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

Pérez-Ortiz, M.F.

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

Proefschrift

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

door

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

Promotores:

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

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

Promotiecommissie:

Prof.dr. J. Ziegel (Universität Bern)

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

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

Prof.dr. H.J. Hupkes

Prof.dr. E.A. Verbitskiy

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

Universiteit Leiden
The Netherlands

CWI
Research institute for mathematics & computer science in the Netherlands

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

Nicolás y los Fumadores

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

Herbert Robbins

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

Bojack Horseman

Preface

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

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

Invariance. Principles of invariance have turned out to be a very valuable tool in statistics. The t-test, perhaps the most used test on Earth, is the prototypical example of a scale-invariant test, a test that does not depend on the units of measurement of the observations. A large part of the introductory statistical theory for the inference of location parameters can be summed up in the single statement that the likelihood ratio test for the t-statistic is the overall—invariant or not—most powerful fixed-sample test. In Chapter 2, with Rianne de Heide, Tyron Lardy and Peter Grünwald, we tackle the anytime-valid counterpart of this problem, where power maximization is no longer

meaningful, under more general invariances. The main results in this line of work were found during the world-wide lock-downs of 2020, but the final form of the results took much longer to reach their present form.

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

Acknowledgements

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

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

Contents

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

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

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

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

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

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

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

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

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

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

Samenvatting **177**

Summary **179**

Resumen **181**

Curriculum Vitae **183**

List of Publications **185**

Bibliography **187**