic at time $n$ measures the evidence in round $n$ against $\mathcal{H}_0$ conditional on the past data, and the cumulative product of these conditional $e$-statistics $S_n(D^n) = \prod_{i=1}^n E_i(D^i)$ is a measure of the total accumulated evidence against the null hypothesis. Formally, the sequence $(S_n(D^n))_{n\in\mathbb{N}}$ of cumulative products forms a nonnegative supermartingale with starting value bounded by 1, a so-called test martingale, i.e. $\mathbb{E}_P[S_{n+1}(D^{n+1}) \mid D^n] \leq S_n(D^n)$ for all $P \in \mathcal{H}_0$ and $n \in \mathbb{N}$, and $\mathbb{E}_P[S_1(D_1)] \leq 1$. By Ville's inequality (see e.g. Shafer, 2021), such test martingales satisfy for any $\alpha > 0$

$$P(\exists\, n \in \mathbb{N}\colon S_n(D^n) \geq 1/\alpha) \leq \alpha. \tag{6.4}$$

A sequential test can thus be defined by monitoring $S_n(D^n)$ and rejecting if it exceeds $1/\alpha$. The inequality (6.4) ensures that this test retains type-I error control, no matter

how we choose the moments to peek at $S_n(D^n)$. In fact, (6.4) is easily derived from a more basic property: $(S_n(D^n))_{n\in\mathbb{N}}$ satisfies $\mathbb{E}_P[S_\tau(D^\tau)] \leq 1$ for any stopping time $\tau$, that is, the stopped process $S_\tau(D^\tau)$ is again an $e$-statistic, both for data dependent and externally imposed stopping rules (Grünwald et al., 2024, Proposition 2). Tests with this property will be referred to as anytime-valid tests.

Perhaps the most prominent example of a test martingale is the likelihood ratio process $L_n = \prod_{i=1}^n p_1(Y_i)/p_0(Y_i)$ for testing the null hypothesis that independent $(Y_n)_{n\in\mathbb{N}}$ stem from a distribution with density $p_0$ against the alternative density $p_1$. Tests based on likelihood ratio processes $(L_n)_{n\in\mathbb{N}}$ are called sequential probability ratio test (SPRT) and have first been studied by Wald (1947), though Wald, like most of the subsequent sequential analysis literature, uses them with reject/accept rules different from ours (e.g. one rejects if the value exceeds $(1-\beta)/\alpha$ instead of $1/\alpha$) that preclude optional stopping and require specific stopping times/rules. We refer to Grünwald et al. (2024, Section 7) for further discussion of related literature on sequential testing.

Under violations of the null hypothesis, one would hope that an $e$-statistic or test martingale attains high values, which gives evidence to reject the null hypothesis. This requires a suitable analogue of power, or notion of optimality, for $e$-statistics. We follow Grünwald et al. (2024); Shafer (2021) and try to find $e$-statistics that maximize the logarithmic expected value under the alternative. That is, suppose the alternative hypothesis is given by a single distribution $\mathcal{H}_1 = \{P^*\}$, then the growth-rate optimal (GRO) $e$-statistic is defined as the $e$-statistic that maximizes $S_n \mapsto \mathbb{E}_{P^*}[\log S_n(D^n)]$ over all $e$-statistics. At first glance, this notion of optimality seems counterintuitive for sequential tests, since it is defined for individual $e$-statistics and not test martingales. Really, one would hope to find a test martingale $(S_n)_{n\in\mathbb{N}}$ that maximizes $\mathbb{E}_{P^*}[\log S_\tau(D^\tau)]$ simultaneously for all stopping times $\tau$. Remarkably, in the special setting of this chapter, the sequence of GRO statistics that we will consider for a point alternative does have exactly this property, as follows from Theorem 12 of Koolen and Grünwald (2022), providing additional justification for our focus on GRO. This is discussed briefly in Appendix D.3.

## 6.3 Conditional Independence Testing With E-Statistics

The conditional randomization test and its permutation version by Berrett et al. (2020) are defined for batches of data $D^n$. We show here how to create $e$-statistics for sequential testing with a stream of data $(D_n)_{n\in\mathbb{N}}$ under the MX assumption. Our method can be seen as a broad generalization of an $e$-statistic introduced in the proof of the main theorem of Turner et al. (2024), who essentially handle the case that $Y$ and $X$ are both Bernoulli.

**Theorem 6.1.** *Let $h_n : \mathcal{D} \to [0,\infty)$, $n \in \mathbb{N}$, be nonnegative measurable functions such that $h_n$ is determined after seeing data $D^{n-1}$. Then the sequence $(E^{CI}_{h_n}(D_n))_{n\in\mathbb{N}}$ defined by*

$$E^{CI}_{h_n}(D_n) = \frac{h_n(X_n, Y_n, Z_n)}{\int_{\mathcal{X}} h_n(x, Y_n, Z_n)\,\mathrm{d}Q_{Z_n}(x)}. \tag{6.5}$$

*is a sequence of conditional e-statistics for the null hypothesis* (6.1)*. Consequently, the sequence $(S^{CI}_{h^n}(D^n))_{n\in\mathbb{N}}$ defined by $S^{CI}_{h^n}(D^n) = \prod_{i=1}^n E^{CI}_{h_i}(D_i)$ is a test martingale.*

To be explicit, the general workflow of our method is as follows. At each time $n = 1, 2, \ldots$, a test function $h_n$ is chosen, usually depending on the past data $D^{n-1}$. After seeing the data point $D_n$, the conditional $e$-statistic (6.5) is computed and the cumulative product is updated according to $S^{\mathrm{CI}}_{h^n}(D^n) = S^{\mathrm{CI}}_{h^{n-1}}(D^{n-1}) \cdot E^{\mathrm{CI}}_{h_n}(D_n)$. For a test at level $\alpha$, one could stop and reject the null hypothesis as soon as $S^{\mathrm{CI}}_{h^n}(D^n) \geq 1/\alpha$ for the first time.

**Remark.** At this point, it should be noted that the MX assumption is stronger than needed within our sequential setup: the MX assumption requires that the data points $(X_n, Y_n, Z_n)$ are i.i.d. and that the distribution of $X_n$ given $Z_n$ is given by $Q_{Z_n}$. However, in our sequential setting it would actually be allowed for the distribution of $(X_n, Y_n, Z_n)$ to depend on the data $D^{n-1}$. At each time $n$, we would use in the denominator a distribution $Q_{Z_n, D^{n-1}}$. It should be clear that the resulting sequence of random variables still defines a test martingale, but with respect to the altered null hypothesis that $Y_n \perp\!\!\!\perp X_n \mid Z_n, D^{n-1}$ and $X_n \mid Z_n \sim Q_{Z_n, D^{n-1}}$ for all $n \in \mathbb{N}$. Clearly, the assumption that $Q_{Z_n, D^{n-1}}$ is known (or can be estimated with high precision) is not realistic in many settings with temporal dependence. One example where this extension would be useful are clinical trials with so-called covariate- or response-adaptive designs, where the allocation of patients to a treatment depends either on the

imbalance of treatment/control in certain covariate groups or on previously observed responses from patients (Robbins, 1952; Pocock and Simon, 1975; Zhang et al., 2007). To avoid cluttering notation, we will not explicitly denote this potential past data dependence, but it is good to keep in mind.

In most cases in practice, one has to resort to simulations to approximate the integral in the denominator of (6.5). That is, one simulates $M$ independent $\tilde{X}_1, \ldots, \tilde{X}_M$ according to $Q_{Z_n}$ and replaces the integral by the empirical mean $\sum_{i=1}^{M} h_n(\tilde{X}_i, Y_n, Z_n)/M$. The following proposition shows that a slight modification of this procedure is guaranteed to yield an $e$-statistic.

**Proposition 6.2.** *Let $\tilde{X}_1, \ldots, \tilde{X}_M$ be independent with distribution $Q_{Z_n}$. Then*

$$\check{E}^{CI}_{h_n}(D_n) := \frac{h_n(X_n, Y_n, Z_n)}{(h_n(X_n, Y_n, Z_n) + \sum_{i=1}^{M} h_n(\tilde{X}_i, Y_n, Z_n))/(M+1)}$$

*satisfies $\mathbb{E}_P[\check{E}^{CI}_{h_n}(D_n) \mid D^{n-1}] = 1$, $P \in \mathcal{H}_0$, where $h_n$ is as in Theorem 6.1 and the expectation is taken both over the data and $\tilde{X}_1, \ldots, \tilde{X}_M$.*

In fact, the proof of Proposition 6.2 only requires that $X_n, \tilde{X}_1, \ldots, \tilde{X}_M$ are exchangeable, so it will also be applicable in many situations where data is randomly permuted. Furthermore, the naive approach without including $h_n(X_n, Y_n, Z_n)$ in the denominator is anti-conservative and does not define a sequence of conditional $e$-statistics. Indeed, taking expectation over $X_n, \tilde{X}_1, \ldots, \tilde{X}_M$,

$$\mathbb{E}_P\left[\frac{h_n(X_n, Y_n, Z_n)}{\sum_{i=1}^{M} h_n(\tilde{X}_i, Y_n, Z_n)/M}\middle| Y_n, Z_n\right] \geq \frac{\int h_n(x, Y_n, Z_n)\, dQ_{Z_n}(x)}{\sum_{i=1}^{M} \int h_n(x, Y_n, Z_n)\, dQ_{Z_n}(x)/M} = 1, \tag{6.6}$$

Here we use that $X_n, \tilde{X}_1, \ldots, \tilde{X}_M$ are independent with distribution $Q_{Z_n}$, and invoke Jensen's inequality with the strictly convex function $s \mapsto 1/s$. Equality in (6.6) holds if any only if $h_n$ is constant in $x$, and in that trivial case the $e$-statistic is equal to the constant 1. In all our applications, we use the variant proposed in Proposition 6.2 to compute the $e$-statistics.

### 6.3.1 Optimality

In order to accumulate evidence against the null hypothesis, the functions $h_i$ in (6.5) should measure the conditional (in)dependence between $X$ and $Y$. This will ensure that the test martingale $(S^{\mathrm{CI}}_{h^n}(D^n))_{n\in\mathbb{N}}$ will grow if the null hypothesis is violated.

Ideally, we would be able to choose a measure of conditional independence that requires little to no assumptions on the distribution of $(X, Y, Z)$. Many such measures have been proposed in the literature, see e.g. Fukumizu et al. (2007); Shah and Peters (2020); Azadkia and Chatterjee (2021). However, as far as we can tell, none of these measures allow for a sequential decomposition. That is, they are defined for fixed sample size $n$ as a function $T_n : \mathcal{D}^n \to [0, \infty)$, but cannot be decomposed in a nontrivial way as a product $T_n(D^n) = \prod_{i=1}^n T_i(D_i)$. Our only option to use them in our test martingale is therefore to set $h_i(D_i) = T_i(D^{i-1}, D_i)$ in (6.5). This would have two major drawbacks: first, it would be computationally involved, because $T_i$ needs to be recalculated entirely *within* the integral in the denominators in (6.5). Secondly, it would generally be ineffective, because $T_i(D^{i-1}, D_i)$ will depend very little on $D_i$ for large $i$, so $h_i(D_i)$ will generally not be sensitive to changing $X_i$ to $x' \sim Q_{Z_i}$. As a result, the fraction in (6.5) will be close to 1, preventing us from accumulating much evidence against the null.

However, if we are willing to assume that under the alternative, the density of $(X, Y, Z)$ is given by $f$, then the conditional density $f_{Y|X,Z}$ itself is a measure of conditional independence. Moreover, evaluating the density in $n$ data points is equivalent to taking the product of the density evaluated at all single data points $i = 1, \dots, n$, because the data stream is i.i.d. This gives a sequential decomposition as desired above. Furthermore, Katsevich and Ramdas (2022) have shown that an optimal conditionally valid p-value based test can be achieved by running the CRT with the conditional density $f_{Y|X,Z}$ as test function. It turns out that this choice also yields the GRO $e$-statistic among all $e$-statistics defined on $n$ observations, and the expectation of the logarithm of this e-statistic is the conditional mutual information (Cover and Thomas, 1991), an established conditional dependence measure which has also been applied for conditional independence testing (Runge, 2018). Note that the expectation is taken over the entirety of the data, i.e. $D^n = (X^n, Y^n, Z^n)$, as opposed to the result by Katsevich and Ramdas (2022) which only holds conditionally on $(Y^n, Z^n)$; see their article for a more thorough discussion.

**Theorem 6.3.** *The GRO e-statistic for testing $\mathcal{H}_0$ as in* (6.1) *against the alternative distribution with density $f$ is given by*

$$S^{CI}_{f_{Y|X,Z}}(D^n) = \prod_{i=1}^n E^{CI}_{f_{Y|X,Z}}(D_i) = \prod_{i=1}^n \frac{f_{Y|X,Z}(Y_i \mid X_i, Z_i)}{f_{Y|Z}(Y_i \mid Z_i)} \tag{6.7}$$

*and achieves growth rate $\mathbb{E}_f[\log S^{CI}_{f_{Y|X,Z}}(D^n)] = nI_f(X; Y \mid Z)$, where $I_f(X; Y \mid Z)$*

*denotes the conditional mutual information if* $(X, Y, Z)$ *follows the distribution with density* $f$*.*

**Remark.** A simple application of Bayes theorem allows one to rewrite (6.7) to $S^{\mathrm{CI}}_{f_{Y|X,Z}}(D^n) = \prod_{i=1}^n f_{X|Y,Z}(X_i \mid Y_i, Z_i)/f_{X|Z}(X_i \mid Z_i)$, which shows that the resulting test martingale is in fact a likelihood ratio process. That is, it is the ratio between the true density of $X$ given $(Y, Z)$ under the alternative and that under the null (whereas the density in the denominator of (6.7) does not have to correspond to the data generating distribution). The latter follows, because under the null $f_{X|Y,Z}$ is equal to $f_{X|Z}$, which we assume to be well-specified and equal the density of $Q_{Z_i}$. Hence the resulting test for a simple alternative hypothesis $f$ is in fact a generalization of the SPRT, where the distribution under the null hypothesis changes with the variable $Z_i$. It depends on the application at hand which formulation, i.e. (6.7) or conditional densities of $X$, is more suitable for constructing a test. For example, we show in Section 6.3.3 that for binary $Y \in \{0, 1\}$ and $(X, Z) \in \mathbb{R}^p \times \mathbb{R}^q$, one can construct a test based on logistic regression, which is often simpler than directly trying to find a suitable conditional density of $X$ given $Y$ and $Z$, especially when $p > 1$.

The information inequality (Cover and Thomas, 1991) implies that $I_f(X; Y \mid Z) \geq 0$, with equality if and only if $Y \perp\!\!\!\perp X \mid Z$, which shows that $S^{\mathrm{CI}}_{f_{Y|X,Z}}$ has nontrivial power to detect deviations from conditional independence if $f$ is the true density of $(X, Y, Z)$. Assuming a simple (point) alternative $f$ is, of course, an unrealistically strong assumption. We now proceed to develop a method that also gets uniform growth rates for potentially large classes of alternative densities $\mathcal{F}$ by building on the simple alternative case. The following result states that if we do not know the density $f_{Y|X,Z}$, but instead use a different density $g_{Y|X,Z}$, the loss in expected growth rate is directly related to a measure of distance between $f_{Y|X,Z}$ and $g_{Y|X,Z}$.

**Proposition 6.4.** *For any conditional density* $g_{Y|X,Z}$*, the following holds:*

$$\mathbb{E}_f\left[\log E^{CI}_{g_{Y|X,Z}}(D)\right] \geq I_f(X; Y \mid Z) - \mathbb{E}_f[\mathrm{KL}(f_{Y|X,Z} \| g_{Y|X,Z})]. \tag{6.8}$$

Since we are in a sequential setting, this proposition implies that if we do not know the density $f$, we could try to estimate it, using estimates that improve as sample size increases. That is, let $\hat{f}_n$ be an estimator of $f_{Y|X,Z}$ based on data $D^n$, and $\hat{f}_0$ an initial guess. The test martingale we use is then given by $\prod_{i=1}^n E^{\mathrm{CI}}_{\hat{f}_{n-1}}(D^i)$. It follows from the combination of Theorem 6.3 and Proposition 6.4 that if the estimator is consistent in a KL sense, then the expected growth per outcome converges to that of the GRO

$e$-statistic.

**Corollary 6.5.** *(i) Assume that* $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_f[\mathrm{KL}(f_{Y|X,Z}\|\hat{f}_{i-1}) \mid D^{i-1}] \xrightarrow[n\to\infty]{a.s.} 0$*, then*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_f\left[\log E^{CI}_{\hat{f}_{i-1}}(D_i)\middle|D^{i-1}\right] \xrightarrow[n\to\infty]{a.s.} I_f(Y;X \mid Z).$$

*(ii) Assume that for some function* $b(n):\mathbb{N}\to\mathbb{R}_0^+$ *with* $b(n)=o(n)$*, we have*

$$\mathbb{E}_f\left[\sum_{i=1}^{n}\mathbb{E}_f[\mathrm{KL}(f_{Y|X,Z}\|\hat{f}_{i-1}) \mid D^{i-1}]\right] \le b(n). \tag{6.9}$$

*Then*

$$\frac{1}{n}\mathbb{E}_f\left[\sum_{i=1}^{n}\log E^{CI}_{\hat{f}_{i-1}}(D_i)\right] \ge I_f(X;Y \mid Z) - \frac{b(n)}{n}. \tag{6.10}$$

Consequently, to achieve an asymptotic optimal growth rate, we need to use a conditional density estimator $\hat{f}_i$ that converges in KL divergence to $f$, where we may assume $f\in\mathcal{F}$ for some given set of densities $\mathcal{F}$, i.e. our statistical model. Barron (1998) showed that, for a wide variety of parametric and nonparametric models $\mathcal{F}$, we have *convergence in information* (his terminology for (6.9) holding for all $n$) if we set $\hat{f}_i$ to be the Bayes predictive density, under no further conditions, uniformly for all $f\in\mathcal{F}$, as long as a suitable prior is used. This means that, for some fixed function $b$ (6.9) holds uniformly for all $f\in\mathcal{F}$, so that also (6.10) holds uniformly for all $f\in\mathcal{F}$: we could put a $\inf_{f\in\mathcal{F}}$ to the left of (6.10) and the result would still hold. Thus, we get the optimal growth rate (which itself can only be achieved by an oracle that knows the 'true' $f\in\mathcal{F}$) up to an additive term of $b(n)/n$ uniformly for all $f\in\mathcal{F}$. The rate $b(n)/n$ is then usually, up to log factors, equal to the minimax rate in squared Hellinger distance. The same rates (potentially up to further log factors) are available for the Bayesian posterior mean under a weak additional condition on the model introduced by Grünwald and Mehta (2020) under the name *witness-of-badness*; it generalizes a well-known earlier condition of Wong and Shen (1995); see also (Bilodeau et al., 2023) for related results. In Section 6.3.3, we demonstrate our approach with $\mathcal{F}$ set to the logistic regression model with $X\in\mathbb{R}^p$ and $Z\in\mathbb{R}^q$, for which a result by Foster et al. (2018) in combination with (Barron, 1998) implies that, if we use the Bayes predictive distribution as above, then $b(n)$ can be chosen as $O((p+q)\log n)$ as long as the first four moments of all components of $X$ and $Z$ exist, implying a parametric rate. However, in our experiments in Section 6.4, rather than the Bayes predictive distribution, we use a

(slightly regularized) MLE, since it can be computed much more efficiently. Although we suspect that the MLE converges at the same rates as Bayesian methods under the same weak conditions, the methods of the aforementioned papers cannot be used to prove this, and instead in Proposition 6.7 we show almost sure convergence of the MLE, without rates, under a stronger subgaussianity assumption on $(X, Z)$.

### 6.3.2 Worst-Case Bounds on Rejection Rate

Up to now, we have discussed the construction and properties of $e$-statistics when the conditional distributions $Q_z$ are known exactly. In this section, we prove a result analogous to Theorem 4 of Berrett et al. (2020) (see (6.3)) on the error rate of our sequential test under the null hypothesis when the distributions $Q_z$ are only approximations. The approximation of $Q_z$ will be denoted by $\hat{Q}_z$, and the (approximate) $e$-statistic at time $n$ is given by

$$\tilde{E}^{\mathrm{CI}}_{h_n}(D^n) = \frac{h_n(X_n, Y_n, Z_n \mid D^{n-1})}{\int_{\mathcal{X}} h_n(x, Y_n, Z_n \mid D^{n-1}) \, \mathrm{d}\hat{Q}_{Z_n}(x)}.$$

Here the nonnegative function $h_n$ depends on $D^{n-1}$ since it can be constructed sequentially, for example by estimating the conditional density of $Y_n$ given $X_n$ and $Z_n$ with all past data $(X_i, Y_i, Z_i)$, $i = 1, \ldots, n-1$. Recall that $Q^n_{Z^n}$ denotes the product distribution of $Q_{Z_i}$, $i = 1, \ldots, n$, that is, for measurable $A \subseteq \mathcal{X}^n$

$$Q^n_{Z^n}(A) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} 1\{x^n \in A\} \, \mathrm{d}Q_{Z_1}(x_1) \cdots \mathrm{d}Q_{Z_n}(x_n),$$

In particular, $Q^n_{Z^n}(A) = P(X^n \in A \mid Z^n) = P(X^n \in A \mid Y^n, Z^n)$ for $P \in \mathcal{H}_0$, due to conditional independence of $Y_i$ and $X_i$ given $Z_i$, $i = 1, \ldots, n$. The distribution $\hat{Q}^n_{Z^n}$ is defined in the same way as $Q^n_{Z^n}$ but with $\hat{Q}_{Z_i}$ instead of $Q_{Z_i}$.

**Theorem 6.6.** *Assume that $h_1, \ldots, h_N > 0$ are measurable. For any $N \in \mathbb{N}$, $\alpha \in (0, 1)$ and $P \in \mathcal{H}_0$,*

$$P\left(\exists\, n \le N \colon \prod_{i=1}^{n} \tilde{E}^{CI}_{h_i}(D^i) \ge \frac{1}{\alpha} \middle| Y^N, Z^N\right) \le \alpha + d_{\mathrm{TV}}(Q^N_{Z^N}, \hat{Q}^N_{Z^N}). \tag{6.11}$$

Theorem 6.6 gives the same worst case bound on the rejection rate as Theorem 4 in Berrett et al. (2020) for a sample size $N$, but optional stopping at any sample size $n \le N$ is allowed. Berrett et al. (2020, Section 5.1) discuss conditions under which

the total variation distance between $Q_{Z^N}^N$ and $\hat{Q}_{Z^N}^N$, which bounds the excess rejection rate, is small. For example, they obtain an upper bound of the form $\mathcal{O}_p(\sqrt{Nk/m})$ when $(X, Z) \in \mathbb{R}^k$ follow a multivariate Gaussian distribution and the conditional law of $X \mid Z$ is estimated with an unlabeled sample of size $m$. Hence when $m$ remains constant but $N$ diverges to infinity, the bound on the distance between $Q_{Z^N}^N$ and $\hat{Q}_{Z^N}^N$ becomes trivial. Furthermore, the discussion at the end of Section 6.3.1 on estimation in terms of KL divergence (which bounds the total variation distance) can also be applied to the estimation of $Q_z$.

### 6.3.3 Application to Logistic Regression

The general construction strategy for $e$-statistics and all results so far assume no specific model for the outcome $Y$ or covariates $(X, Z)$. In this section, we consider the special but important case of a binary outcome $Y \in \{0, 1\}$ which follows a logistic regression model under the alternative, i.e. $(X, Z) \in \mathbb{R}^{p+q}$, and $Y$ equals $y = 0, 1$ with probability

$$p_\theta(y \mid X, Z) = \frac{\exp(y(\beta^\top X + \gamma^\top Z))}{1 + \exp(\beta^\top X + \gamma^\top Z)}, \tag{6.12}$$

with an unknown $(p + q)$-dimensional coefficient vector $\theta = (\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_q)$. Conditional independence of $Y$ and $X$ given $Z$ holds if and only if $\beta_1 = \cdots = \beta_p = 0$. It turns out that in this setting, one can construct $e$-statistics which not only have power on average, as in Corollary 6.5, but which even reject the null hypothesis with probability one if it is violated and the sample size grows to infinity. At this point, it is important to recall that the validity of the $e$-statistics does not require the logistic model to be correctly specified: the validity only depends on the null hypothesis, which is still the set of *all* distributions under which conditional independence holds in the sense of (6.1), including many distributions that violate the logistic model assumption. The result rather shows that *if* the logistic model is suitable (in the sense that, if the alternative is true, then data are sampled from a distribution in this model), then the $e$-statistic has guaranteed power to detect violations of conditional independence.

Following our general strategy, an $e$-statistic for testing CI is given by

$$S_n^{CI}(D^n) = \prod_{i=1}^n \frac{p_{\hat{\theta}_{i-1}}(Y_i \mid X_i, Z_i)}{\int p_{\hat{\theta}_{i-1}}(Y_i \mid x, Z_i) \, \mathrm{d}Q_{Z_i}(x)}, \tag{6.13}$$

where $\hat{\theta}_k$ may be any estimator for $\theta$ based on the first $k$ samples, $(X_i, Y_i, Z_i)$, $i = 1, \ldots, k$. When the observations $(X_i, Y_i, Z_i)$, $i \in \mathbb{N}$, are independent and identically

distributed, the growth rate optimal $e$-statistic is obtained if $\hat{\theta}_k = \theta$ for all $k \in \mathbb{N}$. Nevertheless, the following proposition shows that tests based on $S_n^{CI}$ have asymptotic power one when $\hat{\theta}_k$ is the maximum likelihood estimator. From now on, $\|v\| = (v^\top v)^{1/2}$ denotes the Euclidean norm of a vector $v$, and we denote by $(X, Z)$ the stacked vector $(X_1, \ldots, X_p, Z_1, \ldots, Z_q)$. We will take $(X, Y, Z)$ as a generic observation which has the same distribution as $(X_i, Y_i, Z_i)$, $i \in \mathbb{N}$, for writing probability statements about elements of this sequence.

**Proposition 6.7.** *Let $(X_i, Y_i, Z_i)$, $i \in \mathbb{N}$, be independent and identically distributed such that* (6.12) *holds with $(\beta_1, \ldots, \beta_p) \neq 0$. Assume furthermore that*

*(i) (a) $(X, Z)$ satisfies $P(u^\top (X, Z) \neq 0) > 0$ for all $u \in \mathbb{R}^{p+q} \setminus \{0\}$, and (b) it is subgaussian with variance parameter $\sigma^2$, that is*

$$\mathbb{E}[\exp(u^\top((X, Z) - \mathbb{E}[(X, Z)]))] \leq \exp(\|u\|^2\sigma^2/2), \ \forall\, u \in \mathbb{R}^{p+q},$$

*(ii) $\hat{\theta}_n$ in* (6.13) *is the logistic MLE based on data $(X_i, Y_i, Z_i)$, $i = 1, \ldots, n$, for all $n \in \mathbb{N}$, with $\hat{\theta}_n$ arbitrarily defined but finite if the MLE does not exist.*

*Then $S_n^{CI}$ satisfies $\liminf_{n\to\infty} \log(S_n^{CI})/n \geq I(X; Y|Z) > 0$ almost surely.*

Assumption (i)(a) ensures that the MLE converges almost surely at a fast rate, as shown by Qian and Field (2002). Instead of subgaussianity ((i)(b)) their result requires only moment assumptions (which are implied by subgaussianity), but subgaussianity is indeed required in our setting: see the proof of Proposition 6.7, given in Appendix D.2.6.

## 6.4 Simulations

To investigate the robustness and power of tests of conditional independence based on $e$-statistics, we compare the $e$-statistic (6.13) for binary $Y$ to other methods applicable in this setting. The covariate $X$ is univariate, $X \in \mathbb{R}$, while $Z = (1, Z_1, \ldots, Z_{q-1})$ is a $q$-dimensional vector containing an intercept term. The distribution of the vector $(X, Z_1, \ldots, Z_{q-1})$ is the $q$-dimensional normal distribution with zero mean and a Toeplitz covariance matrix, $\Sigma_{i,j} = 1/(1 + |i - j|)$ for $i, j = 1, \ldots, q$. Then $Q_Z$ is the Gaussian distribution with mean

$$\mu_Z = \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}(Z_1, \ldots, Z_{q-1})^\top \tag{6.14}$$

where $\Sigma_{-1,-1}$ is the submatrix $(\Sigma_{i,j})_{i,j=2,\dots,q}$ and $\Sigma_{1,-1}$ is the row vector $(\Sigma_{1,2},\dots,\Sigma_{1,q})$. The conditional variance equals $\sigma_Z^2 = 1 - \Sigma_{1,-1}\Sigma_{-1,-1}^{-1}\Sigma_{1,-1}^\top$. The binary response $Y$ has probabilities given by $p_\theta(y \mid X, Z)$, $y \in \{0,1\}$, as in (6.12), where the intercept and the coefficients of $Z$, and $\gamma_1,\dots,\gamma_q$, are drawn independently and uniformly distributed on the interval $[-1,1]$. The coefficient of $X$, i.e. $\beta = \beta_1$, is chosen in $[0,1]$, with 0 corresponding to conditional independence of $Y$ and $X$ given $Z$. Below are implementation details for all methods considered in the simulations. In the following subsections, these methods are compared in terms of type-I error and power. A further study on the robustness of the MX-based methods with respect to violations of the MX assumption is given in Appendix D.1.

**Conditional randomization *e*-statistic (E-CRT).** The parameter vector $\theta$ in (6.13) is re-estimated after each new observation with the maximum likelihood method, starting from a minimal sample size of $5q+1$, so that $5q$ observations are available for the first parameter estimate. In addition, the probabilities $p_{\hat\theta_k}(y \mid X, Z)$ are truncated to $[\varepsilon, 1-\varepsilon]$ for some small $\varepsilon > 0$. This is to account for the fact that at small sample sizes the MLE sometimes yields predicted probabilities in $\{0,1\}$. We also include an oracle version of this *e*-statistic (E-CRT-O), which uses the true $\theta$ starting from the first observation, instead of the maximum likelihood estimator. For both variants, the integral in the *e*-statistic is approximated by an average over 500 Monte Carlo samples.

**Conditional randomization test (CRT).** The CRT is applied nonsequentially with the likelihood of the logistic regression model as test statistic and 500 samples for randomization. That is, $X$ is sampled 500 times from the conditional distribution given $Z$, the logistic regression model is re-estimated with this simulated covariate, and the likelihood achieved with these models with simulated $X$ is compared to the likelihood achieved with the actual values of $X$.

The following methods are for testing whether the coefficient $\beta$ in the logistic regression model equals zero. Unlike the E-CRT and CRT, they are not based on the MX assumption, but their type-I error guarantee does require that the true probabilities of $Y$ are given by the logistic model (6.12). A comparison is of interest since these methods are, to the best of our knowledge, currently the only ones that allow sequential testing in a logistic model.

**Running maximum likelihood (R-MLE).** We apply the running MLE method by Wasserman et al. (2020, Section 7), an instance of the generic method introduced in that paper which they call *universal inference*. Parameter estimation is also started with a minimum $5q$ observations, and we additionally investigate an $L_1$ penalized version for estimation under the alternative hypothesis, abbreviated as R-MLE-P, which, like for the E-CRT, is to prevent predicted probabilities close to 0 or 1 due to divergence of the MLE. In this second variant, the penalization parameter is chosen by 10-fold cross validation on the available data at the given time, with likelihood as optimization criterion. The penalization parameter is only updated every 10 observations, since the cross validation is computationally expensive.

**Likelihood ratio test (LRT).** The classical asymptotic likelihood ratio test for the null hypothesis that $\beta = 0$ is applied with fixed sample size, and group sequential versions of it with $K$ equally sized groups and the methods by Pocock (1977) (LRT-PK) and O'Brien and Fleming (1979)(LRT-OF).

The results shown in this section are for dimension $q = 4$ of the covariate vector $(X, Z)$. A maximum sample size of $n = 2000$ is considered, after which the evaluation is terminated independently of whether the null hypothesis is rejected. The sequential methods apply the most aggressive stopping rule, namely, reject the null hypothesis as soon as the test statistic exceeds $1/\alpha$ once; more discussion on this is provided in Section 6.4.2. All results, i.e. rejection rates and average sample sizes, are computed over 800 simulations of the data generating process. The same simulation but with higher dimension ($q = 8$) or with negative correlations between the covariates ($\Sigma_{i,j} = (-1)^{i-j}/(1 + |i - j|)$) yields similar results. For the running MLE method, we additionally tested whether not penalizing the coefficient of interest, $\beta$, may achieve higher power, which was not the case.

Simulations are performed in R 4.2 (R Core Team, 2022), with the `glm` function for maximum likelihood estimation in logistic regression, the package `glmnet` (Simon et al., 2011) for $L_1$ penalized estimation, and the package `ldbounds` (Casper et al., 2022) for computing critical values for the Pocock and O'Brien-Fleming group sequential tests. Replication material for the simulations and the case study, as well as additional figures, are available on github.com/AlexanderHenzi/eindependence.

### 6.4.1 Sequential Tests Under the Null

Table 6.1 shows the rejection rates of the different sequential methods with a maximum sample size of $n = 2000$. The methods based on $e$-statistics and the running MLE

| | E-CRT | R-MLE | R-MLE-P | LRT-PK | LRT-OF |
|---|---|---|---|---|---|
| $\alpha = 0.01$ | 0.0075 (0.0031) | 0.0038 (0.0022) | 0.0038 (0.0022) | 0.0138 (0.0013) | 0.0127 (0.0040) |
| $\alpha = 0.05$ | 0.0438 (0.0072) | 0.0038 (0.0022) | 0.0038 (0.0022) | 0.0700 (0.0090) | 0.0599 (0.0084) |

**Table 6.1:** Rejection frequencies (and standard errors) of the different methods, with implementation details as described in Section 6.4.1. Frequencies are given in $[0, 1]$, not in percentages.

yield rejection rates below the nominal level. For the $e$-statistics, the chosen truncation level is $\varepsilon = 0.05$, but the rejection rate is also below $\alpha$ for $\varepsilon = 0$, 0.01, 0.1. The group sequential methods with $K = 20$ equally sized groups, each of size 100, are slightly anti-conservative, and similar rejection rates are obtained for $K = 5$, 10, 40.

### 6.4.2 Simulations Under the Alternative

We proceed to compare the different methods under violations of the null. This is commonly done by comparing the achieved power at given sample sizes $n$ for different effect sizes $\beta$, or the inverse of that function, i.e. the minimum sample size required to achieve power $1 - \eta$,

$$N(\beta, \eta) = \min\{n \in \mathbb{N} \colon P_\beta(\phi_n = 1) \geq 1 - \eta\}.$$

For a fixed type-I error probability $\alpha$, the test decision is $\phi_n = 1\{\max_{m \leq n} S_m \geq 1/\alpha\}$ for anytime-valid tests based on $(S_n)_{n \in \mathbb{N}}$, or $\phi_n = 1\{p_n \leq \alpha\}$ for a method based on a fixed sample size p-value $p_n$. For an anytime-valid test, $N(\beta, \eta)$ can be regarded as the worst case sample size a researcher has to plan for in order to achieve power $1 - \eta$; the actual sample size at rejection may be smaller thanks to optional stopping. Therefore, we additionally consider the average sample size of the anytime-valid tests when evaluation is terminated at the latest at $N(\beta, \eta)$,

$$N_{\mathrm{av}}(\beta, \eta) = \mathbb{E}_{P_\beta}[\min(N(\beta, \eta), \inf\{n \in \mathbb{N} \colon S_n \geq 1/\alpha\})].$$

The rationale is that — even though we have seen that anytime-valid methods retain type-I error validity under arbitrary stopping times — in practice, a natural way to proceed with an anytime-valid test is to run the experiment until either a rejection or a given upper bound on the number of samples is reached. The obvious choice for this upper bound is $N(\beta, \eta)$, as it ensures that the test will have a power of $1 - \eta$. Then $N_{\mathrm{av}}(\beta, \eta)$ gives the average sample size of anytime-valid tests designed for power $1 - \eta$.

A comparison of the different methods in terms of $N(\beta, \eta)$ and $N_{\text{av}}(\beta, \eta)$ is given in Figure 6.1. The group sequential methods are excluded from this figure and are analyzed in more detail at the end of the section. The upper two panels of Figure 6.1 depict $N(\beta, \eta)$ for $1 - \eta = 0.8$, 0.95 as a function of the parameter $\beta$ (for clarity note that, as before, $\beta$ stands for the parameter vector in (6.13); we never use $1 - \beta$ for power). It can be seen that $N(\beta, \eta)$ is higher for the anytime-valid methods than for fixed sample size tests. This is to be expected: the sample size $N(\beta, \eta)$ ensures power $1 - \eta$, but the actual sample size of an anytime-valid test is random and often smaller thanks to early stopping. The lower two panels of Figure 6.1 similarly show $N_{\text{av}}(\beta, \eta)$ as a function of $\beta$. It can be seen that the average sample size is not more and sometimes even less than the sample size of the nonsequential tests. These results suggest that the average sample size to reject the null hypothesis with the E-CRT is not higher than the fixed sample size one would have to plan for with the CRT or LRT. Similar observations have already been made for $e$-statistics in other settings, such as comparisons with the t-test (Grünwald et al., 2024), Fisher's Exact Test (Turner et al., 2024), or the logrank test (ter Schure et al., 2024).

Furthermore, the running MLE method requires much more data to achieve a rejection than the other methods, even with the superior penalized estimation under the null hypothesis. The large sample sizes required for the R-MLE to have a power of 0.95 are mainly due to predicted probabilities close or equal to zero or one at early stages, a problem which is remedied by using penalization, as already proposed by Wasserman et al. (2020); even then, more data are required though. We again emphasize that the running MLE methods are based on different assumptions than the randomization based tests: they do not require the MX assumption, but are only valid for a correctly specified logistic model. However, even when compared to the classical likelihood ratio test, which requires almost the same sample sizes as the CRT, one would have to plan for substantially higher sample sizes with the running MLE methods.

For the conditional randomization $e$-statistics, we tested different levels of truncation $[\varepsilon, 1-\varepsilon]$ for the predicted probabilities. Truncating at a small level $\varepsilon = 0.01$, 0.05 is superior to no truncation, $\varepsilon = 0$, since it prevents $e$-values close to or equal to zero, but if the truncation level becomes too high, $\varepsilon = 0.1$, it limits the power of the test for observations with predicted probability close to zero or one. See Figure D.1 in Appendix D.1 for the case $\varepsilon = 0$. The results shown in this and the next section are for $\varepsilon = 0.05$. In principle, one could also apply a penalized estimator to remedy convergence problems of the MLE, but truncation is computationally less demanding

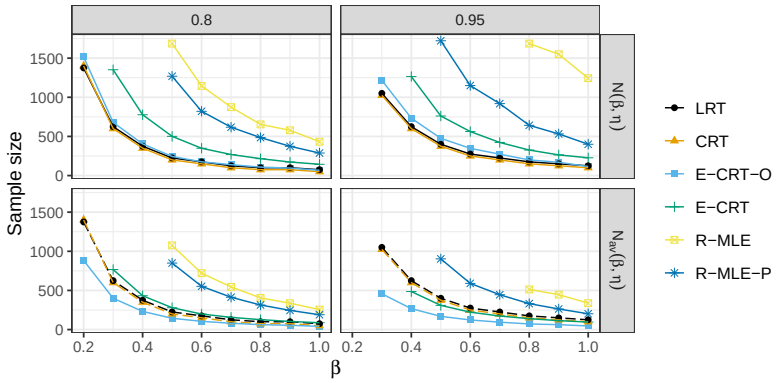



**Figure 6.1:** Sample sizes to achieve power $1-\eta = 0.8, 0.95$ with type-I error 0.05. For the nonsequential methods (LRT, CRT), dashed lines show $N(\beta, \eta)$ also in the lower panels for better comparison. Simulations were conducted up to $n = 2000$, and no results are shown if $N(\beta, \eta) > 2000$ for a given method, $\beta$ and $\eta$, i.e. if more than 2000 observations would be required to achieve a power of $1 - \eta$.

as it does not require the selection of a penalization parameter.

Finally, in Figure 6.2, the conditional randomization $e$-statistic is compared to group sequential methods with the Pocock and the O'Brien and Fleming method with $K = 20$ groups. We see that for small parameter $\beta \in \{0.1, 0.2\}$ these methods achieve a higher power than the $e$-statistics, but as already shown in Table 6.1, the group sequential methods do not control the rejection rate below the nominal level when $\beta = 0$. Also, the E-CRT yields slightly more rejections at small sample sizes than the group sequential methods. As $\beta$ increases, the conditional randomization $e$-statistics tend to outperform the O'Brien and Fleming method, and achieve a rejection rate very close to Pocock's method. Even for large $\beta$, the O'Brien and Fleming requires higher sample sizes due to the fact that the method is designed to yield fewer rejections with small samples. Different numbers of groups for the group sequential methods give similar results, except for the fact that rejecting at small sample sizes becomes impossible if the number of groups is small and the group size large. The performance of the $e$-statistics compared to the group sequential methods is in line with the results of ter Schure et al. (2024) for survival analysis. Also in their study, group sequential methods and alpha-spending approaches, which have to stop at a certain maximum sample size, tend to achieve a higher power than open-ended tests based on $e$-statistics, which do not require a finite upper bound on the sample size.

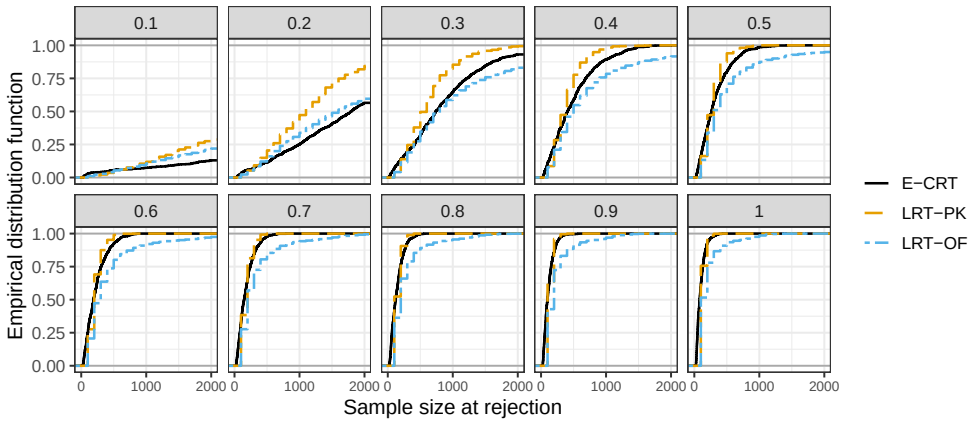



**Figure 6.2:** Empirical distribution of the sample size at rejection, with level $\alpha = 0.05$, for the randomization based $e$-statistics and group sequential methods, and $\beta \in \{0.1, 0.2, \ldots, 1\}$.

## 6.5 Data Application

Berrett et al. (2020) analyze conditional independence relations in the Capital Bikeshare data set. Code and data for their study are available on `https://rinafb.github.io/research/` and `https://ride.capitalbikeshare.com/system-data`. The data set collects information on bike rides with the Capital bikeshare System in Washington DC. One question in the analysis by Berrett et al. (2020) is whether there is dependence of the duration of ride and the binary variable member type, distinguishing casual users from people who purchased a long-term membership, conditional on the start location, destination, and the time of the day during which the bike ride took place.

Let $X$ denote the logarithm of the ride duration and denote by $Z$ the three dimensional vector of starting point, destination of ride, and of the starting time. We model the distribution of $X \mid Z$ as Gaussian, $\mathcal{N}(\mu_Z, \sigma_Z^2)$ in exactly the same way as Berrett et al. (2020). Separately for each combination of starting point and destination, the conditional mean $\mu_Z$ and variance $\sigma_Z^2$ are estimated by a kernel regression of $X$ with the time of day as covariate; details are given in Appendix B of Berrett et al. (2020). We also apply the same preselection criteria, namely, exclude data from weekends and holidays, and values of $Z$ where the estimation might be imprecise due to scarce data. The member type is encoded as $Y \in \{0, 1\}$, with $Y = 0$ referring to casual members.

The data analyzed is from the months September to November in 2011. Unlike

Berrett et al. (2020), who took the months September and November as training data for estimating the distribution of $X \mid Z$ and October as test data, we perform estimation on data of September and October and perform tests on the November data, which would be the natural order for real-time sequential analysis. The test data set, after the application of selection criteria, contains 7173 observations. The training data consists of 158 741 observations. Due to the temporal structure of the data there might be some short lag autocorrelation between the observations, but we did not find any distorting influence of this on the results presented below.

Berrett et al. (2020) apply the conditional randomization test with the test statistic $|\mathrm{cor}(Y, X - \mathbb{E}_{Q_Z}[X])|$. Applied to the November data, this yields a p-value of essentially zero: the observed value of the test statistic was greater than any value obtained with simulated $X$, over 10 000 simulations. For the $e$-statistics, we model the probability that $Y = 1$ with logistic regression, taking $X$ and $\mu_Z$ as covariates and starting with a minimal sample size of 200 observations in the test data. A full model including $Z$ instead of only $\mu_Z$ would be problematic due to the high number of combinations of starting points and destinations relative to the size of the test data. We do not have this limitation in the estimation of the distribution of $X \mid Z$ due to the much larger size of the training data set. However, for a valid test it is not necessary to include $Z$ itself in the model. The probability predictions from the logistic model are then truncated to the interval $[0.01, 0.99]$. For the sequential analysis, the rides are arranged in the order of the start date and time of ride. Figure 6.3 shows how evidence accumulates over time. At the end of the period, an $e$-value of more than $10^6$ is attained, giving decisive evidence against the null hypothesis of conditional independence. An $e$-value of $10^4$ is already reached at the 4380th observation, on November 15th, and hence after about half of the total observation time.

## 6.6 Discussion, Related and Future Work

We have proposed and analyzed anytime-valid tests of conditional independence in the model-X setting. Our method gives a general procedure to transform statistics that measure conditional independence to $e$-statistics. We have shown that for a simple alternative, using the conditional density as statistic leads to the growth rate optimal $e$-statistic, and derived a bound on the inflation of type-I error under violations of the MX assumption.

Duan et al. (2022) and Shaer et al. (2023) have also proposed methods to test CI under MX, but they address the problem in fundamentally different ways than

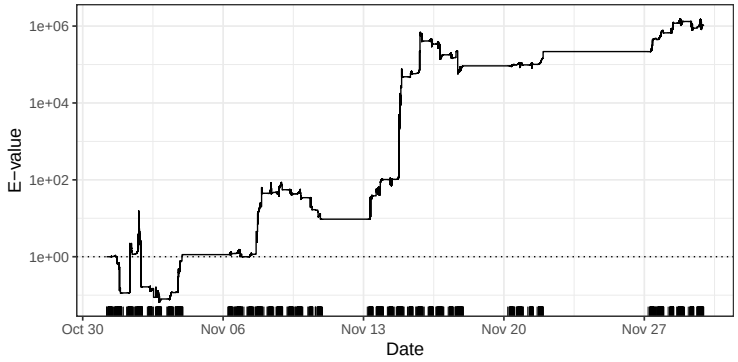



**Figure 6.3:** $E$-value for the bike sharing data set. Lines at the bottom indicate observation times.

we do. In the fully sequential setting, Duan et al. (2022, Appendix E.4) construct a test martingale which grows if a researcher can better than randomly predict a binary $X_n$ given past information and $(Y_n, Z_n)$; notice, however, that this is just a small part of their work, which, for example, also includes tests for a batch setting. Shaer et al. (2023) construct a test martingale which measures how much better a forecaster can predict $Y_n$ based on past data and general $(X_n, Z_n)$, relative to a forecaster only having access to a randomized $\tilde{X}_n \sim Q_{Z_n}$. A derandomization of this procedure is obtained by taking the expectation over $\tilde{X}_n$, which yields similar test martingales as in Theorem 6.1. Both of these methods are very flexible in the choice and tuning of the prediction method, and may therefore behave quite differently from ours. However, ours are justified in terms of the strong GRO optimality criterion by Proposition 6.4 and Corollary 6.5, whenever a fast-converging estimator is used. While this makes our approach in some sense the optimal one, it requires specifying a reasonable (potentially nonparametric) set of densities $\mathcal{F}$ to define the estimator $\hat{f}$. We can basically use any set of densities we like, as long as the estimators converge in information in the sense below Corollary 6.5. However, it may not always be easy to find $\mathcal{F}$ for which estimation is computationally feasible and practically successful, especially if $Z$ is high-dimensional. In that case the approaches of Duan et al. (2022) and Shaer et al. (2023) may have the advantage of being more flexible — whether or not this is the case may be domain-dependent, and is an interesting avenue for future research. Finally, the issue of robustness with respect to misspecification of the distribution of $X|Z$ is only assessed by Shaer et al. (2023) through simulations, whereas our Theorem 6.6 yields the same worst-case bound for our approach as in the batch setting.

For a comparison of our method to the CRT, the following aspects are worth highlighting. Our simulations suggest that our anytime-valid tests with optional stopping do not need more samples, on average, to achieve the same power as the CRT. To be precise, for a given desired power, researchers have to plan for a higher maximum sample size with our anytime-valid tests. But thanks to early stopping, the average sample size of experiments is not more or even less than for the classical CRT. This confirms the findings by Grünwald et al. (2024); ter Schure et al. (2024); Turner et al. (2024) on other anytime-valid tests and their nonsequential counterparts. In terms of computational complexity, our method is not less efficient than the classical CRT if the functions $h_n$ in the test martingale can be updated in constant time; in our application on logistic regression we did not make use of recursive updating, though. A slight advantage of our method compared to the fixed sample size CRT is that the latter requires at least $\lceil 1/\alpha - 1 \rceil$ resamples in order to be able to obtain a p-value below $\alpha$, whereas the test martingale $(S^{\mathrm{CI}}_{h^n})_{n\in\mathbb{N}}$ can exceed any level independently of the number $M$ of resamples. If the strategy of Proposition 6.2 is applied, then each factor in $S^{\mathrm{CI}}_{h^n}$ is bounded by $M+1$, but not their cumulative product.

There are various avenues for future research. With respect to robustness, our Theorem 6.6 shows that for a fixed upper bound $N$ on the sample size, the sequential test achieves the same worst case inflation of rejection rate as the nonsequential CRT with a sample size $N$. It would be of interest to investigate the robustness of the test when this upper bound grows and the approximation $\hat{Q}_Z$ of the conditional distributions of $X \mid Z$ are sequentially updated with new samples. Furthermore, our simulations indicate that our $e$-statistics have competitive power compared to existing methods with relatively small dimension of $Z$, but as stated above, further research is necessary to investigate suitable $e$-statistics and their power when $Z$ is of higher dimension.

# 7 | Tests of Group Invariance

One of the key assumptions underlying our approach to the construction of anytime-valid tests in the previous chapters was that the data were independent. We now consider a setting where this assumption does not necessarily hold. In particular, we study the problem of testing for invariance under a group of transformations, which includes many standard statistical hypothesis tests, such as those for normality and exchangeability. We show that, regardless of any dependence structure in the original data, the invariance properties of the problem can be used to construct a sequence of random variables that are i.i.d. under the null hypothesis. In fact, these transformed data always have the same distribution under the null hypothesis, that is, the null becomes simple. Consequently, it is straightforward to compute the log-optimal $e$-statistic for a given alternative, as it is simply given by the likelihood ratio between the alternative and (now simple) null. It can be shown that the cumulative product of these log-optimal $e$-statistics, the likelihood ratio process, is the log-optimal $e$-process among all $e$-processes that are functions of the transformed data. Remarkably, under some assumptions on the alternative, it is sometimes even log-optimal among all $e$-processes (that is, which are functions of the original data).

We furthermore apply this method to extend recent anytime-valid tests of independence, which leverage exchangeability, to work under general group invariances. Additionally, we show applications to testing for invariance under subgroups of rotations, which corresponds to testing the Gaussian-error assumptions behind linear models.

## 7.1 Introduction

Symmetry plays a crucial role in statistical modeling. Most models, either explicitly or implicitly, introduce assumptions of distributional symmetry about the data. For example, any distribution under which data are independent and identically distributed is symmetric under permutations of the data points and any regression model with Gaussian errors is symmetric under certain rotations of the data. If these symmetries are actually present in the data, employing symmetric models yields advantages for various objectives. These objectives include max-min optimality in hypothesis tests (Lehmann and Romano, 2005) (see also Chapter 8), admissibility of estimators (Brown, 1966), and increased predictive performance of neural networks (Cohen and Welling, 2016). On the other hand, if the symmetries are absent from the data, the use of a symmetric model may lead to poor performance in these same tasks. We address the problem of testing for the presence of symmetries in the data.

The presence of a symmetry is formalized as a null hypothesis of distributional invariance under the action of a group (in the algebraic sense). Perhaps the most prominent example is infinite exchangeability—the hypothesis that the distribution of any finite data sequence is invariant under the group of all permutations. The null hypothesis of exchangeability is at the heart of classic methods such as permutation tests (Fisher, 1936; Pitman, 1937) and rank tests (Sidak et al., 1999). Tests for other symmetries have also been studied, including tests for rotational symmetry, which corresponds to invariance under the orthogonal group (Baringhaus, 1991), symmetries for data taking values on groups (Diaconis, 1988), and more general frameworks (Lehmann and Stein, 1949; Chiu and Bloem-Reddy, 2023). The majority of tests in this line of work are designed for fixed-sample experiments—the amount of data to be collected is determined before the experiment. In this chapter, we focus on testing for the presence of symmetries sequentially and under continuous monitoring.

In the applications that interest us, data are analyzed as they are collected, and the decisions to either stop and reach a conclusion or to continue data collection may depend on what has been observed so far. Hypothesis tests that retain type-I error control under such flexible data collection schemes have been called tests of power one (Robbins and Siegmund, 1974; Lai, 1977), and, more recently, anytime-valid tests (Ramdas et al., 2023). The main insights in this line of work are that a test martingale—a nonnegative martingale with expected value equal to one—can be monitored continuously, and that a test that rejects when the test martingale exceeds a fixed threshold maintains type-I error control uniformly over time (Ramdas et al.,

2020; Shafer, 2021). More generally, the minimum over a family of test martingales—an e-process—can be monitored (Ramdas et al., 2020).

While anytime-valid tests of general symmetries have not received much attention, the specific case of infinite exchangeability has been studied classically. For example, sequential rank tests, which can be interpreted as tests of exchangeability (also called tests of randomness in the literature), have been studied. Sen and Ghosh (1973a,b, 1974) develop asymptotic approximations and law-of-the-iterated-logarithm-type inequalities for linear rank statistics that hold uniformly over the duration of the experiment. More recently, Ramdas et al. (2022) and Saha and Ramdas (2024) developed e-processes for the hypothesis of infinite exchangeability under specific assumptions (binary and paired data, respectively). Anytime-valid tests of exchangeability that do not require any additional assumptions are addressed in the work on conformal prediction (Vovk et al., 2003, 2005). Conformal prediction, perhaps best known as a framework for uncertainty quantification for point predictors, can also be used to produce test martingales to test for exchangeability. In the context of conformal prediction, test martingales are called conformal martingales. Most crucially for our present purposes, Vovk et al. (2005) show that conformal martingales cannot only be used to test for infinite exchangeability, but also to test whether data are generated by a fully general class of sequential data-generating mechanisms, called online compression models (see Section 7.3). It is natural to ask whether distributionally symmetric models define online compression models, as the conformal martingales built by Vovk et al. (2005) would automatically yield tests of distributional symmetry. Unfortunately, this is not true in general.

In this chapter, we show that the above difficulty can be circumvented: Under natural conditions, a distributionally symmetric model does define an online compression model. Furthermore, we show that the resulting conformal martingales are optimal in a specific sense. Indeed, we show that the resulting martingales are likelihood ratios against implicit alternatives and prove that they are optimal—in a sense that is specified in Section 7.2—for testing against that particular alternative. We use these constructions to abstract and generalize existing tests of independence under the assumption of exchangeability (Henzi and Law, 2024) to tests of independence under general symmetries. Finally, we build tests for the Gaussian-error assumptions behind linear models by testing for invariance under subgroups of the orthogonal group.

The rest of this document is organized as follows. Section 7.2 formally introduces the problem of anytime valid testing for distributional invariance, and the optimality criterion that we employ. Then, in Section 7.3, the connection between group-invariant

distributions and online compression models is shown. This connection is used in Section 7.4 to construct test martingales against the hypothesis of distributional invariance; the optimality of this procedure is shown in Section 7.4.2. Section 7.5 shows applications to test the assumptions of linear models, testing sign-invariant exchangeability, and independence testing. Finally, Section 7.6 discusses a potential direction for future work.

## 7.2 Problem Statement

Suppose that we observe data $X_1, X_2, \ldots$ sequentially and that they take values in some topological space $\mathcal{X}$. In our examples, $\mathcal{X} = \mathbb{R}$. Note that we assume neither that these observations are independent nor that they are identically distributed; we only assume that they are sampled from a distribution on infinite sequences. Furthermore, for each $n = 1, 2, \ldots$, we assume that $G_n$ is a compact topological group (in the algebraic sense) that acts continuously on $\mathcal{X}^n$. Here, a topological group is a group that is equipped with a topology under which the group operation, seen as a function $G_n \times G_n \to G_n$, is a continuous map. A (left) group action is a map $\varphi : G_n \times \mathcal{X}^n \to \mathcal{X}^n$ that satisfies, for any $g, h \in G_n$ and $x^n \in \mathcal{X}^n$, that $\varphi(h, \varphi(g, x^n)) = \varphi(hg, x^n)$. To alleviate notation, when the action is clear from context, we write $gx^n$ instead of $\varphi(g, x^n)$. In our examples, the group $G_n$ has a representation as a group of $n \times n$ matrices and the group acts on $\mathbb{R}^n$ by matrix multiplication. We are interested in testing the null hypothesis of invariance of the data under the action of the sequence of groups $(G_n)_{n\in\mathbb{N}}$, that is,

$$\mathcal{H}_0 : gX^n \overset{\mathcal{D}}{=} X^n \quad \text{for all } g \in G_n \text{ and all } n \in \mathbb{N}, \tag{7.1}$$

where $X^n = (X_1, \ldots, X_n)$ and $\overset{\mathcal{D}}{=}$ signifies equality in distribution. At this level of generality, one can build pathological examples of (7.1) that cannot be tested; more structure is needed (see Section 7.3). The next example contains simple instances of the problems that are amenable to our general framework.

**Example 7.1** (Exchangeability, rotational symmetry, and compact matrix groups)**.** For tests of infinite exchangeability, the null hypothesis is given by

$$\mathcal{H}_0 : X_1, \ldots, X_n \text{ are exchangeable for each } n \in \mathbb{N}.$$

By definition, this can be rewritten as

$$\mathcal{H}_0 : (X_{\pi(1)}, \ldots, X_{\pi(n)}) \overset{\mathcal{D}}{=} (X_1, \ldots X_n) \text{ for all } \pi \in S(n) \text{ and } n \in \mathbb{N},$$

where $S(n)$ denotes the group of permutations on $n$ elements. In terms of the notation above, the relevant sequence $(G_n)_{n\in\mathbb{N}}$ of groups is $G_n = S(n)$ and the group action may be written as $\pi X^n = (X_{\pi(1)}, \ldots, X_{\pi(n)})$ for each permutation $\pi \in S(n)$. Note that $S(n)$ can be represented through the group of $n \times n$ permutation matrices (matrices with exactly one entry of 1 in each row and each column and 0 in all other entries). The hypothesis above coincides with that of distributional invariance under multiplication of the data $(X_1, \ldots, X_n)$ by $n \times n$ permutation matrices for all $n$.

Similarly, in the case of tests for sphericity, that is, invariance under rotations of data, the relevant sequence of groups is $G_n = \mathrm{O}(n)$. Here, $O(n)$ denotes the orthogonal group—the group of all $n \times n$ matrices $O$ with orthonormal columns, that is, such that $O^T O = I$. Details are given in Section 7.5.2. The action of permutation and orthogonal groups are special examples of the actions of classic compact matrix groups on $\mathbb{R}^n$ (Meckes, 2019). With adjustments, invariance under any of these classic compact matrix groups is also an instance of the hypothesis in 7.1.

**Anytime-valid tests** We are interested in constructing sequential tests for $\mathcal{H}_0$ as in (7.1) that are anytime-valid at some prescribed level $\alpha \in (0, 1)$. Here, a sequential test is a sequence $(\varphi_n)_{n\in\mathbb{N}}$ of rejection rules $\varphi_n : \mathcal{X}^n \to \{0, 1\}$ and we say that it is anytime valid for $\mathcal{H}_0$ at level $\alpha$ if

$$Q(\exists n \in \mathbb{N} : \varphi_n = 1) \leq \alpha \text{ for any } Q \in \mathcal{H}_0.$$

Notice that this is a type-I error guarantee that is valid uniformly over all sample sizes: the probability that the null hypothesis is ever rejected by $(\varphi_n)_{n\in\mathbb{N}}$ is controlled by $\alpha$. The main tools for constructing anytime-valid tests are test martingales (Ramdas et al., 2020; Shafer, 2021; Grünwald et al., 2024) and minima thereof, e-processes—see Ramdas et al. (2020) for a comprehensive overview. We now define them.

**Test martingales** A sequence of statistics of the data is a test martingale if it is non-negative, starts at one, and is a supermartingale under every element of $\mathcal{H}_0$. Formally, let $\mathbb{G} = (\mathcal{G}_n)_{n\in\mathbb{N}}$ be a filtration of $\sigma$-algebras such that $\mathcal{G}_n \subseteq \sigma(X^n)$, where $\sigma(X^n)$ denotes the $\sigma$-algebra induced by $X^n$. Then a sequence of statistics $(M_n)_{n\in\mathbb{N}}$ that is adapted to $\mathbb{G}$ is a test martingale for $\mathcal{H}_0$ with respect to $\mathbb{G}$ if $\mathbf{E}_Q [M_n \mid \mathcal{G}_{n-1}] \leq M_{n-1}$

for all $Q \in \mathcal{H}_0$ and $M_0 = 1$. The main utility of test martingales is that, under $\mathcal{H}_0$, they take large values with small probability. This is quantified by Ville's inequality (Ville, 1939), which shows that the sequential test given by $\varphi_n = \mathbf{1}\{M_n \geq 1/\alpha\}$ is anytime valid.

**Lemma 7.1** (Ville's inequality)**.** *Let $(M_n)_{n\in\mathbb{N}}$ be a test martingale with respect to some filtration $(\mathcal{G}_n)_{n\in\mathbb{N}}$ under all elements of $\mathcal{H}_0$, then*

$$\sup_{Q\in\mathcal{H}_0} Q(\exists n \in \mathbb{N} : M_n \geq 1/\alpha) \leq \alpha.$$

*Proof.* Fix $Q \in \mathcal{H}_0$. Doob's optional stopping theorem states that $\mathbb{E}_Q[M_\tau] \leq 1$ for any stopping time $\tau$ that is adapted to $(\mathcal{G}_n)_{n\in\mathbb{N}}$ (Durrett, 2019, Theorem 5.7.6). Markov's inequality implies that $Q(M_\tau > \frac{1}{C}) \leq C$ for any $C > 0$. Applying this to the stopping time $\tau^* = \inf\{n \in \mathbb{N} : M_n \geq \frac{1}{\alpha}\}$ shows the result. □

Test martingales that make use of external randomization will also prove useful; we will call them randomized test martingales. For randomized test martingales, we append an independent random number $\theta_n \sim \text{Uniform}([0, 1])$ to each $X_n$, that is, we let $Y_n = (X_n, \theta_n)$ and consider test martingales that are functions of $Y_n$ rather than $X_n$.

Test martingales are part of a broader class of processes, *e*-processes (Ramdas et al., 2023). An *e*-process is any nonnegative stochastic process $E$ such that $\mathbb{E}_Q[E_\tau] \leq 1$ for all $Q \in \mathcal{H}_0$ and (a subset of) all stopping times $\tau$. Any *e*-process can be turned into an anytime-valid test by thresholding it, that is, $\phi_n = \mathbf{1}\left\{E_\tau \geq \frac{1}{\alpha}\right\}$ is an anytime-valid test for any stopping time $\tau$. This property is often referred to as safety under optional stopping (Grünwald et al., 2024). Relatedly, the product of *e*-processes based on independent data is again an *e*-process. That is, suppose some *e*-processes $E$ and $E'$ are used for independent experiments, yielding stopped process $E_{\tau_1}$ and $E'_{\tau_2}$. Then the product again has the property that $\mathbb{E}_Q[E_{\tau_1}E'_{\tau_2}] \leq 1$ for all $Q \in \mathcal{H}_0$, which is referred to as safety under optional continuation.

**Log-optimality** This type of evidence aggregation by multiplication of e-processes motivates a natural optimality criterion. Indeed, suppose we were to repeatedly run a single experiment, using a fixed *e*-process $E$ and stopping time $\tau$. If we measure the total evidence by the cumulative product of the individual *e*-processes, then the asymptotic growth rate of our evidence under true distribution $P$ will be $\mathbb{E}_P[\log E_\tau]$. It is therefore custom to look for *e*-processes that maximize this asymptotic growth

rate, as can be traced back to Kelly betting (Kelly, 1956). Variants of this criterion have more recently been studied under numerous monikers (Shafer, 2021; Koolen and Grünwald, 2022; Grünwald et al., 2024), but here we shall simply refer to maximizers of this criterion as "log-optimal".

## 7.3 Sequential Group Actions are Online Compression Models

The hypothesis in (7.1) can only be meaningfully tested if the statements regarding group invariance for each $n \in \mathbb{N}$ are consistent with each other; without any further restrictions, invariance of the data at one time may contradict the invariance of the data at a later time. To avoid such situations, we assume that there is a certain structure to the action of sequence of groups $(G_n)_{n\in\mathbb{N}}$ on the sample space, which we will refer to as a sequential group action. After the statement of this definition, we discuss its meaning.

**Definition 7.2** (Sequential group action)**.** We say that the action of the sequence of groups $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is sequential if the following conditions hold.

(i) The sequence $(G_n)_{n\in\mathbb{N}}$ is ordered by inclusion: for each $n$, there is an inclusion map $\imath_{n+1} : G_n \to G_{n+1}$ such that $\imath_{n+1}$ is a continuous group isomorphism between $G_n$ and its image, and the image of $G_n$ under $\imath_{n+1}$ is closed in $G_{n+1}$.

(ii) For all $g_n \in G_n$ and all $x^{n+1} \in \mathcal{X}^{n+1}$ , $\mathrm{proj}_{\mathcal{X}^n}(\imath_{n+1}(g_n)x^{n+1}) = g_n(\mathrm{proj}_{\mathcal{X}^n}(x^{n+1}))$, where $\mathrm{proj}_{\mathcal{X}^n}$ is the canonical projection map $\mathrm{proj}_{\mathcal{X}^n} : \mathcal{X}^{n+1} \to \mathcal{X}^n$ given by $\mathrm{proj}_{\mathcal{X}^n}(x_1, \ldots, x_n, x_{n+1}) = (x_1, \ldots, x_n)$.

(iii) Let $n \geq 1$, $g_n \in G_n$, and $g_{n+1} \in G_{n+1}$. For $x^{n+1} = (x_1, \ldots, x_{n+1}) \in \mathcal{X}^{n+1}$, denote $(x^{n+1})_{n+1} = x_{n+1}$. Then, $g_{n+1} = \imath_{n+1}(g_n)$ if and only if, for all $x^{n+1} \in \mathcal{X}^{n+1}$, $(g_{n+1}x^{n+1})_{n+1} = x_{n+1}$.

In Definition 7.2, item (i) gives an ordering of the sequence of groups by inclusion, (ii) ensures that this inclusion does not change the action of the groups on past data, and (iii) implies that the groups do not act on future data. As a result, invariance of $X^{n-1}$ under $G_{n-1}$ is implied by invariance of $X^n$ under $G_n$ and the individual statements of invariance in (7.1) for each $n$ do not contradict each other. The instances of (7.1) discusssed in Example 7.1 satisfy this assumption; a simpler situation where this is satisfied is given in the next example.

**Example 7.2** (Within-batch invariance)**.** Perhaps the simplest example is when, for each $n$, $G_n$ has a product structure and acts on $\mathcal{X}^n$ componentwise. This is when

$$G_n = H_1 \times H_2 \times \cdots \times H_n$$

for some sequence of topological groups $(H_n)_{n\in\mathbb{N}}$, each $H_n$ acting continuously on $\mathcal{X}$ by $(h, x) \mapsto hx$ and $(g_n, X^n) \mapsto (h_1X_1, \ldots, h_nX_n)$ for each $g_n = (h_1, \ldots, h_n) \in G_n$. This covers the setting where batches of data are observed sequentially and the interest is in testing group invariance within each batch. For example, assume that $\mathcal{X} = \mathbb{R}^k$, each $H_i$ is a fixed group $H$ acting on $\mathbb{R}^k$, and data $X_1^k, X_2^k, \ldots$ are assumed to be i.i.d. copies of a random variable $X^k$. Then (7.1) becomes the problem of testing sequentially whether $X^k \overset{\mathcal{D}}{=} hX^k$ for all $h \in H$, that is, whether the distribution of $X^k$ is $H$-invariant. Koning (2023) treats this batch-by-batch setting for general compact groups.

In a addition to this example, sequential group actions also include more complicated situations where there is "cross-action" between the different data points, as in Example 7.1 (corresponding to exchangeability and sphericity). The details for the case of testing rotational symmetry are given in Section 7.5.2.

We now show that, under the assumption that the group action is sequential, the null hypothesis of invariance is an online compression model. The latter are models for computing online summaries, or compressed representations, of the observed data. When the data is generated by an online compression model, the techniques developed for conformal prediction can be used to construct a sequence of statistics that has a Uniform$([0, 1])$ distribution under the null hypothesis. These statistics can be used to build a conformal (test) martingale, as we will discuss in Section 7.4. Vovk et al. define online compression models in abstract terms; we use a simplified definition here.

**Definition 7.3** (Online compression model, Vovk et al. (2005))**.** An online compression model on $\mathcal{X}$ is a 3-tuple of sequences $((\sigma_n)_{n\in\mathbb{N}}, (F_n)_{n\in\mathbb{N}}, (Q_n)_{n\in\mathbb{N}})$, where

1. $(\sigma_n)_{n\in\mathbb{N}}$ is a sequence of statistics $\sigma_n = \sigma_n(X^n)$; we call $\sigma_n$ a summary of $X^n$,
2. $(F_n)_{n\in\mathbb{N}}$ is a sequence of functions such that $F_n(\sigma_{n-1}, X_n) = \sigma_n$,
3. $(Q_n)_{n\in\mathbb{N}}$ is a sequence of conditional distributions for $(\sigma_{n-1}, X_n)$ given $\sigma_n$.

To show how sequential group invariance defines an online compression model, we first recall some group theory. First, the orbit $G_nX^n$ of $X^n$ under the action of $G_n$ is the set of all values that are reached by the action of $G_n$ on $X^n$, i.e., $G_nX^n =$

$\{gX^n : g \in G_n\}$. In order to identify each orbit, we pick a single element of $\mathcal{X}^n$ in each orbit—an orbit representative—and consider the map $\gamma_n : \mathcal{X}^n \to \mathcal{X}^n$ that takes each $X^n$ to its orbit representative. We call $\gamma_n$ an orbit selector (see Section 7.5 for examples), and we assume that it is measurable. Such measurable orbit selectors exist under weak regularity conditions on $\mathcal{X}^n$ and $G_n$ (see Bondar, 1976, Theorem 2) that hold in all the examples of this chapter. Furthermore, because $G_n$ is a compact group, there exists a unique $G_n$-invariant probability distribution $\mu_n$, called the Haar (probability) measure (Bourbaki, 2004, Chapter VII). The Haar measure plays the role of a uniform probability distribution on compact groups. Finally, it is a fact that the data is uniformly distributed on its orbit conditionally on the orbit where it lays; formally, $X^n \mid \gamma_n(X^n) \overset{\mathcal{D}}{=} U\gamma_n(X^n) \mid \gamma_n(X^n)$, where $U \sim \mu_n$ independently of $X$ (Eaton, 1989, Theorem 4.4).

We now show that if a sequence of groups $(G_n)_{n\in\mathbb{N}}$ acts sequentially on the data, any distribution that is invariant under the action of said sequence defines an online compression model. We use the orbit representative as summary statistic, i.e., we use $\sigma_n = \gamma_n(X^n)$. Then, since the distribution of the data is $(G_n)_{n\in\mathbb{N}}$-invariant, the sequence of conditional distributions of $(\sigma_{n-1}, X_{n-1})$ given $\sigma_n$ is uniform over the orbits as remarked earlier. In this way, we fix $\sigma_n$ and $Q_n$ in Definition 7.3. Furthermore, the next proposition shows that, for sequential group actions, $\sigma_n$ can be computed as a function of $\sigma_{n-1}$ and $X_n$. The proof of this proposition can be found in Appendix E.1 and it uses crucially the assumption that the group action is sequential. This discussion and the following proposition prove that a sequential group-invariant model indeed defines an online compression model, as we state in Corollary 7.5.

**Proposition 7.4.** *If the action of $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is sequential, then there exists a sequence $(F_n)_{n\in\mathbb{N}}$ of measurable functions $F_n : \mathcal{X}^{n-1} \times \mathcal{X} \to \mathcal{X}^n$ such that $F_n(\gamma_{n-1}(X^{n-1}), X_n) = \gamma_n(X^n)$ and $F_n(\,\cdot\,, X_n)$ is a one-to-one function of $\gamma_{n-1}(X^{n-1})$.*

**Corollary 7.5.** *Assume that the action of $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is sequential, let $\widetilde{\mu}_n$ be the uniform distribution on $G_nX^n$ induced by the Haar measure $\mu_n$ on $G_n$, and let $(F_n)_{n\in\mathbb{N}}$ be as guaranteed by Proposition 7.4. Then the tuple*

$$((\gamma_n(X^n))_{n\in\mathbb{N}}, (F_n)_{n\in\mathbb{N}}, (\widetilde{\mu}_n)_{n\in\mathbb{N}}),$$

*defines an online compression model on $\mathcal{X}$.*

## 7.4 Testing Group Invariance With Conformal Martingales

We now construct test martingales for the null hypothesis of distributional symmetry in (7.1) any time that a sequence of groups $(G_n)_{n\in\mathbb{N}}$ acts sequentially on the data $(X_n)_{n\in\mathbb{N}}$. To this end, the invariant structure of the null hypothesis $\mathcal{H}_0$ is used in tandem with conformal prediction to build a sequence of independent random variables $(R_n)_{n\in\mathbb{N}}$ with the following three properties:

1. The sequence $(R_n)_{n\in\mathbb{N}}$ is adapted to the data sequence with external randomization $(X_n, \theta_n)_{n\in\mathbb{N}}$, that is, for each $n \in \mathbb{N}$, $R_n = R_n(X_n, \theta_n)$.

2. Under any element of the null hypothesis $\mathcal{H}_0$ from (7.1), $(R_n)_{n\in\mathbb{N}}$ is a sequence of independent and identically distributed Uniform$[0, 1]$ random variables.

3. The distribution of $(R_n)_{n\in\mathbb{N}}$ is not uniform when departures from symmetry are present in the data.

The construction of these random variables is the subject of Section 7.4.1—additional definitions are needed—and their optimality is the subject of Section 7.4.2. In order to guide intuition, Example 7.3 shows a first example for testing exchangeability, which has previously also been studied by Vovk et al. (2005) and Fedorova et al. (2012). They call the statistics $R_1, R_2, \ldots$ p-values owing to their uniformity. We opt against that terminology here, because typically only small p-values are interpreted evidence against the null hypothesis. However, in the context of testing for symmetry, it is any deviation from uniformity that we interpret as evidence against the null hypothesis. For reasons that will become apparent soon, we call $R_1, R_2, \ldots$ (smoothed) orbit ranks (see Definition 7.7).

With the sequence $(R_n)_{n\in\mathbb{N}}$ at hand, test martingales against distributional invariance are built by testing against the uniformity of $(R_n)_{n\in\mathbb{N}}$. Indeed, any time that $(f_n)_{n\in\mathbb{N}}$ is a sequence of functions $f_i : [0, 1] \to \mathbb{R}$ such that $\int f_i(r)\mathrm{d}r = 1$, the process $(M_n)_{n\in\mathbb{N}}$ given by

$$M_n := \prod_{i\leq n} f_i(R_i) \tag{7.2}$$

is a test martingale for $\mathcal{H}_0$ with respect to $\mathbb{F}$, where $\mathbb{F} = (\sigma(R^n))_{n\in\mathbb{N}}$ and $\sigma(R^n)$ is the $\sigma$-algebra generated by $R^n$. This follows from the fact that $\mathbf{E}_Q\left[M_n \mid \sigma(R^{n-1})\right] = M_{n-1}\int f_n(r)\mathrm{d}r = M_{n-1}$, where we leverage independence and uniformity. The functions $(f_n)_{n\in\mathbb{N}}$ are known as calibrators (Vovk and Wang, 2021). They can be taken

to be any sequence of predictable estimators of the density of $R_1, R_2, \ldots$ (Fedorova et al., 2012), so that the test martingale is expected to grow if the true distribution of the orbit ranks is not uniform, i.e., the null hypothesis is violated. The optimality of this procedure is discussed in Section 7.4.2.

**Example 7.3** (Sequential Ranks)**.** Consider the case of testing exchangeability as discussed in Example 7.1, that is, the case when $\mathcal{X} = \mathbb{R}$ and $G_n = S(n)$ . For each $n$, define the random variables $\widetilde{R}_n = \sum_{i \leq n} \mathbf{1}\{X_i \leq X_n\}$—the rank of $X_n$ among $X_1, \ldots, X_n$. The random variables $\widetilde{R}_1, \widetilde{R}_2, \ldots$ are called sequential ranks (Malov, 1996). It is a classic observation that each $\widetilde{R}_n$ is uniformly distributed on $\{1, \ldots, n\}$, and that $(\widetilde{R}_n)_{n\in\mathbb{N}}$ is a sequence of independent random variables (Rényi, 1962). After rescaling and adding external randomization, a sequence of random variables $(R_n)_{n\in\mathbb{N}}$ can be built from $(\widetilde{R}_n)_{n\in\mathbb{N}}$ such that $(R_n)_{n\in\mathbb{N}}$ satisfies items 1, 2 and 3 at the start of this section. Furthermore, if we denote the uniform measure on $S(n)$ by $\mu_n$, then $\widetilde{R}_n$ can also be obtained from $n^{-1}\widetilde{R}_n = \mu_n\{g : (gX_n)_n \leq X_n\}$. While this rewriting may seem esoteric at this point, it turns out to be the correct point of view for generalization.

### 7.4.1 Conformal Prediction Under Invariance

In general, the statistics $R_n$ will be designed to measure how strange the observations $X^n$ are in contrast to what would be expected under distributional invariance. To this end, the values of $X^n$ are compared to those in the orbit of $X^n$ under the action of $G_n$. In order to measure the "strangeness" of the observations in their orbit, we use an adaptation of the conformity measures introduced by Vovk et al. (2005).

**Definition 7.6** (Conformity measure of invariance)**.** We say that the function $A : \mathcal{X} \times \bigcup_{n=1}^{\infty} \mathcal{X}^n \to \mathbb{R}$ is a conformity measure of invariance if the following holds: if there are $X^n, X'^n \in \mathcal{X}^n$, such that $A(X_i, \gamma_n(X^n)) = A(X'_i, \gamma_n(X'^n))$ for all $i \in \{1, \ldots, n\}$, then, for all $g \in G_n$, we also have that $A((gX^n)_i, \gamma_n(X^n)) = A((gX'^n)_i, \gamma_n(X'^n))$ for all $i \in \{1, \ldots, n\}$.

The group-related condition on $A$ that appears in Definition 7.6 is an addition to that of Vovk et al. (2005); it ensures that the action of $G_n$ on $\mathcal{X}^n$ induces an action on the conformity measure. The intuition of the definition is that when $A$ is properly chosen, $A(X_n, \gamma_n(X^n))$ is a numerical score that indicates how similar $X_n$ is to the other values in its orbit. Therefore, the statistic $\alpha_n = A(X_n, \gamma_n(X^n))$ is called a conformity score. The easiest example is when $\mathcal{X}^n \subseteq \mathbb{R}^n$ because then $A$ defined

by $A(X_n, \gamma_n(X^n)) = X_n$ is a conformity measure of invariance—this is the case in Example 7.3. However, perhaps a more intuitive choice would be $A(X_n, \gamma_n(X^n)) = |X_n - \int_G g\gamma_n(X^n)\,\mathrm{d}\mu_n(g)|^{-1}$, since this quantity is large whenever $X_n$ is close to the average value within the orbit, which is given by $\int_G g\gamma_n(X^n)\,\mathrm{d}\mu_n(g)$. For a more involved example, consider the case where the data points are given by $X_i = (Y_i, Z_i)$ for some outcome $Y_i \in \mathbb{R}$ and a covariate $Z_i \in \mathbb{R}$. Then one might consider $A(X_i, \gamma_n(X^n)) = |Y_i - \hat{Y}_i|^{-1}$, where $\hat{Y}_i$ is the prediction of some regression method that was trained on the orbit of $X_i$. In this case, the intuition is that, if the label is very close to the prediction that is made using all of the values in the same orbit, then $X_i$ must have been very typical of the orbit. For a more detailed discussion on different conformity measures, we refer to Fontana et al. (2023).

Since the scale of the conformity scores is arbitrary—they can be scaled at will—, only comparisons between them are meaningful. Therefore, similar to what happened in Example 7.3, we will rank the observed value of the conformity score $\alpha_n$ among all its possible values on the orbit of the data. To this end, we obtain the distribution of the conformity scores under the null hypothesis using the assumed distributional invariance. Indeed, as discussed in Section 7.3, the distribution of $X^n$ conditional on $\gamma(X^n)$ is uniform on its orbit. This idea gives rise to the (smoothed) orbit ranks $(R_n)_{n\in\mathbb{N}}$ in the next definition.

**Definition 7.7** (Smoothed Orbit Ranks)**.** Fix $n \in \mathbb{N}$, let $A$ be a conformity measure, and let $\alpha_n = A(X_n, \gamma_n(X^n))$ be the associated conformity score. We call $R_n$, defined by

$$\begin{aligned} R_n =& \mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n < \alpha_n\}) + \\ & \theta_n \mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n)) = \alpha_n\}), \end{aligned} \tag{7.3}$$

a (smoothed) orbit rank, where $\mu_n$ denotes the Haar probability measure on $G_n$ and $\theta_n \sim \mathrm{Uniform}[0,1]$ is independent of the data $X^n$.

The simplest case is when the group $G_n$ is finite of size $k$ and $A(X_i, \gamma_n(X^n)) = X_i$. In that case, $\mu_n$ is the discrete uniform distribution on $G_n$ and $R_n = \frac{1}{k}\#\{g \in G_n : (gX^n)_n < X_n\} + \frac{\theta_n}{k}\#\{g \in G_n : (gX^n)_n = X_n\}$.

An important intuition is that the statistic $R_n$ is the CDF of the distribution of $\alpha_n$ conditional on $\gamma_n(X^n)$ evaluated in $\alpha_n$ (with added randomization) under $\mathcal{H}_0$. It follows that if said CDF is continuous, smoothing plays no role in (7.3) and $R_n \perp \theta_n$. It also follows—and this is shown in Theorem 7.8—that each $R_n$ is uniformly distributed on $[0,1]$. Vovk et al. (2005, Theorem 11.2) show that, if the data is generated by an

online compression model, and $\theta_1, \theta_2, \ldots$ are independent, then $R_1, R_2, \ldots$ are also independent. Since Corollary 7.5 shows that a sequential group invariance structure defines an online compression model, it follows that the smoothed orbit ranks form an i.i.d. uniform sequence under the null hypothesis. This is stated in the next theorem, for which we provide a direct proof in Appendix E.1 for completeness .

**Theorem 7.8.** *Suppose that the action of $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is sequential, that $(X_n)_{n\in\mathbb{N}}$ is generated by an element of $\mathcal{H}_0$, and that $\theta_1, \theta_2, \ldots$ are independent. Then $R^n \perp \gamma_n(X^n)$ for each $n$ and $(R_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d.* Uniform$[0,1]$ *random variables.*

### 7.4.2 Optimality

We now show that any martingale based on the smoothed orbit ranks as in (7.2) is a likelihood ratio process and that it is log-optimal against the implicit alternative for which it is built. To this end, let $P$ be a distribution such that for all $n$, conditionally on $R^{n-1}$, $R_n$ has density $f_n$ (with respect to the Lebesgue measure). Technically, the conditional density of $R_n$ is not defined only by $P$, but also by the external randomization. To make this explicit in the following, we will use $\widetilde{P}$ to denote $P$ with external randomization added, that is, $\widetilde{P} = P \times \mathcal{U}^\infty$, where $\mathcal{U}^\infty$ is the uniform distribution on $[0,1]^\infty$. Analogously, for each $Q \in \mathcal{H}_0$, define $\widetilde{Q} = Q \times \mathcal{U}^\infty$.

The discussion below (7.2) shows that $M_n = \prod_{i\le n} f_i(R_i)$ is a test martingale. In fact, $M_n$ is the likelihood ratio for the orbit ranks $R^n$ between $\widetilde{P}$ and $\widetilde{Q}$, since the distribution of $R^n$ under $\widetilde{Q}$ equals the uniform distribution for any $Q \in \mathcal{H}_0$ by Theorem 7.8. Surprisingly, if $\widetilde{P}$ is such that $R_n \perp \gamma_n(X^n)$, then $M_n$ is also the likelihood ratio for the full data $X^n$ between $P$ and an appropriately chosen distribution $Q^* \in \mathcal{H}_0$, as shown in the following proposition. For the sake of brevity, the action of $(G_n)_{n\in\mathbb{N}}$ on $(\mathcal{X}^n)_{n\in\mathbb{N}}$ is assumed to be sequential throughout.

**Proposition 7.9.** *Suppose that $A(\,\cdot\,, \gamma_n(X^n))$ is a one-to-one function for each $n \in \mathbb{N}$, suppose that $P$ is any distribution under which $R_n \perp \gamma_n(X^n)$ for each $n$, and let $f_i$ denote the conditional density of $R_i$ given $R^{i-1}$ under $P$. Let $M_n = \prod_{i\le n} f_i(R_i)$. Then, for $Q \in \mathcal{H}_0$,*

$$\widetilde{Q}\left(M_n = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^n)\right) = 1, \tag{7.4}$$

*where $Q^*$ denotes the distribution under which the marginal of $\gamma_n(X^n)$ coincides with that under $P$, and such that $X^n \mid \gamma_n(X^n) \overset{\mathcal{D}}{=} U\gamma_n(X^n) \mid \gamma_n(X^n)$, where $U \sim \mu_n$ independently from $\gamma_n(X^n)$.*

The distribution $Q^*$ can be thought of as a symmetrization of $P$, since the marginal of $\gamma_n(X^n)$ is the same, but the distribution conditional on $\gamma_n(X^n)$ is defined by symmetry. Proposition 7.9 therefore shows that, if the orbit ranks are independent of the orbit selectors under $P$, then $(M_n)_{n\in\mathbb{N}}$ is the likelihood ratio process between $P$ and a symmetrization thereof. The next theorem uses this representation to show the log-optimality of $(M_n)_{n\in\mathbb{N}}$. Its proof follows that of Theorem 12 of Koolen and Grünwald (2022).

**Theorem 7.10.** *Assume that $A(\,\cdot\,, \gamma_n(X^n))$ is one-to-one for all $n \in \mathbb{N}$ and let $P$ be such that, under $P$, the distribution of $X^n \mid \gamma_n(X^n)$ is absolutely continuous with respect to the uniform distribution. Denote $f_i$ for the density of $R_i \mid R^{i-1}$ under $\widetilde{P}$ and let $M_n = \prod_{i\leq n} f_i(R_i)$. Let $\tau$ be any stopping time and $(E_n)_{n\in\mathbb{N}}$ any e-process for $\mathcal{H}_0$, both with respect to $\mathbb{F}$—the filtration generated by the smoothed ranks. Then it holds that*

$$\mathbf{E}_{\widetilde{P}}\left[\ln M_\tau\right] = \mathbf{E}_{\widetilde{P}}\left[\ln \prod_{i=1}^{\tau} f_i(R_i)\right] \geq \mathbf{E}_{\widetilde{P}}\left[\ln E_\tau\right]. \tag{7.5}$$

*Moreover, if $\widetilde{P}$ is such that $R^n \perp \gamma_n(X^n)$ for all $n$, then for any e-process $E'$ for $\mathcal{H}_0$ w.r.t. $(\sigma(X^n, \theta^n))_{n\in\mathbb{N}}$—the full-data filtration—, it also holds that*

$$\mathbf{E}_{\widetilde{P}}\left[\ln M_\tau\right] \geq \mathbf{E}_{\widetilde{P}}\left[\ln E'_\tau\right]. \tag{7.6}$$

The first part of Theorem 7.10, Equation (7.5), establishes that, under some assumptions on $P$, $(M_n)_{n\in\mathbb{N}}$ is log-optimal for testing group invariance among all e-processes defined only on the orbit ranks. Moreover, the second part of Theorem 7.10, Equation (7.6), states that if the orbit ranks are also independent of the orbit selector under $P$, then $(M_n)_{n\in\mathbb{N}}$ is log-optimal for testing group invariance among all e-processes defined on the full data. The additional assumption of independence between $R^n$ and $\gamma_n(X^n)$ is necessary for (7.6) to hold: if $\widetilde{P}$ is a distribution under which $R_1, \ldots, R_n \not\perp \gamma_n(x^n)$, then the conformal martingale is not in general a likelihood ratio as in (7.4). For the deterministic stopping time $\tau = n$, the log-optimal statistic is instead given by $S_n = \prod_{i=1}^n f_n(R_1, \ldots, R_n \mid \gamma_n(X^n))$, as it can be written as a likelihood ratio (see also Grünwald et al., 2024; Koning, 2023). However, the sequence $(S_n)_{n\in\mathbb{N}}$ does not necessarily define a test martingale or e-process, so that it might not be possible to use it in the construction of an anytime-valid test. Using tests based on the sequential ranks circumvents this issue for such alternatives.

The optimality of $M_n$ in Theorem 7.10 is contingent on oracle knowledge of the true distributions $f_1, f_2, \ldots$, which are unknown in practice. To counter this, past

data can be used to sequentially estimate the true density. This idea has previously been applied for testing exchangeability (Vovk et al., 2005; Fedorova et al., 2012). More precisely, for each $n$, let $\hat{f}_n$ be an estimator of $f_n$ based on $R^{n-1}$, and consider the martingale defined by $\prod_{i=1}^{n} \hat{f}_i(R_i)$. In general, this is suboptimal with respect to an oracle that knows the true density. However, in the case that there exists a density $f$ such that $f_i \equiv f$ for all $i$, i.e. data are i.i.d. under $P$, there is limited loss asymptotically if $\hat{f}_i$ is a good estimator of $f$. In order to judge if an estimator is good for the task at hand, consider the difference in expected growth per outcome for fixed $n$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{P}}[\log f(R_i) - \log \hat{f}_i(R_i)] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\widetilde{P}}[\mathrm{KL}(f\|\hat{f}_i)], \tag{7.7}$$

where $\mathrm{KL}(f\|\hat{g}) = \int_0^1 f(r)\log(f(r)/g(r))\mathrm{d}r$ denotes the Kullback-Leibler divergence whenever $f$ is absolutely continuous with respect to $g$, and the expectation on the right-hand side of (7.7) is over past data (on which $\hat{f}_i$ depends). If (7.7) tends to zero as $n$ grows large, the expected growth per outcome converges to that of the log-optimal test martingale. This motivates the use of density estimation algorithms for which this always happens. Under stringent assumptions—for example, if the density $f$ belongs to an exponential family—sequential Bayesian-update-type algorithms are known to guarantee that (7.7) converges to zero (Kotłowski and Grünwald, 2011). Under weaker assumptions, specialized algorithms exist with the same guarantees (Haussler and Opper, 1997; Cesa-Bianchi and Lugosi, 2001; Grünwald and Mehta, 2019).

## 7.5 Applications and Extension

In this section, we discuss applications and an extension of the theory developed in the previous sections.

### 7.5.1 Sign-Invariant Exchangeability

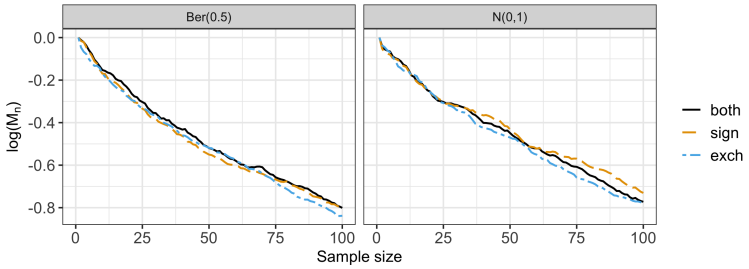
In this subsection, we consider testing for sign-invariant exchangeability (Berman, 1965; Fraiman et al., 2024) with the purpose of illustrating our method on a concrete, basic example, and show its performance through numeric simulation. Real-valued data are sign-invariant if $(X_1, \ldots, X_n) \overset{\mathcal{D}}{=} (\epsilon_1 X_1, \ldots, \epsilon_n X_n)$ for all signs $(\epsilon_1, \ldots, \epsilon_n) \in \{-1, 1\}^n$. We consider $\{-1, 1\}^n$ as a group with componentwise multiplication as operation. Data are sign-invariant exchangeable if they are both sign-invariant and exchangeable. That is, if $(X_1, \ldots, X_n) \overset{\mathcal{D}}{=} (\epsilon_1 X_{\pi(1)}, \ldots, \epsilon_n X_{\pi(n)})$ for all signs $(\epsilon_1, \ldots, \epsilon_n) \in$

$\{-1,1\}^n$ and all permutations $\pi \in S(n)$. The null hypothesis of sign-invariant exchangeability is therefore equivalent to the distributional invariance of $X^n$ under the action of $G_n = \{-1,1\}^n \times S(n)$. The orbit of $X^n$ under $G_n$ is given by the set $\{(\epsilon_1 X_{\pi(1)}, \ldots, \epsilon_n X_{\pi(n)}) : (\epsilon_1, \ldots, \epsilon_n) \in \{-1,1\}^n, \pi \in S(n)\}$. Since $G_n$ is finite, the Haar measure is the discrete uniform distribution. Furthermore, because the data are assumed to be real, we can take $A(X_n, \gamma_n(X^n)) = X_n$. Let $X'_1, \ldots, X'_{2n}$ be given, for $i = 1, \ldots, n$, by $X'_i = X_i$ and $X'_{n+i} = -X_i$. The smoothed orbit rank in (7.3) becomes

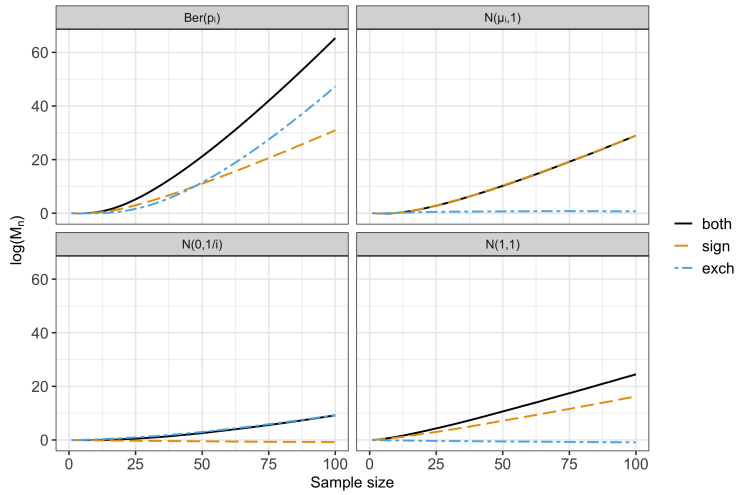
$$R_n = \frac{\#\{i \leq 2n : X'_i < X_n\}}{2n} + \theta_n \frac{\#\{i \leq 2n : X'_i = X_n\}}{2n}. \tag{7.8}$$

These statistics can be computed upon observing the data, and standard density estimation algorithms can used to estimate their density. Following Section 7.4, for each $n$ we use $R_1, \ldots, R_{n-1}$ to build an estimate $\hat{f}_n$ of the density and use $M_n = \prod_{i \leq n} \hat{f}_i(R_i)$ as a test martingale.

We investigate how martingales obtained in this manner behave through simulations. For these experiments, we used the R language (R Core Team, 2022). The density estimation was performed using the kernel density estimation that is implemented in the `Stats` package. However, standard kernel density estimation can lead to poor performance around the boundaries. This is because these algorithms are designed to estimate densities supported on $\mathbb{R}$ and not just on $[0,1]$. Following Fedorova et al. (2012), a solution to this problem is (in the case of testing exchangeability) to reflect the sequence of orbit ranks to the left from zero and to the right from one. Then, the estimate is computed using the extended sample $\cup_{i=1}^n \{-R_i, R_i, 2 - R_i\}$. Finally, the estimated density is set to zero outside of the unit interval and then normalized. We have used this same procedure here. Furthermore, for the sake of comparison, we also include the conformal martingale that would be obtained if testing either for exchangeability exclusively or for sign-invariance. The results are shown in Figure 7.1. We see that, if data are sign-invariant exchangeable—i.i.d. Rademacher or i.i.d. Normal$(0,1)$ in our experiments—, the conformal martingales are indeed martingales and do not take large values, as expected based on the discussion in Section 7.2. Under the alternative, the statistic $M_n$ is no longer a martingale, and it does grow. However, the methods that test for only one of the two symmetries (either exchangeability or sign invariance separately) do not detect alternatives for which that particular symmetry is not violated, but the other is (see Figure 7.1b). On the other hand, the conformal martingale based on $R_n$ as described in (7.8) detects all of the alternatives. In fact, for the alternative where each $X_i \in \{-1,1\}$ and $X_i = 1$ with probability $p_i = 1 - 1/i$




**(a)** Under null models: $X_i = \pm 1$ w.p. 0.5 and $X_i \sim N(0,1)$.




**(b)** Statistics under alternative models (all independent). Upper left corner: $X_i = 1$ w.p. $p_i = 1 - 1/i$ and $X_i = 1$ with probability $1 - p_i$. Upper right corner: $X_i \sim N(i/10, 1)$. Lower left corner: $X_i \sim N(0, 1/i)$. Lower right corner: $X_i \sim N(1, 1)$.

**Figure 7.1:** The logarithm of the conformal martingale against sample size for three different methods: testing for both sign-invariance and exchangeability, or only one of the two. A test built against sign-invariance and exchangeability can detect the absence of either of those two invariances while a test that is built to detect only one of then cannot achieve the same goal (see Section 7.5.1). Data are independent under all considered models. The results were averaged over 500 repetitions.

and $X_i = -1$ with probability $1 - p_i$ indepently, the corresponding test martingale is even log-optimal among all e-processes. This is due to the fact that, regardless of the observed data, the orbit of $X_i$ is always the set $\{-1, 1\}$. Therefore, the orbit selector can be chosen to be $\gamma_n(X^n) = (1, \ldots, 1)$ independently of the data, such that $R_n \perp \gamma_n(X^n)$. The log-optimality then follows from Theorem 7.10.

### 7.5.2 The Orthogonal Group and Linear Models

Consider testing whether the data we observe are drawn from a spherically symmetric distribution, i.e., $\mathcal{X} = \mathbb{R}$ and $G_n = \mathrm{O}(n)$, where $\mathrm{O}(n)$ is the orthogonal group in dimension $n$. Testing for spherical symmetry is equivalent to testing whether the data are generated by a zero-mean Gaussian distribution. Indeed, any distribution on $\mathbb{R}^\infty$ for which the marginal of the first $n$ coordinates is spherically symmetric for any $n$, is a mixture of i.i.d. zero-mean Gaussian distributions (Bernardo and Smith, 2009, Proposition 4.4). It follows that any process that is a supermartingale under all zero-mean Gaussian distributions is also a supermartingale under spherical symmetry and vice-versa. This implies that, for the purpose of testing with test (super)martingales, the two hypotheses are equivalent. We show how this fits in our setting, and defer the application to regression to Appendix E.2.

We now check that testing spherical symmetry fits in our setting, i.e., that Definition 7.2 is fulfilled. Consider the inclusion of $\mathrm{O}(n)$ in $\mathrm{O}(n+1)$ given by

$$\imath_{n+1}(O_n) = \begin{pmatrix} O_n & 0 \\ 0 & 1 \end{pmatrix}$$

for each $O_n \in \mathrm{O}(n)$. Using the canonical projections in $\mathbb{R}^n$, Definition 7.2 is readily checked. Since the data are real, we can consider the simple measure of conformity $A(X_n, \gamma_n(X^n)) = X_n$. An orbit selector is given by $\gamma_n(X^n) = \|X^n\| e_1$, where $e_1$ is the unit vector $e_1 = (1, 0, \ldots, 0)$. For simplicity, we assume that the distribution of $X^n$ has a density with respect to the Lebesgue measure for each $n$, so that $R_n = \mu_n(\{O_n \in \mathrm{O}(n) : (O_n X^n)_n < X_n\})$—no external randomization is needed. Rather than thinking of $\mu_n$ as a measure on $O(n)$, one can think of it as the uniform measure on $S^{n-1}(\|X^n\|)$. This way, $R_n$ can be recognized to be the relative surface area of the hyper-spherical cap with co-latitude angle $\varphi_n = \pi - \cos^{-1}(X_n/\|X^n\|)$. Li (2010) shows that an explicit expression for this area is given by

$$R_n = \begin{cases} 1 - \frac{1}{2} I_{\sin^2(\pi-\varphi_n)}\left(\frac{n-1}{2}, \frac{1}{2}\right) & \text{if } \varphi_n > \frac{\pi}{2}, \\ \frac{1}{2} I_{\sin^2(\varphi_n)}\left(\frac{n-1}{2}, \frac{1}{2}\right) & \text{else,} \end{cases} \tag{7.9}$$

where $I_x(a, b)$ denotes the regularized beta function, $I_x(a, b) = \frac{B(x,a,b)}{B(1,a,b)}$ for $B(x, a, b) = \int_0^x t^{a-1}(1-t)^{b-1}$ for $0 \le x \le 1$.

Note that $\varphi_n > \frac{\pi}{2}$ if and only if $X_n > 0$ and that $\sin^2(\varphi_n) = 1 - \frac{X_n^2}{\|X^n\|^2}$, so that (7.9) equals the CDF of the t-distribution with $n-1$ degrees of freedom evaluated

in $t = \sqrt{n-1}X_n/\|X^{n-1}\|$. If $X^n \sim \mathcal{N}(0, \sigma^2 I_n)$, then $t$ is the ratio of a normally distributed random variable and an independent chi-squared-distributed random variable. Therefore, $t$ has a t-distribution with $n-1$ degrees of freedom. The test thus obtained is a type of sequential t-test that has, to the best of our knowledge, not been considered previously.

This example can be extended to testing for centered spherical symmetry, i.e., whether $X^n = \mu \mathbf{1}_n + \epsilon^n$, where $\mathbf{1}_n$ is the all-ones $n$-vector, $\mu \in \mathbb{R}$ and $\epsilon^n$ is spherically symmetric for every $n \in \mathbb{N}$. By similar reasoning as above, this is equivalent to testing whether the data is i.i.d. Gaussian with any mean/variance. Even more, by considering different isotropy groups, one can also cover the case where the mean $\mu$ is not fixed, but depends on covariates. The ideas needed in that case are similar; we show how deal with the added complexity in Appendix E.2.

### 7.5.3 Modification for Independence Testing

We now propose a minor modification of the conformal martingales from the previous section that can be used to test for independence. Formally, fix $K \in \mathbb{N}$ and suppose that at each time point $n \in \mathbb{N}$, a $K$-dimensional vector $X_n = (X_{1,n}, \ldots, X_{K,n}) \in \mathcal{X}^K$ is observed. We are interested in testing the null hypothesis that states that: (1) for each $k = 1, \ldots, K$ and each $n$ the vectors $(X_{k,1}, \ldots, X_{k,n})$ are $G_n$-invariant, and (2) $(X_{k,1}, \ldots, X_{k,n}) \perp (X_{k',1}, \ldots, X_{k',n})$ for all $k \neq k' \in \{1, \ldots, K\}$. Under this hypothesis, the data is invariant under the sequential action of $(\widetilde{G}_n)_{n\in\mathbb{N}}$ given by $\widetilde{G}_n = G_n^K$, acting on $\mathcal{X}^{K\times n}$ rowwise. That is, the first copy of the group acts on $(X_{1,1}, \ldots, X_{1,n})$, the second on $(X_{2,2}, \ldots, X_{2,n})$, etc. This action is sequential anytime that the action of $(G_n)_{n\in\mathbb{N}}$ is sequential on each of the $K$ data streams.

Based on the discussion above, a first idea to test for invariance under $\widetilde{G}_n$ is to create $K$ test martingales and combine them through multiplication. More specifically, we can treat each of the sequences $(X_{k,n})_{n\in\mathbb{N}}$, $k \in \{1, \ldots, K\}$ as a separate data stream and compute the corresponding statistics in (7.3), leading to $K$ sequences of uniformly distributed random variables $(R_{k,n})_{n\in\mathbb{N}}$. If, for all $n \in \mathbb{N}$ and $k \in \{1, \ldots, K\}$, $f_{k,n}$ is a density on $[0,1]$ then, by independence, the sequence $(M'_n)_{n\in\mathbb{N}}$ defined by $M'_n = \prod_{i=1}^n \prod_{k=1}^K f_{k,i}(R_{k,i})$ is a martingale under the null hypothesis. However, this martingale would not be able to detect alternatives under which the marginals are group invariant, but not independent. This stems from the fact that it only uses that the marginals are uniform under the null, while in fact a stronger claim is true: for each $n$, the joint distribution of $R_{k,n}$, $k \in \{1, \ldots, K\}$, is uniform on $[0,1]^K$. As a result

of this observation, one can choose any sequence of joint density (estimators) $f_1, f_2, \ldots$ on $[0,1]^K$ and create a test martingale by considering $M_n = \prod_{i=1}^{n} f_i(R_{1,i}, \ldots, R_{K,i})$.

In the case that $K = 2$ and $G_n = S(n)$, this is the procedure that was recently employed by Henzi and Law (2024). They discuss a specific choice of $f_n$, a histogram density estimator, that is able to detect departures from independence consistently under the stronger assumption that data are i.i.d. One of their key insights is that independence of the data streams not only implies joint uniformity of the sequential ranks in their setting, but that independence and joint uniformity are actually equivalent. This equivalence breaks down if one does not assume that $X_{k,1}, X_{k,2} \ldots$ are i.i.d. for all $k$. Finding conditions under which the independence of the streams and the joint uniformity of the rank distributions are equivalent so that a histogram density estimator might reliably detect independence in the more general setting is an interesting avenue of research.

## 7.6 Discussion

We have discussed how the theory of conformal prediction can be applied to test for symmetry of infinite sequences of data. Here we discuss two topics. First, the relationship to noninvariant conformal martingales. Second, whether smoothing is necessary when defining orbit ranks.

### 7.6.1 Noninvariant Conformal Martingales

Not all online compression models correspond to a compact-group invariant null hypothesis. An interesting example of this phenomenon is when the data are i.i.d. and exponentially distributed. This distribution is invariant under reflections in any 45° line (not necessarily through the origin), but these reflections do not define a compact group and therefore do not fit the setting discussed in this chapter. Nevertheless, the sum of data points is a sufficient statistic for the data, so this model can still be seen as an online compression model with the sum being the summary. More work is needed to find out whether conformal martingales are log-optimal against certain alternatives in such settings.

### 7.6.2 The Need for Smoothing

In situations when, conditionally on the orbit selector $\gamma_n(X^n)$, the conformity measure $\alpha^n(X^n)$ has a continuous distribution, the smoothing plays no role in (7.7). This is the

case for the rotations discussed in Section 7.5.2. In certain other scenarios, smoothing can be avoided as well. Indeed, one can always define nonsmoothed orbit ranks, in opposition to the smoothed ranks $R_n$ from Definition 7.7, by $\widetilde{R}_n := \mu_n(\{g \in G_n : (g\alpha^n)_n \leq \alpha_n\})$. Notice that this nonsmooth version satisfies $\widetilde{R}_n \leq R_n$. For a particular choice of increasing densities $f_1, f_2, \ldots,$ on $[0,1]$—in the sense that $u \mapsto f_i(u)$ is increasing—, we have that the process $\widetilde{M}_n := \prod_{i=1}^n f_i(\widetilde{R}_i)$ is bounded from above by the conformal martingale $M_n = \prod_{i=1}^n f_i(R_i)$. Such a choice of increasing $f_i$ is natural when high values of $R_i$ (or $\widetilde{R}_i$) are associated with departures from the null hypothesis. Then, any sequential test based on an upper threshold on $\widetilde{M}_n$ inherits the anytime-valid type-I error guarantees of $M_n$—exactly because $\widetilde{M}_n \leq M_n$. This was previously noted by Vovk et al. (2003). However, the process $\widetilde{M}_n$ may not be a martingale itself. Instead, a test martingale can sometimes directly be associated to $\widetilde{R}_n$. For instance, in the setting of Example 7.3 (testing exchangeability), the distribution of $\widetilde{R}_n$ under the null hypothesis is known—it is uniformly distributed on $\{1, \ldots, n\}$. Therefore, we can construct likelihood ratio processes for the sequence of nonsmoothed ranks. Even more, there are parametric alternatives under which the exact distributions of the nonsmoothed ranks can be computed. This is the case for Lehmann alternatives where, under the null, each $X_i$ is assumed to be sampled from some continuous distribution with c.d.f. $F_i(x) = F_0(x)$ for some fixed $F_0$; under the alternative, $F_i(x) = 1 - (1 - F_0(x))^{\theta_i}$ for some $\theta_i$. From Theorem 7.a.1 of Savage (1956) the distribution of $\widetilde{R}_i$ can be derived, so that the likelihood ratio process of $\widetilde{R}_i$ can be used for testing, thus avoiding external randomization.

# 8 | Testing With Group-Invariant Models

So far, in Chapters 4-7, the approach has been to construct log-optimal $e$-statistics against simple alternatives. As discussed in Section 2.2, this can be extended to composite alternatives by using the method of mixtures or using prequential plug-in estimators. In this chapter, we instead take a worst-case approach for the alternative and study worst-case-growth-rate-optimal (GROW) $e$-statistics. In particular, we consider GROW $e$-statistics for hypothesis testing problems between two group models. That is, in the previous chapter, we tested whether the distribution that generated the data was invariant under a certain group of transformations. Here, we instead assume that there is such an invariance and consider test statistics that use this assumption. To this end, it is known that under a mild condition on the action of the underlying group $G$ on the data, there exists a maximally invariant statistic. We show that among all $e$-statistics, invariant or not, the likelihood ratio of the maximally invariant statistic is GROW, both in an absolute and in a relative sense, and that an anytime-valid test can be based on it. The GROW $e$-statistic is equal to a Bayes factor with a right Haar prior on $G$. Our treatment avoids nonuniqueness issues that sometimes arise for such priors in Bayesian contexts. A crucial assumption on the group $G$ is its amenability, a well-known group-theoretical condition, which holds, for instance, in scale-location families. Our results also apply to finite-dimensional linear regression.

## 8.1 Introduction

We develop $e$-statistics and anytime-valid methods (Ramdas et al., 2023) for composite hypothesis testing problems where both null and alternative models remain unchanged under a group of transformations. Assume that the parameter of interest is a function $\delta = \delta(\theta)$ that is invariant under these transformations. Here, $\theta \in \Theta$ is the parameter of a probabilistic model $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ on an observation space $\mathcal{X}$. In the simplest case that we address, we are interested in testing whether the invariant parameter $\delta$ takes one of two values, that is,

$$\mathcal{H}_0 : \delta(\theta) = \delta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \delta(\theta) = \delta_1. \tag{8.1}$$

A prototypical example is the one-sample t-test where $\mathcal{P} = \{N(\mu, \sigma) : (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$ and the parameter of interest is the effect size $\delta(\mu, \sigma) = \mu/\sigma$, an invariant function of the model parameters under changes of scale. Other examples include tests about the correlation coefficient, which is invariant under affine transformations, and the variance of the principal components, an invariant under rotations around the origin (for more examples, see Berger et al., 1998). Data can be reduced by only considering its invariant component. Roughly speaking, by replacing the data $X^n = (X_1, \ldots, X_n)$ with an invariant statistic $M_n = m_n(X^n)$, one discards all information that is not relevant to the parameter $\delta$ (see the formal definitions in Section 8.2). For example, for the one-sample t-test, we can set $M_n$ equal to the t-statistic $M_{\mathcal{S},n} \propto \hat{\mu}_n/\hat{\sigma}_n$ but also to $M_n = (X_1/|X_1|, \ldots, X_n/|X_1|)$. Both are invariant functions under rescaling of all data points by the same factor that retain, as we will see, as much information as possible about the data.

By reducing the data through an invariant function, an invariant test can be obtained. Through the lens of the invariance-reduced data $M_n$, the composite hypotheses about $\theta$ simplify and (8.1) becomes simple-vs.-simple in terms of $\delta$. Indeed, because $M_n$ is an invariant function, its density depends only on $\delta$. Let us denote $p^{M_n}$ and $q^{M_n}$ the densities of $M_n$ under $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Both fixed-sample-size and sequential tests can be based on assessing the value of the likelihood ratio

$$T^{M_n} := \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))}. \tag{8.2}$$

However, it is not clear a priori whether this reduction affects the optimality of the resulting tests. In other words, does the family of invariant tests, i.e. tests that can

be written as a function of (8.2), contain the best ones?

For fixed-sample size tests, with power as a criterion, the answer is positive: a celebrated theorem of Hunt and Stein shows that, when looking for a test that has max-min power, no loss is incurred by looking only among group-invariant tests (Lehmann and Romano, 2005, Section 8.5). In classic sequential testing, the principle of invariance has been used (Cox, 1952; Hall et al., 1965), but no optimality results are known. In this chapter we address this question and provide an analogue of the Hunt-Stein theorem within the setting of *anytime-valid* tests. We replace power by GROW (see again below), the natural optimality criterion in this context, and we show that, under some regularity conditions, $T^{M_n}$ is the *optimal e-statistic* for testing (8.1).

The $e$-statistic (also known as $e$-variable or $e$-value) is a central concept within the theory of *anytime-valid* testing (Vovk and Wang, 2021; Shafer, 2021; Grünwald et al., 2024; Ramdas et al., 2020), interest in which has recently exploded —Ramdas et al. (2023) provide a comprehensive overview. The main objective that is achieved by testing with $e$-statistics is finite-sample type-I error control in two common situations: when experiments are optionally stopped—sampling is stopped at a data-dependent sample size—, and when aggregating the evidence of interdependent experiments. In the latter case, called optional continuation (Grünwald et al., 2024, GHK from now on), the decision to start a new experiment may depend in unknowable ways on the outcome of previous experiments (Vovk and Wang, 2021). We will use the qualifier anytime-valid as an umbrella term that covers both optional stopping and continuation, and study invariance reductions for anytime-valid tests; we stress that, as elaborated in Appendix F.3, anytime-valid testing, while taking place in a sequential setting, is different from classical, Wald-style sequential testing, in which power *is* meaningful. While $e$-statistics have also found applications beyond anytime-validity, for example in multiple testing (Wang and Ramdas, 2022; Ren and Barber, 2024) and when not just the stopping time but also the relevant loss function or significance level may depend in unknowable ways on the data itself (decision-theoretic robustness, Grünwald (2023)), our results focus on optimality in the anytime-valid context. In this context, power is not a meaningful measure of optimality (see Section 8.2.4). A natural replacement of power is the *GROW* criterion, which stands for *growth rate optimal in the worst case*. Informally, among all $e$-statistics, those that are GROW accumulate evidence against the null as fast as possible (in terms of sample size). Some other authors refer to GROW as "maximal $e$-power" (Zhang et al., 2024) or as "optimizing the Kelly criterion" (Ramdas et al., 2023). Sometimes, it is beneficial to consider instead the growth rate relative to an oracle that knows the distribution of the data, not in

absolute terms. $e$-statistics that are optimal in this relative sense are called relatively GROW. Especially this relative criterion (or closely related variations of it) has often been used to design $e$-statistics; recent examples include the work of Henzi et al. (2023) and Waudby-Smith and Ramdas (2024)—see Ramdas et al. (2023) for a more comprehensive list.

Under regularity conditions, the GROW $e$-statistic can be found by minimizing the Kullback-Leibler (KL) divergence between the convex hull of the null and alternative models (GHK). Indeed, the likelihood ratio of the distributions that achieve this minimum KL is a GROW $e$-statistic, and the GROW $e$-statistic is then essentially unique in the sense that any two GROW $e$-statistics agree almost surely under all distributions in $\mathcal{H}_0$ and $\mathcal{H}_1$. As such, $e$-statistics can be seen as composite generalizations of likelihood ratios. In particular, any likelihood ratio of a statistic that has the same distribution under all elements of the null and another single distribution under the alternative is an $e$-statistic (GHK). As a consequence, for any invariant function of the data $M_n$, the likelihood ratio statistic $T^{M_n}$ from (8.2) is an $e$-statistic for the testing problem (8.1). As our main contribution, we show that, under regularity conditions, if $M_n$ is a maximally invariant statistic of the data or of a sufficient statistic for $\theta$, then the KL divergence between $q^{M_n}$ and $p^{M_n}$ equals the minimum KL divergence between the convex hulls of the null and alternative models. By the result of GHK mentioned above that links KL minimization to GROW $e$-statistics, $T^{M_n}$ is GROW. A maximally invariant statistic, informally, loses as little information as possible about the data while being invariant. For example, with $V_n = (X_1/|X_1|, \ldots, X_n/|X_1|)$, setting $M_n := V_n$ as in the beginning of the introduction for the t-test gives a maximal invariant, while using $M'_n := V_{n-1}$ gives an invariant that is not maximal. Furthermore, the t-statistic is not maximally invariant for the raw data, but it is a maximally invariant function of $(\hat{\mu}_n, \hat{\sigma}_n)$ which is a sufficient statistic. As we will see, the likelihood ratio statistic $T^{M_{\mathcal{S},n}}$, where $M_{\mathcal{S},n}$ is the t-statistic and $T^{M_n}$ with $M_n = V_n$ coincide (see Appendix F.1), and it will follow from our results that both are GROW.

Additionally, we show that any GROW $e$-statistic is also relatively GROW in the group-invariant setting. Hence, $T^{M_n}$ is relatively GROW as well. This growth rate optimality motivates the use of $T^{M_n}$ in optional continuation settings. As a further contribution, we show that every time that $M_n$ is a maximal invariant, the sequence $T = (T^{M_n})_{n\in\mathbb{N}}$ is a nonnegative martingale. This extends its use and optimality to optional stopping.

The rest of this chapter is organized as follows. In Section 8.2 we introduce notation, formally lay the groundwork for group-invariant testing, review $e$-statistics and

their optimality criteria, and discuss related work. Section 8.3 is devoted to stating our main results: showing that the $e$-statistic $T^{M_n}$ for a maximally invariant function $M_n = m_n(X^n)$ is both GROW and relatively GROW, proving that $T^{M_n}$ is suited for both optional continuation and optional stopping, and extending these results to composite hypotheses, i.e. sets $\Delta_1$ and $\Delta_0$ of $\delta$'s, both with and without a prior distribution imposed on them (for general discussion on how to choose $\delta_j, \Delta_j$ or such priors, we refer to GHK, Section 6). Next, in Section 8.4, we apply our results to two examples. We end this chapter with Section 8.5, where we discuss further the technical conditions that our results require and related work on group-invariant testing. Section 8.6 contains all proofs that were omitted earlier.

## 8.2 Preparation for the Main Results

This section is structured as follows. We first introduce notation. Then, in section 8.2.2, we introduce the formal setup and our running example, the t-test. In Section 8.2.3, we define $e$-statistics, our main objects of study, and in Section 8.2.4 we define our optimality criteria. Finally, Section 8.2.5 highlights previous work.

### 8.2.1 Notation

All spaces that we consider are assumed to be topological spaces with an additional measurable structure given by the respective $\sigma$-algebra of Borel sets. We write $X$ for a random variable taking values in the observation space $\mathcal{X}$, and $X^n := (X_1, \ldots, X_n)$ for $n$ independent copies of $X$ under the distributions that are to be considered.

Statistics of the data $X^n$ are denoted as $T = t(X^n)$, where $t$ is a measurable map $t : \mathcal{X}^n \to \mathcal{T}_n$. We use letters $\mathbf{P}$ and $\mathbf{Q}$ to refer to distributions of $X$. For a statistic $T = t(X^n)$, we write $\mathbf{P}^T$ for the image measure of $\mathbf{P}$ under $t$, that is, $\mathbf{P}^T\{T \in A\} = \mathbf{P}\{t(X^n) \in A\}$ for measurable $A \subseteq \mathcal{T}_n$ (note that we may think of $T$ as a random variable on the space $\mathcal{X}^n$). When writing conditional expectations, we write $\mathbf{E}^{\mathbf{P}}[f(X)|Y]$, and $\mathbf{P}^{X|y}$ for the conditional distribution of $X$ given $Y = y$. We only deal with situations where such conditional distributions exist. If we are considering a set of distributions parameterized in terms of a parameter space $\Theta$, we write $\mathbf{E}^{\mathbf{P}}_{\theta}[f(X)]$ rather than $\mathbf{E}^{\mathbf{P}_\theta}[f(X)]$ for the sake of readability. Furthermore, for a prior distribution $\mathbf{\Pi}$ on $\Theta$, 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 $B \subseteq \mathcal{X}$. For the posterior distribution of $\theta$ given $X$ we write $\mathbf{\Pi}^{\theta|X}$. The Kullback-Leibler (KL)

divergence between $\mathbf{Q}$ and $\mathbf{P}$ is denoted by $\mathrm{KL}(\mathbf{Q},\mathbf{P}) = \mathbf{E}^{\mathbf{Q}}[\ln(\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P})]$ (Kullback and Leibler, 1951) whenever the Radon–Nikodym derivative $\mathrm{d}\mathbf{Q}/\mathrm{d}\mathbf{P}$ exists. Given two subsets $H, K$ of a group $G$ we write $HK = \{hk : h \in H, k \in K\}$ for the set of all products between elements of $H$ and elements of $K$. Similarly, for $g \in G$ and $K \subseteq G$, we write $gK = \{gk : k \in K\}$ for the translation of $K$ by $g$, and $K^{-1} = \{k^{-1} : k \in K\}$ for the set of inverses of $K$. We say that $K$ is symmetric if $K = K^{-1}$. If $G$ acts on $\mathcal{X}$, then we denote the action of $G$ on $\mathcal{X}$ by $(g, x) \mapsto gx$ for $g \in G$ and $x \in \mathcal{X}$, and extend the action to $\mathcal{X}^n$ component-wise; that is, $(g, x^n) \mapsto gx^n := (gx_1, \ldots, gx_n)$ for $g \in G$ and $x^n \in \mathcal{X}^n$. We write $gB = \{gb : b \in B\}$ for the left translate of a subset $B \subseteq \mathcal{X}$ by $g$. If $G$ acts on $\Theta$, the notation is completely analogous.

### 8.2.2 Group Invariance

We consider a group $G$ that acts freely on both the observation space $\mathcal{X}$ and the parameter space $\Theta$. Recall that $G$ acts freely on a set $\mathcal{Z}$ if anytime that $gz = z$ for some $g \in G$ and $z \in \mathcal{Z}$, then $g$ is the identity element of the group $G$. A probabilistic model $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ on $\mathcal{X}$ is said to be invariant under the action of $G$ if the distribution $\mathbf{P}_\theta$ satisfies

$$\mathbf{P}_\theta\{X \in B\} = \mathbf{P}_{g\theta}\{X \in gB\} \tag{8.3}$$

for any $g \in G$, any measurable $B \subseteq \mathcal{X}$, and any $\theta \in \Theta$. Furthermore, a function $m(x)$ is said to be invariant under the action of $G$ if $m(gx) = m(x)$ for all $x \in \mathcal{X}$ and all $g \in G$; in other words, $m$ is constant on the orbits of $G$. Moreover, $m$ is said to be maximally invariant if it indexes the orbits of $\mathcal{X}$ under the action of $G$; that is, $m(x) = m(x')$ for $x, x' \in \mathcal{X}$ if and only if there exists a $g \in G$ such that $x = gx'$. A statistic is called (maximally) invariant if the corresponding function is. These definitions are completely analogous for functions defined on $\Theta$. In particular, we study situations where the parameter of interest $\delta = \delta(\theta)$ is a maximally invariant function of the parameter $\theta$. We then say that $\delta$ is a maximally invariant parameter.

We now reparametrize the problem described in (8.1) using the group $G$. Using that the action of the group on the parameter space is free, we can reparametrize each orbit in $\Theta/G$ with $G$. Indeed, we can pick an arbitrary but fixed element in the orbit $\theta_0 \in \delta^{-1}(\delta_0)$ and, for any other element $\theta \in \delta^{-1}(\delta_0)$, we can identify $\theta$ with the group element $g(\theta) \in G$ that transports $\theta_0$ to $\theta$, that is, such that $g(\theta)\theta_0 = \theta$. Hence, with a slight abuse of notation, we can identify $\theta \in \delta^{-1}(\delta_0)$ with $g = g(\theta) \in G$ and identify $\mathbf{P}_\theta = \mathbf{P}_{g(\theta)\theta_0}$ with $\mathbf{P}_g$. Define $\mathbf{Q}_g$ using the same construction in the

alternative model by an analogous choice of $\theta_1 \in \delta^{-1}(\delta_1)$. The starting problem (8.1) may now be rewritten in the form

$$\mathcal{H}_0 : X^n \sim \mathbf{P}_g,\ g \in G, \text{ vs. } \mathcal{H}_1 : X^n \sim \mathbf{Q}_g,\ g \in G. \tag{8.4}$$

To make notation more succinct, we use $\mathcal{Q} = \{\mathbf{Q}_g\}_{g\in G}$ to denote the alternative hypothesis and $\mathcal{P} = \{\mathbf{P}_g\}_{g\in G}$ for the null. As will follow from our discussion, our results are insensitive to the choices of $\theta_0 \in \delta^{-1}(\delta_0)$ and $\theta_1 \in \delta^{-1}(\delta_1)$.

As mentioned in the introduction, tests for (8.4) are classically based on the likelihood ratio $T^{M_n}$ of a maximally invariant statistic $M_n = m_n(X^n)$, as in (8.2). While the distribution of $M_n$ might be unknown, it is well-known that its likelihood ratio can be computed by integration over the group $G$ whenever the following three conditions—which will be explained in brief—hold: (1) the action is continuous and proper, (2) $G$ is a $\sigma$-compact locally compact topological group, and (3) for all $g$, $\mathbf{P}_g$ and $\mathbf{Q}_g$ are dominated by a relatively left invariant measure $\nu$. In (1), an action is proper if 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 any compact set is compact. In (2), a topological group is a group equipped with a topology, such that the group operation, seen as a function $G \times G \to G$, is continuous. Under (3), we assume the existence of densities $p_g$ and $q_g$ for $\mathbf{P}_g$ and $\mathbf{Q}_g$, respectively, with respect to $\nu$ for each $g \in G$. Furthermore, since $G$ is assumed to be locally compact, there exists a measure $\rho$ on $G$ that is right invariant (see Bourbaki, 2004, VII,§1,n° 2). This means that for any $g \in G$ and any $B \subseteq G$ that is measurable, it holds that $\rho\{Bg\} = \rho\{B\}$. The measure $\rho$, called the right Haar measure, is unique up to a multiplicative factor and is finite if and only if $G$ is compact. Using disintegration-of-measure results from Bourbaki (2004, VIII.27), Andersson (1982) shows that $T^{M_n}$ can be computed as

$$T^{M_n} = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))} = \frac{\int_G q_g(X^n)\mathrm{d}\rho(g)}{\int_G p_g(X^n)\mathrm{d}\rho(g)}, \tag{8.5}$$

This is known as Wijsman's representation theorem (for extended statement and discussion, see Eaton, 1989, Theorem 5.9). Note that (8.5) implies that the likelihood ratio $T^{M_n}$ is independent of the choice of maximal invariant $M_n$. Remarkably, work by Stein, reported by Hall et al. (1965), shows that it does not even matter whether we consider a maximal invariant of the original data, or whether we first reduce the data through sufficiency and then consider a maximal invariant of the sufficient statistic. In the t-test example, this shows that the likelihood ratio of the t-statistic is equal

to that of $M_n$ as in the start of the introduction. We further discuss this result in Appendix F.1.

Finally, the classic theorem of Hunt and Stein (Lehmann and Romano, 2005, Section 8.5) shows that, under some regularity conditions, when looking for a test that is max-min optimal in the sense of power, it is sufficient to look among invariant tests, i.e. tests that can be written as a function of $T^{M_n}$ as in (8.2). One of the crucial assumptions underlying their result is the *amenability* of $G$. 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 $g \in G$

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

Amenable groups have been thoroughly studied (Paterson, 1988) and include, among others, all finite, compact, commutative, and solvable groups. The easiest example of a nonamenable group is the free group in two elements and any group containing it. Another prominent example of a nonamenable group is that of invertible $d \times d$ matrices with matrix multiplication.

**Example 8.1** (t-test under Gaussian assumptions)**.** Consider an i.i.d. sample $X^n = (X_1, \ldots, X_n)$ of size $n \in \mathbb{N}$ from an unknown Gaussian distribution $N(\mu, \sigma)$, with $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. In the 1-sample t-test, we are interested in testing whether $\mu/\sigma = \delta_0$ or $\mu/\sigma = \delta_1$ for some $\delta_0, \delta_1 \in \mathbb{R}$. For $c \in \mathbb{R}^+$, we have that $cX \sim N(c\mu, c\sigma)$, so it follows that the Gaussian model is invariant under scale transformations. The corresponding group is $G = (\mathbb{R}^+, \cdot\,)$, which acts on $\mathcal{X}^n$ by component-wise multiplication and on $\Theta$ by $(c, (\mu, \sigma)) \mapsto (c\mu, c\sigma)$ for each $c \in G$ and $(\mu, \sigma) \in \Theta$. The parameter of interest, $\delta = \mu/\sigma$, is scale-invariant and indexes the orbits of the action of $G$ on $\Theta$. A maximally invariant statistic is $M_n := (X_1/|X_1|, \ldots, X_n/|X_1|)$. The right Haar measure $\rho$ on $G$ is given by $\mathrm{d}\rho(\sigma) = \mathrm{d}\sigma/\sigma$, so that the likelihood ratio of $M_n$ can be expressed, as in (8.5), by

$$T^{M_n} = \frac{\int_{\sigma>0} \frac{1}{\sigma^n} \exp\left(-\frac{n}{2}\left[\left(\frac{\bar{X}_n}{\sigma} - \delta_1\right)^2 + \frac{1}{n}\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2\right]\right) \frac{\mathrm{d}\sigma}{\sigma}}{\int_{\sigma>0} \frac{1}{\sigma^n} \exp\left(-\frac{n}{2}\left[\left(\frac{\bar{X}_n}{\sigma} - \delta_0\right)^2 + \frac{1}{n}\sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2\right]\right) \frac{\mathrm{d}\sigma}{\sigma}}, \tag{8.6}$$

where $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. The results by Stein, discussed in Appendix F.1, show that the likelihood ratio of the t-statistic, i.e. $M_{\mathcal{S},n} \propto \hat{\mu}_n/\hat{\sigma}_n$, is equal to the expression obtained in (8.6).

### 8.2.3 The Family of E-Statistics

We now define *e*-statistics, our measure of evidence against the null hypothesis. The family of *e*-statistics comprises all nonnegative real statistics whose expected value is bounded by one under all elements of the null, that is, all statistics $T_n = t_n(X^n)$ such that $T_n \geq 0$ and

$$\sup_{g \in G} \mathbf{E}_g^{\mathbf{P}}[T_n] \leq 1. \tag{8.7}$$

An example of an *e*-statistic is the likelihood ratio statistic in any simple-vs-simple testing problem (see e.g. GHK, Section 1 or Ramdas et al. (2023)). In particular, (8.2) is an *e*-statistic for the hypotheses in (8.4). *e*-statistics are appropriate in optional continuation contexts because of the following two properties that are consequences of (8.7).

1. The type-I error of the test that rejects the null hypothesis anytime that $T_n \geq 1/\alpha$ is smaller than $\alpha$, a consequence of (8.7) and Markov's inequality.

2. Suppose that $X^n$ and $X^m$ are the independent outcomes of two subsequent experiments. Let $T_n = t_n(X^n)$ be an *e*-statistic for $X^n$ and let $\{T_{m,\varphi} : \varphi \in \Phi\}$ be a family of *e*-statistics for $X^m$ indexed by some set $\Phi$. Suppose further that, after observing the first sample $X^n$, the specific $T_{m,\varphi}$ used to measure evidence for the second sample is chosen as a function of $X^n$, that is, we use $T_{m,\hat{\varphi}}$ where $\hat{\varphi} = \hat{\varphi}(X^n)$ is some function of $X^n$. Then $T_{n+m} := T_n T_{m,\hat{\varphi}}$ is also an *e*-statistic, irrespective of the definition of $\hat{\varphi}$. In particular, this includes the scenario where we only continue to the second experiment if a certain outcome is observed in the first one. Indeed, $\Phi$ may contain a special value $\mathbf{1}$ so that $t_m(X^m; \mathbf{1}) = 1$ is constant, irrespective of $X^m$. Then, $T_{n+m} = T_n$ every time that $\hat{\varphi} = \mathbf{1}$.

Together, these two properties imply that the test that rejects the null if $T_{n+m} \geq 1/\alpha$ has type-I error bounded by $\alpha$, no matter the definition of $\hat{\varphi}$. Such type-I error guarantees are essentially impossible using p-values (GHK, Section 1.3). Some—not all—types of *e*-statistics can additionally be used in two related settings: (a) *optional stopping*, when there is a single sequence of data $X_1, X_2, \ldots$ and we want to do a test with type-I error guarantees based on all data seen so far, irrespective of when we stop; and (b) optional continuation as in 2. above, but with individual *e*-statistics whose sample size is itself not fixed but determined by some stopping rule. As is well-known, for both (a) and (b) it is sufficient that $(T_n)_{n \in \mathbb{N}}$ is a nonnegative martingale with respect to some filtration $\mathcal{F}$ (see e.g. Ramdas et al., 2023, or GHK). The first

part follows from Ville's inequality for nonnegative martingales: the probability that there will *ever* be a sample size $n$ at which $T_n \geq 1/\alpha$ is bounded by $\alpha$. We thus have type-I error control under optional stopping, which takes care of (a) above. The optional stopping theorem implies that for every stopping time $\tau$ adapted to $\mathcal{F}$, $T_\tau$ is also an $e$-statistic, taking care of (b). For completeness, we provide more details in Appendix F.3, including a subtlety regarding (b): while they seem unlikely to arise in practice, there do exist stopping times $\tau'$ relative to the data that are not stopping times relative to $\mathcal{F}$. We show an example where $T_{\tau'}$ is not an $e$-statistic and (b) breaks.

### 8.2.4 Optimality Criteria for E-Statistics

The standard optimality criterion for hypothesis tests satisfying a certain type-I error guarantee is worst-case power maximization for a fixed-sample-size or, with classic sequential tests, for a fixed stopping rule. This criterion cannot be used when the stopping rule is unknown because knowledge of the stopping rule is required by the definition of power. Additionally, an $e$-statistic that optimizes power at fixed stopping time will take the value zero with positive probability, making it useless for optional continuation by multiplication. A more sensible criterion for $e$-statistics under optional continuation is growth rate optimality in the worst case (GHK). 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

$$T_n \mapsto \inf_{g \in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n] \tag{8.8}$$

over all $e$-statistics. The following theorem, stated in our notation for group-invariant problems, shows that in most cases the GROW $e$-statistic takes the form of a particular Bayes factor.

**Theorem 8.1** (GHK Theorem 1 in Section 4.3)**.** *Suppose that there exists a statistic $V_n = v_n(X^n)$ such that*

$$\inf_{\mathbf{\Pi}_0, \mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^g \mathbf{Q}_g, \mathbf{\Pi}_0^g \mathbf{P}_g) = \min_{\mathbf{\Pi}_0, \mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^g \mathbf{Q}_g^{V_n}, \mathbf{\Pi}_0^g \mathbf{P}_g^{V_n}) < \infty, \tag{8.9}$$

*where $\mathbf{\Pi}_0$ and $\mathbf{\Pi}_1$ are probability distributions on $G$. Let $\mathbf{\Pi}_0^\star$ and $\mathbf{\Pi}_1^\star$ be probability*

*distributions that achieve the minimum on the right hand side. Then*

$$\max_{T_n \ e\text{-stat.}} \inf_{g\in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n] = \mathrm{KL}(\mathbf{\Pi}_1^{\star g}\mathbf{Q}_g^{V_n}, \mathbf{\Pi}_0^{\star g}\mathbf{P}_g^{V_n}),$$

*and the maximum on the left is achieved, essentially uniquely, by $T_n^*$ as given by*

$$T_n^* := \frac{\int q_g^{V_n}(v_n(X^n))\mathrm{d}\mathbf{\Pi}_1^\star(g)}{\int p_g^{V_n}(v_n(X^n))\mathrm{d}\mathbf{\Pi}_0^\star(g)}.$$

*Here 'essentially uniquely' means that any other e-statistic achieving the maximum must coincide with $T_n^*$ almost surely, under all $\mathbf{P}_g$ and $\mathbf{Q}_g$ with $g \in G$. In words, the e-statistic $T_n^*$ is GROW for testing $\{\mathbf{P}_g\}_{g\in G}$ against $\{\mathbf{Q}_g\}_{g\in G}$.*

The statistic $V_n$ may be any measurable function taking values in any set $\mathcal{V}_n$ (equipped with its corresponding $\sigma$-algebra), but in all our examples we can take $\mathcal{V}_n = \mathbb{R}^m$ for some $m \leq n$. By allowing $V_n \neq X^n$, the theorem also covers cases in which the infimum on the left in (8.9) is not achieved. This might be the case when the group $G$ is not compact, as in the t-test example. Corollary 8.3 in the next section, which gives the GROW *e*-statistic when $G$ is possibly noncompact, uses crucially this feature of Theorem 8.1.

Given their worst-case nature, GROW *e*-statistic, while appropriate in some scenarios (e.g. testing exponential families with given minimum effect sizes and no nuisance parameters), are too conservative in others (GHK). GHK propose, for those cases, to maximize a relative form of (8.8), leading to less conservative *e*-statistics. 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,

$$T_n \mapsto \inf_{g\in G} \left\{ \mathbf{E}_g^{\mathbf{Q}}[\ln T_n] - \sup_{T_n' \ e\text{-stat.}} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n'] \right\}. \tag{8.10}$$

As we will see and contrary to the general case, in the group-invariant setting, any GROW *e*-statistic is also relatively GROW. Hence, both criteria coincide and the differences that have been observed between them (raising the sometimes difficult question: which one to choose?) are not a concern for our purposes (Ramdas et al., 2023).

### 8.2.5 Previous and Related Work

Group-invariant problems have a long tradition in statistics. They have been studied both for fixed-sample-size experiments Eaton (1989); Lehmann and Romano (2005) and classical, Wald-type sequential experiments (Rushton, 1950; Cox, 1952). For fixed-sample-size tests, our main result can be viewed, to some extent, as an anytime-valid analogue of the Hunt-Stein theorem. The proof techniques that are needed for our result are, however, distinct. At the core of the proof of the Hunt-Stein theorem lies 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 out 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 the optimal $e$-statistics that we consider (see Section 8.2.4). As for sequential tests with group invariance, most previous work (including the pioneering Rushton (1950); Cox (1952) and in fact, as far as we could ascertain, all work predating Robbins (1970)) dealt, like Wald's original SPRT, with a priori fixed stopping rules and is not directly comparable to our anytime-valid work (see Appendix F.3 for elaboration of this point). Notable exceptions are the works of Robbins (1970) and Lai (1976), who do consider what we now call anytime validity. Lai (1976) also used the expression in (8.6) for the t-test, which, in our terminology, is using the fact that it gives an $e$-statistic. However, our main concern, optimality of $e$-statistics, has not been explored in this context.

Related ideas can also be found in the Bayesian literature, where group-invariant inference with right Haar priors has been 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 that the underlying group (and hence the right Haar prior) is not uniquely defined in some situations, and that different choices lead to different conclusions (Sun and Berger, 2007; Berger and Sun, 2008). Interestingly, as we briefly discuss in Section 8.5 and at length in Appendix F.2, this issue cannot arise in our setting. In the same section, we point out similarities and the main difference to the information-theoretic 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. In a nutshell, despite some similarities, the precise min-max result that they prove is not comparable to the results presented here.

## 8.3 Main Results

In this section, we state the main results of this chapter. In Section 8.3.1 we show that the likelihood ratio $T^{M_n}$ for a maximal invariant $M_n$ is simultaneously GROW and relatively GROW. Next, in Section 8.3.2, we show that $T^{M_n}$ can be used to build an anytime-valid test. Finally, in Section 8.3.3 we extend these results to the case that the hypotheses remain composite after reduction by invariance.

### 8.3.1 GROW for simple invariant hypotheses

In order to build intuition, we first demonstrate our line of reasoning using the very special case of finite groups. So, assume for now that $G$ is a finite group, for instance, a group of permutations. Since the uniform probability distribution $\mathbf{\Pi}_{\mathrm{U}(G)}$ on $G$ is right invariant, the right Haar measure $\rho$ coincides with $\mathbf{\Pi}_{\mathrm{U}(G)}$ up to scaling. By Wijsman's representation theorem (8.5), the likelihood ratio for any maximal invariant $M_n = m_n(X^n)$ can be written as

$$T^{M_n} = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))} = \frac{\frac{1}{|G|}\sum_{g\in G} q_g(X^n)}{\frac{1}{|G|}\sum_{g\in G} p_g(X^n)}. \tag{8.11}$$

Furthermore, Theorem 8.1 above takes a simple form for finite parameter spaces, as is the case here, namely

$$\max_{T_n \text{ e-stat.}} \min_{g\in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n] = \min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^g \mathbf{Q}_g, \mathbf{\Pi}_0^g \mathbf{P}_g), \tag{8.12}$$

where the minimum on the right hand side is taken over all pairs of distributions on $G$. We now employ the information processing inequality (Cover and Thomas, 1991, Section 2.8) which says that KL divergence decreases when taking functions of the data (i.e. if $\mathbf{A}$ and $\mathbf{B}$ are distributions for $X$ and $U = u(X)$, then $\mathrm{KL}(\mathbf{A}\|\mathbf{B}) \geq \mathrm{KL}(\mathbf{A}^U\|\mathbf{B}^U)$). In our setting, the information processing equality implies that for any pair $(\mathbf{\Pi}_0, \mathbf{\Pi}_1)$ of probability distributions on $G$,

$$\mathrm{KL}(\mathbf{\Pi}_1^g \mathbf{Q}_g, \mathbf{\Pi}_0^g \mathbf{P}_g) \geq \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}). \tag{8.13}$$

This lower bound can be rewritten as $\mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}) = \mathrm{KL}(\mathbf{\Pi}_{\mathrm{U}(G)}^g \mathbf{Q}_g, \mathbf{\Pi}_{\mathrm{U}(G)}^g \mathbf{P}_g)$ because of the second equality in (8.11). Therefore, the minimum KL on the right hand side of (8.12) is achieved for the particular choice of two uniform priors on $G$.

Finally, we have that $\mathbf{E}_g^{\mathbf{Q}}[\ln T^{M_n}] = \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n})$ for all $g \in G$. Putting everything together

$$\max_{T_n \ e\text{-stat.}} \min_{g \in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n] = \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}) = \min_{g \in G} \mathbf{E}_g^{\mathbf{Q}}[\ln T^{M_n}];$$

in other words, $T^{M_n}$ is a GROW $e$-statistic. A natural question is whether this same reasoning can be reproduced for infinite groups. If the Haar measure $\rho$ could always be chosen to be a probability measure, we could replace $\mathbf{\Pi}_{\mathrm{U}(G)}$ by $\rho$ everywhere in the reasoning above and conclude that $T^{M_n}$ is GROW in general. However, $\rho$ is finite if and only if $G$ is compact (see e.g. 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 8.1). The main technical contribution of this chapter is the extension of the above optimality result to amenable groups (see Section 8.2.2). Setting technical details aside, the core of the proof of the main Theorem 8.2 below is replacing the Haar measure above by a sequence of almost-right-invariant probability measures and showing that the KL converges to its infimum. Our arguments require the following additional assumptions.

**Assumption 8.1.** Let $G$ be a topological group acting on a topological space $\mathcal{X}^n$, both equipped with their Borel $\sigma$-algebra. 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 $\mu$ on $\mathcal{X}^n$ that is relatively left invariant with some multiplier $\chi$—$\mu\{gB\} = \chi(g)\mu\{B\}$ for any measurable set $B \subseteq \mathcal{X}^n$ and $g \in G$. All densities have a single common support.

Assumption 8.1 holds in most cases of interest for the purpose of parametric inference; some examples where it holds are given in Section 8.4. The topological assumptions on $G$ and $\mathcal{X}$ have two purposes. The first is to ensure that Wijsman's representation theorem (8.5) holds. Though (8.5) requires slightly weaker assumptions than those presented here, see Section 8.2.2, the strengthened conditions are needed for the second purpose: to ensure that the observation space $\mathcal{X}^n$ can be put

in bijective and bimeasurable[1] correspondence with a subset of $G \times \mathcal{X}^n/G$, where the group $G$ acts naturally by multiplication on the first component (Bondar, 1976). This will be used extensively in the proofs given in Section 8.6. With these assumptions, everything is in place to state the main results of this chapter.

**Theorem 8.2.** *Let $M_n = m_n(X^n)$ be a maximally invariant statistic under the action of the group $G$ on $\mathcal{X}^n$. Assume that $G$ is amenable, that Assumption 8.1 holds, and that there is $\varepsilon > 0$ such that*

$$\mathbf{E}_1^{\mathbf{Q}}\left[\left|\ln \frac{q_1(X^n)}{p_1(X^n)}\right|^{1+\varepsilon}\right], \mathbf{E}^{\mathbf{Q}^{M_n}}\left[\left|\ln \frac{q^{M_n}(M_n)}{p^{M_n}(M_n)}\right|^{1+\varepsilon}\right] < \infty, \tag{8.14}$$

*where the subindex 1 refers to the unit element of $G$. 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 over all pairs $(\mathbf{\Pi}_0, \mathbf{\Pi}_1)$ of probability distributions on $G$.*

**Corollary 8.3.** *Under the assumptions of Theorem 8.2, a GROW e-statistic for testing $\mathcal{H}_1$ against $\mathcal{H}_0$ as in (8.4) is given by the likelihood ratio of any maximally invariant statistic $M_n = m_n(X^n)$, i.e.*

$$T^{M_n} = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))}.$$

Corollary 8.3 follows from the combination of Theorem 8.2 with Theorem 8.1. The results are stated in terms of the likelihood ratio of any maximal invariant for the original data. However, as mentioned briefly in Section 8.2.2 and in detail in Appendix F.1, one can use instead any maximal invariant for a sufficient statistic of the original data, rather than for the data itself. The resulting likelihood ratio is identical and the optimality results therefore remain valid. Next, we show that in the group-invariant setting, any statistic that is GROW is also relatively GROW, meaning that any $e$-statistic that maximizes (8.8) also maximizes (8.10). This is not true in general; the result relies crucially on the invariance of the models. For example, for contingency tables, the two $e$-statistics are vastly different (Turner et al., 2024).

**Theorem 8.4.** *Suppose that Part 3 of Assumption 8.1 is satisfied and that, for each*

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

*$g \in G$, there exists $h \in G$ such that $\mathrm{KL}(\mathbf{Q}_g, \mathbf{P}_h)$ is finite. Then the map defined by*

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

*is constant. Consequently, any maximizer of* (8.8) *also maximizes* (8.10)*, that is, an e-statistic is GROW if and only if it is relatively GROW for the hypothesis testing problem* (8.4)*.*

**Corollary 8.5.** *$T^{M_n}$ from Corollary 8.3 is not only GROW, it is also relatively GROW.*

**Example 8.2** (continues=ex:t-test)**.** It is known that the group $G = (\mathbb{R}^+, \cdot)$ of the t-test is amenable—the sequence of probability distributions $(\mathrm{Uniform}([-n, n]))_{n\in\mathbb{N}}$ is almost right invariant. It is readily verified that Assumption 8.1 and condition (8.14) are also satisfied. Hence, Corollary 8.3 implies that the likelihood ratio for the t-statistic, given in (8.6), is a GROW *e*-statistic. Moreover, it follows from Corollary 8.5 that it is also relatively GROW.

### 8.3.2 Anytime-Validity

As discussed in Section 8.2.3, any *e*-statistic can be used in the context of optional continuation with fixed sample sizes, but not all *e*-statistics are suitable for optional stopping and optional continuation with data-dependent sample sizes. A sufficient condition that allows us to engage in these two additional uses is that the sequence of *e*-statistics is a nonnegative martingale. We now show that this is the case for the sequence $(T^{M_n})_{n\in\mathbb{N}}$.

**Proposition 8.6.** *If $(M_n)_{n\in\mathbb{N}}$ is a sequence of maximally invariant statistics $M_n = m_n(X^n)$ for the action of $G$ on $\mathcal{X}^n$, then the process $(T^{M_n})_{n\in\mathbb{N}}$ is a nonnegative martingale with respect to the filtration $(\sigma(M_1, \ldots, M_n))_{n\in\mathbb{N}}$ under any of the elements of the null hypothesis.*

In particular, Proposition 8.6 implies that under every stopping time $\tau$ defined relative to the filtration induced by $(M_n)_{n\in\mathbb{N}}$, $T^{M_\tau}$ is itself an *e*-statistic; see Appendix F.3 for the (standard) proof. There is an interesting subtlety here however: if $\tau'$ is a stopping time relative to the filtration induced by $(X_n)_{n\in\mathbb{N}}$ but not relative to the coarser filtration induced by $(M_n)_{n\in\mathbb{N}}$, then $T^{M_{\tau'}}$ is not necessarily an *e*-statistic anymore. Thus, with such $T^{M_{\tau'}}$, we cannot engage in optional continuation. This is generally not a problem, since most stopping times encountered in practice are stopping times

relative to the filtration induced by $(M_n)_{n\in\mathbb{N}}$. This includes the aggressive stopping time "stop at the smallest $n$ at which $T^{M_n} \geq 1/\alpha$". However, in Appendix F.3.1 we give an explicit example of a stopping time $\tau'$ relative to the filtration induced by $(X_n)_{n\in\mathbb{N}}$ in the t-test such that $T^{M_{\tau'}}$ is not an $e$-statistic.

### 8.3.3 GROW for 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 $\delta$. In this section we consider hypotheses that are composite in the maximally invariant parameter. We also consider problems in which a fixed prior is placed on the maximally invariant parameter $\delta$. This implements the method of mixtures, a standard method to combine test martingales (Wald, 1945; Darling and Robbins, 1968), which was already used in the context of the anytime-valid t-test (Lai, 1976).

Suppose that the initial hypotheses are not defined by a single value of the maximally invariant parameter $\delta = \delta(\theta)$, as in (8.1), but are instead given by

$$\mathcal{H}_0 : \delta(\theta) = \delta, \quad \delta \in \Delta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \delta(\theta) = \delta, \quad \delta \in \Delta_1, \tag{8.15}$$

where $\Delta_0$ and $\Delta_1$ are two sets of possible values of $\delta = \delta(\theta)$. In Section 8.2.2, we reparametrized $\{\mathbf{P}_\theta\}_{\theta\in\Theta:\delta(\theta)=\delta_0}$ and $\{\mathbf{Q}_\theta\}_{\theta\in\Theta:\delta(\theta)=\delta_1}$ in terms of $G$, and denoted the resulting models as $\{\mathbf{P}_g\}_{g\in G}$ and $\{\mathbf{Q}_g\}_{g\in G}$ respectively. Instead of only considering $\delta_0$ and $\delta_1$, we can do the same for all $\delta \in \Delta_0$ and $\delta \in \Delta_1$. We denote the resulting models as $\{\mathbf{P}_{g,\delta}\}_{g\in G,\delta\in\Delta_0}$ and $\{\mathbf{Q}_{g,\delta}\}_{g\in G,\delta\in\Delta_1}$. As an example, $\mathbf{P}_{g,\delta_0}$ and $\mathbf{Q}_{g,\delta_1}$ correspond to what were previously simply $\mathbf{P}_g$ and $\mathbf{Q}_g$. The problem (8.15) may now be rewritten as

$$\mathcal{H}_0 : X^n \sim \mathbf{P}_{g,\delta}, \quad \delta \in \Delta_0,\ g \in G \quad \text{vs.} \quad \mathcal{H}_1 : X^n \sim \mathbf{Q}_{g,\delta}, \quad \delta \in \Delta_1,\ g \in G. \tag{8.16}$$

Since the distribution of a maximally invariant function of the data $M_n = m_n(X^n)$ depends on the parameter $\delta$, these hypotheses 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 (8.16), it is enough to do so for the invariance-reduced problem

$$\mathcal{H}_0 : M_n \sim \mathbf{P}_\delta^{M_n}, \quad \delta \in \Delta_0 \quad \text{vs.} \quad \mathcal{H}_1 : M_n \sim \mathbf{Q}_\delta^{M_n}, \quad \delta \in \Delta_1. \tag{8.17}$$

We follow the same steps that we followed in Section 8.3.1, and begin by showing that

if there exists a minimizer for the KL minimization problem associated to (8.17), then it has the same value as that associated to (8.16).

**Proposition 8.7.** *Assume that there exists a pair of probability distributions* $\mathbf{\Pi}_0^\star, \mathbf{\Pi}_1^\star$ *on* $\Delta_0$ *and* $\Delta_1$ *that satisfy*

$$\mathrm{KL}(\mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}) = \min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{\delta}\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^{\delta}\mathbf{P}_\delta^{M_n}). \tag{8.18}$$

*For each* $g \in G$*, define the probability distributions* $\mathbf{P}_g^\star = \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_{g,\delta}$ *and* $\mathbf{Q}_g^\star = \mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_{g,\delta}$ *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 8.2, then*

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

From this proposition, using Theorem 8.1 and the steps used for Corollaries 8.3 and 8.5, we can conclude that the ratio of the Bayes marginals for the invariance-reduced data $M_n$ using the optimal priors $\mathbf{\Pi}_0^\star$ and $\mathbf{\Pi}_1^\star$ is both a GROW and a relatively GROW $e$-statistic for (8.16). We now state the corollary and apply it to to our running example, the t-test.

**Corollary 8.8.** *Under the assumptions of Proposition 8.7, the statistic given by*

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

*is a (both absolute and relative) GROW e-statistic for (8.16).*

**Example 8.3** (continues=ex:t-test)**.** Suppose, in the t-test setting, that we are interested in testing

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

for some $\delta_0 < \delta_1$, where, recall, $\delta = \mu/\sigma$ is the maximally invariant parameter. Corollary 8.8 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. Since the density of the t-statistic is monotone in $\delta$, we can use Proposition 3 of GHK, Section 3.1. to infer that the minimum in (8.18) is achieved by the probability distributions $\mathbf{\Pi}_0^\star$ and $\mathbf{\Pi}_1^\star$ that put all of their mass on $\delta_0$ and $\delta_1$, respectively. Corollary 8.8 yields that $T_n^* = p_{\delta_1}^{M_n}/p_{\delta_0}^{M_n}$ 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.

Another approach to deal with the unknown parameter values is to employ proper prior distributions, as is standard practice both within Bayesian statistics and with $e$-statistics. That is, we may want to use specific priors $\tilde{\mathbf{\Pi}}_0$ and $\tilde{\mathbf{\Pi}}_1$ on $\Delta_0$ and $\Delta_1$ respectively. 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 8.3, the proof of Proposition 8.7 also provides the following corollary.

**Corollary 8.9.** *Let $\tilde{\mathbf{\Pi}}_0$ and $\tilde{\mathbf{\Pi}}_1$ be two probability distributions on $\Delta_0$ and $\Delta_1$, 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 8.3, then the e-statistic*

$$\tilde{T}_n = \frac{\int q_\delta(m_n(X^n))\mathrm{d}\tilde{\mathbf{\Pi}}_1(\delta)}{\int p_\delta(m_n(X^n))\mathrm{d}\tilde{\mathbf{\Pi}}_0(\delta)} \tag{8.19}$$

*is both GROW and relatively GROW for testing $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ against $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$.*

**Example 8.4** (continues=ex:t-test)**.** Jeffreys (1961) proposed a Bayesian version of the t-test based on the Bayes factor (8.6) with $\delta_0$ to 0 and a Cauchy prior centered at 0 on $\delta_1$. Popularized as the *Bayesian t-test* (Rouder et al., 2009), it is an instance of (8.19) with $\tilde{\mathbf{\Pi}}_1$ set to aforementioned Cauchy prior and $\tilde{\mathbf{\Pi}}_0$ putting mass 1 on $\delta_0 = 0$. It is itself an $e$-statistic (GHK), but condition (8.14) of Theorem 8.2 does not hold because the Cauchy distribution does not have any moments. Thus, we cannot verify whether (8.19) has the relative GROW property. However, as soon as we replace the Cauchy prior by any prior centered at 0 for which, for some $\varepsilon > 0$, the $(2+\varepsilon)$-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 8.10 in the next section (applied with $d = 1$) to infer that assumption (8.14) holds. Finally, Proposition 8.9 can be applied to conclude that the corresponding Bayes factor is then relatively GROW.

## 8.4 Testing Multivariate Normal Distributions

We show how the theory developed in the previous sections can be applied to hypothesis testing under normality assumptions. The latter is particularly suited for the group-invariant setting, because the family of normal distributions carries a natural invariance under scale-location transformations, as we have already seen in Example 8.1. Different subsets of scale-location transformations correspond to different parameters of interest. We develop two examples in detail. The first is an alternative to Hotelling's

$T^2$ for testing whether the (multivariate) mean of the distribution is identically zero. The corresponding group is that of lower triangular matrices with positive entries on the diagonal. This test is in direct relation with the step-down procedure of Roy and Bargmann (1958)[2] (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. In this case, the group is a subset of the affine linear group.

### 8.4.1 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 $\Sigma$. We consider a test for whether the mean $\mu$ of the distribution is zero. To formalize the test, recall that the Cholesky decomposition of a positive definite matrix $\Sigma$ is $\Sigma = \Lambda\Lambda'$ for a unique $\Lambda \in \mathrm{LT}^+(d)$. Here, $\mathrm{LT}^+(d)$ denotes the group of lower triangular matrices with positive entries on the diagonal, which is amenable. We can therefore parametrize the Gaussians in terms of $(\mu, \Lambda)$, taking the parameter space to be $\Theta = \mathbb{R}^d \times \mathrm{LT}^+(d)$. In this parametrization, consider the following hypothesis testing problem, which generalizes the t-test (Example 8.1) to dimensions $d \geq 1$:

$$\mathcal{H}_0 : \Lambda^{-1}\mu = \delta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \Lambda^{-1}\mu = \delta_1. \tag{8.20}$$

A test for whether $\mu$ is zero can be obtained by setting $\delta_0 = 0$. The group $\mathrm{LT}^+(d)$ acts freely and continuously on $\mathcal{X}^n$ through component-wise matrix multiplication, i.e. $(L, X^n) \mapsto (LX_1, \ldots, LX_n)$ for any $L \in \mathrm{LT}^+(d)$. This action is continuous and free, and can be shown to be proper on the restriction of $\mathcal{X}^n$ to matrices of rank $d$ if $n \geq d + 1$. If $X_i \sim N(\mu, \Lambda)$, then $LX_i \sim N(L\mu, L\Lambda)$, so that $\mathrm{LT}^+(d)$ acts on $\Theta$ by $(L, (\mu, \Lambda)) \mapsto (L\mu, L\Lambda)$ for each $(\mu, \Lambda) \in \Theta$ and $L \in \mathrm{LT}^+(d)$. A maximally invariant parameter under this action is $\delta(\mu, \Lambda) = \Lambda^{-1}\mu$, so that (8.20) is indeed a test of the form described in Section 8.2.2. Furthermore, 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 $\chi(L) = |\det(L)|^n$. It follows that Assumption 8.1 holds and therefore, the likelihood ratio of any maximally invariant statistic is GROW by Corollary 8.3.

By the results of Hall et al. (1965), recapped in Appendix F.1, this likelihood

[2] 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 is shown by Subbaiah and Mudholkar (1978)

ratio must coincide with that of an invariantly sufficient statistic for $\delta$. We now proceed to compute one such statistic. Recall that the pair $S_n = s_n(X^n) = (\bar{X}_n, \bar{V}_n)$, consisting of the unbiased estimators $\bar{X}_n$ and $\bar{V}_n$ for the mean and covariance matrix respectively, is a sufficient statistic for $(\mu, \Sigma)$. Analogous to the technique we used for the parameter space, we can perform the Cholesky decomposition $\bar{V}_n = L_n L_n'$. The statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(S_n) = \sqrt{n/(n-1)} L_n^{-1} \bar{X}_n$ is maximally invariant under the action of $\mathrm{LT}^+(d)$ on $S_n$; in other words, $M_{\mathcal{S},n}$ is invariantly sufficient for $\delta$. Hence, the GROW $e$-statistic can be written as $T^{M_{\mathcal{S},n}} = q^{M_{\mathcal{S},n}} / p^{M_{\mathcal{S},n}}$. Since it was used in Example 8.1 (underneath Corollary 8.9), we give an explicit expression for the likelihood ratio $T^{M_{\mathcal{S},n}}$ when $\delta_0 = 0$, from which values for other $\delta_0$ can be computed. It is based on a more general computation in Appendix F.4.

**Lemma 8.10.** *For the maximally invariant statistic $M_{\mathcal{S},n} = \sqrt{\frac{n}{n-1}} L_n^{-1} \bar{X}_n$, we have*

$$\frac{q^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}{p^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))} = \mathrm{e}^{-\frac{n}{2}\|\delta_1\|^2} \int \mathrm{e}^{n\langle \delta_1, T A_n^{-1} M_{\mathcal{S},n} \rangle} \mathrm{d}\mathbf{P}_{n,I}(T), \tag{8.21}$$

*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.* This follows from Proposition F.5 in Appendix F.4 with $\gamma = \sqrt{n}\delta_1$, $X = \sqrt{n}\bar{X}_n$, $m = n - 1$, and $S = \bar{V}_n$. □

### 8.4.2 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 $X_i, Y_i \in \mathbb{R}$ and $Z_i \in \mathbb{R}^d$ for each $i$. 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

$$\mathcal{H}_0 : \gamma/\sigma = \delta_0 \quad \text{vs.} \quad \mathcal{H}_1 : \gamma/\sigma = \delta_1. \tag{8.22}$$

A test for whether $\gamma = 0$ is readily obtained by taking $\delta_0 = 0$. This problem is invariant under the action of the group $G = \mathbb{R}^+ \times \mathbb{R}^d$ given by $((c, v), (X, Y, Z)) \mapsto (X, cY + v'Z, Z)$ (Kariya, 1980; Eaton, 1989). The corresponding action of $G$ on the parameter space is given by $((c, v), (\gamma, \beta, \sigma)) \mapsto (c\gamma, c\beta + v, c\sigma)$. A maximally invariant parameter is $\delta(\gamma, \beta, \sigma) = \gamma/\sigma$, so that the problem in (8.22) is of the form described in Section 8.2.2. Furthermore, it can be shown that the action of $G$ on $\mathcal{X}$ is continuous and proper, and that $G$ is amenable. Since the Lebesgue measure is again relatively left invariant, it follows that Assumption 8.1 holds. All that remains is to find a maximally invariant function of the data. To this end, define the vectors $Y^n = (Y_1, \ldots, Y_n)'$ and $X^n = (X_1, \ldots, X_n)'$, and the $n \times d$ matrix $\boldsymbol{Z}_n = [Z_1, \ldots, Z_n]'$ whose rows are the vectors $Z_1, \ldots, Z_n$. Assume that $\boldsymbol{Z}_n$ has full rank. A maximally invariant function of the data is given by $M_n = \left(\frac{\boldsymbol{A}_n' Y^n}{\|\boldsymbol{A}_n' Y^n\|}, X^n, \boldsymbol{Z}_n\right)$, where $\boldsymbol{A}_n$ is an $(n - d) \times n$ matrix whose columns form an orthonormal basis for the orthogonal complement of the column space of $\boldsymbol{Z}_n$ (Kariya, 1980; Bhowmik and King, 2007). In order to compute the likelihood ratio for $M_n$, we assume that the mechanism that generates $X^n$ and $\boldsymbol{Z}_n$ is the same under both hypotheses, so that we only need to consider the distribution of $\boldsymbol{U}_n = \frac{\boldsymbol{A}_n' Y^n}{\|\boldsymbol{A}_n' Y^n\|}$ conditionally on $X^n$ and $\boldsymbol{Z}_n$. Bhowmik and King (2007) show that for arbitrary effect size $\delta$, the density of this distribution is given by

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

where $k = n - d$, $u$ is a unit vector in $\mathbb{R}^k$, ${}_1F_1$ is the confluent hypergeometric function, $a(u, \delta) = \delta X^{n\prime} \boldsymbol{A}_n u$, and $c(\delta) = -\frac{1}{2}\delta^2 X^{n\prime} \boldsymbol{A}_n \boldsymbol{A}_n' X^n$. This can be used to compute the likelihood ratio for $M_n$, which is the relatively GROW $e$-statistic for testing (8.22). In fact, Bhowmik and King compute in more generality the density of the maximally invariant statistic when $X$ is allowed to have a nonlinear 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.

## 8.5 Discussion and Future Work

In this concluding section we bring up an issue that deserves further discussion and may inspire future work. We also use this issue to highlight the differences between

our work and related work in a Bayesian context.

### 8.5.1 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 therefore whether amenability is also a necessary condition for the latter to hold. This is not only of theoretical relevance: some groups that are important for statistical practice are not amenable. For instance, the general linear group $\mathrm{GL}(d)$, which is the relevant group in Hotelling's test, is nonamenable. The setup of Hotelling's test is similar to that in Section 8.4.1, except that the hypotheses are given by

$$\mathcal{H}_0 : \|\Lambda^{-1}\mu\|^2 = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \|\Lambda^{-1}\mu\|^2 = \gamma. \tag{8.23}$$

A maximally invariant statistic is the $T^2$-statistic $n\bar{X}_n' \bar{V}_n^{-1} \bar{X}_n$, where, as in Section 8.4.1, $\bar{X}_n$ and $\bar{V}_n$ are the unbiased estimators of the mean and the covariance matrix, respectively. Notice that this test is equivalent to (8.20) 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 (8.21) with a proper prior on $\delta \in \Delta$ as defined there. It follows as a result of Proposition 8.7 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). An interesting question is whether amenability can be replaced by a weaker condition, and/or whether a counterexample to Theorem 8.2 for nonamenable groups can be given.

### 8.5.2 Nonuniqueness Issues Do Not Arise

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. As we explain in full detail in Appendix F.2, this is generally unproblematic: as soon as the assumptions of Theorem 8.2 hold for at least one of the two groups, we can construct the GROW $e$-statistic, and it is uniquely defined. Superficially, this may seem to contradict Sun and Berger (2007) who point out that in some settings, the underlying group is not uniquely determined and then the right Haar prior for the considered model $\mathcal{P}$ is not uniquely defined.

Then, different choices of right Haar prior give different Bayesian posteriors—a fact that has sometimes been taken as a criticism of objective Bayesian approaches. Such nonuniqueness is avoided in our approach. The reason is, essentially, that whereas the GROW $e$-statistic $T_n^*$ is a ratio between Bayes marginals for different models $\mathcal{H}_0$ and $\mathcal{H}_1$ at the same sample size $n$, the Bayes predictive distribution based on a single model $\mathcal{P}$ is a ratio between Bayes marginals for the same $\mathcal{P}$ at different sample sizes $n$ and $n-1$. The role of "same" and "different" being interchanged, it turns out that this Bayes predictive distribution *can* depend on the group on which the right Haar prior for $\mathcal{P}$ is based. Since the Bayes predictive distribution can be rewritten as a marginal over the Bayes posterior, which is Sun and Berger (2007)'s quantity of interest, it is then not surprising that this Bayes posterior may also change if the underlying group is changed. Instead, one may quantify uncertainty by the *e-posterior*, an $e$-statistic-based measure of uncertainty recently put forward by Grünwald (2023): if one replaces the standard Bayes posterior on $\delta$ by the $e$-posterior based on the GROW $e$-statistic $T_n^*$, the nonuniqueness issue disappears as well.

## 8.6 Proofs

In this section, we give all the proofs that were omitted earlier. We first provide two remarks that will be useful throughout the proofs.

**Remark.** Without loss of generality, we may modify 3 in Assumption 8.1 as follows:

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 $\nu$ 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 $\mu$ with multiplier $\chi$, a left-invariant measure $\nu$ can be constructed. Indeed, Bourbaki (2004, Chap. 7, §2 Proposition 7) shows that, under our assumptions, for any multiplier $\chi$ 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 $\varphi$, probability densities with respect to $\mu$ are readily transformed into probability densities with respect to $\nu$. The invariance of the models implies that the densities of $\mathbf{P}_g$ and $\mathbf{Q}_g$ with respect to $\nu$ take the form $p_g(x^n) = p_1(g^{-1}x^n)$ and $q_g(x^n) = q_1(g^{-1}x^n)$ for any $x^n \in \mathcal{X}^n$, where 1 denotes the unit element of the group $G$. It follows that for any $g, h \in G$ it holds that $p_g(x^n) = p_h(hg^{-1}x^n)$ for all $x^n \in \mathcal{X}^n$. A similar statement can be made for $q_g$.

**Remark.** So far, we have only considered the right Haar measure $\rho$ on $G$, however on any locally compact group $G$ there also exists a left-invariant measure $\lambda$, called the left Haar measure. It can be shown that $\lambda$ is relatively right invariant with a multiplier $\Delta$, that is, for any measurable $B \subseteq G$ and $g \in G$ it holds that $\lambda\{Bg\} = \Delta(g)\lambda\{B\}$ for any $g \in G$. Moreover, a computation shows that the measure $\rho'$ defined by $\rho'\{B\} = \lambda\{B^{-1}\}$ for each measurable $B \subseteq G$, is right invariant; in other words, $\rho'$ is a right Haar measure. We may therefore choose $\rho$ to be equal to $\rho'$ and 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 $\int f(h^{-1})\mathrm{d}\lambda(h) = \int f(h)\mathrm{d}\rho(h)$ hold (see Eaton, 1989, Section 1.3).

### 8.6.1 Proofs of Theorem 8.4, Proposition 8.6, Proposition 8.7

Here we prove all results in the main text except the main Theorem 8.2, which is deferred to the next subsection.

*Proof of Theorem 8.4.* Let $g$ be a fixed group element of $G$. Recall from Remark 8.6 that we may assume that both models are dominated by a left invariant measure $\nu$ on $\mathcal{X}$. Theorem 1 by GHK (its simplest instantiation in their Section 2) implies that

$$\sup_{T_n \ e\text{-stat.}} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n] = \inf_{\mathbf{\Pi}_0} \mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}_0^{g'}\mathbf{P}_{g'}), \tag{8.24}$$

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

$$\mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}^{g'}\mathbf{P}_{g'}) = \mathrm{KL}(\mathbf{Q}_h, \tilde{\mathbf{\Pi}}^{g'}\mathbf{P}_{g'}). \tag{8.25}$$

From this, our claim will follow: by symmetry, the previous display implies that $g \mapsto \sup_{T_n \ e\text{-stat.}} \mathbf{E}_g^{\mathbf{Q}}[\ln T_n]$ is constant over $G$ because of its relation to the KL minimization in (8.24). Let $\bar{p} = \int p_{g'}\mathrm{d}\mathbf{\Pi}(g')$, use both the invariance of $\nu$ 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 $\tilde{\mathbf{\Pi}}\{H \in B\} = \mathbf{\Pi}\{H \in$

$gh^{-1}B\}$ 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 $\nu$, 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 (8.25), as was our objective. By our previous discusion, the result follows. □

*Proof of Proposition 8.6.* Let $g \in G$ be arbitrary but fixed. We start by showing that $T^{M_n}$ equals the likelihood ratio for $M^n = (M_1, \ldots, M_n)$ 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 $M^{n-1} = (M_1, \ldots, M_{n-1})$, all the maximally invariant statistics at previous times. Two consequences fall from these observations. First, no additional information about $T^{M_n}$ is gained by knowing the value of $M^{n-1} = (M_1, \ldots, M_{n-1})$ with respect to only knowing $M_{n-1}$, that is, $\mathbf{E}_g^{\mathbf{P}}\left[T^{M_n}|M_{n-1}\right] = \mathbf{E}_g^{\mathbf{P}}\left[T^{M_n}|M^{n-1}\right]$. Second, the likelihood ratio between $\mathbf{P}_g$ and $\mathbf{Q}_g$ for the sequence $M_1, \ldots, M_n$ equals the likelihood ratio for $M_n$ alone, that is,

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

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

*Proof of Proposition 8.7.* Let $\mathbf{\Pi}_0^{g,\delta}, \mathbf{\Pi}_1^{g,\delta}$ be two probability distributions on $G \times \Delta_0$ and $G \times \Delta_1$, respectively. If we call $\mathbf{\Pi}_0^{\delta}$ and $\mathbf{\Pi}_1^{\delta}$ their respective marginals on $\Delta_0$ and $\Delta_1$, then the information processing inequality implies that

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

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

our target infimum, that is,

$$\inf_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}\mathbf{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) \geq \mathrm{KL}(\mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}). \tag{8.26}$$

To show that this is indeed an equality, it suffices to prove it when taking the infimum over a smaller subset of probability distributions $\mathbf{\Pi}_0, \mathbf{\Pi}_1$. We proceed to build such a subset. Let $\mathcal{P}(\mathbf{\Pi}_0^{\star\delta})$ be the set of probability distributions on $G \times \Delta_0$ with marginal distribution $\mathbf{\Pi}_0^{\star\delta}$. Define analogously the set of probability distributions $\mathcal{P}(\mathbf{\Pi}_1^{\star\delta})$ on $G \times \Delta_1$. By our assumptions, Theorem 8.2 can be readily used to conclude that

$$\inf_{(\mathbf{\Pi}_0,\mathbf{\Pi}_1)\in\mathcal{P}(\mathbf{\Pi}_0^{\star\delta})\times\mathcal{P}(\mathbf{\Pi}_1^{\star\delta})} \mathrm{KL}(\mathbf{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}, \mathbf{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) = \mathrm{KL}(\mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}) \tag{8.27}$$

holds; (8.26) and (8.27) together imply the result that we were after. □

### 8.6.2 Proof of the Main Theorem, Theorem 8.2

For the proof of the main result, we use an equivalent definition of amenability to the one that was already anticipated in Section 8.2.2. We take the one that suits our purposes best (see Bondar and Milnes, 1981, p. 109, Condition $A_1$). That is, a group $G$ is amenable if there exists an increasing sequence of symmetric compact subsets $C_1 \subseteq C_2 \subseteq \cdots \subset G$ such that, for any compact set $K \subseteq G$,

$$\frac{\rho\{C_i\}}{\rho\{C_i K\}} \to 1, \quad \text{as } i \to \infty.$$

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 8.2.* Under our assumptions, Theorem 2 of Bondar (1976) implies the existence of a bimeasurable one-to-one map $r : \mathcal{X}^n \to G \times \mathcal{X}^n/G$ 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 $m(x^n) \in \mathcal{X}^n/G$. Hence, by a change of variables, we can take densities with respect to the image measure $\mu$ of $\nu$ under the map $r$ 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 $\mu = \lambda \times \beta$ where $\lambda$ is the left Haar measure on $G$ and $\beta$ is some measure

on $\mathcal{X}^n/G$ (see Remark 8.6). Note that rewriting the data in this way does not affect our objective because the KL divergence remains unchanged under bijective transformations of the data. 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 $\lambda$.

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

$$\mathrm{KL}(\mathbf{\Pi}^g\mathbf{Q}_g, \mathbf{\Pi}^g\mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}^M, \mathbf{P}^M) + \int \mathrm{KL}(\mathbf{\Pi}^g\mathbf{Q}_g^{H|m}, \mathbf{\Pi}^g\mathbf{P}_g^{H|m})\mathrm{d}\mathbf{Q}^M(m). \quad (8.28)$$

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 (8.28) pertaining the conditional distributions given $M$—the second term on the right hand side—goes to zero, that is, such that

$$\int \mathrm{KL}(\mathbf{\Pi}_i^g\mathbf{Q}_g^{H|m}, \mathbf{\Pi}_i^g\mathbf{P}_g^{H|m})\mathrm{d}\mathbf{Q}^M(m) \to 0 \quad \text{as} \quad i \to \infty. \quad (8.29)$$

We define the distributions $\mathbf{\Pi}_i$ as the normalized restriction of the right Haar measure $\rho$ 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

$$\mathbf{\Pi}_i\{g \in B\} := \frac{\rho\{B \cap C_i\}}{\rho\{C_i\}}, \quad (8.30)$$

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

**Lemma 8.11.** *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*

$$\frac{\rho\{J_i\}}{\rho\{J_iK_iL_i\}} \to 1 \quad \textit{as } i \to \infty.$$

The proof of this lemma is given in Appendix F.4.1. There is no risk of dividing by $\infty$ in (8.30): by the continuity of the group operation each $C_i$ is compact, hence $\rho\{C_i\} < \infty$. Lemma 8.11 ensures that $\mathbf{\Pi}_i\{g \in J_i\} \to 1$ 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 (8.29). To this end, notice that for any function $f = f(h, m)$, the expected value $\mathbf{E}_g^{\mathbf{Q}}[f(H, M)] = \mathbf{E}_1^{\mathbf{Q}}[f(gH, M)]$. Indeed,

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

Use this fact to obtain that

$$\int \mathrm{KL}(\mathbf{\Pi}_i^g\mathbf{Q}_g^{H|m}, \mathbf{\Pi}_i^g\mathbf{P}_g^{H|m})\mathrm{d}\mathbf{Q}(m) = \int \mathbf{E}_1^{\mathbf{Q}}\left[\ln\frac{q_i(gH|M)}{p_i(gH|M)}\right]\mathrm{d}\mathbf{\Pi}_i(g) \tag{8.31}$$

$$= \underbrace{\int \mathbf{E}_1^{\mathbf{Q}}\left[\mathbf{1}\{gH \in J_iK_i\}\ln\frac{q_i(gH|M)}{p_i(gH|M)}\right]\mathrm{d}\mathbf{\Pi}_i(g)}_{\mathrm{A}} + \underbrace{\int \mathbf{E}_1^{\mathbf{Q}}\left[\mathbf{1}\{gH \notin J_iK_i\}\ln\frac{q_i(gH|M)}{p_i(gH|M)}\right]\mathrm{d}\mathbf{\Pi}_i(g)}_{\mathrm{B}}. \tag{8.32}$$

We separate the rest of the proof in two steps, one for bounding each term in (8.31). These steps use two technical lemmas that we prove in Appendix F.4.1.

**Bound for A in** (8.31)**:** Recall 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 8.12.** *Let $N$ and $L$ be compact subsets of $G$. Assume that $L$ is symmetric. Then, for each $m \in \mathcal{X}^n/G$ it holds that*

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

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

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

At the same time this implies that A in (8.31) 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 $\mathbf{P}_1\{H \in L_i \mid M = m\} \to 1$ 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 (8.31). This ends the first step of the proof. Now, we turn to the second term in (8.31).

**Bound for B in** (8.31)**:** Our strategy at this point is to show that, as $i \to \infty$,

$$\int \mathbf{Q}_1 \{gH \notin J_iK_i\} \,\mathrm{d}\mathbf{\Pi}_i(g) \to 0, \tag{8.33}$$

and to use (8.14) to show our goal, that B in (8.31) tends to zero. To show (8.33), notice that if $g \in J_i$ and $h \in K_i$, then $gh \in J_iK_i$, which implies that

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

Since the sets $K_i$ increase to cover $G$, we have $\mathbf{Q}_1 \{H \in K_i\} \to 1$ as $i \to \infty$, and by our initial choice of sets $J_i, K_i, L_i$, the probability $\mathbf{\Pi}_i \{g \in J_i\} \to 1$, as $i \to \infty$. Hence (8.33) holds. To bound the second term, we use the following lemma with $\mathbf{\Pi} = \mathbf{\Pi}_i$.

**Lemma 8.13.** *Let* $\mathbf{\Pi}$ *be a distribution on* $G$*. Then, for each* $h \in G$ *and* $m \in \mathcal{X}^n/G$*, setting* $\mathrm{d}\mathbf{\Pi}(g|h,m) = \frac{q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}{\int q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}$*, it holds that*

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

After invoking the previous lemma, apply Hölder's and Jensen's inequality consecutively to bound B in (8.31) by

$$\iint \left[\mathbf{1}\{gh \notin J_iK_i\} \int \ell\,(gh|m)\mathrm{d}\mathbf{\Pi}_i(g'|h,m)\right] \mathrm{d}\mathbf{Q}_1(h,m)\mathrm{d}\mathbf{\Pi}_i(g) \leq \tag{8.34}$$

$$\underbrace{\left(\int \mathbf{Q}_1 \{gH \notin J_iK_i\} \,\mathrm{d}\mathbf{\Pi}_i(g)\right)^{1/q}}_{\to 0 \text{ as } i\to\infty \text{ by (8.33)}} \left(\iint \left|\int \ell\,(gh|m)\mathrm{d}\mathbf{\Pi}_i(g'|h,m)\right|^p \mathrm{d}\mathbf{Q}_1(h,m)\mathrm{d}\mathbf{\Pi}_i(g)\right)^{1/p}$$

where here and in the sequel, $\ell\,(gh|m)$ abbreviates $\ln \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)}$, and $p = 1+\varepsilon$ and $q$ is $p$'s Hölder conjugate, that is, $1/p + 1/q = 1$. Next, we show that the second factor on

the right of (8.34) remains bounded as $i \to \infty$. By Jensen's inequality, this quantity is smaller than

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

After a series of rewritings and using our Assumption (8.14), we will show that this quantity is bounded. First, we deduce that

$$\begin{aligned}
\iint | \ell\,(gh|m)|^p \, \mathrm{d}\mathbf{\Pi}_i(g'|h,m)\mathrm{d}\mathbf{Q}_1(h,m)\mathrm{d}\mathbf{\Pi}_i(g) = \\
\iint | \ell\,(h|m)|^p \, \mathrm{d}\mathbf{\Pi}_i(g'|h,m)\mathrm{d}\mathbf{Q}_g(h,m)\mathrm{d}\mathbf{\Pi}_i(g) = \\
\iint | \ell\,(h|m)|^p \, \mathrm{d}\mathbf{\Pi}_i(g'|h,m)\mathrm{d}\mathbf{Q}_i(h,m) = \mathbf{E}_1^{\mathbf{Q}} \left| \ln \frac{q_1(H|M)}{p_1(H|M)} \right|^p,
\end{aligned}$$

where we used again the change of variable that we used to obtain (8.31)—but now in the opposite direction—and in the final equality, we used Bayes' theorem. Hence, as

$$\begin{aligned}
\left( \mathbf{E}_1^{\mathbf{Q}} \left[ \left| \ln \frac{q_1(H|M)}{p_1(H|M)} \right|^p \right] \right)^{1/p} &\leq \left( \mathbf{E}_1^{\mathbf{Q}} \left[ \left| \ln \frac{q_1(H,M)}{p_1(H,M)} \right|^p \right] \right)^{1/p} + \left( \mathbf{E}_1^{\mathbf{Q}} \left[ \left| \ln \frac{q_1(M)}{p_1(M)} \right| \right]^p \right)^{1/p} \\
&< \infty
\end{aligned}$$

by (8.14), we have shown that (8.34) tends to 0 as $i \to \infty$ and that consequently B in (8.31) tends to 0 in the same limit.

After completing these two steps, we have shown that both A and B in (8.31) tend to 0 as $i \to \infty$, and that consequently the claim of the theorem follows. All that is left is to prove lemmas 8.11, 8.12, and 8.13. The proofs being straightforward but tedious, we delegated these to Appendix F.4. □

# 9 Discussion

From a wider perspective, the content of this thesis was focused on the construction of optimal anytime-valid hypothesis tests. The premise established in the introduction was that these methods are valuable because they allow researchers to analyze and learn from data in real time. This is not only advantageous due to the additional flexibility in the design of experiments; it also solves the problem of inflated type-I error rates due to optional stopping. Here, we discuss two implicit assumptions that form the basis of this premise.

## 9.1 Publish or perish

The first assumption is that the adoption of anytime-valid methods will reduce the inflation of type-I error rates. Although this sounds reasonable, it may overlook the deeper reasons that explain why optional stopping is a problem in the first place. In many academic disciplines, researchers are pushed to publish papers in order to secure funding and maintain a good reputation, as the volume of publications is often used as a measure of success. This culture, sometimes referred to as "publish or perish", creates strong incentives for researchers to find positive results. This is amplified by the fact that papers are often not considered for publication unless they contain a statistical analysis in which some null hypothesis is rejected. Statistical analyses have thus ceased to play a supporting role in academia and have instead become the main target. The problem with this is captured by Goodhart's law: "When a measure becomes a target, it ceases to be a good measure". By themselves, the test statistics that hypothesis tests are based on are reasonable measures of evidence. However, as soon as the goal becomes finding evidence, statistical methods are bound to be abused. This applies to anytime-valid methods as well. For example, researchers could start considering multiple different test statistics and choosing the one that gives the most evidence, or

leaving out data points that negatively impact the strength of the evidence. In fact, in the same survey that was cited in Chapter 1 to show the scope of the problem with optional stopping, 38% of participants admitted to "deciding whether to exclude data after looking at the impact of doing so on the results" (John et al., 2012). While it will likely help to a certain extent, there is little reason to believe that switching to anytime-valid methods will fix the problem of inflated type-I errors entirely; the way in which the system is being cheated will simply change.

Of course, this is only one side of the story. There are also many situations where the only incentive for researchers performing statistical analyses is to draw accurate conclusions. Think of, for example, farmers that are testing the impact of a new crop variety on yield or companies that are testing whether some software version is better than another. In such cases, the flexibility that anytime-valid tests offer might indeed prevent researchers from accidentally misusing statistical tools, leading to a reduction in the inflation of type-I error rates.

## 9.2 How Much Freedom Is Too Much?

The second assumption underlying the premise of this work is that more flexibility in experimental design is always better. Indeed, anytime-valid methods offer experimenters great flexibility in the sense that, no matter how data are collected, the resulting conclusions will be valid (the type-I error probability is controlled). Nevertheless, the construction based on $e$-processes gives an idea about when to stop the data collection: stop when the evidence is large enough to reject. However, as noted by Schmitz (1993): "... a stopping rule gives, at any time point, only the advice whether to stop or not but not how the next observations have to be made, which is a design aspect of the experiment." That is, while the construction hints at when to stop data collection, it does not tell experimenters how to proceed with data collection. The idea is that there is no need to prescribe how this should be done, as one can just proceed fully sequentially. That is, by collecting data one-by-one. However, it was already noted by Wald (1947, p. 101) (the founder of the SPRT) that: "For practical reasons, it may sometimes be preferable to take the observations in groups, rather than singly." A fitting example is given by Barnard (1946), who was working on sequential methods at the same time as Wald: "If we are experimenting with the growth of trees, for example, it may take years for a tree to grow to maturity; in such a case it would be absurd, in our present system of society, to grow one tree first, and see what happened, and, then to plant another, and so on." On the other hand, there are

also applications where data do naturally come in one-by-one. In the current theory of anytime-valid testing, it is left to the experimenter to decide which situation they are in and how they should proceed with the data collection. However, the number of data points to be collected at any moment should ideally be based on a trade-off between how close the $e$-process already is to the desired threshold, and the costs of data collection (see also Schmitz, 1993; Novikov, 2024). For example, if the $e$-process is very close to $1/\alpha$, it might be beneficial to only collect a few data points, but only if the cost of collecting a small number of data points is not disproportionately high compared to collecting a larger batch. It seems unfair to leave such decisions entirely up to the experimenters, and a theory of optimal data collection may be a promising direction for future research that could lead to a wider adoption of anytime-valid methods.

# Summary

In this dissertation, we study hypothesis testing: evaluating whether sample data support a claim regarding a broader population. The approach is to assume that the claim is false and to examine whether this assumption, called the null hypothesis, holds up in light of the data. For example, researchers in a clinical trial wish to use the data to refute the hypothesis that their medication does not work. Hypothesis tests help guide decision-making in all walks of life, from finance to agriculture, by offering a structured framework to evaluate the strength of evidence against hypotheses.

This dissertation is a contribution to the theory of anytime-valid hypothesis tests, which are tools that are compatible with flexible experimental design. That is, anytime-valid methods allow researchers to stop or continue their experiment based on observed data, which is not possible with traditional methods. Anytime-valid methods work by keeping track of a numerical measure of evidence—the $e$-process—against the null hypothesis. In particular, we consider $e$-processes obtained through the combination of $e$-statistics, which measure the evidence that can be derived from each data point separately. The hypothesis can be refuted if the combined evidence against it grows too large. The goal is therefore to construct $e$-statistics that give as much evidence as possible if the hypothesis is not true. These are called log-optimal $e$-statistics.

In Chapter 3, we discuss the abstract problem of finding log-optimal $e$-statistics, which leads to a general recipe for their construction. In Chapters 4–7, we use this recipe to find the log-optimal $e$-statistics for specific hypotheses of interest (exponential families, conditional independence and group invariance). A key assumption underlying the optimality results in these chapters is that we know (or can learn) what exactly happens if the null hypothesis is not true. In Chapter 8, we take a different approach and consider the worst case over a set of possible alternative hypotheses. That is, we study a setting where it is possible to compute the $e$-statistic that maximizes the rate at which evidence is accumulated in the worst case over the alternative.

# Samenvatting

In dit proefschrift, getiteld *Optimale Toetsstatistieken Voor Altijd-Valide Hypothese Toetsen*, bestuderen we hypothese toetsen. Dit zijn methodes om te bepalen of een bepaalde veronderstelling over een populatie ondersteund wordt door een steekproef. Bij een hypothese toets neemt men aan dat de veronderstelling in kwestie onjuist is en gaat dan na of deze aanname, de nulhypothese genoemd, stand houdt met oog op de gevonden resultaten. Onderzoekers in klinische studies kunnen zo bijvoorbeeld de data uit hun studie gebruiken om de hypothese dat hun medicatie niet werkt te verwerpen. Zo kunnen zij concluderen dat de medicatie wél werkt. Hypothese toetsen helpen bij het maken van besluiten in allerlei facetten van het leven, van financiën tot agricultuur, door het bieden van een gestructureerde methode om bewijs tegen hypotheses te evalueren.

Dit proefschrift levert een bijdrage aan de theorie van altijd-valide hypothese toetsen. Dit zijn methodes die geschikt zijn voor flexibele onderzoeksopzetten. Deze toetsen staan onderzoekers namelijk toe om hun experiment te stoppen of voort te zetten op basis van geobserveerde data, wat niet mogelijk is met traditionele methodes. Altijd-valide methodes werken door een maat van bewijs—het $e$-proces—tegen de nulhypothese over de tijd bij te houden. In dit proefschrift, beschouwen we in het bijzonder $e$-processen die het resultaat zijn van de combinatie van $e$-statistieken. Een $e$-statistiek meet het bewijs dat uit één afzonderlijk datapunt kan worden gewonnen. De hypothese kan verworpen worden als het bewijs, gecombineerd over alle datapunten, ertegen te groot wordt. Het doel is zodoende om $e$-statistieken te construeren die zo veel mogelijk bewijs geven als de hypothese niet waar is. Dat worden log-optimale $e$-statistieken genoemd.

In Hoofdstuk 3 beschouwen we het abstracte probleem van het vinden van log-optimale $e$-statistieken, wat leidt tot een algemeen recept voor hun constructie. In Hoofdstukken 4–7 gebruiken we dit recept om de log-optimale $e$-statistieken voor spec-

ifieke nulhypotheses te vinden (exponentiële families, conditionele onafhankelijkheid en groepsinvariantie). Een belangrijke aanname achter de optimaliteits resultaten in deze hoofdstukken is dat we weten (of kunnen leren) wat er precies gebeurt als de nulhypothese niet waar is. In Hoofdstuk 8 volgen we een andere aanpak en beschouwen we het ergste geval over een verzameling alternatieve hypotheses. Dat wil zeggen, we bestuderen een context waarin het mogelijk is om de $e$-statistiek vast te stellen die het meeste bewijs accumuleert in het ergste geval over het alternatief.

# Bibliography


Adams, R. (2020). Safe hypothesis tests for the 2× 2 contingency table. Master's thesis, Delft University of Technology.

Agrawal, A. (2018). Lecture notes on Loewner order. `https://www.akshayagrawal.com/lecture-notes/html/loewner-order.html`.

Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

Andersson, S. (1982). Distributions of maximal invariants using quotient measures. *The Annals of Statistics*, 10(3):955–961.

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, 10(1):89–100.

Awad, Y., Bar-Lev, S. K., and Makov, U. (2022). A new class of counting distributions embedded in the Lee–Carter model for mortality projections: A Bayesian approach. *Risks*, 10(6):111.

Azadkia, M. and Chatterjee, S. (2021). A Simple Measure of Conditional Dependence. *The Annals of Statistics*, 49(6):3070 – 3102.

Balsubramani, A. and Ramdas, A. (2016). Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, Uncertainty in Artificial Intelligence 2016, pages 42—-51, Arlington, Virginia, USA. Uncertainty in Artificial Intelligence Press.

Bar-Lev, S. K. (2020). Independent, though identical results: the class of Tweedie on power variance functions and the class of Bar-Lev and Enis on reproducible natural exponential families. *International Journal of Statistics and Probability*, 9(1):30–35.

Bar-Lev, S. K., Letac, G., and Ridder, A. (2024). A delineation of new classes of exponential dispersion models supported on the set of nonnegative integers. *Annals of the Institute of Statistical Mathematics*, 76(4):679–709.

Bar-Lev, S. K. and Ridder, A. (2021). New exponential dispersion models for count data – the ABM and LM classes. *ESAIM: Probability and Statistics*, 25:31–52.

Bar-Lev, S. K. and Ridder, A. (2023). Exponential dispersion models for overdispersed zero-inflated count data. *Communications in Statistics-Simulation and Computation*, 52(7):3286–3304.

Baringhaus, L. (1991). Testing for spherical symmetry of a multivariate distribution. *The Annals of Statistics*, pages 899–917.

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Supplement to the Journal of the Royal Statistical Society*, 8(1):1–21.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester, UK.

Barnett, A. (2018). It's a sign of how bad things have got that researchers think it's acceptable to write this in a nature journal: "we continuously increased the number of animals until statistical significance was reached to support our conclusions.". `https://x.com/aidybarnett/status/1036392482139865088`. X user @aidybarnett, id: 1036392482139865088, Accessed: 23-10-2024.

Barron, A. (1998). Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems. In *Bayesian Statistics*, volume 6, pages 27–52. Oxford University Press, Oxford.

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A*, 60:307–321.

Berger, J. O. and Sun, D. (2008). Objective priors for the bivariate normal model. *The Annals of Statistics*, 36(2):963–982.

Berk, R. H. (1972). A note on sufficiency and invariance. *The Annals of Mathematical Statistics*, 43(2):647–650.

Berman, S. M. (1965). Sign-invariant random variables and stochastic processes with sign-invariant increments. *Transactions of the American Mathematical Society*, 119(2):216–243.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The Conditional Permutation Test for Independence While Controlling for Confounders. *Journal of the Royal Statistical Society: Series B*, 82(1):175–197.

Bhowmik, J. L. and King, M. L. (2007). Maximal invariant likelihood based testing of semi-linear models. *Statistical Papers*, 48(3):357–383.

Bilodeau, B., Foster, D. J., and Roy, D. M. (2023). Minimax rates for conditional density estimation via empirical entropy. *The Annals of Statistics*, 51(2):762–790.

Bondar, J. V. (1976). Borel cross-sections and maximal invariants. *The Annals of Statistics*, 4(5):866–877.

Bondar, J. V. and Milnes, P. (1981). Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57(1):103–128.

Bourbaki, N. (2004). *Integration II: Chapters 7–9*. Elements of Mathematics. Springer-Verlag, Berlin Heidelberg, 1st edition.

Brinda, W. D. (2018). *Adaptive estimation with Gaussian radial basis mixtures*. PhD thesis, Yale University.

Brown, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *The Annals of Mathematical Statistics*, 37(5):1087–1136.

Brown, L. D. (1986). Fundamentals of statistical exponential families with applications in statistical decision theory. *Lecture Notes-Monograph Series*, 9:i–279.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577.

Carney, D. R. (2016). My position on "power poses". `https://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf`. Accessed: 23-10-2024.

Carney, D. R., Cuddy, A. J., and Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, 21(10):1363–1368.

Casper, C., Cook, T., and on FORTRAN program ld98., O. A. P. B. (2022). *ldbounds: Lan-DeMets Method for Group Sequential Boundaries*. R package version 2.0.0.

Cesa-Bianchi, N. and Lugosi, G. (2001). Worst-case bounds for the logarithmic loss of predictors. *Machine Learning*, 43:247–264.

Chang, J. T. and Pollard, D. (1997). Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317.

Chen, P., Chen, Y., and Rao, M. (2008). Metrics defined by Bregman divergences. *Communications in Mathematical Sciences*, 6(4):915–926.

Chiu, K. and Bloem-Reddy, B. (2023). Non-parametric hypothesis tests for distributional group symmetry. ArXiv preprint arXiv:2307.15834.

Cohen, T. and Welling, M. (2016). Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, Inc., New York.

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2):290–299.

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der ergodizität von Markoffschen Ketten. *A Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 8(1-2):85–108.

Csiszár, I. and Matúš, F. (2003). Information projections revisited. *IEEE Transactions on Information Theory*, 49(6):1474–1490.

Csiszár, I. and Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statistics and Decisions*, Supplementary Issue 1:205–237.

Cukier, K. and Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Affairs*, 92(3):20–32.

Darling, D. and Robbins, H. (1967). Confidence Sequences for Mean, Variance, and Median. *Proceedings of the National Academy of Sciences*, 58(1):66–68.

Darling, D. A. and Robbins, H. (1968). Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences*, 61(3):804–809.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society: Series B*, 41(1):1–15.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society, Series B (Methodological)*, 35(2):189–233.

Diaconis, P. (1988). Group representations in probability and statistics. *Lecture notes-monograph series*, 11:i–192.

Dodge, H. F. and Romig, H. G. (1929). A method of sampling inspection. *The Bell System Technical Journal*, 8(4):613–631.

Duan, B., Ramdas, A., and Wasserman, L. (2022). Interactive rank testing by betting. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 201–235.

Durrett, R. (2019). *Probability: Theory and examples*. Number 49 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5th edition.

Eaton, M. L. (1989). Group invariance applications in statistics. *Regional Conference Series in Probability and Statistics*, 1:i–133.

Eaton, M. L. and Sudderth, W. D. (1999). Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli*, 5(5):833–854. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability.

Eaton, M. L. and Sudderth, W. D. (2002). Group invariant inference and right Haar measure. *Journal of Statistical Planning and Inference*, 103(1):87–99.

Efron, B. (2022). *Exponential Families in Theory and Practice*. Institute of Mathematical Statistics Textbooks. Cambridge University Press.

Fedorova, V., Gammerman, A., Nouretdinov, I., and Vovk, V. (2012). Plug-in martingales for testing exchangeability on-line. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1639–1646, New York, NY, USA. Omnipress.

Feller, W. K. (1940). Statistical aspects of esp. *The Journal of Parapsychology*, 4(2):271.

Fisher, R. A. (1936). Design of experiments. *British Medical Journal*, 1(3923):554.

Fontana, M., Zeni, G., and Vantini, S. (2023). Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23.

Foster, D. J., Kale, S., Luo, H., Mohri, M., and Sridharan, K. (2018). Logistic regression: The Importance of Being Improper. In *Conference on learning theory*, pages 167–208.

Fraiman, R., Moreno, L., and Ransford, T. (2024). Application of the cramér–wold theorem to testing for invariance under group actions. *TEST*, 33(2):379–399.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007). Kernel Measures of Conditional Dependence. In *Advances in Neural Information Processing Systems*, volume 20.

Giri, N., Kiefer, J., and Stein, C. (1963). Minimax character of Hotelling's $T^2$ test in the simplest case. *The Annals of Mathematical Statistics*, 34(4):1524 – 1535.

Greenwood, J. A. (1938). An empirical investigation of some sampling problems. *The Journal of Parapsychology*, 2(3):222.

Greenwood, J. A. and Greville, T. (1939). On the probability of attaining a given standard deviation ratio in an infinite series of trials. *The Annals of Mathematical Statistics*, 10(3):297–298.

Grünwald, P. (2007). *The minimum description length principle*. MIT press.

Grünwald, P. (2023). The E-posterior. *Philosophical Transactions of the Royal Society of London, Series A*.

Grünwald, P., de Heide, R., and Koolen, W. (2024). Safe testing. *Journal of the Royal Statistical Society, Series B*.

Grünwald, P. and Harremoës, P. (2009). Finiteness of redundancy, regret, Shtarkov sums, and Jeffreys integrals in exponential families. In *Proceedings for the International Symposium for Information Theory, Seoul, 2009*, pages 714–718. IEEE.

Grünwald, P., Henzi, A., and Lardy, T. (2023). Anytime valid tests of conditional independence under model-X. *Journal of the American Statistical Association*, 119(546):1554–1565.

Grünwald, P. D. (2024). Beyond neyman–pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39).

Grünwald, P. D. and Mehta, N. A. (2019). A tight excess risk bound via a unified pac-bayesian–rademacher–shtarkov–mdl complexity. In *Algorithmic Learning Theory*, pages 433–465. PMLR.

Grünwald, P. D. and Mehta, N. A. (2020). Fast rates for general unbounded loss functions: from ERM to generalized Bayes. *The Journal of Machine Learning Research*, 21(1):2040–2119.

Grünwald, P., Lardy, T., Hao, Y., Bar-Lev, S. K., and De Jong, M. (2024). Optimal e-values for exponential families: the simple case. ArXiv preprint, arXiv:2404.19465. A revised version of this manuscript has been accepted as a contribution to the Springer festschrift "Information Theory, Probability and Statistical Learning: A Festschrift in Honor of Andrew Barron.".

Hall, W. J., Wijsman, R. A., and Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *The Annals of Mathematical Statistics*, 36(2):575–614.

Hall, W. J., Wijsman, R. A., and Ghosh, J. K. (1995). Correction: The relationship between sufficiency and invariance with applications in sequential analysis. *The Annals of Statistics*, 23(2):705–705.

Ham, D. W., Imai, K., and Janson, L. (2024). Using machine learning to test causal hypotheses in conjoint analysis. *Political Analysis*, 32(3):329–344.

Hao, Y., Grünwald, P., Lardy, T., Long, L., and Adams, R. (2024). E-values for k-sample tests with exponential families. *Sankhya A*, 86(1):596–636.

Haussler, D. and Opper, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492.

Henzi, A. and Law, M. (2024). A rank-based sequential test of independence. *Biometrika*.

Henzi, A., Puke, M., Dimitriadis, T., and Ziegel, J. (2023). A safe hosmer-lemeshow test. *The New England Journal of Statistics in Data Science*, 2(2):175–189.

Henzi, A. and Ziegel, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika*, 109(3):647–663.

Howard, S. R., Ramdas, A., Mcauliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080.

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press, London, 3rd edition.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532.

Jørgensen, B. (1997). *The Theory of Exponential Dispersion Models*, volume 76 of *Monographs on Statistics and Probability*. Chapman and Hall, London.

Kariya, T. (1980). Locally robust tests for serial correlation in least squares regression. *The Annals of Statistics*, 8(5):1065–1070.

Katsevich, E. and Ramdas, A. (2022). On the Power of Conditional Independence Testing Under Model-X. *Electronic Journal of Statistics*, 16(2):6348 – 6394.

Kelly, J. L. (1956). A new interpretation of information rate. *The Bell System Technical Journal*, 35(4):917–926.

Koning, N. W. (2023). Online permutation tests: $e$-values and likelihood ratios for testing group invariance. ArXiv preprint arXiv:2310.01153.

Koolen, W. M. and Grünwald, P. (2022). Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 141:69–82.

Kotłowski, W. and Grünwald, P. (2011). Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 457–476. JMLR Workshop and Conference Proceedings.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Kumar, M. A. and Sason, I. (2016). Projection theorems for the Rényi divergence on $\alpha$-convex sets. *IEEE Transactions on Information Theory*, 62(9):4924–4935.

Lai, T. L. (1976). On confidence sequences. *The Annals of Statistics*, 4(2):265–280.

Lai, T. L. (1977). Power-one tests based on sample sums. *The Annals of Statistics*, 5(5):866–880.

Lardy, T. (2021). E-values for hypothesis testing with covariates. Master's Thesis, Leiden University.

Lardy, T., Grünwald, P., and Harremoës, P. (2024). Reverse information projections and optimal e-statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631.

Lardy, T. and Pérez-Ortiz, M. F. (2024). Anytime-valid tests of group invariance through conformal prediction. ArXiv preprint arXiv:2401.15461.

Larsson, M., Ramdas, A., and Ruf, J. (2024). The numeraire e-variable and reverse information projection. ArXiv preprint arXiv:2402.18810.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer-Verlag, New York, 3rd edition.

Lehmann, E. L. and Stein, C. (1949). On the theory of some non-parametric hypotheses. *The Annals of Mathematical Statistics*, 20(1):28–45.

Levin, L. A. (1976). Uniform tests of randomness. In *Doklady Akademii Nauk*, volume 227, pages 33–35. Russian Academy of Sciences.

Lhéritier, A. and Cazals, F. (2018). A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370.

Li, J. (1999). *Estimation of mixture models*. PhD thesis, Yale University, New Haven, CT.

Li, J. and Barron, A. (1999). Mixture density estimation. *Advances in neural information processing systems*, 12.

Li, S. (2010). Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70.

Li, S. and Liu, M. (2023). Maxway crt: improving the robustness of the model-x inference. *Journal of the Royal Statistical Society, Series B*, 85(5):1441–1470.

Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory*, 50(11):2708–2726.

Liang, H. (2023). Stratified safe sequential testing for mean effect. Master's Thesis, University of Amsterdam.

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*. Teubner, Leipzig.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.

Lindon, M., Ham, D. W., Tingley, M., and Bojinov, I. (2024). Anytime-valid linear models and regression adjusted causal inference in randomized experiments. ArXiv preprint arXiv:2210.08589.

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2021). Fast and Powerful Conditional Randomization Testing via Distillation. *Biometrika*, 109(2):277–293.

Malov, S. (1996). Sequential ranks and order statistics. *Journal of Mathematical Sciences*, 81:2434–2441.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models (2nd ed.)*. CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Meckes, E. S. (2019). *The random matrix theory of the classical compact groups*, volume 218. Cambridge University Press.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10(1):65–80.

Niu, Z., Chakraborty, A., Dukes, O., and Katsevich, E. (2024). Reconciling model-x and doubly robust approaches to conditional independence testing. *The Annals of Statistics*, 52(3):895–921.

Nogales, A. G. and Oyola, J. A. (1996). Some remarks on sufficiency, invariance and conditional independence. *The Annals of Statistics*, 24(2):906–909.

Novikov, A. (2024). Group sequential hypothesis tests with variable group sizes: Optimal design and performance evaluation. *Communications in Statistics-Theory and Methods*, 53(16):5744–5760.

O'Brien, P. C. and Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics*, 35(3):549–556.

Pandeva, T., Bakker, T., Naesseth, C. A., and Forré, P. (2022). E-valuating classifier two-sample tests. *ArXiv preprint arXiv:2210.13027*.

Paterson, A. L. (1988). *Amenability*. Number 29 in Mathematical surveys and monographs. American Mathematical Soc., 1st edition.

Pérez-Ortiz, M. F., Lardy, T., De Heide, R., and Grünwald, P. D. (2024). E-statistics, group invariance and anytime-valid testing. *The Annals of Statistics*, 52(4):1410–1432.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, 64(2):191–199.

Pocock, S. J. and Simon, R. (1975). Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. *Biometrics*, 31(1):103–115.

Qian, G. and Field, C. (2002). Law of Iterated Logarithm and Consistent Model Selection Criterion in Logistic Regression. *Statistics & Probability Letters*, 56(1):101–112.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. ArXiv preprint arXiv:2009.03167.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2022). Testing Exchangeability: Fork-Convexity, Supermartingales and E-Processes. *International Journal of Approximate Reasoning*, 141:83–109.

Ranehill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., and Weber, R. A. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26(5):653–656.

Reiter, H. and Stegeman, J. D. (2000). *Classical harmonic analysis and locally compact groups*. London Mathematical Society Monographs. Oxford University Press, Oxford, New York, 2nd edition.

Ren, Z. and Barber, R. F. (2024). Derandomised knockoffs: leveraging e-values for false discovery rate control. *Journal of the Royal Statistical Society, Series B*, 86(1):122–154.

Rényi, A. (1962). On the extreme elements of observations. *MTA III. Oszt. Közl*, 12:105–121.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58:527–535.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409.

Robbins, H. and Siegmund, D. (1974). The expected sample size of some tests of power one. *The Annals of Statistics*, 2(3):415–436.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., and Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2):225–237.

Roy, S. N. and Bargmann, R. E. (1958). Tests of multiple independence and the associated confidence bounds. *The Annals of Mathematical Statistics*, 29(2):491–503.

Runge, J. (2018). Conditional Independence Testing Based on a Nearest-Neighbor Estimator of Conditional Mutual Information. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 938–947.

Rushton, S. (1950). On a sequential t-test. *Biometrika*, 37(3–4):326–333.

Sagiroglu, S. and Sinanc, D. (2013). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)*, pages 42–47. IEEE.

Saha, A. and Ramdas, A. (2024). Testing exchangeability by pairwise betting. In *International Conference on Artificial Intelligence and Statistics*, pages 4915–4923. PMLR.

Savage, I. R. (1956). Contributions to the Theory of Rank Order Statistics-the Two-Sample Case. *The Annals of Mathematical Statistics*, 27(3):590–615. Publisher: Institute of Mathematical Statistics.

Schmitz, N. (1993). *Optimal sequentially planned decision procedures*, volume 79 of *Lecture Notes in Statistics*. Springer-Verlag New York.

Sen, P. K. and Ghosh, M. (1973a). A chernoff-savage representation of rank order statistics for stationary $\varphi$-mixing processes. *Sankhyā: The Indian Journal of Statistics, Series A*, 35(2):153–172. Publisher: Springer.

Sen, P. K. and Ghosh, M. (1973b). A Law of Iterated Logarithm for One-Sample Rank Order Statistics and an Application. *The Annals of Statistics*, 1(3):568–576. Publisher: Institute of Mathematical Statistics.

Sen, P. K. and Ghosh, M. (1974). Sequential Rank Tests for Location. *The Annals of Statistics*, 2(3):540–552. Publisher: Institute of Mathematical Statistics.

Shaer, S., Maman, G., and Romano, Y. (2023). Model-x sequential testing for conditional independence via testing by betting. In *International Conference on Artificial Intelligence and Statistics*, pages 2054–2086. PMLR.

Shafer, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431.

Shah, R. D. and Peters, J. (2020). The Hardness of Conditional Independence Testing and the Generalized Covariance Measure. *The Annals of Statistics*, 48(3):1514–1538.

Shalaevskii, O. V. (1971). Minimax character of Hotelling's T2 test. I. In Kalinin, V. M. and Shalaevskii, O. V., editors, *Investigations in Classical Problems of Probability Theory and Mathematical Statistics: Part I*, pages 74–101. Springer US, Boston, MA, 1st edition.

Shiryaev, A. N. (2016). *Probability-1*, volume 95. Springer.

Sidak, Z., Sen, P. K., and Hajek, J. (1999). *Theory of Rank Tests*. Elsevier.

Simmons, J. P. and Simonsohn, U. (2017). Power posing: P-curving the evidence. *Psychological Science*, 28(5):687–693.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5):1–13.

Smith, A. F. (1981). On random sequences with centred spherical symmetry. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(2):208–209.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737.

Subbaiah, P. and Mudholkar, G. S. (1978). A comparison of two tests for the significance of a mean vector. *Journal of the American Statistical Association*, 73(362):414–418.

Sun, D. and Berger, J. O. (2007). Objective Bayesian analysis for the multivariate normal model. *Bayesian Statistics*, 8:525–562.

Ter Schure, J. and Grünwald, P. (2019). Accumulation bias in meta-analysis: the need to consider time in error control. *F1000Research*, 8.

Ter Schure, J. and Grünwald, P. (2022). All-in meta-analysis: breathing life into living systematic reviews. *F1000Research*, 11.

ter Schure, J., Pérez-Ortiz, M. F., Ly, A., and Grünwald, P. D. (2024). The anytime-valid logrank test: Error control under continuous monitoring with unlimited horizon. *The New England Journal of Statistics in Data Science*, 2(2):190–214.

Topsøe, F. (2007). Information theory at the service of science. In Csiszár, I., Katona, G. O. H., and Tardos, G., editors, *Entropy, Search, Complexity*, volume 16 of *Bolyai Society Mathematical Studies*, pages 179–207. János Bolyai Mathematical Society and Springer-Verlag.

Turner, R. and Grünwald, P. (2022). Safe sequential testing and effect estimation in stratified count data. In *Proceedings of the Twenty-Sixth International Conference on Artificial Intelligence and Statistics (AISTATS) 2023*, volume 206 of *Proceedings of Machine Learning Research*.

Turner, R., Ly, A., and Grünwald, P. (2024). Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *Statistical Planning and Inference*, 230.

Turner, R. J. and Grünwald, P. D. (2023). Exact anytime-valid confidence intervals for contingency tables and beyond. *Statistics & Probability Letters*, 198:109835.

van Erven, T., Grünwald, P. D., Mehta, N. A., Reid, M. D., and Williamson, R. C. (2015). Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861.

van Erven, T. and Harremoës, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

De Jong, M. (2021). Tests of significance for linear regression using E-values. Master's Thesis, Leiden University.

Ville, J. (1939). *Etude critique de la notion de collectif*. Gauthier-Villars Paris.

Vovk, V. (2002). On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 187–196. IEEE.

Vovk, V. (2004). A universal well-calibrated algorithm for on-line classification. *The Journal of Machine Learning Research*, 5:575–604.

Vovk, V. (2023). The power of forgetting in statistical hypothesis testing. In *Conformal and Probabilistic Prediction with Applications*, pages 347–366. PMLR.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Vovk, V., Nouretdinov, I., and Gammerman, A. (2003). Testing exchangeability on-line. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 768–775.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wald, A. (1947). *Sequential analysis*. John Wiley & Sons, Inc.

Wallis, W. A. (1980). The statistical research group, 1942–1945. *Journal of the American Statistical Association*, 75(370):320–330.

Wang, Q., Wang, R., and Ziegel, J. (2024). E-backtesting. ArXiv preprint arXiv:200209.00991.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society, Series B*, 84:822–852.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society, Series B*, 86(1):1–27.

Weber, F., Hoang Do, J. P., Chung, S., Beier, K. T., Bikov, M., Saffari Doost, M., and Dan, Y. (2018). Regulation of rem and non-rem sleep by periaqueductal gabaergic neurons. *Nature Communications*, 9(1):354.

Wennerholm, U.-B., Saltvedt, S., Wessberg, A., Alkmark, M., Bergh, C., Wendel, S. B., Fadl, H., Jonsson, M., Ladfors, L., Sengpiel, V., et al. (2019). Induction of labour at 41 weeks versus expectant management and induction of labour at 42 weeks (SWEdish Post-term Induction Study, swepis): multicentre, open label, randomised, superiority trial. *British Medical Journal*, 367.

Williams, D. (1991). *Probability with martingales*. Cambridge university press.

Wong, W. H. and Shen, X. S. (1995). Probability Inequalities for Likelihood Ratios and Convergence Rates of Sieve MLES. *The Annals of Statistics*, 23(2):339—362.

Young, W. H. (1912). On classes of summable functions and their Fourier series. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87(594):225–229.

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2007). Asymptotic Properties of Covariate-Adjusted Response-Adaptive Designs. *The Annals of Statistics*, 35(3):1166–1182.

Zhang, Z., Ramdas, A., and Wang, R. (2024). On the existence of powerful p-values and e-values for composite hypotheses. *The Annals of Statistics*, 52(5):2241–2267.

# Appendices

# A | Appendix to Chapter 3

## A.1 Proofs

### A.1.1 Proofs for Section 3.3

Before giving the intended results, we note that we introduced $m_P$ as the averaged Bregman divergence associated with $\gamma(x) = x - 1 - \ln(x)$. For the proof, it will be useful to also define the Bregman divergence associated with $\gamma(x) = x-1-\ln(x)$ itself, which is the so-called Itakura-Saito divergence. For $f, g \in \mathcal{M}(\Omega, \mathbb{R}_{>0})$, it is given by

$$IS_P(f,g) = \int_\Omega \left( \frac{f}{g} - 1 - \ln \frac{f}{g} \right) \mathrm{d}P.$$

By definition, it holds that

$$m_P^2(f,g) = \frac{1}{2} IS\left(f, \frac{f+g}{2}\right) + \frac{1}{2} IS\left(g, \frac{f+g}{2}\right).$$

Furthermore, for $Q \in \mathcal{C}$, we have $IS_P(q,p) = D(P\|Q)$. We now state some auxiliary results before giving the proofs for Section 3.3.

**Lemma A.1.** *For $x, y \in \mathbb{R}_{>0}$, we have*

$$|\ln(x) - \ln(y)| = g(m_\gamma^2(x,y)),$$

*where $g$ denotes the function*

$$g(t) = 2t + 2\ln\left(1 + (1 - \exp(-2t))^{1/2}\right).$$

*The function $g$ is concave and satisfies $g(t) \geq 2t$.*

*Proof.* Let $m = \frac{x+y}{2}$. Our goal is to determine the function $g$ function such that

$$|\ln(x) - \ln(y)| = g(m_\gamma^2(x, y)).$$

We first rewrite the right-hand side

$$\begin{aligned} g(m_\gamma^2(x,y)) &= g\left(\ln(m) - \frac{1}{2}\ln(x) - \frac{1}{2}\ln(y)\right) \\ &= g\left(\frac{1}{2}\ln\left(\frac{m^2}{x\cdot y}\right)\right) \\ &= g\left(\frac{1}{2}\ln\left(\frac{\left(\frac{m}{y}\right)^2}{\frac{x}{y}}\right)\right) \\ &= g\left(\frac{1}{2}\ln\left(\frac{\left(\frac{1+\frac{x}{y}}{2}\right)^2}{\frac{x}{y}}\right)\right). \end{aligned}$$

Plugging this back in and replacing $\frac{x}{y}$ by $w$ leads to

$$|\ln(w)| = g\left(\frac{1}{2}\ln\left(\frac{\left(\frac{1+w}{2}\right)^2}{w}\right)\right)$$

Then we solve the equation

$$\frac{1}{2}\ln\left(\frac{\left(\frac{1+w}{2}\right)^2}{w}\right) = t,$$

which gives

$$w = 2\exp(2t) - 1 + 2\cdot(\exp(4t) - \exp(2t))^{1/2}$$

$$\begin{aligned} g(t) &= \ln\left(2\exp(2t) - 1 + 2\cdot(\exp(4t) - \exp(2t))^{1/2}\right) \\ &= 2t + \ln\left(2 - \exp(-2t) + 2\cdot(1 - \exp(-2t))^{1/2}\right) \\ &= 2t + 2\ln\left(1 + (1 - \exp(-2t))^{1/2}\right). \end{aligned}$$

The derivatives of $g$ are

$$\begin{aligned} g'(t) &= 2 + 2\frac{\left(1-\exp\left(-2t\right)\right)^{-1/2}\exp\left(-2t\right)}{1+\left(1-\exp\left(-2t\right)\right)^{1/2}} \\ &= \frac{2}{(1-\exp(-2t))^{1/2}} \\ g''(t) &= \frac{-\exp\left(-t/2\right)}{2^{1/2}\left(\sinh t\right)^{3/2}}. \end{aligned}$$

We see that $g''(t) < 0$ and conclude that $g$ is concave. Finally, we have

$$g\left(t\right) = 2t + 2\ln\left(1+\left(1-\exp\left(-2t\right)\right)^{1/2}\right) \geq 2t,$$

because $1 - \exp\left(-2t\right) \geq 0$. □

**Lemma A.2.** *Let $(f_n)_{n\in\mathbb{N}}$ be a sequence of elements of $\mathcal{M}(\Omega, \mathbb{R}_{>0})$, then*

$$\limsup_{m,n\to\infty} m_P(f_m, f_n) = 0 \Leftrightarrow \limsup_{m,n\to\infty} \int_\Omega \left|\ln\left(\frac{f_m}{f_n}\right)\right| \,\mathrm{d}P = 0.$$

*Proof.* By Lemma A.1, we have for $m, n \in \mathbb{N}$,

$$\begin{aligned} m_P^2(f_n, f_m) &= \int_\Omega m_\gamma^2(f_n, f_m)\,\mathrm{d}P \\ &\leq \frac{1}{2}\int_\Omega \left|\ln\left(\frac{f_m}{f_n}\right)\right| \,\mathrm{d}P, \end{aligned}$$

as well as

$$\begin{aligned} \int_\Omega \left|\ln\left(\frac{f_n}{f_m}\right)\right| \,\mathrm{d}P &= \int_\Omega g(m_\gamma^2(f_n, f_m))\,\mathrm{d}P \\ &\leq g\left(\int_\Omega m_\gamma^2(f_n, f_m)\,\mathrm{d}P\right) \\ &= g\left(m_P^2\left(f_n, f_m\right)\right). \end{aligned}$$

The result then follows by continuity of $g$. □

**Lemma A.3.** *For $Q_1, Q_2 \in \mathcal{C}$ such that $P \ll Q_i$ for $i \in \{1, 2\}$, we have*

$$m_P^2(q_1, q_2) \leq \frac{D(P\|Q_1 \rightsquigarrow \mathcal{C}) + D(P\|Q_2 \rightsquigarrow \mathcal{C})}{2}.$$

*Proof.* Let $\bar{Q}$ denote the midpoint between $Q_1$ and $Q_2$. Then we have

$$\begin{aligned}
&\frac{D(P\|Q_1 \rightsquigarrow \mathcal{C}) + D(P\|Q_2 \rightsquigarrow \mathcal{C})}{2} \\
&= \frac{\sup_{Q\in\mathcal{C}} D(P\|Q_1 \rightsquigarrow Q) + \sup_{Q\in\mathcal{C}} D(P\|Q_2 \rightsquigarrow Q)}{2} \\
&\geq \frac{D(P\|Q_1 \rightsquigarrow \bar{Q}) + D(P\|Q_2 \rightsquigarrow \bar{Q})}{2} = m_P^2(q_1, q_2).
\end{aligned}$$

□

*Proof of Proposition 3.4.* This follows as a direct corollary of Lemma A.2. □

We now deviate slightly from the order of the results in Section 3.3 and first state the proof of Proposition 3.6, so that we can use its results in the proof of Theorem 3.5.

*Proof of Proposition 3.6.* The implications (3) → (2) → (1) are obvious, so we show here only the implication (1) → (3). Assume that $P'$ is a measure such that $-\infty < D(P\|P' \rightsquigarrow \mathcal{C}) < \infty$. Then there exists a sequence of measures $Q_n \in \mathcal{C}$ such that

$$D(P\|P' \rightsquigarrow Q_n) \to D(P\|P' \rightsquigarrow \mathcal{C})$$

for $n \to \infty$. Without loss of generality we may assume that $-\infty < D(P\|P' \rightsquigarrow Q_n) < \infty$ for all $n$. The result follows because

$$D(P\|P' \rightsquigarrow \mathcal{C}) = D(P\|P' \rightsquigarrow Q_n) + D(P\|Q_n \rightsquigarrow \mathcal{C})$$

and all involved quantities are finite. □

*Proof of Theorem 3.5 (1).* Let $(Q_n)_{n\in\mathcal{C}}$ denote a sequence in $\mathcal{C}$ such that

$$\lim_{n\to\infty} D(P\|Q_n \rightsquigarrow \mathcal{C}) = \inf_{Q\in\mathcal{C}} D(P\|Q \rightsquigarrow \mathcal{C}) = 0,$$

where the last equality follows from Proposition 3.6. Without loss of generality, we may assume that $D(P\|Q_n \rightsquigarrow \mathcal{C}) < \infty$ for all $n$, so that $P \ll Q_n$ for all $n$. It then follows from Lemma A.3 that for $m, n \in \mathbb{N}$ we have

$$m_P^2(q_m, q_n) \leq \frac{D(P\|Q_m \rightsquigarrow \mathcal{C}) + D(P\|Q_n \rightsquigarrow \mathcal{C})}{2}.$$

It follows that $(q_n)_{n\in\mathbb{N}}$ is a Cauchy sequence with respect to $m_P$, so that $(q_n)_{n\in\mathbb{N}}$ converges to some function $\hat{q}$ in $m_P$. The latter follows from the completeness of

$(\mathcal{M}(\Omega,(0,\infty)),m_P)$, i.e. Proposition 3.4.

Furthermore, suppose that $(Q'_n)_{n\in\mathcal{C}}$ is another sequence in $\mathcal{C}$ such that

$$\lim_{n\to\infty} D(P\|Q'_n \rightsquigarrow \mathcal{C}) = 0.$$

Then, by the same reasoning as before, $Q_1, Q'_1, Q_2, Q'_2, Q_3, Q'_3, \dots$ is also a Cauchy sequence that converges and since a Cauchy sequence can only converge to a single element this implies the desired uniqueness. □

*Proof of Theorem 3.5 (2).* The equality

$$\int_\Omega \ln \frac{p'}{\hat{q}}\,\mathrm{d}P = \lim_{n\to\infty}\int_\Omega \ln \frac{p'}{q_n}\,\mathrm{d}P$$

follows from Theorem 3.5 (1) together with the fact that convergence of $q_n$ in $m_P$ implies convergence of the logarithms in $L_1(P)$. □

*Proof of Theorem 3.5 (3).* Let $(Q_n)_{n\in\mathcal{C}}$ denote a sequence in $\mathcal{C}$ such that

$$\lim_{n\to\infty} D(P\|Q_n \rightsquigarrow \mathcal{C}) = 0.$$

Without loss of generality, we may assume that $D(P\|Q_n \rightsquigarrow \mathcal{C}) < \infty$ for all $n$ and that $q_n$ converges to $\hat{q}$ $P$-almost surely. The latter is valid, because convergence in $m_P$ implies convergence of the logarithms in $L_1(P)$ by Lemma A.2, which gives the existence of an almost surely converging sub-sequence.

Let $\tilde{Q} = (1-t)Q_1 + tQ$ for fixed $Q \in \mathcal{C}$ and fixed $0 < t < 1$. Let $Q_{n,s}$ denote the convex combination $Q_{n,s} = (1-s_n)Q_n + s_n\tilde{Q}$ and $s_n \in [0,1]$. By Theorem 3.5 (1), we know that there exists some $\hat{Q}$ such that $q_n \to \hat{q}$ in $m_P$.

Since $Q_{n,s} \in \mathcal{C}$ by convexity, we have that $D(P\|Q_n \rightsquigarrow Q_{n,s}) \le D(P\|Q_n \rightsquigarrow \mathcal{C})$. We also have

$$\begin{aligned} D(P\|Q_n \rightsquigarrow Q_{n,s}) &= s_n D(P\|Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P(\tilde{q}, q_{n,s}) \\ &\quad + (1-s_n) IS_P(q_n, q_{n,s}) \\ &\ge s_n D(P\|Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P(\tilde{q}, q_{n,s}). \end{aligned}$$

Hence

$$s_n D(P\|Q_n \rightsquigarrow \tilde{Q}) + s_n IS_P(\tilde{q}, q_{n,s}) \le D(P\|Q_n \rightsquigarrow \mathcal{C}).$$

Division by $s_n$ gives

$$D(P\|Q_n \rightsquigarrow \tilde{Q}) + IS_P(\tilde{q}, q_{n,s}) \leq \frac{D(P\|Q_n \rightsquigarrow \mathcal{C})}{s_n}.$$

Choosing $s_n = D(P\|Q_n \rightsquigarrow \mathcal{C})^{1/2}$, this gives

$$D(P\|Q_n \rightsquigarrow \tilde{Q}) + IS_P(\tilde{q}, q_{n,s}) \leq s_n^{1/2}.$$

Then we get

$$IS_P(\tilde{q}, q_{n,s}) \leq D(P\|\tilde{Q} \rightsquigarrow Q_n) + s_n^{1/2}.$$
$$\int_\Omega \left( \frac{\tilde{q}}{q_{n,s}} + \ln \frac{q_{n,s}}{q_n} \right) \mathrm{d}P \leq P(\Omega) + \tilde{Q}(\Omega) - Q_n(\Omega) + s_n^{1/2}.$$

Writing $q_n$ as $\frac{q_{n,s} - s_n \tilde{q}}{1 - s_n}$, we see

$$\begin{aligned}
\ln \frac{q_{n,s}}{q_n} &= \ln \frac{q_{n,s}}{\frac{q_{n,s} - s_n \tilde{q}}{1 - s_n}} \\
&= \ln(1 - s_n) - \ln \frac{q_{n,s} - s_n \tilde{q}}{q_{n,s}} \\
&= \ln(1 - s_n) - \ln \left( 1 - s_n \frac{\tilde{q}}{q_{n,s}} \right) \\
&\geq \ln(1 - s_n) + s_n \frac{\tilde{q}}{q_{n,s}}.
\end{aligned}$$

Hence

$$\ln(1 - s_n) + (1 + s_n) \int_\Omega \frac{\tilde{q}}{q_{n,s}} \mathrm{d}P \leq P(\Omega) + \tilde{Q}(\Omega) - Q_n(\Omega) + s_n^{1/2}.$$

As $\lim_{n\to\infty} s_n = 0$, taking the limit inferior as $n \to \infty$ on both sides gives

$$\liminf_{n\to\infty} \int_\Omega \frac{\tilde{q}}{q_{n,s}} \mathrm{d}P \leq P(\Omega) + \tilde{Q}(\Omega) - \liminf_{n\to\infty} Q_n(\Omega).$$

An application of Fatou's lemma gives

$$\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \mathrm{d}\tilde{Q} \leq P(\Omega) + \tilde{Q}(\Omega) - \liminf_{n\to\infty} Q_n(\Omega).$$

Since $\tilde{Q} = (1-t)Q_1 + tQ$ we get the inequality

$$\begin{aligned}
&\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}\left((1-t)\,Q_1 + tQ\right) \\
&\qquad \leq P(\Omega) + (1-t)Q_1(\Omega) + tQ(\Omega) - \liminf_{n\to\infty} Q_n(\Omega), \\
&(1-t)\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}Q_1 + t\int_\Omega \frac{\mathrm{d}P}{\mathrm{d}\hat{Q}} \,\mathrm{d}Q \\
&\qquad \leq P(\Omega) + (1-t)Q_1(\Omega) + tQ(\Omega) - \liminf_{n\to\infty} Q_n(\Omega).
\end{aligned}$$

Finally we let $t$ tend to one and obtain the desired result. □

*Proof of Proposition 3.7.* Let $Q \in \mathcal{C}$ arbitrarily. Then there exists a sequence $(w_i)_{i=1}^n$ in $[0,1]$ with $\sum_i w_i = 1$ such that $Q = \sum_{i=1}^n w_i Q_i$. It follows that

$$\begin{aligned}
D\left(P \| \frac{1}{n}\sum_i Q_i \rightsquigarrow Q\right) &= \int_\Omega \ln \frac{\sum_i w_i Q_i}{\frac{1}{n}\sum_i Q_i} \,\mathrm{d}P \\
&\leq \int_\Omega \ln \frac{\max_i w_i \sum_i Q_i}{\frac{1}{n}\sum_i Q_i} \,\mathrm{d}P \\
&= \ln(n) + \ln(\max_i w_i) \leq \ln(n).
\end{aligned}$$

The proposition follows by taking the supremum over $Q$ on both sides. □

*Proof of Proposition 3.8.* Since $Q^*$ is the normalized maximum likelihood distribution we have $\sup_Q \sup_\omega \ln \frac{\mathrm{d}Q}{\mathrm{d}Q^*} < \infty$. In particular

$$\begin{aligned}
\sup_{Q\in\mathcal{C}} D(P\|Q^* \rightsquigarrow Q) &= \sup_{Q\in\mathcal{C}} \int_\Omega \ln \frac{\mathrm{d}Q}{\mathrm{d}Q^*} \,\mathrm{d}P \\
&\leq \sup_{Q\in\mathcal{C}} \sup_\omega \ln \frac{\mathrm{d}Q}{\mathrm{d}Q^*}(\omega) < \infty.
\end{aligned}$$

□

*Proof of Proposition 3.10.* We can write

$$D(P\|Q_\theta \rightsquigarrow Q^*) = D(P\|Q_\theta \rightsquigarrow Q) + D(P\|Q \rightsquigarrow Q^*).$$

By assumption all terms are finite so that minimising $D(P\|Q_\theta \rightsquigarrow Q^*)$ over $\theta$ must be equivalent to minimising $D(P\|Q_\theta \rightsquigarrow Q)$ over $\theta$. The same argument holds for step 5 in Algorithm 1. The result then follows from (Brinda, 2018, Theorem 3.0.13). Whereas the algorithm described there works by choosing $\theta_k$ to minimize $\int_\Omega \log((1-\alpha_k)q_{\theta_{k-1}} +$

$\alpha_k q_\theta)\,\mathrm{d}P$, the proof relies on (Li, 1999, Lemma 5.9), which indeed uses minimization of $D(P\|(1-\alpha_k)Q_{\theta_{k-1}} + \alpha_k Q_\theta \rightsquigarrow Q)$ as described here. □

*Proof of Theorem 3.9.* For any $a \in \mathbb{R}$ we have

$$f_0(i) + a \cdot f_1(i) = f_0(i) \cdot \left(1 + a \cdot \frac{f_1(i)}{f_0(i)}\right). \tag{A.1}$$

Since $\frac{f_1(i)}{f_0(i)} \to 0$ for $i \to \infty$ we have that $f_0(i) + a \cdot f_1(i) \geq 0$ for $i$ sufficiently large. Therefore, we can apply Fatou's lemma to the function and obtain

$$\begin{aligned}
&\sum f_0(i) \cdot q^*(i) + a \cdot \sum f_1(i) \cdot q^*(i) \\
&\quad = \sum \left(f_0(i) + a \cdot f_1(i)\right) \cdot q^*(i) \\
&\quad = \sum \liminf_{n\to\infty} \left(f_0(i) + a \cdot f_1(i)\right) \cdot q_n(i) \\
&\quad \leq \liminf_{n\to\infty} \sum_i \left(f_0(i) + a \cdot f_1(i)\right) \cdot q_n(i) \\
&\quad = \liminf_{n\to\infty} \left(\sum_i f_0(i) \cdot q_n(i) + a \cdot \sum_i f_1(i) \cdot q_n(i)\right) \\
&\quad = \liminf_{n\to\infty} \left(\lambda_0 + a \cdot \lambda_1\right) = \lambda_0 + a \cdot \lambda_1.
\end{aligned}$$

Hence

$$a \cdot \left(\sum f_1(i) \cdot q^*(i) - \lambda_1\right) \leq \lambda_0 - \sum f_0(i) \cdot q^*(i). \tag{A.2}$$

This inequality should hold for all $a \in \mathbb{R}$, which is only possible if

$$\begin{aligned}
\sum f_1(i) \cdot q^*(i) - \lambda_1 &= 0. \\
\sum f_1(i) \cdot q^*(i) &= \lambda_1.
\end{aligned}$$

□

### A.1.2 Proofs for Section 3.4

*Proof of Proposition 3.12.* Assume that $E_1, E_2, E_3, \ldots$ is a sequence of $e$-variables such that

$$\int_\Omega \ln\left(\frac{E_n}{E'}\right) \mathrm{d}P \to \sup_E \int_\Omega \ln\left(\frac{E}{E'}\right) \mathrm{d}P$$

for $n \to \infty$. Then $E_{n,m} = (E_m + E_n)/2$ are also $e$-variables and by convexity

$$\int_\Omega \ln\left(\frac{E_{m,n}}{E'}\right) \mathrm{d}P \to \sup_E \int_\Omega \ln\left(\frac{E}{E'}\right) \mathrm{d}P\,,$$

which implies that $m_\gamma^2(E_m, E_n) \to 0$ for $m, n \to \infty$. By completeness $E_n$ converges to some $e$-variable $E_\infty$. Using Lemma A.2 we see that $m_\gamma(E_n, E_\infty) \to 0$ implies that

$$\int_\Omega \ln\left(\frac{E_m}{E'}\right) \mathrm{d}P \to \int_\Omega \ln\left(\frac{E_\infty}{E'}\right) \mathrm{d}P$$

so that

$$\sup_E \int_\Omega \ln\left(\frac{E}{E'}\right) \mathrm{d}P = \int_\Omega \ln\left(\frac{E_\infty}{E'}\right) \mathrm{d}P\,.$$

Hence

$$\sup_E \int_\Omega \ln\left(\frac{E}{E_\infty}\right) \mathrm{d}P = 0$$

Therefore $E_\infty$ is a strongest $e$-statistic.

Assume that both $E_1$ and $E_2$ are strongest $e$-variables. Then they are both stronger than the average $\bar{E} = (E_1 + E_2)/2$. Hence

$$0 \le m_\gamma^2(E_1, E_2) = \frac{1}{2}\int \left(\ln\left(\frac{\bar{E}}{E_1}\right) + \ln\left(\frac{\bar{E}}{E_2}\right)\right) \mathrm{d}P \le 0.$$

Therefore $E_1 = E_2$ $P$-almost surely. □

*Proof of Theorem 3.13.* Firstly, since $\hat{E} > 0$ holds $P$-almost surely, we have that $\hat{E}$ is stronger than any $E' \in \mathcal{E}_\mathcal{C}$ with $P(E' = 0) > 0$.

Secondly, let $E \in \mathcal{E}_\mathcal{C}$ be an $e$-statistic for which $E > 0$ holds $P$-almost surely. Furthermore, let $Q_n$ be a sequence of measures in $\mathcal{C}$ such that $D(P\|Q_n \rightsquigarrow \mathcal{C}) \to 0$. We can define a sequence of sub-probability measures $R_n$ by $R_n(F) = \int_F E \,\mathrm{d}Q_n$, which satisfies $\mathrm{d}R_n/\mathrm{d}Q_n = E$. We see

$$\begin{aligned}\int_\Omega \ln\left(\frac{\hat{E}}{E}\right) \mathrm{d}P &= \int_\Omega \ln\left(\frac{\mathrm{d}Q_n}{\mathrm{d}\hat{Q}}\right) \mathrm{d}P + D(P\|R_n) \\ &\quad + (P(\Omega) - R_n(\Omega)) \\ &\ge \int_\Omega \ln\left(\frac{\mathrm{d}Q_n}{\mathrm{d}\hat{Q}}\right) \mathrm{d}P.\end{aligned}$$

The last expression goes to zero as $n \to \infty$, so we see that $\hat{E}$ is stronger than $E$. □

*Proof of Proposition 3.14.* Using the fact that $\ln(x) \leq x - 1$ for $x > 0$, we see

$$\begin{aligned} D(P\|Q^* \rightsquigarrow Q) &= \int_\Omega \ln \frac{\mathrm{d}Q}{\mathrm{d}Q^*} \,\mathrm{d}P \\ &\leq \int_\Omega \left( \frac{\mathrm{d}Q}{\mathrm{d}Q^*} - 1 \right) \mathrm{d}P \\ &= \int_\Omega \frac{\mathrm{d}P}{\mathrm{d}Q^*} \,\mathrm{d}Q - 1 \leq 0, \end{aligned}$$

where the last inequality follows from the fact that $\mathrm{d}P/\mathrm{d}Q^*$ is an $e$-statistic. □

*Proof of Theorem 3.16.* Without loss of generality, assume that $\int_\Omega q'/q \,\mathrm{d}P = 1 + \epsilon$ for some $\epsilon > 0$. For the sake of brevity, we write $c_\beta := \|q'/q\|_{1+\beta}^{1+\beta}$. We now define a function $g : [0,1] \to \mathbb{R}_{\geq 0}$ as

$$g(\alpha) := D\left(P\|(1-\alpha)Q + \alpha Q' \rightsquigarrow \mathcal{C}\right).$$

Notice that $g(0) = \delta$ and $g(\alpha) \geq 0$, since $(1-\alpha)Q + \alpha Q' \in \mathcal{C}$. This function and its derivatives will guide the rest of the proofs, and we now list some properties that we will need:

$$g'(\alpha) := \frac{\mathrm{d}}{\mathrm{d}\alpha} g(\alpha) = \int_\Omega \frac{q - q'}{(1-\alpha)q + \alpha q'} \,\mathrm{d}P, \tag{A.3}$$

so that

$$g'(0) = \int_\Omega \left(1 - \frac{q'}{q}\right) \mathrm{d}P = -\epsilon, \tag{A.4}$$

$$g''(\alpha) := \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2} g(\alpha) = \int_\Omega \left( \frac{q' - q}{(1-\alpha)q + \alpha q'} \right)^2 \mathrm{d}P, \tag{A.5}$$

so that

$$g''(0) = \int_\Omega \left(1 - \frac{q'}{q}\right)^2 \mathrm{d}P = 1 - 2(1+\epsilon) + c_1$$

and

$$0 \leq g''(\alpha) \leq \frac{1}{(1-\alpha)^2} g''(0). \tag{A.6}$$

We now prove (3.10). We start with the case $\beta = 1$ and will use the result for $\beta = 1$ to prove the case for $\beta < 1$. The proof for the case $\beta > 1$ comes later; it requires a completely different proof.

Case $\beta = 1$. The general idea is simple: at $\alpha = 0$ the function $g(\alpha)$ is equal to $\delta$ and has derivative $-\epsilon$. Its second derivative is positive and bounded by constant times $g''(0) \leq c_1$ for all $\alpha \leq 1/2$. Thus, if $\epsilon$ is larger then a certain threshold, $g(\alpha)$ will become negative at some $\alpha \leq 1/2$, but this is not possible since $g$ is a description gain and we would arrive at a contradiction. The details to follow simply amount to calculating the threshold as a function of $\delta$.

By Taylor's theorem, we have for any $\alpha \in [0, 1/2]$ that

$$\begin{aligned} g(\alpha) &= g(0) + g'(0)\alpha + \max_{0 \leq \alpha^\circ \leq \alpha} \frac{g''(\alpha^\circ)}{2}\alpha^2 \\ &\leq g(0) + g'(0)\alpha + 2g''(0)\alpha^2 \\ &\leq \delta - \epsilon\alpha + 2\alpha^2 c_1, \end{aligned}$$

where we use the properties derived above. This final expression has a minimum in $\alpha^* = \min\{\epsilon/4c_1, 1/2\}$. By nonnegativity of $g$, we know that $\delta - \epsilon\alpha^* + 2\alpha^{*2}c_1 \geq 0$. This gives $\epsilon \leq (8c_1\delta)^{1/2}$ in the case that $\alpha^* = \epsilon/4c_1 < 1/2$, and $\epsilon \leq 2\delta + c_1$ otherwise. In the latter case, it holds that $c_1 < \epsilon/2$, so the bound can be loosened slightly to find the simplification $\epsilon \leq 4\delta$. This concludes the proof for $\beta = 1$, which we now use to prove Case $\beta < 1$.

Case $\beta < 1$. For any $a > 0$, it holds that

$$\int_\Omega \frac{q'}{q}\,\mathrm{d}P = \int_\Omega \frac{q'}{q}\mathbf{1}_{\{q'/q \leq a\}}\,\mathrm{d}P + \int_\Omega \frac{q'}{q}\mathbf{1}_{\{q'/q > a\}}\,\mathrm{d}P. \tag{A.7}$$

We write $q'' := q'\mathbf{1}_{\{q'/q \leq a\}}$ and we will bound the first term on the right-hand side of (A.7) using the proof above with $Q'$ replaced by $Q''$. Since $Q''$ is not necessarily an element of $\mathcal{C}$, we need to verify nonnegativity, which follows because for each $\alpha \in (0,1)$, we have that $D(P\|(1-\alpha)Q + \alpha Q'' \rightsquigarrow \mathcal{C}) \geq D(P\|(1-\alpha)Q + \alpha Q' \rightsquigarrow \mathcal{C}) \geq 0$.

Furthermore, it holds that

$$\begin{aligned}\left\|\frac{q''}{q}\right\|_2^2 &= \int_\Omega \left(\frac{q''}{q}\right)^2 \mathrm{d}P \\ &= \int_\Omega \left(\frac{q''}{q}\right)^{1+\beta} \left(\frac{q''}{q}\right)^{1-\beta} \mathrm{d}P \\ &\leq a^{1-\beta} c_\beta\end{aligned}$$

The results above therefore give

$$\int_\Omega \frac{q''}{q} \mathrm{d}P \leq 1 + \max\{(8a^{1-\beta}c_\beta\delta)^{1/2}, 2\delta\}.$$

For the second term on the right-hand side of (A.7), we use a Markov-type bound, i.e.

$$\begin{aligned}\int_\Omega \frac{q'}{q} \mathbf{1}_{\{q'/q > a\}} \mathrm{d}P &\leq \int_\Omega \frac{q'}{q} \left(\frac{q'/q}{a}\right)^\beta \mathbf{1}_{\{q'/q > a\}} \mathrm{d}P \\ &\leq a^{-\beta} c_\beta.\end{aligned}$$

Putting this together gives

$$\int_\Omega \frac{q'}{q} \mathrm{d}P \leq 1 + \max\{(8a^{1-\beta}c_\beta\delta)^{1/2}, 4\delta\} + a^{-\beta}c_\beta.$$

Since this holds for any $a$, we now pick it to minimize this bound. To this end, consider

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}a}(8a^{1-\beta}c_\beta\delta)^{1/2} + a^{-\beta}c_\beta \\ &\qquad = \frac{(1-\beta)(8c_\beta\delta)^{1/2}}{2} a^{-(1+\beta)/2} - \beta a^{-(1+\beta)} c_\beta.\end{aligned}$$

Setting this to zero, we find

$$a^* = \left(\frac{\beta c_\beta^{1/2}}{(1-\beta)(2\delta)^{1/2}}\right)^{\frac{2}{1+\beta}}.$$

The proof is concluded by noting that

$$(8a^{*1-\beta}c_\beta\delta)^{1/2} = \left(8\left(\frac{\beta c_\beta^{1/2}}{(1-\beta)(2\delta)^{1/2}}\right)^{2\frac{1-\beta}{1+\beta}} c_\beta\delta\right)^{1/2}$$
$$= 2c_\beta^{1/(\beta+1)}(2\delta)^{\beta/(\beta+1)}\left(\frac{\beta}{1-\beta}\right)^{\frac{1-\beta}{1+\beta}}$$

and

$$a^{*-\beta}c_\beta = \left(\frac{\beta c_\beta^{1/2}}{(1-\beta)(2\delta)^{1/2}}\right)^{\frac{-2\beta}{1+\beta}} c_\beta$$
$$= c_\beta^{1/(\beta+1)}\left(\frac{\beta}{1-\beta}\right)^{\frac{-2\beta}{1+\beta}}(2\delta)^{\beta/(1+\beta)}.$$

Case $\beta > 1$. We now prove the result for $\beta \in (1,\infty)$; the proof for $\beta = \infty$ follows by a minor modification of (A.9). If $\epsilon \leq 0$ there is nothing to prove, so without loss of generality we can write $\epsilon = \gamma\delta$ for some $\gamma > 0$; we will bound $\gamma$. Whereas the previous proof exploited the fact that the second derivative $g''(\alpha)$ was bounded above in terms of $\delta$ and hence 'not too large', the proof below uses the condition that $c_\beta$ is finite to show first, (a), that $g''(\alpha)$ can also be bounded *below* in terms of $(\gamma,\delta)$. Therefore, if $\epsilon$ exceeds a certain threshold, as $\alpha$ moves away from the $\alpha^*$ at which $g(\alpha)$ achieves its minimum in the direction of the furthest boundary point (i.e. if $\alpha^* < 1/2$, we consider $\alpha \uparrow 1$, if $\alpha^* \geq 1/2$ we consider $\alpha \downarrow 0$), $g(\alpha)$ will become larger than $K\delta$ or $\delta$ respectively, and we arrive at a contradiction. (b) below gives the detailed calculation of this threshold.

*Proof of (a).* Fix some $0 \leq \tilde{\alpha} < 1$ (we will derive a bound for any such $\tilde{\alpha}$ and later optimize for $\tilde{\alpha}$; for a sub-optimal yet easier derivation take $\tilde{\alpha} = 1/2$). By Taylor's theorem, we have $0 \leq g(\tilde{\alpha}) = \delta - \tilde{\alpha}\epsilon + (1/2)\tilde{\alpha}^2 g''(\alpha^\circ)$ for some $0 \leq \alpha^\circ \leq \tilde{\alpha}$. Plugging in $\epsilon = \gamma\delta$ we find that

$$g''(\alpha^\circ) \geq \frac{2}{\tilde{\alpha}^2}(\tilde{\alpha}\gamma - 1)\delta.$$

This gives a lower bound on $g''(\alpha^\circ)$ for *some* $\alpha^\circ$ in terms of $(\gamma,\delta)$. We now turn this into a weaker lower bound on *all* $\alpha$. First, using (A.6) and then $\alpha^\circ \leq \tilde{\alpha}$ and then the

above lower bound, we find

$$\begin{aligned} g''(0) &\geq \max_{\alpha\in[0,\tilde{\alpha}]} (1-\alpha)^2 g''(\alpha) \geq (1-\alpha^\circ)^2 g''(\alpha^\circ) \\ &\geq (1-\tilde{\alpha})^2 g''(\alpha^\circ) \geq 2 f_{\tilde{\alpha}}(\gamma,\delta), \end{aligned} \tag{A.8}$$

where $f_{\tilde{\alpha}}(\gamma,\delta) := ((1-\tilde{\alpha})/\tilde{\alpha})^2(\tilde{\alpha}\gamma-1)\delta$ is a function that is linear in $\gamma$ and $\delta$. We have now lower bounded $g''(0)$ in terms of $\gamma,\delta$. We next show that, under our condition that $c_\beta < \infty$, this implies a (weaker) lower bound on $g''(\alpha)$ for all $\alpha$. For this, fix any $C > 1$. We have for all $0 < \alpha \leq 1$:

$$\begin{aligned} g''(\alpha) &\geq \int_\Omega \mathbf{1}_{q'\leq Cq} \cdot \left(\frac{q'-q}{(1-\alpha)q+\alpha q'}\right)^2 \mathrm{d}P \\ &\geq \int_\Omega \mathbf{1}_{q'\leq Cq} \cdot \left(\frac{q'-q}{(1-\alpha)q+\alpha Cq}\right)^2 \mathrm{d}P \\ &= \int_\Omega \mathbf{1}_{q'\leq Cq} \cdot \left(\frac{q'-q}{q}\right)^2 \mathrm{d}P \cdot \frac{1}{(1+\alpha(C-1))^2} \\ &\geq \frac{1}{(1+(C-1))^2}\left(g''(0) - \int_\Omega \mathbf{1}_{q'>Cq}\left(\frac{q'}{q}-1\right)^2 \mathrm{d}P\right) \\ &\geq \frac{1}{C^2}\left(2f_{\tilde{\alpha}}(\gamma,\delta) - C^{1-\beta}c_\beta\right), \end{aligned} \tag{A.9}$$

where in the fourth line we used the definition of $g''(0)$, and in the fifth line we used (A.8) and a Markov-type bound on the integral, i.e. we used that $\int_\Omega \mathbf{1}_{q'>Cq} \cdot (q'/q-1)^2 \,\mathrm{d}P$ is bounded by

$$\begin{aligned} \int_\Omega \mathbf{1}_{q'>Cq} \cdot \left(\frac{q'}{q}\right)^2 \mathrm{d}P &\leq \int_\Omega \left(\frac{q'/q}{C}\right)^{\beta-1} \cdot \left(\frac{q'}{q}\right)^2 \mathrm{d}P \\ &= C^{1-\beta}c_\beta. \end{aligned}$$

By differentiation we can determine the $C$ that maximizes the bound (A.9). This gives $C^{1-\beta} = f_{\tilde{\alpha}}(\gamma,\delta)(4/c_\beta(1+\beta))$. and with this choice of $C$, (A.9) becomes

$$g''(\alpha) \geq f_{\tilde{\alpha}}(\gamma,\delta)^{(\beta+1)/(\beta-1)} c_\beta^{2/(1-\beta)} h(\beta) \tag{A.10}$$

where $h(\beta) = (4/(1+\beta))^{2/(\beta-1)} \cdot 2(\beta-1)/(1+\beta)$. We are now ready to continue to:

*Proof of (b).* Let $\alpha^* \in [0,1]$ be the point at which $g(\alpha)$ achieves its minimum. If

$\alpha^* \leq 1/2$, a second-order Taylor approximation of $g(1)$ around $\alpha^*$ gives that

$$\begin{aligned} K\delta \geq g(1) &\geq \frac{1}{2}(1-\alpha^*)^2 \min_{\alpha \in [\alpha^*, 1]} g''(\alpha) \\ &\geq \frac{1}{8} f_{\tilde{\alpha}}(\gamma, \delta)^{(\beta+1)/(\beta-1)} c_\beta^{2/(1-\beta)} h(\beta), \end{aligned}$$

so that after some manipulations

$$f_{\tilde{\alpha}}(\gamma, \delta)^{(1+\beta)/(\beta-1)} \leq 8K' c_\beta^{2/(\beta-1)} \cdot h(\beta)^{-1} \delta, \tag{A.11}$$

with $K' = K$. If $\alpha^* > 1/2$, we perform a completely analogous second-order Taylor approximation of $g(0)$ around $\alpha^*$, which will then give (A.11) again but with $K'$ replaced by 1. We thus always have (A.11) with $K' = \max\{K, 1\}$. Unpacking $f_{\tilde{\alpha}}$ in (A.11) and rearranging gives:

$$\gamma \leq \frac{\tilde{\alpha}}{(1-\tilde{\alpha})^2} \cdot V + \frac{1}{\tilde{\alpha}}$$

with

$$V = c_\beta^{2/(1+\beta)} \cdot \left( \frac{8K'}{h(\beta)} \right)^{\frac{\beta-1}{1+\beta}} \delta^{\frac{-2}{1+\beta}}.$$

We now pick the $\tilde{\alpha}$ that makes both terms on the right equal, so that the right-hand side becomes equal to $2/\tilde{\alpha}$. This is the solution to the equation $(\tilde{\alpha}/(1-\tilde{\alpha}))^2 V = 1$ which must clearly be obtained for some $0 < \tilde{\alpha} < 1$, so this $\tilde{\alpha}$ satisfies our assumptions. Basic calculation gives

$$\gamma \leq \frac{2}{\tilde{\alpha}} = 2 \cdot \left( V^{1/2} + 1 \right)$$

and unpacking $V$ we obtain

$$\epsilon = \gamma\delta \leq c^* \cdot \delta^{\frac{\beta}{1+\beta}} + 2\delta.$$

where

$$c^* = c_\beta^{1/(1+\beta)} \cdot \left( \frac{8K'}{h(\beta)} \right)^{\frac{\beta-1}{2(1+\beta)}}.$$

Unpacking $h(\beta)$ gives the desired result.

□

## A.2 RIPr Strict Sub-Probability Measure

In this appendix, we discuss a general way to construct a measure $P$ and convex set of distributions $\mathcal{C}$ such that the reverse information projection of $P$ on $\mathcal{C}$ is a strict sub-probability measure. For simplicity, we take $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\mathbb{N}}$, though the idea should easily translate to more general settings.

**Proposition A.4.** *Let $g : \mathbb{N} \to \mathbb{R}_{>0}$ be a function, and let $\mathcal{C}$ denote the set of measures $\{Q : \sum_i g(i)\, q(i) \leq \nu\}$ for some $\nu > 0$. Then for any $P$ that is not in $\mathcal{C}$ we have that $E(i) = g(i)/\nu$ is the optimal e-statistic.*

*Proof.* The extreme points in $\mathcal{C}$ are the measure with total mass 0 and measures of the form $\frac{\nu}{g(i)}\delta_i$, i.e. measures concentrated in single points. An $e$-statistic $E$ must satisfy

$$\sum_j E(j)\, \frac{\nu}{g(i)} \delta_i(j) \leq 1$$

or, equivalently, $E(i)\, \frac{\nu}{g(i)} \leq 1$. Hence $E \leq g/\nu$ so the optimal $e$-statistic is $g/\nu$. □

Let $g : \mathbb{N} \to \mathbb{R}_{>0}$ be any function that satisfies

$$\lim_{n\to\infty} g(n) = 0.$$

Furthermore, let $P$ denote a probability measure on the natural numbers such that

$$\sum_i \frac{p(i)}{g(i)} = c$$

for some $c \in \mathbb{R}_{>0}$. Fix $\nu^* \in (0, 1/c)$ and let $\mathcal{C}_{\nu^*}$ denote the set of measures $\{Q : \sum_i g(i)\, q(i) \leq \nu^*\}$. Note that we do not yet require all measures in $\mathcal{C}_{\nu^*}$ to be probability measures so that the set $\mathcal{C}_{\nu^*}$ is compact. It follows that there exists a unique element of $\mathcal{C}_{\nu^*}$ that minimizes $\sum_i p(i) \ln(p(i)/q(i))$.

The optimal $e$-statistic is $E_{\nu^*} = g/\nu^*$, and we may define the measure $Q_{\nu^*}$ by

$$q_{\nu^*}(i) = \frac{p(i)}{E_{\nu^*}(i)} = \nu^* p(i)/g(i),$$

and we can check that $Q_{\nu^*} \in \mathcal{C}_{\nu^*}$. Hence $Q_{\nu^*}$ minimizes $\sum_i p(i) \ln(p(i)/q(i))$.

This is a strict sub-probability measure:

$$\begin{aligned}\sum_i q_{\nu^*}(i) &= \nu^* \sum_i \frac{p(i)}{g(i)} \\ &= \nu^* c \\ &< 1,\end{aligned}$$

where we use that $\nu^* < 1/c$.

The next step is to prove that the information projection does not change if we restrict to the set of probability measures in $\mathcal{C}_{\nu^*}$, which we denote by $\tilde{\mathcal{C}}_{\nu^*}$. To this end, note first that for $\nu < \nu^*$, we have that $\sum g(i)\, q_\nu(i) < \nu^*$, so that for all $\nu < \nu^*$ there exists $n_\nu \in \mathbb{N}$ such that the probability measure defined by

$$q_\nu(i) + \left(1 - \sum_j q_\nu(j)\right) \delta_{n_\nu}(i)$$

is an element of $\tilde{\mathcal{C}}_{\nu^*}$. Hence

$$\begin{aligned}D(P\|\tilde{C}_{\nu^*}) &\leq D\left(P \middle\| Q_\nu + \left(1 - \sum_{j\in\mathbb{N}} q_\nu(j)\right)\delta_{n_\nu}\right) \\ &= \sum_{i\in\mathbb{N}} p(i) \ln\left(\frac{p(i)}{Q_\nu(i) + \left(1 - \sum_{j\in\mathbb{N}} q_\nu(j)\right)\delta_{n_\nu}(i)}\right) \\ &= -p(n_\nu)\ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) \\ &\quad + p(n_\nu)\ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j\in\mathbb{N}} q_\nu(j)}\right) \\ &\quad + \sum_{i\in\mathbb{N}} p(i)\ln\left(\frac{p(i)}{q_\nu(i)}\right).\end{aligned}$$

The first term can be written as

$$\begin{aligned}p(n_\nu)\ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) &= q_\nu(n_\nu)\frac{p(n_\nu)}{q_\nu(n_\nu)}\ln\left(\frac{p(n_\nu)}{q_\nu(n_\nu)}\right) \\ &= q_\nu(n_\nu)\frac{g(n_\nu)}{\nu}\ln\left(\frac{g(n_\nu)}{\nu}\right)\end{aligned}$$

Then notice that for $\nu \to \nu^*$, we must have that $n_\nu \to \infty$. Using that $c\ln(c) \to 0$ for

$c \to 0$ we see the first term tends to 0 for $\nu \to \nu^*$. Similarly, the second term can be written as

$$\begin{aligned} & p(n_\nu) \ln \left( \frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)} \right) \\ & = \left( q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j) \right) \frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)} \\ & \quad \cdot \ln \left( \frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_{j \in \mathbb{N}} q_\nu(j)} \right). \end{aligned}$$

We also have

$$\frac{p(n_\nu)}{q_\nu(n_\nu) + 1 - \sum_i q_\nu(i)} \to 0$$

for $\nu \to \nu^*$ and using that $c \ln(c) \to 0$ for $c \to 0$ we get the second term tends to 0 for $\nu \to \nu^*$. Therefore we see

$$\begin{aligned} D(P \| \tilde{\mathcal{C}}_{\nu^*}) & \leq \lim_{\nu \to \nu^*} D\left( P \middle\| Q_\nu + \left( 1 - \sum_i q_\nu(i) \right) \delta_{n_\nu} \right) \\ & \leq \sum_i p(i) \ln \left( \frac{p(i)}{q_{\nu^*}(i)} \right) \\ & = \inf_{Q \in \mathcal{C}_{\nu^*}} \sum_i p(i) \ln \left( \frac{p(i)}{q(i)} \right). \end{aligned}$$

The inequality trivially also holds the other way around, so we find that

$$D(P \| \tilde{\mathcal{C}}_{\nu^*}) = \inf_{Q \in \mathcal{C}_{\nu^*}} \sum_i p(i) \ln \left( \frac{p(i)}{q(i)} \right).$$

It follows that $Q_{\nu^*}$ is a strict sub-probability measure, and at the same time it is the reverse information projection of $P$ onto $\tilde{\mathcal{C}}_{\nu^*}$.

## A.3 Convexity

One of the main assumptions made throughout the main text is that the set of measures $\mathcal{C}$ is convex, i.e. closed under finite mixtures. However, one can also consider stronger notions of convexity, such as $\sigma$-convexity and Choquet-convexity. In this appendix, we investigate whether considering different levels of convexity can change the reverse

information projection.

**Definition A.5.** A set $\mathcal{C}'$ of measures is said to be *$\sigma$-convex* if $Q_1, Q_2, Q_3 \cdots \in \mathcal{C}'$ implies that $\sum_{i=1}^{\infty} w_i Q_i \in \mathcal{C}'$ when $w_i \geq 0$ and $\sum_{i=1}^{\infty} w_i = 1$. The $\sigma$-convex hull of a set of measures $\mathcal{C}$, denoted by $\sigma\text{-conv}(\mathcal{C})$, is the smallest $\sigma$-convex set containing $\mathcal{C}$.

In order to avoid topological complications we will restrict the discussion of Choquet-convexity to Polish spaces, i.e. spaces for which there exists a complete metric that generates the topology. That is, assume that $\Omega$ is a Polish space equipped with the Borel $\sigma$-algebra. Let $\Theta$ be another Polish space and let $\{Q_\theta : \theta \in \Theta\}$ denote a parameterized set of probability measures on $\Omega$ such that $\theta \to \int_\Omega f \, \mathrm{d}Q_\theta$ is Borel measurable for any measurable function $f : \Omega \to \mathbb{R}$. Then for any probability measure $\nu$ on $\Theta$ the *Choquet-convex mixture* $\mu_\nu$ can be defined by

$$\int_\Omega f \, \mathrm{d}\mu_\nu = \int_\Theta \left( \int_\Omega f \, \mathrm{d}\mu_\theta \right) \mathrm{d}\nu,$$

for any measurable function $f : \Omega \to \mathbb{R}$.

**Definition A.6.** A set $\mathcal{C}'$ of measures is said to be *Choquet-convex* if it is closed under Choquet convex mixtures. The Choquet-convex hull of a set of measures $\mathcal{C}$ is the smallest Choquet-convex set that contains $\mathcal{C}$.

So far, we have assumed that all of the measures in $\mathcal{C}$ are finite. However, a countable or Choquet convex mixture of finite measures may not be finite. It follows that our results on the existence of the RIPr might not be applicable to the $\sigma$-convex and Choquet-convex hull of $\mathcal{C}$. We therefore assume for the remainder of this section that all involved measures are sub-probability measures, in which case this problem does not arise. With all of this in place, it is relatively straightforward to construct examples where the RIPr of $P$ on a convex set does not exist, whereas the RIPr of $P$ on its $\sigma$-convex hull does exist.

**Example A.1.** Let $P$ denote a geometric distribution on $\mathbb{N}_0$ and let $\mathcal{C}$ denote the set of probability measures on $\mathbb{N}_0$ with finite support. Then $D(P\|Q \rightsquigarrow \mathcal{C}) = -\infty$ for any $Q \in \mathcal{C}$. Therefore the reverse information projection of $P$ on $\mathcal{C}$ is not defined according to the definitions given in Chapter 3. However, the $\sigma$-convex hull of $\mathcal{C}$ consists of all probability measures on $\mathbb{N}_0$, which implies that the reverse information projection on the $\sigma$-convex hull is well-defined and equals $P$.

However, as the following results show, if the RIPr of $P$ on $\mathcal{C}$ does exist, then it must coincide with the RIPr of $P$ on $\sigma\text{-conv}(\mathcal{C})$.

**Lemma A.7.** *Let $P$ and $Q$ be sub-probability measures and let $Q_1, Q_2, \ldots$ be a sequence of sub-probability measures such that $D(P\|Q \rightsquigarrow Q_1) > -\infty$, and let $w_1, w_2, \ldots$ be a sequence of positive numbers with sum 1. Then*

$$D\left(P \middle\| Q \rightsquigarrow \frac{\sum_{i=1}^{n} w_i \cdot Q_i}{\sum_{i=1}^{n} w_i}\right) \to D\left(P \middle\| Q \rightsquigarrow \sum_{i=1}^{\infty} w_i \cdot Q_i\right)$$

*for $n \to \infty$.*

*Proof.* Firstly, note that

$$\ln \frac{\mathrm{d}\sum_{i=1}^{n+1} w_i Q_i}{\mathrm{d}Q} \geq \ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q}$$

and

$$\begin{aligned}\int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q}\,\mathrm{d}P &\geq \int_\Omega \ln \frac{\mathrm{d}w_1 Q_1}{\mathrm{d}Q}\,\mathrm{d}P \\ &= D(P\|Q \rightsquigarrow Q_1) + \ln w_1 \\ &\quad + (Q_1(\Omega) - Q(\Omega)) \\ &> -\infty.\end{aligned}$$

Since $\sum_{i=1}^{n} w_i q_i \to \sum_{i=1}^{\infty} w_i q_i$ pointwise, applying the monotone convergence theorem to the sequence

$$\left(\ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q} - \ln \frac{\mathrm{d}w_1 Q_1}{\mathrm{d}Q}\right)_{n\in\mathbb{N}}$$

gives that

$$\begin{aligned}\int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q} - \ln \frac{\mathrm{d}w_1 Q_1}{\mathrm{d}Q}\,\mathrm{d}P \\ \to \int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{\infty} w_i Q_i}{\mathrm{d}Q} - \ln \frac{\mathrm{d}w_1 Q_1}{\mathrm{d}Q}\,\mathrm{d}P.\end{aligned}$$

We get

$$\int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{n} w_i Q_i}{\mathrm{d}Q}\,\mathrm{d}P \to \int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^{\infty} w_i Q_i}{\mathrm{d}Q}\,\mathrm{d}P$$

for $n \to \infty$. Finally, we see that

$$\begin{aligned} & D\left(P \,\middle\|\, Q \rightsquigarrow \frac{\sum_{i=1}^n w_i \cdot Q_i}{\sum_{i=1}^n w_i}\right) \\ &= \int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^n w_i Q_i}{\mathrm{d}Q} \,\mathrm{d}P - (Q_n(\Omega) - Q(\Omega)) - \ln \sum_{i=1}^n w_i \\ &\to \int_\Omega \ln \frac{\mathrm{d}\sum_{i=1}^\infty w_i Q_i}{\mathrm{d}Q} \,\mathrm{d}P - (Q_\infty(\Omega) - Q(\Omega)) \\ &= D(P\|Q \rightsquigarrow Q_\infty), \end{aligned}$$

where $Q_\infty := \sum_{i=1}^\infty w_i Q_i$ and we use that $\ln \sum_{i=1}^n w_i \to 0$ and $Q_n(\Omega) \to Q_\infty(\Omega)$. To see the latter, note that

$$Q_n(\Omega) = \int_\Omega \frac{\sum_{i=1}^n q_i(\omega) w_i}{\sum_{i=1}^n w_i} \,\mathrm{d}\mu(\omega),$$

and $0 \le \sum_{i=1}^n q_i(\omega) w_i / \sum_{i=1}^n w_i \le q_\infty(\omega)/w_1$, where the RHS integrates, so that the desired convergence follows from the dominated convergence theorem. □

**Theorem A.8.** *Let $P$ be a finite measure and $\mathcal{C}$ a convex set of sub-probability measures such that $D(P\|Q \rightsquigarrow \mathcal{C}) = 0$. If $Q_1, Q_2, \ldots$ is a sequence of measures in $\mathcal{C}$ such that $D(P\|Q_n \rightsquigarrow \mathcal{C}) \to 0$, then $D(P\|Q_n \rightsquigarrow \sigma\text{-conv}(\mathcal{C})) \to 0$.*

*Proof.* Fix $Q^* \in \mathcal{C}$ such that $D(P\|Q^* \rightsquigarrow \mathcal{C}) \le \varepsilon$ and let $\bar{Q} = \sum_{i=1}^\infty w_i Q_i \in \sigma\text{-conv}(\mathcal{C})$ arbitrarily. Let $s \in (0,1)$ and consider $\tilde{Q} := s \cdot Q^* + (1-s) \cdot \bar{Q} = \sum_{i=0}^\infty \tilde{w}_i Q_i$, where $Q_0 := Q^*$, $\tilde{w}_0 = s$ and $\tilde{w}_i = (1-s) \cdot w_i$ for $i = 1, 2, \ldots$. Note that $D(P\|Q^* \rightsquigarrow Q_0) = 0$, so it follows from Lemma A.7 that

$$\lim_{n\to\infty} D\left(P \,\middle\|\, Q^* \rightsquigarrow \frac{\sum_{i=0}^n \tilde{w}_i Q_i}{\sum_{i=0}^n \tilde{w}_i}\right) = D(P\|Q^* \rightsquigarrow \tilde{Q}).$$

The left hand side is, by definition of $Q^*$, bounded by $\varepsilon$ since $\sum_{i=0}^n \tilde{w}_i Q_i / \sum_{i=0}^n \tilde{w}_i \in \mathcal{C}$, so that we find $D(P\|Q^* \rightsquigarrow \tilde{Q}) \le \varepsilon$. Furthermore, by concavity of the log,

$$\begin{aligned} \varepsilon &\ge D(P\|Q^* \rightsquigarrow \tilde{Q}) \\ &\ge s \cdot D(P\|Q^* \rightsquigarrow Q_0) + (1-s) \cdot D(P\|Q^* \rightsquigarrow \bar{Q}) \\ &= (1-s) \cdot D(P\|Q^* \rightsquigarrow \bar{Q}). \end{aligned}$$

Taking the limit of $s \to 0$, we see $D(P\|Q^* \rightsquigarrow \bar{Q}) \le \varepsilon$. Finally, the result follows by taking the supremum over $\bar{Q}$. □

We conjecture that if $\mathcal{C}$ is a $\sigma$-convex set of sub-probability measures and $\mathcal{C}'$ is the Choquet-convex hull of $\mathcal{C}$ then $D(P\|Q \rightsquigarrow \mathcal{C}) = D(P\|Q \rightsquigarrow \mathcal{C}')$ for any sub-probability measures $P$ and $Q$ such that $P, Q$, and the sub-probability measures in $\mathcal{C}$ all have densities with respect to a common $\sigma$-finite measure.

# B | Appendix to Chapter 4

## B.1 Application in Practice: $k$ Separate I.I.D. Data Streams

In the simplest practical applications, we observe one block at a time, i.e. at time $n$, we have observed $\boldsymbol{X}_{(1)}, \ldots, \boldsymbol{X}_{(n)}$, where each $\boldsymbol{X}_{(i)} = (X_{i,1}, \ldots, X_{i,k})$ is a block, i.e. a vector with one outcome for each of the $k$ groups. This is a rather restrictive setup, but we can easily extend it to blocks of data in which each group has a different number of outcomes. For example, if data comes in blocks with $m_j$ outcomes in group $j$, for $j = 1 \ldots k$, $X_{(i)} = (X_{i,1,1}, \ldots, X_{i,1,m_1}, X_{i,2,1}, \ldots, X_{i,2,m_2}, \ldots, X_{i,k,1}, \ldots, X_{i,k,m_k})$, we can re-organize this having $k' = \sum_{j=1}^{k} m_j$ groups, having 1 outcome in each group, and having an alternative in which the first $m_1$ entries of the outcome vector share the same mean $\mu'_1 = \ldots = \mu'_{m_1} = \mu_1$; the next $m_2$ entries share the same mean $\mu'_{m_1+1} = \ldots = \mu'_{m_1+m_2} = \mu_2$, and so on.

Even more generally though, we will be confronted with $k$ separate i.i.d streams and data in each stream may arrive at a different rate. We can still handle this case by pre-determining a multiplicity $m_1, \ldots, m_k$ for each stream. As data comes in, we fill virtual 'blocks' with $m_j$ outcomes for group $j$, $j = 1 \ldots k$. Once a (number of) virtual block(s) has been filled entirely, the analysis can be performed as usual, restricted to the filled blocks. That is, if for some integer $B$ we have observed $Bm_j$ outcomes in stream $j$, for all $j = 1 \ldots k$, but for some $j$, we have not yet observed $(B+1)m_j$ outcomes, and we decide to stop the analysis and calculate the evidence against the null, then we output the product of $e$-variables for the first $B$ blocks and ignore any additional data for the time being. Importantly, if we find out, while analyzing the streams, that some streams are providing data at a much faster rate than others, we may adapt $m_1, \ldots, m_k$ dynamically: whenever a virtual block has been finished, we

may decide on alternative multiplicities for the next block; see Turner et al. (2024) for a detailed description for the case that $k = 2$.

## B.2 Proofs for Section 4.2

In the proofs we freely use, without specific mention, basic facts about derivatives of (log-) densities of exponential families. These can all be found in, for example, Barndorff-Nielsen (1978).

### B.2.1 Proof of Proposition 4.6

*Proof.* Since $S_{\mathrm{GRO}(\mathcal{M})}$ was already shown to be an E-variable in Lemma 4.4, the 'if' part of the statement holds. The 'only-if' part follows directly from Corollary 2 to Theorem 1 in (Grünwald et al., 2024), which states that there can be at most one E-variable of the form $p_{\boldsymbol{\mu}}(X^k)/r(X^k)$ where $r$ is a probability density for $X^k$. □

### B.2.2 Proof of Proposition 4.7

*Proof.* Define $g(\mu_0) := \mathbb{E}_{p_{\langle\mu_0\rangle}}\left[S_{\mathrm{PSEUDO}(\mathcal{M})}\right]$ and $B(\mu_i) := A\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right)$.

$$
\begin{aligned}
g(\mu_0) &= \mathbb{E}_{p_{\langle\mu_0\rangle}}\left[\prod_{i=1}^{k} \frac{p_{\mu_i}(X_i)}{p_{\mu_0^*}(X_i)}\right] = \prod_{i=1}^{k} \mathbb{E}_{Y\sim p_{\mu_0}}\left[\frac{p_{\mu_i}(Y)}{p_{\mu_0^*}(Y)}\right] \\
&= \prod_{i=1}^{k} \int \exp\left(\lambda(\mu_0)y - A\left(\lambda(\mu_0)\right)\right) \cdot \frac{\exp\left(\lambda(\mu_i)y - A\left(\lambda(\mu_i)\right)\right)}{\exp\left(\lambda(\mu_0^*)y - A\left(\lambda(\mu_0^*)\right)\right)} d\rho(y) \\
&= \prod_{i=1}^{k} \int \exp\left(\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right) y - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right) + A\left(\lambda(\mu_0^*)\right)\right) d\rho(y) \\
&= \prod_{i=1}^{k} \exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right) \exp\left(B(\mu_i)\right) \\
&\qquad \cdot \int \exp\left(\left(\lambda(\mu_i) + \lambda(\mu_0) - \lambda(\mu_0^*)\right) y - B(\mu_i)\right) d\rho(y) \\
&= \prod_{i=1}^{k} \exp\left(A\left(\lambda(\mu_0^*)\right) - A\left(\lambda(\mu_i)\right) - A\left(\lambda(\mu_0)\right)\right) \exp\left(B(\mu_i)\right) \cdot 1 \\
&= \exp\left(kA\left(\lambda(\mu_0^*)\right) - \sum_{i=1}^{k} A\left(\lambda(\mu_i)\right) - kA\left(\lambda(\mu_0)\right) + \sum_{i=1}^{k} B(\mu_i)\right). \qquad \text{(B.1)}
\end{aligned}
$$

Taking first and second derivatives with respect to $\mu_0$, we find

$$\frac{d}{d\mu_0} g(\mu_0) = g(\mu_0) \cdot \frac{d}{d\mu_0} \left( \sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right) \right) \tag{B.2}$$

and

$$\begin{aligned} \frac{d^2}{d\mu_0^2} g(\mu_0) =& \left( \frac{d}{d\mu_0} g(\mu_0) \right) \cdot \frac{d}{d\mu_0} \left( \sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right) \right) \\ &+ g(\mu_0) \cdot \frac{d^2}{d\mu_0^2} \left( \sum_{i=1}^{k} B(\mu_i) - kA\left(\lambda(\mu_0)\right) \right) \\ =& g(\mu_0) \left( \sum_{i=1}^{k} (\mu_i + \mu_0 - \mu_0^*) - k\mu_0 \right)^2 \\ &+ g(\mu_0) \left( \sum_{i=1}^{k} \mathrm{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\mathrm{VAR}_{P_{\mu_0}}[X] \right) \\ =& g(\mu_0) \left( \sum_{i=1}^{k} \mathrm{VAR}_{P_{\mu_i + \mu_0 - \mu_0^*}}[X] - k\mathrm{VAR}_{P_{\mu_0}}[X] \right) = g(\mu_0) \cdot f(\mu_0). \end{aligned} \tag{B.3}$$

where the second equality holds because of (B.2), $(d/d\lambda(\mu))A(\lambda(\mu)) = \mathbb{E}_{P_\mu}[X]$ and $(d^2/d\lambda(\mu)^2)A(\lambda(\mu)) = \mathrm{VAR}_{P_\mu}[X]$. (B.3) is continuous with respect to $\mu_0$. Therefore, if $f(\mu_0^*) > 0$ holds, it means that there exists an interval $\mathtt{M}^* \subset \mathtt{M}$ with $\mu_0^*$ in the interior of $\mathtt{M}^*$ on which (B.1) is strictly convex. Then there must exist a point $\mu_0' \in \mathtt{M}^*$ satisfying $\mathbb{E}_{P_{\langle \mu_0' \rangle}}\left[S_{\mathrm{PSEUDO}(\mathcal{M})}\right] > \mathbb{E}_{P_{\langle \mu_0^* \rangle}}\left[S_{\mathrm{PSEUDO}(\mathcal{M})}\right] = 1$, i.e. $S_{\mathrm{PSEUDO}(\mathcal{M})}$ is not an E-variable. Conversely, $f(\mu_0^*) < 0$ means that there exists an interval $\mathtt{M}^* \subset \mathtt{M}$ with $\mu_0^*$ in the interior of $\mathtt{M}^*$, on which (B.1) is strictly concave. The result follows. □

### B.2.3 Proof of Theorem 4.8

To prepare for the proof of Theorem 4.8, let us first recall Young's [1912] inequality:

**Lemma B.1. [Young's inequality]** *Let $p, q$ be positive real numbers satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then if $a, b$ are nonnegative real numbers, $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$.*

The proof of Theorem 4.8 follows exactly the same argument as the one used by Turner et al. (2024) to prove this statement in the special case that $\mathcal{M}$ is the Bernoulli model.

*Proof.* We first show that $S_{\mathrm{GRO}(\mathrm{IID})}$ as defined in the theorem statement is an E-variable.

For this, we set $p_0^*(X) = \frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X)$. We have:

$$\mathbb{E}_{X^k \sim P_{\langle \mu_0 \rangle}}\left[S_{\text{GRO(IID)}}\right] = \mathbb{E}_{X_1 \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right] \cdot \ldots \cdot \mathbb{E}_{X_k \sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right]. \tag{B.4}$$

We also have

$$\begin{aligned} &\frac{1}{k}\mathbb{E}_{X_1 \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X_1)}{p_0^*(X_1)}\right] + \cdots + \frac{1}{k}\mathbb{E}_{X_k \sim P_{\mu_0}}\left[\frac{p_{\mu_k}(X_k)}{p_0^*(X_k)}\right] \\ =&\frac{1}{k}\mathbb{E}_{X \sim P_{\mu_0}}\left[\frac{p_{\mu_1}(X)}{\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X)} + \cdots + \frac{p_{\mu_k}(X)}{\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X)}\right] = 1. \end{aligned} \tag{B.5}$$

We need to show that (B.4) $\leq 1$, for which we can use (B.5). Stated more simply, it is sufficient to prove $\prod_{i=1}^{k} r_i \leq 1$ with $\frac{1}{k}\sum_{i=1}^{k} r_i \leq 1$, $r_i \in \mathbb{R}^+$. But this is easily established:

$$\begin{aligned} \frac{1}{k}\sum_{i=1}^{k} r_i = \frac{k-1}{k} \cdot \frac{\sum_{i=1}^{k-1} r_i}{k-1} + \frac{r_k}{k} &\geq \left(\frac{\sum_{i=1}^{k-1} r_i}{k-1}\right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}} \\ &= \left(\frac{k-2}{k-1} \cdot \frac{\sum_{i=1}^{k-2} r_i}{k-2} + \frac{r_{k-1}}{k-1}\right)^{\frac{k-1}{k}} r_k^{\frac{1}{k}} \\ &\geq \left(\frac{\sum_{i=1}^{k-2} r_i}{k-2}\right)^{\frac{k-2}{k}} r_{k-1}^{\frac{1}{k}} r_k^{\frac{1}{k}} \\ &\vdots \\ &\geq \left(\frac{r_1 + r_2}{2}\right)^{\frac{2}{k}} \prod_{i=3}^{k} r_i^{\frac{1}{k}} \geq \prod_{i=1}^{k} r_i^{\frac{1}{k}} \end{aligned} \tag{B.6}$$

where the first inequality holds because of Young's inequality, by setting $\frac{1}{p} := \frac{k-1}{k}$, $\frac{1}{q} := \frac{1}{k}$, $a^p := \frac{\sum_{i=1}^{k-1} r_i}{k-1}$, $b^q := r_k$ in Lemma B.1. The other inequalities are established in the same way. It follows that $\prod_{i=1}^{k} r_i^{\frac{1}{k}} \leq 1$ and further $\prod_{i=1}^{k} r_i \leq 1$.

This shows that $S_{\text{GRO(IID)}}$ is a e-variable. It remains to show that $S_{\text{GRO(IID)}}$ is indeed the GRO e-variable relative to $\mathcal{H}_0(\text{IID})$; once we have shown this, it follows by Lemma 2 that it is the unique such e-variable and therefore by Lemma 1 that $P_0^*$ achieves the minimum in Lemma 1. Since we already know that $S_{\text{GRO(IID)}}$ is an e-variable, the fact

that it is the GRO e-variable relative to $\mathcal{H}_0(\text{IID})$ follows immediately from Corollary 2 of Theorem 1 in Grünwald et al. (2024), which states that there can be at most one e-variable of form $p_{\boldsymbol{\mu}}(X^k)/r(X^k)$ where $r$ is a probability density. Since $S_{\text{GRO(IID)}}$ is such an e-variable, Lemma 1 gives that it must be the GRO e-variable. □

### B.2.4 Proof of Proposition 4.11

*Proof.* The observed values of $X_1, X_2, \ldots, X_k$ are denoted as $x^k$ ($:= x_1, \ldots, x_k$). With $X_k(x^{k-1}, z) := z - \sum_{i=1}^{k-1} x_i$ and $\mathcal{C}(z)$ as in (4.12) and $p_{\boldsymbol{\mu};[Z]}(z)$ and $\rho(x^{k-1})$ as in (4.11), we get:

$$
\begin{aligned}
p_{\boldsymbol{\mu}}\left(x^{k-1} \middle| Z = z\right) &= \frac{p_{\boldsymbol{\mu}}\left(x^k\right)}{p_{\boldsymbol{\mu};[Z]}(z)} \\
&= \frac{\exp\left(\sum_{i=1}^{k} (\lambda(\mu_i)x_i - A(\lambda(\mu_i)))\right)}{\int_{\mathcal{C}(z)} \exp\left(\sum_{i=1}^{k-1} (\lambda(\mu_i)y_i - A(\lambda(\mu_i)) + \lambda(\mu_k)X_k(y^{k-1}, z)) - A(\lambda(\mu_k)))\right) d\rho(y^{k-1})} \\
&= \frac{\exp\left(\lambda(\mu_k)z + \sum_{i=1}^{k-1} (\lambda(\mu_i) - \lambda(\mu_k))x_i)\right)}{\int_{\mathcal{C}(z)} \exp\left(\lambda(\mu_k)z + \sum_{i=1}^{k-1} (\lambda(\mu_i) - \lambda(\mu_k))y_i\right) d\rho(y^{k-1})} \\
&= \frac{\exp\left(\sum_{i=1}^{k-1} (\lambda(\mu_i) - \lambda(\mu_k))x_i\right)}{\int_{\mathcal{C}(z)} \exp\left(\sum_{i=1}^{k-1} (\lambda(\mu_i) - \lambda(\mu_k))y_i\right) d\rho(y^{k-1})}.
\end{aligned}
$$

□

## B.3 Proofs for Section 4.3

### B.3.1 Proof of Theorem 4.12

*Proof.* We prove the theorem using an elaborate Taylor expansion of $F(\delta)$, defined below, around $\delta = 0$. We first calculate the first four derivatives of $F(\delta)$. Thus we define and derive, with $\mu_i = \mu_0 + \alpha_i\delta$ and $f_y(\delta) = \sum_{i=1}^{k} p_{\mu_i}(y)$ defined as in the theorem

statement,

$$\begin{aligned}
F(\delta) :=& \mathbb{E}_{P_{\langle\mu_0\rangle+\boldsymbol{\alpha}\delta}}\left[\log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{GRO(IID)}}\right] \\
=& \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \prod_{j=1}^{k}\left(\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X_j)\right) - \log p_{\langle\mu_0\rangle}(X^k)\right] \\
=& \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\sum_{j=1}^{k}\log f_{X_j}(\delta) - \sum_{j=1}^{k}\log p_{\mu_0}(X_j)\right] - k\log k \\
\overset{(a)}{=}& \sum_{j=1}^{k}\mathbb{E}_{X\sim P_{\mu_j}}\left[\log f_X(\delta) - \log p_{\mu_0}(X)\right] - k\log k \\
\overset{(b)}{=}& \overbrace{\int_{y\in\mathcal{X}} f_y(\delta)\log f_y(\delta)d\rho(y)}^{F_1(\delta)} + \overbrace{\left(-\int_{y\in\mathcal{X}} f_y(\delta)\log p_{\mu_0}(y)d\rho(y)\right)}^{F_2(\delta)} - k\log k, \qquad \text{(B.7)}
\end{aligned}$$

where we define $F_1(\delta)$ to be equal to the leftmost term in (B.7) and $F_2(\delta)$ to be equal to the second, and $(a)$ and (b) both hold provided that

$$\text{for all } j \in \{1,\dots,k\}\text{: } \mathbb{E}_{X_j\sim P_{\mu_j}}\left[\,|\log f_{X_j}(\delta) - \log p_{\mu_0}(X_j)\,|\,\right] < \infty \qquad \text{(B.8)}$$

is finite. In Appendix B.6 we verify that this condition, as well as a plethora of related finiteness-of-expectation-of-absolute-value conditions hold for all $\delta$ sufficiently close to 0. Together these not just imply (a) and (b), but also (c) that we can freely exchange integration over $y$ and differentiation over $\delta$ for all such $\delta$ when computing the first $k$ derivatives of $F_1(\delta)$ and $F_2(\delta)$, for any finite $k$ and (d) that all these derivatives are finite for $\delta$ in a compact interval including 0 (since the details are straightforward but quite tedious and long-winded we deferred these to Appendix B.6). Thus, using (c), we will freely differentiate under the integral sign in the remainder of the proof below, and using (d), we will be able to conclude that the final result is finite.

For each derivative, we first compute the derivative of $F_1(\delta)$ and then that of $F_2(\delta)$.

$$\begin{aligned}
F_1'(\delta) &= \int f_y'(\delta)d\rho(y) + \int f_y'(\delta)\log f_y(\delta)d\rho(y) = 0, \\
F_2'(\delta) &= -\int f_y'(\delta)\log p_{\mu_0}(y)d\rho(y) = 0, \text{ so } \; F'(0) = F_1'(0) + F_2'(0) = 0, \qquad \text{(B.9)}
\end{aligned}$$

where the above formulas hold since $f_x'(0) = 0$ for all $x \in \mathcal{X}$, which can be obtained

by

$$
\begin{aligned}
f'_x(\delta^\circ) =& \sum_{j=1}^{k} \frac{dp_{\mu_j}(x)}{d\mu_j} \frac{d\mu_j}{d\delta}(\delta^\circ), \\
f'_x(0) =& \frac{dp_{\mu_0}(x)}{d\mu_0} \sum_{j=1}^{k} \frac{d\mu_j}{d\delta}(0) = \frac{dp_{\mu_0}(x)}{d\mu_0} \sum_{j=1}^{k} \alpha_j = 0,
\end{aligned} \tag{B.10}
$$

where we used that all $\mu_j$ are equal to $\mu_0$ at $\delta = 0$. We turn to the second derivatives:

$$
\begin{aligned}
F''_1(\delta) =& \int f''_y(\delta) d\rho(y) + \int \left( f''_y(\delta) \log f_y(\delta) + \frac{\left(f'_y(\delta)\right)^2}{f_y(\delta)} \right) d\rho(y) \\
=& \int \left( f''_y(\delta) \log f_y(\delta) + \frac{\left(f'_y(\delta)\right)^2}{f_y(\delta)} \right) d\rho(y) \\
F''_1(0) =& \int \left( f''_y(0) \log f_y(0) + \frac{\left(f'_y(0)\right)^2}{f_y(0)} \right) d\rho(y); \\
=& \int f''_y(0) \log p_{\mu_0}(y) d\rho(y) + \int_{y \in \mathcal{X}} \left( f''_y(0) \log k \right) d\rho(y) \\
=& \int \left( f''_y(0) \log p_{\mu_0}(y) \right) d\rho(y),
\end{aligned} \tag{B.11}
$$

where $\int f''_y(\delta) d\rho(y) = 0$ because $\int f_y(\delta) d\rho(y) = k$, in which $k$ is a constant that does not depend on $\delta$. Then $F''_2(\delta)$ is given by

$$
\begin{aligned}
F''_2(\delta) =& - \int f''_y(\delta) \log p_{\mu_0}(y) d\rho(y) \ ; \ F''_2(0) = - \int f''_y(0) \log p_{\mu_0}(y) d\rho(y), \text{ so} \\
F''(0) =& F''_1(0) + F''_2(0) = 0.
\end{aligned} \tag{B.12}
$$

Now we compute the third derivative of $F(\delta)$, denoted as $F^{(3)}(\delta)$.

$$
\begin{aligned}
F^{(3)}_1(\delta) =& \int \left( f^{(3)}_y(\delta) \log f_y(\delta) + \frac{f''_y(\delta) f'_y(\delta)}{f_y(\delta)} + \frac{2 f''_y(\delta) f'_y(\delta) f_y(\delta) - (f'_y(\delta))^3}{(f_y(\delta))^2} \right) d\rho(y) \\
F^{(3)}_1(0) =& \int f^{(3)}_y(0) \log f_y(0) d\rho(y) = \int f^{(3)}_y(0) \log p_{\mu_0}(y) d\rho(y) + \int f^{(3)}_y(0) \log k d\rho(y) \\
=& \int f^{(3)}_y(0) \log p_{\mu_0}(y) d\rho(y) \\
F^{(3)}_2(\delta) =& - \int f^{(3)}_y(\delta) \log p_{\mu_0}(y) d\rho(y)
\end{aligned}
$$

$$F_2^{(3)}(0) = -\int f_y^{(3)}(0)\log p_{\mu_0}(y)d\rho(y),\ \text{so}\ F^{(3)}(0) = F_1^{(3)}(0) + F_2^{(3)}(0) = 0,$$

which holds since $f_y'(0) = 0$ and $\int f_y(0)d\rho(y) = k$.

The fourth derivative of $F(\delta)$ can be computed as follows:

$$\begin{aligned} F_1^{(4)}(\delta) = &\int \left( f_y^{(4)}(\delta)\log f_y(\delta) + \frac{f_y^{(3)}(\delta)f_y'(\delta)}{f_y(\delta)} \right) d\rho(y) \\ &+ \int 3\cdot \frac{\left( f_y^{(3)}(\delta)f_y'(\delta) + (f_y''(\delta))^2 \right) f_y(\delta) - f_y''(\delta)\left(f_y'(\delta)\right)^2}{\left(f_y(\delta)\right)^2} d\rho(y) \\ &- \int \frac{3\left(f_y(\delta)f_y'(\delta)\right)^2 \cdot f_y''(\delta) - 2\left(f_y'(\delta)\right)^4 \cdot f_y(\delta)}{\left(f_y(\delta)\right)^4} d\rho(y)\ ; \end{aligned} \tag{B.13}$$

$$\begin{aligned} F_1^{(4)}(0) &= \int \left( f_y^{(4)}(0)\log f_y(0) + \frac{3\left(f_y''(0)\right)^2}{f_y(0)} \right) d\rho(y) \\ &= \int f_y^{(4)}(0)\log p_{\mu_0}(y)d\rho(y) + \log k \int_{y\in\mathcal{X}} f_y^{(4)}(0)d\rho(y) + \int_{y\in\mathcal{X}} \frac{3\left(f_y''(0)\right)^2}{f_y(0)} d\rho(y) \\ &= \int f_y^{(4)}(0)\log p_{\mu_0}(y)d\rho(y) + \int_{y\in\mathcal{X}} \frac{3\left(f_y''(0)\right)^2}{f_y(0)} d\rho(y), \end{aligned}$$

and $F_2^{(4)}(\delta)$ can be computed by

$$\begin{aligned} &F_2^{(4)}(\delta) = -\int f_y^{(4)}(\delta)\log p_{\mu_0}(y)d\rho(y),\ F_2^{(4)}(0) = -\int f_y^{(4)}(0)\log p_{\mu_0}(y)d\rho(y),\ \text{so} \\ &F^{(4)}(0) = F_1^{(4)}(0) + F_2^{(4)}(0) = \int \frac{3\left(f_y''(0)\right)^2}{f_y(0)} d\rho(y) > 0. \end{aligned}$$

Based on the above derivatives, we can now do a fourth-order Taylor expansion of $F(\delta)$ around $\delta = 0$, which gives:

$$\begin{aligned} \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{GRO(IID)}}\right] &= \frac{1}{4!}F^{(4)}(0)\delta^4 + o(\delta^4) \\ &= \frac{1}{8}\int_{y\in\mathcal{X}} \frac{\left(f_y''(0)\right)^2}{f_y(0)} d\rho(y)\cdot\delta^4 + o\left(\delta^4\right), \end{aligned}$$

where $f_y(0) = \sum_{i=1}^k p_{\mu_0}(y) = k p_{\mu_0}(y)$ and $f_y''(0) = \left(\sum_{i=1}^k \alpha_i^2\right) \cdot \frac{d^2}{d\mu^2} p_\mu(y) \mid_{\mu=\mu_0} = \frac{d^2}{d\mu^2} p_\mu(y) \mid_{\mu=\mu_0}$. □

### B.3.2 Proof of Theorem 4.13

*Proof.* We obtain the result using an even more involved Taylor expansion than in the previous theorem. As in that theorem, we will freely differentiate (with respect to $\delta$) under the integral sign — that this is allowed is again verified in Appendix B.6.

Let $\boldsymbol{\mu}$, $\boldsymbol{\alpha}, \mathcal{C}(z), \rho(x^{k-1}), P_{\boldsymbol{\mu}}$ etc. be as in the theorem statement. We have:

$$
\begin{aligned}
f(\delta) &:= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{COND}}\right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log \frac{p_{\boldsymbol{\mu}}\left(X^{k-1} \mid Z\right)}{p_{\langle\mu_0\rangle}\left(X^{k-1} \mid Z\right)}\right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} - \log \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0\rangle}\left(X^k\right)} + \log \frac{\int_{\mathcal{C}(z)} p_{\boldsymbol{\mu}}\left(x^k\right) d\rho(x^{k-1})}{\int_{\mathcal{C}(z)} p_{\langle\mu_0\rangle}\left(x^k\right) d\rho(x^{k-1})}\right] \\
&= D\left(P_{\langle\mu_0\rangle + \boldsymbol{\alpha}\delta;[Z]} \| P_{\langle\mu_0\rangle;[Z]}\right).
\end{aligned}
$$

We will prove the result by doing a Taylor expansion for $f(\delta)$ around $\delta = 0$. It is obvious that $f(0) = 0$ and the first derivative $f'(0) = 0$ since $f(0)$ is the minimum of $f(\delta)$ over an open set, and $f(\delta)$ is differentiable. We proceed to compute the second derivative of $f(\delta)$, using the notation $g_z(\delta) = p_{\langle\mu_0\rangle + \boldsymbol{\alpha}\delta;[Z]}(z)$ as in the theorem statement, with $g_z'$ and $g_z''$ denoting first and second derivatives.

$$
\begin{aligned}
f'(\delta) &= \int g_z'(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z) + \int g_z'(\delta) d\rho_{[Z]}(z) = \int g_z'(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z). \\
f''(\delta) &= \int g_z''(\delta) \log \frac{g_z(\delta)}{g_z(0)} d\rho_{[Z]}(z) + \int \frac{\left(g_z'(\delta)\right)^2}{g_z(\delta)} d\rho_{[Z]}(z),
\end{aligned}
$$

where in the first line, the second equality follows since the second term does not change if we interchanging differentiation and integration and the fact that $\int g_z(\delta) dz = 1$ is constant in $\delta$. We obtain

$$
f''(0) = \int \frac{\left(g_z'(0)\right)^2}{g_z(0)} d\rho_{[Z]}(z), \tag{B.14}
$$

and, with $x_k$ set to $X_k(x^{k-1}, z)$ and recalling that $\boldsymbol{\mu} = \langle \mu_0 \rangle + \boldsymbol{\alpha}\delta$ and $\mu_j = \mu_0 + \alpha_j \delta$,

$$
\begin{aligned}
g'_z(\delta) &= \int_{\mathcal{C}(z)} \frac{d}{d\delta} p_{\langle \mu_0 \rangle + \boldsymbol{\alpha}\delta}(x^k) d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} \prod_{i \in \{1,\ldots,k\} \setminus j} p_{\mu_i}(x_i) \frac{dp_{\mu_j}(x_j)}{d\delta} d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\mu_1,\ldots,\mu_{j-1},\mu_{j+1},\ldots,\mu_k}(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k) \frac{dp_{\mu_j}(x_j)}{d\mu_j} \frac{d\mu_j}{d\delta} d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\boldsymbol{\mu}}(x^k) \frac{d \log p_{\mu_j}(x_j)}{d\mu_j} \alpha_j d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\boldsymbol{\mu}}(x^k) \left( I(\mu_j) x_j - \mu_j I(\mu_j) \right) \alpha_j d\rho(x^{k-1})
\end{aligned}
$$

where $I(\mu_j)$ is the Fisher information. The final equality follows because, with $\lambda(\mu_j)$ denoting the canonical parameter corresponding to $\mu_j$, we have $d\lambda(\mu_j)/d\mu_j = I(\mu_j)$ and $dA(\beta)/d\beta) \mid_{\beta=\lambda(\mu_j)} = \mu_j$; see e.g. (Grünwald, 2007, Chapter 18). Now

$$
\begin{aligned}
g'_z(0) &= \int_{\mathcal{C}(z)} \sum_{j=1}^{k} p_{\langle \mu_0 \rangle}(x^k) \left( I(\mu_0) x_j - \mu_0 I(\mu_0) \right) \alpha_j d\rho(x^{k-1}) \\
&= \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) I(\mu_0) \sum_{j=1}^{k} x_j \alpha_j d\rho(x^{k-1}) \qquad \text{(B.15)} \\
&= I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) \sum_{j=1}^{k} x_j \alpha_j d\rho(x^{k-1}) \qquad \text{(B.16)}
\end{aligned}
$$

where the second equality follows from $\sum_{j=1}^{k} \alpha_j = 0$. Because $X^k$ i.i.d. $\sim P_{\mu_0}$ under $P_{\langle \mu_0 \rangle}$ and the integral in (B.15) is over a set of exchangeable sequences, (For understanding the statement, we can consider the simple case $k = 2$, $X_1$ and $X_2$ can be exchangeable because they are 'symmetric' for given $\mathcal{C}(z)$.) we must have that (B.15) remains valid if we re-order the $\alpha_j$'s in round-robin fashion, i.e. for all $i = 1..k$, we have, with $\alpha_{j,i} = \alpha_{(j+i-1) \mod k}$,

$$
g'_z(0) = I(\mu_0) \cdot \int_{\mathcal{C}(z)} p_{\langle \mu_0 \rangle}(x^k) \sum_{j=1}^{k} x_j \alpha_{j,i} d\rho(x^{k-1}).
$$

Summing these $k$ equations we get, using that $\sum_{i=1}^{k} \alpha_i = 0$, that $kg'_z(0) = 0$ so that $g'_z(0) = 0$. From (B.14) we now see that

$$f''(0) = 0.$$

Now we compute the third derivative of $f(\delta)$, denoted as $f^{(3)}(\delta)$:

$$\begin{aligned} f^{(3)}(\delta) = &\int \left( g_z^{(3)}(\delta) \log \frac{g_z(\delta)}{g_z(0)} + \frac{g''_z(\delta) g'_z(\delta)}{g_z(\delta)} \right) d\rho_{[Z]}(z) \\ &+ \int \left( \frac{2g''_z(\delta) g'_z(\delta) g_z(\delta) - (g'_z(\delta))^3}{(g_z(\delta))^2} \right) d\rho_{[Z]}(z) \end{aligned}$$

So since $g'_z(0) = 0$ we must also have

$$f^{(3)}(0) = 0.$$

The fourth derivative of $f(\delta)$ is now computed as follows:

$$\begin{aligned} f^{(4)}(\delta) = &\int \left( g_z^{(4)}(\delta) \log \frac{g_z(\delta)}{g_z(0)} + \frac{g_z^{(3)}(\delta) \cdot g'_z(\delta)}{g_z(\delta)} \right) d\rho_{[Z]}(z) \\ &+ \int 3 \cdot \frac{\left( g_z^{(3)}(\delta) \cdot g'_z(\delta) + (g''_z(\delta))^2 \right) g_z(\delta) - g''_z(\delta) \cdot (g'_z(\delta))^2}{(g_z(\delta))^2} d\rho_{[Z]}(z). \end{aligned}$$

Then

$$f^{(4)}(0) = \int \frac{3 \left( g''_z(0) \right)^2}{g_z(0)} d\rho_{[Z]}(z) > 0.$$

We now have all ingredients for a fourth-order Taylor expansion of $f(\delta)$ around $\delta = 0$, which gives:

$$\mathbb{E}_{P_\mu} \left[ \log S_{\text{PSEUDO}(\mathcal{M})} - \log S_{\text{COND}} \right] = \frac{1}{8} \int \frac{\left( g''_z(0) \right)^2}{g_z(0)} d\rho_{[Z]}(z) \cdot \delta^4 + o\left(\delta^4\right)$$

which is what we had to prove. □

## B.4 Proofs for Section 4.4

In this section, we prove all the statements in Table 4.1.

### B.4.1 Bernoulli Family

We prove that for $\mathcal{M}$ equal to the Bernoulli family, we have $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})} = S_{\text{GRO(IID)}} \succ S_{\text{COND}}$.

*Proof.* We set $\mu_0^* = \frac{1}{k}\sum_{i=1}^{k} \mu_i$.

$$S_{\text{GRO(IID)}} := \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k}\left(\frac{1}{k}\sum_{i=1}^{k} p_{\mu_i}(X_j)\right)} = \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k}\left(\frac{1}{k}\sum_{i=1}^{k}\left(\mu_i^{X_j}(1-\mu_i)^{1-X_j}\right)\right)} \tag{B.17}$$

$$= \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k}\left((\mu_0^*)^{X_j}(1-\mu_0^*)^{1-X_j}\right)}$$

$$= \frac{p_{\boldsymbol{\mu}}(X^k)}{\prod_{j=1}^{k} p_{\mu_0^*}(X_j)} = S_{\text{PSEUDO}(\mathcal{M})} \tag{B.18}$$

where the third equality holds since $X_i \in \{0,1\}$. So $S_{\text{PSEUDO}(\mathcal{M})}$ is an E-variable and $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})}$ according to Theorem 4.6. Then the claim follows using (4.9) together with the fact that when $Z = 0$ or $Z = 2$, we have $S_{\text{COND}} = 1$, while this is not true for the other $e$-variables, so that $S_{\text{COND}} \neq S_{\text{GRO}(\mathcal{M})} = S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO(IID)}}$. The result then follows from (4.9). □

### B.4.2 Poisson and Gaussian Family With Free Mean and Fixed Variance

We prove that for $\mathcal{M}$ equal to the family of Gaussian distributions with free mean and fixed variance $\sigma^2$, we have $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})} = S_{\text{COND}} \succ S_{\text{GRO(IID)}}$. The proof that the same holds for $\mathcal{M}$ equal to the family of Poisson distributions is omitted, as it is completely analogous.

*Proof.* Note that if we let $Z := \sum_{i=1}^{k} X_i$, then we have that $Z \sim \mathcal{N}(\sum_{i=1}^{k}\mu_i, k\sigma^2)$ if $X^k \sim P_{\boldsymbol{\mu}}$. Let $\mu_0^*$ be given by (4.8) relative to fixed alternative $P_{\boldsymbol{\mu}}$ as in the definition of $S_{\text{PSEUDO}(\mathcal{M})}$ underneath (4.8). Since $k\mu_0^* = \sum_{i=1}^{k}\mu_i$, we have that $Z$ has the same distribution for $X^k \sim P_{\langle\mu_0^*\rangle}$. This can be used to write

$$S_{\text{COND}} = \frac{p_{\boldsymbol{\mu}}\left(X^k \mid Z\right)}{p_{\langle\mu_0^*\rangle}\left(X^k \mid Z\right)} = \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0^*\rangle}\left(X^k\right)}\frac{p_{\langle\mu_0^*\rangle}(Z)}{p_{\boldsymbol{\mu}}(Z)} = \frac{p_{\boldsymbol{\mu}}\left(X^k\right)}{p_{\langle\mu_0^*\rangle}\left(X^k\right)} = S_{\text{PSEUDO}(\mathcal{M})}.$$

Therefore, $S_{\text{PSEUDO}(\mathcal{M})}$ is also an $e$-variable, so we derive that $S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{GRO}(\mathcal{M})}$ by Theorem 4.6. Furthermore, we have that the denominator of $S_{\text{GRO(IID)}}$ is given by a different distribution than $p_{\langle \mu_0^* \rangle}$, so that $S_{\text{GRO(IID)}} \neq S_{\text{GRO}(\mathcal{M})} = S_{\text{PSEUDO}(\mathcal{M})} = S_{\text{COND}}$. The result then follows from (4.9).

□

### B.4.3 The Families for Which $S_{\text{pseudo}(\mathcal{M})}$ Is Not an E-variable

Here, we prove that $S_{\text{PSEUDO}(\mathcal{M})}$ is not an $e$-variable for $\mathcal{M}$ equal to the family of beta distributions with free $\beta$ and fixed $\alpha$. It then follows from (4.9) that $S_{\text{PSEUDO}(\mathcal{M})} \succ S_{\text{GRO}(\mathcal{M})}$. (4.9) also gives $S_{\text{GRO}(\mathcal{M})} \succeq S_{\text{GRO(IID)}}$ and $S_{\text{GRO}(\mathcal{M})} \succeq S_{\text{COND}}$. The same is true for $\mathcal{M}$ equal to the family of geometric distributions and the family of Gaussian distributions with free variance and fixed mean, as the proof that $S_{\text{PSEUDO}(\mathcal{M})}$ is not an $e$-variable is entirely analogous to the proof for the beta distributions given below. In all of these cases, one easily shows by simulation that in general, $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{GRO(IID)}}$ and $S_{\text{GRO}(\mathcal{M})} \neq S_{\text{COND}}$, so then $S_{\text{GRO}(\mathcal{M})} \succ S_{\text{GRO(IID)}}$ and $S_{\text{GRO}(\mathcal{M})} \succ S_{\text{COND}}$ follow.

*Proof.* First, let $Q_{\alpha,\beta}$ represent a beta distribution in its standard parameterization, so that its density is given by

$$q_{\alpha,\beta}(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1}(1-u)^{\beta-1}, \qquad \alpha, \beta > 0; u \in [0,1].$$

To simplify the proof, we assume $\alpha = 1$ here. Then

$$q_{1,\beta}(u) = \frac{\Gamma(1+\beta)}{\Gamma(\beta)}(1-u)^{\beta-1} = \frac{1}{1-u} \exp\left(\beta \log(1-u) - \log \frac{1}{\beta}\right)$$

where the first equality holds since $\Gamma(1+\beta) = \beta\Gamma(\beta)$. Comparing this to (4.1), we see that $\beta$ is the canonical parameter corresponding to the family $\{Q_{1,\beta} : \beta > 0\}$, and we have

$$\lambda(\mu) = \beta, \quad t(u) = \log(1-u), \quad A(\beta) = \log \frac{1}{\beta}.$$

To prove the statement, according to Proposition 4.7, we just need to show, for any $\mu_1, \ldots, \mu_k$ that are not all equal to each other, that, with $X = t(U) = \log(1-U)$ and $\mu_0^* = \frac{1}{k}\sum_{i=1}^{k} \mu_i$ defined as in (4.8), we have

$$\sum_{i=1}^{k} \text{VAR}_{P_{\mu_i}}[X] - k\text{VAR}_{P_{\mu_0^*}}[X] > 0. \tag{B.19}$$

Straightforward calculation gives

$$\text{VAR}_{P_{\mu_i}}[X] = \text{VAR}_{Q_{1,\beta_i}}[X] = \frac{d^2}{d^2\beta_i}(\log\frac{1}{\beta_i}) = \frac{1}{\beta_i^2} \text{ in particular } \text{VAR}_{P_{\mu_0^*}}[X] = \frac{1}{(\beta_0^*)^2} \tag{B.20}$$

where $\beta_i$ corresponds to $\mu_i$, i.e. $\mathbb{E}_{Q_{1,\beta_i}}[(X)] = \mu_i$. We also have:

$$\mathbb{E}_{P_{\beta_0^*}}[(X)] = \mu_0^* = \frac{1}{k}\sum_{i=1}^{k}\mu_i = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{P_{\beta_i}}[(X)]\,. \tag{B.21}$$

While $\mathbb{E}_{P_{\beta_i}}[(X)] = \frac{d}{d\beta_i}(\log\frac{1}{\beta_i}) = -\frac{1}{\beta_i}$, therefore $\frac{1}{\beta_0^*} = \frac{1}{k}\sum\limits_{i=1}^{k}\frac{1}{\beta_i}$. We obtain, together with (B.20) and (B.21), that

$$\sum_{i=1}^{k}\text{VAR}_{P_{\mu_i}}[(X)] - k\text{VAR}_{P_{\mu_0^*}}[(X)] = \sum_{i=1}^{k}\frac{1}{(\beta_i)^2} - k\left(\frac{1}{k}\sum_{i=1}^{k}\frac{1}{\beta_i}\right)^2. \tag{B.22}$$

Jensen's inequality now gives that (B.22) is strictly positive, whenever at least one of the $\mu_i$ is not equal to $\mu_0^*$, which is what we had to show. □

## B.5 Graphical Depiction of RIPr-Approximation

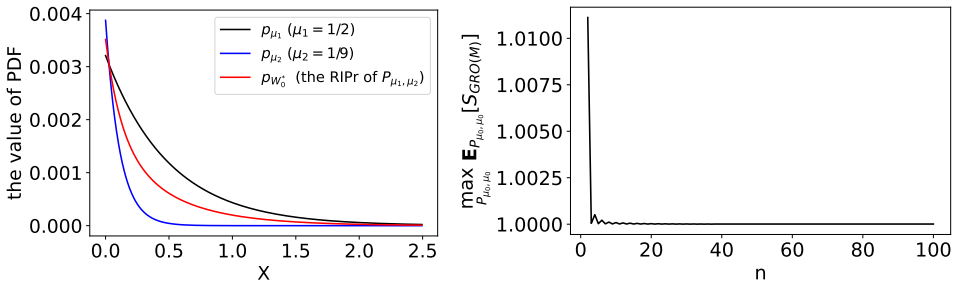



**Figure B.1:** Exponential distribution. On the right, $n$ represents number of iterations with Li's algorithm, starting at iteration 2

We illustrate RIPr-approximation and convergence of Li's algorithm with four distributions: exponential, beta with free $\beta$ and fixed $\alpha$, geometric and Gaussian with free variance and fixed mean, each with one particular (randomly chosen) setting of the parameters. The pictures on the left in Figure B.1– B.4 give the probability density functions (for geometric distributions, discrete probability mass functions) after

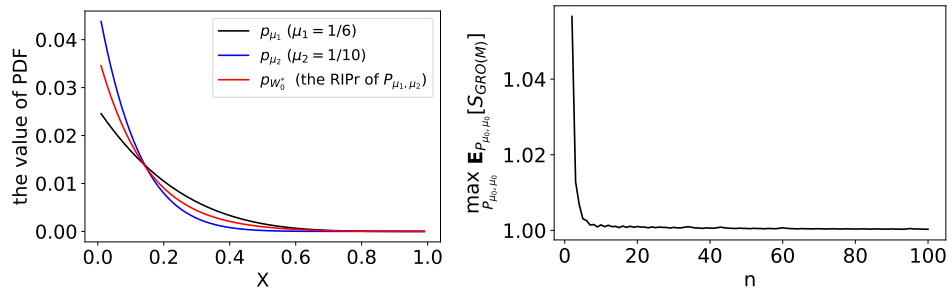



**Figure B.2:** beta with free $\beta$ and fixed $\alpha$. On the right, $n$ represents number of iterations with Li's algorithm, starting at iteration 2

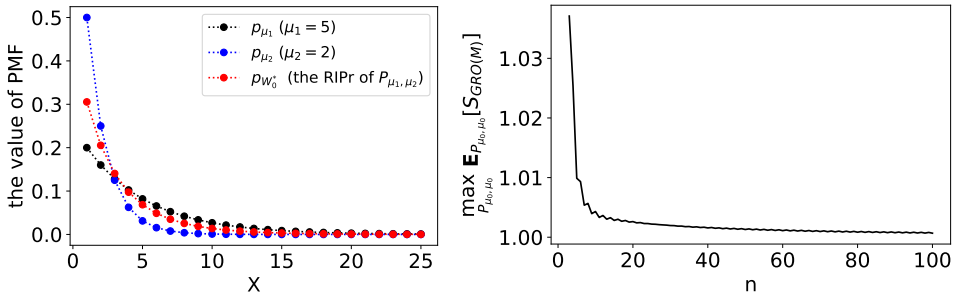



**Figure B.3:** geometric distribution. On the right, $n$ represents number of iterations with Li's algorithm, starting at iteration 3

$n = 100$ iterations of Li's algorithm. The pictures on the right illustrate the speed of convergence of Li's algorithm. The pictures on the right do not show the first (or the first two, for geometric and Gaussian with free variance) iteration(s), since the worst-case expectation $\sup_{\mu_0 \in \mathtt{M}}[S_{\mathrm{GRO}(\mathcal{M})}]$ is invariably incomparably larger in these initial steps. We empirically find that Li's algorithm converges quite fast for computing the true $S_{\mathrm{GRO}(\mathcal{M})}$. In each step of Li's algorithm, we searched for the best mixture weight $\alpha$ in $P_{(m)}$ over a uniformly spaced grid of 100 points in $[0, 1]$, and for the novel component $P' = P_{\mu',\mu'}$ by searching for $\mu'$ in a grid of 100 equally spaced points inside the parameter space $\mathtt{M}$ where the left- and right- endpoints of the grid were determined by trial and error. While with this ad-hoc discretization strategy we obviously cannot guarantee any formal approximation results, in practice it invariably worked well: in all cases, we found that $\max_{\mu_0 \in \mathtt{M}} \mathbb{E}_{P_{\mu_0,\mu_0}}[S_{\mathrm{GRO}(\mathcal{M})}] \leq 1.005$ after 15 iterations. For comparison, we show the best approximation that can be obtained by brute-force combining of just two components, for the same parameter values, in Table B.1.

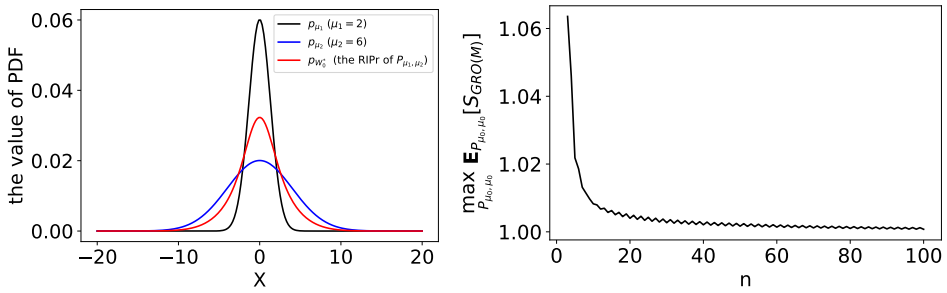



**Figure B.4:** Gaussian with free variance and fixed mean. On the right, $n$ represents number of iterations with Li's algorithm, starting at iteration 3

| Distributions | $(\mu_1, \mu_2)$ | $\alpha$ | $(\mu_{01}, \mu_{02})$ | $\sup_{\mu_0 \in \mathtt{M}} \mathbb{E}_{X_1,X_2 \sim P_{\mu_0,\mu_0}}[S]$ |
|---|---|---|---|---|
| beta | $(\frac{1}{6}, \frac{1}{10})$ | 0.57 | (0.12, 0.16) | 1.00071 |
| geometric | (5, 2) | 0.39 | (2.52, 4.21) | 1.00035 |
| Exponential | $(\frac{1}{2}, \frac{1}{9})$ | 0.53 | (0.13, 0.51) | 1.00083 |
| Gaussian with free variance and fixed mean | (2, 6) | 0.41 | (5.82, 3.36) | 1.00035 |

**Table B.1:** Analogue of Table 4.2 for $\mu_1, \mu_2$ corresponding to the parameters used in Figures B.1–B.4

## B.6 Further Details

In this section, we verify that all conditions are met for the implicit use of Fubini's theorem and differentiation under the integral sign in the proofs of Theorem 2 and 3, and that all derivatives of interest are bounded.

### B.6.1 Theorem 2

In the chapter, notation is as follows:

$$\begin{aligned} \mu_j &= \mu_0 + \delta\alpha_j \\ \lambda(\mu_j) &= \text{nat. param. } \lambda \text{ corresponding to mean } \mu = \mu_j \\ p_\mu(y) &= e^{\lambda(\mu)y - A(\lambda(\mu))} \\ f_y(\delta) &= \sum_{j=1}^{k} p_{\mu_i}(y). \end{aligned}$$

As this will simplify the notation for the derivatives, we write $g_y(\lambda) = e^{\lambda y - A(\lambda)}$, so that

$$f_y(\delta) = \sum_{j=1}^{k} g_y(\lambda(\mu_j)) \text{ and } p_{\mu_0}(y) = g_y(\lambda(\mu_0)). \tag{B.23}$$

To stress dependence on $\delta$, we write $\mu_j(\delta)$ instead of $\mu_j$ in the following.

**Step 1** We first establish the finiteness condition (B.8). We note that

$$\begin{aligned}
\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) &\leq \log(\max_j g_y(\lambda(\mu_j(\delta)))k) \\
&= \max_j \log(g_y(\lambda(\mu_j(\delta)))) + \log k \\
&\leq \max_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k \\
&\leq \sum_j \log(\max\{g_y(\lambda(\mu_j(\delta))), 1\}) + \log k \\
&\leq \sum_j |\lambda(\mu_j(\delta))y - \log A(\lambda(\mu_j(\delta)))| + \log k.
\end{aligned}$$

and

$$\begin{aligned}
\log \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) &= \log \frac{1}{k} \sum_{j=1}^{k} g_y(\lambda(\mu_j(\delta))) + \log k \\
&\geq \frac{1}{k} \sum_{j=1}^{k} \log g_y(\lambda(\mu_j(\delta))) + \log k \\
&= \frac{1}{k} \sum_{j=1}^{k} \lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta))) + \log k.
\end{aligned}$$

Putting these together, we see that

$$\begin{aligned}
&|\log f_y(\delta)| \leq \\
&\max \left\{ \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k, \left| \frac{1}{k} \sum_{j=1}^{k} (\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))) + \log k \right| \right\} \\
&\leq \sum_j |\lambda(\mu_j(\delta))y - A(\lambda(\mu_j(\delta)))| + \log k,
\end{aligned} \tag{B.24}$$

and, more trivially,

$$|\log g_y(\lambda(\mu_0))| \leq |\lambda(\mu_0)y - A(\lambda(\mu_0)|\,. \tag{B.25}$$

We know that $\lambda(\mu_j(\delta))$ and $A(\lambda(\mu_j(\delta)))$ are smooth, hence finite functions for $\mu_j(\delta)$ in the interior of the mean-value parameter space $\mathtt{M}$ (see (Barndorff-Nielsen, 1978, Chapter 9, Theorem 9.1 and Eq. (2))). Since $\mathtt{M}$ is open and for all $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathtt{M}$, it follows that $|\log f(y)(\delta) - \log g_y(\lambda(\mu_0))|$ can be written as a smooth, in particular finite function of $|y|$ for all $\delta$ in a compact subset of $\mathbb{R}$ with 0 in its interior. Since $|y| \leq 1 + y^2$ has finite expectation under all $P_\mu$ with $\mu \in \mathtt{M}$, finiteness of (B.8) follows by (B.23).

**Step 2** We now proceed to establish that we can differentiate with respect to $\delta$ for $\delta$ in a compact subset of $\mathbb{R}$ with 0 in its interior. The proof will make use of (B.24) and (B.25). We denote derivatives of functions $f_y$ and $g_y$ as

$$g_y^s(\lambda) = \frac{\mathrm{d}^s}{\mathrm{d}\lambda^s} g_y(\lambda) \quad \text{and} \quad f_y^s(\delta) = \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta).$$

We will argue that, for any $s \in \mathbb{N}$, the family $\{\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) : \delta \in \Delta\}$ is uniformly integrable for any compact $\Delta \subset \mathbb{R}$, so that we are allowed to interchange differentiation and integration (see e.g. Williams, 1991, Chapter A16).

Using standard results for exponential families, we have, for $\lambda$ in the interior of the canonical parameter space,

$$\begin{aligned} g_y^{(1)}(\lambda) &= (y - \mu(\lambda)) g_y(\lambda) \\ g_y^{(2)}(\lambda) &= -I(\lambda) g_y(\lambda) + (y - \mu(\lambda))^2 g_y(\lambda), \end{aligned}$$

where $\mu(\lambda)$ denotes the mean-value parameter corresponding to $\lambda$ and $I(\lambda)$ the corresponding Fisher information.

Continuing this using the fact that $(d^s/d\lambda^s)A(\lambda)$ is continuous for all $s$, gives

$$g_y^{(s)}(\lambda) = g_y(\lambda) \cdot h_{y,s}(\lambda) \text{ with } h_{y,s}(\lambda) = \sum_{t=1}^{s} h_{[t,s]}(\lambda)(y - \mu(\lambda))^t \tag{B.26}$$

for some smooth functions $h_{[1,s]}, h_{[2,s]}, \ldots, h_{[s,s]}$ of $\lambda$ (we do not need to know precise

definitions of these functions). Similarly

$$f_y^{(1)}(\delta) = \sum_j g_y^{(1)}(\lambda_{\mu_j(\delta)}) \cdot (\lambda(\mu_j(\delta)))'$$

where $\lambda(\mu_j(\delta))' = \frac{\mathrm{d}}{\mathrm{d}\delta}\lambda(\mu_j(\delta))$. We know that $\lambda'(\mu_j(\delta))$ and further derivatives are smooth functions for $\mu_j(\delta)$ in the interior of the mean-value parameter space $\mathtt{M}$ (see (Barndorff-Nielsen, 1978, Chapter 9, Theorem 9.1 and Eq. (2))). Since this space is open and for all $j = 1..k$, $\mu_j(0) = \mu_0 \in \mathtt{M}$, it follows that $\lambda'(\mu_j(\delta))$ are smooth functions of $\delta$ for $\delta$ in a compact subset of $\mathbb{R}$ with 0 in its interior. Thus, analogously to what we did above with $g^{(s)}$, we get that

$$f_y^{(s)}(\delta) = \sum_j \sum_{t=1}^{s} g_y^{(t)}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j) \tag{B.27}$$

for some smooth functions $r_{t,s}$, the details of which we do not need to know. In particular this gives, with

$$b_y^{(s)} := \frac{f_y^{(s)}(\delta)}{f_y(\delta)}$$

that

$$\begin{aligned}\left|b_y^{(s)}\right| &\leq \frac{\sum_j g_y(\lambda(\mu_j(\delta))) \cdot (\sum_{t=1}^{s} |h_{y,t}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j(\delta))|)}{\sum_j g_y(\lambda(\mu_j(\delta)))} \\ &\leq \sum_j \sum_{t=1}^{s} |h_{y,t}(\lambda(\mu_j(\delta))) \cdot r_{t,s}(\mu_j(\delta))|.\end{aligned}$$

Inspecting the proof in the main text, we informally note that all terms without logarithms in the first four derivatives of $F_0(\delta)$ and $F_1(\delta)$ can be written as products $f_y(\delta) \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta)$ for the $b_y^{(s)}$ we just bounded in terms of polynomials in $|y|$; similarly, the terms involving logarithms can be bounded in terms of such polynomials as well using (B.24) and (B.25), suggesting that all terms inside all integrals can be

such bounded. This is indeed the case: formalizing the reasoning, we see that

$$\int \left( \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) \right)^2 d\rho(y) =$$

$$\int \left( f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))) + f_y(\delta) \sum_u c_u \cdot b_y^{(s_2)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) \right)^2 d\rho(y)$$

$$= \int (f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))))^2 + \left( f_y(\delta) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) \right)^2$$
$$+ f_y(\delta) f_y^{(s)}(\log f_y(\delta) - \log g_y(\lambda(\mu_0))) \sum_u c_u \cdot b_y^{(s_1)}(\delta) \cdot \ldots \cdot b_y^{(s_u)}(\delta) d\rho(y).$$

By (B.24) and (B.25) and the bound on $|b_y^{(s)}|$ given above, all the terms within the integral can be bounded by polynomials in $y$ (or $|y|$), so the integral is given by linear functions of moments of $\rho$ and $P_\mu$. Therefore, using also that $\rho$ is itself a probability measure and a member of the exponential family under consideration (equal to $P_\mu$ with $\lambda(\mu) = 0$), the integral can be uniformly bounded over $\delta$ in a compact subset of the mean-value parameter space. It follows that the family $\{ \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f_y(\delta) \log f_y(\delta) - f_y(\delta) \log g_y(\lambda(\mu_0)) : \delta \in \Delta \}$ is uniformly integrable (see e.g. Williams, 1991, Chapter 13.3), so integration and differentiation may be interchanged freely (see e.g. Williams, 1991, Chapter A16). It also follows that the quantity on the right-hand side in the theorem statement is bounded.

### B.6.2 Theorem 3

As in the proof of Theorem 3, let $f(\delta) = \mathbb{E}_{P_{\boldsymbol{\mu}}} \left[ \log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} - \log \frac{p_{\boldsymbol{\mu}}(X^k|Z)}{p_{\langle \mu_0 \rangle}(X^k|Z)} \right]$.

To validate the proof in the main text we merely need to show that $f(\delta)$ is finite, and that we can interchange differentiation and expectation with respect to $\delta$ in a compact interval containing $\delta = 0$. Thus, we want to show that, for any $s \in \mathbb{N}$, we have that

$$\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f(\delta) = \mathbb{E} \left[ \frac{\mathrm{d}^s}{\mathrm{d}\delta^s} \left( \log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} - \log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle \mu_0 \rangle}(X^k \mid Z)} \right) \right].$$

To show this, first note that both $\mathbb{E}_{P_{\boldsymbol{\mu}}} \left[ \log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle \mu_0 \rangle}(X^k)} \right]$ and $\mathbb{E}_{P_{\boldsymbol{\mu}}} \left[ \log \frac{p_{\boldsymbol{\mu}}(X^k|Z)}{p_{\langle \mu_0 \rangle}(X^k|Z)} \mid Z \right]$ are KL divergences between members of exponential families (the fact that conditioning on a sum of sufficient statistics results in a new, derived full exponential family is shown by, for example, Brown (1986)), which are finite as long as $\delta$ is in a sufficiently

small interval containing 0 in its interior (since then $\boldsymbol{\mu}$ is in the interior of the mean-value parameter space). This already shows that $f(\delta)$ is finite, and it also allows us to rewrite

$$f(\delta) = \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right].$$

Furthermore, (Brown, 1986, Theorem 2.2) in combination with Theorem 9.1. and Chapter 9, Eq.2. of Barndorff-Nielsen (1978) shows that for any full exponential family, for any finite $k > 0$, the $k$-th derivative of the KL divergence with respect to its first argument, given in the mean-value parameterization, exists, is finite, and can be obtained by differentiating under the integral sign, at any $\mu$ in the interior of the mean-value parameter space. We are therefore allowed to interchange expectation and differentiation for such terms separately for all $\delta$ in any compact interval containing 0. Thus, starting with the previous display, we can write

$$\begin{aligned}
&\frac{\mathrm{d}^s}{\mathrm{d}\delta^s} f(\delta) = \frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k \mid Z)}{p_{\langle\mu_0\rangle}(X^k \mid Z)}\right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)} + \log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle\mu_0\rangle;[Z]}(Z)}\right] = \\
&\mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] - \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu}}(X^k)}{p_{\langle\mu_0\rangle}(X^k)}\right] + \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle\mu_0\rangle;[Z]}(Z)}\right] \\
&= \mathbb{E}_{P_{\boldsymbol{\mu}}}\left[\frac{\mathrm{d}^s}{\mathrm{d}\delta^s}\log \frac{p_{\boldsymbol{\mu};[Z]}(Z)}{p_{\langle\mu_0\rangle;[Z]}(Z)}\right],
\end{aligned}$$

where in the last line we use that all involved terms are finite. This is what we had to show.

# C | Appendix to Chapter 5

## C.1 Details for Section 5.4.4

We need to establish that $\Sigma_p(\boldsymbol{\mu}) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu}) = \Sigma_q^{(0)}(\boldsymbol{\mu}) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu})$ is positive semidefinite for all $\boldsymbol{\mu} \in \mathbb{R}^d$.

Thus, take any $\boldsymbol{\mu}^* \in \mathbb{R}^d$. By (5.26), we have that $q_{\boldsymbol{\mu}^*}^{(\theta^\circ)} = f_{\lambda^\circ, \boldsymbol{\beta}^\circ}^{(\theta^\circ)}$ and $p_{\boldsymbol{\mu}^*} = q_{\boldsymbol{\mu}^*}^{(0)} = f_{\lambda^*, \boldsymbol{\beta}^*}^{(0)}$ for some $\lambda^\circ, \boldsymbol{\beta}^\circ$ and $\lambda^*, \boldsymbol{\beta}^*$ that are related to each other via the normal equations (5.27). Based on the sufficient statistics (5.25), we can thus write, for $\theta \in \{0, \theta^\circ\}$, that

$$\Sigma_q^{(\theta)}(\boldsymbol{\mu}^*) = \begin{pmatrix} A^{(\theta)} & B^{(\theta)} \\ (B^{(\theta)})^T & C^{(\theta)} \end{pmatrix}$$

where $A^{(\theta^\circ)}$ is the variance of $\sum Y_i^2$ according to distribution $F_{\lambda^\circ, \boldsymbol{\beta}^\circ}^{(\theta^\circ)}$ and $C^{(\theta^\circ)}$ is the $d \times d$ covariance matrix of the $t_j(Y^n)$ according to this distribution and

$$B^{(\theta^\circ)} = \left( \operatorname{COV}\left( \sum Y_i^2, t_1(Y^n) \right), \ldots, \operatorname{COV}\left( \sum Y_i^2, t_d(Y^n) \right) \right)$$

where the covariances are again under this distribution. Similarly, $A^{(0)}$ is the variance of $\sum Y_i^2$ according to distribution $F_{\lambda^*, \boldsymbol{\beta}^*}^{(0)}$ and $B^{(0)}$, $C^{(0)}$ are defined accordingly.

Positive semidefiniteness of $\Sigma_q^{(0)}(\boldsymbol{\mu}^*) - \Sigma_q^{(\theta^\circ)}(\boldsymbol{\mu}^*)$ is easily seen to be implied[1] if we can show that $C^{(0)} - C^{(\theta^\circ)}$ is positive definite and that

$$(A^{(0)} - A^{(\theta^\circ)}) - (B^{(0)} - B^{(\theta^\circ)})^T (C^{(0)} - C^{(\theta^\circ)})^{-1} (B^{(0)} - B^{(\theta^\circ)}) \geq 0. \tag{C.1}$$

To show that $C^{(0)} - C^{(\theta^\circ)}$ is positive definite, note that $C^{(\theta^\circ)}$ (as is readily established, for example, by twice differentiating $\log Z_q^{(\theta^\circ)}(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$ at $\lambda = 0, \boldsymbol{\beta} = 0$) is simply

[1] For an explicit derivation see `https://math.stackexchange.com/questions/2280671/definiteness-of-a-general-partitioned-matrix-mathbf-m-left-beginmatrix-bf`.

the standard covariance matrix in linear regression scaled by $1/\sigma^{\circ 2}$, i.e. $C^{(\theta^\circ)} = \sigma^{\circ 2}\sum \mathbf{x}_i\mathbf{x}_i^T$ which by the maximal rank assumption is positive definite. Similarly $C^{(0)} = \sigma^{*2}\sum \mathbf{x}_i\mathbf{x}_i^T$ so that, since by assumption $\theta^\circ \neq 0$ and using the normal equations (5.27), we have that $C^{(0)} - C^{(\theta)} = cC^{(\theta)}$ for $c = \sigma^{*2} - \sigma^{\circ 2} > 0$ is also positive definite.

It only remains to show (C.1). As again easily established (for example, by twice differentiating $\log Z_q^{(\theta^\circ)}(\lambda, \boldsymbol{\beta}; \boldsymbol{\mu}^*)$ at $\lambda = 0, \boldsymbol{\beta} = 0$), we have that $A^{(\theta^\circ)} = 2\sigma^{\circ 2}\left(2(\sum \nu_i^{\circ 2}) + n\sigma^{\circ 2}\right)$ and similarly we find $A^{(0)} = 2\sigma^{*2}\left(2(\sum \nu_i^{*2}) + n\sigma^{*2}\right)$ and $B_j^{(\theta^\circ)} = -2\sigma^{\circ 2}(\sum \nu_i^\circ x_{i,j})$ and similarly $B_j^{(0)} = -2\sigma^{*2}(\sum \nu_i^* x_{i,j})$. By the normal equations (5.27) we find that $B_j^{(0)} - B_j^{(\theta^\circ)} = -2(\sigma^{*2} - \sigma^{\circ 2})\sum \nu_i^* x_{i,j}$. After some matrix multiplications (where we may use the cyclic property of the trace of a matrix product) we get that (C.1) is equivalent to

$$(A^{(0)} - A^{(\theta)^\circ}) - 4(\sigma^{*2} - \sigma^{\circ 2})\sum \nu_i^{*2} \geq 0.$$

But this is easily verified: it is equivalent to

$$2\sigma^{*2}\left(2\left(\sum \nu_i^{*2}\right) + n\sigma^{*2} - 2\left(\sum \nu_i^{*2}\right)\right) - 2\sigma^{\circ 2}\left(2\left(\sum \nu_i^{\circ 2}\right) + n\sigma^{\circ 2} - 2\left(\sum \nu_i^{*2}\right)\right) \geq 0$$

which in turn is equivalent to

$$2n\sigma^{*4} - 2n\sigma^{\circ 4} + 4(\sum \nu_i^{*2} - \sum \nu_i^{\circ 2})\sigma^{\circ 2} \geq 0$$

which by the normal equations is equivalent to

$$\sigma^{*4} - \sigma^{\circ 4} + 2(\sigma^{*2} - \sigma^{\circ 2})\sigma^{\circ 2} \geq 0$$

but this must be the case since by the normal equations, $\sigma^{*2} > \sigma^{\circ 2}$.

# D | Appendix to Chapter 6

## D.1 Additional Simulations

### D.1.1 Effect of Truncation on Power

Figure D.1 shows the same plot as Figure 1 in Chapter 6 but without truncation for the probabilities in the $e$-statistics, i.e. $\varepsilon = 0$.

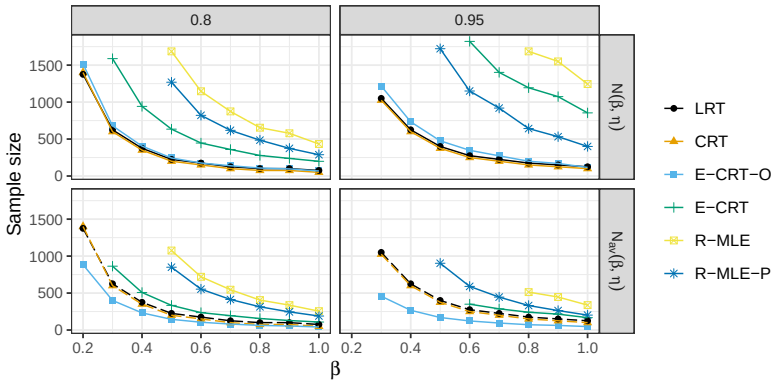



**Figure D.1:** Sample sizes for different methods as in Figure 1 in Chapter 6 but with $\varepsilon = 0$ for the $e$-statistics.

### D.1.2 Robustness With Respect to Misspecification

We test the robustness of the randomization based $e$-statistics with respect to misspecification of the conditional distribution of $X$ in the same way as in the simulation study of Berrett et al. (2020). All simulations in this section are under the null hypothesis, i.e. $\beta = 0$. Rejection rates of the $e$-statistics are again computed with a

maximal sample size of 2000 and with optional stopping, i.e. rejection if the level $1/\alpha$ is exceeded at least once, and with truncation level $\varepsilon = 0.05$. For comparison, the conditional randomization test is applied to a sample of fixed size, for sizes 200, 1000, and 2000, and additionally with the unconditional absolute correlation $|\mathrm{cor}(X, Y)|$ as test statistic, as in Berrett et al. (2020), for sample sizes 200 and 2000.

First, instead of sampling $X$ with conditional mean $\mu_Z$ as defined in (6.14), we set the mean to

$$\begin{aligned} &\mu_Z - \xi\mu_Z^3 && \text{(cubic misspecification)},\\ &\mu_Z + \xi\mu_Z^2 && \text{(quadratic misspecification)},\\ &\tanh(\xi\mu_Z)/\xi && \text{(hyperbolic tangent)}, \end{aligned}$$

which are the same misspecifications as in Berrett et al. (2020, Section 6.1.1). They are illustrated in Figure D.2 for different values of $\xi$, the range of which has been selected for each misspecification type in such a way that the relative misspecification compared to the true mean approximately matches the one in the simulations by Berrett et al. (2020). When the parameter $\xi$ equals zero, understood as limit $\xi \to 0$ for the hyperbolic tangent, the model is correctly specified. Panel (a) of Figure D.3 shows that both the CRT and the $e$-statistics are robust with respect to slight misspecifications of the conditional mean. The CRT based on the likelihood is much more robust than the other two tests, due to the fact that re-estimating the logistic regression model with simulated $X$ is invariant under affine transformations of $X$ and $Z$ and hence able to correct much of the misspecification. The $e$-statistic based test is less robust than this variant of the CRT, since it does not re-estimate the logistic model with simulated $X$, but still substantially more robust than the CRT based on unconditional correlation, which already with $n = 200$, as compared to $n = 2000$ for the $e$-values, has rejection rates strongly exceeding the nominal level as $\xi$ increases.

In panel (b) of Figure D.3, the rejection rates of the tests are shown when the distribution of $X_p$ is estimated on an independent unlabeled data set, for different sizes of this data set. The estimation of the conditional distribution is by linear regression, with the maximum likelihood estimator for the conditional variance. Here the $e$-statistics have rejection rates below the nominal level, even for unlabeled sample size as small as 50. Also the CRT with logistic likelihood as test statistic has rejection rate close to the nominal level.

Finally, in panel (c) of Figure D.3 the rejection rates are depicted for the case when the same data is used both for estimating the distribution of $X$ and for testing. The

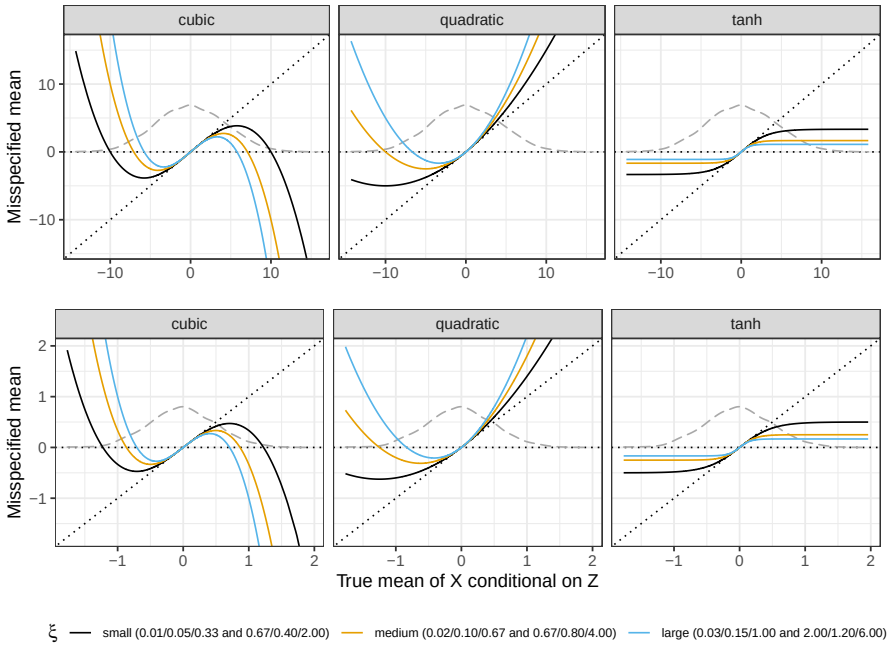



**Figure D.2:** Misspecification in the conditional mean of $X$ given $Z$ for the three different functions from Section D.1.2. Upper row of plots: $X \mid Z$ generated as in Berrett et al. (2020, Section 6.1.1). Lower row: $X \mid Z$ generated as in Section 6.4 with $q = 4$. The dashed line shows the (height adjusted) density of the conditional expectation of $X$ given $Z$. The values for $\xi$ given in the legend refer to the misspecifications in the same order as the panel colums (cubic/quadratic/tanh), with the first triple giving $\xi$ for $X \mid Z$ as simulated by Berrett et al. (2020) (upper three figures), and the second triple the values of $\xi$ applied when $X \mid Z$ is generated as in Section 6.4 (lower three figures).

estimation is as described in the previous paragraph. For the CRT, the distribution of $X$ is estimated on the same data to which the test is applied, like in the simulation study of Berrett et al. (2020). For the $e$-statistics, a slightly different approach is taken, tailored to sequential settings. We start with a potentially small unlabeled sample, and each time a new instance is observed, the estimate of the distribution of $X$ is updated with all the data available so far. Again, all tests except for the correlation based CRT with sample size 2000 have rejection rate close to the nominal level.

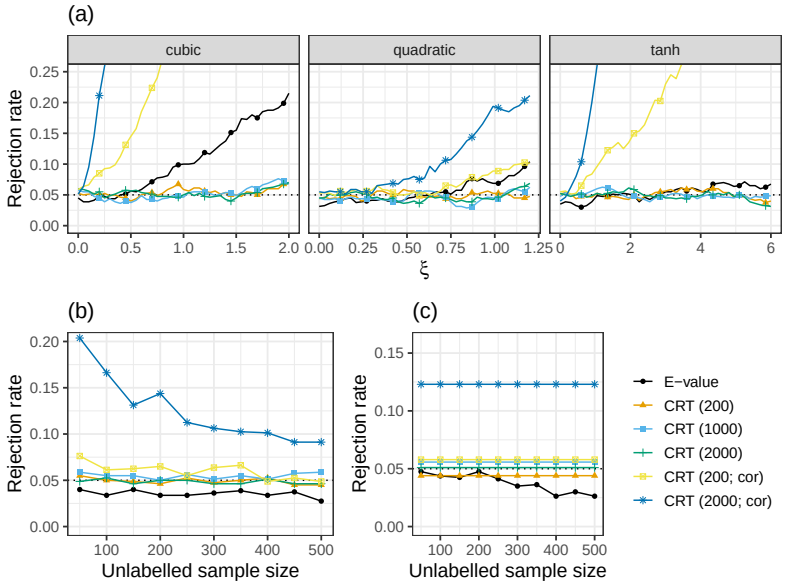



**Figure D.3:** (a) Rejection rates of $e$-values and conditional randomization test (with sample sizes $n = 200$, 1000, 2000 and likelihood as test statistic, and $n = 200$, 2000 and correlation as test statistic) at the level $\alpha = 0.05$, under different misspecifications for the conditional mean of $X$. (b) Rejection rates when the distribution of $X$ is estimated on a separate sample, for varying sample sizes. (c) Rejection rates when the same data is used both for estimating the conditional distribution of $X$ and applying the test, as described in the text.

# D.2 Proofs of Main Results

## D.2.1 Proof of Theorem 6.1

*Proof.* Let $P \in \mathcal{H}_0$ arbitrarily. The proof relies on the simple insight that we can separate the expectation with respect to $(Y_n, Z_n)$ from that with $X_n$,

$$\begin{aligned}
&\mathbb{E}_P[E_{h_n}^{\mathrm{CI}}(X_n, Y_n, Z_n) \mid D^{n-1}] \\
&\quad = \mathbb{E}_P\left[\mathbb{E}_P\left[\frac{h_n(X_n, Y_n, Z_n)}{\int_{\mathcal{X}} h_n(x, Y_n, Z_n)\,\mathrm{d}Q_{Z_n}(x)}\middle| Y_n, Z_n, D^{n-1}\right]\middle| D^{n-1}\right] \\
&\quad = \mathbb{E}_P\left[\frac{\int_{\mathcal{X}} h_n(x', Y_n, Z_n)\,\mathrm{d}Q_{Z_n}(x')}{\int_{\mathcal{X}} h_n(x, Y_n, Z_n)\,\mathrm{d}Q_{Z_n}(x)}\middle| D^{n-1}\right] = 1,
\end{aligned}$$

where in the last step we use that $P_{X_n|Y_n,Z_n} = P_{X_n|Z_n} = Q_{Z_n}$. □

### D.2.2 Proof of Proposition 6.2

*Proof.* Define $\tilde{X}_0 = X_n$. The random variables $\tilde{X}_0, \ldots, \tilde{X}_M$ are exchangeable, so

$$\check{E}^{\mathrm{CI}}_{h_n;j}(D_n) := \frac{h_n(\tilde{X}_j, Y_n, Z_n)}{\sum_{i=0}^{M} h_n(\tilde{X}_i, Y_n, Z_n)/(M+1)},\ j = 0, \ldots, M,$$

have the same expected value as $\check{E}^{\mathrm{CI}}_{h_n}(D_n) = \check{E}^{\mathrm{CI}}_{h_n;0}(D_n)$. Since $\sum_{i=0}^{M} \check{E}^{\mathrm{CI}}_{h_n;i}(D_n) \equiv M+1$, this implies $\mathbb{E}_P[\check{E}^{\mathrm{CI}}_{h_n}(D_n) \mid D^{n-1}] = 1$. □

### D.2.3 Proof of Theorem 6.3

*Proof.* Let $f = f_{X,Y,Z}(x, y, z)$ be the density of $(X, Y, Z)$ with respect to a measure $\sigma \times \mu \times \nu$ on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Then the conditional density $f_{Y|X,Z}$ equals

$$f_{Y|X,Z}(y \mid x, z) = \frac{f(x, y, z)}{\int_{\mathcal{Y}} f(x, s, z)\, \mathrm{d}\sigma(s)}.$$

The density of $Q_Z$ must equal the conditional density $f_{X|Z}$, which is given by

$$f_{X|Z}(x \mid z) = \frac{\int_{\mathcal{Y}} f(x, s, z)\, \mathrm{d}\sigma(s)}{\int_{\mathcal{X}} \int_{\mathcal{Y}} f(r, s, z)\, \mathrm{d}\sigma(s)\, \mathrm{d}\mu(r)},$$

so that, with $h(x, y, z) = f_{Y|X,Z}(y|x, z)$,

$$\int_{\mathcal{X}} h(x, y, z)\, \mathrm{d}Q_z(x) = \int_{\mathcal{X}} \frac{f(r, y, z)}{\int_{\mathcal{X}} \int_{\mathcal{Y}} f(r', s, z)\, \mathrm{d}\sigma(s)\, \mathrm{d}\mu(r')}\, \mathrm{d}\mu(r) = f_{Y|Z}(y \mid z).$$

Hence the $e$-statistic with this choice of $h$ is equal to

$$E^{\mathrm{CI}}_{f_{Y|X,Z}}(X_i, Y_i, Z_i) = \frac{f_{Y|X,Z}(Y_i \mid X_i, Z_i)}{f_{Y|Z}(Y_i \mid Z_i)} = \frac{f_{X,Y,Z}(X_i, Y_i, Z_i)}{f_{Y|Z}(Y_i \mid Z_i) f_{X|Z}(X_i \mid Z_i) f_Z(Z_i)}.$$

The denominator is the density of an element of $\mathcal{H}_0$ as in (6.1). Theorem 1 by Grünwald et al. (2024) states that this $e$-statistic must therefore be the GRO $e$-statistic for a single data point $(X_i, Y_i, Z_i)$ and the same argument can be applied to the product of these $e$-statistics. Finally, a slight rewriting shows that the $e$-statistic corresponds to the ratio of the joint conditional density of $(X, Y)$ given $Z$ divided by the product of its marginals. For all $i$, the expected value of $\log E^{\mathrm{CI}}_{f_{Y|X,Z}}(X_i, Y_i, Z_i)$ conditional on $Z$ is therefore equal to the conditional mutual information of $X$ and $Y$ given $Z$. □

### D.2.4 Proof of Proposition 6.4

*Proof.* Since the distribution $Q_Z$ is well-specified, we denote $g_{Y|Z}$ for the density $\int g_{Y|X,Z}\,\mathrm{d}Q_Z$. Then the quantity of interest is given by

$$\begin{aligned}
\mathbb{E}_f\left[\log E^{\mathrm{CI}}_{g_{Y|X,Z}}(x,y,z)\right] &= \mathbb{E}_f\left[\log \frac{g_{Y|X,Z}(y \mid x,z)}{g_{Y|Z}(y \mid z)}\right] \\
&= I_f(X;Y \mid Z) + \mathbb{E}_f\left[\log \frac{g_{Y|X,Z}(y \mid x,z)}{g_{Y|Z}(y \mid z)} - \log \frac{f(x,y,z)}{f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z) f_Z(z)}\right] \\
&= I_f(X;Y \mid Z) + \mathbb{E}_f\left[\log \frac{g_{Y|X,Z}(y \mid x,z)}{g_{Y|Z}(y \mid z)} - \log \frac{f_{Y|X,Z}(y \mid x,z)}{f_{Y|Z}(y \mid z)}\right] \\
&= I_f(X;Y \mid Z) + \mathbb{E}_f[\mathrm{KL}(f_{Y|Z} \| g_{Y|Z})] - \mathbb{E}_f[\mathrm{KL}(f_{Y|X,Z} \| g_{Y|X,Z})].
\end{aligned}$$

The desired result follows from the nonnegativity of KL divergence. □

### D.2.5 Proof of Theorem 6.6

*Proof.* Fix $N \in \mathbb{N}$ and $\alpha \in (0,1)$. Conditional on $Y_i, Z_i$, $i = 1, \ldots, N$, the randomness of the process $S_n = S_n(X^n) = \prod_{i=1}^n \tilde{E}^{\mathrm{CI}}_{h_n}$, $n = 1, \ldots, N$, solely stems from $X_1, \ldots, X_N$, and we will write $Y_i$, $Z_i$ with lower case letters $y_i$, $z_i$ to reflect that all statements are conditional on their values. So the $e$-value at time $n$ writes as

$$\tilde{E}^{\mathrm{CI}}_{h_n} = \frac{h_n(X_n, y_n, z_n \mid X^{n-1}, y^{n-1}, z^{n-1})}{\int_{\mathcal{X}} h_n(x, y_n, z_n \mid X^{n-1}, y^{n-1}, z^{n-1})\,\mathrm{d}\hat{Q}_{z_n}(x)}.$$

The condition $h_n > 0$ ensures that this $e$-value is well-defined. For $n > N$, set $h_n \equiv 1$, so that $S_n = S_N$ for $n > N$. If $X^N$ has distribution $\hat{Q}^N_{Z^N}$, then the process $(S_n)_{n\in\mathbb{N}}$ is a nonnegative martingale with respect to the filtration $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$, because

$$\mathbb{E}\left[S_n | X^{n-1}\right] = \int_{\mathcal{X}} \frac{h_n(x, y_n, z_n \mid X^{n-1}, y^{n-1}, z^{n-1})}{\int_{\mathcal{X}} h_n(x, y_n, z_n \mid X^{n-1}, y^{n-1}, z^{n-1})\,\mathrm{d}\hat{Q}_{Z_n}(x)}\,\mathrm{d}\hat{Q}_{Z_n}(x) = 1$$

almost surely. Hence by Ville's inequality, $P\left(\exists\, n \le N \colon S_n \ge 1/\alpha\right) \le \alpha$. Let

$$A = \left\{x^n \in \mathcal{X}^n \colon \exists\, n \le N \text{ s.t. } S_n(x^n) \ge 1/\alpha\right\}.$$

Then, since $Q^N_{Z^N}(A) = P(\exists\, n \le N \colon S_n \ge 1/\alpha \mid Y^N = y^N, Z^N = z^N)$,

$$P(\exists\, n \le N \colon S_n \ge 1/\alpha \mid Y^N, Z^N) \le \hat{Q}^N_{z_N}(A) + d_{\mathrm{TV}}(Q^N_{Z^N}, \hat{Q}^N_{Z^N}) \le \alpha + d_{\mathrm{TV}}(Q^N_{Z^N}, \hat{Q}^N_{Z^N}).$$

□

### D.2.6 Proof of Proposition 6.7

*Proof.* The subgaussianity assumption (i) implies that

$$P\left(|u^\top((X,Z)-\mathbb{E}[(X,Z)])|\geq\eta\right)\leq 2\exp(-\eta^2/(2\|u\|^2\sigma^2))\,,\ \eta>0,\ u\in\mathbb{R}^{p+q},\quad\text{(D.1)}$$

and that $\mathbb{E}[\|(X,Z)\|^k]<\infty$ for all $k\in\mathbb{N}$. As a consequence of the latter and of assumption (i)(a), Theorem 1 of Qian and Field (2002) implies that the MLE $\hat{\theta}_n$ exists with asymptotic probability one and satisfies $\|\hat{\theta}_n-\theta\|=\mathcal{O}(n^{-1/2}\log(\log(n))^{1/2})$ almost surely.

We now study the properties of the function $\theta\mapsto\log(p_\theta(y\mid x,z))$ for $\theta\in\mathbb{R}^{p+q}$. The derivative of $\log(p_\theta(y\mid x,z))$ with respect to $\theta_j$ equals

$$\frac{d}{d\theta_j}\log(p_\theta(y\mid x,z))=\begin{cases}yx_j-x_jp_\theta(1\mid x,z) & \text{if } j\leq p\\ yz_{j-p},-z_{j-p}p_\theta(1\mid x,z) & \text{else.}\end{cases}$$

Consequently, for any $\theta,\theta'\in\mathbb{R}^{p+q}$,

$$|\log(p_\theta(y\mid x,z))-\log(p_{\theta'}(y\mid x,z))|\leq\|(x,z)\|\|\theta-\theta'\|.\quad\text{(D.2)}$$

This implies that

$$\begin{aligned}\frac{1}{n}\Big|\sum_{i=1}^n\log(p_{\hat{\theta}_{i-1}}(Y_i\mid X_i,Z_i))-\log(p_\theta(Y_i\mid X_i,Z_i))\Big|&\leq\frac{1}{n}\sum_{i=1}^n\|\hat{\theta}_{i-1}-\theta\|\|(X_i,Z_i)\|\\&\leq\left(\frac{1}{n}\sum_{i=1}^n\|\hat{\theta}_{i-1}-\theta\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n\|(X_i,Z_i)\|^2\right)^{1/2}.\end{aligned}$$

Since $\|(X_i,Z_i)\|^2$, $i\in\mathbb{N}$, are independent with expectation $\mathbb{E}[\|(X,Z)\|^2]<\infty$, the law of large number implies that $\sum_{i=1}^n\|(X_i,Z_i)\|^2/n\to\mathbb{E}[\|(X,Z)\|^2]<\infty$ almost surely, and $\sum_{i=1}^n\|\hat{\theta}_{i-1}-\theta\|^2/n\to 0$ since $\|\hat{\theta}_n-\theta\|\to 0$ almost surely as $n\to\infty$. It remains to show an analogous convergence result for the denominator in $S_n^{CI}$. Define

$$r_n=\frac{\int p_\theta(Y_n\mid x,Z_n)\,\mathrm{d}Q_{Z_n}(x)}{\int p_{\hat{\theta}_{n-1}}(Y_n\mid x,Z_n)\,\mathrm{d}Q_{Z_n}(x)}.$$

We want to show that $\liminf_{n\rightarrowtail\infty}\sum_{i=1}^n\log(r_i)/n\geq 0$ almost surely. To this end,

write

$$r_n = \frac{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)}{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x) + \int (p_{\hat{\theta}_{n-1}}(Y_n \mid x, Z_n) - p_\theta(Y_n \mid x, Z_n))\,\mathrm{d}Q_{Z_n}(x)}$$
$$\geq \frac{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)}{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x) + \int |p_{\hat{\theta}_{n-1}}(Y_n \mid x, Z_n) - p_\theta(Y_n \mid x, Z_n)|\,\mathrm{d}Q_{Z_n}(x)}.$$

Since $\log(1+x) \leq x$, we have $\log(1/(1+x)) = -\log(1+x) \geq -x$, for $x > -1$. So

$$\log(r_n) \geq -\frac{\int |p_{\hat{\theta}_{n-1}}(Y_n \mid x, Z_n) - \int p_\theta(Y_n \mid x, Z_n)|\,\mathrm{d}Q_{Z_n}(x)}{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)}$$

The function $\theta \mapsto p_\theta(y \mid x, z)$ is Lipschitz continuous, because for $k = 1, \ldots, p+q$,

$$\Big|\frac{d}{d\theta_k} p_\theta(y \mid x, z)\Big| = \begin{cases} |x_k| p_\theta(1 \mid x, z)(1 - p_\theta(1 \mid x, z)) \leq |x_k| & \text{if } k = 1, \ldots, p \\ |z_{k-p}| p_\theta(1 \mid x, z)(1 - p_\theta(1 \mid x, z)) \leq |z_{k-p}| & \text{else.} \end{cases}$$

This implies that

$$\log(r_n) \geq -\frac{\|\hat{\theta}_{n-1} - \theta\| \int \|(x, Z_n)\|\,\mathrm{d}Q_{Z_n}(x)}{\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)}.$$

To bound this from below, we now show that the denominator $\int p_\theta(Y_n \mid x, Z_n)\,\mathrm{d}Q_{Z_n}(x)$ is small only with a small probability. Let $\kappa_n = n^{-\delta}/2$ for $\delta > 0$. Define the events

$$A_n = \left\{ \min_{y=0,1} p_\theta(y \mid X_n, Z_n) \leq \kappa_n \right\}.$$

Let $\mathrm{logit}(p) = \log(p/(1-p))$. Then,

$$\min_{y=0,1} p_\theta(y \mid x, z) \leq \kappa_n \iff |\theta^\top (x, z)| \geq |\mathrm{logit}(\kappa_n)|,$$

and therefore, since $|\mathrm{logit}(p)| \geq |\log(2p)|$ for $p \in (0, 1/2]$,

$$A_n \subseteq \{|\theta^\top (X_n, Z_n)| \geq |\log(2\kappa_n)|\} = \{|\theta^\top (X_n, Z_n)| \geq \delta \log(n)\},$$

The above derivations yield $P(A_n) \leq P(|\theta^\top (X_n, Z_n)| \geq \delta \log(n))$, and (D.1) implies,

with $B = \|\theta\|$,

$$P(|\theta^\top (X,Z)| \geq \delta \log(n)) \leq P\left(|\theta^\top ((X,Z) - \mathbb{E}[(X,Z)])| \geq \delta \log(n)) - |\theta^\top \mathbb{E}[(X,Z)]|\right) \\ \leq 2\exp(-\delta^2 \log(n)^2/(8B^2\sigma^2)),$$

for $n$ large enough such that $\delta \log(n)/2 \geq |\theta^\top \mathbb{E}[(X,Z)]|$. In a next step, we use this to bound $\min_{y=0,1} \int p_\theta(y \mid x, Z_n)\, \mathrm{d}Q_{Z_n}(x)$. First, note that for $y \in \{0,1\}$,

$$\begin{aligned}\int p_\theta(y \mid x, Z_n)\, \mathrm{d}Q_{Z_n}(x) &= \int p_\theta(y \mid x, Z_n) 1\{p_\theta(y \mid x, Z_n) \geq 1 - \kappa_n\}\, \mathrm{d}Q_{Z_n}(x) \\ &\quad + \int p_\theta(y \mid x, Z_n) 1\{p_\theta(y \mid x, Z_n) < 1 - \kappa_n\}\, \mathrm{d}Q_{Z_n}(x) \\ &\leq Q_{Z_n}(p_\theta(y \mid X_n, Z_n) \geq 1 - \kappa_n) + 1 - \kappa_n.\end{aligned}$$

It follows that for $\eta > 0$, if $\int p_\theta(y \mid x, Z_n)\, \mathrm{d}Q_{Z_n}(x) \geq 1 - n^{-\eta}$, then $Q_{Z_n}(p_\theta(y \mid X_n, Z_n) \geq 1 - \kappa_n) \geq \kappa_n - n^{-\eta}$. Recall that $\kappa_n = n^{-\delta}/2$ with $\delta > 0$ unspecified so far. For $n$ large enough such that $n^{-\eta/2} \leq 1/4$, choosing $\delta = \eta/2$ implies $\kappa_n - n^{-\eta} = n^{-\eta/2}(1/2 - n^{-\eta/2}) \geq n^{-\eta/2}/4$. Consequently, for large $n$, by Markov's inequality,

$$\begin{aligned}&P\left(\int p_\theta(y \mid x, Z_n)\, \mathrm{d}Q_{Z_n}(x) \geq 1 - n^{-\eta}\right) \\ &\leq P\left(Q_{Z_n}(p_\theta(y \mid X_n, Z_n) \geq 1 - \kappa_n) \geq n^{-\eta/2}/4\right) \\ &\leq 4n^{\eta/2}\mathbb{E}[Q_{Z_n}(p_\theta(y \mid X_n, Z_n) \geq 1 - \kappa_n)] \\ &= 4n^{\eta/2} P(p_\theta(y \mid X_n, Z_n) \geq 1 - \kappa_n). \qquad \text{(D.3)}\end{aligned}$$

But it has already been shown that

$$P(p_\theta(y \mid X_n, Z_n) \geq 1 - \kappa_n) = P(p_\theta(1-y \mid X_n, Z_n) \leq \kappa_n) \leq 2\exp(-\delta^2 \log(n)^2/(8B^2\sigma^2))$$

for large $n$, which in (D.3) gives an upper bound of

$$8\exp\left(-\log(n)(\eta^2 \log(n)/(32B^2\sigma^2) - \eta/2)\right).$$

Since $\eta^2 \log(n)/(32B^2\sigma^2) - \eta/2 \to \infty$ as $n \to \infty$, it holds that $\eta^2 \log(n)/(32B^2\sigma^2) - \eta/2 > 1$ for $n$ large enough, and we can conclude

$$\sum_{n=1}^{\infty} P\left(\min_{y=0,1} \int p_\theta(y \mid x, Z_n)\, \mathrm{d}Q_{Z_n}(x) \leq n^{-\eta}\right) < \infty.$$

Thus the Borel-Cantelli Lemma implies that $\min_{y=0,1} \int p_\theta(y \mid x, Z_n) \, \mathrm{d}Q_{Z_n}(x) \leq n^{-\eta}$ holds for only finitely many $n$ with probability one. Now

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \log(r_i) &\geq -\frac{1}{n} \sum_{i=1}^{n} \frac{\|\hat{\theta}_{i-1} - \theta\| \int \|(x, Z_i)\| \, \mathrm{d}Q_{Z_i}(x)}{\int p_\theta(Y_i \mid x, Z_i) \, \mathrm{d}Q_{Z_i}(x)} \\
&\geq -\frac{M}{n} - \sum_{i=1}^{n} i^{\eta} \|\hat{\theta}_{i-1} - \theta\| \int \|(x, Z_i)\| \, \mathrm{d}Q_{Z_i}(x) \\
&= -\frac{M}{n} - \frac{1}{n} \sum_{i=1}^{n} i^{\eta} \|\hat{\theta}_{i-1} - \theta\| \mathbb{E}[\|(X_i, Z_i)\| \mid Z_i] \\
&\geq -\frac{M}{n} - \left( \frac{1}{n} \sum_{i=1}^{n} i^{2\eta} \|\hat{\theta}_{i-1} - \theta\|^2 \right)^{1/2} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\|(X_i, Z_i)\| \mid Z_i]^2 \right)^{1/2},
\end{aligned} \tag{D.4}$$

where

$$M = \sum_{i=1}^{\infty} 1 \left\{ \int p_\theta(Y_i \mid x, Z_i) \, \mathrm{d}Q_{Z_i}(x) \leq i^{-\eta} \right\} \frac{\|\hat{\theta}_{i-1} - \theta\| \int \|(x, Z_i)\| \, \mathrm{d}Q_{Z_i}(x)}{\int p_\theta(Y_i \mid x, Z_i) \, \mathrm{d}Q_{Z_i}(x)}$$

is the sum of $\log(r_i)$ over all almost surely finitely many $i$ s.t. $\int p_\theta(Y_i \mid x, Z_i) \, \mathrm{d}Q_{Z_i}(x) \leq i^{-\eta}$. Since $(X_i, Z_i)$, $i \in \mathbb{N}$, are independent and identically distributed with

$$\mathbb{E}[\mathbb{E}[\|(X, Z)\| | Z]^2] \leq \mathbb{E}[\|(X, Z)\|^2] < \infty,$$

the law of large numbers implies

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\|(X_i, Z_i)\| \mid Z_i]^2 \leq \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\|(X_i, Z_i)\|^2 \mid Z_i] \to \mathbb{E}[\|(X, Z)\|^2] < \infty$$

almost surely as $n \to \infty$. On the other hand, $n^{2\eta} \|\hat{\theta}_{n-1} - \theta\|^2 = \mathcal{O}(n^{2\eta-1} \log(\log(n)))$ almost surely, so that for $\eta < 1/2$, we have $n^{2\eta} \|\hat{\theta}_{n-1} - \theta\|^2 \to 0$ almost surely as $n \to \infty$. Finally, since $M$ only takes finite values, also $M/n \to 0$ for $n \to \infty$. Hence (D.4) converges to 0 almost surely. It follows that

$$\liminf_{n \to \infty} \frac{1}{n} \left( \log(S_n^{CI}) - \log \left( \prod_{i=1}^{n} \frac{p_\theta(Y_i \mid X_i, Z_i)}{\int p_\theta(Y_i \mid x, Z_i) \, \mathrm{d}Q_{Z_i}(x)} \right) \right) \geq 0$$

almost surely. Since

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{p_\theta(Y_i \mid X_i, Z_i)}{\int p_\theta(Y_i \mid x, Z_i)\,\mathrm{d}Q_{Z_i}(x)}\right) \to I(X;Y \mid Z) > 0,\ n \to \infty,$$

almost surely, by the law of large numbers, this proves the theorem. □

## D.3 Anytime-Valid E-Statistics

In this section, we discuss an alternative way to define anytime-valid tests using *e*-statistics and show that, in the setting of Chapter 6, this method coincides with the method discussed in Section 6.2.2. In Section 6.2.2, we mentioned that a sequence of conditional *e*-statistics gives rise to a test martingale $(S_n(D^n))_{n\in\mathbb{N}}$, which satisfies $\mathbb{E}_P[S_\tau(D^\tau)] \leq 1$ for any stopping time $\tau$ and $P \in \mathcal{H}_0$. Rather than taking the latter as a consequence, Koolen and Grünwald (2022) take this as the definition of what they call *anytime-valid* *e*-statistics. That is, they call a nonnegative process $(E_n(D^n))_{n\in\mathbb{N}}$ an anytime-valid *e*-statistic if $\mathbb{E}_P[E_\tau(D^\tau)] \leq 1$ for any stopping time $\tau$ and $P \in \mathcal{H}_0$. The same object is referred to as *e*-process in Ramdas et al. (2022), and it can be shown that the class of anytime-valid *e*-statistics (or *e*-processes) is strictly larger than the class of test martingales. A priori it is not obvious whether the GRO criterion, which maximizes the expected growth rate without referring to any particular stopping time, also yields powerful *e*-statistics when specific stopping rules $\tau$ are applied. Therefore, Koolen and Grünwald (2022) propose, for fixed alternative distribution $\mathcal{H}_1 = \{P^*\}$ and stopping time $\tau$, to look for the anytime-valid *e*-statistic that maximizes

$$(E_n)_{n\in\mathbb{N}} \mapsto \mathbb{E}_{P^*}[\log E_\tau(D^\tau)]. \tag{D.5}$$

It turns out that there are settings in which the optimal anytime-valid *e*-statistic is actually equal to the GRO test martingale. One of the settings in which this happens is given in their Theorem 12. We present a slightly rephrased version of this theorem here.

**Theorem D.1** (Koolen and Grünwald (2022))**.** *Assume that the data is given by an i.i.d. stream $(D_i)_{i\in\mathbb{N}}$ and that the alternative is given by $\mathcal{H}_1 = \{P^*\}$, where $P^*$ admits a density $p^*$. Suppose further that the GRO e-statistic is given by the likelihood ratio $p^*/q$, where $q$ is the density of an element of $\mathcal{H}_0$. Then the process $(p^*(D_i)/q(D_i))_{i\in\mathbb{N}}$ also maximizes* (D.5) *for any stopping time $\tau$.*

In the proof of our Theorem 6.3 (see Section D.2.3), we show that the GRO $e$-variable is exactly of the form described in Theorem D.1. It therefore follows that the test martingale that we give in (6.7) is actually also the optimal anytime-valid $e$-statistic. We therefore chose to focus on the GRO property in Chapter 6.

# E | Appendix to Chapter 7

## E.1 Proofs

### E.1.1 Proof of Proposition 7.4

*Proposition 7.4.* For $X^{n-1} \in \mathcal{X}^{n-1}$ and $X_n \in \mathcal{X}$, define $F_n(\gamma_{n-1}(X^{n-1}), X_n) = \gamma_n((\gamma_{n-1}(X^{n-1}), X_n))$, where, with a slight abuse of notation, we use $(\gamma_{n-1}(X^{n-1}), X_n)$ to refer to the concatenation of $\gamma_{n-1}(X^{n-1})$ and $X_n$. We will show that $F_n$ has the claimed properties. First, we will show that the vectors $(\gamma_{n-1}(X^{n-1}), X_n)$ and $X^n$ are in the same orbit, so that also $\gamma_n((\gamma_{n-1}(X^{n-1}), X_n)) = \gamma_n(X^n)$. To this end, let $g' \in G_{n-1}$ denote the group element such that $g'X^{n-1} = \gamma_{n-1}(X^{n-1})$. Then it holds that

$$\begin{aligned}\{g(\gamma_{n-1}(X^{n-1}), X_n) : g \in G_n\} &= \{g(g'X^{n-1}, X_n) : g \in G_n\} \\ &= \{g\iota_n(g')X^n : g \in G_n\} \\ &= \{gX^n : g \in G_n\},\end{aligned}$$

where we used (iii) of Definition 7.2 for the second equality and called $X^n$ the concatenation of $X^{n-1}$ and $X_n$. This shows the first claim. For the second claim, that $F_n(\,\cdot\,, X_n)$ is one-to-one for each fixed $X_n$, we show that we can reconstruct $\gamma_{n-1}(X^{n-1})$ from $X_n$ and $\gamma_n(X^n)$.

Pick any $g_{X_n} \in G_n$ such that $(g_{X_n}\gamma_n(X^n))_n = X_n$. We furthermore know that there exists some $g \in G_n$ such that $gX^n = \gamma_n(X^n)$. Note that $g_{X_n}g$ does nothing to the final coordinate of $X^n$, so by item (iii) of Definition 7.2 there is a $g^*_{n-1} \in G_{n-1}$

such that $g_{X_n} g X^n = \imath(g_{n-1}^*) X^n$. Then we see

$$\begin{aligned}\{\imath(g_{n-1}) g_{X^n} \gamma_n(X^n) : g_{n-1} \in G_{n-1}\} &= \{\imath(g_{n-1}) g_{X^n} g X^n : g_{n-1} \in G_{n-1}\} \\ &= \{\imath(g_{n-1}) \imath(g_{n-1}^*) X^n : g_{n-1} \in G_{n-1}\} \\ &= \{\imath(g_{n-1}) X^n : g_{n-1} \in G_{n-1}\}.\end{aligned}$$

It follows from item (ii) of Definition 7.2 that $G_{n-1}\text{proj}_{n-1}(g_{X_n}\gamma_n(X^n)) = G_{n-1}X^{n-1}$. It therefore follows that $\gamma_{n-1}(\text{proj}_{n-1}(g_{X_n}\gamma_n(X^n))) = \gamma_{n-1}(X^{n-1})$. □

## E.1.2 Proof of Theorem 7.8

*Theorem 7.8.* The proof can be divided in two main steps: (1) to show that, conditionally on $\gamma_n(X^n)$, $R_n$ is uniformly distributed for each $n$ and (2) to show that $R_1, R_2, \dots$ are also independent. The second step is completely analogous to the proof of Theorem 3 by Vovk (2002). For each $n$, define the $\sigma$-algebra $\mathcal{G}_n = \sigma(\gamma_n(X^n), X_{n+1}, X_{n+2}, \dots)$. Notice that $\mathcal{G}_n$ contains—among others—all $G_n$-invariant functions of $X^n$ because $\gamma_n$ is a maximally invariant function of $X^n$—any other $G_n$-invariant function of $X^n$ is a function of $\gamma_n(X^n)$. Let $g' \in G_n$ such that $\gamma_n(X^n) = g'X^n$, then we have that $\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n < \alpha_n\} = \{g \in G_n : A((g\gamma_n(X^n)_n, \gamma_n(X^n))_n < \alpha_n\}g'$. Here, we define $Bg = \{bg : b \in B\}$ for a subset $B \subseteq G_n$. By the invariance of $\mu_n$—it is the Haar probability measure—, it follows that

$$\begin{aligned}&\mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n < \alpha_n\}) \\ &= \mu_n(\{g \in G_n : A((g\gamma_n(X^n))_n, \gamma_n(X^n))_n < \alpha_n\}).\end{aligned}$$

An analogous identity can be derived for the second term in (7.3). We have $\alpha_n \mid \mathcal{G}_n \overset{\mathcal{D}}{=} A((U\gamma_n(X^n))_n, \gamma_n(X^n))_n \mid \mathcal{G}_n$.

We will denote $F(b) := \mu(\{g \in G_n : A((g\gamma_n(X^n))_n, \gamma_n(X^n))_n < b\})$ and define $G(\delta) = \sup\{b \in \mathbb{R} : F(b) \leq \delta\}$. If $\alpha_n \mid \mathcal{G}_n$ is continuous, then $F$ is the CDF of that distribution, otherwise it is the CDF minus the probability of equality. In any case, $F$ is is increasing and right-continuous. For any $\delta \in (0,1)$, we have that $F(G(\delta)) = \delta'$ for some $\delta' \leq \delta$, with equality if $F$ is continuous in $G(\delta)$. Then we can write

$$\mathbb{P}(R_n \leq \delta \mid \mathcal{G}_n) = \mathbb{P}(R_n \leq \delta' \mid \mathcal{G}_n) + \mathbb{P}(\delta' < R_n \leq \delta \mid \mathcal{G}_n). \tag{E.1}$$

For any $\theta \in (0,1]$, we have that $R_n = F(\alpha_n) + \theta(F(\alpha_n^+) - F(\alpha_n)) \leq \delta'$ if and only if either $F(\alpha_n) < \delta'$ or $F(\alpha_n^+) - F(\alpha_n) = 0$, which happens precisely when $\alpha_n < G(\delta)$.

We therefore see

$$\mathbb{P}(R_n \leq \delta' \mid \mathcal{G}_n) = \mathbb{P}(\alpha_n < G(\delta') \mid \mathcal{G}_n) = F(G(\delta')) = \delta'.$$

If $F$ is continuous in $G(\delta)$, then this shows that $\mathbb{P}(R_n \leq \delta \mid \mathcal{G}_n) = \delta$, since $\delta' = \delta$ in that case. If $F$ is not continuous in $G(\delta)$, then we have that

$$\mathbb{P}(\delta' < R_n \leq \delta \mid \mathcal{G}_n) = \mathbb{P}(\delta' < F(\alpha_n) + \theta(F(\alpha_n^+) - F(\alpha_n)) \leq \delta \mid \mathcal{G}_n).$$

Notice that $\delta' < F(\alpha_n) + \theta(F(\alpha_n^+) - F(\alpha_n)) \leq \delta$ if and only if $\alpha_n = G(\delta)$ and $\theta < (\delta - \delta')/(F(\alpha_n^+) - F(\alpha_n))$, so that we can write

$$\begin{aligned}
\mathbb{P}(\delta' < R_n \leq \delta \mid \mathcal{G}_n) &= \mathbb{P}(\alpha_n = G(\delta) \mid \mathcal{G}_n)\mathbb{P}\left(\theta \leq \frac{\delta - \delta'}{F(G(\delta')^+) - F(G(\delta'))} \mid \mathcal{G}_n\right) \\
&= (F(G(\delta')^+) - F(G(\delta')))\frac{\delta - \delta'}{(F(G(\delta')^+) - F(G(\delta')))} \\
&= \delta - \delta'.
\end{aligned}$$

Putting everything together, we see that $\mathbb{P}(R_n \leq \delta \mid \mathcal{G}_n) = \delta$. This shows the first part, that $R_n$ has a conditional uniform distribution on $[0, 1]$.

For the second part of the proof, we show that the sequence $R_1, R_2, \ldots$ is also an independent sequence. We have that $R_n$ is $\mathcal{G}_{n-1}$-measurable because it is invariant under transformations of the form $X^n \mapsto (gX^{n-1}, X_n)$ for $g \in G_{n-1}$ (see also Vovk, 2004, Lemma 2). We proceed (implicitly) by induction:

$$\begin{aligned}
\mathbb{P}(R_n \leq \delta_n, \ldots, R_1 \leq \delta_1 \mid \mathcal{G}_n) &= \mathbf{E}\left[\mathbf{1}\left\{R_n \leq \delta_n, \ldots, R_1 \leq \delta_1\right\} \mid \mathcal{G}_n\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}\left\{R_n \leq \delta_n, \ldots, R_1 \leq \delta_1\right\} \mid \mathcal{G}_{n-1}\right] \mid \mathcal{G}_n\right] \\
&= \mathbf{E}\left[\mathbf{1}\left\{R_n \leq \delta_n\right\}\mathbf{E}\left[\mathbf{1}\left\{p_{n-1} \leq \delta_{n-1}, \ldots, R_1 \leq \delta_1\right\} \mid \mathcal{G}_{n-1}\right] \mid \mathcal{G}_n\right] \\
&= \mathbf{E}\left[\mathbf{1}\left\{R_n \leq \delta_n\right\}\right]\delta_{n-1} \cdots \delta_1 \\
&= \delta_n \cdots \delta_1.
\end{aligned}$$

It follows by the law of total expectation that

$$\mathbb{P}(R_n \leq \delta_n, \ldots, R_1 \leq \delta_1) = \delta_n \cdots \delta_1,$$

which shows that $R_1, R_2, \ldots, R_n$ are independent and uniformly distributed on $[0, 1]$ for any $n \in \mathbb{N}$. This implies that the distribution of $R_1, R_2, \ldots$ coincides with $U^\infty$

by Kolmogorov's extension theorem (see e.g. Shiryaev, 2016, Theorem II.3.3). This shows the claim of the theorem. □

### E.1.3 Proof of Proposition 7.9

The proof of Proposition 7.9 follows directly from Lemma E.1. It states that, with probability one, enough of the original data can be recovered using the smoothed ranks and the orbit representative. We state Lemma E.1, prove Proposition 7.9 and then prove Lemma E.1.

**Lemma E.1.** *Suppose, for each $n \in \mathbb{N}$, that $A(\,\cdot\,, \gamma_n(X^n))$ is a one-to-one function of $X_n$, then there exists a map $D_n : [0,1]^n \times \mathcal{X}^n \to [0,1]^n \times \mathcal{X}^n$ s.t. for any $Q \in \mathcal{H}_0$, $\widetilde{Q}(D_n(R^n, \gamma_n(X^n)) = (\widetilde{\theta}^n, X^n)) = 1$. Here, $\widetilde{\theta}^n = (\widetilde{\theta}_n)_{n\in\mathbb{N}}$ is the sequence given by $\widetilde{\theta}_n = \theta_n \mathbf{1}\{\mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n = \alpha_n\}) \neq 0\}$.*

*Proposition 7.9.* Consider, without loss of generality, the case that $A(X^n, \gamma_n(X^n)) = X_n$. Because of the independence of $R_n$ and $\gamma_n$ under $P$ and the assumption that the marginal distribution of $\gamma_n$ under $Q^*$ and under $P$ are equal, $M_n = \frac{\mathrm{d}\widetilde{P}(R^n, \gamma_n(X^n))}{\mathrm{d}\widetilde{Q}^*(R^n, \gamma_n(X^n))}$. Using the sequence of functions $(D_n)_{n\in\mathbb{N}}$ from Lemma E.1 and that the external randomization is independent of $X^n$, the claim follows. □

*Lemma E.1.* As in the proof of Theorem 7.8, we will denote $F(b) = \mu_n(\{g \in G_n : A((gX^n)_n, \gamma_n(X^n))_n < b\})$ and define $G(\delta) = \sup\{b \in \mathbb{R} : F(b) \leq \delta\}$. Furthermore, we will write $\mathbb{P}_{\alpha_n|\gamma_n(X^n)}$ for the distribution of $\alpha_n$ given $\gamma_n(X^n)$ and denote its support by

$$\operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}) := \{x \in \mathbb{R} \mid \text{for all } I \text{ open, if } x \in I \text{ then } \mathbb{P}_{\alpha_n|\gamma_n(X^n)}(I) > 0\},$$

If $b \in \operatorname{int}(\operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}))$, then there exists an open interval $B$ with $b \in B$ and $B \subseteq \operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}))$. For all $c \in B$ with $c > b$, we have that $F(c) - F(b) = \mathbb{P}_{\alpha_n|\gamma_n(X^n)}([b, c)) > 0$, since $[b, c)$ contains an open neighborhood of an interior point of the support. It follows that $F(c) > F(b)$. In words, there are no points $c$ to the right of $b$ such that $F(c) > F(b)$. Consequently, we have

$$G(F(b)) = \sup\{a \in \mathbb{R} : F(a) \leq F(b)\} = b.$$

In a similar fashion, we can conclude that the same identity holds whenever $b \in \operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}) \backslash \operatorname{int}(\operatorname{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)}))$. Notice furthermore that $G(R_n) = G(F(\alpha_n)+$

$\theta_n(F(\alpha_n^+) - F(\alpha_n))) = G(F(\alpha_n))$ whenever $\theta_n < 1$, which happens with probability one. Together with the fact that $\mathbb{P}_{\alpha_n|\gamma_n(X^n)}(\mathrm{supp}(\mathbb{P}_{\alpha_n|\gamma_n(X^n)})) = 1$, this gives $\mathbb{P}_{\alpha_n|\gamma_n(X^n)}(G(R_n) = \alpha_n) = 1$, so also $\mathbb{P}(G(R_n) = \alpha_n) = 1$. If $(F(G(R_n)^+) - F(G(R_n))) = \mu_n(\{g \in G_n : (g\alpha^n)_n = \alpha_n\}) = 0$, set $\widetilde{\theta}_n = 0$. If $\mu_n(\{g \in G_n : (g\alpha^n)_n = \alpha_n\}) > 0$, then it follows that $\mathbb{P}(\theta_n = (R_n - F(G(R_n)))/(F(G(R_n)^+) - F(G(R_n)))) = 1$, so set $\widetilde{\theta}_n = (R_n - F(G(R_n)))/(F(G(R_n)^+) - F(G(R_n)))$. Since $A(\cdot, \gamma_n(X^n))$ is one-to-one by assumption, its inverse maps $\alpha_n$ to $X_n$. By Proposition 7.4, there also exists a map from $X_n$ and $\gamma_n(X^n)$ to $\gamma_{n-1}(X^{n-1})$. At this point, we can repeat the procedure above to recover $X_{n-1}$ from $(R_{n-1}, \gamma_{n-1}(X^{n-1}))$, from which we can then recover $\gamma_{n-2}(X^{n-2})$, etc. Together, all of the maps involved give the function as in the statement of the proposition. □

### E.1.4 Proof of Theorem 7.10

*Theorem 7.10.* We first show (7.6). Assume that $\widetilde{P}$ is such that $R^n \perp \gamma_n(X^n)$ for all $n$. Let $Q^*$ denote the distribution under which the marginal of $\gamma_n(X^n)$ coincides with that under $P$, and such that $X^n \mid \gamma_n(X^n) \overset{\mathcal{D}}{=} U\gamma_n(X^n) \mid \gamma_n(X^n)$, where $U \sim \mu_n$ is uniform on $G_n$ and independent from $\gamma_n(X^n)$. First note that

$$\begin{aligned}
\widetilde{Q}^*\left(\prod_{i=1}^{\tau} f_i(R_i) = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^\tau)\right) &\geq \widetilde{Q}^*\left(\forall t : \prod_{i=1}^{t} f_i(R_i) = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^t)\right) \\
&= 1 - \widetilde{Q}^*\left(\exists t : \prod_{i=1}^{t} f_i(R_i) \neq \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^t)\right) \\
&= 1 - \widetilde{Q}^*\left(\bigcup_{t=1}^{\infty}\left\{\prod_{i=1}^{t} f_i(R_i) \neq \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^t)\right\}\right) \\
&\geq 1 - \sum_{t=1}^{\infty} \widetilde{Q}^*\left(\left\{\prod_{i=1}^{t} f_i(R_i) \neq \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^t)\right\}\right) = 1.
\end{aligned}$$

In the last inequality, we used Lemma E.1. By assumption, we have $\widetilde{P} \ll \widetilde{Q}^*$, so we also have $\widetilde{P}\left(\prod_{i=1}^{\tau} f_i(R_i) = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^\tau)\right) = 1$. We have shown that $M_\tau$ is a modification of the likelihood ratio evaluated at $X^\tau$. We now show that the latter is optimal.

Denote $\ell_n = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(X^n)$ and let $f(\alpha) = \mathbf{E}_{\widetilde{P}}[\ln((1-\alpha)\ell_\tau + \alpha E'_\tau)]$; a concave function. We will show that the derivative of $f$ in 0 is negative, which implies that $f$

attains its maximum in $\alpha = 0$. This in turn implies our claim. Indeed,

$$\begin{aligned} f'(0) &= \mathbf{E}_{\widetilde{P}}\left[\frac{E'_\tau - \ell_\tau}{\ell_\tau}\right] \\ &= \sum_{i=1}^{\infty} \mathbf{E}_{\widetilde{P}}\left[\frac{E'_i}{\ell_i}\mathbf{1}\left\{\tau = i\right\}\right] - 1 \\ &= \sum_{i=1}^{\infty} \mathbf{E}_{\widetilde{Q}^*}\left[E'_i\mathbf{1}\left\{\tau = i\right\}\right] - 1 \\ &= \mathbf{E}_{\widetilde{Q}^*}\left[E'_\tau\right] - 1 \leq 0, \end{aligned}$$

where we use that differentiation and integration can be interchanged, because

$$|f'(\alpha)| = \left|\frac{E'_\tau - \ell_\tau}{(1-\alpha)\ell_\tau + \alpha E'_\tau}\right| \leq \max\left\{\frac{1}{1-\alpha}, \frac{1}{\alpha}\right\},$$

so that the dominated convergence theorem is applicable. Finally, this gives that $\mathbf{E}_{\widetilde{P}}[\ln \prod_{i=1}^{\tau} f(R_i)] = \mathbf{E}_{\widetilde{P}}[\ln E'_\tau] \geq \mathbf{E}_{\widetilde{P}}[\ln E'_\tau]$. The proof of (7.5) follows from the same argument, but using $\ell'_n = \frac{\mathrm{d}P}{\mathrm{d}Q^*}(R^n)$. □

## E.2 Linear Models and Isotropy Groups

The rotational symmetry described in Section 7.5.2 is that of symmetry around the origin, which we argued is equivalent to testing whether $X_i \sim \mathcal{N}(0, \sigma)$ for some $\sigma \in \mathbb{R}^+$. Of course, there are many applications where it is not reasonable to assume that the data is zero-mean and it is more interesting to test whether the data is spherically symmetric around some point other than the origin. One particular instance of such noncentered sphericity is to test whether, for each $n$, the data can be written as $X^n = \mu\mathbf{1}_n + \epsilon^n$, where $\mu \in \mathbb{R}$, the error $\epsilon^n$ is spherically symmetric and $\mathbf{1}_n$ is the $n$-vector of all ones. If $\mu$ is known, we can test for spherical symmetry of $X^n - \mu\mathbf{1}_n$ under $\mathrm{O}(n)$ and the problem reduces to that of the previous section. It is still possible treat the more realistic case where $\mu$ is unknown because the null model is still symmetric under a family of rotations. Notice the following: for any $O_n \in \mathrm{O}(n)$ it holds that $O_n X^n = \mu O_n \mathbf{1}_n + O_n\epsilon^n$. Unless $\mu = 0$, it follows that $X^n \overset{\mathcal{D}}{=} O_n X^n$ every time that $O_n\mathbf{1}_n = \mathbf{1}_n$. That is, the null distribution of $X^n$ is invariant under the isotropy group of $\mathbf{1}_n$, i.e. $G_n = \{O_n \in \mathrm{O}(n) : O_n\mathbf{1}_n = \mathbf{1}_n\}$. Invariance under the action of $G_n$ has previously appeared in the literature as centered spherical symmetry (Smith, 1981). Through the lens of test martingales, testing sequentially for centered spherical

symmetry is equivalent to testing whether the data was generated by any Gaussian. This holds because any probability distribution on $\mathbb{R}^\infty$ for which the marginal of the first $n$ coordinates is centered spherically symmetric for any $n$ can be written as a mixture of Gaussians (Smith, 1981; Eaton, 1989, Theorem 8.13).

Using some geometry, a test is readily obtained. Note that we can write $X^n = X^n_{\mathbf{1}_n} + X^n_{\perp \mathbf{1}_n}$, where $X^n_{\mathbf{1}_n} = \frac{\langle X^n, \mathbf{1} \rangle}{n} \mathbf{1}_n$ is the projection of $X^n$ onto the span of $\mathbf{1}_n$, and $X^n_{\perp \mathbf{1}_n}$ the projection onto its orthogonal complement. We have that $gX^n = X^n_{\mathbf{1}_n} + gX^n_{\perp \mathbf{1}_n}$ for any $g \in G_n$. Consequently, the orbit of $X^n$ under $G_n$ is given by the intersection of $S^{n-1}(\|X^n\|)$ and the hyperplane $H_n(X^n)$ defined by $H_n(X^n) = \{x'^n \in \mathbb{R}^n : \langle x'^n, \mathbf{1}_n \rangle = \langle X^n, \mathbf{1}_n \rangle\}$. There is a unique line that is perpendicular to $H_n(X^n)$ and passes through the origin $0_n = (0, \ldots, 0)$; it intersects $H_n(X^n)$ in the point $0_{H_n} := \frac{\langle X^n, \mathbf{1}_n \rangle}{n} \mathbf{1}_n$. For any $x'^n \in S^{n-1}(\|X^n\|) \cap H_n(X^n)$, Pythagoras' theorem gives that $\|x'^n - 0_{H_n}\|^2 = \|X^n\|^2 - \|0_{H_n} - 0_n\|^2$. In other words, $S^{n-1}(\|X^n\|) \cap H_n(X^n)$ forms an $(n-2)$-dimensional sphere of radius $(\|X^n\|^2 - \|0_{H_n} - 0_n\|^2)^{1/2}$ around $0_{H_n}$. If one considers the projection of this sphere on the $n$-th coordinate, then the highest possible value is given by $\|X^n\|$, and the lowest value therefore by $\frac{\langle X^n, \mathbf{1}_n \rangle}{n} - \frac{1}{2}(\|X^n\| - \frac{\langle X^n, \mathbf{1}_n \rangle}{n})$. The relative value of $X_n$ is therefore given by $\widetilde{X}_n := X_n - \frac{\langle X^n, \mathbf{1}_n \rangle}{n} + \frac{1}{2}(\|X^n\| - \frac{\langle X^n, \mathbf{1}_n \rangle}{n})$. As a result, $R_n$ is the relative surface area of the $(n-2)$-dimensional hyper-spherical cap with co-latitude angle $\varphi = \pi - \cos^{-1}(\widetilde{X}_n / (\|X^n\|^2 - \|0_{H_n} - 0_n\|^2)^{1/2})$, so that equation (7.9) can again be used to determine $R_n$. With this construction, we recover what Vovk (2023) refers to as the "full Gaussian model", which is an online compression model that is defined in terms of the summary statistic $\sigma_n = (\langle X^n, \mathbf{1}_n \rangle, \|X^n\|)$.

This model can be extended to the case in which there are covariates, i.e. $X_n = (Y_n, Z^d_n)$ for some $Y_n \in \mathbb{R}$ and $Z^d_n \in \mathbb{R}^d$. Denote $Z_n$ for the matrix with row-vectors $Z^d_n$ and, as is a standard assumption in regression, assume that $Z_n$ is full rank for every $n$. The model of interest is $Y^n = Z_n \beta + \epsilon^n$ where $\beta \in \mathbb{R}^d$ and $\epsilon^n$ is spherically symmetric for each $n$. Similar to the reasoning above, this model is invariant under the intersection of the isotropy groups of the column vectors of $Z_n$, i.e. $G_n = \{O_n \in \mathrm{O}(n) : O_n Z_n = Z_n\}$. The orbit of $X^n$ under $G_n$ is given by the intersection of $S^{n-1}(\|X^n\|)$ with the intersection of the $d$ hyperplanes defined by the columns of $Z_n$, so that for $\alpha^n(Y^n, Z_n) = Y^n$, computing $R_n$ is analogous. Interestingly, however, it does not always hold that testing for invariance under $G_n$ is equivalent to testing for normality with mean $Z_n \beta^d$. A sufficient condition for the equivalence to hold is that $\lim_{n \to \infty} (Z'_n Z_n)^{-1} = 0$, which is essentially the condition that the parameter vector $\beta$ can be consistently estimated by means of least squares (Eaton, 1989, Section 9.3).

# F | Appendix to Chapter 8

## F.1 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—that is, if the sufficient statistic is equivariant—, 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 $\Theta$ be the parameter space, and let $\delta = \delta(\theta)$ be a maximally invariant function of $\theta$ for the action of $G$ on $\Theta$. 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 (8.1). Assume further that $G$ acts freely and continuously on the image space $\mathcal{S}_n$ of the sufficient statistic $S_n = s_n(X^n)$. Denote by $(g, s) \mapsto gs$ the action of $G$ on $\mathcal{S}_n$. We assume that $s_n$ is equivariant, that is, $s_n$ 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. 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 $\delta$. 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 $\delta$. 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 (see Remark F.1).

**Theorem F.1** (C. Stein)**.** *If there exists a right Haar measure on the group $G$ and $G$ is $\sigma$-finite, 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 $\delta$.*

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 F.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 $\delta$. 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 F.1, $M_{\mathcal{S},n}$ is sufficient for $\delta$. 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 F.3.** *Under the assumptions of Proposition F.2, if $S_n = s_n(X^n)$,*

$$\frac{q^{M_{\mathcal{X},n}}(m_{\mathcal{X},n}(X^n))}{p^{M_{\mathcal{X},n}}(m_{\mathcal{X},n}(X^n))} = \frac{q^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}{p^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}.$$

*Hence, if assumptions of Corollary 8.3 also hold, the likelihood ratio for the invariantly sufficient statistic $M_{\mathcal{S},n}$ is (relatively) GROW.*

**Example F.1** (continues=ex:t-test)**.** We have seen that a maximally invariant function of the data is $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n) = (X_1/|X_1|, \ldots, X_n/|X_1|)$ 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 F.1) shows that the t-statistic $M_{\mathcal{S},n}$ is sufficient for the maximally invariant parameter $\delta = \mu/\sigma$. Corollary F.3 shows that the likelihood ratio for the t-statistic is relatively GROW.

**Remark.** In the present form, the assumptions in Theorem F.1 avoid issues that may arise with almost-invariant functions (see Lehmann and Romano, 2005, Section 6.5). Almost-invariant functions are functions that are invariant under the action of a group almost surely up to a null set that may depend on the group element in question. Under the assumptions in Theorem F.1, every almost invariant function is equivalent to an invariant one (Lehmann and Romano, 2005, Theorem 6.5.1). In turn, the assumptions in Theorem F.1 are implied by Assumption 8.1, so that the same is true in the general setting of Chapter 8. See also Hall et al. (1965, discussion in p. 581).

## F.2 Detailed Comparison to Sun and Berger (2007) and Liang and Barron (2004)

As the example in Section 8.5.1 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 $\Delta_a$ and priors $\mathbf{\Pi}_j^{*\delta_a}$ on $\Delta_a$ achieving (8.18) relative to group $G_a$, for $j = 0, 1$; group $G_b$ has parameter of interest in $\Delta_b$ and priors $\mathbf{\Pi}_j^{*\delta_b}$ achieving (8.18) 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 8.7 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 8.7 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, $G_a$ was $\mathrm{GL}(d)$ and the priors $\mathbf{\Pi}_0^{*\delta_a}, \mathbf{\Pi}_1^{*\delta_a}$ were degenerate priors on 0 and $\gamma$ as in (8.23), respectively; $G_b$ was the lower triangular group with a specific prior as indicated in the example. 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 8.8 can be used to identify the GROW $e$-statistic as soon as the assump-

tions of Proposition 8.7 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 8.8 with that tuple; then $T^*$ as defined in the corollary must be GROW, irrespective of whether $T^*$ 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 8.8 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 (8.3) 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, can depend on the choice.

In contrast, our quantity of interest, the GROW $e$-statistic $T_n^*$, is uniquely defined as soon as there exists one group $G$ with $\mathcal{H}_0$ and $\mathcal{H}_1$ as in (8.1) for which the assumptions of Theorem 8.2 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 8.7 hold, even if there exist other such tuples.

To reconcile uniqueness of the GROW $e$-statistic $T_n^*$ with nonuniqueness of the Bayes posterior, note that the former is a ratio between Bayes marginals for different models $\mathcal{H}_0$ and $\mathcal{H}_1$ at the same sample size $n$. In contrast, the Bayes predictive distribution based on a single model $\mathcal{P}$ is a ratio between Bayes marginals for the same $\mathcal{P}$ at different sample sizes $n$ and $n-1$. The role of 'same' and 'different' being interchanged, it turns out that this Bayes predictive distribution *can* depend on the group on which the right Haar prior for $\mathcal{P}$ is based. Since the Bayes predictive distribution can be rewritten as a marginal over the Bayes posterior for $\mathcal{P}$, it is then not surprising that this Bayes posterior may also change if the underlying group is changed.

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 (8.3) 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 8.4.2). Consider then such a $\mathcal{P}$ and let $p^{M_n}(m_n(X^n))$ be as in (8.5). 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.

## F.3 Anytime-Valid Testing Under Optional Stopping and Optional Continuation

Consider the setting of Section 8.2.2. 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 $\mathcal{X}^n$.

Suppose that data $X_1, X_2, \ldots$ 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 $X_1, X_2, \ldots$. We consider the test defined by $\xi_n = \mathbf{1}\left\{T^{M_n} \geq 1/\alpha\right\}$ for some value $\alpha$. We note that Wald-style—Sequential Probability Ratio Tests—tests are different because they would output "no decision" until a particular sample size $n$. Afterwards, they would output 1 ("reject the null") or 0 ("there is no evidence to reject the null") forever. In contrast, in the present setting $\xi_n = 1$ means "if you stop now, for whatever reason, it is safe to reject the null". Below we prove the anytime validity of $\xi$. Additionally, we show that, for certain stopping times $\tau \leq \infty$, the optionally stopped $e$-statistic $T^{M_\tau}$ remains an $e$-statistic. This fact validates the use of the stopped $T^{M_\tau}$ for optional continuation because we can multiply the $e$-statistics $T^{M_\tau}$ across studies while retaining type-I error control. This result is not new and we add it merely for

completeness; it follows by standard arguments as Ramdas et al. (2023) or GHK.

**Proposition F.4.** *Let $T^* = (T^{M_n})_{n\in\mathbb{N}}$, where, for each $n$, $T^{M_n}$ 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 $\xi_n = \mathbf{1}\left\{T^{M_n} \geq 1/\alpha\right\}$. Then, the following two properties hold:*

1. *The sequential test $\xi$ is anytime valid at level $\alpha$, that is,*

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

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

$$\sup_{\theta_0\in\Theta_0} \mathbf{E}^{\mathbf{P}}_{\theta_0}[T^{M_\tau}] \leq 1. \tag{F.1}$$

It is natural to ask whether (F.1) also holds for stopping times that are adapted to the full data $(X^n)_{n\in\mathbb{N}}$ but not to the reduced $(M_n)_{n\in\mathbb{N}}$. In our t-test example, this could be a stopping time $\tau^*$ such as "$\tau^* := 1$ if $|X_1| \notin [a, b]$; $\tau^* = 2$ otherwise" for some $0 < a < b$. The answer is negative: after proving Proposition F.4, we show that, for appropriate choice of $a$ and $b$, this $\tau^*$ is a counterexample. This means that such nonadapted $\tau^*$ cannot be safely used under optional continuation. However, using such a stopping time has no repercussions for optional stopping, since the time $N$ in part 1 of the proposition above is not even required to be a stopping time—$N$ is not restricted by the filtration induced by $M$ and it is even allowed to depend on future observations.

*Proof of Proposition F.4.* From Proposition 8.6, we know that $T^* = (T^{M_n})_{n\in\mathbb{N}}$ is a nonnegative martingale with expected value equal to one. Let $\xi = (\xi_n)_n$ be the sequential test given by $\xi_n = \mathbf{1}\left\{T^{M_n} \geq 1/\alpha\right\}$. The anytime-validity at level $\alpha$ of $\xi$, is a consequence of Ville's inequality, and the fact that the distribution of each $T^{M_n}$ does not depend on $g$. Indeed, these two, together, imply that

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

This implies the first statement. Now, let $\tau \leq \infty$ be a stopping time with respect to the filtration induced by $M$. If the stopping time $\tau$ is almost surely bounded, $T^{M_\tau}$ 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 $T^*_\infty$. Even when $\tau$ might be infinite with positive probability, Theorem 4.8.4 of Durrett (2019) implies that $T^{M_\tau}$ is still an $e$-statistic. □

### F.3.1 Importance of the Filtration for Randomly Stopped E-Statistics

Consider the t-test as in Example 8.1. Fix some $0 < a < b$, and define the stopping time $\tau^* := 1$ if $|X_1| \notin [a, b]$. $\tau^* = 2$ otherwise. Then $\tau^*$ 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 $T^{M_n}(X^n)$ as in (8.6); here we make explicit its dependence on $\delta_1$. 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 $\kappa$. We can now apply Corollary 8.9 (with prior $\tilde{\mathbf{\Pi}}_0^\delta$ putting mass 1 on $\delta = \delta_0 = 0$), which gives that $\tilde{T}_n = \tilde{t}_n(X^n)$ 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 $\tilde{T}_1$ 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:

$$\tilde{T}_2 = \frac{\sqrt{2\kappa^2 + 1} \cdot (x_1^2 + x_2^2)}{\kappa^2 (x_1 - x_2)^2 + (x_1^2 + x_2^2)}.$$

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 $\tau^*$ as above, $\tilde{T}_{\tau^*}$ is not an E-variable for some specific choices of $a, b$ and $\kappa$. 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

$$\mathbf{E}_{X_1,X_2\sim N(0,1)}[\tilde{T}_{\tau^*}] = \mathbf{P}_{X_1\sim N(0,1)}\{|X_1| \notin [a,b]\} + \mathbf{E}_{X_1\sim N(0,1)}[\mathbf{1}_{|X_1|\in[a,b]} f(X_1)] > 1.$$

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 $b \approx 1.70$. 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 8.9 (it is even a GROW one), $\tilde{T}_{\tau^*}$ is not an $e$-statistic (its expectation is 0.19 too large), providing the claimed counterexample.

# F.4 Further Derivations, Computations and Proofs

In this appendix, we prove the technical lemmas whose proof was omitted from the main text. In Section F.4.1, we prove the lemmas used in the proof of Theorem 8.2. In Section F.4.2, we show the computations omitted from Section 8.4.1.

## F.4.1 Proof of Technical Lemmas 8.11, 8.12, and 8.13 for Theorem 8.2

*Proof of Lemma 8.11.* 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 $\rho\{J_l\}/\rho\{J_lK_iL_i\} \to 1$ 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 8.12.* 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

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

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 8.13.* The result follows from a rewriting and an application of Jensen's inequality. Indeed,

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

as it was to be shown. □

### F.4.2 Derivation and Computation for Section 8.4.1

We now provide Proposition F.5, giving the derivation underlying Lemma 8.10 in the main text about the likelihood ratio $T^*_{\mathcal{S},n}$ for $\delta_0 = 0$, followed by details about numerical computation.

**Proposition F.5.** *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 $\Sigma$. 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 these rewritings, 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

$$\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T) = \frac{2^d}{K}\frac{\det(\Lambda^{-1}T)^m}{\det(\Lambda)}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}X-\gamma)(\Lambda^{-1}X-\gamma)'\right)$$

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

$$\frac{\det(\Lambda^{-1}T)^{m+1}}{K''}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}TM-\gamma)(\Lambda^{-1}TM-\gamma)'\right).$$

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,

$$\begin{aligned} p^M_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\int_{T\in\mathcal{L}^+} \det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}TT' - \frac{1}{2}(TM-\gamma)(TM-\gamma)'\right)\mathrm{d}\nu(T) \\ &= \frac{2^d}{K}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\int_{T\in\mathcal{L}^+} \det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}T(I+MM')T' + \gamma(TM)'\right)\mathrm{d}\nu(T). \end{aligned}$$

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

$$p^M_{\gamma,\Lambda}(M) = \frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}}{K}\int_{T\in\mathcal{L}^+} \det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}(TA)(TA)' + \gamma(TM)'\right)\mathrm{d}\nu(T).$$

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

$$\begin{aligned} p^M_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\int_{T\in\mathcal{L}^+} \det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}TT' + \gamma(TA^{-1}M)'\right)\mathrm{d}\nu(T) \\ &= \frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)}\frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\mathbf{P}^T_{m+1}\left[\mathrm{e}^{\langle\gamma,TA^{-1}M\rangle}\right], \end{aligned}$$

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** (Numerical computation)**.** Computing the optimal $e$-statistic is feasible nu-

merically. 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,

$$\mathbf{E}_{m+1}^{\mathbf{P}}[\mathrm{e}^{a_{ij}T_{ij}}] = \exp\left(\frac{1}{2}a_{ij}^2\right).$$

For the diagonal elements the situation is not as simple, but a numerical solution is possible. Indeed, for $a_{ii} \geq 0$, and $k_i = (m+1) - i + 1$

$$\begin{aligned}\mathbf{E}_m^{\mathbf{P}}[\mathrm{e}^{a_{ii}T_{ii}}] &= \frac{1}{2^{\frac{k_i}{2}}\Gamma\left(\frac{k_i}{2}\right)}\int_0^\infty x^{\frac{k_i}{2}-1}\exp\left(-\frac{1}{2}x + a_{ii}\sqrt{x}\right)\mathrm{d}x \\ &= {}_1F_1\left(\frac{k_i}{2}, \frac{1}{2}, \frac{a_{ii}^2}{2}\right) + \frac{\sqrt{2}a_{ii}\Gamma\left(\frac{k_i+1}{2}\right)}{\Gamma\left(\frac{k_i}{2}\right)}{}_1F_1\left(\frac{k_i+1}{2}, \frac{3}{2}, \frac{a_{ii}^2}{2}\right),\end{aligned}$$

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.

# Curriculum Vitae

Tyron Darnell Lardy was born in Haarlem on December 6, 1996. He completed his bilingual high school education at the Mendelcollege in 2014, after which he pursued a bachelor's degree in both mathematics and physics at Leiden University. Ultimately, Tyron decided to focus on mathematics, earning a master's degree in the field (cum laude) while working part-time as a software developer to gain practical experience. Alongside his academic activities, Tyron competed internationally in karate as a member of the Dutch national team, achieving third place at the World U21 Championships in 2015, first place at the European U21 Championships in 2016, and third place at the European Championships in 2018. He retired as an athlete to pursue a PhD at the Mathematical Institute of Leiden University under the supervision of Professor Peter Grünwald and Professor Wouter Koolen, completing his doctoral studies in 2025.