

**Carnegie Mellon University**  
Department of Statistics & Data Science



**Advances in anytime-valid sequential inference**

Ian Waudby-Smith

**Thesis Committee:**

Aaditya Ramdas (Chair)

Edward H. Kennedy

Larry Wasserman

Susan Murphy (Harvard University)

Stefan Wager (Stanford University)

*Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.*

*For my parents.*

## Abstract

This thesis contains some advances in the field of “anytime-valid sequential inference”, a paradigm of statistical inference where confidence intervals,  $p$ -values, and hypothesis tests are valid for all sample sizes simultaneously, including data-dependent stopping times. In more practical terms, anytime-valid procedures allow an analyst to collect data sequentially over time and stop sampling for any data-dependent reason without inflating type-I error rates.

Even in the non-sequential (“batch”) setting, there are two broad categories of statistical procedures: nonasymptotic and asymptotic ones. Neither is universally preferable to the other, with nonasymptotic methods enjoying stronger guarantees in finite samples, and with asymptotic ones being more widely applicable and simpler to implement. This thesis studies anytime-valid inference in both regimes and is correspondingly divided into two parts.

The first part focuses on nonasymptotic inference and concentration inequalities. Here, we introduce new methods for both anytime-valid and batch inference for means of bounded random variables when sampling with and without replacement. These computationally and statistically efficient algorithms find several applications in risk-limiting election audits, off-policy evaluation in contextual bandits, and concentration inequalities under differential privacy constraints. Each application has a dedicated chapter.

The second part studies asymptotic anytime-valid inference, a far less mature corner of the literature. As such, there is an increased focus on articulating the right definitions of anytime-valid procedures and their guarantees, as well as laying some of the requisite probabilistic foundations. In particular, we develop distribution-uniform strong laws of large numbers and strong Gaussian coupling inequalities which are then used to provide a framework for asymptotic anytime-valid inference. As one illustrative application of this framework, we develop the first sequential test of conditional independence that does not rely on the Model-X assumption.

## Acknowledgments

Starting with my thesis advisor, there is insufficient space here to enumerate all the ways that Aaditya Ramdas has shaped me as a researcher so I will only mention a select few. He taught me how to thoughtfully craft papers, confidently give talks, and attack hard mathematical problems with the perfect balance of playfulness and tenacity. It is hard to put into words the amount of gratitude I feel for all the mentorship that he provided over the course of my PhD. Most importantly, working alongside him these past five years has been an absolute blast.

In addition to Aaditya, I want to thank my other committee members — Edward H. Kennedy, Larry Wasserman, Susan A. Murphy, and Stefan Wager — for providing invaluable feedback on this thesis. An additional thank you to Edward for many years of mentorship and collaboration.

Let me now briefly highlight my “behind-the-scenes committee” — those people that have greatly influenced the contents of this thesis but in the context of collaboration and/or informal one-on-one conversations. This list includes Philip B. Stark, Sivaraman Balakrishnan, Tudor Manole, Arun Kuchibhotla, Ruodu Wang, Martin Larsson, Johannes Ruf, Peter Grünwald, David Childers, and Zhiwei Steven Wu. I would also like to append to this list my coauthors with whom I worked during summer internships: David Arbour, Ritwik Sinha, Paul Mineiro, Lili Wu, and Nikos Karampatziakis. Chapters 6 and 7 both began as internship collaborations.

I am fortunate to have made so many brilliant and kind friends at CMU, whether in Statistics, CSD, S3D, MLD, Heinz, Robotics, Tepper, or Philosophy. To prevent this document from becoming one-half “thesis” and one-half “acknowledgments”, let me just say this: You know who you are; thank you for the memories. A special thanks goes to my close friend and office-mate Tudor Manole for — in addition to the countless great times shared — being the one that pointed me to the literature on strong Gaussian approximation a few years ago, a topic that turned out to play a central role in my work as exemplified by Part II of this thesis.

Let me now rewind the clock nearly a decade so I can express my gratitude to Zihui Amy Liu, my first mentor in the field of statistics. Amy is the person who initially encouraged me to pursue a PhD, a path that seemed foreign and out of reach at the time, but advice that I nevertheless followed without a single regret. She also introduced me to Eleanor Pullenayegum, another early mentor I would like to thank for bringing me on as an intern at SickKids the summer prior to my start at CMU. To round out the list of pre-PhD mentors, I am grateful to Joel A. Dubin, Joon Lee, Pengfei Li, and Yu-Ru Liu for introducing me to the ways of research when I was an exploratory and wide-eyed undergraduate. In particular, thank you to Joel and Joon for guiding me through the paper-writing and publication processes for the first time.

I’d also like to mention the close friends I made when I lived in Seattle, especially Arjun Sondhi for visiting me here in Pittsburgh multiple times and my former roommate Charlie Wolock for moving to Pennsylvania just to hang out with me more (perhaps with some other professional and life circumstances aligning well for him too, I suppose).

A special thank you goes to my best buds Karl Robinson, Brian McCrindle, and Liam Brown Guerra-Acero for being such important figures in my life for decades and being incredibly supportive of every path I’ve taken. Regularly spending time with you three on a handful of different continents over the years has made this journey infinitely more fulfilling.

Finally, thank you to my family for the limitless source of encouragement and support throughout my life. Most of all, thank you Mom and Dad for absolutely everything and more. This thesis is dedicated to you two.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	A selective primer on anytime-valid sequential inference . . . . .	2
1.1.1	Nonasymptotic anytime-valid inference . . . . .	2
1.1.2	Asymptotic anytime-valid inference . . . . .	7
1.2	A detailed outline of this thesis . . . . .	9
<b>I</b>	<b>Nonasymptotic inference and time-uniform concentration</b>	<b>14</b>
<b>2</b>	<b>Confidence sequences for sampling without replacement</b>	<b>15</b>
2.1	Introduction . . . . .	15
2.1.1	Notation, supermartingales and the model for sampling WoR . . . . .	15
2.2	Discrete categorical setting . . . . .	17
2.2.1	The prior-posterior-ratio (PPR) martingale . . . . .	17
2.2.2	CSs for binary settings using the hypergeometric distribution . . . . .	18
2.2.3	Role of the ‘prior’ in the prior-posterior CS . . . . .	19
2.3	Bounded real-valued setting . . . . .	20
2.3.1	Hoeffding-type bounds . . . . .	20
2.3.2	Empirical Bernstein-type bounds . . . . .	22
2.4	Testing hypotheses about finite sets of nonrandom numbers . . . . .	25
2.5	Summary . . . . .	26
2.A	Four prototypical examples . . . . .	28
2.B	Proofs of the main results . . . . .	30
2.B.1	Proof of Proposition 2.2.1 . . . . .	30
2.B.2	Proof of Theorem 2.3.1 . . . . .	32
2.B.3	Proof of Theorem 2.3.2 . . . . .	34
2.C	Sampling multivariate binary variables WoR . . . . .	35
2.D	Coupling the ‘prior’ with the stopping rule to improve power . . . . .	37
2.E	Choosing a $\lambda$ -sequence for Hoeffding and empirical Bernstein CSs . . . . .	37
2.F	Comparing our CSs to those implied by Bardenet & Maillard . . . . .	38
2.G	Time-uniform versus fixed-time bounds . . . . .	41
2.H	Computational considerations . . . . .	42

2.I	Simple experiments for computing miscoverage rates . . . . .	42
<b>3</b>	<b>Estimating means of bounded random variables by betting</b>	<b>44</b>
3.1	Introduction . . . . .	44
3.2	Concentration inequalities via nonnegative supermartingales . . . . .	47
3.2.1	Confidence sequences and the method(s) of mixtures . . . . .	48
3.2.2	Nonparametric confidence sequences via sequential testing . . . . .	49
3.2.3	Connections to the Chernoff method . . . . .	50
3.3	Warmup: exponential supermartingales and predictable plug-ins . . . . .	51
3.3.1	Predictable plug-in Cramer-Chernoff supermartingales . . . . .	52
3.3.2	Application: closed-form empirical Bernstein confidence sets . . . . .	53
3.3.3	Guiding principles for deriving predictable plug-ins . . . . .	56
3.4	The capital process, betting, and martingales . . . . .	58
3.4.1	Connections to betting . . . . .	59
3.4.2	Connections to likelihood ratios . . . . .	60
3.4.3	Adaptive, constrained adversaries . . . . .	61
3.4.4	The hedged capital process . . . . .	61
3.5	Betting while sampling without replacement (WoR) . . . . .	65
3.5.1	Existing (super)martingale-based confidence sequences or tests . . . . .	66
3.5.2	The capital process for sampling without replacement . . . . .	66
3.5.3	Powerful betting schemes . . . . .	67
3.5.4	Relationship to composite null testing . . . . .	69
3.6	A brief selective history on betting and its mathematical applications . . . . .	70
3.7	Summary . . . . .	71
3.A	Proofs of main results . . . . .	74
3.A.1	Proof of Proposition 3.3.1 . . . . .	74
3.A.2	Proof of Theorem 3.3.1 . . . . .	75
3.A.3	Proof of Proposition 3.4.1 . . . . .	76
3.A.4	Proof of Proposition 3.4.2 . . . . .	77
3.A.5	Proof of Theorem 3.4.1 . . . . .	77
3.A.6	Proof of Lemma 3.B.1 . . . . .	79
3.A.7	Proof of Proposition 3.B.1 . . . . .	80
3.A.8	Proof of Proposition 3.5.1 . . . . .	82
3.A.9	Proof of Theorem 3.5.1 . . . . .	82
3.B	How to bet: deriving adaptive betting strategies . . . . .	83
3.B.1	Predictable plug-ins yield good betting strategies . . . . .	83
3.B.2	Growth rate adaptive to the particular alternative (GRAPA) . . . . .	84
3.B.3	Approximate GRAPA (aGRAPA) . . . . .	85
3.B.4	Lower-bound on the wealth (LBOW) . . . . .	86
3.B.5	Online Newton Step (ONS- $m$ ) . . . . .	87
3.B.6	Diversified Kelly betting (dKelly) . . . . .	88
3.B.7	Confidence Boundary (ConBo) . . . . .	90
3.B.8	Sequentially Rebalanced Portfolio (SRP) . . . . .	91

3.C	Simulations . . . . .	92
3.C.1	Time-uniform confidence sequences (with replacement) . . . . .	93
3.C.2	Fixed-time confidence intervals (with replacement) . . . . .	94
3.C.3	Time-uniform confidence sequences (without replacement) . . . . .	95
3.C.4	Fixed-time confidence intervals (without replacement) . . . . .	96
3.D	Simulation details . . . . .	97
3.D.1	Time-uniform confidence sequences (with replacement) . . . . .	97
3.D.2	Fixed-time confidence intervals (with replacement) . . . . .	98
3.D.3	Time-uniform confidence sequences (without replacement) . . . . .	100
3.D.4	Fixed-time confidence intervals (without replacement) . . . . .	102
3.D.5	Betting “confidence distributions”: confidence sets at several resolutions	103
3.E	Additional theoretical results . . . . .	104
3.E.1	Betting confidence sets are tighter than Hoeffding . . . . .	104
3.E.2	Optimal convergence of betting confidence sets . . . . .	106
3.E.3	On the width of empirical Bernstein confidence intervals . . . . .	112
3.E.4	aGRAPA sublevel sets need not be intervals: a worst-case example . . . . .	116
3.E.5	Betting confidence sequences for non-iid data . . . . .	117
3.E.6	Owen’s empirical likelihood ratio and Mykland’s dual likelihood ratio	118
3.F	An extended history of betting and its applications . . . . .	120
<b>4</b>	<b>RiLACS: Risk-limiting audits via confidence sequences</b>	<b>124</b>
4.1	Introduction . . . . .	124
4.1.1	SHANGRLA Reduces Election Auditing to Sequential Testing . . . . .	125
4.1.2	Confidence Sequences . . . . .	126
4.1.3	Contributions and Outline . . . . .	127
4.2	Confidence sequences are risk-limiting . . . . .	127
4.2.1	Relationship to Sequential Hypothesis Testing . . . . .	129
4.2.2	Auditing Multiple Contests . . . . .	129
4.3	Designing powerful confidence sequences for RLAs . . . . .	131
4.3.1	Designing Martingales and Tests from Reported Vote Totals . . . . .	132
4.3.2	Designing Martingales and Tests without Vote Totals . . . . .	134
4.4	Illustration: auditing Canada’s 43rd federal election . . . . .	136
4.5	Risk-limiting tallies via confidence sequences . . . . .	139
4.6	Summary . . . . .	141
4.A	Maximizing a proxy for $M_N(1/2)$ . . . . .	141
<b>5</b>	<b>Nonparametric extensions of randomized response for private confidence sets</b>	<b>143</b>
5.1	Introduction . . . . .	143
5.1.1	Background: Local Differential Privacy . . . . .	144
5.1.2	Contributions and Outline . . . . .	144
5.1.3	Related Work . . . . .	145
5.2	Extending Warner’s randomized response . . . . .	146
5.3	Private confidence intervals for bounded data . . . . .	149
5.3.1	What is a Locally Private Confidence Set? . . . . .	150

5.3.2	A Locally Private Hoeffding CI via NPPR . . . . .	151
5.3.3	Time-uniform Confidence Sequences for $\mu^*$ . . . . .	153
5.3.4	Confidence Sequences for Time-varying Means . . . . .	154
5.4	Illustration: Private online A/B testing . . . . .	155
5.5	Additional results & summary . . . . .	157
5.A	Proofs of main results . . . . .	159
5.A.1	Prelude: filtrations, supermartingales, and Ville's inequality . . . . .	159
5.A.2	Proof of Theorem 5.2.1 . . . . .	160
5.A.3	Proof of Theorem 5.3.2 . . . . .	160
5.A.4	Proof of Theorem 5.3.3 . . . . .	161
5.A.5	Proof of Theorem 5.3.4 . . . . .	162
5.B	Additional results . . . . .	166
5.B.1	Confidence sets under randomized response . . . . .	166
5.B.2	Confidence sets for sample means . . . . .	167
5.B.3	Why one should set $G = 1$ for Hoeffding-type methods . . . . .	167
5.B.4	Confidence sets under the sequentially interactive Laplace mechanism . . . . .	168
5.B.5	Variance-adaptive confidence intervals and sequences . . . . .	169
5.B.6	Choosing $(r, G)$ for variance-adaptive confidence sets . . . . .	173
5.B.7	One-sided time-varying . . . . .	174
5.B.8	Private hypothesis testing and $p$ -values . . . . .	175
5.B.9	A/B testing the weak null . . . . .	177
5.B.10	Locally private adaptive online A/B testing . . . . .	178
5.C	Proofs of additional results . . . . .	179
5.C.1	Proof of Proposition 5.B.1 . . . . .	179
5.C.2	A lemma for Theorems 5.B.1 and 5.B.2 . . . . .	181
5.C.3	Proof of Theorem 5.B.1 . . . . .	183
5.C.4	Proof of Theorem 5.B.2 . . . . .	184
5.C.5	Proof of Proposition 5.B.3 . . . . .	186
5.C.6	Proof of Proposition 5.B.5 . . . . .	188
5.C.7	Proof of Proposition 5.B.4 . . . . .	190
5.C.8	Proof of Theorem 5.B.3 . . . . .	195
5.D	A more detailed survey of related work . . . . .	197
6	<b>Anytime-valid off-policy inference for contextual bandits</b>	199
6.1	Introduction . . . . .	199
6.1.1	Off-policy inference, confidence intervals, and confidence sequences . . . . .	200
6.1.2	Desiderata for anytime-valid off-policy inference . . . . .	202
6.1.3	Outline and contributions . . . . .	204
6.1.4	Related work . . . . .	204
6.1.5	Notation: supermartingales, filtrations, and stopping times . . . . .	205
6.2	Warmup: Off-policy inference for constant policy values . . . . .	206
6.2.1	Tighter confidence sequences via doubly robust pseudo-outcomes . . . . .	207
6.2.2	Tuning, truncating, and mirroring . . . . .	209

6.2.3	Closed-form confidence sequences . . . . .	211
6.2.4	Fixed-time confidence intervals . . . . .	213
6.3	Inference for time-varying policy values . . . . .	215
6.3.1	A remark on policy value differences . . . . .	218
6.3.2	Time-varying treatment effects in adaptive experiments . . . . .	220
6.3.3	Sequential testing and anytime $p$ -values for off-policy inference . . . . .	221
6.4	Time-uniform inference for the off-policy CDF . . . . .	224
6.5	Summary & extensions . . . . .	227
6.A	Proofs of the main results . . . . .	229
6.A.1	A technical lemma . . . . .	229
6.A.2	Proof of Theorem 6.2.1 . . . . .	230
6.A.3	Proof of Theorem 6.3.1 . . . . .	232
6.A.4	Proof of Proposition 6.2.2 . . . . .	233
6.A.5	Proof of Proposition 6.3.1 . . . . .	234
6.A.6	Proof of Theorem 6.4.1 . . . . .	236
6.B	A causal view of contextual bandits via potential outcomes . . . . .	242

<b>II</b>	<b>Asymptotic inference, strong laws, and Gaussian approximation</b>	<b>246</b>
<b>7</b>	<b>Time-uniform central limit theory and asymptotic confidence sequences</b>	<b>247</b>
7.1	Introduction . . . . .	247
7.1.1	Contributions and outline . . . . .	249
7.2	Asymptotic confidence sequences . . . . .	249
7.2.1	Defining asymptotic confidence sequences . . . . .	249
7.2.2	Warmup: AsympCSs for the mean of i.i.d. random variables . . . . .	252
7.2.3	A general recipe for deriving asymptotic confidence sequences . . . . .	255
7.2.4	Lindeberg- and Lyapunov-type AsympCSs for time-varying means . . . . .	256
7.2.5	Asymptotic coverage and type-I error control . . . . .	259
7.2.6	Asymptotic confidence sequences using Robbins' delayed start . . . . .	261
7.3	Illustration: Causal effects and semiparametric estimation . . . . .	263
7.3.1	Sequential sample splitting and cross fitting . . . . .	264
7.3.2	Asymptotic confidence sequences in randomized experiments . . . . .	266
7.3.3	Asymptotic confidence sequences in observational studies . . . . .	268
7.3.4	The running average of individual treatment effects . . . . .	270
7.3.5	Extensions to general semiparametric estimation and the delta method	272
7.4	Simulation studies: Widths and empirical coverage . . . . .	273
7.5	Real data application: effects of IV fluid caps in sepsis patients . . . . .	277
7.6	Conclusion . . . . .	278
7.A	Proofs of the main results . . . . .	279
7.A.1	Proof of Theorem 7.2.2 . . . . .	279
7.A.2	Proof of Theorem 7.2.3 . . . . .	282
7.A.3	Proof of Proposition 7.2.2 . . . . .	282
7.A.4	Proof of Corollary 7.2.1 . . . . .	285

7.A.5	Proof of Theorems 7.3.1 and 7.3.2 . . . . .	286
7.A.6	Proof of Theorem 7.3.3 . . . . .	292
7.A.7	Proof of Theorem 7.2.5 . . . . .	297
7.A.8	Proof of Proposition 7.3.1 . . . . .	301
7.A.9	Proof of Proposition 7.2.3 . . . . .	301
7.B	Additional discussions . . . . .	305
7.B.1	One-sided asymptotic confidence sequences . . . . .	305
7.B.2	Optimizing Robbins' normal mixture for $(t, \alpha)$ . . . . .	312
7.B.3	Time-uniform convergence in probability is equivalent to almost sure convergence . . . . .	313
7.B.4	Comparing AsympCSs to group-sequential repeated confidence intervals . . . . .	314
7.B.5	The Lyapunov-type condition implies the Lindeberg-type condition . . . . .	316
7.B.6	On martingale AsympCSs for running average treatment effects . . . . .	316
7.B.7	Explicit connections between "delayed start" boundaries and other works . . . . .	317
7.B.8	A brief review of efficient estimators . . . . .	319
7.B.9	On the sharpness of AsympCSs using efficient influence functions . . . . .	320
7.B.10	Multivariate asymptotic confidence sequences . . . . .	321
<b>8</b>	<b>Distribution-uniform anytime-valid sequential inference</b> . . . . .	<b>325</b>
8.1	Introduction . . . . .	325
8.1.1	Outline of the chapter . . . . .	327
8.1.2	Notation . . . . .	328
8.2	What is distribution-uniform anytime-valid inference? . . . . .	328
8.2.1	Our primary goal: Inference for the mean . . . . .	329
8.2.2	Time- and $\mathcal{P}$ -uniform central limit theory for partial sums . . . . .	331
8.3	Almost-sure consistency and time-uniform asymptotics . . . . .	332
8.3.1	What is $\mathcal{P}$ -uniform almost-sure consistency? . . . . .	332
8.3.2	$\mathcal{P}$ -uniform almost-sure variance estimation . . . . .	335
8.3.3	The main result: $(\mathcal{P}, n, \alpha)$ -uniform statistical inference . . . . .	335
8.4	Illustration: Sequential conditional independence testing . . . . .	336
8.4.1	Prelude: weak regression consistency . . . . .	337
8.4.2	A brief refresher on batch conditional independence testing . . . . .	337
8.4.3	On the hardness of anytime-valid conditional independence testing . . . . .	339
8.4.4	SeqGCM: The sequential generalized covariance measure test . . . . .	340
8.5	Distribution-uniform strong Gaussian approximation . . . . .	343
8.6	Summary & discussion . . . . .	346
8.A	Proofs of the main results . . . . .	347
8.A.1	Proof of Proposition 8.2.2 . . . . .	347
8.A.2	Proof of Lemma 8.3.1 . . . . .	348
8.A.3	Proof of Theorem 8.3.3 . . . . .	352
8.A.4	Proof of Proposition 8.4.1 . . . . .	355
8.A.5	Proof of Theorem 8.4.1 . . . . .	356
8.A.6	Proof of Corollary 8.5.1 . . . . .	365

8.A.7	Proof of Lemma 8.5.1 and Theorem 8.5.2 . . . . .	366
8.B	Additional theoretical discussions and results . . . . .	371
8.B.1	The Robbins-Siegmund distribution . . . . .	371
8.B.2	Uniform convergence of perturbed random variables . . . . .	373
<b>9</b>	<b>Distribution-uniform strong laws of large numbers</b>	<b>379</b>
9.1	Introduction . . . . .	379
9.1.1	Notation and conventions . . . . .	383
9.1.2	Outline and summary of contributions . . . . .	383
9.2	Distribution-uniform strong laws of large numbers . . . . .	384
9.3	Other distribution-uniform convergence results . . . . .	389
9.3.1	A distribution-uniform Khintchine-Kolmogorov convergence theorem . . . . .	389
9.3.2	A distribution-uniform Kolmogorov three-series theorem . . . . .	390
9.3.3	A three-series theorem for stochastic nonincreasingness . . . . .	392
9.3.4	A distribution-uniform stochastic generalization of Kronecker's lemma . . . . .	393
9.3.5	Distribution-uniform Borel-Cantelli lemmas . . . . .	396
9.4	Proof details for Theorems 9.2.1 and 9.2.2 . . . . .	397
9.4.1	Proof details for Theorem 9.2.1 . . . . .	397
9.4.2	Proof details for Theorem 9.2.2 . . . . .	412
9.5	Application to uniformly consistent variance estimation . . . . .	415
9.6	Summary . . . . .	417
9.A	A note on de la Vallée-Poussin's criterion . . . . .	418

# Chapter 1

## Introduction

This thesis contains a variety of investigations into different statistical problems spanning concentration inequalities, gambling, election audits, privacy, bandit algorithms, laws of large numbers, and strong Gaussian couplings. However, one central theme that each of the following chapters revolves around — whether directly or indirectly — is a simple question:

*“What is the right way to perform statistical inference when data are collected sequentially over time rather than in a single batch with a fixed sample size?”*

and this question can be thought of as an informal summary of what “anytime-valid sequential inference” — found in the title of this thesis — aims to give an answer to. To illustrate how easily this question can naturally arise, consider the toy problem of trying to discern whether a coin is unbiased. That is, whether the outcome of a coin flip is “heads” or “tails” with equal probability  $1/2$ . As statisticians, we can formalize this problem by positing the null hypothesis

$$H_0 : P(\text{heads}) = 1/2 \quad \text{versus} \quad H_1 : P(\text{heads}) \neq 1/2. \tag{1.1}$$

To proceed, we flip the coin  $n = 100$  times (say) and we compute a 90% confidence interval (CI)  $C_n$  for  $P(\text{heads})$ . Suppose that we find our CI to still include  $1/2$ , but only *barely* — e.g.  $C_n = [0.49, 0.7]$ . Intuitively, this might suggest that  $P(\text{heads}) > 1/2$ , but from a formal hypothesis testing perspective, no such conclusion can be made. Logistically speaking, it would be easy to flip a few more coins to see whether  $1/2 \notin C_N$  for some slightly enlarged sample size  $N$ , but the mere decision to do so *completely* invalidates the resulting statistical analysis since  $N$  is now a *data-dependent stopping time* (to be made more formal shortly).<sup>1</sup> When performed carelessly or out of dishonesty, statisticians would classify this type of continued sampling as a form of “*p-hacking*”. To summarize the previous discussion more mathematically, let

---

<sup>1</sup>Worse still, it may be tempting to believe that we can just start from scratch and sample a new batch of coins (not including the original  $n$ ) of size  $n^* \gg n$ , but this does not remedy the issue since those  $n^*$  coins never would have been sampled if the original sample  $n$  was large enough from the outset. Confidence sets, *p*-values, and so on are only valid if all potential sources of randomness are taken into account, including random decisions of whether or not to draw a new sample.

$\theta_H := P(\text{heads})$ ; it is typically the case that

$$P(\theta_H \in C_n) \geq 90\% \quad \text{but} \quad P_{H_0}(\theta_H \in C_N) \ll 90\%. \quad (1.2)$$

In other words,  $C_N$  is not a 90% CI at all and has no meaningful statistical interpretation. Indeed, if  $N$  is enlarged indefinitely, it is the case with the vast majority of familiar statistical methods that  $P(\theta_H \in C_N) = 0$ , a simple consequence of the law of the iterated logarithm which will be discussed later. While the discussion thus far has centered around an inconsequential toy example involving coin flips, these same phenomena can be seen in several high-stakes statistical problems. For example, sequential sampling of data arises in (a) randomized controlled trials studying the effectiveness of a treatment over placebo, (b) internet companies understanding the effects of modifying a feature in an app or on a website, (c) continuously monitoring the effects of interventions on health outcomes [289, 182], and (d) rigorously auditing the outcome of an election (as we will study in some depth in Chapter 4), to name a few. As statisticians, we now find ourselves at a crossroads with two options:

- (i) Always ensure that classical statistical methods are used in the way that they were intended and remind practitioners that  $n$  must be a data-*independent* sample size, or
- (ii) Develop new statistical tools that allow – and even encourage – adaptive and sequential sampling, enabling valid inference at highly data-dependent sample sizes.

Despite both (i) and (ii) being valid approaches to combat the aforementioned “*p*-hacking” and deserving attention from the statistical community, this thesis focuses primarily on (ii). The statistical tools that we will develop will take the form of so-called *confidence sequences*, *anytime p-values*, and *sequential hypothesis tests*, all of which will be reviewed in the following section.

## 1.1 A selective primer on anytime-valid sequential inference

Turning to a more formal discussion, consider the problem of estimating the population mean  $\mu = \mathbb{E}(X_1)$  from a sequence of i.i.d. data  $(X_t)_{t=1}^\infty \equiv (X_1, X_2, \dots)$  that are observed sequentially over time from a distribution  $P$ . There are several ways to attack this problem depending on the types of assumptions that one wants to make about  $P$ . These different approaches fall under two broad categories: nonasymptotic and asymptotic methods. The former will be the subject of Part I of this thesis while the latter will be the subject of Part II, introduced in Sections 1.1.1 and 1.1.2, respectively.

### 1.1.1 Nonasymptotic anytime-valid inference

In the fixed- $n$  setting, a nonasymptotic  $(1 - \alpha)$ -CI for  $\mu$  is an interval  $\dot{C}_n \equiv \dot{C}(X_1, \dots, X_n)$  with the property that

$$\forall n \in \mathbb{N}, \mathbb{P}(\mu \in \dot{C}_n) \geq 1 - \alpha, \quad \text{or equivalently,} \quad \forall n \in \mathbb{N}, \mathbb{P}(\mu \notin \dot{C}_n) \leq \alpha. \quad (1.3)$$

---

<sup>2</sup>Throughout this thesis, we use  $\mathbb{N}$  for  $\{1, 2, \dots\}$  and  $\mathbb{N}_0$  for  $\mathbb{N} \cup \{0\}$ .

The coverage guarantee (1.3) of a CI is only valid at some *prespecified* sample size  $n$ , which must be decided in advance of seeing any data — peeking at the data in order to determine the sample size would constitute “*p-hacking*” as previously discussed. However, it may be restrictive to fix  $n$  beforehand, and even if clever sample size calculations are carried out based on prior knowledge, it is impossible to know *a priori* whether  $n$  will be large enough to detect some signal of interest: after collecting the data, one may regret collecting too little data or collecting much more than necessary.

On the other hand, a *confidence sequence* (CS) provides the flexibility to choose sample sizes data-adaptively without ever compromising the type-I error rate (see Figure 1.1). CSs were introduced and studied in depth in a series of papers by Herbert Robbins and several colleagues and students [76, 218, 217, 168], and has received renewed interest in the past decade [17, 132, 18, 135, 125, 123]. Formally, a nonasymptotic CS is a sequence of CIs  $(\bar{C}_t)_{t=1}^\infty$  such

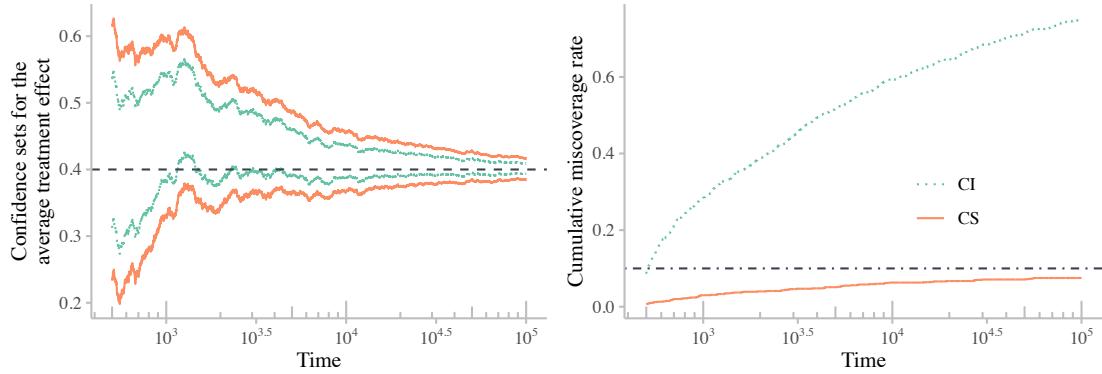


Figure 1.1: This figure is taken from an application in Chapter 7 where the left plot shows one run of a single experiment: an asymptotic CS alongside an asymptotic CI for a parameter of interest (in this case, the average treatment effect (ATE) of 0.4, an example we expand on in Section 7.3). The true value of the ATE is covered by the CS simultaneously from time 30 to 10000. On the other hand, the CI fails to cover the true ATE at several points in time. By repeating such an experiment hundreds of times, one obtains the right plot which displays the cumulative probability of miscoverage — i.e. the probability of the CS or CI failing to capture the true ATE at any time up to  $t$ . Notice that the CI error rate begins at  $\alpha = 0.1$  and quickly grows, while the CS error rate never exceeds  $\alpha = 0.1$ .

that

$$\mathbb{P}(\forall t \in \mathbb{N}, \mu \in \bar{C}_t) \geq 1 - \alpha, \quad \text{or equivalently,} \quad \mathbb{P}(\exists t \in \mathbb{N} : \mu \notin \bar{C}_t) \leq \alpha. \quad (1.4)$$

The statements (1.3) and (1.4) look similar but are markedly different from the data analyst’s or experimenter’s perspective. In particular, employing a CS has the following implications:

- (a) The CS can be (optionally) updated whenever new data become available;
- (b) Experiments can be continuously monitored, adaptively stopped, or continued;
- (c) The type-I error is controlled at all stopping times, including data-dependent times.

In fact, CSs may be equivalently defined as CIs that are valid at arbitrary stopping times, i.e.

$$\mathbb{P}(\mu \in \bar{C}_\tau) \geq 1 - \alpha \quad \text{for any stopping time } \tau,^3 \quad (1.5)$$

and a proof of this equivalence can be found in Howard et al. [125, Lemma 3]. Much like CSs are anytime-valid analogues of CIs, so-called *anytime p-values* and *sequential hypothesis tests* serve as such analogues of classical fixed- $n$  p-values and hypothesis tests. For the sake of completeness, let us provide (somewhat informal) definitions of all three of these nonasymptotic anytime-valid inference tools as they will repeatedly appear throughout Part I of this document.

**Definition 1.1.1** (Nonasymptotic anytime-valid inference). Let  $(\Omega, \mathcal{F}, \mathcal{P}) \equiv (\Omega, \mathcal{F}, P)_{P \in \mathcal{P}}$  be a collection of filtered probability spaces with filtration  $\mathcal{F} \equiv (\mathcal{F}_t)_{t=0}^\infty$ . For a given  $\alpha \in [0, 1]$ , we say that a sequence of intervals  $(C_t)_{t=1}^\infty$  is a  $(1 - \alpha)$ -confidence sequence for a parameter  $\mu \in \mathbb{R}$  if

$$P(\exists t \geq 1 : \mu \notin C_t) \leq \alpha. \quad (1.6)$$

Furthermore, we say that a  $[0, 1]$ -valued process  $(p_t)_{t=1}^\infty$  is an anytime p-value – sometimes called a p-process – if for all  $\alpha \in [0, 1]$ ,

$$P(\exists t \geq 1 : p_t \leq \alpha) \leq \alpha. \quad (1.7)$$

Finally, we say that a binary  $\{0, 1\}$ -valued process  $(\phi_t)_{t=1}^\infty$  is a level- $\alpha$  sequential hypothesis test if

$$P(\exists t \geq 1 : \phi_t = 1) \leq \alpha. \quad (1.8)$$

Clearly, omitting the uniformity over  $t \geq 1$  in Definition 1.1.1 – i.e. dropping the existential quantifiers in (1.6)–(1.8) – recovers the usual definitions of nonasymptotic confidence intervals, p-values, and hypothesis tests, respectively. It should be noted that these tools can be extended to more general filtrations not necessarily indexed by  $\mathbb{N}_0$  and random variables taking values in spaces other than  $\mathbb{R}$ , and so on. The purpose of Definition 1.1.1 is to prime the reader’s intuition for the more in-depth discussions that will be found in the following chapters.

Making matters more concrete, let us now give some examples of CSs for a simplification of the problem posed at the outset of Section 1.1. Suppose that  $(X_t)_{t=1}^\infty$  are i.i.d. random variables with support  $[0, 1]$  and we would like to estimate their mean  $\mu := \mathbb{E}X_1 \in [0, 1]$ . One of the most well-known fixed- $n$  CIs for  $\mu$  in this setting is that based on Hoeffding’s sub-Gaussian concentration inequality:

$$\dot{C}_n := \left[ \frac{1}{n} \sum_{i=1}^n X_i \pm \sqrt{\frac{\log(2/\alpha)}{2n}} \right], \quad (1.9)$$

---

<sup>3</sup>Here, the stopping time is with respect to some implicit filtration which will be made more explicit in the chapters that follow. For the sake of intuition, one can think about this filtration  $(\mathcal{F}_t)_{t=0}^\infty$  as the one generated by a sequence of data  $(X_t)_{t=1}^\infty$  so that  $\mathcal{F}_t := \sigma(X_1, \dots, X_t)$  for  $t = 1, 2, \dots$  and  $\mathcal{F}_0$  is the trivial sigma-algebra.

and indeed,  $P(\mu \in \dot{C}_n) \geq 1 - \alpha$  for any fixed  $n \in \mathbb{N}$ . An analogous anytime-valid CS  $(\bar{C}_t)_{t=1}^\infty$  due to Robbins [217] (see also Howard et al. [125] and Chapter 7) can be constructed from a *time-uniform* sub-Gaussian concentration boundary via mixture supermartingales:

$$\bar{C}_t := \left[ \frac{1}{t} \sum_{i=1}^t X_i \pm \frac{1}{2} \cdot \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log \left( \frac{t\rho^2 + 1}{\alpha^2} \right)} \right], \quad (1.10)$$

where  $\rho > 0$  is some pre-specified constant whose effect on the shape of the above CS is discussed in Section 7.B.2 of Chapter 7. Inspecting  $\dot{C}_n$  and  $\bar{C}_t$  side-by-side, notice that  $\dot{C}_n \asymp 1/\sqrt{n}$  (like many CIs for the mean in various settings) while  $\bar{C}_t \asymp \sqrt{\log t/t}$ . This juxtaposition raises a natural question: “*is it possible to construct a CS for  $\mu$  with a width of  $1/\sqrt{t}$ , and if not, is there a ‘fastest possible’ rate?*”. The short and informal answer to this is “*no*”, and “*yes, it is  $\sqrt{\log \log t/t}$* ”. The reasoning behind both of these answers follows straightforwardly from the law of the iterated logarithm [158, 159] which states that for i.i.d. random variables  $(X_t)_{t=1}^\infty$  with mean  $\mu$  and variance  $\sigma^2$ ,

$$\limsup_{t \rightarrow \infty} \frac{\sum_{i=1}^t (X_i - \mu)}{\sqrt{2\sigma^2 \log \log t}} = 1 \quad \text{almost surely.} \quad (1.11)$$

In other words, the centered sample mean  $\frac{1}{t} \sum_{i=1}^t (X_i - \mu)$  will oscillate between and hit the boundaries  $+\sqrt{2\sigma^2 \log \log t/t}$  and  $-\sqrt{2\sigma^2 \log \log t/t}$  infinitely often with probability one, and hence no sequence of intervals centered at the sample mean  $\frac{1}{t} \sum_{i=1}^t X_i$  and shrinking at a rate of  $1/\sqrt{t}$  can contain  $\mu$  uniformly over time  $t \geq 1$  with any positive probability.

One might then wonder if a CS with a rate of  $\sqrt{\log \log t/t}$  is possible to construct. Indeed, using the so-called “stitching” technique, Howard et al. [125] provide a clean and closed-form solution to this problem:

$$\frac{1}{t} \sum_{i=1}^t X_i \pm \frac{1.7}{2} \cdot \sqrt{\frac{\log \log(2t) + 0.72 \log(10.4/\alpha)}{t}}, \quad (1.12)$$

but there are various practical and theoretical reasons why one would prefer the bound in (1.10), but we defer those discussions until later, for instance in Chapters 3 and 7. Notice, however, that neither (1.10) nor (1.12) has the ability to adapt to the true variance  $\text{Var}_P(X)$  of the distribution  $P$ . Indeed, depending on the goals of the statistician, one may wish to eschew closed-form bounds like (1.10) and (1.12) altogether and use CSs (or even CIs) that can be much tighter via variance-adaptivity at the expense of analytic expressions. Such approaches are given a detailed treatment in Chapter 3 as well as the followup work of Chapter 6. These techniques also find applications in risk-limiting election audits and differentially private concentration inequalities, studied in Chapters 4 and 5, respectively.

Let us now briefly give an overview of how CSs, anytime  $p$ -values, and sequential tests are typically derived, but a more detailed discussion can be found in Section 3.2 within Chapter 3. In short, there are two key ingredients: test supermartingales (or  $e$ -processes more generally)

and Ville's inequality [264] which we elaborate on now. Provided a filtration  $\mathcal{F} \equiv (\mathcal{F}_t)_{t=0}^\infty$  indexed by  $\mathbb{N}_0$  for simplicity, a process  $(M_t)_{t=0}^\infty$  is said to be a *supermartingale* if it is adapted to  $\mathcal{F}$  – meaning that  $M_t$  is  $\mathcal{F}_t$ -measurable for each  $t \in \mathbb{N}_0$  – and

$$\mathbb{E}(M_{t+1} | \mathcal{F}_t) \leq M_t \quad (1.13)$$

for each  $t \in \mathbb{N}_0$ . If the above inequality ( $\leq$ ) is replaced by an inequality ( $=$ ), then  $(M_t)_{t=0}^\infty$  is said to be a *martingale*. We say that a (super)martingale  $(M_t)_{t=0}^\infty$  is a *test* (super)martingale if  $M_t \geq 0$  almost surely and  $\mathbb{E}M_0 = 1$  [211]. Test supermartingales have just enough structure so that they cannot become too large except with small probability, a result due to Ville [264] and correspondingly named “Ville's inequality” which states that for any  $\alpha \in (0, 1)$ ,

$$P\left(\sup_{t \in \mathbb{N}_0} M_t \geq 1/\alpha\right) \leq \alpha. \quad (1.14)$$

How can Ville's inequality be translated into inferential procedures? Let us give one example that illustrates the general recipe later outlined in Theorem 3.2.1, ultimately culminating in the CS discussed in (1.10).

Continuing with the running example from the previous paragraphs, let  $(X_t)_{t=1}^\infty$  be i.i.d. random variables with support  $[0, 1]$  and suppose that we aim to estimate their mean  $\mu$ . By Hoeffding [120],  $[0, 1]$ -bounded random variables are sub-Gaussian with variance proxy  $1/4$  meaning that for any  $\lambda \in \mathbb{R}$ ,

$$\mathbb{E} \exp\{\lambda(X_1 - \mu)\} \leq \exp\{\lambda^2/8\}. \quad (1.15)$$

Looking at the partial sum process  $S_t(\mu) := \sum_{i=1}^t (X_i - \mu)$  and applying the above, it is routine to check that

$$M_t(\mu) := \exp(\lambda S_t(\mu) - t\lambda^2/8) \quad (1.16)$$

forms a supermartingale when  $M_0 \equiv 1$ , and is indeed nonnegative. As an immediate consequence of Ville's inequality, we have that  $p_t := 1/M_t(m)$  is an anytime  $p$ -value for the null hypothesis  $H_0^{(m)} : \mu = m$ , for any  $m \in [0, 1]$ . Viewing  $(H_0^{(m)})_{m \in [0, 1]}$  as a family of hypotheses, we have that

$$\bar{C}_t := \{m \in \mathbb{R} : M_t(m) < 1/\alpha\}^4 \quad (1.17)$$

forms a  $(1 - \alpha)$ -CS for  $\mu$ . As a brief aside, setting  $\lambda := \sqrt{8 \log(2/\alpha)/n}$  for a fixed  $n$  and using (1.17) to obtain  $\bar{C}_n$  recovers Hoeffding's inequality (1.9) exactly. Alternatively, one can appeal to Fubini's theorem to observe that any mixture  $\int_{\lambda \in \mathbb{R}} M_t(\mu) dF(\lambda)$  over  $\lambda \in \mathbb{R}$  for some distribution  $F$  yields another test supermartingale. Taking  $F$  to be a Gaussian distribution with mean zero and variance  $\rho^2 > 0$  is precisely how Robbins' mixture CS in (1.10) is obtained from (1.17); see Robbins [217], Howard et al. [125], or Chapter 7.

---

<sup>4</sup>Technically, one need only search over  $m \in [0, 1]$  rather than all of  $\mathbb{R}$ , but the latter leads to simpler expressions for the sake of exposition.

### 1.1.2 Asymptotic anytime-valid inference

One property that all of the CSs from the previous section (as well as their implicit anytime  $p$ -values and sequential tests) have in common is that they are nonasymptotically valid. That is, the miscoverage rates and type-I errors in Definition 1.1.1 are controlled in finite samples, and there is no “approximate” or “asymptotic” validity that one would see when using methods based on the central limit theorem for example. This might sound like all upside with no downside — why settle for asymptotic validity when you can have nonasymptotic (*and* time-uniform<sup>5</sup>) validity? The reason is quite simple: asymptotic inference — via confidence sets,  $p$ -values and so on — is quite easy in a plethora of settings where nonasymptotic inference is not just challenging, but *impossible*. Let us elaborate on this point now.

Consider the running example from the previous section where we observe i.i.d. observations  $(X_n)_{n=1}^\infty$  with variance  $\sigma^2$  and we would like to estimate  $\mu := \mathbb{E}_P X_1$  (or test for it) but now without the assumption that  $X_1$  is supported on  $[0, 1]$  or any other *a priori* known interval  $[a, b]$  for that matter. The classical impossibility results of Bahadur and Savage [14] tell us that without further assumptions on  $P$ , the only nonasymptotic confidence intervals,  $p$ -values, and hypothesis tests that exist are trivial ones (e.g. a level- $\alpha$  test that randomly rejects with probability  $\alpha \in (0, 1)$  irrespective of the data). Nevertheless, the central limit theorem (CLT) makes it easy to construct nontrivial *asymptotic* confidence intervals. Indeed, setting

$$\dot{C}_n := \left[ \frac{1}{n} \sum_{i=1}^n X_i \pm \hat{\sigma}_n \cdot \frac{\Phi^{-1}(1 - \alpha)}{\sqrt{n}} \right] \quad (1.18)$$

where  $\hat{\sigma}_n^2$  is the sample variance and  $\Phi^{-1}$  is the standard Gaussian quantile, we have that  $\lim_{n \rightarrow \infty} P(\mu \notin \dot{C}_n) = \alpha$ . In other words, simply changing the benchmark from nonasymptotic to *asymptotic* validity makes statistical inference for the mean  $\mu$  not just possible, but rather straightforward via the closed-form expression in (1.18).

Of course, the previous paragraph and the confidence interval (1.18) are in the context of asymptotic *fixed-n* (non-sequential) inference. Chapters 7 and 8 are interested in addressing the question of whether analogous *confidence sequences* exist with similar simplicity and leanness of assumptions. We omit the precise definitions of asymptotic confidence sequences and their corresponding coverage guarantees here for simplicity but in short, we have that under the same finite moment assumptions as those imposed on CLT-based confidence intervals that

$$\bar{C}_t := \left[ \frac{1}{t} \sum_{i=1}^t X_i \pm \hat{\sigma}_t \cdot \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)} \right] \quad (1.19)$$

forms a  $(1 - \alpha)$  *asymptotic confidence sequence* for  $\mu$ , where  $\rho > 0$  is a prespecified constant. The resemblance to the sub-Gaussian CS in (1.10) is no coincidence, and the precise connection relies on almost-sure approximation of the centered partial sums  $\sum_{i=1}^t (X_i - \mu)$  by an implicit

---

<sup>5</sup>Here and throughout, “time-uniformity” is used to refer to uniformity over sample sizes. The use of the word “time” comes from the stochastic processes literature [124].

partial sum of Gaussian random variables, the details of which can be found in Chapter 7. Moreover, the parameter  $\rho$  can be tuned to obtain a sequence  $(\rho_m)_{m=1}^\infty$  so that when  $\rho_m$  is plugged into  $\bar{C}_t$  — now indexed by  $m \geq 1$  to obtain  $(\bar{C}_t^{(m)})_{t=m}^\infty$  — we have that

$$\forall P \in \mathcal{P} \quad \lim_{m \rightarrow \infty} P \left( \exists t \geq m : \mu \notin \bar{C}_t^{(m)} \right) = \alpha, \quad (1.20)$$

where  $\mathcal{P}$  is the collection of distributions with a bounded  $(1 + \delta)^{\text{th}}$  moment — i.e.  $\mathbb{E}_P|X - \mathbb{E}_P X|^{1+\delta} < \infty$  for some  $\delta > 0$ . In other words,  $\bar{C}_t^{(m)}$  forms a sequence with approximate time-uniform  $(1 - \alpha)$ -coverage for  $\mu$ , as long as the burn-in time  $m \gg 1$  is quite large, closely resembling the asymptotic coverage guarantees of CLT-based confidence intervals. These types of guarantees were first studied in a simpler setting by Robbins and Siegmund [220], were recently considered by Bibaut et al. [34], and are studied further in Chapter 7.

While (1.20) provides a satisfying analogue of asymptotic fixed- $n$  coverage, this guarantee is *distribution-pointwise* in the sense that the limit only holds for every fixed  $P \in \mathcal{P}$  but *not* uniformly in the collection  $\mathcal{P}$ . By contrast, fixed- $n$  CIs can be shown to have the stronger guarantee that if  $\sup_{P \in \mathcal{P}} \mathbb{E}_P|X - \mathbb{E}_P X|^{2+\delta} < \infty$  for some  $\delta > 0$  and  $\inf_{P \in \mathcal{P}} \text{Var}_P(X) > 0$ , then the CLT-based CI of (1.18) has the  $\mathcal{P}$ -uniform guarantee

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \mu \notin \dot{C}_n \right) = \alpha. \quad (1.21)$$

This begs the question of whether it is possible to obtain a uniform analogue of (1.20). This is the central question posed and answered in Chapter 8, where it is shown that under the same conditions as the  $\mathcal{P}$ -uniform CI above, one can construct sets  $(\bar{C}_t^{(m)})_{t=m}^\infty$  indexed by  $m \geq 1$  so that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \exists t \geq m : \mu \notin \bar{C}_t^{(m)} \right) = \alpha. \quad (1.22)$$

It turns out that deriving such a collection of sets  $\bar{C}_t^{(m)}$  satisfying (1.22) is deceptively hard for the following reason. As previously alluded to, the derivation of bounds with time-uniform (but distribution-pointwise) guarantees as in (1.20) heavily relies on some powerful results in probability theory called *strong Gaussian approximations*, such as those due to Strassen [248, 249] or Komlós, Major, and Tusnády [160, 161] where partial sums can be almost surely approximated by implicit partial sums of Gaussian random variables. However, all of these results are fundamentally  $P$ -pointwise as they provide statements which hold  $P$ -almost surely for a fixed  $P$  but are not *distribution-uniform* in any sense. As such, one of the main contributions of Chapter 8 is the definition and construction of distribution-uniform (as well as nonasymptotic) strong Gaussian approximations that can then be used to derive anytime-valid statistical methods with the guarantee found in (1.22).

There is one other key ingredient needed to derive distribution-uniform coverage guarantees of the type in (1.22) — the ability to estimate certain nuisance functions *almost surely* and *uniformly* in a collection of distributions  $\mathcal{P}$ . This topic is given a detailed treatment in Chapter 9 where we give sufficient (and in some cases *necessary*) conditions for strong laws of

large numbers to hold uniformly in a class of distributions  $\mathcal{P}$ .

## 1.2 A detailed outline of this thesis

As previously alluded to, this thesis is divided into two parts: *nonasymptotic* and *asymptotic* anytime-valid inference. Each part contains a few chapters and we give a summary of each of their contents below.

### Part I: Nonasymptotic inference and time-uniform concentration.

- **Chapter 2: Confidence sequences for sampling without replacement.**

We begin by extending Hoeffding’s inequality and an empirical Bernstein inequality for bounded random variables to the setting of *sampling without replacement* where a finite list of (nonrandom) numbers are sampled (whether sequentially or in a batch) in a random permutation. These improve on some existing concentration inequalities due to Serfling [232] and Bardenet and Maillard [19] and yield sharp confidence intervals and time-uniform confidence sequences. Separately, we briefly take a frequentist view of Bayesian methods to show that the ratio of a prior distribution to the resulting posterior distribution forms a nonnegative martingale when evaluated at the true parameter value. Consequently, these can be used to derive confidence sequences, anytime-valid  $p$ -values, and sequential tests. This result is then used to obtain computationally tractable (frequentist) confidence sets for the parameters of the multivariate hypergeometric distribution for sampling without replacement by exploiting conjugacy of the Dirichlet-multinomial prior.

*Chapter 2 is based on Waudby-Smith and Ramdas [281] which appeared in the Conference on Neural Information Processing Systems (NeurIPS) and was selected for a Spotlight talk.*

- **Chapter 3: Estimating means of bounded random variables by betting**

This chapter continues with the theme of concentration inequalities and confidence sets, ultimately providing sharp solutions to the problem of estimating means of bounded random variables in the following four settings: at fixed sample sizes and at data-dependent stopping times, both with and without replacement. The perspective taken in this chapter is one where the estimation problem is viewed as an inversion of a family of testing problems, each of which are re-interpreted as a *betting* or *gambling* problem, both for intuitive and mathematical reasons. In short, we set up a family of simple mathematical games, each of which represents a particular null hypothesis against which an imagined gambler with an initial wealth of one dollar is allowed to bet. The larger the gambler’s wealth becomes over successive rounds of the game, the more evidence against that particular null. Crucially, we design the game so that if the null is true, then the gambler’s wealth forms a nonnegative martingale so that it cannot become too large except with small probability, a phenomenon can be directly translated to a statement about type-I error control via Ville’s inequality [264]. Deriving powerful statistical tests (and hence downstream confidence sets) then reduces to the task of coming up with smart betting

strategies for the imagined gamblers. We show through extensive simulation that the resulting tests and confidence sets outperform all known prior work in the literature in a variety of settings.

*Chapter 3 is based on Waudby-Smith and Ramdas [282] which appeared in the Journal of the Royal Statistical Society (Series B) as a discussion paper.*

- **Chapter 4: RiLACS: Risk-limiting audits via confidence sequences.**

This chapter focuses on a particular application of some of the ideas contained in Chapter 3 to *risk-limiting post-election audits*, typically shortened to “RLA”. There are several reasons why it is of interest to audit the asserted outcome of an election. As one example, consider electronic voting machines which are often used to automatically tally paper ballots cast by voters. While this automation is convenient since manual tallying is time-consuming and expensive, such machines may be subject to software bugs or outright malicious tampering, and hence it is crucial to have some way of auditing the assertions that they make. The work of Stark [246] reduces the problem of auditing an election to one of *sequential hypothesis testing when sampling without replacement*, making the time-uniform and without-replacement techniques of Chapters 2 and 3 immediately applicable.<sup>6</sup> The improved statistical power of those methods directly translates to election audits that can be stopped earlier, requiring fewer ballots to be checked. This chapter contains several other advances, including betting strategies tailored to the auditing problem, an observation that the perspective of estimation rather than testing allows for one to audit several assertions simultaneously without needing multiplicity corrections, among others.

*Chapter 4 is based on Waudby-Smith, Stark, and Ramdas [283] which appeared in The International Conference for Electronic Voting (E-Vote-ID) and received a Best Paper Award.*

- **Chapter 5: Nonparametric extensions of randomized response for private confidence sets.**

The main goal of this chapter is to derive statistical procedures for the mean both in the fixed- $n$  and sequential settings under the constraint of *local differential privacy*. The methods of the previous chapters (and indeed the majority of statistical methods in the literature) rely on having access to the raw data  $X_1, \dots, X_n$ , but what if this data consists of sensitive information that individuals are not keen to hand over to data collector that may have less-than-perfect security practices for storing data at rest, or who may simply be untrustworthy in the eyes of that individual. Local differential privacy can be thought of as an information-theoretic framework for adding noise to an individual’s data point  $X_i$  to obtain a privatized  $Z_i$  that contains very little information about that specific individual, but may still contain enough to gain population-level insights when taken together with  $Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n$ . However, the literature on constructing nonasymptotic confidence sets for population parameters is relatively sparse, especially under nonparametric assumptions and in sequential settings. To develop such methods, we first introduce a privacy mechanism that generalizes Warner’s famous “randomized

---

<sup>6</sup>Stark’s reduction is applicable to a wide array of election types, not just plurality elections that are more common in the United States, Canada, and the United Kingdom, for example.

response” mechanism [278] to arbitrary bounded random variables. We then apply concentration techniques to obtain confidence intervals and sequences, as well as  $e$ -processes, anytime  $p$ -values, and sequential tests that cleanly generalize Hoeffding’s inequality and the methods found in Chapter 3 to *privatized* data. The explicit widths of the former match the minimax rates for locally private point estimation due to Duchi et al. [92].

*Chapter 5 is based on Waudby-Smith, Wu, and Ramdas [286] which appeared in the International Conference on Machine Learning (ICML) and was selected for an Oral presentation.*

- **Chapter 6: Anytime-valid off-policy inference for contextual bandits.**

One aspect that the previous chapters have in common is that they implicitly operate in the so-called “on-policy” setting where the functional of interest (e.g. the mean) is a functional of sampling distribution. Said differently, if the data  $(Y_n)_{n=1}^{\infty}$  come from a distribution  $P$ , then parameter of interest is the mean  $\mathbb{E}_P Y$  under  $P$ . By contrast, in causal inference and contextual bandits, it is typically of interest to estimate a functional such as  $\mathbb{E}_Q Y$  under the distribution  $Q$  when given samples from a different distribution  $P$ . For example, subjects in a clinical trial may be assigned treatment with probability 1/2 but we ideally want to understand the resulting health outcomes *if those subjects were assigned treatment (or control) with probability one*. The contextual bandit framework is a generalization of this setting where contexts  $X$  (“covariates”), actions  $A$  (“treatments”), and rewards  $R$  (“outcomes”) are observed in triplets  $(X_t, A_t, R_t)_{t=1}^{\infty}$ , and actions are taken with respect to a conditional distribution over actions  $h(\cdot | X)$  — referred to as a “policy”. The aim of *off-policy inference* is to understand properties of the reward distribution  $R$  *if* actions were taken with respect to some other policy  $\pi$ . We develop confidence sequences for the expected reward under  $\pi$  as well as time-uniform simultaneous confidence bands for the entire reward distribution function under  $\pi$ . These methods satisfy two key desiderata. First, the policies  $(h_t)_{t=1}^{\infty}$  that generate actions  $(A_t)_{t=1}^{\infty}$  can vary over time and can depend on all previous observations (which is the case for online learning algorithms, for example). Second, unlike many nonasymptotic methods in the off-policy literature, the maximal importance weight  $\text{ess sup}_{t,a,x} \pi(a | x)/h_t(a | x)$  need not be known, or even uniformly bounded. This not only allows these methods to be applied in a wider range of problems, but also allows our methods to be purely variance adaptive without needing to suffer from conservative bounds on this maximal value.

*Chapter 6 is based on Waudby-Smith, Wu, Ramdas, Karampatziakis, and Mineiro [284] which appeared in the ACM/IMS Journal of Data Science. Much of the work in this chapter was performed while an intern at Microsoft Research, hosted by Paul Mineiro.*

## Part II: Asymptotic inference, strong laws, and Gaussian approximation.

- **Chapter 7: Time-uniform central limit theory and asymptotic confidence sequences.**

Following the discussion in Section 1.1.1, all of the chapters in Part I take a nonasymptotic perspective of anytime-valid inference. While nonasymptotic methods are often desirable due to their strong finite-sample guarantees, they correspondingly tend to rely on

stronger assumptions and cannot be used in the same range of problems that asymptotic methods can. While methods such as central limit theorem-based confidence intervals are mature and well-studied in the fixed- $n$  regime, the literature on *anytime-valid* asymptotic methods was sparse-to-nonexistent. This work takes a first step in expanding this literature by both defining and deriving so-called *asymptotic confidence sequences* as well as corresponding asymptotic time-uniform coverage guarantees. Satisfyingly, the assumptions needed to obtain *time-uniform* asymptotic guarantees are effectively the same as those needed in the fixed- $n$  regime, forming a true anytime-valid parallel of central limit theorem-based inference. A special emphasis is placed on applications to sequential causal inference with nuisance parameters, especially in observational studies where *nonasymptotic* inference is impossible. All of the methods in this chapter centrally rely on strong Gaussian approximation theorems, a rich corner of the probability literature that had previously seen limited interface with statistical inference.

*Chapter 7 is based on Waudby-Smith, Arbour, Sinha, Kennedy, and Ramdas [287] which has been accepted for publication in the Annals of Statistics. Part of this work was performed while an intern at Adobe Research, hosted by David Arbour and Ritwik Sinha. Applications of methods herein to statistical problems at Adobe are summarized in Maharaj et al. [180], but those discussions have been omitted from the present document.*

- **Chapter 8: Distribution-uniform anytime-valid sequential inference.**

In this chapter, we follow up and improve on Chapter 7 by deriving statistical procedures with asymptotic time-uniform coverage and type-I error control guarantees that hold *uniformly* in a collection of distributions. While Chapter 7 manages to bring many of the benefits of asymptotic inference to the anytime-valid regime, the methods therein were only shown to satisfy *distribution-pointwise* asymptotic guarantees. To elaborate, Chapter 7 shows that for the mean estimation problem (say), asymptotic  $(1 - \alpha)$  coverage holds in the limit, meaning that as the first peeking time (or the “burn-in” time)  $m \geq 1$  approaches infinity, the probability of simultaneously covering the mean for all sample sizes larger than  $m$  approaches  $(1 - \alpha)$ . While this guarantee holds for any distribution  $P \in \mathcal{P}$  satisfying some central limit theorem-like moment finiteness assumptions, it was *not* shown that the time-uniform coverage probability would converge to  $(1 - \alpha)$  *uniformly* in that collection of distributions  $\mathcal{P}$ . The reason for this stems from the fact that the proofs found in Chapter 7 rely on strong Gaussian approximation theorems like those due to Strassen [248, 249] and Komlós, Major, and Tusnády [160, 161], but these existing results are fundamentally distribution-pointwise. As such, one of the focal points of Chapter 8 is in the derivation of  $\mathcal{P}$ -uniform strong Gaussian approximation theorems which may be of purely probabilistic interest. When paired with several other lemmas concerning  $\mathcal{P}$ - and time-uniform convergence of random variables, Chapter 8 yields a framework for  $\mathcal{P}$ -uniform anytime-valid inference. As one application of this framework, we provide the first anytime-valid test of conditional independence that does not rely on the Model-X assumption. That is, given i.i.d. triplets  $(X_n, Y_n, Z_n)_{n=1}^{\infty}$  taking values in  $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ , we are interested in deriving powerful sequential tests for the conditional independence null  $H_0 : X \perp\!\!\!\perp Y | Z$  versus the alternative that some conditional dependence exists. Existing anytime-valid tests of  $H_0$  rely on the Model-X

assumption so that  $X \mid Z$  is known exactly. We instead take a fully nonparametric approach in the spirit of Shah and Peters [238], allowing for powerful tests of  $H_0$  to be derived under assumptions about the estimability of certain regression functions. These analyses heavily rely on some new distribution-uniform strong laws of large numbers for independent random variables, which form a small part of the following chapter.

*Chapter 8 is based on the recent work of Waudby-Smith, Kennedy, and Ramdas [285].*

- **Chapter 9: Distribution-uniform strong laws of large numbers.**

In this chapter, we revisit the question of whether the strong law of large numbers (SLLN) holds uniformly in a rich family of distributions, culminating in a distribution-uniform generalization of the Marcinkiewicz-Zygmund SLLN. These results can be viewed as extensions of Chung's distribution-uniform SLLN to random variables with uniformly integrable  $q^{\text{th}}$  absolute central moments for  $0 < q < 2$ ;  $q \neq 1$ . Furthermore, we show that uniform integrability of the  $q^{\text{th}}$  moment is both sufficient and necessary for the SLLN to hold uniformly at the Marcinkiewicz-Zygmund rate of  $n^{1/q-1}$ . These proofs centrally rely on novel distribution-uniform analogues of some familiar almost sure convergence results including the Khintchine-Kolmogorov convergence theorem, Kolmogorov's three-series theorem, a stochastic generalization of Kronecker's lemma, and the Borel-Cantelli lemmas. We also consider the non-identically distributed case and an application to strongly consistent variance estimation, both of which are used in the proofs for distribution-uniform anytime-valid inference procedures found in Chapter 8.

*Chapter 9 is based on the recent work of Waudby-Smith, Larsson, and Ramdas [288].*

## **Part I**

# **Nonasymptotic inference and time-uniform concentration**

# Chapter 2

## Confidence sequences for sampling without replacement

### 2.1 Introduction

This chapter derives closed-form CSs when samples are drawn without replacement (WoR) from a finite population. The technical underpinnings are novel (super)martingales for both categorical (Section 2.2) and continuous (Section 2.3) observations. In the latter setting, our results unify and improve on the time-uniform with-replacement extensions of Hoeffding's [120] and empirical Bernstein's inequalities by Maurer and Pontil [187] that have been derived recently [124, 125], with several related inequalities for sampling WoR by Serfling [232] and extensions by Bardenet and Maillard [19] and Greene and Wellner [110].

**Outline.** In Section 2.2, we use Bayesian ideas to obtain frequentist CSs for categorical observations. In Section 2.3, we construct CSs for the mean of a finite set of bounded real numbers. We discuss implications for testing in Section 2.4. Some prototypical applications are described in Appendix 2.A. The other appendices contain proofs, choices of tuning parameters, and computational considerations.

#### 2.1.1 Notation, supermartingales and the model for sampling WoR

Everywhere in this chapter, the  $N$  objects in the finite population  $\{x_1, \dots, x_N\}$  are fixed and nonrandom. In the discrete setting (Section 2.2) with  $K \geq 2$  categories  $\{c_k\}_{k=1}^K$ , we have  $x_i \in \{c_1, c_2, \dots, c_K\}$ . In the continuous setting (Section 2.3),  $x_i \in [\ell, u]$  for some known bounds  $\ell < u$ . What is random is only the order of observation; the model for sampling uniformly at random WoR posits that

$$X_t \mid \{X_1, \dots, X_{t-1}\} \sim \text{Uniform}(\{x_1, \dots, x_N\} \setminus \{X_1, \dots, X_{t-1}\}). \quad (2.1)$$

---

<sup>1</sup>Code to reproduce plots is available at [github.com/wannabesmith/confseq\\_wor](https://github.com/wannabesmith/confseq_wor).

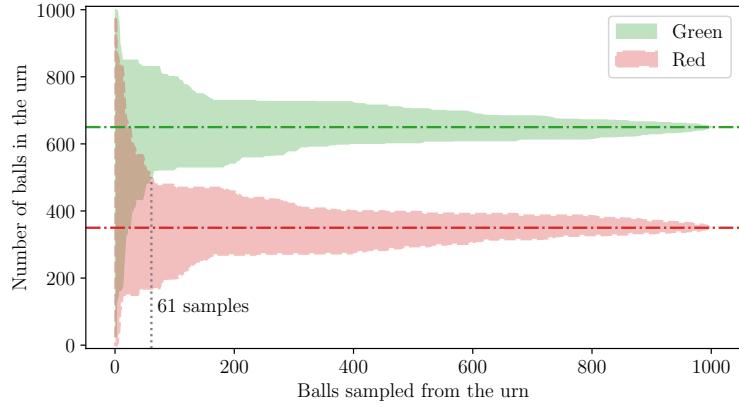


Figure 2.1: 95% CS for the number of green and red balls in an urn by sampling WoR<sup>1</sup>. Notice that the true totals (650 green, 350 red) are captured by the CSs uniformly over time from the initial sample until all 1000 balls are observed. After sampling 61 balls in this example, the CSs cease to overlap, and we can conclude with 95% confidence that there are more green than red balls in the urn.

All probabilities in this chapter are to be understood as solely arising from observing fixed entities in a random order, with no distributional assumptions being made on the finite population. It is worth remarking on the power of this randomization—as demonstrated in our experiments, one can estimate the average of a deterministic set of numbers to high accuracy without observing a large fraction of the set.

The results in this chapter draw from the theory of *supermartingales*. While they can be defined in more generality, we provide a definition of supermartingales which will suffice for the theorems that follow.

A filtration is an increasing sequence of sigma fields. For the entirety of this chapter, we consider the ‘canonical’ filtration  $(\mathcal{F}_t)_{t=0}^N$  defined by  $\mathcal{F}_t := \sigma(X_1, \dots, X_t)$ , with  $\mathcal{F}_0$  is the empty or trivial sigma field. For any fixed  $N \in \mathbb{N}$ , a stochastic process  $(M_t)_{t=0}^N$  is said to be a *supermartingale* with respect to  $(\mathcal{F}_t)_{t=0}^N$  if for all  $t \in \{0, 1, \dots, N-1\}$ ,  $M_t$  is measurable with respect to  $\mathcal{F}_t$  (informally,  $M_t$  is a function of  $X_1, \dots, X_t$ ), and

$$\mathbb{E}(M_{t+1} | \mathcal{F}_t) \leq M_t.$$

If the above inequality is replaced by an equality for all  $t$ , then  $(M_t)_{t=0}^N$  is said to be a *martingale*.

For succinctness, we use the notation  $a_1^t := \{a_1, \dots, a_t\}$  and  $[a] := \{1, \dots, a\}$ . Using this terminology, one can rewrite model (2.1) as positing that  $X_t | \mathcal{F}_{t-1} \sim \text{Uniform}(x_1^N \setminus X_1^{t-1})$ .

## 2.2 Discrete categorical setting

When observations are of this discrete form, the variables can be rewritten in such a way that they follow a hypergeometric distribution. In such a setting, the following “prior-posterior-ratio martingale” can be used to obtain CSs for parameters of the hypergeometric distribution which shrink to a single point after all data have been observed.

### 2.2.1 The prior-posterior-ratio (PPR) martingale

While the PPR martingale will be particularly useful for obtaining CSs when sampling discrete categorical random variables WoR from a finite population, it may be employed whenever one is able to compute a posterior distribution, and is certainly *not limited to this chapter’s setting*. Moreover, this posterior distribution need not be computed in closed form, and computational techniques such as Markov Chain Monte Carlo may be employed when a conjugate prior is not available or desirable.

To avoid confusion, we emphasize that while we make use of terminology from Bayesian inference such as posteriors and conjugate priors, all of the probability statements with regards to CSs should be read in the frequentist sense, and are not interpreted as sequences of credible intervals.

Consider any family of distributions  $\{F_\theta\}_{\theta \in \Theta}$  with density  $f_\theta$  with respect to some underlying common measure (such as Lebesgue for continuous cases, counting measure for discrete cases). Let  $\theta^* \in \Theta$  be a fixed parameter and let  $\mathcal{T} = [N]$  where  $N \in \mathbb{N} \cup \{\infty\}$ . Suppose that  $X_1 \sim f_{\theta^*}(x)$  and

$$X_{t+1} \sim f_{\theta^*}(x | X_1^t) \quad \text{for all } t \in \mathcal{T}.$$

Let  $\pi_0(\theta)$  be a prior distribution on  $\Theta$ , with posterior given by

$$\pi_t(\theta) = \frac{\pi_0(\theta) f_\theta(X_1^t)}{\int_{\eta \in \Theta} \pi_0(\eta) f_\eta(X_1^t) d\eta}.$$

To prepare for the result that follows, define the *prior-posterior ratio (PPR)* evaluated at  $\theta \in \Theta$  as

$$R_t(\theta) := \frac{\pi_0(\theta)}{\pi_t(\theta)}.$$

**Proposition 2.2.1** (Prior-posterior-ratio martingale). *For any prior  $\pi_0$  on  $\Theta$  that assigns nonzero mass everywhere, the sequence of prior-posterior ratios evaluated at the true  $\theta^*$ , that is  $(R_t(\theta^*))_{t=0}^N$ , is a nonnegative martingale with respect to  $(\mathcal{F}_t)_{t=0}^N$ . Further, the sequence of sets*

$$C_t := \{\theta \in \Theta : R_t(\theta) < 1/\alpha\}$$

*forms a  $(1 - \alpha)$ -CS for  $\theta^*$ , meaning that  $\Pr(\exists t \in \mathcal{T} : \theta^* \notin C_t) \leq \alpha$ .*

The proof is given in Appendix 2.B.1.

Going forward, we adopt the label *working* before ‘prior’ and ‘posterior’ and encase them in ‘quotes’ to emphasize that they constitute part of a Bayesian ‘working model’, to contrast it against an assumed Bayesian model; the latter would be inappropriate given the discussion in Section 2.1.1. Next, we apply this result to the hypergeometric distribution. We will later examine the practical role of this working prior.

### 2.2.2 CSs for binary settings using the hypergeometric distribution

Recall that a random variable  $X$  has a hypergeometric distribution with parameters  $(N, N^+, n)$  if it represents the number of “successes” in  $n$  random samples WoR from a population of size  $N$  in which there are  $N^+$  such successes, and each observation is either a success or failure (1 or 0). The probability of a particular number of successes  $x \in \{0, 1, \dots, \min(N^+, n)\}$  is

$$\Pr(X = x) = \binom{N^+}{x} \binom{N-N^+}{n-x} / \binom{N}{n}.$$

For notational simplicity, we consider the case when  $n = 1$ , that is we make one observation at a time, but this is not a necessary restriction. In fact, one would obtain the same CS at time ten if we repeatedly make one observation ten times, or make ten observations in one go. For a moment, let us view this problem from the Bayesian perspective, treating the fixed parameter  $N^+$  as a random parameter, which we call  $\tilde{N}^+$  to avoid confusion. We choose a beta-binomial ‘working prior’ on  $\tilde{N}^+$  as it is conjugate to the hypergeometric distribution up to a shift in  $\tilde{N}^+$  [104]. Concretely, suppose

$$\begin{aligned} X_t | (\tilde{N}^+, X_1, \dots, X_{t-1}) &\sim \text{HyperGeo}\left(N - (t-1), \tilde{N}^+ - \sum_{i=1}^{t-1} X_i, 1\right), \\ \tilde{N}^+ &\sim \text{BetaBin}(N, a, b), \end{aligned}$$

for some  $a, b > 0$ . Then for any  $t \in [N]$ , the ‘working posterior’ for  $\tilde{N}^+$  is given by

$$\tilde{N}^+ - \sum_{i=1}^t X_i | X_1^t \sim \text{BetaBin}\left(N - t, a + \sum_{i=1}^t X_i, b + t - \sum_{i=1}^t X_i\right).$$

Now that we have ‘prior’ and ‘posterior’ distributions for  $\tilde{N}^+$ , an application of the prior-posterior martingale (Proposition 2.2.1) yields a CS for the true  $N^+$ , summarized in the following theorem.

**Theorem 2.2.1** (CS for binary observations). *Suppose  $x_1^N \in \{0, 1\}^N$  is a nonrandom set with the number of successes  $\sum_{i=1}^N x_i \equiv N^+$  fixed and unknown. Under observation model (2.1), we have*

$$X_t | X_1^{t-1} \sim \text{HyperGeo}\left(N - (t-1), N^+ - \sum_{i=1}^{t-1} X_i, 1\right).$$

For any beta-binomial ‘prior’  $\pi_0$  for  $N^+$  with parameters  $a, b > 0$  and induced ‘posterior’  $\pi_t$ ,

$$C_t := \left\{ n^+ \in [N] : \frac{\pi_0(n^+)}{\pi_t(n^+)} < \frac{1}{\alpha} \right\}$$

is a  $(1-\alpha)$ -CS for  $N^+$ . Further, the running intersection,  $(\bigcap_{s \leq t} C_s)_{t \in [N]}$  is also a valid  $(1-\alpha)$ -CS.

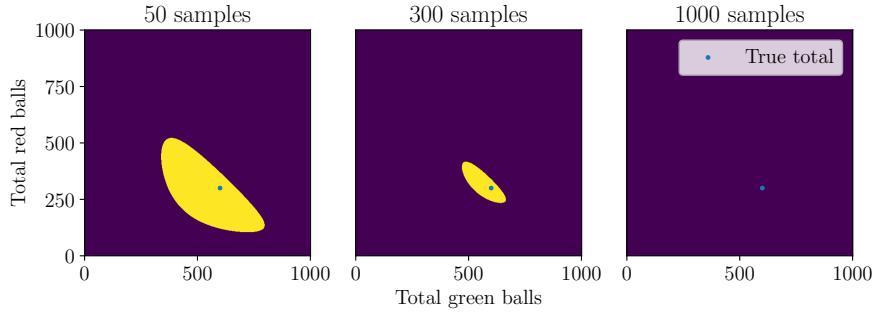


Figure 2.2: Consider sampling balls from an urn WoR with three distinct colors (red, green, and purple). In this example, the urn contains 1000 balls with 300 red, 600 green, and 100 purple. We only require a two-dimensional confidence sequence (yellow region) to capture uncertainty about all three totals. After around 300 balls have been sampled, we are quite confident that the urn is made up mostly of green; after 1000 samples, we know the totals for each color with certainty.

The proof of Theorem 2.2.1 is a direct application of Proposition 2.2.1. Note that for any ‘prior’, the ‘posterior’ at time  $t = N$  is  $\pi_N(n^+) = \mathbb{1}(n^+ = N^+)$ , so  $C_t$  shrinks to a point, containing only  $N^+$ . For  $K > 2$  categories, Theorem 2.2.1 can be extended to use a multivariate hypergeometric with a Dirichlet-multinomial prior to yield higher-dimensional CSs, but we leave the (notationally heavy) derivation to Appendix 2.C. See Figure 2.2 to get a sense of what these CSs can look like when  $K = 3$ .

### 2.2.3 Role of the ‘prior’ in the prior-posterior CS

The prior-posterior CSs discussed thus far have valid (frequentist) coverage for any ‘prior’ on  $N^+$ , and in particular are valid for a beta-binomial ‘prior’ with any data-independent choices of  $a, b > 0$ . Importantly, the corresponding CS always shrinks to zero width. How, then, should the user pick  $(a, b)$ ? Figure 2.3 provides some visual intuition.

These are our takeaway messages: (a) if the ‘prior’ is very accurate (coincidentally peaked at the truth), the resulting CS is narrowest, (b) even if the ‘prior’ is horribly inaccurate (placing almost no mass at the truth), the resulting CS is well-behaved and robust, albeit wider, (c) if we do not actually have any idea what the underlying truth might be, we suggest using a uniform ‘prior’ to safely balance the two extremes. However, a more risky ‘prior’ pays a relatively low statistical price.

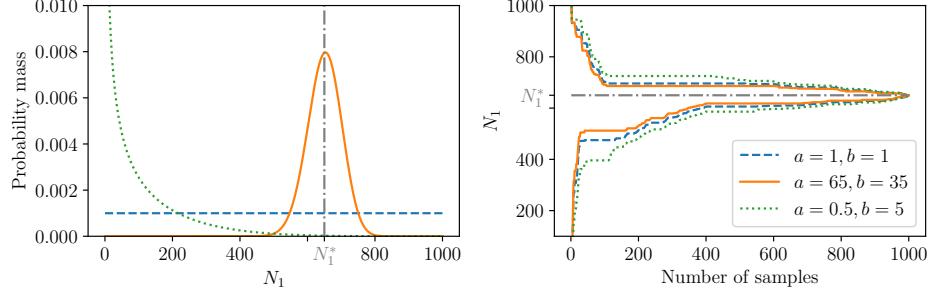


Figure 2.3: Beta-binomial probability mass function as a ‘prior’ on  $N_1^*$  with different choices of  $(a, b)$ , and the resulting PPR CS for the parameter  $N_1^*$  of a hypergeometric distribution when  $(N_1^*, N_2^*) = (650, 350)$ .

## 2.3 Bounded real-valued setting

Suppose now that observations are real-valued and bounded as in Examples C and D of Appendix 2.A. Here we introduce Hoeffding- and empirical Bernstein-type inequalities for sampling WoR.

### 2.3.1 Hoeffding-type bounds

Recalling Section 2.1.1, we deal with a fixed batch  $x_1^N$  of bounded real numbers  $x_i \in [\ell, u]$  with mean  $\mu := \frac{1}{N} \sum_{i=1}^N x_i$ . Our CS for  $\mu$  will utilize a novel WoR mean estimator,

$$\hat{\mu}_t := \frac{\sum_{i=1}^t X_i + \sum_{i=1}^t \frac{1}{N-i+1} \sum_{j=1}^{i-1} X_j}{t + \sum_{i=1}^t \frac{i-1}{N-i+1}}. \quad (2.2)$$

More generally, if  $\lambda_1, \dots, \lambda_N$  is a predictable sequence (meaning  $\lambda_t$  is  $\mathcal{F}_{t-1}$ -measurable for  $t \in \{1, \dots, N\}$ ), then we may define the weighted WoR mean estimator,

$$\hat{\mu}_t(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i (X_i + \frac{1}{N-i+1} \sum_{j=1}^{i-1} X_j)}{\sum_{i=1}^t \lambda_i (1 + \frac{i-1}{N-i+1})}, \quad (2.3)$$

where it should be noted that if  $\lambda_1 = \dots = \lambda_N$  then  $\hat{\mu}_t(\lambda_1^t)$  recovers  $\hat{\mu}_t$ . Past WoR works [232, 19, 110] base their bounds on the sample average  $\sum_i X_i / t$ . Both  $\hat{\mu}_t$  and the sample average are conditionally biased and unconditionally unbiased (see Appendix 2.B.2 for more details). As frequently encountered in Hoeffding-style inequalities for bounded random variables [120], define

$$\psi_H(\lambda) := \frac{\lambda^2(u - \ell)^2}{8}. \quad (2.4)$$

Setting  $M_0^H := 1$ , we introduce a new exponential Hoeffding-type process for a predictable sequence  $\lambda_1^N$ ,

$$M_t^H := \exp \left\{ \sum_{i=1}^t \left[ \lambda_i \left( X_i - \mu + \frac{1}{N-i+1} \sum_{j=1}^{i-1} (X_j - \mu) \right) - \psi_H(\lambda_i) \right] \right\}. \quad (2.5)$$

**Theorem 2.3.1** (A time-uniform Hoeffding-type CS for sampling WoR). *Under the observation model and filtration  $(\mathcal{F}_t)_{t=0}^N$  of Section 2.1.1, and for any predictable sequence  $\lambda_1^N$ , the process  $(M_t^H)_{t=0}^N$  is a nonnegative supermartingale, and thus,*

$$\Pr \left( \exists t \in [N] : \mu - \hat{\mu}_t(\lambda_1^t) \geq \frac{\sum_{i=1}^t \psi_H(\lambda_i) + \log(1/\alpha)}{\sum_{i=1}^t \lambda_i \left( 1 + \frac{i-1}{N-i+1} \right)} \right) \leq \alpha.$$

Consequently,

$$C_t^H := \hat{\mu}_t(\lambda_1^t) \pm \frac{\sum_{i=1}^t \psi_H(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left( 1 + \frac{i-1}{N-i+1} \right)} \quad \text{forms a } (1-\alpha)\text{-CS for } \mu.$$

The proof in Appendix 2.B.2 combines ideas from the with-replacement, *time-uniform* extension of Hoeffding's inequality of Howard et al. [125, 124] with the fixed-time, *without-replacement* extension of Hoeffding's by Bardenet & Maillard [19], to yield a bound that improves on both. When  $\lambda := \lambda_1 = \dots = \lambda_N$  is a constant, the term

$$A_t := \sum_{i=1}^t \frac{i-1}{N-i+1} \quad (2.6)$$

captures the ‘advantage’ over the classical Hoeffding’s inequality; we discuss this term more soon.

In order to use the aforementioned CS, one needs to choose a predictable  $\lambda$ -sequence. First, consider the simpler case of a fixed real-valued  $\lambda := \lambda_1 = \dots = \lambda_N$  as this will aid our intuition in choosing a more complex  $\lambda$ -sequence. In this case,  $\lambda$  corresponds to a time  $t_0 \in [N]$  for which the CS is tightest. If the user wishes to optimize the width of the CS for time  $t_0$ , then the corresponding  $\lambda$  to be used is given by

$$\lambda := \sqrt{\frac{8 \log(2/\alpha)}{t_0(u-\ell)^2}}. \quad (2.7)$$

Alternatively, if the user does not wish to commit to a single time  $t_0$ , they can choose a  $\lambda$ -sequence akin to (2.7) but which spreads its width optimization over time. For example, one

can use the sequence for  $t \in \{1, \dots, N\}$ ,

$$\lambda_t := \sqrt{\frac{8 \log(2/\alpha)}{t \log(t+1)(u-\ell)^2}} \wedge \frac{1}{u-\ell}, \quad (2.8)$$

where the minimum was taken to prevent the CS width from being dominated by early terms. Note however that any predictable  $\lambda$ -sequence yields a valid CS (see Appendix 2.E for more examples).

Optimizing a real-valued  $\lambda = \lambda_1 = \dots = \lambda_N$  for a particular time is in fact the typical strategy used to obtain the tightest fixed-time (i.e. non-sequential) Chernoff-based confidence intervals (CIs) such as those based on Hoeffding's inequality [125, 120]. This same strategy can be used with our WoR CSs to obtain tight fixed-time CIs for sampling WoR. Specifically, plugging (2.7) into Theorem 2.3.1 for a fixed sample size  $n \in [N]$ , we obtain the following corollary.

**Corollary 2.3.1** (Hoeffding-type CI for sampling WoR). *For any  $n \in [N]$ ,*

$$\hat{\mu}_n \pm \frac{\sqrt{\frac{1}{2}(u-\ell)^2 \log(2/\alpha)}}{\sqrt{n} + A_n/\sqrt{n}} \text{ forms a } (1-\alpha) \text{ CI for } \mu. \quad (2.9)$$

Notice that the classical Hoeffding confidence interval is recovered exactly, including constants, by dropping the  $A_n$  term and using the usual sample mean estimator instead of  $\hat{\mu}_t$ . To get a sense of how large the advantage is, note that

$$\begin{aligned} \text{for small } n \ll N, \quad A_n &\asymp \sum_{i=1}^{n-1} i/N \asymp n^2/N, \\ \text{for large } n \approx N, \quad A_n &\asymp A_N = \sum_{i=1}^{N-1} \frac{i}{N-i} = \sum_{j=1}^{N-1} \frac{N-j}{j} \asymp N \log N - (N-1). \end{aligned}$$

Thus, the advantage is negligible for  $n = O(\sqrt{N})$ , while it is substantial for  $n = O(N)$ , but it is clear that the CI of (2.9) is strictly tighter than Hoeffding's inequality for any  $n$ .

### 2.3.2 Empirical Bernstein-type bounds

Hoeffding-type bounds like the one in Theorem 2.3.1 only make use of the fact that observations are bounded, and they can be loose if only some observations are near the boundary of  $[\ell, u]$  while the rest are concentrated near the middle of the interval. More formally, the CS of Theorem 2.3.1 has the same width whether the underlying population  $x_1^N$  has large or small variance  $\sum_{i=1}^N (x_i - \mu)^2$ —thus, they are tightest when the  $x_i$ s equal  $\ell$  or  $u$ , and they are loosest when  $x_i \approx (\ell + u)/2$  for all  $i$ . As an alternative that adaptively takes a variance-like term into account [187, 18], we introduce a sequential, WoR, empirical Bernstein CS. As is typical in

empirical Bernstein bounds [125], we use a different ‘subexponential’-type function,

$$\psi_E(\lambda) := (-\log(1 - c\lambda) - c\lambda)/4 \quad \text{for any } \lambda \in [0, 1/c)$$

where  $c := u - \ell$ .  $\psi_E$  seems quite different from  $\psi_H$ , but Taylor expanding  $\log$  yields  $\psi_E(\lambda) \approx c^2\lambda^2/8$ . Indeed,

$$\lim_{\lambda \rightarrow 0} \psi_E(\lambda)/\psi_H(\lambda) = 1. \quad (2.10)$$

Note that one typically picks small  $\lambda$ , e.g.: set  $t_0 = N/2, \ell = -1, u = 1$  in (2.7) to get  $\lambda_1 \propto 1/\sqrt{N}$ .

In what follows, we derive a time-uniform empirical-Bernstein inequality for sampling WoR. Similar to Theorem 2.3.1, underlying the bound is an exponential supermartingale. Set  $M_0^E = 1$ , and recall that  $c = u - \ell$  to define a novel exponential process for any  $[0, 1/c]$ -valued predictable sequence  $\lambda_1, \dots, \lambda_N$ :

$$M_t^E := \exp \left\{ \sum_{i=1}^t \left[ \lambda_i \left( X_i - \mu + \frac{1}{N-i+1} \sum_{j=1}^{i-1} (X_j - \mu) \right) - \left( \frac{c}{2} \right)^{-2} (X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i) \right] \right\}. \quad (2.11)$$

**Theorem 2.3.2** (A time-uniform empirical Bernstein-type CS for sampling WoR). *Under the observation model and filtration  $(\mathcal{F}_t)_{t=0}^N$  of Section 2.1.1, and for any  $[0, 1/c]$ -valued predictable sequence  $\lambda_1^N$ , the process  $(M_t^E)_{t=0}^N$  is a nonnegative supermartingale, and thus,*

$$\Pr \left( \exists t \in [N] : \mu - \hat{\mu}_t(\lambda_1^t) \geq \frac{\sum_{i=1}^t (c/2)^{-2} (X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i) + \log(1/\alpha)}{\sum_{i=1}^t \lambda_i \left( 1 + \frac{i-1}{N-i+1} \right)} \right) \leq \alpha.$$

Consequently,

$$C_t^E := \hat{\mu}_t(\lambda_1^t) \pm \frac{\sum_{i=1}^t (c/2)^{-2} (X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left( 1 + \frac{i-1}{N-i+1} \right)} \quad \text{forms a } (1-\alpha)\text{-CS for } \mu.$$

The proof in Appendix 2.B.3 involves modifying the proof of Theorem 4 in Howard et al. [125] to use our WoR versions of  $\hat{\mu}_t$  and to include predictable values of  $\lambda_t$ .

As before one must choose a  $\lambda$ -sequence to use  $C_t^E$ . We will again consider the case of a real-valued  $\lambda := \lambda_1 = \dots = \lambda_N$  to help guide our intuition on choosing a more complex  $\lambda$ -sequence. Unlike earlier, we cannot optimize the width of  $C_t^E$  in closed-form since  $\psi_E$  is less analytically tractable. Once more, fact (2.10) comes to our rescue: substituting  $\psi_H$  for  $\psi_E$  and optimizing the width yields an expression like (2.7):

$$\lambda^* := \sqrt{\frac{2 \log(2/\alpha)}{\hat{V}_t}}, \quad (2.12)$$

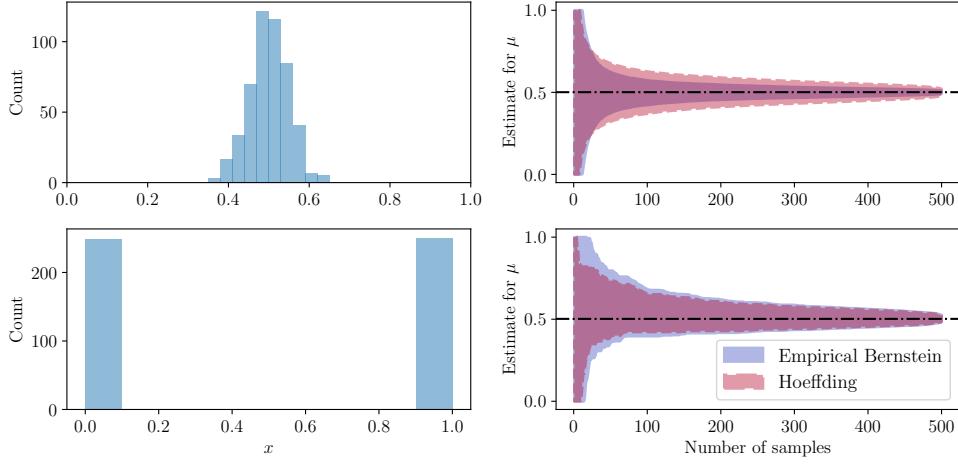


Figure 2.4: Left-most plots show the histogram of the underlying set of numbers  $x_1^N \in [0, 1]^N$ , while right-most plots compare empirical Bernstein- and Hoeffding-type CSs for  $\mu$ . Specifically, the Hoeffding and empirical Bernstein CSs use the  $\lambda$ -sequences in (2.8) and (2.14), respectively. As expected, in low-variance settings (top),  $C_t^E$  is superior, but in a high-variance setting (bottom),  $C_t^H$  has a slight edge.

where  $\hat{V}_t := \sum_{i=1}^t (X_i - \hat{\mu}_{i-1})^2$  is a variance process. However, we cannot use this choice of  $\lambda^*$  since it depends on  $X_1^t$ . Instead, we construct a predictable  $\lambda$ -sequence which mimics  $\lambda^*$  and adapts to the underlying variance as samples are collected. To heuristically optimize the CS for a particular time  $t_0$ , take an estimate  $\hat{\sigma}_{t-1}^2$  of the variance which only depends on  $X_1^{t-1}$ , and set

$$\lambda_t := \sqrt{\frac{2 \log(2/\alpha)}{\hat{\sigma}_{t-1}^2 t_0}} \wedge \frac{1}{2c}. \quad (2.13)$$

Alternatively, to spread the CS width optimization over time as in (2.8), one can use the  $\lambda$ -sequence,

$$\lambda_t := \sqrt{\frac{2 \log(2/\alpha)}{\hat{\sigma}_{t-1}^2 t \log(t+1)}} \wedge \frac{1}{2c}, \quad (2.14)$$

but again, any predictable sequence will suffice.

Similarly to the Hoeffding-type CS, we may instantiate the empirical Bernstein-type CS at a particular time to obtain tight CIs for sampling WoR. However, ensuring that the resulting fixed-time CI is valid when using a data-dependent  $\lambda$ -sequence requires some additional care. Suppose now that  $X_1^n$  is a simple random sample WoR from the finite population,  $x_1^N \in [\ell, u]^N$ . If we randomly permute  $X_1, \dots, X_n$  to obtain the sequence,  $\tilde{X}_1, \dots, \tilde{X}_n$ , we have recovered the observation model of Section 2.1.1, and thus Theorem 2.3.2 applies. We choose a  $\lambda$ -sequence which sequentially estimates the variance, but heuristically optimizes for the sample size  $n$  as

in (2.13). For  $t \in [n]$ , define

$$\tilde{\lambda}_t := \sqrt{\frac{2 \log(2/\alpha)}{n \tilde{\sigma}_{t-1}^2}} \wedge \frac{1}{2c} \quad \text{where} \quad \tilde{\sigma}_t^2 := \frac{c^2/4 + \sum_{i=1}^t (\tilde{X}_i - \tilde{\mu}_i)^2}{t+1} \quad \text{and} \quad \tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \tilde{X}_i. \quad (2.15)$$

Here, an extra  $c^2/4$  was added to  $\tilde{\sigma}_t^2$  so that it is defined at time 0, but this is simply a heuristic and any other choice of  $\tilde{\sigma}_0^2$  will suffice. The resulting CI can be summarized in the following corollary.

**Corollary 2.3.2.** *Let  $X_1^n$  be a simple random sample WoR from the finite population  $x_1^N$  and let  $\tilde{X}_1^n$  be a random permutation of  $X_1^n$ . Let  $\tilde{\lambda}_t$  be a predictable sequence such as the one in (2.15) for each  $t \in [n]$ . Then for any  $n \in [N]$ ,*

$$\hat{\mu}_n(\tilde{\lambda}_1^n) \pm \frac{\sum_{i=1}^n (c/2)^{-2} (\tilde{X}_i - \tilde{\mu}_{i-1})^2 \psi_E(\tilde{\lambda}_i) + \log(2/\alpha)}{\sum_{i=1}^n \tilde{\lambda}_i \left(1 + \frac{i-1}{N-i+1}\right)} \text{ forms a } (1-\alpha) \text{ CI for } \mu.$$

The aforementioned CSs and CIs have a strong relationship with corresponding hypothesis tests. In the following section, we discuss how one can use the techniques developed here to sequentially test hypotheses about finite sets of nonrandom numbers.

## 2.4 Testing hypotheses about finite sets of nonrandom numbers

In classical hypothesis testing, one has access to i.i.d. data from some underlying distribution(s), and one wishes to test some property about them; this includes sequential tests dating back to Wald [271]. However, it is not often appreciated that it is possible to test hypotheses about a finite list of numbers that do not have any distribution attached to them. Recalling the setup of Section 2.1.1, this is the nonstandard setting we find ourselves in. For instance in the same example as Figure 2.1, we may wish to test:

$$H_0 : N_1^* \leq 550 \quad (\text{At most 550 of the balls are green}).$$

If we had access to each ball in advance, then we could accept or reject the null without any type-I or type-II error, but this is tedious, and so we sequentially take samples in a random order to test this hypothesis. The main question then is: *how do we calculate a p-value  $P_t$  that we can track over time, and stop sampling when  $P_t \leq 0.05$ ?*

Luckily, we do not need any new tools for this, and our CSs provide a straightforward answer. Though we left it implicit, each confidence sequence  $C_t$  is really a function of confidence level  $\alpha$ . Consider the family  $\{C_t(q)\}_{q \in (0,1)}$  indexed by  $q$ , which we only instantiated at  $q = \alpha$ . Now, define

$$P_t := \inf\{q : C_t(q) \cap H_0 = \emptyset\}, \quad (2.16)$$

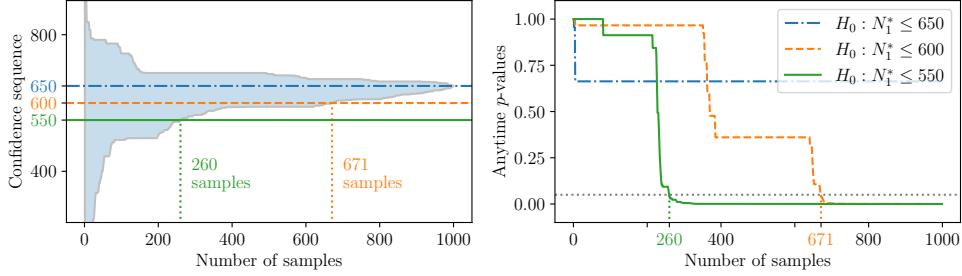


Figure 2.5: The duality between anytime  $p$ -values and CSs for three null hypotheses:  $H_0 : N_1^* \leq D$  for  $D \in \{550, 600, 650\}$ . The first null is rejected at a 5% significance level after 260 samples, exactly when the 95% CS stops intersecting the null set  $[0, 550]$ . However,  $H_0 : N_1^* \leq 650$  is never rejected since 650, the ground truth, is contained in the CS at all times from 0 to 1000.

which is the smallest error level  $q$  at which  $C_t(q)$  just excludes the null set  $H_0$ . This ‘duality’ is familiar in non-sequential settings, and in our case it yields an anytime-valid  $p$ -value [135, 125],

$$\text{Under } H_0, \quad \Pr(\exists t \in [N] : P_t \leq \alpha) \leq \alpha \text{ for any } \alpha \in [0, 1].$$

In words, if the null hypothesis is true, then  $P_t$  will remain above  $\alpha$  through the whole process, with probability  $\geq 1 - \alpha$ . To more clearly bring out the duality to CSs, define the stopping time

$$\tau := \inf\{t \in [N] : P_t \leq \alpha\}, \text{ and we set } \tau = N \text{ if the inf is not achieved.}$$

Then under the null,  $\tau = N$  (we never stop early) with probability  $\geq 1 - \alpha$ . If we do stop early, then  $\tau$  is exactly the time at which  $C_t(\alpha)$  excluded the null set  $H_0$ . The manner in which anytime-valid  $p$ -values and CSs are connected through stopping times is demonstrated in Figure 2.5.

In summary, our CSs directly yield  $p$ -values (2.16) for composite null hypotheses. These  $p$ -values can be tracked, and are valid simultaneously at all times, including at arbitrary stopping times. Aforementioned type-I error probabilities are due to the randomness in the ordering, not in the data.

It is worth noting that our (super)martingales  $(R_t)$ ,  $(M_t^H)$  and  $(M_t^E)$  also immediately yield ‘e-values’ [236] and hence ‘safe tests’ [113], meaning that under nulls of the form in Figure 2.5, they satisfy  $\mathbb{E}M_\tau \leq 1$  for any stopping time  $\tau$ .

## 2.5 Summary

WoR sampling and inference naturally arise in a variety of applications such as finite-population studies and permutation-based statistical methods as outlined in Appendix 2.A. Furthermore, several machine learning tasks involve random samples from finite ‘populations’, such as sampling (a) points for a stochastic gradient method, (b) covariates in a random order for

coordinate descent, (c) columns of a matrix, or (d) edges in a graph.

In order to quantify uncertainty when sequentially sampling WoR from a finite set of objects, this chapter developed three new confidence sequences: one in the discrete setting and two in the continuous setting (Hoeffding, empirical-Bernstein). Their construction was enabled by the development of new technical tools—the prior-posterior-ratio martingale, and two exponential supermartingales—which may be of independent interest. We clarified how these can be tuned (role of ‘prior’ or  $\lambda$ -sequence), and demonstrated their advantages over naive sampling with replacement. Our CSs can be inverted to yield anytime-valid  $p$ -values to sequentially test arbitrary composite hypotheses. Importantly, these CSs can be efficiently updated, continuously monitored, and adaptively stopped without violating their uniform validity, thus merging theoretical rigor with practical flexibility.

## 2.A Four prototypical examples

The following examples are meant to demonstrate situations where we might care about sequentially quantifying uncertainty for parameters of finite populations (see Figure 2.6).

### A. Opinion surveys (discrete categorical)

Imagine you have access to a registry of phone numbers of a group of 1000 people, such as all residents of a neighborhood, voters in a township, or occupants of a university building. You wish to quickly determine the majority opinion on a categorical question, like preference of Biden vs. Trump. You pick names uniformly at random, call and ask. Obviously, you never call the same person twice. When can you confidently stop? In a typical run on a hypothetical ground truth of 650/350, our method stopped after 123 calls (Figure 2.6A).

In the example of opinion surveys, the data are discrete and consist of 650 responses showing preference for Biden and 350 showing preference for Trump (encoded as ones and zeros, respectively). The observed data is thus a random permutation of 650 ones and 350 zeros. The CS used was the PPR CS for the hypergeometric distribution with a uniform ‘working prior’ (i.e.  $a = b = 1$  in the beta-binomial pmf).

### B. Permutation $p$ -values (discrete binary)

Statistical inference is often performed using permutation tests. Formally, the permutation  $p$ -value is defined as  $P_{\text{perm}} := \frac{1}{m!} \sum_{\pi \in S_m} I(T_m \geq T_{\pi(m)})$ , where  $T_m, T_{\pi(m)}$  are the original and permuted test statistics on  $m$  datapoints, and  $S_m$  is the set of all  $m$ -permutations (size  $N = m!$ ).  $P_{\text{perm}}$  is intractable to calculate for large  $m$ , so it is often approximated by randomly sampling  $\pi$  with replacement (often 1000 times, fixed and arbitrary). Instead, our tools allow a user to construct a CS for  $P_{\text{perm}}$  and sequentially sample WoR until the CS is confident about whether  $P_{\text{perm}}$  is below or above (say) 0.05. In one example (small, so we can calculate  $P_{\text{perm}} = 0.04$  to verify accuracy), we stopped after 876 steps (Figure 2.6B).

The permutation test used in this example is a slight modification of the famous ‘Lady Tasting Tea’ experiment [106]. The experiment proceeds as follows.

There are 12 cups of tea with milk, half of which had the tea poured first, and the other half had milk poured first. The tea expert is told that half of the cups are milk-first and the other half are tea-first and is tasked with determining which ones are which. The null hypothesis is that the tea expert has no ability to distinguish between tea-first and milk-first (i.e. their guesses are independent of the order of milk/tea). Suppose they guess 10 out of 12 cups correctly. The statistical question becomes, “what is the probability of guessing 10 or more cups correctly if the expert is guessing randomly?”. This probability is exactly the permutation  $p$ -value that the statistician is interested in.

To calculate this permutation  $p$ -value, we consider the set of all possible random guesses that the tea expert could have made, and compute the fraction of those which identify 10 or more cups correctly. If we randomly sample a sequence of possible guesses from the set of  $\binom{12}{6}$

possible guesses and record whether 10 or more cups are correctly identified, then observations are a random stream of ones and zeros. We then construct a PPR CS with a uniform ‘working prior’ for the number of ones,  $N^+$  in this set to arrive at a CS for the permutation  $p$ -value,  $P := \frac{N^+}{\binom{12}{6}}$ .

### C. Shapley values (bounded real-valued)

First developed in game theory, Shapley values have been recently proposed as a measure of variable or data-point importance for supervised learning. Given a set of players  $\{1, \dots, B\}$  and a reward function  $\nu$ , the Shapley value  $\phi_b$  for player  $b$  can be written as an average of  $B!$  function evaluations, one for each permutation of  $\{1, \dots, B\}$ . As above,  $\phi_b$  is intractable to compute and Monte-Carlo techniques are popular. This real-valued setting requires different CS techniques from the categorical setting. As Figure 2.6C unfolds from left to right (with  $B = 7$ ), it can be stopped adaptively with valid confidence bounds on all  $\{\phi_b\}_{b=1}^B$ . In this example, we consider a simple cost allocation problem. Suppose there are  $n$  people that wish to share transportation to get from point A to their respective destinations, which are all in succession on the same street. Suppose that the cost of going from point A to the  $i^{\text{th}}$  person’s destination costs  $c_i$ , and without loss of generality suppose  $c_1 < c_2 < \dots < c_n$ . In this particular example, we used  $n = 7$  with costs of 1, 10, 40, 80, 130, 175, and 200. The ‘cost’,  $\nu : 2^{[n]} \rightarrow \mathbb{R}$  of a trip is defined in the following natural way,

$$\begin{aligned}\nu(\emptyset) &= 0 \\ \nu(\{i\}) &= c_i \\ \nu(S) &= c_j \text{ where } c_j \geq c_k \text{ for all } k, j \in S\end{aligned}$$

The *Shapley value*,  $\phi_i$  for person  $i$  can be written as,

$$\phi_i = \frac{1}{n!} \sum_{\pi} [\nu(S_{\pi,i} \cup \{i\}) - \nu(S_{\pi,i})] \quad (2.17)$$

where the sum is taken over all permutations  $\pi$  of  $[n]$ , and  $S_{\pi,i}$  is the set of numbers to the left of  $i$  in the permutation  $\pi([n])$ .

Since the Shapley value  $\phi_i$  is an average of  $n!$  numbers, it may be tedious to compute for large  $n$  especially when  $\nu$  cannot be computed quickly. In our case, the summands have a crude upper bound of  $c_n$  and a lower bound of 0 so we can randomly sample WoR from the set of permutations on  $[n]$  to construct the empirical Bernstein CS of Theorem 2.3.2 with the  $\lambda$ -sequence of (2.14). After 1252 permutations, we are able to conclude with high confidence which player has the highest Shapley value.

### D. Tracking interventions (bounded real-valued)

Suppose a state school board is interested in introducing a new program to help students improve their standardized testing skills. Before deploying it to each of their 3000 public schools, the board decides to incrementally introduce the program to randomly selected

schools, measuring standardized test scores before and after its introduction. The board can construct a CS for the overall percentage increase in test scores (which could get worse), and stop the experiment once they are confident about the program's effectiveness. In Figure 2.6D, with effect size 20%, the board can confidently decide to mandate the program statewide after 260 random schools have been trialed, but they may also continue tracking progress and stop later. In this example, we simply generated 3000 observations from a Beta(3, 2) distribution, appropriately scaled to be between -100 and 100 (representing percentage changes in test scores). To construct a CS for the average change in test scores, we used the Hoeffding-type CS optimized for times 10, 100, and 1000. Note that this CS would be tighter if the empirical Bernstein CS were used as the Beta(3, 2) has a relatively small variance.

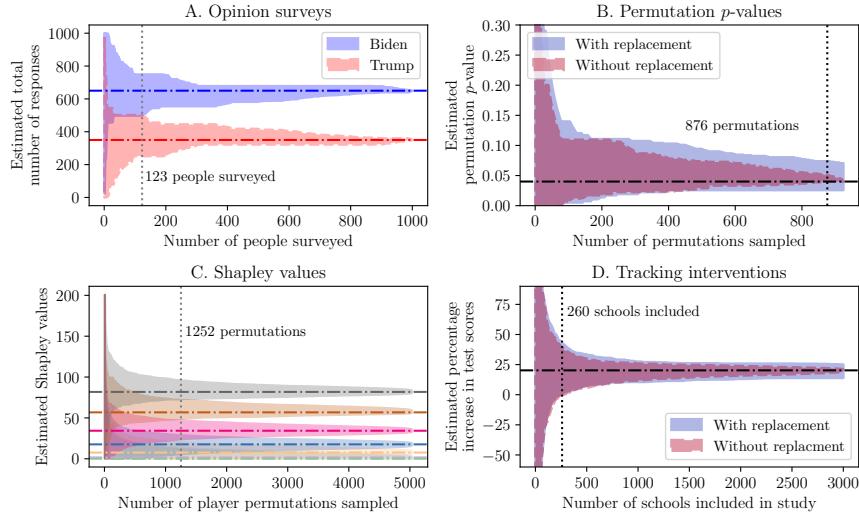


Figure 2.6: Typical simulation runs for the aforementioned examples, with more details in the Supplement. All experiments can be proactively monitored, optionally continued and adaptively stopped.

## 2.B Proofs of the main results

### 2.B.1 Proof of Proposition 2.2.1

The proof is broken into two steps. First, we prove that with respect to the filtration  $(\mathcal{F}_t)_{t=0}^N$  outlined in Section 2.1.1, the prior-posterior ratio (PPR) evaluated at the true  $\theta^* \in \Theta$ ,

$$R_t(\theta^*) := \frac{\pi_0(\theta^*)}{\pi_t(\theta^*)}, \quad (2.18)$$

is a nonnegative martingale with initial value one. Later, we invoke Ville's inequality [264, 124] for nonnegative supermartingales to construct the CS.

**Step 1.** Let  $\pi_0$  be any prior on  $\Theta$  that assigns nonzero mass everywhere. Define the prior-posterior ratio,  $R_t(\theta)$  as in (2.18). Writing the conditional expectation of  $R_{t+1}(\theta^*)$  given  $X_1^t$  for any  $t \in \{1, \dots, N\}$  in its integral form,

$$\begin{aligned}
\mathbb{E}(R_{t+1}(\theta^*) \mid X_1^t) &= \int_{\mathcal{X}_{t+1}} \frac{\pi_0(\theta^*)}{\pi_{t+1}(\theta^*)} p_{\theta^*}(x_{t+1} \mid X_1^t) dx_{t+1} \\
&= \int_{\mathcal{X}_{t+1}} \frac{\pi_0(\theta^*) \int_{\Theta} p_{\eta}(X_1^t, x_{t+1}) \pi_0(\eta) d\eta}{p_{\theta^*}(X_1^t, x_{t+1}) \pi_0(\theta^*)} p_{\theta^*}(x_{t+1} \mid X_1^t) dx_{t+1} && \text{(Bayes' rule)} \\
&= \int_{\mathcal{X}_{t+1}} \frac{\pi_0(\theta^*) \int_{\Theta} p_{\eta}(X_1^t, x_{t+1}) \pi_0(\eta) d\eta}{p_{\theta^*}(X_1^t) \pi_0(\theta^*)} dx_{t+1} && \text{(Bayes' rule again)} \\
&= \int_{\mathcal{X}_{t+1}} \frac{\pi_0(\theta^*) \int_{\Theta} p_{\eta}(X_1^t, x_{t+1}) \pi_0(\eta) d\eta}{\pi_t(\theta^*) \int_{\Theta} p_{\lambda}(X_1^t) \pi_0(\lambda) d\lambda} dx_{t+1} && \text{(Bayes' rule again)} \\
&= \frac{\pi_0(\theta^*)}{\pi_t(\theta^*)} \int_{\mathcal{X}_{t+1}} \frac{\int_{\Theta} p_{\eta}(X_1^t, x_{t+1}) \pi_0(\eta) d\eta}{\int_{\Theta} p_{\lambda}(X_1^t) \pi_0(\lambda) d\lambda} dx_{t+1} \\
&= \frac{\pi_0(\theta^*)}{\pi_t(\theta^*)} \frac{\int_{\Theta} \int_{\mathcal{X}_{t+1}} p_{\eta}(X_1^t, x_{t+1}) dx_{t+1} \pi_0(\eta) d\eta}{\int_{\Theta} p_{\lambda}(X_1^t) \pi_0(\lambda) d\lambda} && \text{(Fubini's theorem)} \\
&= \frac{\pi_0(\theta^*)}{\pi_t(\theta^*)} \frac{\int_{\Theta} p_{\eta}(X_1^t) \pi_0(\eta) d\eta}{\int_{\Theta} p_{\lambda}(X_1^t) \pi_0(\lambda) d\lambda} = R_t(\theta^*).
\end{aligned}$$

Furthermore, for the case when  $t = 0$ ,

$$\begin{aligned}
\mathbb{E}(R_1(\theta^*)) &= \int_{\mathcal{X}_1} \frac{\pi_0(\theta^*) \int_{\Theta} p_{\eta}(X_1) \pi_0(\eta) d\eta}{p_{\theta^*}(X_1) \pi_0(\theta^*)} p_{\theta^*}(X_1) dx_1 \\
&= \frac{\pi_0(\theta^*)}{\pi_0(\theta^*)} \int_{\mathcal{X}_1} \int_{\Theta} p_{\eta}(X_1) \pi_0(\eta) d\eta dx_1 && \text{(Bayes' rule)} \\
&= \frac{\pi_0(\theta^*)}{\pi_0(\theta^*)} \int_{\Theta} \int_{\mathcal{X}_1} p_{\eta}(X_1) dx_1 \pi_0(\eta) d\eta && \text{(Fubini's theorem)} \\
&= \frac{\pi_0(\theta^*)}{\pi_0(\theta^*)} \int_{\Theta} \pi_0(\eta) d\eta = \frac{\pi_0(\theta^*)}{\pi_0(\theta^*)} = R_0 = 1.
\end{aligned}$$

Establishing that  $R_t(\theta^*)$  is a nonnegative martingale with initial value one completes the first step.

**Step 2.** Ville's inequality for nonnegative supermartingales [264, 124] implies that for any  $\beta > 0$ ,

$$\Pr(\exists t \in [N] : R_t(\theta^*) \geq \beta) \leq \frac{\mathbb{E}(R_0(\theta^*))}{\beta}.$$

In particular, for a threshold  $\alpha \in (0, 1)$ ,

$$\Pr\left(\exists t \in [N] : R_t(\theta^*) \geq 1/\alpha\right) \leq \alpha. \quad (2.19)$$

Define the sequence of sets for  $t \in [N]$ ,

$$C_t := \{\theta : R_t(\theta) \leq 1/\alpha\}.$$

As a consequence of (2.19), we have that

$$\Pr(\forall t \in [N], \theta^* \in C_t) \geq 1 - \alpha,$$

as desired, which completes the proof.

## 2.B.2 Proof of Theorem 2.3.1

*Proof.* Similar to the proof of Proposition 2.2.1, we proceed in two steps. First, we show that the exponential Hoeffding-type process (2.5) is a nonnegative supermartingale with respect to the filtration outlined in Section 2.1.1. We then apply Ville's inequality to this supermartingale and ultimately obtain the bound stated in the theorem.

We prove the bound for  $[0, 1]$ -bounded random variables but the general result holds by taking any  $[\ell, u]$ -bounded random variable,  $X_i$  and applying the transformation,  $X_i \mapsto (X_i - \ell)/(u - \ell)$

**Step 1.** Let  $(\mathcal{F}_t)_{t=0}^N$  be the filtration defined in Section 2.1.1. Furthermore, let  $\lambda_t = \lambda_t(X_1, \dots, X_{t-1})$  be a sequence of  $\mathcal{F}_{t-1}$ -measurable random variables. Consider the exponential Hoeffding-type process  $(M_t^H)_{t=0}^N$  with a 'predictable mixture',

$$M_t^H := \exp \left\{ \sum_{i=1}^t \left[ \lambda_i (X_i - \mu + Z_{i-1}^*) - \frac{\lambda_i^2}{8} \right] \right\} \equiv \prod_{i=1}^t \exp \left\{ \lambda_i (X_i - \mu + Z_{i-1}^*) - \frac{\lambda_i^2}{8} \right\}$$

where  $Z_i^* = \frac{1}{N-i} \sum_{j=1}^i (X_j - \mu)$  and  $M_0^H = 0$  by convention. Writing the conditional expectation of this process for any  $t \geq 1$ ,

$$\begin{aligned} \mathbb{E}(M_{t+1}^H \mid \mathcal{F}_t) &= \mathbb{E} \left( \prod_{i=1}^{t+1} \exp \left\{ \lambda_i (X_i - \mu + Z_{i-1}^*) - \frac{\lambda_i^2}{8} \right\} \mid \mathcal{F}_t \right) \\ &= M_t^H \cdot \mathbb{E} \left( \exp \left\{ \lambda_{t+1} (X_{t+1} - \mu + Z_t^*) - \frac{\lambda_{t+1}^2}{8} \right\} \mid \mathcal{F}_t \right). \end{aligned}$$

Using the fact that  $\mathbb{E}(X_{t+1} - \mu + Z_t^* \mid \mathcal{F}_t) = 0$ , the fact that  $X_{t+1} \in [0, 1]$ , and that  $\lambda_{t+1}$  is  $\mathcal{F}_t$ -measurable, we have by sub-Gaussianity of bounded random variables,

$$\mathbb{E} \left( \exp \{ \lambda_{t+1} (X_{t+1} - \mu + Z_t^*) \} \mid \mathcal{F}_t \right) \leq \exp \left\{ \frac{\lambda_{t+1}^2}{8} \right\}$$

and thus  $\mathbb{E}(M_{t+1}^H \mid \mathcal{F}_t) \leq M_t^H$ . Therefore, with respect to the filtration  $(\mathcal{F}_t)_{t=0}^N$ , we have that  $M_t^H$  is a nonnegative supermartingale.

**Step 2.** Now that we have shown that  $M_t^H$  is a nonnegative supermartingale, we may apply Ville's inequality to obtain,

$$\Pr \left( \exists t \in [N] : M_t^H \geq \frac{1}{\alpha} \right) \leq \alpha.$$

In particular, with probability at least  $(1 - \alpha)$ , we have that for all  $t \in [N]$ ,  $M_t^H < \frac{1}{\alpha}$ .

**Step 3.** ‘Inverting’ the above statement and solving for  $\hat{\mu}_t(\lambda_1^t) - \mu$ , we get that with probability at least  $(1 - \alpha)$ , for all  $t \in [N]$ ,

$$\hat{\mu}_t(\lambda_1^t) - \mu < \frac{\sum_{i=1}^t \lambda_i^2 / 8 + \log(1/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)}.$$

Applying all of the aforementioned logic to  $-X_1, \dots, -X_t$  and  $-\mu$ , and taking a union bound, we arrive at the desired result,

$$\Pr \left( \exists t \in [N] : |\hat{\mu}_t(\lambda_1^t) - \mu| \geq \frac{\sum_{i=1}^t \lambda_i^2 / 8 + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)} \right) \leq \alpha,$$

which completes the proof. □

**Remark:  $\hat{\mu}_t$  is unconditionally unbiased.** Recalling the advantage term  $A_t := \sum_{i=1}^t \frac{i-1}{N-i+1}$ , a short calculation shows that  $\hat{\mu}_t$  (2.2) has conditional expectation equaling a convex combination of  $\hat{\mu}_t, \mu$ :

$$\mathbb{E}[\hat{\mu}_{t+1}|X_1^t] = \frac{1 + A_{t+1} - A_t}{t + 1 + A_{t+1}} \mu + \frac{t + A_t}{t + 1 + A_{t+1}} \hat{\mu}_t.$$

Multiplying both sides by  $t + 1 + A_{t+1}$ , we can write it in a recursive, telescoping form:

$$\mathbb{E}[(t + 1 + A_{t+1})\hat{\mu}_{t+1}|X_1^t] = \mu + (A_{t+1} - A_t)\mu + (t + A_t)\hat{\mu}_t.$$

Taking expectation with respect to  $X_t|X_1^{t-1}$ , and using the above equation to evaluate the last term,

$$\mathbb{E}[(t + 1 + A_{t+1})\hat{\mu}_{t+1}|X_1^{t-1}] = 2\mu + (A_{t+1} - A_{t-1})\mu + (t - 1 + A_{t-1})\hat{\mu}_{t-1}.$$

Unrolling this process out, we see that  $\mathbb{E}[(t + 1 + A_{t+1})\hat{\mu}_{t+1}] = (t + 1)\mu + (A_{t+1} - A_0)\mu$ . Since  $A_0 \equiv 0$ , we conclude that  $\hat{\mu}_{t+1}$  is an unconditionally unbiased estimator of  $\mu$ .

Interestingly, the without-replacement mean estimator is not necessarily ‘consistent’ (in

the sense of recovering  $\mu$  after all  $N$  samples are drawn). However, the concept of consistency is subtle for finite populations as there is no longer any uncertainty after all samples are drawn. In any case, the without-replacement mean estimator was not introduced to replace the usual sample mean estimator in all without-replacement settings, but was simply the quantity that resulted from attempting to develop exponential supermartingales within this sample scheme.

### 2.B.3 Proof of Theorem 2.3.2

*Proof.* Much like the proof of Theorem 2.3.1, the proof proceeds in three steps: (1) showing that an exponential empirical Bernstein-type process is a supermartingale, (2) applying Ville's inequality, and (3) inverting the process and taking a union bound. Again, we prove the result for  $[0, 1]$ -bounded random variables since for an  $[\ell, u]$ -bounded random variable  $X_i$ , one can make the transformation  $X_i \mapsto (X_i - \ell)/(u - \ell)$

**Step 1.** Let  $(\mathcal{F}_t)_{t=0}^N$  be the filtration defined in Section 2.1.1. Let  $\lambda_t \equiv \lambda_t(X_1, \dots, X_{t-1})$  be a sequence of  $\mathcal{F}_{t-1}$ -measurable random variables. Consider the exponential empirical Bernstein-type process,  $(M_t^E)_{t=0}^N$  with a ‘predictable mixture’,

$$\begin{aligned} M_t^E &:= \exp \left\{ \sum_{i=1}^t [\lambda_i (X_i - \mu + Z_{i-1}^*) - 4(X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i)] \right\} \\ &\equiv \prod_{i=1}^t \exp \{ \lambda_i (X_i - \mu + Z_{i-1}^*) - 4(X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i) \} \end{aligned}$$

where  $M_0^E := 0$ . Writing out the conditional expectation of  $M_{t+1}^E$  given  $\mathcal{F}_t$  for  $t \in [N]$ ,

$$\mathbb{E}(M_{t+1}^E | \mathcal{F}_t) = M_t^E \cdot \mathbb{E} \left( \exp \left\{ \lambda_{t+1} (X_{t+1} - \mu + Z_t^*) - 4\psi_E(\lambda_{t+1}) (X_{t+1} - \hat{\mu}_t)^2 \right\} \middle| \mathcal{F}_t \right).$$

Therefore, it suffices to show that for any  $t \in [N]$ ,

$$\mathbb{E} \left( \exp \left\{ \lambda_{t+1} (X_{t+1} - \mu + Z_t^*) - 4\psi_E(\lambda_{t+1}) (X_{t+1} - \hat{\mu}_t)^2 \right\} \middle| \mathcal{F}_t \right) \leq 1.$$

For succinctness, denote

$$Y_{t+1} := X_{t+1} + \frac{1}{N-t} \sum_{j=1}^t X_j - \frac{N}{N-t} \mu \quad \text{and} \quad \delta_t := \hat{\mu}_t + \frac{1}{N-t} \sum_{j=1}^t X_j - \frac{N}{N-t} \mu.$$

Note that  $Y_{t+1}$  is conditionally mean zero. It then suffices to prove that for any  $(0, 1)$ -bounded,  $\mathcal{F}_t$ - measurable  $\lambda_{t+1} \equiv \lambda_{t+1}(X_1, \dots, X_t)$ ,

$$\mathbb{E} \left[ \exp \left\{ \lambda_{t+1} Y_{t+1} - 4(Y_{t+1} - \delta_t)^2 \psi_E(\lambda_{t+1}) \right\} \middle| \mathcal{F}_t \right] \leq 1.$$

Indeed, in the proof of Proposition 4.1 in Fan et al. [102],  $\exp\{\xi\lambda - 4\xi^2\psi_E(\lambda)\} \leq 1 + \xi\lambda$  for any  $\lambda \in [0, 1)$  and  $\xi \geq -1$ . Setting  $\xi := Y_{t+1} - \delta_t = X_{t+1} - \hat{\mu}_t$ ,

$$\begin{aligned} & \mathbb{E}\left[\exp\left\{\lambda_{t+1}Y_{t+1} - 4(Y_{t+1} - \delta_t)^2\psi_E(\lambda_{t+1})\right\} \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\exp\left\{\lambda_{t+1}(Y_{t+1} - \delta_t) - 4(Y_{t+1} - \delta_t)^2\psi_E(\lambda_{t+1})\right\} \mid \mathcal{F}_t\right] \exp(\lambda_{t+1}\delta_t) \\ &\leq \mathbb{E}\left[1 + (Y_{t+1} - \delta_t)\lambda_{t+1} \mid \mathcal{F}_t\right] \exp(\lambda_{t+1}\delta_t) \stackrel{(i)}{=} \mathbb{E}\left[1 - \delta_t\lambda_{t+1} \mid \mathcal{F}_t\right] \exp(\lambda_{t+1}\delta_t) \stackrel{(ii)}{\leq} 1, \end{aligned}$$

where equality (i) follows from the fact that  $Y_{t+1}$  is conditionally mean zero as mentioned earlier, and inequality (ii) follows from the inequality  $1 - x \leq \exp(-x)$  for all  $x \in \mathbb{R}$ .

**Step 2.** Now that we have established that  $M_t^E$  is a nonnegative supermartingale, we apply Ville's inequality to obtain,

$$\Pr\left(\exists t \in [N] : M_t^E \geq \frac{1}{\alpha}\right) \leq \alpha.$$

**Step 3.** Solving for  $\hat{\mu}_t - \mu$  in the inequality in the above probability statement, we get that

$$\Pr\left(\exists t \in [N] : \hat{\mu}_t - \mu \geq \frac{\sum_{i=1}^t 4\psi_E(\lambda_i)(X_i - \hat{\mu}_{i-1})^2 + \log(1/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)}\right) \leq \alpha.$$

Applying the same logic to  $-X_1, \dots, -X_t$  and  $-\mu$ , and taking a union bound, we arrive at the desired result,

$$\Pr\left(\exists t \in [N] : |\hat{\mu}_t - \mu| \geq \frac{\sum_{i=1}^t 4\psi_E(\lambda_i)(X_i - \hat{\mu}_{i-1})^2 + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)}\right) \leq \alpha.$$

□

## 2.C Sampling multivariate binary variables WoR

The prior-posterior martingale from Section 2.2.2 extends naturally to the multivariate case as follows. Suppose we have  $N$  objects, each belonging to one of  $K \geq 2$  categories, and there are  $N_1^*, \dots, N_K^*$  objects from each category, respectively. Let  $c$  denote the category of a randomly sampled object, and let

$$\mathbf{X} := \begin{pmatrix} \mathbb{1}(c=1) & \mathbb{1}(c=2) & \cdots & \mathbb{1}(c=K) \end{pmatrix}.$$

Then  $\mathbf{X}$  is said to follow a multivariate hypergeometric distribution with parameters  $N$ ,  $(N_1^*, \dots, N_K^*)$ , and  $n = 1$  and has probability mass function,

$$\Pr(\mathbf{X} = x) = \frac{\prod_{k=1}^K \binom{N_k^*}{x_k}}{\binom{N}{n}}.$$

Note that  $\sum_{k=1}^K x_k = 1$  and  $x_k \in \{0, 1\}$  for each  $k \in \{1, \dots, K\}$ . More generally, if  $n \geq 2$  objects are sampled WoR, then  $\mathbf{X}$  would have the same probability mass function with  $x_1, \dots, x_K \in \{1, \dots, n\}$  such that  $\sum_{k=1}^K x_k = n$ . As in Section 2.2.2, we will consider the case where  $n = 1$  for notational simplicity.

Let us now view this random variable and the fixed multivariate parameter  $\mathbf{N}^* := (N_1^*, \dots, N_K^*)$  from the Bayesian perspective as in Section 2.2.2 by treating  $\mathbf{N}^*$  as a random variable which we denote by  $\tilde{\mathbf{N}}^*$  to avoid confusion. Suppose that

$$\mathbf{X}_t | (\tilde{\mathbf{N}}^*, \mathbf{X}_1, \dots, \mathbf{X}_{t-1}) \sim \text{MultHyperGeo}\left(N - (t-1), \tilde{\mathbf{N}}^* - \sum_{i=1}^{t-1} \mathbf{X}_i, 1\right), \quad \text{and}$$

$$\tilde{\mathbf{N}}^* \sim \text{DirMult}(N, \mathbf{a})$$

for some  $\mathbf{a} := (a_1, \dots, a_K)$  with  $a_k > 0$  for each  $k \in \{1, \dots, K\}$ . Then for any  $t \in \{1, 2, \dots, N\}$ ,

$$\tilde{\mathbf{N}}^* - \sum_{i=1}^t \mathbf{X}_i | (\mathbf{X}_1, \dots, \mathbf{X}_t) \sim \text{DirMult}\left(N - t, \mathbf{a} + \sum_{i=1}^t \mathbf{X}_i\right).$$

With these prior and posterior distributions, we're ready to invoke Proposition 2.2.1 to obtain a sequence of confidence sets for  $\mathbf{N}^*$ .

**Theorem 2.C.1** (Confidence sequences for multivariate hypergeometric parameters). *Suppose that*

$$\mathbf{X}_t | (\mathbf{X}_1, \dots, \mathbf{X}_{t-1}) \sim \text{MultHyperGeo}\left(N - (t-1), \mathbf{N}^* - \sum_{i=1}^{t-1} \mathbf{X}_i, 1\right).$$

*Let  $\pi_0$  and  $\pi_t$  be the Dirichlet-multinomial prior with positive parameters  $\mathbf{a} = (a_1, \dots, a_K)$  and corresponding posterior,  $\pi_t$ , respectively. Then the sequence of sets  $(C_t)_{t=0}^N$  defined by*

$$C_t := \left\{ \mathbf{n} \in \{0, \dots, N\}^K : \sum_{k=1}^K \mathbf{n}_k = N \text{ and } \frac{\pi_0(\mathbf{n})}{\pi_t(\mathbf{n})} < \frac{1}{\alpha} \right\}$$

*is a  $(1 - \alpha)$ -CS for  $\mathbf{N}^*$ . Furthermore, the running intersection,  $\bigcap_{s \leq t} C_t$  is a  $(1 - \alpha)$ -CS for  $\mathbf{N}^*$ .*

*Proof.* This is a direct consequence of Theorem 2.2.1 applied to the multivariate hypergeometric

distribution with a Dirichlet-multinomial prior.  $\square$

## 2.D Coupling the ‘prior’ with the stopping rule to improve power

Somewhat at odds with their intended use-case, working ‘priors’ need not always be chosen to reflect the user’s prior information. When approximating  $p$ -values for permutation tests, for example, it is of primary interest to conclude whether  $P_{\text{perm}}$  is above or below some prespecified  $\alpha_{\text{perm}} \in (0, 1)$  with high confidence as quickly as possible. As discussed in Theorem 2.2.1, the CS for  $P_{\text{perm}}$  will shrink to a single point regardless of the prior, so if  $P_{\text{perm}}$  is much larger or much smaller than  $\alpha_{\text{perm}}$ , we expect to discover the decision rule, “reject” versus “do not reject” rather quickly. It is when  $P_{\text{perm}}$  is very close to  $\alpha_{\text{perm}}$  that the user desires sharper confidence intervals, so that they can make decisions sooner (see Figure 2.7). In this case, they simply need to place more mass near the decision boundary, with a necessary tradeoff between the sharpness of confidence sets near  $\alpha_{\text{perm}}$  and the size of the neighborhood around  $\alpha_{\text{perm}}$  for which this sharpness is realized.

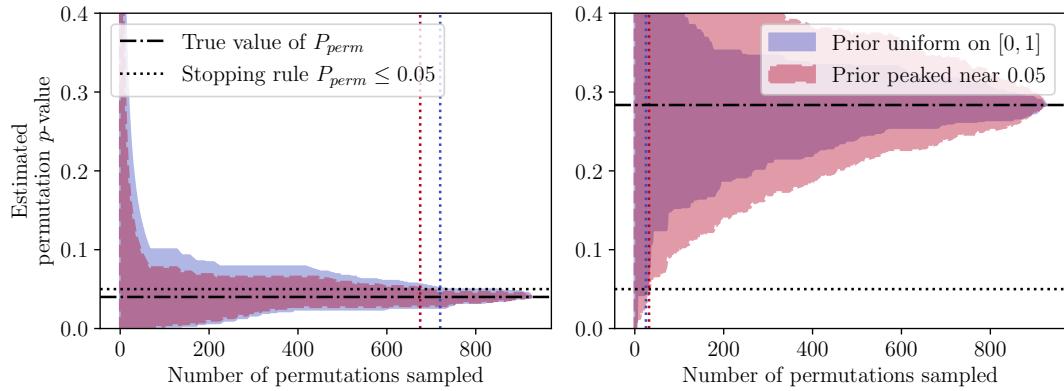


Figure 2.7: Comparing priors for Example B: using a uniform prior versus a prior peaked near 0.05. When the decision rule is to stop whenever the CS is entirely on one side of 0.05, coupling the prior to the decision rule leads to earlier stopping.

## 2.E Choosing a $\lambda$ -sequence for Hoeffding and empirical Bernstein CSs

Recall the Hoeffding-type CS of Theorem 2.3.1,

$$C_t^H := \hat{\mu}_t(\lambda_1^t) \pm \underbrace{\frac{\sum_{i=1}^t \psi_H(\lambda_i) + \log(2/\alpha)}{\sum_i \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)}}_{\text{width } W_t}$$

In Section 2.3, we presented the  $\lambda$ -sequence,

$$\lambda_t := \sqrt{\frac{8 \log(2/\alpha)}{t \log(t+1)(u-\ell)^2}} \wedge \frac{1}{u-\ell}. \quad (2.20)$$

This is visually similar to the single value of  $\lambda \in \mathbb{R}$ ,

$$\lambda := \sqrt{\frac{8 \log(2/\alpha)}{t_0(u-\ell)^2}}$$

which optimizes the bound for time  $t_0$ . Two natural questions arise: (1) where did the extra  $\log(t)$  in (2.20) come from, and (2) why this particular  $\lambda$ -sequence and not others? The answers to these questions are based on some heuristics derived in Chapter 3 in the with-replacement setting. To make matters simpler, ignore the  $\left(1 + \frac{i-1}{N-i+1}\right)$  term in the CS and consider the scaling of the width  $W_t$ ,

$$W_t \asymp \frac{\sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i} \asymp \frac{\sum_{i=1}^t \lambda_i^2}{\sum_{i=1}^t \lambda_i}.$$

When the method of mixtures is used to obtain CSs in the with-replacement setting, their widths often follow a  $\sqrt{\log t/t}$  rate [125]. Following the approximations in Table 2.1, we may opt to pick a sequence  $(\lambda_i)_{i=1}^\infty$  which scales like  $1/\sqrt{i \log i}$  to obtain a width  $W_t \asymp \sqrt{\log t/t}$ . In particular, scaling  $\lambda_i$  as  $1/\sqrt{i \log i}$  is simply an effort to obtain CSs with reasonable widths. The same arguments combined with (2.10) can be applied to the empirical Bernstein CS to obtain (2.14).

Furthermore, we truncate the  $\lambda$ -sequence in 2.20 to prevent the CS width from being dominated by large  $\lambda_t$  at small  $t$ . It is important to keep in mind that *any* sequence would have yielded a valid CS. The choice presented here was derived based on a heuristic argument and kept because of its reasonable empirical performance.

## 2.F Comparing our CSs to those implied by Bardenet & Maillard

Bardenet & Maillard [19, Theorem 2.4] provide the following two time-uniform Hoeffding-Serfling inequalities when sampling bounded real numbers WoR from a finite population. For any  $n \in [N]$ ,

$$\Pr \left( \exists t \in \{1, \dots, n\} : \frac{1}{N-t} \sum_{i=1}^t (X_i - \mu) \geq \frac{n\epsilon}{N-n} \right) \leq \exp \left\{ -\frac{2n\epsilon^2}{(1-(n-1)/N)(u-\ell)^2} \right\} \quad \text{and}$$

$$\Pr \left( \exists t \in \{n, \dots, N-1\} : \frac{1}{t} \sum_{i=1}^t (X_i - \mu) \geq \epsilon \right) \leq \exp \left\{ -\frac{2n\epsilon^2}{(1-n/N)(1+1/n)(u-\ell)^2} \right\}.$$

Sequence $(\lambda_i)_{i=1}^\infty$	$\sum_{i=1}^t \lambda_i$	$\sum_{i=1}^t \lambda_i^2$	Width $W_t$
$\asymp 1/i$	$\asymp \log t$	$\asymp 1$	$1/\log t$
$\asymp \sqrt{\log i/i}$	$\asymp \sqrt{t \log t}$	$\asymp \log^2 t$	$\asymp \log^{3/2} t / \sqrt{t}$
$\asymp 1/\sqrt{i}$	$\asymp \sqrt{t}$	$\asymp \log t$	$\asymp \log t / \sqrt{t}$
$\asymp 1/\sqrt{i \log i}$	$\asymp \sqrt{t/\log t}$	$\asymp \log \log t$	$\asymp \sqrt{\log t/t}$
$\asymp 1/\sqrt{i \log i \log \log i}$	$\asymp \sqrt{t/\log t}$	$\asymp \log \log \log t$	$\asymp \sqrt{\log t/t}$

Table 2.1: Above, we think of  $\log x$  as  $1 \vee \log(1 \vee x)$  to avoid trivialities. The claimed rates are easily checked by approximating the sums as integrals, and taking derivatives. For example,  $\frac{d}{dx} \log \log x = 1/x \log x$ , so the sum of  $\sum_{i \leq t} 1/i \log i \asymp \log \log t$ . It is worth remarking that for  $t = 10^{80}$ , the number of atoms in the universe,  $\log \log t \approx 5.2$ , which is why we treat  $\log \log t$  as a constant when expressing the rate for  $W_t$ . The iterated logarithm pattern in the last two lines of the table can be continued indefinitely.

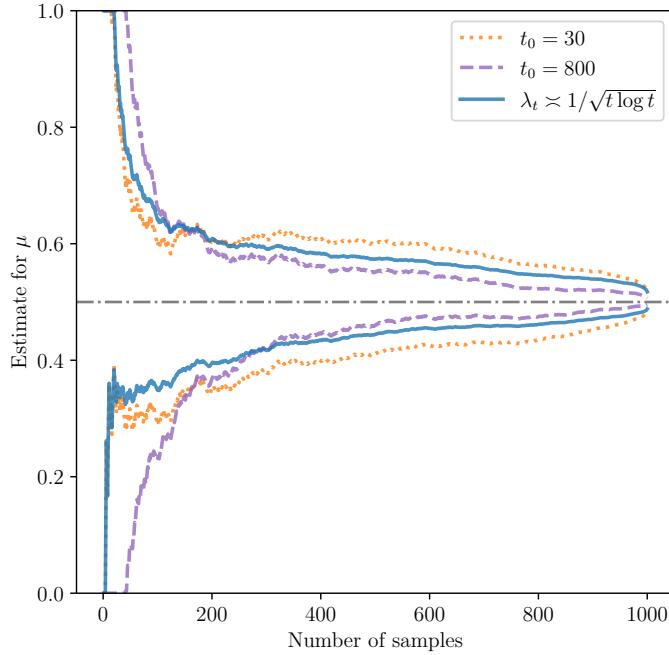


Figure 2.8: Hoeffding CSs based on fixed  $\lambda$  values optimized for times 30 and 800, respectively alongside the CS based on the  $\lambda$ -sequence in (2.20). Notice that no CS uniformly dominates the others, but that the sequence in (2.20) acts as a middle ground between the other two.

Inverting these inequalities and taking a union bound to get two-sided inequalities, we have

$$\frac{1}{t} \sum_{i=1}^t X_i \pm \frac{n(N-t)}{t(N-n)} \sqrt{\frac{\log(4/\alpha)(1-(n-1)/N)(u-\ell)^2}{2n}} \quad \text{when } t \leq n \quad (2.21)$$

$$\frac{1}{t} \sum_{i=1}^t X_i \pm \sqrt{\frac{\log(4/\alpha)(1-n/N)(1+1/n)(u-\ell)^2}{2n}} \quad \text{when } t \geq n \quad (2.22)$$

is a  $(1 - \alpha)$  CS for  $\mu$ . We term the CS defined by (2.21) and (2.22) as the Bardenet-Maillard CS for simplicity.

A comparison of the aforementioned CS to our Hoeffding-type CS is displayed in Figure 2.9, where we see that our bound is roughly as tight as the Bardenet-Maillard CS at the time of optimization, while our bounds are (much) tighter everywhere else. This phenomenon was observed and studied in the with-replacement setting, attributing the benefits of confidence bounds like our Hoeffding CS to an underlying ‘line-crossing’ inequality being uniformly tighter than an underlying Freedman-type inequality. For more information on the with-replacement analogy, we direct the reader to the pair of papers by Howard et al. [125, 124]. Returning back to the WoR setting, we remark that (2.21) uses the standard sample mean, but we use a more sophisticated sample mean (2.2).

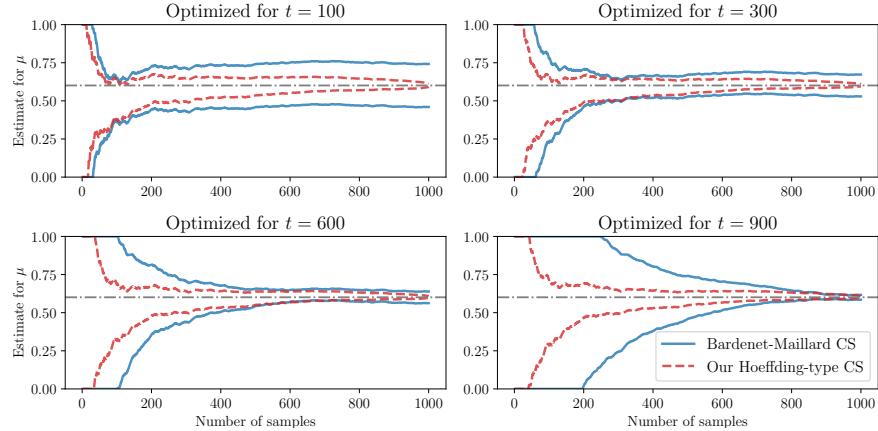


Figure 2.9: A comparison of our Hoeffding-type CS against the Hoeffding-Serfling CS of Bardenet & Maillard [19]. Our Hoeffding CS appears to be as tight as the Hoeffding-Serfling bound at the time of optimization, but tighter at all other times.

## 2.G Time-uniform versus fixed-time bounds

A natural question to ask is, ‘how much does one sacrifice by using a time-uniform CS instead of a fixed-time confidence interval?’ The answer to this question will depend largely on the type of bound used, the underlying finite population, and other factors. However, in the case of sampling binary numbers from a finite population, it seems that the answer is ‘not much’. In Figure 2.10, we display the fixed-time Hoeffding confidence interval of Corollary 2.3.1 alongside its time-uniform counterpart from Theorem 2.3.1 and the prior-posterior ratio CS from Theorem 2.2.1. In terms of the width of confidence bounds, we find that not much is lost by using the two aforementioned CSs over the fixed-time Hoeffding confidence interval. For this small price, the user is awarded the flexibility that comes with using CSs such as properties (a), (b), and (c) described in the Introduction.

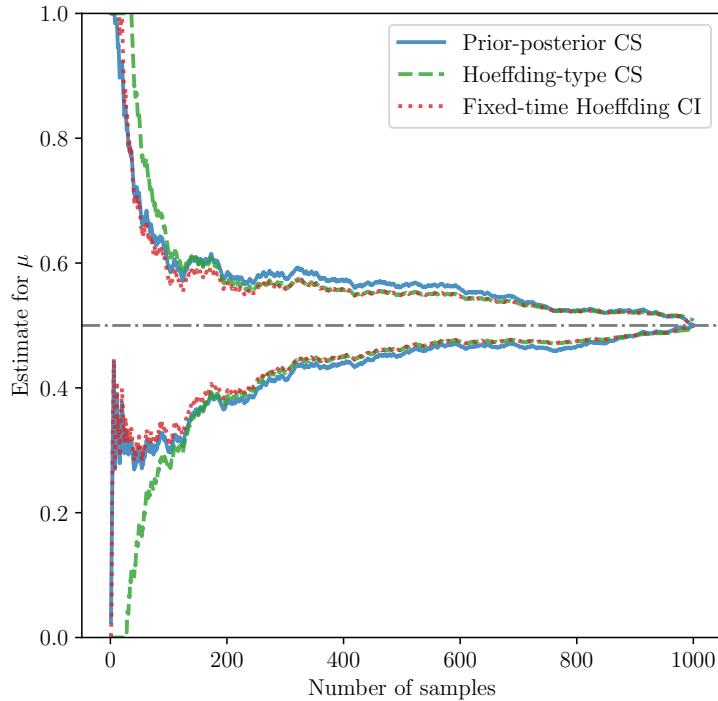


Figure 2.10: Comparing fixed-time and time-uniform confidence bounds for sampling binary numbers from a population of size 1000, consisting of 500 ones and 500 zeros. The dotted red line shows the fixed-time Hoeffding bound of Corollary 2.3.1, while the dashed green and solid blue lines refer to the time-uniform Hoeffding-type CS and the prior-posterior ratio CS, respectively. Notice that the increase in confidence bound width that results from using a time-uniform bound is relatively minor.

## 2.H Computational considerations

When using the CSs of Theorems 2.2.1, 2.3.1, and 2.3.2 in practice, it is important to keep in mind the computational costs associated with each method. For fixed values of  $\lambda$ , updating the Hoeffding and empirical Bernstein CSs at each time  $t$  takes constant time and constant memory, since all calculations involve cumulative sums (or averages). Furthermore, optimal values of  $\lambda$  can be computed as in (2.7) for Hoeffding-type bounds and approximated as in (2.12) for empirical Bernstein-type bounds, all in constant time. On the other hand, the prior-posterior ratio (PPR) CS of Theorem 2.2.1 is the more computationally expensive method among those presented, but can still be computed quickly for many problems. In order to find the CS,

$$C_t := \left\{ n^+ \in [N] : \frac{\pi_0(n^+)}{\pi_t(n^+)} < \frac{1}{\alpha} \right\},$$

one must find all values in  $\{0, \dots, N\}$  which, when provided as an input to  $\frac{\pi_0(\cdot)}{\pi_t(\cdot)}$  are less than  $1/\alpha$ .

Therefore, computing the entire CS takes  $O(PN^2)$  time where  $P$  is the time required to compute  $\pi_0(n)/\pi_t(n)$ . In all of the PPR CSs presented in this chapter, we used computationally tractable conjugate priors, so  $P = 1$ . We believe more sophisticated root-finding methods can be employed to arrive at a time of  $O(N \log(PN))$ , but these methods are reasonably fast in our experience. Moreover, the PPR CS can be computed on a subset of  $[N]$  if needed, and is parallelizable.

For reference, we provide average computation times in Table 2.2. All calculations were measured using Python’s default `time` package and were performed in Python 3.8.3 using the `numpy` and `scipy` packages on a quad-core CPU with 8 threads at 1.8GHz each. However, no parallel processing was performed aside from the default multithreading provided by Python.

	Time in seconds (std. dev.)
Hoeffding	$2.13 \times 10^{-4}$ ( $2.88 \times 10^{-5}$ )
Empirical Bernstein	$2.35 \times 10^{-4}$ ( $3.24 \times 10^{-5}$ )
Prior-posterior ratio	0.306 (0.0115)

Table 2.2: Average time taken to compute the various CSs for  $N = 1000$  discrete observations with equal numbers of ones and zeros, with standard deviations for 100 repeated experiments.

## 2.I Simple experiments for computing miscoverage rates

Typically, in nonparametric testing, there is no ‘uniformly most powerful’ test: any test achieving high power against some class of alternatives must necessarily be less powerful

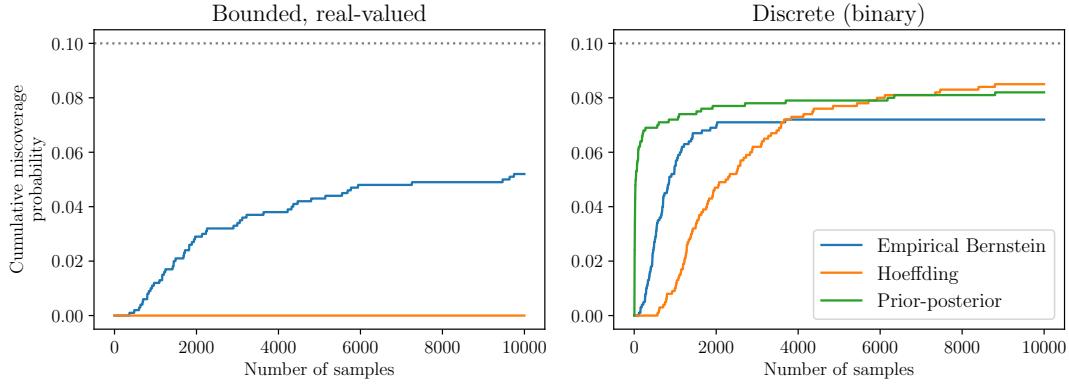


Figure 2.11: Empirical miscoverage probabilities for our empirical Bernstein, Hoeffding, and prior-posterior CSs. The left plot compares empirical Bernstein and Hoeffding for a population of  $N = 10,000$  consisting of bounded, real-valued observations uniformly distributed on the unit interval. The plot on the right-hand side compares all three for a population of the same size containing discrete elements with zeros and ones in equal proportions. Notice that while the empirical Bernstein CS does reasonably well in both settings, none of the three methods uniformly dominates the others.

against some other class of alternatives, while a different test may display the opposite behavior. An analogous story holds for nonparametric estimation as well: the class of bounded random variables (or sequences of bounded random numbers) is nonparametric, and in such a setting, no single estimation technique can uniformly dominate all others (that is, always have lower width for any bounded sequence). This phenomenon is easy to exemplify for our confidence sequences: we can construct settings where the Hoeffding-type CS is less conservative (tighter estimation, more powerful as a test) than the empirical-Bernstein CS, and other settings in which the opposite is true. Figure 2.11 considers two such ‘opposite’ scenarios: the binary setting which maximizes the variance of the sequence, and another setting in which the observations are uniformly distributed on  $[0, 1]$ . In the first setting, there is no point in ‘estimating’ the variance (empirical-Bernstein) as opposed to just assuming that it is the maximum possible variance (Hoeffding-type), and so the former is more conservative than the latter. In the second setting, the Hoeffding CS is far more conservative, as expected. With no prior knowledge on the type of sequence to be encountered, the empirical Bernstein CS seems like a safer choice.

# Chapter 3

## Estimating means of bounded random variables by betting

### 3.1 Introduction

This work presents a new approach to two fundamental problems: (Q1) how do we produce a confidence interval for the mean of a distribution with (known) bounded support using  $n$  independent observations? (Q2) given a fixed list of  $N$  (nonrandom) numbers with known bounds, how do we produce a confidence interval for their mean by sampling  $n \leq N$  of them without replacement in a random order? We work in a nonasymptotic and nonparametric setting, meaning that we do not employ asymptotics or parametric assumptions. Both (Q1) and (Q2) are well studied questions in probability and statistics, but we bring new conceptual tools to bear, resulting in state-of-the-art solutions to both.

We also consider sequential versions of these problems where observations are made one-by-one; we derive time-uniform confidence sequences, or equivalently, confidence intervals that are valid at arbitrary stopping times. In fact, we first describe our techniques in the sequential regime, because the employed proof techniques naturally lend themselves to this setting. We then instantiate the derived bounds for the more familiar setting of a fixed sample size when a batch of data is observed all at once. Our supermartingale techniques can be thought of as generalizations of classical methods for deriving concentration inequalities, but we prefer to present them in the language of betting, since this is a more accurate reflection of the authors' intuition.

Arguably the most famous concentration inequality for bounded random variables was derived by Hoeffding [120]. What is now referred to as “Hoeffding’s inequality” was in fact improved upon in the same paper where he derived a Bernoulli-type upper bound on the moment generating function of bounded random variables [120, Equation (3.4)]. While these bounds are already reasonably tight in a worst-case sense, the resulting confidence intervals do not adapt to non-Bernoulli distributions with lower variance. Inequalities by Bennett [27],

Bernstein [30] and Bentkus [28] improve upon Hoeffding’s, but such improvements require knowledge of nontrivial upper bounds on the variance. This led to the development of so-called “empirical Bernstein inequalities” by Audibert et al. [12] and Maurer and Pontil [187], which outperform Hoeffding’s method for low-variance distributions at large sample sizes by estimating the variance from the data. Our new, and arguably quite simple, approaches to developing bounds significantly outperform these past works (e.g. Figure 3.1).<sup>1</sup> We also show that the same conceptual (betting) framework extends to without-replacement sampling, resulting in significantly tighter bounds than classical ones by Serfling [232], improvements by Bardenet and Maillard [19] and previous state-of-the-art methods found in Chapter 2.

For providing intuition, our approach can be described in words as follows: *If we are allowed to repeatedly bet against the mean being  $m$ , and if we make a lot of money in the process, then we can safely exclude  $m$  from the confidence set.* The rest of this chapter makes the above claim more precise by showing smart, adaptive strategies for (automated) betting, quantifying the phrase “a lot of money”, and explaining why such an exclusion is mathematically justified. At the risk of briefly losing the unacquainted reader, here is a slightly more detailed high-level description:

For each  $m \in [0, 1]$ , we set up a “fair” multi-round game of statistician against nature whose payoff rules are such that if the true mean happened to equal  $m$ , then the statistician can neither gain nor lose wealth in expectation (their wealth in the  $m$ -th game is a nonnegative martingale), but if the mean is not  $m$ , then it is possible to bet smartly and make money. Each round involves the statistician making a bet on the next observation, nature revealing the observation and giving the appropriate (positive or negative) payoff to the statistician. The statistician then plays all these games (one for each  $m$ ) in parallel, starting each with one unit of wealth, and possibly using a different, adaptive, betting strategy in each. The  $1 - \alpha$  confidence set at time  $t$  consists of all  $m \in [0, 1]$  such that the statistician’s money in the corresponding game has not crossed  $1/\alpha$ . The true mean  $\mu$  will be in this set with high probability.

Our choice of language above stems from a game-theoretic approach towards probability, as developed in the books by Shafer and Vovk [235, 236] and a recent paper by Shafer [234], but from a purely mathematical viewpoint, our results are extensions of a unified supermartingale approach towards nonparametric concentration and estimation described in Howard et al. [124, 125]; related supermartingale approaches were studied by [149], [139]. We elaborate on this viewpoint in Section 3.4.1. The most directly related works to our own are by Hendriks [119], whose preprint has initial explorations of methods similar to ours for with-replacement sequential testing and estimation, and [246], who credits Kaplan for a computationally intractable variant of our approach for sequential testing in the without-replacement case. Apart from several novel results, the present chapter extends these past works in *depth, breadth and unity*: our work contains a deeper empirical and theoretical investigation from statistical

---

<sup>1</sup>[github.com/wannabesmith/betting-paper-simulations](https://github.com/wannabesmith/betting-paper-simulations) has code to reproduce figures. The `betting` module of the Python package in [github.com/gostevehoward/confseq](https://github.com/gostevehoward/confseq) has the main algorithms, but the package also contains implementations from other papers.

and computational viewpoints, places our work in a broader context of related work in both settings, and unifies the with- and without-replacement methodology for both testing and estimation in both fixed-time and sequential settings.

We now have the appropriate context for a concrete formalization of our problem, which is slightly more general than introduced above. After that, we describe the game, why the rules of engagement result in valid statistical inference, and derive computationally and statistically efficient betting strategies.

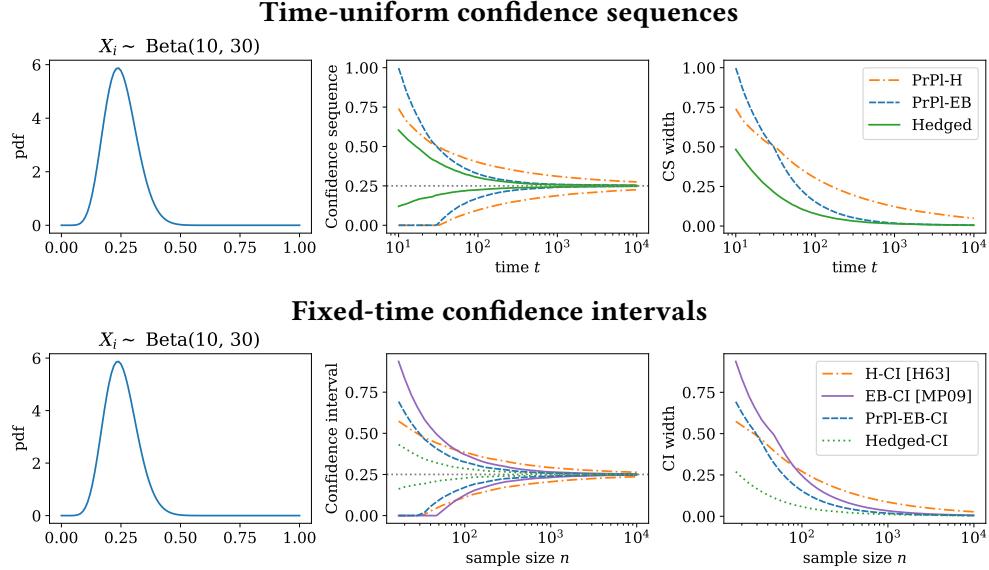


Figure 3.1: Time-uniform 95% confidence sequences (upper row) and fixed-time 95% confidence intervals (lower row) for the mean of independent and identically distributed (iid) draws from a Beta(10, 30) distribution (unknown to the methods). The betting approaches (Hedged and Hedged-CI) adapt to both the small variance and asymmetry of the data, outperforming the other methods. For a detailed empirical comparison under a larger variety of settings, see Section 3.C; for additional comparisons under non-iid data, see Section 3.E.5.

**Outline.** We summarize the broad approach in Section 3.2. As a warmup, we derive a new predictable plug-in method for deriving confidence sequences using exponential supermartingales (Section 3.3), which already leads to computationally efficient and visually appealing empirical Bernstein confidence intervals and sequences. We then further improve on the aforementioned methods by developing a new martingale approach to deriving time-uniform and fixed-time confidence sets for means of bounded random variables, and connect the developed ideas to betting (Section 3.4). Section 3.B discusses some principles to derive powerful betting strategies to obtain tight confidence sets. We then show how our techniques also extend to sampling without replacement (Section 3.5). Revealing simulations are performed along the way to demonstrate the efficacy of the new methods, with a more extensive comparison with past work in Section 3.C. Section 3.6 summarizes how betting ideas have shaped mathematics,

outside of this chapter's focus on statistical inference. We postpone proofs to Section 3.A and further theoretical insights to Section 3.E.

## 3.2 Concentration inequalities via nonnegative supermartingales

To set the stage, let  $\mathcal{Q}^m$  be the set of all distributions on  $[0, 1]$ , where each distribution has mean  $m$ . Note that  $\mathcal{Q}^m$  is a convex set of distributions and it has no common dominating measure, since it consists of both discrete and continuous distributions.

Consider the setting where we observe a (potentially infinite) sequence of  $[0, 1]$ -valued random variables with conditional mean  $\mu$  for some unknown  $\mu \in [0, 1]$ . We write this as  $(X_t)_{t=1}^\infty \sim P$  for some  $P \in \mathcal{P}^\mu$ , where  $\mathcal{P}^\mu$  is the set of all distributions  $P$  on  $[0, 1]^\infty$  such that  $\mathbb{E}_P(X_t | X_1, \dots, X_{t-1}) = \mu$ . This includes familiar settings such as independent observations, where  $X_i \sim Q_i \in \mathcal{Q}^\mu$ , or i.i.d. observations where all  $Q_i$ 's are identical, but captures more general settings where the conditional distribution of  $X_t$  given the past is an element of  $\mathcal{Q}^\mu$ . When one only observes  $n$  outcomes, it suffices to imagine throwing away the rest, so that in what follows, we avoid new notation for distributions  $P$  over finite length sequences.

We are interested in deriving tight confidence sets for  $\mu$ , typically intervals, with no further assumptions. Specifically, for a given error tolerance  $\alpha \in (0, 1)$ , a  $(1 - \alpha)$  confidence interval (CI) is a random set  $C_n \equiv C(X_1, \dots, X_n) \subseteq [0, 1]$  such that

$$\forall n \geq 1, \inf_{P \in \mathcal{P}^\mu} P(\mu \in C_n) \geq 1 - \alpha. \quad (3.1)$$

As mentioned earlier, the inequality by [120] implies that we can choose

$$C_n := \left( \bar{X}_n \pm \sqrt{\frac{\log(2/\alpha)}{2n}} \right) \cap [0, 1]. \quad (3.2)$$

Above, we write  $(a \pm b)$  to mean  $(a - b, a + b)$  for brevity.

This inequality is derived by what is now known as the Chernoff method [39], involving an analytic upper bound on the moment generating function of a bounded random variable. However, we will proceed differently; we adopt a hypothesis testing perspective, and couple it with a generalization of the Chernoff method. As mentioned in the introduction, we first consider the sequential regime where data are observed one after another over time, since nonnegative supermartingales – the primary mathematical tools used throughout this chapter – naturally arise in this setup. As we will see, these sequential bounds can be instantiated for a fixed sample size, yielding tight confidence intervals for this more familiar setting. These will be much tighter than the Hoeffding confidence interval (3.2), which is itself one such fixed-sample-size instantiation [124, Figures 4 and 6].

Let us briefly review some terminology. For succinctness, we use the notation  $X_1^t := (X_1, \dots, X_t)$ . Define the sigma-field  $\mathcal{F}_t := \sigma(X_1^t)$  generated by  $X_1^t$  with  $\mathcal{F}_0$  being the trivial sigma-field. The *canonical filtration*  $\mathcal{F} := (\mathcal{F}_t)_{t=0}^\infty$  refers to the increasing sequence of sigma-fields  $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots$ . A stochastic process  $(M_t)_{t=0}^\infty$  is called a *test supermartingale* for

$P$  if  $(M_t)_{t=0}^\infty$  is a nonnegative process adapted to  $\mathcal{F}$ ,  $M_0 = 1$ , and

$$\mathbb{E}_P(M_t | \mathcal{F}_{t-1}) \leq M_{t-1} \text{ for each } t \geq 1. \quad (3.3)$$

$(M_t)_{t=0}^\infty$  is called a *test martingale* for  $P$  if the above “ $\leq$ ” is replaced with “ $=$ ”. We sometimes shorten  $(M_t)_{t=0}^\infty$  to just  $(M_t)$  for brevity. If the above property holds simultaneously for all  $P \in \mathcal{P}$ , we call  $(M_t)$  a test (super)martingale for  $\mathcal{P}$ . We say that a sequence  $(\lambda_t)_{t=1}^\infty$  is *predictable* if  $\lambda_t$  is  $\mathcal{F}_{t-1}$ -measurable for each  $t \geq 1$ , meaning  $\lambda_t$  can only depend on  $X_1^{t-1}$ . (In)equalities are interpreted in an almost sure sense.

### 3.2.1 Confidence sequences and the method(s) of mixtures

Even though the concentration inequalities thus far have been described in a setting where the sample size  $n$  is fixed in advance, all of our ideas stem from a sequential approach towards uncertainty quantification. The goal there is not to produce one confidence set  $C_n$ , but to produce an infinite sequence  $(C_t)_{t=1}^\infty$  such that

$$\sup_{P \in \mathcal{P}^\mu} P(\exists t \geq 1 : \mu \notin C_t) \leq \alpha. \quad (3.4)$$

Such a  $(C_t)_{t=1}^\infty$  is called a *confidence sequence* (CS), and preferably  $\lim_{t \rightarrow \infty} C_t = \{\mu\}$ . It is known [125, Lemma 3] that (3.4) is equivalent to requiring that  $\sup_{P \in \mathcal{P}^\mu} P(\mu \notin C_\tau) \leq \alpha$  for arbitrary stopping times  $\tau$  with respect to  $\mathcal{F}$ .

As detailed in the next subsection, one general way to construct a CS is to invert a family of sequential tests based on applying Ville’s maximal inequality [264] to a test (super)martingale. In fact, Ramdas et al. [209] proved that this is (in some formal sense) a universal method to construct CSs, meaning that any other approach can in principle be recovered or dominated by the aforementioned one.

Designing test supermartingales is nontrivial, and the task of making it have “power one” against composite alternatives is often accomplished via the *method of mixtures*. This can arguably be traced back (in a nonstochastic context) to Ville’s 1939 thesis and (in a stochastic context) to Wald [271]. Robbins and collaborators [218, 217, 73] applied the method to derive CSs, and these ideas have been extended to a variety of nonparametric settings by Howard et al. [124, 125]. The latter paper describes several variants: conjugate mixtures, discrete mixtures, stitching and inverted stitching.

These works form our vantage point for the rest of the chapter, but we extend them in several ways. First, we describe a “predictable plug-in” technique that is implicit in the work of Ville. It can be viewed as a nonparametric extension of a passing remark in the parametric setting in the textbook by Wald [1945] and later explored in the parametric case by [222].

Like Ville’s work in the binary setting, the predictable plug-in method connects the game-theoretic approach and the aforementioned mixture methods — succinctly, the plugged-in value determines the bet, where each bet is implicitly targeting a different alternative (much like the components of a mixture). Following this translation, prior work on using the method mixtures for confidence sequences can be viewed as using the same betting strategy (mixture

distribution) for every value of  $m$ . We find that there is significant statistical benefit to betting differently for each  $m$  (but tied together in a specific way, not in an ad hoc manner). One must typically specify the mixture distribution in advance of observing data, but betting can be viewed as building up a data-dependent mixture distribution on the fly (this led us to previously name our approach as the “predictable mixture” method). These sequential perspectives are powerful, even if one is only interested in fixed-sample CIs.

### 3.2.2 Nonparametric confidence sequences via sequential testing

As seen above, it is straightforward to derive a confidence interval for  $\mu$  by resorting to a nonparametric concentration inequality like Hoeffding’s. In contrast, it is also well known that CIs are inversions of families of hypothesis tests (as we will see below), so one could presumably derive CIs by first specifying tests. However, the literature on nonparametric concentration inequalities, such as Hoeffding’s, has not commonly utilized a hypothesis testing perspective to derive concentration bounds; for example the excellent book on concentration by Boucheron, Lugosi, and Massart [39] has no examples of such an approach. This is presumably because the underlying nonparametric, composite hypothesis tests may be quite challenging themselves, and one may not have nonasymptotically valid solutions or closed-form analytic expressions for these tests. This is in contrast to simple parametric nulls, where it is often easy to calculate a  $p$ -value based on likelihood ratios. In abandoning parametrics, and thus abandoning likelihood ratios, it may be unclear how to define a powerful test or calculate a nonasymptotically valid  $p$ -value. This is where betting and test (super)martingales come to the rescue. Ramdas et al. [209, Proposition 4] prove that not only do likelihood ratios form test martingales, but every (nonparametric, composite) test martingale is also a (nonparametric, composite) likelihood ratio.

**Theorem 3.2.1** (4-step procedure for supermartingale confidence sets). *On observing  $(X_t)_{t=1}^\infty \sim P$  from  $P \in \mathcal{P}^\mu$  for some unknown  $\mu \in [0, 1]$ , do*

- (a) Consider the composite null hypothesis  $H_0^m : P \in \mathcal{P}^m$  for each  $m \in [0, 1]$ .
- (b) For each index  $m \in [0, 1]$ , construct a nonnegative process  $M_t^m \equiv M^m(X_1, \dots, X_t)$  such that the process  $(M_t^m)_{t=0}^\infty$  indexed by  $\mu$  has the following property: for each  $P \in \mathcal{P}^\mu$ ,  $(M_t^m)_{t=0}^\infty$  is upper-bounded by a test (super)martingale for  $P$ , possibly a different one for each  $P$ .
- (c) For each  $m \in [0, 1]$  consider the sequential test  $(\phi_t^m)_{t=1}^\infty$  defined by

$$\phi_t^m := \mathbb{1}(M_t^m \geq 1/\alpha),$$

where  $\phi_t^m = 1$  represents a rejection of  $H_0^m$  after  $t$  observations.

- (d) Define  $C_t$  as the set of  $m \in [0, 1]$  for which  $\phi_t^m$  fails to reject  $H_0^m$ :

$$C_t := \{m \in [0, 1] : \phi_t^m = 0\}.$$

Then  $(C_t)_{t=1}^\infty$  is a  $(1 - \alpha)$ -confidence sequence for  $\mu$ :  $\sup_{P \in \mathcal{P}^\mu} P(\exists t \geq 1 : \mu \notin C_t) \leq \alpha$ .

The above result relies centrally on Ville's inequality [264], which states that if  $(L_t) \equiv (L_t)_{t=1}^\infty$  is (upper bounded by) a test martingale for  $P$ , then we have  $P(\exists t \geq 1 : L_t \geq 1/\alpha) \leq \alpha$ . See [124, Section 6] for a short proof.

*Proof of Theorem 3.2.1.* By Ville's inequality,  $\phi_t^m$  is a level- $\alpha$  sequential hypothesis test, in the sense that for any  $P \in \mathcal{P}^\mu$ , we have  $P(\exists t \geq 1 : \phi_t^\mu = 1) \leq \alpha$ . Now, by definition of the sets  $(C_t)_{t=1}^\infty$ , we have that  $\mu \notin C_t$  at some time  $t \geq 1$  if and only if there exists a time  $t \geq 1$  such that  $\phi_t^\mu = 1$ , and hence

$$\sup_{P \in \mathcal{P}^\mu} P(\exists t \geq 1 : \mu \notin C_t) = \sup_{P \in \mathcal{P}^\mu} P(\exists t \geq 1 : \phi_t^\mu = 1) \leq \alpha, \quad (3.5)$$

which completes the proof.  $\square$

At a high level, this approach is not new. Composite test supermartingales for  $\mathcal{P}$  have been used in past works on concentration inequalities and/or confidence sequences (which are related but different), from the initial series of works by Robbins and collaborators in the 1960s and 1970s, to [81], to recent work by Jun and Orabona [139, Section 7.2] and Howard et al. [124, 125]. Test martingales have also been explicitly considered in some hypothesis testing problems [267, 237]; the latter paper popularized the term “test martingale” that we borrow, but unlike us, used it primarily for singleton  $\mathcal{P} = \{P\}$ . We highlight an (independently developed) unpublished preprint by Hendriks [119] that has overlaps with the current chapter in the with-replacement setting, and some complementary results. For singleton (parametric) classes  $\mathcal{P}$ , Wald's sequential likelihood ratio statistic is a test martingale, so all of the above methods can be viewed as inverting nonparametric or composite generalizations of Wald's tests.

Nevertheless, we make two additional comments. First, the requirement in step (b) of the algorithm that the process  $(M_t^m)$  be *upper-bounded* by a test (super)martingale for each  $P \in \mathcal{P}$  was posited by [124], and has recently been christened a e-process for  $\mathcal{P}$  [210] (see also [113]). E-processes are strictly more general than test (super)martingales for  $\mathcal{P}$  in the sense that there exist many interesting classes  $\mathcal{P}$  for which nontrivial test (super)martingales do not exist, but one can design powerful e-processes for  $\mathcal{P}$ . Second, one must take care to design test (super)martingales for each  $m$  that are tied together across  $m$  in a nontrivial manner that improves statistical power while maintaining computational tractability. All the confidence sets in this chapter (both in the sequential and batch settings) will be based on this 4-step procedure, but with different carefully chosen processes  $(M_t^m)$ . In the language of betting, we will come up with new, powerful ways to bet for each  $m$ , and also tie together the betting strategies for different  $m$ .

### 3.2.3 Connections to the Chernoff method

By virtue of  $(C_t)_{t=1}^\infty$  being a time-uniform confidence sequence, we also have that  $C_n$  is a  $(1 - \alpha)$ -confidence interval for  $\mu$  for any fixed sample size  $n$ . In fact, the celebrated Chernoff

method results in such a confidence interval. So, how exactly are the two approaches related? The answer is simple: Theorem 3.2.1 generalizes and improves on the Chernoff method. To elaborate, recall that Hoeffding proved that

$$\sup_{P \in \mathcal{P}^\mu} \mathbb{E}_P[\exp(\lambda(X - \mu) - \lambda^2/8)] \leq 1, \text{ for any } \lambda \in \mathbb{R}, \quad (3.6)$$

and so if  $X_1^n$  are independent (say), the following process can be used in Step (b):

$$M_t^m := \prod_{i=1}^t \exp(\lambda(X_i - m) - \lambda^2/8). \quad (3.7)$$

Usually, the only fact that matters for the Chernoff method is that  $\mathbb{E}_P[M_t^m] \leq 1$ , and Markov's inequality is applied (instead of Ville's) in Step (c). To complete the story, the Chernoff method then involves a smart choice for  $\lambda$ . Setting  $\lambda := \sqrt{8 \log(1/\alpha)/n}$  recovers the familiar Hoeffding inequality for the batch sample-size setting. Taking a union bound over  $X_1^n$  and  $-X_1^n$  yields the Hoeffding confidence interval (3.2) exactly. Using our 4-step approach, the resulting confidence sequence is a time-uniform generalization of Hoeffding's inequality, recovering the latter precisely including constants at time  $n$ ; see [124] for this and other generalizations.

In recent parlance, a statistic like  $M_t^m$ , which has at most unit expectation under the null, has been called a betting score [234] or an *e*-value [265] and their relationship to sequential testing [113] and estimation [209] as an alternative to *p*-values has been recently examined. In parametric settings with singleton nulls and alternative hypotheses, the likelihood ratio is an *e*-value. For composite null testing, the split likelihood ratio statistic [280] (and its variants) are *e*-values. However, our setup is more complex:  $\mathcal{P}^m$  is highly composite, there is no common dominating measure to define likelihood ratios, but Hoeffding's result yields an *e*-value. (In fact, it yields test supermartingale and hence an *e*-process, which is an *e*-value even at stopping times.)

In summary, the Chernoff method is simply one powerful, but as it turns out, rather limited way to construct an *e*-value. This chapter provides better constructions of  $M_t^m$ , whose expectation is exactly equal to one, thus removing one source of looseness in the Hoeffding-type approach above, as well as better ways to pick the tuning parameter  $\lambda$ , which will correspond to our bet.

### 3.3 Warmup: exponential supermartingales and predictable plug-ins

A central technique for constructing confidence sequences (CSs) is Robbins' *method of mixtures* [217], see also [73, 218, 220, 221, 222]. Related ideas of “pseudo-maximization” or Laplace's method were further popularized and extended by de la Peña et al. [80, 81, 82], and has led to several other followup works [2, 17, 124, 149].

However, beyond the case when the data are (sub)-Gaussian, the method of mixtures rarely leads to a closed-form CS; it yields an *implicit* construction for  $C_t$  which can sometimes be

computed efficiently (e.g. using conjugate mixtures [125]), but is otherwise analytically opaque and computationally tedious. Below, we provide an alternative construction — called the “predictable plug-in” — that is exact, explicit and efficient (computationally and statistically).

In the next section, our CSs avoid exponential supermartingales, and are much tighter than the recent state-of-the-art in [125]. The ones in this section match the latter but are simpler to compute, so we present them first.

### 3.3.1 Predictable plug-in Cramer-Chernoff supermartingales

Suppose  $(X_t)_{t=1}^{\infty} \sim P$  for some  $P \in \mathcal{P}^{\mu}$  where  $\mathcal{P}^{\mu}$  is the set of all distributions on  $\prod_{i=1}^{\infty} [0, 1]$  so that  $\mathbb{E}_P(X_t \mid \mathcal{F}_{t-1}) = \mu$  for each  $t$ . The Hoeffding process  $(M_t^H(m))_{t=0}^{\infty}$  for a given candidate mean  $m \in [0, 1]$  is given by

$$M_t^H(m) := \prod_{i=1}^t \exp(\lambda(X_i - m) - \psi_H(\lambda)) \quad (3.8)$$

with  $M_0^H(m) \equiv 1$  by convention. Here  $\psi_H(\lambda) := \lambda^2/8$  is an upper bound on the cumulant generating function (CGF) for  $[0, 1]$ -valued random variables with  $\lambda \in \mathbb{R}$  chosen in some strategic way. For example, to maximize  $M_n^H(m)$  at a fixed sample size  $n$ , one would set  $\lambda := \sqrt{8 \log(1/\alpha)/n}$  as in the classical fixed-time Hoeffding inequality [120].

Following Howard et al. [125], we have that  $(M_t^H(\mu))_{t=0}^{\infty}$  is a nonnegative supermartingale with respect to the canonical filtration. Therefore, by Ville’s maximal inequality for nonnegative supermartingales [264, 124],

$$P(\exists t \geq 1 : M_t^H(\mu) \geq 1/\alpha) \leq \alpha. \quad (3.9)$$

Robbins’ method of mixtures proceeds by noting that  $\int_{\lambda \in \mathbb{R}} M_t^H(m) dF(\lambda)$  is also a supermartingale for any “mixing” probability distribution  $F(\lambda)$  on  $\mathbb{R}$  and thus

$$P\left(\exists t \geq 1 : \int_{\lambda \in \mathbb{R}} M_t^H(\mu) dF(\lambda) \geq 1/\alpha\right) \leq \alpha. \quad (3.10)$$

In this particular case, if  $F(\lambda)$  is taken to be the Gaussian distribution, then the above integral can be computed in closed-form [124]. For other distributions or altogether different supermartingales (i.e. other than Hoeffding), the integral may be computationally tedious or intractable.

To combat this, instead of fixing  $\lambda \in \mathbb{R}$  or integrating over it, consider constructing a sequence  $\lambda_1, \lambda_2, \dots$  which is predictable, and thus  $\lambda_t$  can depend on  $X_1^{t-1}$ . Then,

$$M_t^{\text{PrPl-H}}(m) := \prod_{i=1}^t \exp(\lambda_i(X_i - m) - \psi_H(\lambda_i)) \quad (3.11)$$

is also a test supermartingale for  $\mathcal{P}^m$  (and hence Ville’s inequality applies). We call such a sequence  $(\lambda_t)_{t=1}^{\infty}$  a *predictable plug-in*. While not always explicitly referred to by this exact name, predictable plug-ins have appeared in works on parametric sequential analysis by Wald

[272, Eq. (10:10)], Robbins and Siegmund [222, Eq. (4)], Dawid [78], and Lorden and Pollak [178] as well as in the information theory literature [214]. As we will see, these techniques also prove useful in nonparametric testing and estimation problems both in sequential and batch settings.

Using  $M_t^{\text{PrPl-H}}(m)$  as the process in Step (b) of Theorem 3.2.1 results in a lower CS for  $\mu$ , while constructing an analogous supermartingale using  $(-X_t)_{t=1}^\infty$  yields an upper CS. Combining these by taking a union bound results in the predictable plug-in Hoeffding CS which we introduce now.

**Proposition 3.3.1** (Predictable plug-in Hoeffding CS [**PrPl-H**]). *Suppose that  $(X_t)_{t=1}^\infty \sim P$  for some  $P \in \mathcal{P}^\mu$ . For any chosen real-valued predictable  $(\lambda_t)_{t=1}^\infty$ ,*

$$C_t^{\text{PrPl-H}} := \left( \frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i} \right) \text{ forms a } (1 - \alpha)\text{-CS for } \mu,$$

as does its running intersection,  $\bigcap_{i \leq t} C_i^H$ .

A sensible choice of predictable plug-in is given by

$$\lambda_t^{\text{PrPl-H}} := \sqrt{\frac{8 \log(2/\alpha)}{t \log(t+1)}} \wedge 1, \quad (3.12)$$

for reasons which will be discussed in Section 3.3.3. The proof of Proposition 3.3.1 is provided in Section 3.A.1. As alluded to earlier, predictable plug-ins are actually the *least* interesting when using Hoeffding's sub-Gaussian bound because of the available closed form Gaussian-mixture boundary. However, the story becomes more interesting when either (a) the method of mixtures is computationally opaque or complex, or (b) the optimal choice of  $\lambda$  is based on unknown but estimable quantities. Both (a) and (b) are issues that arise when computing empirical Bernstein-type CSs and CIs. In the following section, we present predictable plug-in empirical Bernstein-type CSs and CIs which are both computationally and statistically efficient.

### 3.3.2 Application: closed-form empirical Bernstein confidence sets

To prepare for the results that follow, consider the empirical Bernstein-type process,

$$M_t^{\text{PrPl-EB}}(m) := \prod_{i=1}^t \exp \{ \lambda_i (X_i - m) - v_i \psi_E(\lambda_i) \} \quad (3.13)$$

where, following Howard et al. [124, 125], we have defined  $v_i := 4(X_i - \hat{\mu}_{i-1})^2$  and

$$\psi_E(\lambda) := (-\log(1 - \lambda) - \lambda)/4 \quad \text{for } \lambda \in [0, 1]. \quad (3.14)$$

As we revisit later, the appearance of the constant 4 is to facilitate easy comparison to  $\psi_H$ , since  $\lim_{\lambda \rightarrow 0^+} \psi_E(\lambda)/\psi_H(\lambda) = 1$ . In short,  $\psi_E$  is nonnegative, increasing on  $[0, 1]$ , and grows quadratically near 0.

Using  $M_t^{\text{PrPl-EB}}(m)$  in Step (b) in Theorem 3.2.1 – and applying the same procedure but with  $(X_t)_{t=1}^\infty$  and  $m$  replaced by  $(-X_t)_{t=1}^\infty$  and  $-m$  combined with a union bound over the resulting CSs – we get the following CS.

**Theorem 3.3.1** (Predictable plug-in empirical Bernstein CS [PrPl-EB]). *Suppose  $(X_t)_{t=1}^\infty \sim P$  for some  $P \in \mathcal{P}^\mu$ . For any  $(0, 1)$ -valued predictable  $(\lambda_t)_{t=1}^\infty$ ,*

$$C_t^{\text{PrPl-EB}} := \left( \frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t v_i \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i} \right) \text{ forms a } (1 - \alpha)\text{-CS for } \mu,$$

as does its running intersection,  $\bigcap_{i \leq t} C_i^{\text{PrPl-EB}}$ .

In particular, we recommend the predictable plug-in  $(\lambda_t^{\text{PrPl-EB}})_{t=1}^\infty$  given by

$$\lambda_t^{\text{PrPl-EB}} := \sqrt{\frac{2 \log(2/\alpha)}{\hat{\sigma}_{t-1}^2 t \log(1+t)}} \wedge c, \quad \hat{\sigma}_t^2 := \frac{\frac{1}{4} + \sum_{i=1}^t (X_i - \hat{\mu}_i)^2}{t+1}, \quad \hat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1} \quad (3.15)$$

for some  $c \in (0, 1)$  (a reasonable default being  $1/2$  or  $3/4$ ). This choice was inspired by the fixed-time empirical Bernstein as well as the widths of time-uniform CSs (more details are provided in Section 3.3.3). The sequences of estimators  $(\hat{\mu}_t)_{t=1}^\infty$  and  $(\hat{\sigma}_t^2)_{t=1}^\infty$  can be interpreted as predictable, regularized sample means and variances. This technique was employed by Kothlowski et al. [162] for misspecified exponential families in the so-called *maximum likelihood plug-in strategy*.

The proof of Theorem 3.3.1 relies on establishing that  $M_t^{\text{PrPl-EB}}(m)$  is a test supermartingale for  $\mathcal{P}^m$ . This latter fact is related to, but cannot be derived directly from, a powerful deterministic inequality for bounded numbers due to Fan et al. [102]. One needs an additional trick from Howard et al. [125, Section A.8] which swaps  $(X_i - m)^2$  with  $(X_i - \hat{\mu}_{i-1})^2$ , for any predictable  $\hat{\mu}_{i-1}$ , within the variance term  $v_i$ . It is this additional piece which yields both tighter and *closed-form* CSs; details are in Section 3.A.2. We remark that before taking the running intersection, the above intervals are symmetric around the weighted sample mean, but this symmetry will not carry forward to other CSs in the chapter.

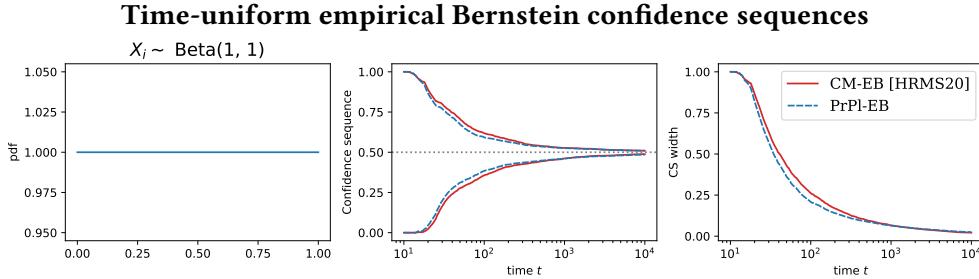


Figure 3.2: Empirical Bernstein CSs produced via a predictable plug-in (PrPl) with  $(\lambda_t)_{t=1}^\infty$  from (3.15) match (or slightly improve) those obtained via conjugate mixtures (CM) by [125]; the former is closed-form, but the latter is not and requires numerical methods.

Figure 3.2 compares the conjugate mixture empirical-Bernstein CS (CM-EB) due to Howard

et al. [125] with our predictable plug-in empirical-Bernstein CS (PrPl-EB). The two CSs perform similarly, but our closed-form PrPl-EB is over 500 times faster to compute than CM-EB (in our experience) which requires root finding at each step. However, our later bounds will be tighter than both of these.

*Remark 1.* Theorem 3.3.1 yields computationally and statistically efficient empirical Bernstein-type CIs for a fixed sample size  $n$ . Recalling (3.15), we recommend using  $\bigcap_{i \leq n} C_i^{\text{PrPl-EB}}$  along with the predictable sequence

$$\lambda_t^{\text{PrPl-EB}(n)} := \sqrt{\frac{2 \log(2/\alpha)}{n \hat{\sigma}_{t-1}^2}} \wedge c. \quad (3.16)$$

We call the resulting confidence interval the ‘‘predictable plug-in empirical Bernstein confidence interval’’ or **[PrPl-EB-CI]** for short; see Figure 3.3.

If  $X_1, \dots, X_n$  are independent, then at the expense of computation, the above CI can be effectively derandomized to remove the effect of the ordering of variables. One can randomly permute the data  $B$  times to obtain  $(\tilde{X}_{1,b}, \dots, \tilde{X}_{n,b})$  and correspondingly compute  $\tilde{M}_{n,b}^{\text{PrPl-EB}}(m)$ , one for each permutation  $b \in \{1, \dots, B\}$ . Averaging over these permutations, define  $\tilde{M}_n^{\text{PrPl-EB}}(m) := \frac{1}{B} \sum_{b=1}^B \tilde{M}_{n,b}^{\text{PrPl-EB}}(m)$ . For each  $b$ ,  $M_{n,b}^{\text{PrPl-EB}}(\mu)$  has expectation at most one (by linearity of expectation). Thus,  $\tilde{M}_n^{\text{PrPl-EB}}(\mu)$  is a  $e$ -value (i.e. it has expectation at most 1). By Markov’s inequality,  $\tilde{C}_n^{\text{PrPl-EB}} := \{m \in [0, 1] : \tilde{M}_n^{\text{PrPl-EB}}(m) < 1/\alpha\}$  is a  $(1 - \alpha)$ -CI for  $\mu$ . This set is not available in closed-form and the intersection  $\bigcap_{i \leq n} \tilde{C}_i^{\text{PrPl-EB}}$  no longer yield a valid CI. In our experience, this derandomization procedure neither helps nor hurts. In any case, both  $\bigcap_{i \leq n} C_i$  and  $\tilde{C}_n$  will be significantly improved in Section 3.4.4.

In Section 3.E.3, we show that in iid settings the width of [PrPl-EB-CI] scales with the true (unknown) standard deviation:

$$\sqrt{n} \left( \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i} \right) \xrightarrow{\text{a.s.}} \sigma \sqrt{2 \log(2/\alpha)}. \quad (3.17)$$

Notice that (3.17) is the same asymptotic behavior that one would observe for CIs based on Bernstein’s or Bennett’s inequalities, both of which require knowledge of the true variance  $\sigma^2$ , while [PrPl-EB-CI] does not. This is in contrast to the empirical Bernstein CIs of Maurer and Pontil [187] whose limit would be  $\sigma \sqrt{2 \log(4/\alpha)}$ . In the maximum variance case where  $\sigma = 1/2$ , (3.17) yields the same asymptotic behavior as Hoeffding’s CI (3.2).

Until now, we presented various predictable plug-ins —  $(\lambda_t^{\text{PrPl-H}})_{t=1}^\infty$ ,  $(\lambda_t^{\text{PrPl-EB}})_{t=1}^\infty$ , and  $(\lambda_t^{\text{PrPl-EB}(n)})_{t=1}^n$  — but have not provided intuition for why these are sensible choices. Next, we discuss guiding principles for deriving predictable plug-ins.

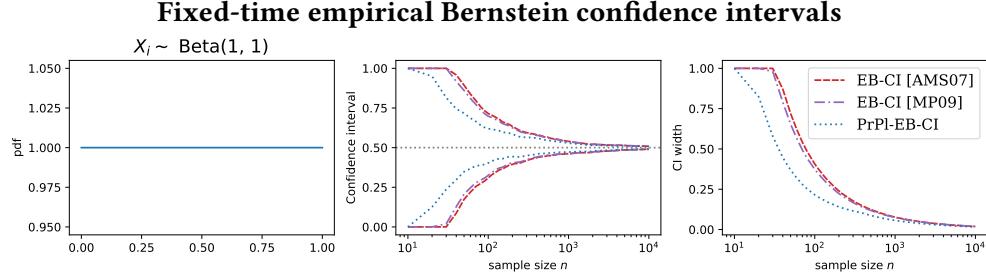


Figure 3.3: Our predictable plug-in (PrPl) empirical Bernstein (EB) CI is significantly tighter than those of [187] and [12].

Table 3.1: Below, we think of  $\log x$  as  $\log(x + 1)$  to avoid trivialities. The claimed rates are easily checked by approximating the sums as integrals, and taking derivatives. For example,  $\frac{d}{dx} \log \log x = 1/x \log x$ , so the sum of  $\sum_{i \leq t} 1/i \log i \asymp \log \log t$ . It is worth remarking that for  $t = 10^{80}$ , the number of atoms in the universe,  $\log \log t \approx 5.2$ , which is why we treat  $\log \log t$  as a constant when expressing the rate for  $W_t$ . The iterated logarithm pattern in the last two lines can be continued indefinitely.

Strategy $(\lambda_i)_{i=1}^\infty$	$\sum_{i=1}^t \lambda_i$	$\sum_{i=1}^t \lambda_i^2$	Width $W_t$
$\asymp 1/i$	$\asymp \log t$	$\asymp 1$	$1/\log t$
$\asymp \sqrt{\log i/i}$	$\asymp \sqrt{t \log t}$	$\asymp \log^2 t$	$\asymp \log^{3/2} t / \sqrt{t}$
$\asymp 1/\sqrt{i}$	$\asymp \sqrt{t}$	$\asymp \log t$	$\asymp \log t / \sqrt{t}$
$\asymp 1/\sqrt{i \log i}$	$\asymp \sqrt{t/\log t}$	$\asymp \log \log t$	$\asymp \sqrt{\log t/t}$
$\asymp 1/\sqrt{i \log i \log \log i}$	$\asymp \sqrt{t/\log t}$	$\asymp \log \log \log t$	$\asymp \sqrt{\log t/t}$

### 3.3.3 Guiding principles for deriving predictable plug-ins

Let us begin our discussion with the predictable plug-in Hoeffding process (3.11) and the resulting CS in Proposition 3.3.1, which has a half-width

$$W_t = \frac{\log(2/\alpha) + \sum_{i=1}^t \lambda_i^2/8}{\sum_{i=1}^t \lambda_i}$$

To ensure that  $W_t \rightarrow 0$  as  $t \rightarrow \infty$ , it is clear that we want  $\lambda_t \xrightarrow{\text{a.s.}} 0$ , but at what rate? As a sensible default, we recommend setting  $\lambda_t \asymp 1/\sqrt{t \log t}$  so that  $W_t = \tilde{O}(\sqrt{\log t/t})$  which matches the width of the conjugate mixture Hoeffding CS [124, Proposition 2] (here  $\tilde{O}$  treats  $O(\log \log t)$  factors as constants). See Table 3.1 for a comparison between rates for  $\lambda_t$  and their resulting CS widths.

Now consider the predictable plug-in empirical Bernstein process (3.13) and the resulting CS of Theorem 3.3.1, which has a half-width

$$W_t = \frac{\log(2/\alpha) + \sum_{i=1}^t 4(X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i}$$

By two applications of L'Hôpital's rule, we have that

$$\frac{\psi_E(\lambda)}{\psi_H(\lambda)} \xrightarrow{\lambda \rightarrow 0^+} 1. \quad (3.18)$$

Performing some approximations for small  $\lambda_i$  to help guide our choice of  $(\lambda_t)_{t=1}^\infty$  (without compromising validity of resulting confidence sets) we have that

$$W_t \approx \frac{\log(2/\alpha) + \sum_{i=1}^t 4(X_i - \mu)^2 \lambda_i^2 / 8}{\sum_{i=1}^t \lambda_i}. \quad (3.19)$$

Thus, in the special case of i.i.d.  $X_i$  with variance  $\sigma^2$ , for large enough  $t$ ,

$$\mathbb{E}_P(W_t \mid \mathcal{F}_{t-1}) \lesssim \frac{\log(2/\alpha) + \sigma^2 \sum_{i=1}^t \lambda_i^2 / 2}{\sum_{i=1}^t \lambda_i}. \quad (3.20)$$

If we were to set  $\lambda_1 = \lambda_2 = \dots = \lambda^* \in \mathbb{R}$  and minimize the above expression for a specific time  $t^*$ , this amounts to minimizing

$$\frac{\log(2/\alpha) + \sigma^2 t^* \lambda^{*2} / 2}{t^* \lambda^*}, \quad (3.21)$$

which is achieved by setting

$$\lambda^* := \sqrt{\frac{2 \log(2/\alpha)}{\sigma^2 t^*}}. \quad (3.22)$$

This is precisely why we suggested the predictable plug-in  $(\lambda_t^{\text{PrPl}})_{t=1}^\infty$  given by (3.15), where the additional  $\log(t+1)$  is included in an attempt to enforce  $W_t = \tilde{O}(\sqrt{\log t/t})$ .

The above calculations are only used as guiding principles to sharpen the confidence sets, but *all* such schemes retain the validity guarantee. As long as  $(\lambda_t)_{t=1}^\infty$  is  $[0, 1]$ -valued and predictable, we have that  $(M_t^E(\mu))_{t=0}^\infty$  is a test supermartingale for  $\mathcal{P}^\mu$  which can be used in Theorem 3.2.1 to obtain different valid CSs for  $\mu$ .

Foreshadowing our attempt to generalize this procedure in the next section, notice that the exponential function was used throughout to ensure nonnegativity, but that any other test supermartingale would have sufficed. In fact, if a martingale is used in place of a supermartingale, then Ville's inequality is tighter.

Next, we present a test *martingale*, removing a source of looseness in the confidence sets derived thus far. We discuss its betting interpretation, provide other guiding principles for setting  $\lambda_i$  (equivalently, for betting), which will involve attempting to maximize the expected log-wealth in the betting game.

### 3.4 The capital process, betting, and martingales

In Section 3.3, we generalized the Cramer-Chernoff method to derive predictable plug-in exponential supermartingales and used this result to obtain tight empirical Bernstein CSs and CIs. In this section, we consider an alternative process which can be interpreted as the wealth accumulated from a series of bets in a game. This process is a central object of study in the game-theoretic probability literature where it is referred to as the *capital process* [235]. We discuss its connections to the purely statistical goal of constructing CSs and CIs and demonstrate how these sets improve on Cramer-Chernoff approaches, including the empirical Bernstein confidence sets of the previous section.

Consider the same setup as in Section 3.3: we observe an infinite sequence of conditionally mean- $\mu$  random variables,  $(X_t)_{t=1}^\infty \sim P$  from some distribution  $P \in \mathcal{P}^\mu$ . Define the *capital process*  $\mathcal{K}_t(m)$  for any  $m \in [0, 1]$ ,

$$\mathcal{K}_t(m) := \prod_{i=1}^t (1 + \lambda_i(m) \cdot (X_i - m)), \quad (3.23)$$

with  $\mathcal{K}_0(m) := 1$  and where  $(\lambda_t(m))_{t=1}^\infty$  is a  $(-1/(1-m), 1/m)$ -valued predictable sequence, and thus  $\lambda_t(m)$  can depend on  $X_1^{t-1}$ . Note that for each  $t \geq 1$ , we have  $X_t \in [0, 1]$ ,  $m \in [0, 1]$  and  $\lambda_t(m) \in (-1/(1-m), 1/m)$ . Here and below,  $1/m$  should be interpreted as  $\infty$  when  $m = 0$  and similarly for  $1/(1-m)$  and  $m = 1$ , respectively. Importantly,  $(1 + \lambda_t(m) \cdot (X_t - m)) \in [0, \infty)$ , and thus  $\mathcal{K}_t(m) \geq 0$  for all  $t \geq 1$ . Following similar techniques to the previous section, the reader may easily check that  $\mathcal{K}_t(\mu)$  is a test martingale. Moreover, we have the stronger result summarized in the following central proposition.

**Proposition 3.4.1.** *Suppose a draw from some distribution  $P$  yields a sequence  $X_1, X_2, \dots$  of  $[0, 1]$ -valued random variables, and let  $\mu \in [0, 1]$  be a constant. The following four statements imply each other:*

- (a)  $\mathbb{E}_P(X_t | \mathcal{F}_{t-1}) = \mu$  for all  $t \in \mathbb{N}$ , where  $\mathcal{F}_{t-1} = \sigma(X_1, \dots, X_{t-1})$ .
- (b) There exists a constant  $\lambda \in \mathbb{R} \setminus \{0\}$  for which  $(\mathcal{K}_t(\mu))_{t=0}^\infty$  is a strictly positive test martingale for  $P$ .
- (c) For every fixed  $\lambda \in (-\frac{1}{1-\mu}, \frac{1}{\mu})$ ,  $(\mathcal{K}_t(\mu))_{t=0}^\infty$  is a test martingale for  $P$ .
- (d) For every  $(-\frac{1}{1-\mu}, \frac{1}{\mu})$ -valued predictable sequence  $(\lambda_t)_{t=1}^\infty$ ,  $(\mathcal{K}_t(\mu))_{t=0}^\infty$  is a test martingale for  $P$ .

Further, the intervals  $(-\frac{1}{1-\mu}, \frac{1}{\mu})$  mentioned above can be replaced by any subinterval containing at least one nonzero value, like  $[-1, 1]$  or  $(-0.5, 0.5)$ . Finally, every test martingale for  $\mathcal{P}^\mu$  is of the form  $(\mathcal{K}_t(\mu))$  for some predictable sequence  $(\lambda_t)$ .

The proof can be found in Section 3.A.3. While the subsequent theorems will primarily make use of (a)  $\implies$  (d), the above proposition establishes a core fact: the assumption of the (conditional) means being identically  $\mu$  is an *equivalent restatement* of our capital process being a test martingale. Thus, test martingales are not simply “technical tools” to deal with

means of bounded random variables, they are fundamentally at the very heart of the problem definition itself.

Proposition 3.4.1 can be generalized to another remarkable, yet simple, result: for any set of distributions  $\mathcal{S}$ , every test martingale for  $\mathcal{S}$  has the same form.

**Proposition 3.4.2** (Universal representation). *For any arbitrary set of (possibly unbounded) distributions  $\mathcal{S}$ ,  $(M_t)$  is a test martingale for  $\mathcal{S}$  if and only if  $M_t = \prod_{i=1}^t (1 + \lambda_i Z_i)$  for some  $Z_i \geq -1$  such that  $\mathbb{E}_S[Z_i | \mathcal{F}_{i-1}] = 0$  for every  $S \in \mathcal{S}$ , and some predictable  $\lambda_i$  such that  $\lambda_i Z_i \geq -1$ . The same claim also holds for test supermartingales for  $\mathcal{S}$ , with the aforementioned “= 0” replaced by “ $\leq 0$ ”.*

The proof can be found in Section 3.A.4. The above proposition immediately makes this chapter’s techniques actionable for a wide class of nonparametric testing and estimation problems. We give an example relating to quantiles later.

### 3.4.1 Connections to betting

It is worth pausing to clarify how the capital process  $\mathcal{K}_t(m)$  and Proposition 3.4.1 can be viewed in terms of betting. We imagine that nature implicitly posits a hypothesis  $H_0^m$  – which we treat as a game providing us a chance to make money if the hypothesis is wrong, by repeatedly betting some of our capital against  $H_0^m$ . We start the game with a capital of 1 (i.e.  $\mathcal{K}_0(m) := 1$ ), and design a bet of  $b_t := s_t |\lambda_t^m|$  at each step, where  $s_t \in \{-1, 1\}$ . Setting  $s_t := 1$  indicates that we believe that  $\mu > m$  while  $s_t := -1$  indicates the opposite.  $|\lambda_t^m|$  indicates the amount of our capital that we are willing to put at stake at time  $t$ : setting  $\lambda_t^m = 0$  results in neither losing nor gaining any capital regardless of the outcome, while setting  $\lambda_t^m \in \{-1/(1-m), 1/m\}$  means that we are willing to risk all of our capital on the next outcome.

However, if  $H_0^m$  is true (i.e.  $m = \mu$ ), then by Proposition 3.4.1, our capital process is a martingale. In betting terms, no matter how clever a betting strategy  $(\lambda_t^m)_{t=1}^\infty$  we devise, we cannot expect to make (or lose) money at each step. If on the other hand,  $H_0^m$  is false, then a clever betting strategy will make us a lot of money. In statistical terms, when our capital exceeds  $1/\alpha$ , we can confidently reject the hypothesis  $H_0^m$  since if it were true (and the game were fair) then by Ville’s inequality [264], the a priori probability of this ever occurring is at most  $\alpha$ . We imagine simultaneously playing this game with  $H_0^{m'}$  for each  $m' \in [0, 1]$ . At any time  $t$ , the games  $m' \in [0, 1]$  for which our capital is small ( $< 1/\alpha$ ) form a CS.

Both the Cramer-Chernoff processes of Section 3.3 and  $\mathcal{K}_t(m)$  are nonnegative and tend to increase when  $\mu > m$ . However, only  $\mathcal{K}_t(m)$  is a *test martingale* when  $m = \mu$ ; the others are test supermartingales. A test martingale is the wealth accumulated in a “fair game” where our capital stays constant in expectation, while a test supermartingale is the wealth accumulated in a game where our capital is expected to decrease (not strictly). Larger values of capital correspond to rejecting  $H_0^m$  more readily. Therefore, test supermartingales tend to yield conservative tests compared to their martingale counterparts.

More generally, every nonnegative supermartingale can be regarded as the wealth pro-

cess of a gambler playing a game with odds that are fair or stacked against them. In other words, there is a one-to-one correspondence between wealths of hypothetical gamblers and nonnegative supermartingales. Taking this perspective, every statement involving nonnegative supermartingales (and thus likelihood ratios) are statements about betting, and vice versa. Mixture methods that combine nonnegative supermartingales are simply strategies to hedge across various instruments available to the gambler. Thus, the gambling analogy can be entirely dropped, and our results would find themselves comfortably nestled in the rich literature on martingale methods for concentration inequalities, but we mention the betting analogy for intuition so that the mathematics are animated and easier to absorb.

Ville introduced martingales into modern mathematical probability theory, and centered them around their betting interpretation. Since then, ideas from betting have appeared in various fields, including probability theory, statistical testing and estimation, information theory, and online learning theory. While this chapter focuses on the utility of betting in some statistical inference tasks, Section 3.F provides a brief overview of the use of betting in other mathematical disciplines.

### 3.4.2 Connections to likelihood ratios

As alluded to in the previous subsection, useful intuition is provided via the connection to likelihood ratios.  $\mathcal{K}_t(m)$  is a “composite” test martingale for  $\mathcal{P}^m$ , meaning that it is a nonnegative martingale starting at one for every  $P \in \mathcal{P}^m$  (recall that  $P$  is a distribution over infinite sequences of observations with conditional mean  $m$ ).

If we were dealing with a single distribution such as  $Q^\infty$ , meaning a product distribution where every observation is drawn iid from  $Q$ , then one may pick any alternative  $Q'$  that is absolutely continuous with respect to  $Q$ , to observe that the likelihood ratio  $\prod_{i=1}^t Q'(X_i)/Q(X_i)$  is a test martingale for  $Q^\infty$ .

However, since  $\mathcal{P}^m$  is highly composite and nonparametric and is not even dominated by a single measure (as it contains atomic measures, continuous measures, and all their mixtures), it is unclear how one can even begin to write down a likelihood ratio. Nevertheless, Ramdas et al. [209, Proposition 4] show that if  $(M_t)$  is a composite test martingale for any  $\mathcal{S}$ , then for every distribution  $Q \in \mathcal{S}$ ,  $M_t$  equals the likelihood ratio of some  $Q'$  against  $Q$  (where  $Q'$  depends on  $Q$ ).

Thus, not only is every likelihood ratio a test martingale, but every (composite) test martingale can also be represented as a likelihood ratio. Hence, in a formal sense, test martingales are nonparametric composite generalizations of likelihood ratios, which are at the very heart of statistical inference. When this observation is combined with Proposition 3.4.1, it should be no surprise any longer that the capital process  $\mathcal{K}_t(m)$  (even devoid of any betting interpretation) is fundamental to the problem at hand. In Section 3.E.6 we also observe connections to the empirical likelihood of [200] and the dual likelihood of [192].

### 3.4.3 Adaptive, constrained adversaries

Despite the analogies to betting, the game described so far appears to be purely stochastic in the sense that nature simply commits to a distribution  $P \in \mathcal{P}^\mu$  for some unknown  $\mu \in [0, 1]$  and presents us observations from  $P$ . However, Proposition 3.4.1 can be extended to a more adversarial setup, but with a constrained adversary.

To elaborate, recall the difference between  $\mathcal{Q}$  and  $\mathcal{P}$  from the start of Section 3.2 and consider a game with three players: an adversary, nature, and the statistician. First, the adversary commits to a  $\mu \in [0, 1]$ . Then, the game proceeds in rounds. At the start of round  $t$ , the statistician publicly discloses the bets for every  $m$ , which could depend on  $X_1, \dots, X_{t-1}$ . The adversary picks a distribution  $Q_t \in \mathcal{Q}^\mu$ , which could depend on  $X_1, \dots, X_{t-1}$  and the statistician's disclosed bets, and hands  $Q_t$  to nature. Nature simply acts like an arbitrator, first verifying that the adversary chose a  $Q_t$  with mean  $\mu$ , and then draws  $X_t \sim Q_t$  and presents  $X_t$  to the statistician.

In this fashion, the adversary does not need to pick  $\mu$  and  $P \in \mathcal{P}^\mu$  at the start of the interaction, which is the usual stochastic setup, but can instead build the distribution  $P$  in a data-dependent fashion over time. In other words, the adversary does not commit to a distribution  $P$ , but instead to a *rule for building  $P$*  from the data. Of course, they do not need to disclose this rule, or even be able express what this rule would do on any other hypothetical outcomes other than the one observed. The results in this chapter, which build on the central Proposition 3.4.1, continue to hold in this more general interaction model.

A geometric reason why we can move from the stochastic model first described to the above (constrained) adversarial model, is that the above distribution  $P$  lies in the “fork convex hull” of  $\mathcal{P}^\mu$ . Fork-convexity is a sequential analogue of convexity [210]. Informally, the fork-convex hull of a set of distributions over sequences is the set of predictable plug-ins of these distributions, and is much larger than their convex hull (mixtures). If a process is a nonnegative martingale under every distribution in a set, then it is also a nonnegative martingale under every distribution in the fork convex hull of that set. No results about fork convexity are used anywhere in this chapter, and we only mention it for the mathematically curious.

### 3.4.4 The hedged capital process

We now return to the purely statistical problem of using the capital process  $\mathcal{K}_t(m)$  to construct time-uniform CSs and fixed-time CIs. We might be tempted to use  $\mathcal{K}_t(\mu)$  as the nonnegative martingale in Theorem 3.2.1 to conclude that  $\mathfrak{B}_t := \{m \in [0, 1] : \mathcal{K}_t(m) < 1/\alpha\}$  forms a  $(1 - \alpha)$ -CS for  $\mu$ . Unlike the empirical Bernstein CS of Section 3.3,  $\mathfrak{B}_t$  cannot be computed in closed-form. Instead, we theoretically need to compute the family of processes  $\{\mathcal{K}_t(m)\}_{m \in [0, 1]}$  and include those  $m \in [0, 1]$  for which  $\mathcal{K}_t(m)$  remains below  $1/\alpha$ . This is not practical as the parameter space  $[0, 1]$  is uncountably infinite. But if we know a priori that  $\mathfrak{B}_t$  is guaranteed to produce an interval for each  $t$ , then it is straightforward to find a superset of  $\mathfrak{B}_t$  by either performing a grid search on  $(0, 1/g, 2/g, \dots, (g-1)/g, 1)$  for some large  $g \in \mathbb{N}$ , or by employing root-finding

algorithms. This motivates the *hedged capital process*, defined for any  $\theta, m \in [0, 1]$  as

$$\begin{aligned} \mathcal{K}_t^\pm(m) &:= \max \{\theta \mathcal{K}_t^+(m), (1 - \theta) \mathcal{K}_t^-(m)\}, \\ \text{where } \mathcal{K}_t^+(m) &:= \prod_{i=1}^t (1 + \lambda_i^+(m) \cdot (X_i - m)), \\ \text{and } \mathcal{K}_t^-(m) &:= \prod_{i=1}^t (1 - \lambda_i^-(m) \cdot (X_i - m)), \end{aligned} \quad (3.24)$$

and  $(\lambda_t^+(m))_{t=1}^\infty$  and  $(\lambda_t^-(m))_{t=1}^\infty$  are predictable sequences of  $[0, \frac{1}{m}]$ - and  $[0, \frac{1}{1-m}]$ -valued random variables, respectively.

$\mathcal{K}_t^\pm(m)$  can be viewed from the betting perspective as dividing one's capital into proportions of  $\theta$  and  $(1 - \theta)$  and making two series of simultaneous bets, positing that  $\mu \geq m$ , and  $\mu < m$ , respectively which accumulate capital in  $\mathcal{K}_t^+(m)$  and  $\mathcal{K}_t^-(m)$ . If  $\mu \neq m$ , then we expect that one of these strategies will perform poorly, while we expect the other to make money in the long term. If  $\mu = m$ , then we expect neither strategy to make money. The maximum of these processes is upper-bounded by their convex combination,

$$\mathcal{M}_t^\pm := \theta \mathcal{K}_t^+ + (1 - \theta) \mathcal{K}_t^-.$$

Both  $\mathcal{K}_t^\pm$  and  $\mathcal{M}_t^\pm$  can be used for Step (b) of Theorem 3.2.1 to yield a CS. Empirically, both yield intervals, but only the former provably so.

**Theorem 3.4.1** (Hedged capital CS [**Hedged**]). *Suppose  $(X_t)_{t=1}^\infty \sim P$  for some  $P \in \mathcal{P}^\mu$ . Let  $(\tilde{\lambda}_t^+)_{t=1}^\infty$  and  $(\tilde{\lambda}_t^-)_{t=1}^\infty$  be real-valued predictable sequences not depending on  $m$ , and for each  $t \geq 1$  let*

$$\lambda_t^+(m) := |\tilde{\lambda}_t^+| \wedge \frac{c}{m}, \quad \lambda_t^-(m) := |\tilde{\lambda}_t^-| \wedge \frac{c}{1-m}, \quad (3.25)$$

for some  $c \in [0, 1)$  (some reasonable defaults being  $c = 1/2$  or  $3/4$ ). Then

$$\mathfrak{B}_t^\pm := \{m \in [0, 1] : \mathcal{K}_t^\pm(m) < 1/\alpha\} \quad \text{forms a } (1 - \alpha)\text{-CS for } \mu,$$

as does its running intersection  $\bigcap_{i \leq t} \mathfrak{B}_i^\pm$ . Further,  $\mathfrak{B}_t^\pm$  is an interval for each  $t \geq 1$ . Finally, replacing  $\mathcal{K}_t^\pm(m)$  by  $\mathcal{M}_t^\pm(m)$  yields a tighter  $(1 - \alpha)$ -CS for  $\mu$ .

For reasons given in Section 3.B.1, we recommend setting  $\tilde{\lambda}_t^+ = \tilde{\lambda}_t^- = \lambda_t^{\text{PrPl}\pm}$  as

$$\lambda_t^{\text{PrPl}\pm} := \sqrt{\frac{2 \log(2/\alpha)}{\hat{\sigma}_{t-1}^2 t \log(t+1)}}, \quad \hat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \hat{\mu}_i)^2}{t+1}, \quad \text{and } \hat{\mu}_t := \frac{1/2 + \sum_{i=1}^t X_i}{t+1}, \quad (3.26)$$

for each  $t \geq 1$ , and truncation level  $c := 1/2$  or  $3/4$ ; see Figure 3.4. A reasonable point estimator for  $\mu$  is  $\operatorname{argmin}_{m \in [0, 1]} \mathcal{K}_t^\pm(m)$  or  $\operatorname{argmin}_{m \in [0, 1]} \mathcal{M}_t^\pm(m)$  (see Figure 3.18).

*Remark 2.* Since  $\mathcal{K}_t^\pm(m) \leq \mathcal{M}_t^\pm(m)$ , the latter confidence sequence is tighter. In the proof of Theorem 3.4.1, we use a property of the max function to establish quasiconvexity of  $\mathcal{K}_t^\pm(m)$ , implying that  $\mathfrak{B}_t^\pm$  is an interval. We find the difference in empirical performance negligible

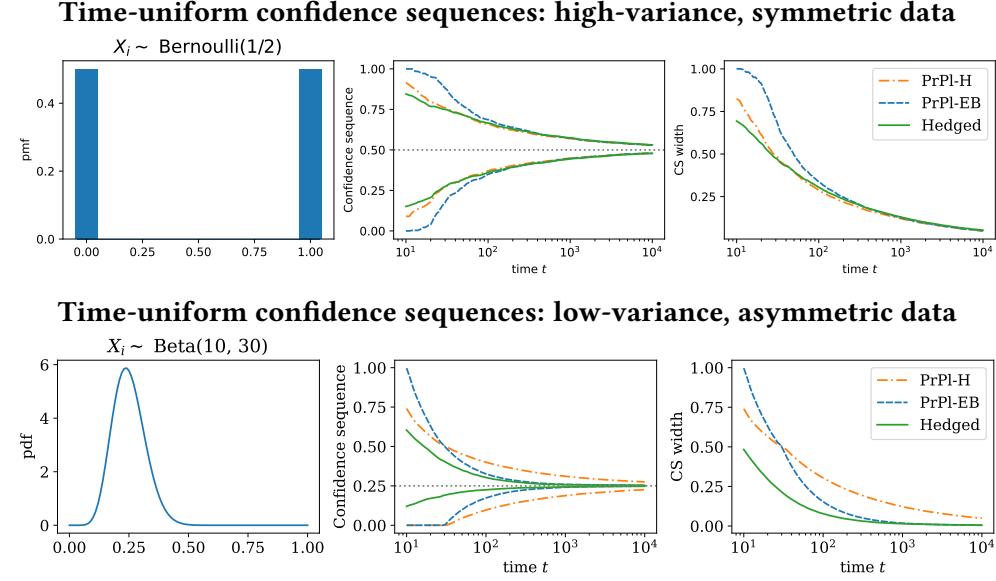


Figure 3.4: Predictable plug-in Hoeffding, empirical Bernstein, and hedged capital CSs under two distributional scenarios. Notice that the latter roughly matches the others in the  $\text{Bernoulli}(1/2)$  case, but shines in the low-variance, asymmetric scenario.

(Figure 3.5). For the interested reader, Section 3.E.4 constructs a (pathological) CS that is almost surely not an interval.

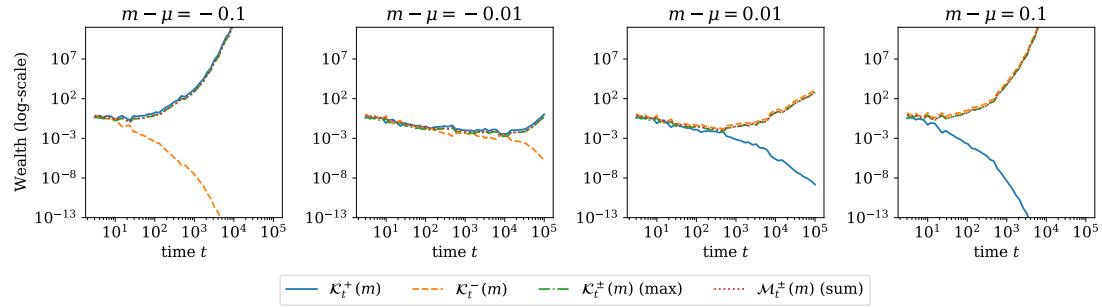


Figure 3.5: A comparison of capital processes  $\mathcal{K}_t^+(m)$ ,  $\mathcal{K}_t^-(m)$ , the hedged capital process  $\mathcal{K}_t^\pm(m)$ , and its upper-bounding nonnegative martingale,  $\mathcal{M}_t^\pm(m)$  under four alternatives (from left to right):  $m \ll \mu$ ,  $m < \mu$ ,  $m > \mu$ ,  $m \gg \mu$ . When  $m < \mu$ , we see that  $\mathcal{K}_t^+(m)$  increases, while  $\mathcal{K}_t^-(m)$  approaches zero, but the opposite is true when  $m > \mu$ . Notice that not much is gained by taking a sum  $\mathcal{M}_t^\pm(m)$  rather than a maximum  $\mathcal{K}_t^\pm(m)$ , since one of  $\mathcal{K}_t^+(m)$  and  $\mathcal{K}_t^-(m)$  vastly dominates the other, depending on whether  $m > \mu$  or  $m < \mu$ .

*Remark 3.* Theorem 3.4.1 yields tight hedged CIs for a fixed sample size  $n$ . Recalling (3.26), we

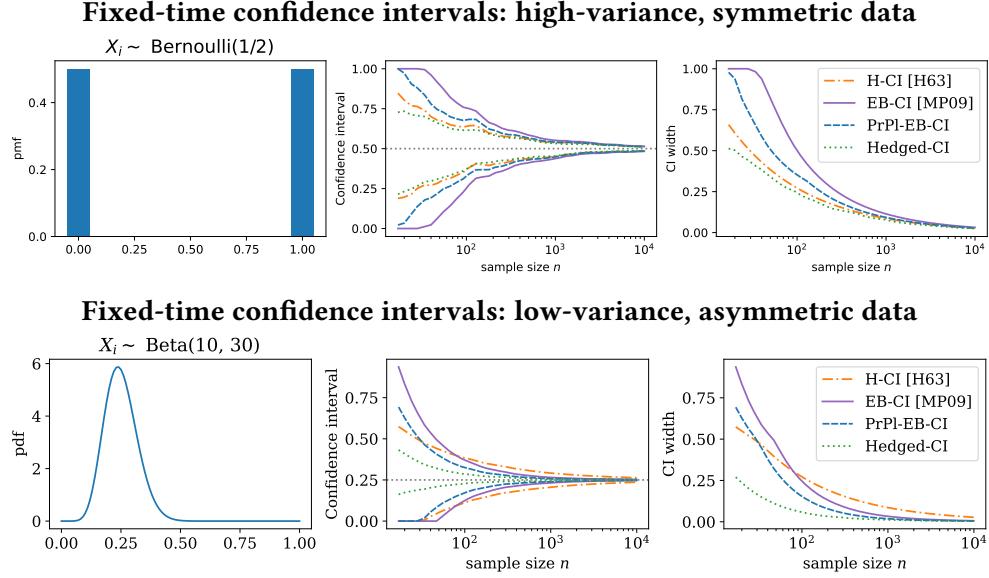


Figure 3.6: Hoeffding (H), empirical Bernstein (EB), and hedged capital CIs under two distributional scenarios. Similar to the time-uniform setting, the betting approach tends to outperform the other bounds, especially for low-variance, asymmetric data.

recommend using  $\bigcap_{i \leq n} \mathfrak{B}_i^\pm$ , and setting  $\tilde{\lambda}_t^+ = \tilde{\lambda}_t^- = \tilde{\lambda}_t^\pm$  given by

$$\tilde{\lambda}_t^\pm := \sqrt{\frac{2 \log(2/\alpha)}{n \hat{\sigma}_{t-1}^2}}. \quad (3.27)$$

We refer to the resulting CI as the ‘‘hedged capital confidence interval’’ or **[Hedged-CI]** for short, and demonstrate its superiority to past work in Figure 3.6.

Similar to the discussion after Remark 1, if  $X_1, \dots, X_n$  are independent, then one can permute the data many times and average the resulting capital processes to effectively derandomize the procedure.

The proof of Theorem 3.4.1 is in Section 3.A.5. Unlike the empirical Bernstein-type CSs and CIs of Section 3.3, those based on the hedged capital process are not necessarily symmetric. In fact, we empirically find through simulations that these CSs and CIs are able to adapt and benefit from this asymmetry (see Figures 3.4 and 3.6). While it is not obvious from the definition of  $\mathfrak{B}_t^\pm$ , bets can be chosen such that hedged capital CSs and CIs converge at the optimal rates of  $O(\sqrt{\log \log t/t})$  and  $O(1/\sqrt{n})$ , respectively (see Section 3.E.2) and such that for sufficiently large  $n$ , hedged capital CIs almost surely dominate those based on Hoeffding’s inequality (see Section 3.E.1). However, the implications of time-uniform convergence rates are subtle, and optimal rates are not always desirable in practical applications (see [125, Section 3.5]). Nevertheless, we find that hedged capital CSs and CIs significantly outperform past works even for small sample sizes (see Section 3.C). Some additional tools for visualizing CSs across  $\alpha$  and  $t$  are provided in Section 3.D.5.

In Section 3.B, we discuss some guiding principles for deriving powerful betting strategies, presenting the hedged capital CSs and CIs as special cases along with the following game-theoretic betting schemes:

- Growth rate adaptive to the particular alternative (GRAPA),
- Approximate GRAPA (aGRAPA),
- Lower-bound on the wealth (LBOW),
- Online Newton step- $m$  (ONS- $m$ ),
- Diversified Kelly betting (dKelly),
- Confidence boundary bets (ConBo), and
- Sequentially rebalanced portfolio (SRP).

Each of these betting strategies have their respective benefits, whether computational, conceptual, or statistical which are discussed further in Section 3.B.

### 3.5 Betting while sampling without replacement (WoR)

This section tackles a slightly different problem, that of sampling without replacement (WoR) from a finite set of real numbers in order to estimate its mean. Importantly, the  $N$  numbers in the finite population  $(x_1, \dots, x_N)$  are fixed and nonrandom. What is random is only the order of observation; the model for sampling uniformly at random without replacement (WoR) posits that at time  $t \geq 1$ ,

$$X_t \mid (X_1, \dots, X_{t-1}) \sim \text{Uniform}((x_1, \dots, x_N) \setminus (X_1, \dots, X_{t-1})). \quad (3.28)$$

All probabilities are thus to be understood as solely arising from observing fixed entities in a random order, with no distributional assumptions being made on the finite population. We consider the same canonical filtration  $\mathcal{F} = (\mathcal{F}_t)_{t=0}^N$  as before. For  $t \geq 1$ , let  $\mathcal{F}_t := \sigma(X_1^t)$  be the sigma-field generated by  $X_1, \dots, X_t$  and let  $\mathcal{F}_0$  be the empty sigma-field. For succinctness, we use the notation  $[a] := \{1, \dots, a\}$ .

For each  $m \in [0, 1]$ , let  $\mathcal{L}^m := \{x_1^N \in [0, 1]^N : \sum_{i=1}^N x_i/N = m\}$  be the set of all unordered lists of  $N \geq 2$  real numbers in  $[0, 1]$  whose average is  $m$ . For instance,  $\mathcal{L}^0$  and  $\mathcal{L}^1$  are both singletons, but otherwise  $\mathcal{L}^m$  is uncountably infinite. Let  $\mathcal{P}^m$  be the set of all measures on  $\mathcal{F}_N$  that are formed as follows: pick an arbitrary element of  $\mathcal{L}^m$ , apply a uniformly random permutation, and reveal the elements one by one. Thus, every element of  $\mathcal{P}^m$  is a uniform measure on the  $N!$  permutations of some element in  $\mathcal{L}^m$ , so there is a one-to-one mapping between  $\mathcal{L}^m$  and  $\mathcal{P}^m$ .

Define  $\mathcal{P} := \bigcup_m \mathcal{P}^m$  and let  $\mu$  represent the true unknown mean, meaning that the data is drawn from some  $P \in \mathcal{P}^\mu$ . For every  $m \in [0, 1]$ , we posit a composite null hypothesis  $H_m^0 : P \in \mathcal{P}^m$ , but clearly only one of these nulls is true. We will design betting strategies to test these nulls and thus find efficient confidence intervals or sequences for  $\mu$ . It is easier to present the sequential case first, since that is arguably more natural for sampling WoR, and

discuss the fixed-time case later.

### 3.5.1 Existing (super)martingale-based confidence sequences or tests

Several papers have considered estimating the mean of a finite set of nonrandom numbers when sampling WoR, often by constructing concentration inequalities [120, 232, 19, 281]. Notably, Hoeffding [120] showed that the same bound for sampling with replacement (3.2) can be used when sampling WoR. Serfling [232] improved on this bound, which was then further refined by Bardenet and Maillard [19]. While test supermartingales appeared in some of the aforementioned works, Chapter 2 identified better test supermartingales which yield explicit Hoeffding- and empirical Bernstein-type concentration inequalities and CSs for sampling WoR that significantly improved on previous bounds. Consider their exponential Hoeffding-type supermartingale,

$$M_t^{\text{H-WoR}} := \exp \left\{ \sum_{i=1}^t \left[ \lambda_i \left( X_i - \mu + \frac{1}{N-(i-1)} \sum_{j=1}^{i-1} (X_j - \mu) \right) - \psi_H(\lambda_i) \right] \right\}, \quad (3.29)$$

and their exponential empirical Bernstein-type supermartingale,

$$M_t^{\text{EB-WoR}} := \exp \left\{ \sum_{i=1}^t \left[ \lambda_i \left( X_i - \mu + \frac{1}{N-(i-1)} \sum_{j=1}^{i-1} (X_j - \mu) \right) - v_i \psi_E(\lambda_i) \right] \right\}, \quad (3.30)$$

where  $(\lambda_t)_{t=1}^N$  is any predictable  $\lambda$ -sequence (real-valued for  $M_t^{\text{H-WoR}}$ , but  $[0, 1]$ -valued for  $M_t^{\text{EB-WoR}}$ ),  $v_i = 4(X_i - \hat{\mu}_{i-1})^2$  as before, and  $\psi_H(\cdot)$  and  $\psi_E(\cdot)$  are defined as in Section 3.3. Defining  $M_0^{\text{H-WoR}} \equiv M_0^{\text{EB-WoR}} := 1$ , in Chapter 2 we prove that  $(M_t^{\text{H-WoR}})_{t=0}^N$  and  $(M_t^{\text{EB-WoR}})_{t=0}^N$  are test supermartingales with respect to  $\mathcal{F}$ , and hence can be used in Step (b) of Theorem 3.2.1.

In recent work on election audits, Stark [246] credits Harold Kaplan for proposing

$$M_t^K := \int_0^1 \prod_{i=1}^t \left( 1 + \gamma \left[ X_i \frac{1-(i-1)/N}{\mu - \sum_{j=1}^{i-1} X_j / N} - 1 \right] \right) d\gamma. \quad (3.31)$$

The ‘‘Kaplan martingale’’  $(M_t^K)_{t=0}^N$  was employed for election auditing, but it is a polynomial of degree  $t$  and is computationally expensive for large  $t$  [246].

Next, we mimic the approach of Section 3.4 to derive a capital process for sampling WoR. We then derive WoR analogues of the efficient betting strategies from Section 3.B.

### 3.5.2 The capital process for sampling without replacement

Define the predictable sequence  $(\mu_t^{\text{WoR}})_{t \in [N]}$  where

$$\mu_t^{\text{WoR}} := \mathbb{E}[X_t | \mathcal{F}_{t-1}] = \frac{N\mu - \sum_{i=1}^{t-1} X_i}{N - (t-1)}. \quad (3.32)$$

It is clear that  $\mu_t^{\text{WoR}} \in [0, 1]$ , since it is the mean of the unobserved elements of  $\{x_i\}_{i \in [N]}$ .  $(\mu_t^{\text{WoR}})_{t \in [N]}$  is unobserved since  $\mu$  is unknown, so it is helpful to define

$$m_t^{\text{WoR}} := \frac{Nm - \sum_{i=1}^{t-1} X_i}{N - (t - 1)}. \quad (3.33)$$

Now, let  $(\lambda_t(m))_{t=1}^N$  be a predictable sequence such that  $\lambda_t(m)$  is  $\left(-\frac{1}{1-m_t^{\text{WoR}}}, \frac{1}{m_t^{\text{WoR}}}\right)$ -valued. Define the *without-replacement capital process*  $\mathcal{K}_t^{\text{WoR}}(m)$ ,

$$\mathcal{K}_t^{\text{WoR}}(m) := \prod_{i=1}^t (1 + \lambda_i(m) \cdot (X_i - m_i^{\text{WoR}})) \quad (3.34)$$

with  $\mathcal{K}_0^{\text{WoR}}(m) := 1$ . The following result is analogous to Proposition 3.4.1.

**Proposition 3.5.1.** *Let  $X_1^N$  be a WoR sample from  $x_1^N \in [0, 1]^N$ . The following two statements imply each other:*

1.  $\mathbb{E}_P(X_t | \mathcal{F}_{t-1}) = \mu_t^{\text{WoR}}$  for each  $t \in [N]$ .
2. For every predictable sequence with  $\lambda_t(m) \in \left(-\frac{1}{(1-\mu_t^{\text{WoR}})}, \frac{1}{\mu_t^{\text{WoR}}}\right)$ ,  $(\mathcal{K}_t^{\text{WoR}}(\mu))_{t=0}^\infty$  is a test martingale.

The other claims within Proposition 3.4.1 also hold above with minor modification, but we do not mention them again for brevity. Further, Proposition 3.4.2 technically covers WoR sampling as well. We now present a “hedged” capital process and powerful betting schemes for sampling WoR, to construct a CS for  $\mu = \frac{1}{N} \sum_{i=1}^N x_i$ .

### 3.5.3 Powerful betting schemes

Similar to Section 3.4.4, define the hedged capital process for sampling WoR:

$$\begin{aligned} \mathcal{K}_t^{\pm, \text{WoR}}(m) := \max & \left\{ \theta \prod_{i=1}^t (1 + \lambda_i^+(m) \cdot (X_i - m_i^{\text{WoR}})) , \right. \\ & \left. (1 - \theta) \prod_{i=1}^t (1 - \lambda_i^-(m) \cdot (X_i - m_i^{\text{WoR}})) \right\} \end{aligned}$$

for some predictable  $(\lambda_t^+(m))_{t=1}^N$  and  $(\lambda_t^-(m))_{t=1}^N$  taking values in  $[0, 1/m_t^{\text{WoR}}]$  and  $[0, 1/(1 - m_t^{\text{WoR}})]$  at time  $t$ , respectively. Using  $(\mathcal{K}_t^{\pm, \text{WoR}}(m))_{t=0}^\infty$  as the process in Step (b) of Theorem 3.2.1, we obtain the CS summarized in the following theorem.

**Theorem 3.5.1** (WoR hedged capital CS [**Hedged-WoR**]). *Given a finite population  $x_1^N \in [0, 1]^N$  with mean  $\mu := \frac{1}{N} \sum_{i=1}^N x_i = \mu$ , suppose that  $X_1, X_2, \dots, X_N$  are sampled WoR from  $x_1^N$ . Let  $(\dot{\lambda}_t^+)^{\infty}_{t=1}$  and  $(\dot{\lambda}_t^-)^{\infty}_{t=1}$  be real-valued predictable sequences not depending on  $m$ , and for*

each  $t \geq 1$  let

$$\lambda_t^+(m) := |\dot{\lambda}_t^+| \wedge \frac{c}{m_t^{\text{WoR}}}, \quad \lambda_t^-(m) := |\dot{\lambda}_t^-| \wedge \frac{c}{1 - m_t^{\text{WoR}}},$$

for some  $c \in [0, 1)$  (some reasonable defaults being  $c = 1/2$  or  $3/4$ ). Then

$$\mathfrak{B}_t^{\pm, \text{WoR}} := \left\{ m \in [0, 1] : \mathcal{K}_t^{\pm, \text{WoR}}(m) < 1/\alpha \right\} \text{ forms a } (1 - \alpha)\text{-CS for } \mu,$$

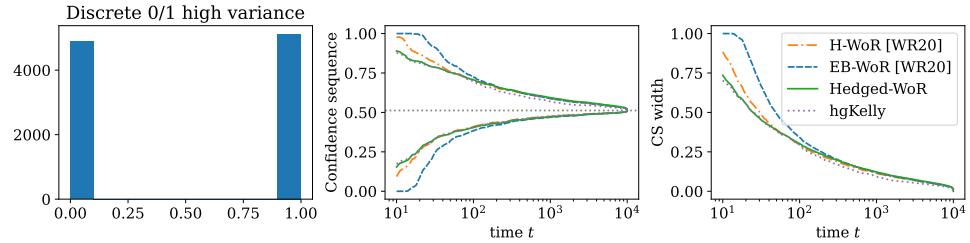
as does  $\bigcap_{i \leq t} \mathfrak{B}_i^{\pm, \text{WoR}}$ . Furthermore,  $\mathfrak{B}_t^{\pm, \text{WoR}}$  is an interval for each  $t \geq 1$ .

The proof of Theorem 3.5.1 is in Section 3.A.9. We recommend setting  $\dot{\lambda}_t^+ = \dot{\lambda}_t^- = \lambda_t^{\text{PrPl}\pm}$  as was done earlier in (3.26); for each  $t \geq 1$ , and  $c := 1/2$ , let

$$\lambda_t^{\text{PrPl}\pm} := \sqrt{\frac{2 \log(2/\alpha)}{\hat{\sigma}_{t-1}^2 \log(t+1)}}, \quad \hat{\sigma}_t^2 := \frac{\frac{1}{4} + \sum_{i=1}^t (X_i - \hat{\mu}_i)^2}{t+1}, \text{ and } \hat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1},$$

See Figure 3.7 for a comparison to the best prior work.

#### WoR time-uniform confidence sequences: high-variance, symmetric data



#### WoR time-uniform confidence sequences: low-variance, asymmetric data

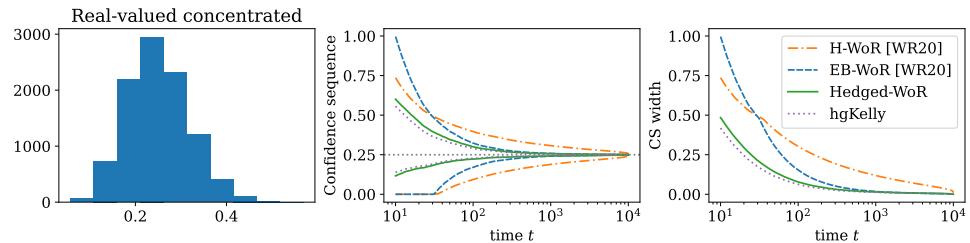


Figure 3.7: Without-replacement betting CSs versus the predictable plug-in supermartingale-based CSs (Chapter 2). Similar to the with-replacement case, the betting approach matches or vastly outperforms past state-of-the art methods.

*Remark 4.* As before, we can use Theorem 3.5.1 to derive powerful CIs for the mean of a nonrandom set of bounded numbers given a fixed sample size  $n \leq N$ . We recommend using  $\bigcap_{i \leq n} \mathfrak{B}_i^{\pm, \text{WoR}}$ , and setting  $\dot{\lambda}_t^+ = \dot{\lambda}_t^- = \dot{\lambda}_t^\pm$  as in (3.27):  $\dot{\lambda}_t^\pm := \sqrt{\frac{2 \log(2/\alpha)}{n \hat{\sigma}_{t-1}^2}}$ . We refer to the resulting CI as **[Hedged-WoR-CI]**; see Figure 3.8.

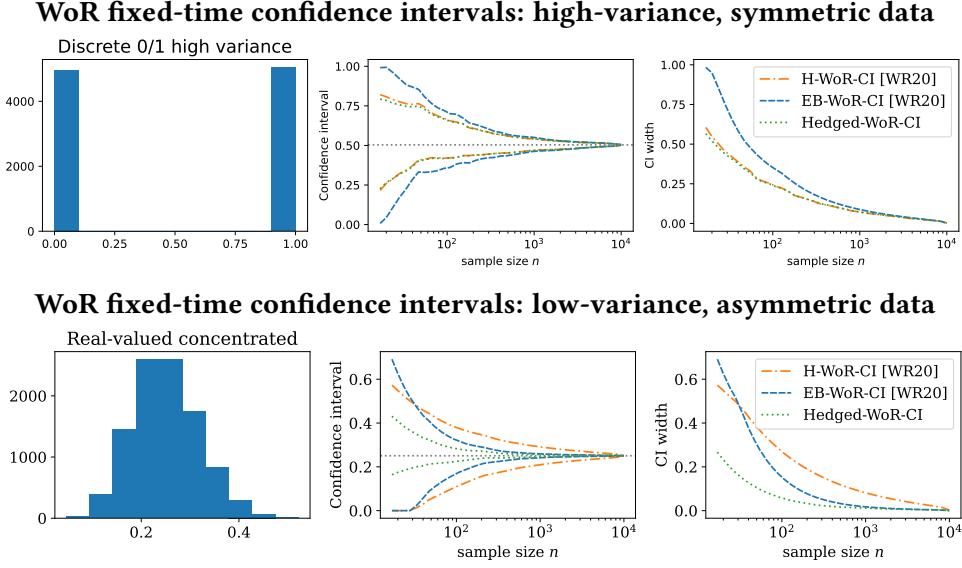


Figure 3.8: WoR analogue of the hedged capital CI versus the WoR Hoeffding- and empirical Bernstein-type CIs (Chapter 2). Similar to with-replacement, the betting approach has excellent performance.

*Remark 5.* For some values of  $m$  near 0 or 1,  $m_t^{\text{WoR}}$  could lie outside of  $[0, 1]$ , leading  $\mathcal{K}_t^{\pm, \text{WoR}}(m)$  to potentially be negative. However, it is impossible for  $\mathcal{K}_t^{\pm, \text{WoR}}(\mu)$  to be negative since  $\mu_t \in [0, 1]$  always. In fact, a negative  $m_t$  implies that the value of  $m$  being tested is impossible, and thus one can reject that null immediately. In particular, when running our method, one can instead use the modified capital process

$$\tilde{\mathcal{K}}_t^{\pm, \text{WoR}}(m) := |\mathcal{K}_t^{\pm, \text{WoR}}(m)| / \mathbb{1}(m_t \in [0, 1]) \quad (3.35)$$

which takes on the value  $+\infty$  if the denominator evaluates to zero. Note that  $\tilde{\mathcal{K}}_t^{\pm, \text{WoR}}(\mu)$  still forms a nonnegative martingale since its denominator is always one when  $m = \mu$ .

Notice that constructing a WoR test martingale only relies on changing the fixed conditional mean  $\mu$  to the time-varying conditional mean  $\mu_t^{\text{WoR}} := \frac{N\mu - \sum_{i=1}^{t-1} X_i}{N-t+1}$  and now designing  $(-1/(1-\mu_t^{\text{WoR}}), 1/\mu_t^{\text{WoR}})$ -valued bets instead of  $(-1/(1-\mu), 1/\mu)$ -valued ones. In this way, it is possible to adapt any of the betting strategies in Section 3.B to sampling WoR, yielding a wide array of solutions to this estimation problem.

### 3.5.4 Relationship to composite null testing

This chapter focuses primarily on estimation, but we end with a note that our CSs (or CIs) yield valid, sequential (or batch) tests for composite null hypotheses  $H_0 : \mu \in S$  for any  $S \subset [0, 1]$ . Specifically, for any of our capital processes  $\mathcal{K}_t(m)$ ,

$$\mathfrak{p}_t := \sup_{m \in S} \frac{1}{\mathcal{K}_t(m)}$$

is an “anytime-valid p-value” for  $H_0$ , as is  $\tilde{p}_t := \inf_{s \leq t} p_s$ , meaning that

$$\sup_{P \in \bigcup_{m \in S} \mathcal{P}^m} P(p_\tau \leq \alpha) \leq \alpha \text{ for arbitrary stopping times } \tau.$$

Alternately,  $p_t$  is also the smallest  $\alpha$  for which our  $(1 - \alpha)$ -CS does not intersect  $S$ . Similarly,  $e_t := \inf_{m \in S} \mathcal{K}_t(m)$  is an “e-process” for  $H_0$ , meaning that

$$\sup_{P \in \bigcup_{m \in S} \mathcal{P}^m} \mathbb{E}_P[e_\tau] \leq 1 \text{ for arbitrary stopping times } \tau.$$

For more details on inference at arbitrary stopping times, we refer the reader to Howard et al. [124, 125], Grünwald et al. [113], Ramdas et al. [209].

### 3.6 A brief selective history on betting and its mathematical applications

From a purely statistical perspective, this chapter could be viewed as tackling the problem of deriving sharp confidence sets for means of bounded random variables. In this pursuit, we find that a technique with excellent empirical performance happens to have strong connections to the topics of betting and gambling. While we provide a more detailed discussion in Section 3.F, here we briefly summarize some of the ways in which betting ideas have appeared in and shaped probability, statistical inference, information theory, and online learning, in the broad context of this chapter.

- **Probability:** The 1939 PhD thesis of [264] brought betting and martingales to the forefront of modern probability theory, by giving actionable interpretations to Kolmogorov’s newly developed measure-theoretic probability, and dealing a near-fatal blow to the theory of collectives by von Mises. Ville showed that for *any* event  $A$  of probability measure zero (like sequences violating the law of large numbers), he could design an explicit betting strategy that never bets more than it has, whose wealth (a test martingale) grows to infinity if the event  $A$  occurs. Ville worked with binary sequences, but his result holds more generally; see [235].

One may view Ville’s result as a theorem in measure-theoretic probability theory; what he effectively proved was: the event that a test (super)martingale exceeds  $1/\alpha$  has probability at most  $\alpha$  (Ville’s inequality in this chapter). This holds for any  $\alpha \in [0, 1]$ , treating  $1/0 \equiv +\infty$ , with the  $\alpha = 0$  case being the most remarkable part. But Ville’s result is also an axiomatic building block for *game-theoretic probability* [268, 235, 236]. Many classical results in probability can be derived in completely game-theoretic terms [235, 236]. The capital processes used for deriving CSs are of the same form as those used to derive these foundational theorems of game-theoretic probability, despite the two goals being quite different.

- **Statistical inference:** The famous book of Wald [272] was the first to thoroughly present and study sequential hypothesis testing. Despite not being presented in this way by Wald,

we know in hindsight that the sequential probability ratio test (SPRT) is quite centrally based on the fact that the likelihood ratio is a nonnegative martingale. Two decades later, Robbins and colleagues built on Wald’s sequential testing work in several ways, including to estimation via confidence sequences [73, 74, 75, 218, 219, 220, 221, 222, 217, 168]. The recent work of Howard et al. [124, 125], Ramdas et al. [210], Wasserman et al. [280] extends the early work of Wald, Robbins and colleagues to a broader class of problems using exponential supermartingales and “e-processes”, which can be seen as nonparametric, composite generalizations of the SPRT martingale. Connections between *betting* and the works of Wald, Robbins et al., and Howard et al. are implicit in those works, but can now be seen in hindsight, and this chapter makes these connections explicit.

- **Information theory:** Working in the new field of information theory, Kelly Jr [152] made direct connections to betting by showing that the capacity of a channel (itself fundamentally related to entropy and the Kullback-Leibler divergence) is given by the maximal rate of growth of wealth of a gambler in a simple game with iid  $\text{Bernoulli}(p)$  observations and known  $p$ . [41] generalized Kelly’s results significantly, and Krichevsky and Trofimov [163] extended these results beyond the case of known  $p$  using a mixture method. Thomas Cover’s interest in these techniques spans several decades [67, 68, 69, 24, 23], culminating in his famous universal portfolio algorithm [70]. The results of Krichevsky-Trofimov and Cover are essentially regret inequalities, leading directly to the final subfield below.
- **Online learning:** The techniques of Krichevsky, Trofimov and Cover found extensive applications to *sequential prediction with the logarithmic loss* [50]. Here, one derives *regret inequalities* for the total loss accumulated when predicting the next observation from a potentially adversarial sequence. This problem is fundamentally connected to online convex optimization, for which Orabona and colleagues use parameter-free betting algorithms to derive regret inequalities [196, 197, 140, 72, 139]. Rakhlin and Sridharan [208] articulated a deep connection between martingale concentration and deterministic regret inequalities, and Jun and Orabona [139, Section 7.1] derive concentration bounds for the general setting of Banach space-valued observations with sub-exponential noise.

### 3.7 Summary

Nonparametric confidence sequences are particularly useful in sequential estimation because they enable valid inference at arbitrary stopping times, but they are underappreciated as powerful tools to provide accurate inference even at fixed times. Recent work [124, 125] has developed several time-uniform generalizations of the Cramer-Chernoff technique utilizing “line-crossing” inequalities and using various variants of Robbins’ method of mixtures (discrete mixtures, conjugate mixtures and stitching) to convert them to “curve-crossing” inequalities.

This work adds new techniques to the toolkit: to complement the aforementioned mixture methods, we develop a “predictable plug-in” approach. When coupled with existing nonparametric supermartingales, it yields (for example) computationally efficient empirical-Bernstein

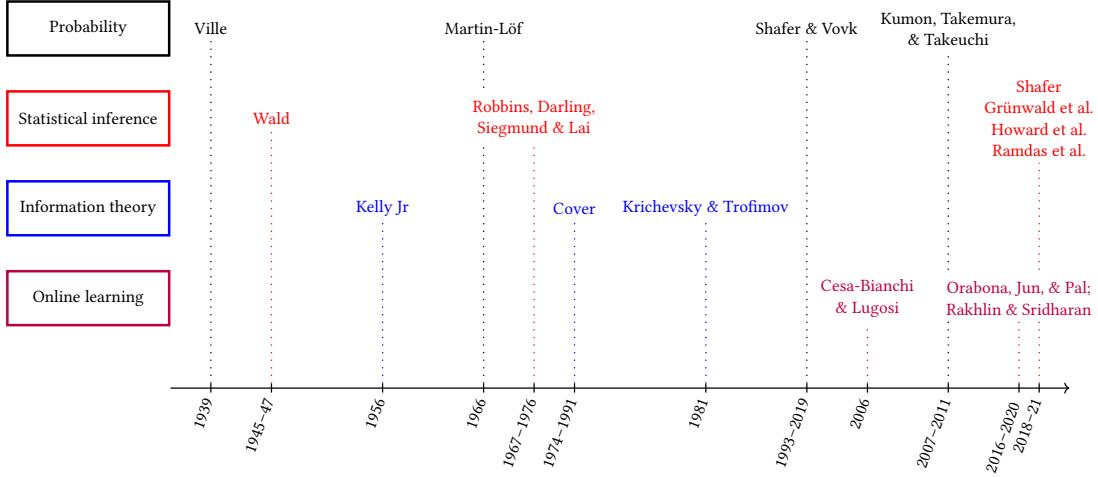


Figure 3.9: A brief selective history of betting ideas appearing (often implicitly) in various literatures. As discussed further in Section 3.F, these subfields have rarely cited each other, but ideas are now beginning to permeate. Several authors did not explicitly use the language of betting, and their inclusion above is due to reinterpreting their work in hindsight.

confidence sequences. One of our major contributions is to thoroughly develop the theory and methodology for a new nonnegative martingale approach to estimating means of bounded random variables in both with- and without-replacement settings. These convincingly outperform all existing published work that we are aware of, for CIs and CSs, both with and without replacement.

Our methods are particularly easy to interpret in terms of evolving capital processes and sequential testing by betting [234] but we go much further by developing powerful and efficient betting strategies that lead to state-of-the-art variance-adaptive confidence sets that are significantly tighter than past work in all considered settings. In particular, Shafer espouses *complementary* benefits of such approaches, ranging from improved scientific communication, ties to historical advances in probability, and reproducibility via continued experimentation (also see [113]), but our focus here has been on developing a new state of the art for a set of classical, fundamental problems.

There appear to be nontrivial connections to online learning theory [162, 166, 197, 72], and to empirical and dual likelihoods (see Section 3.E.6 and an extended historical review of betting in Section 3.F). The reductions from regret inequalities to concentration bounds described in [208] and [139] are fascinating, but existing published bounds are loose in the constants and not competitive in practice compared to our direct approach. Exploring deeper connections may yield other confidence sequences or betting strategies.

It is clear to us, and hopefully to the reader as well, that the ideas behind this work (adaptive statistical inference by betting) form the tip of the iceberg—they lead to powerful, efficient, nonasymptotic, nonparametric inference and can be adapted to a range of other problems.

As just one example, let  $\mathcal{P}^{p,q}$  represent the set of all continuous distributions such that the  $p$ -quantile of  $X_t$ , conditional on the past, is equal to  $q$ . This is also a nonparametric, convex set of distributions with no common reference measure. Nevertheless, for any predictable  $(\lambda_i)$ , it is easy to check that

$$M_t = \prod_{i=1}^t (1 + \lambda_i(\mathbf{1}_{X_i \leq q} - p))$$

is a test martingale for  $\mathcal{P}^{p,q}$ . Setting  $p = 1/2$  and  $q = 0$ , for example, we can sequentially test if the median of the underlying data distribution is the origin. The continuity assumption can be relaxed, and this test can be inverted to get a confidence sequence for any quantile. We do not pursue this idea further in the current chapter because the recent (rather different) nonnegative martingale methods of Howard and Ramdas [123] already provide a challenging benchmark for that problem. Typically, one test martingale-based method cannot uniformly dominate another, and the large gains in this chapter were made possible because all previous published approaches implicitly or explicitly employed test *supermartingales*, while we employ test martingales that are computationally simple to implement.

To conclude, we opine that “game-theoretic statistical inference” is in its nascent stage, and we expect much theoretical and practical progress in coming years. We hope the reader shares our excitement in this regard.

### 3.A Proofs of main results

We first introduce a lemma which will aid in the proofs to follow.

**Lemma 3.A.1** (Predictable plug-in Chernoff supermartingales). *Suppose that  $X_1, X_2, \dots \sim P$ , and for some  $\mu, v_t$  and  $\psi(\lambda)$ , we have that for any  $\lambda \in \Lambda \subseteq \mathbb{R}$ ,*

$$\mathbb{E}_P [\exp(\lambda(X_t - \mu) - v_t\psi(\lambda)) \mid \mathcal{F}_{t-1}] \leq 1 \quad \text{for each } t \geq 1. \quad (3.36)$$

*Then, for any  $\Lambda$ -valued sequence  $(\lambda_t)_{t=1}^\infty$  that is predictable with respect to  $\mathcal{F}$ ,*

$$M_t^\psi(\mu) := \prod_{i=1}^t \exp(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i))$$

*forms a test supermartingale with respect to  $\mathcal{F}$ .*

*Proof.* Writing out the conditional expectation of  $M_t^\psi$  for any  $t \geq 2$ ,

$$\begin{aligned} \mathbb{E}(M_t^\psi(\mu) \mid \mathcal{F}_{t-1}) &= \mathbb{E}\left(\prod_{i=1}^t \exp(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)) \mid \mathcal{F}_{t-1}\right) \\ &\stackrel{(i)}{=} \prod_{i=1}^{t-1} \exp(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i)) \underbrace{\mathbb{E}[\exp(\lambda_t(X_t - \mu) - v_t\psi(\lambda_t)) \mid \mathcal{F}_{t-1}]}_{\leq 1 \text{ by assumption}} \\ &= M_{t-1}^\psi(\mu), \end{aligned}$$

where (i) follows from the fact that  $\exp(\lambda_i(X_i - \mu) - v_i\psi(\lambda_i))$  is  $\mathcal{F}_{t-1}$ -measurable for  $i \leq t-1$ . Since  $\mathcal{F}_0$  was assumed to be trivial, for  $M_1$  we have that

$$\mathbb{E}[M_1^\psi(\mu) \mid \mathcal{F}_0] = \underbrace{\mathbb{E}[\exp(\lambda_1(X_1 - \mu) - v_1\psi(\lambda_1))]}_{\leq 1 \text{ by assumption}},$$

which completes the proof. □

#### 3.A.1 Proof of Proposition 3.3.1

The proof proceeds in three steps. First, apply a standard MGF bound by Hoeffding [120]. Second, we apply Lemma 3.A.1. Finally, we apply Theorem 3.2.1 to obtain a CS and take a union bound.

**Step 1.** By Hoeffding [120], we have that  $\mathbb{E}[\exp(\lambda_t(X_t - \mu) - \psi_H(\lambda_t)) \mid \mathcal{F}_{t-1}] \leq 1$  since  $X_t \in [0, 1]$  almost surely and since  $\lambda_t$  is  $\mathcal{F}_{t-1}$ -measurable.

**Step 2.** By Step 1 and Lemma 3.A.1, we have that

$$M_t^{\text{PrPl-H}}(\mu) := \prod_{i=1}^t \exp(\lambda_i(X_i - \mu) - \psi_H(\lambda_i))$$

forms a test supermartingale.

**Step 3.** By Step 2 combined with Theorem 3.2.1, we have that

$$\begin{aligned} P\left(\exists t \geq 1 : \mu \leq \frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\alpha) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right) \\ = P\left(\exists t \geq 1 : M_t^{\text{PrPl-H}}(\mu) \geq 1/\alpha\right) \leq \alpha. \end{aligned}$$

Applying the same bound to  $(-X_t)_{t=1}^\infty$  with mean  $-\mu$  and taking a union bound, we have the desired result,

$$P\left(\exists t \geq 1 : \mu \notin \left(\frac{\sum_{i=1}^t \lambda_i X_i}{\sum_{i=1}^t \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right)\right) \leq \alpha,$$

which completes the proof.  $\square$

### 3.A.2 Proof of Theorem 3.3.1

By Lemma 3.A.1 combined with Theorem 3.2.1, it suffices to prove that

$$\mathbb{E}_P [\exp\{\lambda_t(X_t - \mu) - v_t \psi_E(\lambda_t)\} \mid \mathcal{F}_{t-1}] \leq 1.$$

For succinctness, denote

$$Y_t := X_t - \mu \quad \text{and} \quad \delta_t := \hat{\mu}_t - \mu.$$

Note that  $\mathbb{E}_P(Y_t \mid \mathcal{F}_{t-1}) = 0$ . It then suffices to prove that for any  $[0, 1)$ -bounded,  $\mathcal{F}_{t-1}$ -measurable  $\lambda_t \equiv \lambda_t(X_1^{t-1})$ ,

$$\mathbb{E}\left[\exp\left\{\lambda_t Y_t - 4(Y_t - \delta_{t-1})^2 \psi_E(\lambda_t)\right\} \mid \mathcal{F}_{t-1}\right] \leq 1.$$

Indeed, in the proof of Proposition 4.1 in Fan et al. [102],  $\exp\{\xi\lambda - 4\xi^2\psi_E(\lambda)\} \leq 1 + \xi\lambda$  for any  $\lambda \in [0, 1)$  and  $\xi \geq -1$ . Setting  $\xi := Y_t - \delta_{t-1} = X_t - \hat{\mu}_{t-1}$ ,

$$\begin{aligned} & \mathbb{E}\left[\exp\left\{\lambda_t Y_t - 4(Y_t - \delta_{t-1})^2 \psi_E(\lambda_t)\right\} \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[\exp\left\{\lambda_t(Y_t - \delta_{t-1}) - 4(Y_t - \delta_{t-1})^2 \psi_E(\lambda_t)\right\} \mid \mathcal{F}_{t-1}\right] \exp(\lambda_t \delta_{t-1}) \end{aligned}$$

$$\leq \mathbb{E}\left[1 + (Y_t - \delta_{t-1})\lambda_t \mid \mathcal{F}_{t-1}\right] \exp(\lambda_t \delta_{t-1}) \stackrel{(i)}{=} \mathbb{E}\left[1 - \delta_{t-1}\lambda_t \mid \mathcal{F}_{t-1}\right] \exp(\lambda_t \delta_{t-1}) \stackrel{(ii)}{\leq} 1,$$

where equality (i) follows from the fact that  $Y_t$  is conditionally mean zero, and inequality (ii) follows from the inequality  $1 - x \leq \exp(-x)$  for all  $x \in \mathbb{R}$ . This completes the proof.  $\square$

### 3.A.3 Proof of Proposition 3.4.1

We proceed by proving  $(d) \implies (c) \implies (b) \implies (a) \implies (d)$ .

**Proof of (d)  $\implies$  (c).** This claim follows from the fact that for  $\lambda \in (-1/(1-\mu), 1/\mu)$ , we have that  $(\lambda, \lambda, \dots)$  is a  $(-1/(1-\mu), 1/\mu)$ -valued predictable sequence.

**Proof of (c)  $\implies$  (b).** By the assumption of (c), we have that for  $\lambda = 0.5$ ,  $\mathcal{K}_t(\mu)$  forms a test martingale. Furthermore, since  $X_i, \mu \in [0, 1]$  for each  $i \in \{1, 2, \dots\}$ , we have that  $1 + 0.5(X_i - \mu) > 0$  almost surely for each  $i$ . Therefore,  $(\mathcal{K}_t(\mu))_{t=1}^\infty$  is a strictly positive test martingale.

**Proof of (b)  $\implies$  (a).** Suppose that there exists  $\lambda \in \mathbb{R} \setminus \{0\}$  such that  $\mathcal{K}_t(\mu)$  forms a strictly positive martingale. Then we must have

$$\begin{aligned}\mathcal{K}_{t-1}(\mu) &= \mathbb{E}(\mathcal{K}_t(\mu) \mid \mathcal{F}_{t-1}) \\ &= \mathcal{K}_{t-1}(\mu) \cdot \mathbb{E}(1 + \lambda(X_i - \mu) \mid \mathcal{F}_{t-1}) \\ &= \mathcal{K}_{t-1}(\mu) \cdot [1 + \lambda(\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu)].\end{aligned}$$

Now since  $\mathcal{K}_{t-1}(\mu) > 0$ , we have that

$$1 + \lambda(\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu) = 1.$$

Since  $\lambda \neq 0$  by assumption, we have that  $\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = \mu$  as required.

**Proof of (a)  $\implies$  (d).** Let  $(\lambda_t(\mu))_{t=1}^\infty$  be a  $(-1/(1-\mu), 1/\mu)$ -valued predictable sequence. Then  $\mathcal{K}_t(\mu)$  is clearly nonnegative and  $\mathcal{K}_0(\mu) = 1$  by definition. Writing out the conditional mean of the capital process for any  $t \geq 1$ ,

$$\begin{aligned}\mathbb{E}(\mathcal{K}_t(\mu) \mid \mathcal{F}_{t-1}) &= \mathcal{K}_{t-1}(\mu) \cdot \mathbb{E}(1 + \lambda_t(\mu)(X_i - m) \mid \mathcal{F}_{t-1}) \\ &= \mathcal{K}_{t-1}(\mu) \cdot [1 + \lambda_t(\mu)(\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu)] \\ &= \mathcal{K}_{t-1}(\mu),\end{aligned}$$

and thus  $\mathcal{K}_t(\mu)$  forms a test martingale.

The proof of the final part of the proposition is simple. Let  $(M_t)$  be a test martingale for  $\mathcal{P}^\mu$ . Define  $Y_t := M_t/M_{t-1}$  if  $M_{t-1} > 0$ , and as  $Y_t := 0$  otherwise. Now note that  $M_t = \prod_{i=1}^t Y_i$  and  $\mathbb{E}_P[Y_t \mid \mathcal{F}_{t-1}] = 1$  for any  $P \in \mathcal{P}^\mu$ . In other words, every test martingale is a product of nonnegative random variables with conditional mean one. Now rewrite  $Y_t$  as  $(1 + f_t(X_t))$  for

some predictable function  $f_t$ . Since  $Y_t$  is nonnegative, we must have  $f_t(X_t) \geq -1$ , and since  $Y_t$  is conditional mean one, we must have  $f_t(X_t)$  is conditional mean zero. Such a representation in fact holds true for any test martingale, and we have not yet used the fact that we are working with test martingales for  $\mathcal{P}^\mu$ . Now, the proof ends by noting that the only predictable functions  $f_t$  with the latter property under every  $P \in \mathcal{P}^\mu$  has the form  $\lambda_t(X_t - \mu)$  for some predictable  $\lambda_t$ ; any nonlinear function of  $X_t$  would not have mean zero under *every distribution* with mean  $\mu$ .

This completes the proof of Proposition 3.4.1 altogether.  $\square$

### 3.A.4 Proof of Proposition 3.4.2

We only prove the martingale part of the proposition, since the supermartingale aspect follows analogously, and as mentioned early in the chapter, inequalities and equalities are meant in an almost sure sense.

First, it is easy to check that if  $(M_t)$  is a test martingale for  $\mathcal{S}$ , then  $M_t$  is the product of nonnegative conditionally unit mean terms, that is  $M_t = \prod_{i=1}^t Y_i$  such that for all  $S \in \mathcal{S}$ , we have  $\mathbb{E}_S[Y_i | \mathcal{F}_{i-1}] = 1$  and  $Y_i \geq 0$ . (Indeed, one can identify  $Y_i := \frac{M_i}{M_{i-1}} \mathbf{1}_{M_{i-1} > 0}$ .) Now, define  $Z'_i := Y_i - 1$ , and note that  $Z'_i \geq -1$ , and  $\mathbb{E}_S[Z'_i | \mathcal{F}_{i-1}] = 0$ . Thus,  $M_t$  has been represented as  $\prod_{i=1}^t (1 + Z'_i)$ . Now, the proof is completed by noting that any such  $Z'_i$  can be written as  $\lambda_i Z_i$  for a predictable  $\lambda_i$  (this step is purely cosmetic).  $\square$

### 3.A.5 Proof of Theorem 3.4.1

First, we present Lemma 3.A.2 which establishes that the hedged capital process is a quasiconvex function of  $m$  (and thus has convex sublevel sets). We then invoke this lemma to prove the main result.

**Lemma 3.A.2.** *Let  $\theta \in [0, 1]$  and*

$$\begin{aligned} \mathcal{K}_t^\pm(m) &:= \max \{ \theta \mathcal{K}_t^+(m), (1 - \theta) \mathcal{K}_t^-(m) \} \\ &\equiv \max \left\{ \theta \prod_{i=1}^t (1 + \lambda_i^+(m) \cdot (X_i - m)), (1 - \theta) \prod_{i=1}^t (1 - \lambda_i^-(m) \cdot (X_i - m)) \right\} \end{aligned}$$

*be the hedged capital process as in Section 3.4. Consider the  $(1 - \alpha)$  confidence set of the same theorem,*

$$\mathfrak{B}_t^\pm \equiv \mathfrak{B}^\pm(X_1, \dots, X_t) := \left\{ m \in [0, 1] : \mathcal{K}_t^\pm(m) < \frac{1}{\alpha} \right\}.$$

*Then  $\mathfrak{B}_t^\pm$  is an interval on  $[0, 1]$ .*

*Proof.* Since sublevel sets of quasiconvex functions are convex, it suffices to prove that  $\mathcal{K}_t^\pm(m)$  is a quasiconvex function of  $m \in [0, 1]$ . The crux of the argument is: the product of nonnegative nonincreasing functions is quasiconvex, the product of nonnegative nondecreasing functions is also quasiconvex, and the maximum of quasiconvex functions is quasiconvex.

To elaborate, we will proceed in two steps. First, we use an induction argument to show that  $\mathcal{K}_t^+(m)$  and  $\mathcal{K}_t^-(m)$  are nonincreasing and nondecreasing, respectively, and hence quasiconvex. Finally, we note that  $\mathcal{K}_t^\pm(m) := \max\{\theta\mathcal{K}_t^+(m), (1-\theta)\mathcal{K}_t^-(m)\}$  is a maximum of quasiconvex functions and is thus itself quasiconvex.

**Step 1.** First, since  $\dot{\lambda}_t^+$  does not depend on  $m$ , we have that

$$1 + \lambda_t^+(m)(X_t - m) := 1 + \left(|\dot{\lambda}_t^+| \wedge \frac{c}{m}\right)(X_t - m)$$

is nonnegative and nonincreasing in  $m$  for each  $t \in \{1, 2, \dots\}$ . (To see this, consider the terms with and without truncation separately.) Suppose for the sake of induction that

$$\prod_{i=1}^{t-1} (1 + \lambda_i^+(m)(X_i - m))$$

is nonnegative and nonincreasing in  $m$ . Then,

$$\begin{aligned} \mathcal{K}_t^+(m) &:= \prod_{i=1}^t (1 + \lambda_i^+(m)(X_i - m)) \\ &= (1 + \lambda_t^+(m)(X_t - m)) \cdot \prod_{i=1}^{t-1} (1 + \lambda_i^+(m)(X_i - m)) \end{aligned}$$

is a product of nonnegative and nonincreasing functions, and is thus itself nonnegative and nonincreasing. By a similar argument,  $\mathcal{K}_t^-(m)$  is nonnegative and *nondecreasing*.  $\mathcal{K}_t^+(m)$  and  $\mathcal{K}_t^-(m)$  are thus both quasiconvex.

**Step 2.** Since the maximum of quasiconvex functions is quasiconvex, we infer that

$$\mathcal{K}_t^\pm(m) := \max\{\theta\mathcal{K}_t^+(m), (1-\theta)\mathcal{K}_t^-(m)\}$$

is quasiconvex. In particular, the sublevel sets of quasiconvex functions is convex, and thus

$$\mathfrak{B}_t^\pm := \left\{m \in [0, 1] : \mathcal{K}_t^\pm(m) < \frac{1}{\alpha}\right\}$$

is an interval, which completes the proof of Lemma 3.A.2.  $\square$

*Proof of Theorem 3.4.1.* The proof proceeds in three steps. First we show that  $\mathcal{K}_t^\pm(\mu)$  is upper-bounded by test martingale. Second, we apply the 4-step procedure in Theorem 3.2.1 to get a CS for  $\mu$ . Third and finally, we invoke Lemma 3.A.2 to conclude that the CS is indeed convex at each time  $t$ .

**Step 1.** We first upper bound  $\mathcal{K}_t^\pm(m)$  as follows:

$$\begin{aligned}\mathcal{K}_t^\pm(m) &:= \max\{\theta\mathcal{K}_t^+(m), (1-\theta)\mathcal{K}_t^-(m)\} \\ &\leq \theta\mathcal{K}_t^+(m) + (1-\theta)\mathcal{K}_t^-(m) =: \mathcal{M}_t^\pm(m).\end{aligned}$$

By Proposition 3.4.1, we have that  $\mathcal{K}_t^+(\mu)$  and  $\mathcal{K}_t^-(\mu)$  are test martingales for  $\mathcal{P}$ . For each  $P \in \mathcal{P}$ , writing out the conditional expectation of  $\mathcal{M}_t^\pm(\mu)$  for any  $t \geq 1$ ,

$$\begin{aligned}\mathbb{E}_P[\mathcal{M}_t^\pm(\mu) | \mathcal{F}_{t-1}] &= \mathbb{E}_P\left[\theta\mathcal{K}_t^+(\mu) + (1-\theta)\mathcal{K}_t^-(\mu) \mid \mathcal{F}_{t-1}\right] \\ &= \theta\mathbb{E}_P(\mathcal{K}_t^+(\mu) | \mathcal{F}_{t-1}) + (1-\theta)\mathbb{E}_P(\mathcal{K}_t^-(\mu) | \mathcal{F}_{t-1}) \\ &= \theta\mathcal{K}_{t-1}^+(\mu) + (1-\theta)\mathcal{K}_{t-1}^-(\mu) \\ &= \mathcal{M}_{t-1}^\pm(\mu),\end{aligned}$$

and  $\mathcal{M}_0^\pm(\mu) = \theta\mathcal{K}_0^+(\mu) + (1-\theta)\mathcal{K}_0^-(\mu) = 1$ . Therefore,  $(\mathcal{M}_t^\pm(\mu))_{t=0}^\infty$  is a test martingale for  $\mathcal{P}$ .

**Step 2.** By Step 1 combined with Theorem 3.2.1 we have that

$$\mathfrak{B}_t^\pm := \left\{ m \in [0, 1] : \mathcal{K}_t^\pm(m) < \frac{1}{\alpha} \right\}$$

forms a  $(1 - \alpha)$ -CS for  $\mu$ .

**Step 3.** Finally, by Lemma 3.A.2, we have that  $\mathfrak{B}_t^\pm$  is an interval for each  $t \in \{1, 2, \dots\}$ , which completes the proof of Theorem 3.4.1.  $\square$

### 3.A.6 Proof of Lemma 3.B.1

Following the proof of Lemma 4.1 in Fan et al. [102], we have that the function

$$f(x) := \begin{cases} \frac{\log(1+x) - x}{x^2/2} & x \in (-1, \infty) \setminus \{0\} \\ -1 & x = 0 \end{cases} \quad (3.37)$$

is an increasing and continuous function in  $x$  (note that  $f(0)$  is defined as  $-1$  because it is a removable singularity). For any  $y \geq -m$  and  $\lambda \in [0, 1/m]$  we have

$$\lambda y \geq -m\lambda > -1. \quad (3.38)$$

Combining (3.37) and (3.38), we have

$$\frac{\log(1 + \lambda y) - \lambda y}{\lambda^2 y^2/2} \geq \frac{\log(1 - m\lambda) + m\lambda}{\lambda^2 m^2/2},$$

$$\text{and thus, } \log(1 + \lambda y) - \lambda y \stackrel{(i)}{\geq} \frac{y^2}{m^2} (\log(1 - m\lambda) + m\lambda).$$

Above, (i) can be quickly verified for the case when  $\lambda y = 0$ , and follows from (3.37) and (3.38) otherwise. Rearranging terms, we obtain the first half of the desired result,

$$\log(1 + \lambda y) \geq \lambda y + \frac{y^2}{m^2} (\log(1 - m\lambda) + m\lambda). \quad (3.39)$$

Now, for any  $y \leq 1 - m$  and  $\lambda \in (-1/(1-m), 0]$ , we have

$$\lambda y \geq (1 - m)\lambda > -1,$$

and proceed similarly to before to obtain

$$\log(1 + \lambda y) \geq \lambda y + \frac{y^2}{(1-m)^2} (\log(1 + (1-m)\lambda) - (1-m)\lambda),$$

which completes the proof.  $\square$

### 3.A.7 Proof of Proposition 3.B.1

Since sublevel sets of convex functions are convex, it suffices to prove that with probability one,  $\mathcal{K}_n^{\text{hgKelly}}(m)$  is a convex function in  $m$  on the interval  $[0, 1]$ .

We proceed in three steps. First, we show that if two functions are (a) both nonincreasing (or both nondecreasing), (b) nonnegative, and (c) convex, then their product is convex. Second, we use Step 1 and an induction argument to prove that  $\prod_{i=1}^t (1 + \gamma(X_i/m - 1))$  is convex for any fixed  $\gamma \in [0, 1]$ . Third and finally, we show that  $\mathcal{K}_n^{\text{hgKelly}}(m)$  is a convex combination of convex functions and is thus itself convex.

**Step 1.** The claim is that if two functions  $f$  and  $g$  are (a) both nonincreasing (or both nondecreasing), (b) nonnegative, and (c) convex on a set  $\mathcal{S} \subseteq \mathbb{R}$ , then their product is also convex on  $\mathcal{S}$ . Let  $x_1, x_2 \in \mathcal{S}$ , and let  $t \in [0, 1]$ . Furthermore, abbreviate  $f(x_1)$  by  $f_1$ ,  $g(x_1)$  by  $g_1$ , and similarly for  $f_2$  and  $g_2$ . Writing out the product  $fg$  evaluated at  $tx_1 + (1-t)x_2$ ,

$$\begin{aligned} (fg)(tx_1 + (1-t)x_2) &= f(tx_1 + (1-t)x_2)g(tx_1 + (1-t)x_2) \\ &= |f(tx_1 + (1-t)x_2)| |g(tx_1 + (1-t)x_2)| \\ &\leq |tf_2 + (1-t)f_1| |tg_2 + (1-t)g_1| \\ &= t^2 f_1 g_1 + t(1-t) (f_1 g_2 + f_2 g_1) + (1-t)^2 f_2 g_2, \end{aligned}$$

where the second equality follows from assumption that  $f$  and  $g$  are nonnegative, and the inequality follows from the assumption that they are both convex. To show convexity of  $(fg)$ , it then suffices to show that,

$$(tf_1 g_1 + (1-t)f_2 g_2) - (t^2 f_1 g_1 + t(1-t) [f_1 g_2 + f_2 g_1] + (1-t)^2 f_2 g_2) \geq 0. \quad (3.40)$$

To this end, write out the above expression and group terms,

$$\begin{aligned}
& \left( t f_1 g_1 + (1-t) f_2 g_2 \right) - \left( t^2 f_1 g_1 + t(1-t) [f_1 g_2 + f_2 g_1] + (1-t)^2 f_2 g_2 \right) \\
&= (1-t) t f_1 g_1 + t(1-t) f_2 g_2 - t(1-t) [f_1 g_2 + f_2 g_1] \\
&= t(1-t) \left( f_1 g_1 + f_2 g_2 - f_1 g_2 - f_2 g_1 \right) \\
&= t(1-t)(f_1 - f_2)(g_1 - g_2).
\end{aligned}$$

Now, notice that  $t(1-t) \geq 0$  since  $t \in [0, 1]$  and that  $(f_1 - f_2)(g_1 - g_2) \geq 0$  by the assumption that  $f$  and  $g$  are both nonincreasing or nondecreasing. Therefore, we have satisfied the inequality in (3.40), and thus  $fg$  is convex on  $\mathcal{S}$ .

**Step 2.** Now, we prove convexity of  $\prod_{i=1}^t (1 + \gamma(X_i/m - 1))$  for a fixed  $\gamma \in [0, 1]$ . First note that for any  $\gamma \in [0, 1]$ ,  $1 + \gamma(X_i/m - 1)$  is a nonincreasing, nonnegative, and convex function in  $m \in [0, 1]$ . Suppose for the sake of induction that conditions (a), (b), and (c) hold for  $\prod_{i=1}^{n-1} (1 + \gamma(X_i/m - 1))$ . By the inductive hypothesis, we have that

$$\prod_{i=1}^n (1 + \gamma(X_i/m - 1)) = (1 + \gamma(X_n/m - 1)) \cdot \prod_{i=1}^{n-1} (1 + \gamma(X_i/m - 1))$$

is a product of functions satisfying (a) through (c). By Step 1,  $\prod_{i=1}^n (1 + \gamma(X_i/m - 1))$  is convex in  $m \in [0, 1]$ . A similar argument can be made for  $\mathcal{K}_n^-(m)$ , but instead of the multiplicands being nonincreasing, they are now nondecreasing.

**Step 3.** Now, notice that for the evenly-spaced points  $(\lambda^{1+}, \dots, \lambda^{G+})$  on  $[0, 1/m]$ , we have that  $(\gamma^{1+}, \dots, \gamma^{G+}) = (m\lambda^{1+}, \dots, m\lambda^{G+})$  are  $G$  evenly-spaced points on  $[0, 1]$ . It then follows that for any  $m$  and any  $g \in \{0, 1, \dots, G\}$ ,

$$m \mapsto \prod_{i=1}^n (1 + \lambda^{g+}(X_i - m))$$

is a nonincreasing, nonnegative, and convex function in  $m \in [0, 1]$ . It follows that

$$\frac{1}{G} \sum_{g=1}^G \prod_{i=1}^n (1 + \lambda^{g+}(X_i - m))$$

is convex in  $m \in [0, 1]$ . A similar argument goes through for  $\frac{1}{G} \sum_{g=1}^G \prod_{i=1}^n (1 + \lambda^{g-}(X_i - m))$ . Finally, since  $\theta \in [0, 1]$ , we have that

$$\frac{\theta}{G} \sum_{g=1}^G \prod_{i=1}^n (1 + \lambda^{g+}(X_i - m)) + \frac{1-\theta}{G} \sum_{g=1}^G \prod_{i=1}^n (1 + \lambda^{g-}(X_i - m))$$

is a convex combination of convex functions in  $m \in [0, 1]$ . It then follows that

$$\{m \in [0, 1] : \mathcal{K}_t^{\text{hgKelly}}(m) < 1/\alpha\}$$

is an interval, which completes the proof.  $\square$

### 3.A.8 Proof of Proposition 3.5.1

**Proof of (1)  $\implies$  (2).** By definition of  $\mathcal{K}_t^{\text{WoR}}(\mu)$ , we have

$$\begin{aligned} \mathbb{E}(\mathcal{K}_t^{\text{WoR}}(\mu) \mid \mathcal{F}_{t-1}) &= \prod_{i=1}^{t-1} (1 + \lambda_i(\mu) \cdot (X_i - \mu_t^{\text{WoR}})) \cdot \mathbb{E}(1 + \lambda_t(\mu) \cdot (X_t - \mu_t^{\text{WoR}}) \mid \mathcal{F}_{t-1}) \\ &= \mathcal{K}_{t-1}^{\text{WoR}}(\mu) \cdot (1 + \lambda_t(\mu) \cdot (\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) - \mu_t^{\text{WoR}})) \\ &= \mathcal{K}_{t-1}^{\text{WoR}}(\mu). \end{aligned}$$

Since  $\mathcal{K}_0^{\text{WoR}}(\mu) \equiv 1$  by convention, we have that  $\mathcal{K}_t^{\text{WoR}}(\mu)$  is a martingale.

Now, note that since  $X_t \in [0, 1]$  and  $\lambda_t^{\text{WoR}}(\mu) \in [-1/(1 - \mu_t^{\text{WoR}}), 1/\mu_t^{\text{WoR}}]$  for each  $t$  by assumption, we have that  $1 + \lambda_t(\mu) \cdot (X_t - \mu_t^{\text{WoR}}) \geq 0$  and thus  $\mathcal{K}_t^{\text{WoR}}(\mu) \geq 0$ . Therefore,  $\mathcal{K}_t^{\text{WoR}}(\mu)$  is a test martingale.

**Proof of (2)  $\implies$  (1).** Suppose that  $\mathcal{K}_t^{\text{WoR}}(\mu)$  is a test martingale for any  $(\lambda_t(\mu))_{t=1}^N$  with  $\lambda_t(\mu) \in [-1/(1 - \mu_t^{\text{WoR}}), 1/\mu_t^{\text{WoR}}]$ , but suppose for the sake of contradiction that  $\mathbb{E}(X_{t^\star} \mid \mathcal{F}_{t^\star-1}) \neq \mu_{t^\star}^{\text{WoR}}$  for some  $t^\star \in \{1, 2, \dots\}$ . Set  $\lambda_1 = \lambda_2 = \dots = \lambda_{t^\star-1} = 0$  and  $\lambda_{t^\star} = 1$ . Then,

$$\mathcal{K}_{t^\star}^{\text{WoR}}(\mu) \equiv \mathcal{K}_{t^\star-1}^{\text{WoR}}(\mu) \cdot (1 + \lambda_{t^\star}(X_{t^\star} - \mu_{t^\star}^{\text{WoR}})) = 1 + X_{t^\star} - \mu_{t^\star}^{\text{WoR}}.$$

By assumption of  $\mathcal{K}_t^{\text{WoR}}(\mu)$  forming a martingale, we have that  $\mathbb{E}(\mathcal{K}_{t^\star}^{\text{WoR}}(\mu) \mid \mathcal{F}_{t^\star-1}) = \mathcal{K}_{t^\star-1}^{\text{WoR}}(\mu) = 1$ . On the other hand, since  $\mathbb{E}(X_{t^\star} \mid \mathcal{F}_{t^\star-1}) \neq \mu_{t^\star}^{\text{WoR}}$ , we have

$$\mathbb{E}(\mathcal{K}_{t^\star}^{\text{WoR}}(\mu) \mid \mathcal{F}_{t^\star-1}) = \mathbb{E}(1 + X_{t^\star} - \mu_{t^\star}^{\text{WoR}} \mid \mathcal{F}_{t^\star-1}) \neq 1,$$

a contradiction. Therefore, we must have that  $\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = \mu_t^{\text{WoR}}$  for each  $t$ , which completes the proof of (2)  $\implies$  (1) and Proposition 3.5.1.  $\square$

### 3.A.9 Proof of Theorem 3.5.1

The proof that  $\mathfrak{B}_t^{\pm, \text{WoR}}$  forms a  $(1 - \alpha)$ -CS for  $\mu$  proceeds in exactly the same manner as Theorem 3.4.1, noting that  $\mathbb{E}(X_t \mid \mathcal{F}_{t-1}) = \mu_t^{\text{WoR}}$  instead of  $\mu$ .

To show that  $\mathfrak{B}_t^{\pm, \text{WoR}}$  is indeed an interval for each  $t \geq 1$ , we note that the proof of Theorem 3.4.1 applies since  $m_t^{\text{WoR}}$  is increasing or decreasing if and only if  $m$  is increasing or decreasing, respectively.  $\square$

## 3.B How to bet: deriving adaptive betting strategies

In Section 3.4.4, we presented CSs and CIs via the hedged capital process. We suggested a specific betting scheme which has strong empirical performance but did not discuss where it came from. In this section, we derive various betting strategies and discuss their statistical and computational properties.

### 3.B.1 Predictable plug-ins yield good betting strategies

First and foremost, we will examine why any predictable plug-in for empirical Bernstein-type CSs and CIs (i.e. those recommended in Theorem 3.3.1 and Remark 1) yield effective betting strategies. Consider the hedged capital process

$$\begin{aligned}\mathcal{K}_t^\pm(m) &:= \max \left\{ \theta \prod_{i=1}^t (1 + \lambda_i^+(X_i - m)), (1 - \theta) \prod_{i=1}^t (1 - \lambda_i^-(X_i - m)) \right\} \\ &\equiv \max \{ \theta \mathcal{K}_t^+(m), (1 - \theta) \mathcal{K}_t^-(m) \},\end{aligned}$$

where  $(\lambda_t^+(m))_{t=1}^\infty$  and  $(\lambda_t^-(m))_{t=1}^\infty$  are  $[0, 1/m]$ -valued and  $[0, 1/(1-m)]$ -valued predictable sequences as in Theorem 3.4.1. First, consider the “positive” capital process,  $\mathcal{K}_t^+(\mu)$  evaluated at  $m = \mu$ . An inequality that has been repeatedly used to derive empirical Bernstein inequalities [124, 125, 281], including the current chapter is the following due to Fan et al. [102, equation 4.12]: for any  $y \geq -1$  and  $\lambda \in [0, 1)$ , we have

$$\log(1 + \lambda y) \geq \lambda y - 4\psi_E(\lambda)y^2. \quad (3.41)$$

where  $\psi_E(\lambda)$  is as defined in (3.14). If the predictable sequence  $(\lambda_t^+(m))_{t=1}^\infty$  is further restricted to  $[0, 1)$ , then by (3.41) we have

$$\begin{aligned}\mathcal{K}_t^+(\mu) &:= \prod_{i=1}^t (1 + \lambda_i^+(X_i - \mu)) \geq \exp \left( \sum_{i=1}^t \lambda_i^+(X_i - \mu) - \sum_{i=1}^t 4(X_i - \mu)^2 \psi_E(\lambda_i^+) \right) \\ &\stackrel{(i)}{\approx} \exp \left( \sum_{i=1}^t \lambda_i^+(X_i - \mu) - \sum_{i=1}^t 4(X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i^+) \right) \\ &= M_t^{\text{PrPl-EB}}(\mu),\end{aligned}$$

where (i) follows from the approximations  $\hat{\mu}_{t-1} \approx \mu$  for large  $t$ . Not only does the approximate inequality  $\mathcal{K}_t^+(\mu) \gtrsim M_t^{\text{PrPl-EB}}(\mu)$  shed light on why a sensible empirical Bernstein predictable plug-in translates to a sensible betting strategy, but also why we might expect  $\mathcal{K}_t^+(m)$  to be more powerful than  $M_t^{\text{PrPl-EB}}(m)$  for the same  $[0, 1)$ -valued predictable sequence  $(\lambda_t^+(m))_{t=1}^\infty$ . Moreover,  $\mathcal{K}_t^+(m)$  has the added flexibility of allowing  $(\lambda_t(m))_{t=1}^\infty$  to take values in  $[0, 1/m] \supset [0, 1)$  which we find – through simulations – tends to improves empirical performance (see Figure 3.19 in Section 3.E.2.2). Finally, a similar story holds for  $\mathcal{K}_t^-(\mu)$  with the added caveat that  $(\lambda_t^-)_{t=1}^\infty$  can instead take values in  $[0, 1/(1-m)] \supset [0, 1)$  which as before, seems to improve empirical performance.

Despite the success of predictable plug-ins as betting strategies, it is natural to wonder whether it is preferable to focus on directly maximizing capital over time. As will be seen in the following section, these capital-maximizing approaches tend to have improved empirical performance, but are not always guaranteed to produce convex confidence sets (i.e. intervals). Nevertheless, it is worth examining some of these strategies both for their intuitive appeal and excellent empirical performance.

### 3.B.2 Growth rate adaptive to the particular alternative (GRAPA)

As alluded to in Section 3.6, Kelly Jr [152] dealt with capital processes, betting strategies, etc. in the fields of information and communication theory in the pursuit of maximizing the information rate over a channel. Kelly suggested that an effective betting strategy is one that maximizes a gambler’s expected *log-capital* – i.e. the growth rate of the gambler’s capital – under a particular alternative.<sup>2</sup> However, Kelly’s setup was a simplified special case of ours: Kelly’s observations were binary, and the exact alternative was assumed known, while ours are merely bounded in  $[0, 1]$  with an unknown alternative. Nevertheless, the principle of maximizing the log-capital can be adapted to our setting under bounded observations and an unknown alternative. We summarize this adaptation here and refer to it as maximizing the “growth rate adaptive to the particular alternative” or “GRAPA” for short.

Write the log-capital process at time  $t$  as

$$\ell_t(\lambda_1^t, m) := \log(\mathcal{K}_t(m)) = \sum_{i=1}^t \log(1 + \lambda_i(m)(X_i - m)), \quad (3.42)$$

for a general  $[-1/(1-m), 1/m]$ -valued sequence  $(\lambda_t(m))_{t=1}^\infty$ . If we were to choose a single value of  $\lambda^{\text{HS}} := \lambda_1 = \dots = \lambda_t$  which maximizes the log-capital  $\ell_t$  “in hindsight” (i.e. based on *all* of the previous data), then this value is given by

$$\frac{\partial \ell_t(\lambda^{\text{HS}}, m)}{\partial \lambda^{\text{HS}}} = \sum_{i=1}^t \frac{X_i - m}{1 + \lambda^{\text{HS}}(X_i - m)} \stackrel{\text{set}}{=} 0.$$

However,  $\lambda^{\text{HS}}$  is clearly not predictable. Following Kumon et al. [166] (who referred to this as the “sequential optimization strategy”), we set  $(\lambda_t^{\text{GRAPA}}(m))_{t=1}^\infty$  such that

$$\frac{1}{t-1} \sum_{i=1}^{t-1} \frac{X_i - m}{1 + \lambda_t^{\text{GRAPA}}(m)(X_i - m)} \stackrel{\text{set}}{=} 0, \quad (3.43)$$

truncated to lie between  $(-c/(1-m), c/m)$  using some  $c \leq 1$ . Importantly,  $\lambda_t^{\text{GRAPA}}(m)$  only depends on  $X_1, \dots, X_{t-1}$ , and is thus predictable.

This rule is a sequentially adaptive version of the worst-case “GROW” criterion of Grünwald

---

<sup>2</sup>This objective has also been arrived at indirectly as the dual in optimization programs for deriving regret bounds for Kullback-Leibler-based UCB algorithms in multi-armed bandit problems [121, 49].

et al. [113]. To see the connection, one can derive (3.43) from a slightly different motivation. At the  $t$ -th step, we want to choose  $\lambda_t(m)$  so that the wealth multiplier  $(1 + \lambda_t(m))(X_t - m)$  is as large as possible. The ideal choice would be

$$\lambda_t^*(m) := \operatorname{argmax}_{\lambda \in [-1/(1-m), 1/m]} \mathbb{E}_{P^\mu} [\log(1 + \lambda(X_t - m)) \mid \mathcal{F}_{t-1}], \quad (3.44)$$

where  $P^\mu$  is the unknown true distribution. Writing down the stationary condition for this optimization problem by differentiating through the expectation, we get

$$\mathbb{E}_{P^\mu} \left[ \frac{X_t - m}{1 + \lambda_t^*(m)(X_t - m)} \mid \mathcal{F}_{t-1} \right] = 0. \quad (3.45)$$

Since  $P^\mu$  is unknown, using a simple empirical plug-in estimator yields (3.43).

CSs constructed from  $(\lambda_t^{\text{GRAPA}}(m))_{t=1}^\infty$  tend to have excellent empirical performance, but can be prohibitively slow due to the required root-finding in (3.43) for each time  $t$  and  $m \in [0, 1]$  (or a sufficiently fine grid of  $[0, 1]$ ). A similar but computationally inexpensive alternative to GRAPA is “approximate GRAPA” (aGRAPA), which we derive now.

### 3.B.3 Approximate GRAPA (aGRAPA)

Rather than solve (3.43), we take the Taylor approximation of  $(1 + y)^{-1}$  by  $(1 - y)$  for  $y \approx 0$  to obtain

$$\begin{aligned} \frac{1}{t-1} \sum_{i=1}^{t-1} \frac{X_i - m}{1 + \lambda_t^{\text{aGRAPA}}(m)(X_i - m)} &\approx \frac{1}{t-1} \sum_{i=1}^{t-1} (1 - \lambda_t^{\text{aGRAPA}}(m)(X_i - m)) (X_i - m) \\ &= \frac{1}{t-1} \sum_{i=1}^{t-1} (X_i - m) - \frac{\lambda_t^{\text{aGRAPA}}(m)}{t-1} \sum_{i=1}^{t-1} (X_i - m)^2 \\ &\stackrel{\text{set}}{=} 0, \end{aligned}$$

which, after appropriate truncation leads what we call the “approximate GRAPA” (aGRAPA) betting strategy,

$$\lambda_t^{\text{aGRAPA}}(m) := -\frac{c}{1-m} \vee \frac{\hat{\mu}_{t-1} - m}{\hat{\sigma}_{t-1}^2 + (\hat{\mu}_{t-1} - m)^2} \wedge \frac{c}{m},$$

for some truncation level  $c \leq 1$ . This expression is quite natural: we bet more aggressively if our empirical mean is far away from  $m$ , and are further emboldened if the empirical variance is small.

As alluded to at the end of Section 3.B.1, CSs derived using the capital process  $\mathcal{K}_t(m)$  with arbitrary betting schemes are not always guaranteed to produce a convex set (interval). In fact, it is possible to construct scenarios where the sublevel sets of  $\mathcal{K}_t^{\text{aGRAPA}}(m)$  are nonconvex in  $m$  (see Section 3.E.4 for an example). In our experience, this type of situation is not common,

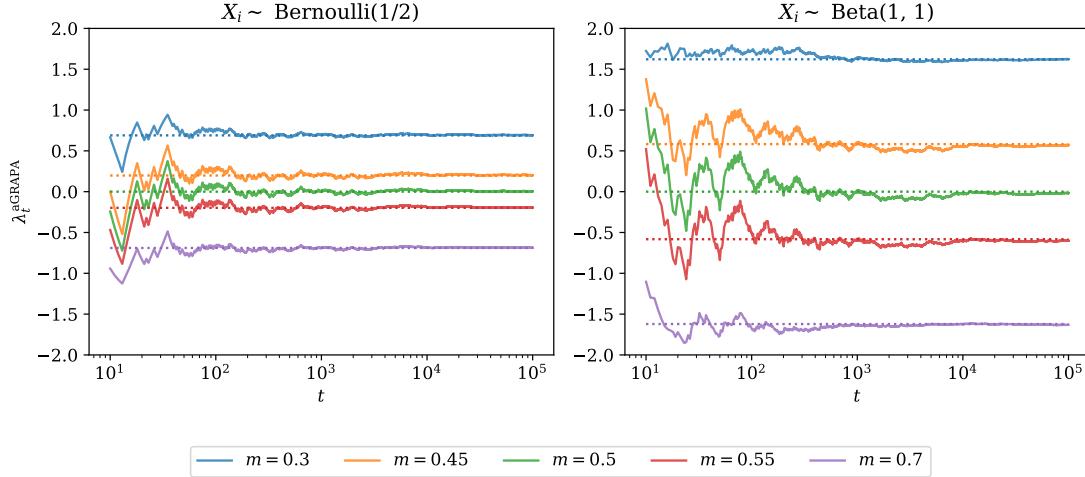


Figure 3.10:  $\lambda_t^{\text{aGRAPA}}$  for various values of  $m$  under two distributions:  $\text{Bernoulli}(1/2)$  and  $\text{Beta}(1, 1)$ . The dotted lines show the “oracle” bets, meaning  $\lambda_t^{\text{aGRAPA}}$  with estimates of the mean and variance replaced by their true values. As time passes, bets stabilize and approach their oracle quantities.

and one must actively search for such pathological examples.

### 3.B.4 Lower-bound on the wealth (LBOW)

Instead of maximizing  $\log(\mathcal{K}_t(m))$ , we may aim to do so for a tight lower-bound on the wealth (LBOW). This technique has proven useful in the game-theoretic probability literature [235, Proof of Lemma 3.3] and [72, Proof of Theorem 1]. Our lower bound will rely on an extension of Fan’s inequality (3.41) to  $\lambda \in (-1/(1-m), 1/m)$ , summarized in the following lemma.

**Lemma 3.B.1.** *If  $y \geq -m$ , then for any  $\lambda \in [0, 1/m]$ , we have*

$$\log(1 + \lambda y) \geq \lambda y + \frac{y^2}{m^2} (\log(1 - m\lambda) + m\lambda).$$

*On the other hand, if  $y \leq 1 - m$ , then for any  $\lambda \in (-1/(1-m), 0]$ , we have*

$$\log(1 + \lambda y) \geq \lambda y + \frac{y^2}{(1-m)^2} (\log(1 + (1-m)\lambda) - (1-m)\lambda).$$

*Thus, for  $y \in [-m, 1 - m]$ , both of the above inequalities hold.*

The proof is an easy generalization of inequality (3.41) by Fan et al. [102], and also follows from similar observations about the subexponential function  $\psi_E$  in Howard et al. [124, 125], but we prove it from first principles in Section 3.A.6 for completeness. Using Lemma 3.B.1, we

have for  $\lambda^{L+} \in [0, 1/m]$ , the following lower-bound on  $\ell(\lambda^{L+}, m)$ ,

$$\begin{aligned}\ell(\lambda^{L+}, m) &:= \log \left( \prod_{i=1}^t (1 + \lambda^{L+}(X_i - m)) \right) \\ &\geq \lambda^{L+} \sum_{i=1}^t (X_i - m) + \frac{\log(1 - m\lambda^{L+}) + m\lambda^{L+}}{m^2} \sum_{i=1}^t (X_i - m)^2,\end{aligned}\quad (3.46)$$

and for  $\lambda^{L-} \in (-1/(1-m), 0]$ , we have

$$\begin{aligned}\ell(\lambda^{L-}, m) &:= \log \left( \prod_{i=1}^t (1 + \lambda^{L-}(X_i - m)) \right) \\ &\geq \lambda^{L-} \sum_{i=1}^t (X_i - m) + \frac{\log(1 + (1-m)\lambda^{L-}) - (1-m)\lambda^{L-}}{(1-m)^2} \sum_{i=1}^t (X_i - m)^2.\end{aligned}\quad (3.47)$$

Importantly, if  $\sum_{i=1}^t (X_i - m)$  is positive, then (3.46) is concave, while if negative, (3.47) is concave. Maximizing (3.46) or (3.47) depending on the sign of  $\sum_{i=1}^t (X_i - m)$  we obtain the following “hindsight” choice for  $\lambda^L$ ,

$$\lambda^L = \begin{cases} \frac{\sum_{i=1}^t (X_i - m)}{m \sum_{i=1}^t (X_i - m) + \sum_{i=1}^t (X_i - m)^2} & \text{if } \sum_{i=1}^t (X_i - m) \geq 0, \\ \frac{\sum_{i=1}^t (X_i - m)}{-(1-m) \sum_{i=1}^t (X_i - m) + \sum_{i=1}^t (X_i - m)^2} & \text{if } \sum_{i=1}^t (X_i - m) \leq 0. \end{cases}$$

Of course, this choice of  $\lambda^L$  is not predictable and thus is not a valid betting strategy in the framework of the current chapter. This motivates the following strategy,  $(\lambda_t^L(m))_{t=1}^\infty$  given by

$$\lambda_t^L(m) := \frac{-c}{1-m} \vee \frac{\hat{\mu}_{t-1} - m}{\omega_{t-1} |\hat{\mu}_{t-1} - m| + \hat{\sigma}_{t-1}^2 + (\hat{\mu}_{t-1} - m)^2} \wedge \frac{c}{m}, \quad (3.48)$$

$$\text{where } \omega_t := \begin{cases} m & \text{if } \hat{\mu}_t - m \geq 0, \\ 1-m & \text{if } \hat{\mu}_t - m < 0. \end{cases}$$

Similarly to the aGRAPA betting procedure, LBOW is computationally-inexpensive but is not guaranteed to produce an interval. The expression also carries similar intuition to the GRAPA case.

### 3.B.5 Online Newton Step (ONS- $m$ )

Betting algorithms play an essential role in online learning as several optimization problems can be framed in terms of coin-betting games [72, 197, 140, 139]. While the downstream application is different, the game-theoretic techniques of maximizing wealth are almost immediately applicable to the problem at hand. Here, we consider a slight modification to the Online Newton

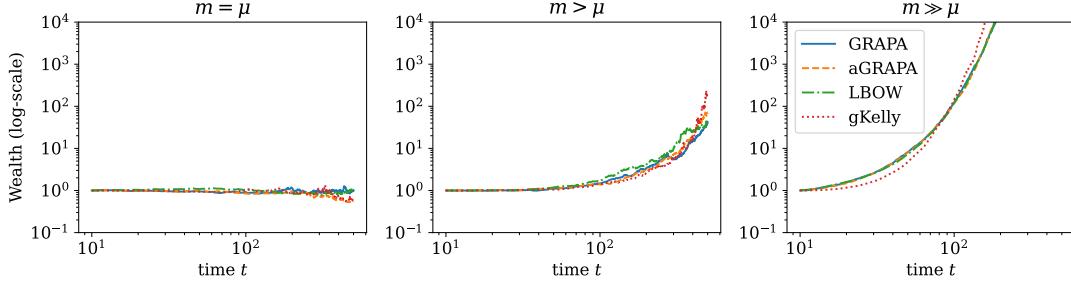


Figure 3.11: Comparison of the wealth process under various game-theoretic betting strategies with 100 repeats. In this example, the 1000 observations are drawn from a Beta(10, 10) distribution, and the candidate means  $m$  being tested are 0.5, 0.51, and 0.55 (from left to right). Notice that these strategies perform similarly, but have varying computational costs (see Table 3.2).

Step (ONS) algorithm due to Cutkosky and Orabona [72].

Algorithm 3.1: Online Newton Step (ONS- $m$ )

```

Result:  $(\lambda_t^O(m))_{t=1}^T$ 
 $\lambda_1^O(m) \leftarrow 1$ 
for  $t = 1$  to  $T - 1$  do
     $y_t \leftarrow X_t - m$ 
     $z_t \leftarrow y_t / (1 - y_t \lambda_t^O(m))$ 
     $A_t \leftarrow 1 + \sum_{i=1}^t z_i^2$ 
     $\lambda_{t+1}^O(m) \leftarrow \max \left( \frac{-c}{1-m}, \min \left( \lambda_t^O(m) - \frac{2}{2-\log(3)} \frac{z_t}{A_t}, \frac{c}{m} \right) \right)$ 
end for

```

Through simulations, we find that ONS- $m$  performs competitively. However, its lack of closed-form expression makes it a slightly more computationally-expensive alternative to aGRAPA and LBOW, but not nearly as expensive as GRAPA (see Table 3.2).

### 3.B.6 Diversified Kelly betting (dKelly)

Instead of committing to one betting strategy such as aGRAPA or LBOW, we can simply take the average capital among  $D$  separate strategies. This follows from the fact that an average of test martingales is itself a test martingale. That is, if  $(\lambda_t^1)_{t=1}^\infty, (\lambda_t^2)_{t=1}^\infty, \dots, (\lambda_t^D)_{t=1}^\infty$  are  $D$  separate betting strategies, then

$$\mathcal{K}_t^{\text{dKelly}}(\mu) := \frac{1}{D} \sum_{d=1}^D \prod_{i=1}^t \left( 1 + \lambda_i^d(\mu)(X_i - \mu) \right)$$

forms a test martingale. Following Kelly's original motivation to maximize (expected) log-capital, notice that by Jensen's inequality,

$$\log \left( \mathcal{K}_t^{\text{dKelly}} \right) > \frac{1}{D} \sum_{d=1}^D \log \left( \prod_{i=1}^t \left( 1 + \lambda_i^d(\mu)(X_i - \mu) \right) \right).$$

In other words, the log-capital of the diversified bets is strictly larger than the average log-capital among the diverse candidate bets.

**Grid Kelly betting (gKelly).** While it is possible to use any finite collection of strategies, we focus our attention on a particularly simple (and useful) example where the bets are constant values on a grid. Specifically, divide the interval  $[-1/(1-m), 1/m]$  up into  $G$  evenly-spaced points  $\lambda^1, \dots, \lambda^G$ . Then define the gKelly capital process  $\mathcal{K}_t^{\text{gKelly}}$  by

$$\mathcal{K}_t^{\text{gKelly}}(m) := \frac{1}{G} \sum_{g=1}^G \prod_{i=1}^t (1 + \lambda^g(X_i - m)).$$

When used to construct confidence sequences for  $\mu$ ,  $\mathcal{K}_t^{\text{gKelly}}$  demonstrates excellent empirical performance. Moreover, this procedure can be slightly modified into "Hedged gKelly" (hgKelly) so that confidence sequences constructed using gKelly are intervals almost surely.

In order to mimic the unknown optimal  $\lambda^*$ ,  $D$  or  $G$  should not be kept constant, but itself grow slowly (say logarithmically) with  $t$ . In game-theoretic terms, one should slowly add more strategies to the portfolio, in order to asymptotically match the performance of the optimal one over time. (When adding a new  $\lambda^g$  to an existing mixture, it obviously only begins to contribute to the wealth from the following step onwards; formally  $G$  would be replaced by  $G_t$ , and  $\prod_{i=1}^t (1 + \lambda^g(X_i - m))$  would be replaced by  $\prod_{i=t_g}^t (1 + \lambda^g(X_i - m))$  if  $\lambda^g$  was first introduced after  $t_g - 1$  steps.)

**Hedged gKelly.** First, divide the interval  $[-1/(1-m), 0]$  and  $[0, 1/m]$  into  $G$  evenly-spaced points:  $(\lambda^{1-}, \dots, \lambda^{G-})$  and  $(\lambda^{1+}, \dots, \lambda^{G+})$ , respectively. Then define the "Hedged grid Kelly capital process"  $\mathcal{K}_t^{\text{hgKelly}}$  given by

$$\mathcal{K}_t^{\text{hgKelly}}(m) := \frac{\theta}{G} \sum_{g=1}^G \prod_{i=1}^t (1 + \lambda^{g+}(X_i - m)) + \frac{1-\theta}{G} \sum_{g=1}^G \prod_{i=1}^t (1 + \lambda^{g-}(X_i - m)),$$

where  $\theta \in [0, 1]$  (a reasonable default being  $\theta = 1/2$ ).

**Proposition 3.B.1.** *If  $(X_t)_{t=1}^\infty \sim P$  for some  $P \in \mathcal{P}^\mu$ , then  $\mathcal{K}_t^{\text{hgKelly}}(\mu)$  forms a test martingale and  $\mathfrak{B}_t^{\text{hgKelly}} := \{m \in [0, 1] : \mathcal{K}_t^{\text{hgKelly}}(m) < 1/\alpha\}$  is a CS for  $\mu$  that forms an interval for each  $t \geq 1$ .*

The proof in Section 3.A.7 proceeds by showing that  $\mathcal{K}_t^{\text{hgKelly}}$  is a convex function of  $m$

and hence its sublevel sets are intervals.

### 3.B.7 Confidence Boundary (ConBo)

The aforementioned strategies benefit from targeting bets against a particular null hypothesis,  $H_0^m$  for each  $m \in [0, 1]$ , but this has the drawback of  $\mathcal{K}_t(m)$  potentially not being quasiconvex in  $m$ . One of the advantages of the hedged capital process as described in Theorem 3.4.1 is that  $\mathcal{K}_t^\pm(m)$  is always quasiconvex, and thus its sublevel sets (and hence the confidence sets  $\mathfrak{B}_t^\pm$ ) are intervals.

In an effort to develop game-theoretic betting strategies which generate confidence sets which are intervals, we present the Confidence Boundary (ConBo) bets. Rather than bet against the null hypotheses  $H_0^m$  for each  $m \in [0, 1]$ , consider two sequences of nulls,  $(H_0^{u_t})_{t=1}^\infty$  and  $(H_0^{l_t})_{t=1}^\infty$  corresponding to upper and lower confidence boundaries, respectively. The ConBo bet  $\lambda_t^{\text{CB}}$  is then targeted against  $u_{t-1}$  and  $l_{t-1}$  using *any* game-theoretic betting strategy (e.g. \*GRAPA, \*Kelly, LBOW, or ONS-m). Letting  $\lambda_t^G(m)$  be any such strategy, we summarize the ConBo betting scheme in Algorithm 3.2.

**Corollary 3.B.1** (Confidence boundary CS [ConBo]). *In Algorithm 3.2,*

$$\mathfrak{B}_t^{\text{CB}} \text{ forms a } (1 - \alpha)\text{-CS for } \mu,$$

as does  $\bigcap_{i \leq t} \mathfrak{B}_i^{\text{CB}}$ . Further,  $\mathfrak{B}_t^{\text{CB}}$  is an interval for any  $t \geq 1$ .

Algorithm 3.2: Confidence boundary (ConBo)

```

1: Result:  $(\mathcal{K}_t^{\text{CB}}(m))_{t=1}^T$ 
2:  $l_0 \leftarrow 0; u_0 \leftarrow 1$ 
3:  $\mathcal{K}_0^{\text{CB+}}(m) \leftarrow \mathcal{K}_0^{\text{CB-}}(m) \leftarrow 1$ 
4: for  $t = 1$  to  $T$  do
5:    $\lambda_t^{\text{CB+}} \leftarrow \max \left\{ \lambda_t^G(l_{t-1}), 0 \right\} \wedge \frac{c}{m}$             $\triangleright$  Compute ConBo bets
6:    $\lambda_t^{\text{CB-}} \leftarrow \left| \min \left\{ \lambda_t^G(u_{t-1}), 0 \right\} \right| \wedge \frac{c}{1-m}$ 
7:    $\mathcal{K}_t^{\text{CB+}}(m) \leftarrow \left[ 1 + \lambda_t^{\text{CB+}}(X_t - m) \right] \cdot \mathcal{K}_{t-1}^{\text{CB+}}(m)$         $\triangleright$  Update capital
8:    $\mathcal{K}_t^{\text{CB-}}(m) \leftarrow \left[ 1 - \lambda_t^{\text{CB-}}(X_t - m) \right] \cdot \mathcal{K}_{t-1}^{\text{CB-}}(m)$ 
9:    $\mathcal{K}_t^{\text{CB}}(m) \leftarrow \max \left\{ \theta \mathcal{K}_t^{\text{CB+}}(m), (1 - \theta) \mathcal{K}_t^{\text{CB-}}(m) \right\}$             $\triangleright$  Hedging
10:   $\mathfrak{B}_t^{\text{CB}} \leftarrow \{m \in [0, 1] : \mathcal{K}_t(m) < 1/\alpha\}$ 
11:   $l_t \leftarrow \inf \mathfrak{B}_t^{\text{CB}}$             $\triangleright$  Update confidence boundaries to bet against
12:   $u_t \leftarrow \sup \mathfrak{B}_t^{\text{CB}}$ 
13: end for

```

We can also adapt the ConBo betting scheme outlined in Algorithm 3.2 to the without-replacement setting by replacing  $m$  by  $m_t^{\text{WoR}}$  for each time  $t$ .

**Corollary 3.B.2** (WoR confidence boundary CS [ConBo-WoR]). *Under the same conditions*

as Theorem 3.5.1, define  $\lambda_t^{\text{CB-WoR}+}$  and  $\lambda_t^{\text{CB-WoR}-}$  as in Algorithm 3.2 but with  $m$  replaced by  $m_t^{\text{WoR}}$ . Then,

$$\mathfrak{B}_t^{\text{CB-WoR}} := \{m \in [0, 1] : \mathcal{K}_t^{\text{CB-WoR}} < 1/\alpha\} \quad \text{forms a } (1 - \alpha)\text{-CS for } \mu,$$

as does  $\bigcap_{i \leq t} \mathfrak{B}_i^{\text{CB-WoR}}$ . Further,  $\mathfrak{B}_t^{\text{CB-WoR}}$  is an interval for each  $t \geq 1$ .

### 3.B.8 Sequentially Rebalanced Portfolio (SRP)

Implicitly, none of the aforementioned strategies take advantage of “rebalancing”, meaning the ability to take ones capital  $\mathcal{K}_t$  at time  $t$ , diversify it in any manner at time  $t + 1$ , and repeat. This has had the mathematical advantage of being able to write the resulting capital process  $(\mathcal{K}_t(m))_{t=1}^\infty$  in the following general, but closed-form expression:

$$\mathcal{K}_t(m) := \sum_{d=1}^D \theta_d \prod_{i=1}^t (1 + \lambda_i^d(m) \cdot (X_i - m)),$$

where  $D \geq 1$  is as in Section 3.B.6,  $(\lambda_t^1(m))_{t=1}^\infty, \dots, (\lambda_t^D(m))_{t=1}^\infty$  are  $[-1/(1-m), 1/m]$ -valued predictable sequences as usual, and  $(\theta_d)_{d=1}^D$  are convex weights such that  $\sum_{d=1}^D \theta_d = 1$ . However, a more general capital process martingale can be written but instead of having a closed-form product expression, it can be written recursively as

$$\mathcal{K}_t^{\text{SRP}}(m) := \sum_{d=1}^{D_t} (1 + \lambda_t^d(m) \cdot (X_t - m)) \cdot \theta_t^d \cdot \mathcal{K}_{t-1}^{\text{SRP}}(m), \quad (3.49)$$

where  $(\lambda_t^d)_{d=1}^{D_t}$  are  $[1/(1-m), 1/m]$ -valued predictable bets,  $(\theta_t^d)_{d=1}^{D_t}$  are predictable convex weights that sum to 1 (conditional on  $X_1^{t-1}$ ), and we have set the initial capital  $\mathcal{K}_0^{\text{SRP}}(m)$  to 1 as usual.

Adopting the betting interpretation, (3.49) is a rather intuitive procedure. At each time step  $t$ , the gambler divides their previous capital  $\mathcal{K}_{t-1}^{\text{SRP}}(m)$  up into  $D_t \geq 1$  portions given by  $\theta_t^1 \cdot \mathcal{K}_{t-1}^{\text{SRP}}(m), \dots, \theta_t^{D_t} \cdot \mathcal{K}_{t-1}^{\text{SRP}}(m)$ , then invests these wealths with bets  $\lambda_t^1(m), \dots, \lambda_t^{D_t}(m)$ , respectively. The gambler’s wealths are then updated to

$$(1 + \lambda_t^1(m) \cdot (X_t - m)) \cdot \theta_t^1 \cdot \mathcal{K}_{t-1}^{\text{SRP}}(m), \dots, (1 + \lambda_t^{D_t}(m) \cdot (X_t - m)) \cdot \theta_t^{D_t} \cdot \mathcal{K}_{t-1}^{\text{SRP}}(m),$$

which are then combined via summation to yield a final capital of (3.49).

It is now routine to check that the process given by (3.49) is a nonnegative martingale when evaluated at  $\mu$  since

$$\mathbb{E}(\mathcal{K}_t^{\text{SRP}}(\mu) | X_1^{t-1}) = \sum_{d=1}^{D_t} \mathcal{K}_{t-1}^{\text{SRP}}(\mu) \cdot \theta_t^d \cdot \left( 1 + \lambda_t(\mu) \left( \underbrace{\mathbb{E}(X_t | X_1^{t-1}) - \mu}_{=0} \right) \right)$$

$$= \mathcal{K}_{t-1}^{\text{SRP}}(\mu) \underbrace{\sum_{d=1}^{D_t} \theta_t^{D_t}}_{=1} = \mathcal{K}_{t-1}^{\text{SRP}}(\mu).$$

Note that SRP is the most general and customizable betting strategy presented in this chapter, since it can be composed of any of the previously discussed strategies, and includes each of them as a special case.

### 3.C Simulations

This section contains a comprehensive set of simulations comparing our new confidence sets presented against previous works. We present simulations for building both time-uniform CSs and fixed-time CIs with or without replacement. Each of these are presented under four distributional “themes”: (1) discrete, high-variance; (2) discrete, low-variance; (3) real-valued, evenly spread; and (4) real-valued, concentrated.

### 3.C.1 Time-uniform confidence sequences (with replacement)

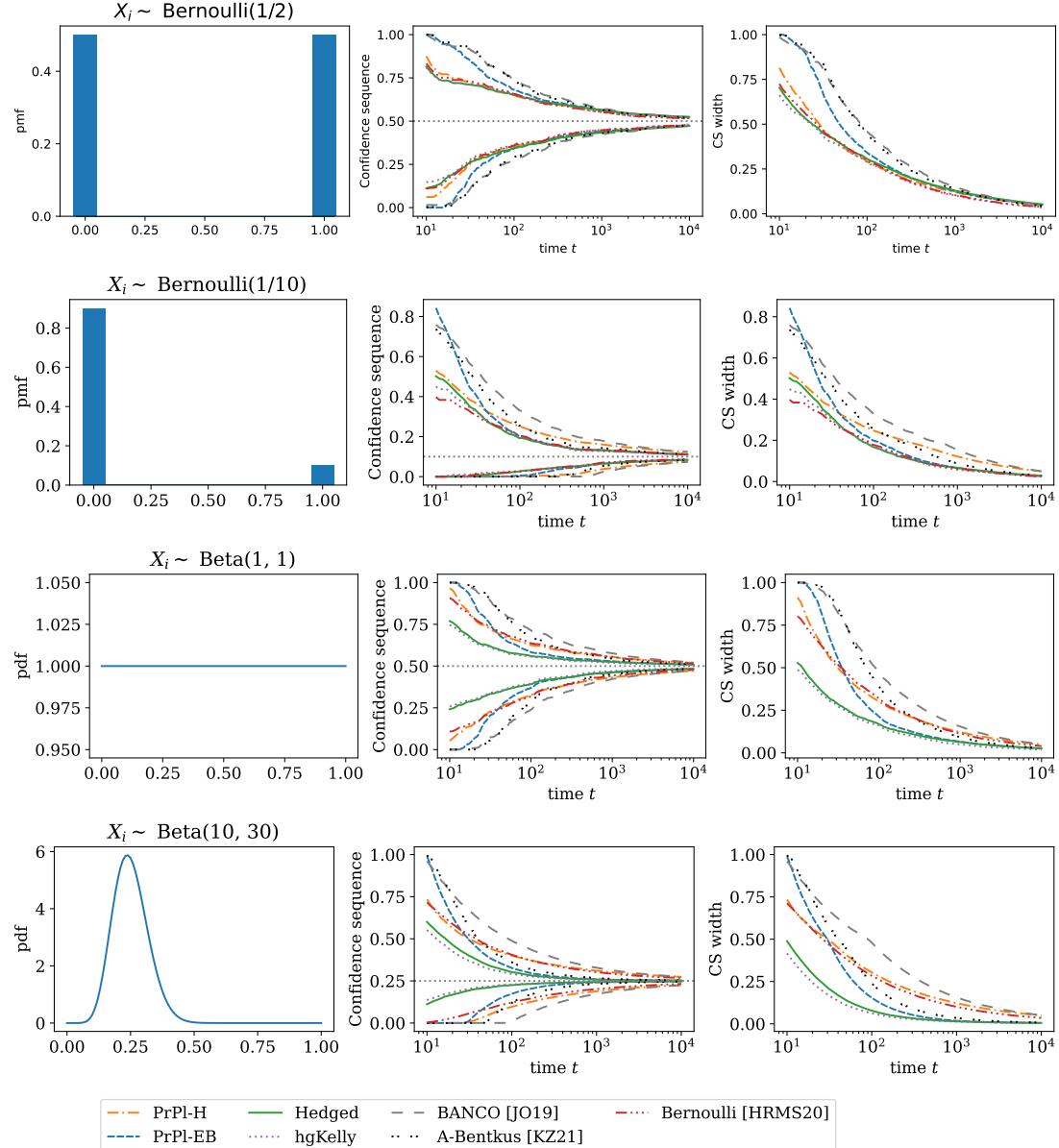


Figure 3.12: Comparing Hedged, hgKelly, PrPl-EB, and PrPl-H CSs alongside other time-uniform confidence sequences in the literature; further details in Section 3.D.1. Clearly, the betting approach is dominant in all settings.

### 3.C.2 Fixed-time confidence intervals (with replacement)

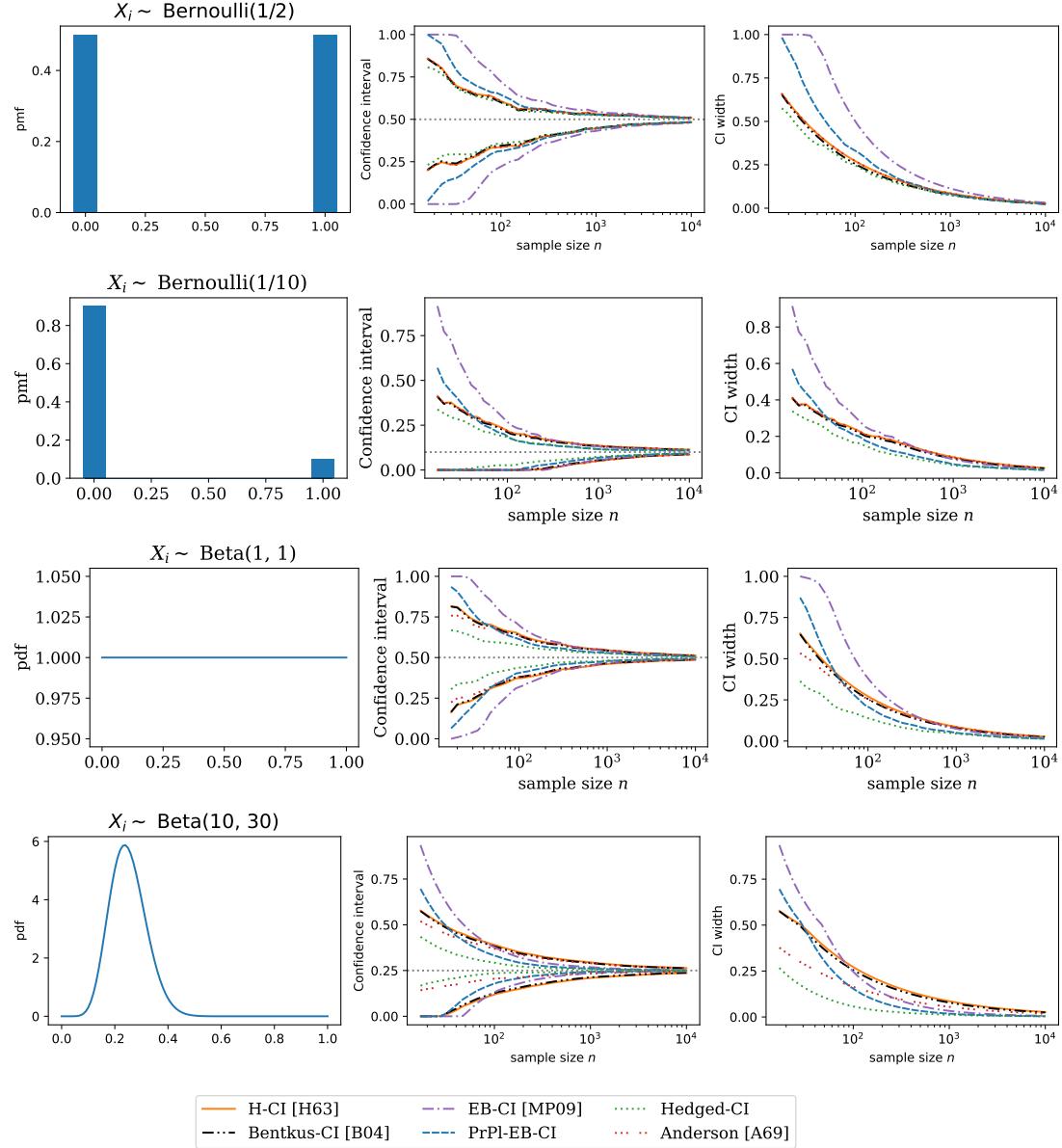


Figure 3.13: Hedged capital, Anderson, Bentkus, Maurer-Pontil empirical Bernstein, and predictable plug-in empirical Bernstein CIs under four distributional scenarios. Further details can be found in Section 3.D.2. Clearly, the betting approach is dominant in all settings.

### 3.C.3 Time-uniform confidence sequences (without replacement)

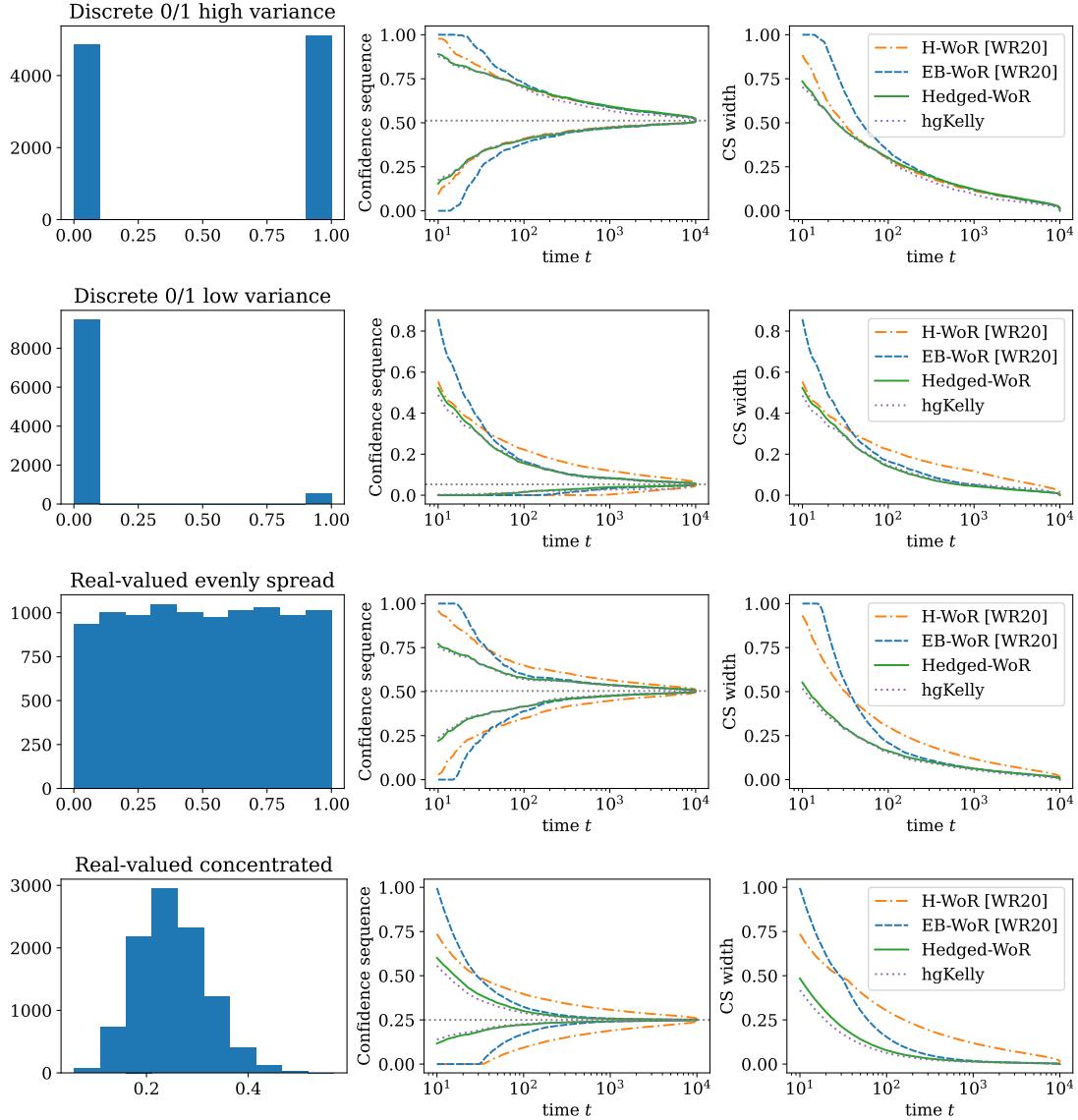


Figure 3.14: Hedged capital, Hoeffding, and empirical Bernstein CSs for the mean of a finite set of bounded numbers when sampling WoR. Further details can be found in Section 3.D.3. Clearly, the betting approach is dominant in all settings.

### 3.C.4 Fixed-time confidence intervals (without replacement)

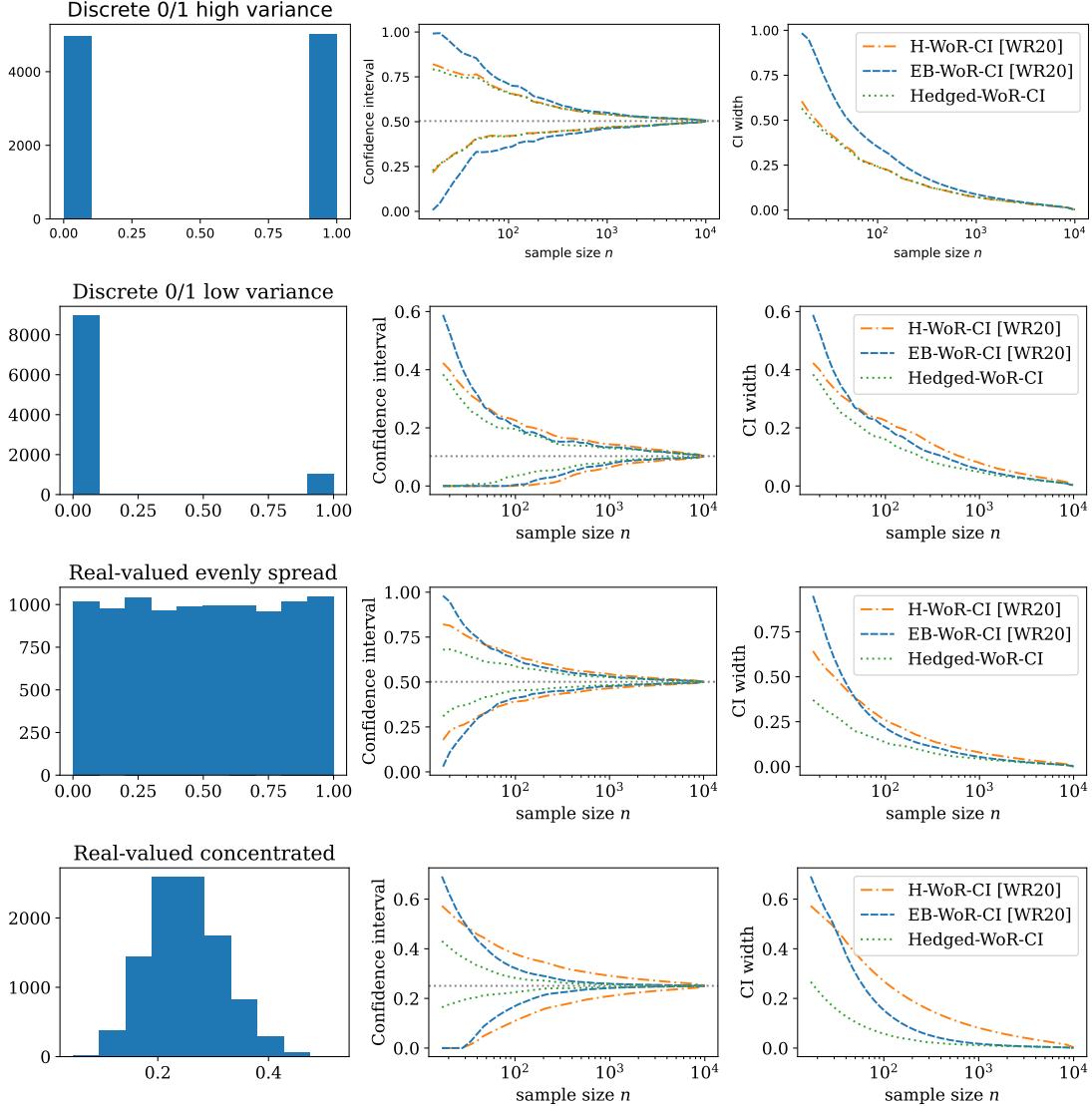


Figure 3.15: Fixed-time hedged capital, Hoeffding-type, and empirical Bernstein-type CIs for the mean of a finite set of bounded numbers when sampling WoR. Further details can be found in Section 3.D.4. Clearly, the two betting approaches (Hedged and ConBo) are dominant in all settings.

Table 3.2: Typical computation time for constructing a CS from time 1 to  $10^3$  for the mean of Bernoulli(1/2)-distributed random variables. The three betting CSs were computed for 1000 evenly-spaced values of  $m$  in  $[0, 1]$ , while a coarser grid would have sped up computation. All CSs were calculated on a laptop powered by a quad-core 2GHz 10th generation Intel Core i5. Parallelization was carried out using the Python library, `multiprocess` [188].

Betting scheme	Interval (a.s.)	Computation time (seconds)
ConBo+LBOW	✓	0.08
Hedged+ $(\lambda_t^{\text{PrPl}\pm})_{t=1}^\infty$	✓	0.25
hgKelly ( $G = 20$ )	✓	1.38
aGRAPA		0.35
LBOW		0.25
ONS- $m$		12.45
Kelly		197.38

## 3.D Simulation details

In each simulation containing confidence sequences or intervals and their widths, we took an average over 5 random draws from the relevant distribution. For example, in the “Time-uniform confidence sequences” plot of Figure 3.1, the CSs (PrPl-H, PrPl-EB, and Hedged) were averaged over 5 random draws from a Beta(10, 30) distribution. Computation times for various strategies are given in Table 3.2.

### 3.D.1 Time-uniform confidence sequences (with replacement)

Each of the CSs considered in the time-uniform (with replacement) case are presented as explicit theorems and propositions throughout the chapter. Specifically,

- **PrPl-H:** Predictable plug-in Hoeffding (Proposition 3.3.1);
- **PrPl-EB:** Predictable plug-in empirical Bernstein (Theorem 3.3.1);
- **Hedged:** Hedged capital process (Theorem 3.4.1); and
- **hgKelly:** Hedged grid-Kelly (Proposition 3.B.1).

**Bernoulli [HRMS20]** Section 3.C compared these against the conjugate mixture sub-Bernoulli confidence sequence by Howard et al. [125], recalled below.

Hoeffding [120, Equation (3.4)], presented the sub-Bernoulli upper-bound on the moment generating function of bounded random variables for any  $\lambda > 0$ :

$$\mathbb{E}_P(\exp\{\lambda(X_i - \mu)\}) \leq 1 - \mu + \mu \exp\{\lambda\},$$

which can be used to construct an  $e$ -value by noting that

$$\mathbb{E}_P\left(\exp\left\{\lambda(X_i - \mu) - \log(1 - \mu + \mu e^\lambda)\right\} \mid \mathcal{F}_{i-1}\right) \leq 1.$$

Then, Howard et al. [125] showed that the cumulative product process

$$\prod_{i=1}^t \left( \exp\left\{\lambda(X_i - \mu) - \log(1 - \mu + \mu e^\lambda)\right\} \right) \quad (3.50)$$

forms a test supermartingale, as does a mixture of (3.50) for any probability distribution  $F(\lambda)$  on  $\mathbb{R}^+$ :

$$\int_{\lambda \in \mathbb{R}^+} \prod_{i=1}^t \left( \exp\left\{\lambda X_i - \log(1 - \mu + \mu e^\lambda)\right\} \right) dF(\lambda). \quad (3.51)$$

In particular, Howard et al. [125] take  $F(\lambda)$  to be a beta distribution so that the integral (3.51) can be computed in closed-form. Using (3.51) in Step (b) in Theorem 3.2.1 yields the “Bernoulli [HRMS20]” confidence sequence.

There are yet other improvements of Hoeffding’s inequality, for example one that goes by the name of Kearns-Saul [151] but was incidentally noted in Hoeffding’s original paper itself. This inequality, and other variants, are looser than the sub-Bernoulli bound and so we exclude them here; see Howard et al. [124] for more details. Most importantly, none of these adapt to the true underlying variance of the random variables, unlike most of our new techniques.

**A-Bentkus [KZ21]** We also compared our bounds against the “adaptive Bentkus confidence sequence” (A-Bentkus) due to Kuchibhotla and Zheng [164, Section 3.5]. These combine a maximal version of Bentkus et al.’s concentration inequality [164, Theorem 1] with the “stitching” technique [295, 190, 125] — a method to obtain infinite-horizon concentration inequalities by taking a union bound over exponentially-spaced *finite* time horizons.

### 3.D.2 Fixed-time confidence intervals (with replacement)

For the fixed-time CIs included from this chapter, we have

- **PrPl-EB-CI:** Predictable plug-in empirical Bernstein CI (Remark 1); and
- **Hedged-CI:** Hedged capital process CI (Remark 3).

These were compared against CIs due to Hoeffding [120], Maurer and Pontil [187], Anderson [9], and Bentkus [28] which we now recall.

**H-CI [H63]** These intervals refer to the CIs based on Hoeffding's classical concentration inequalities [120]. Specifically, for a sample size  $n \geq 1$ , "H-CI [H63]" refers to the CI,

$$\frac{1}{n} \sum_{i=1}^n X_i \pm \sqrt{\frac{\log(2/\alpha)}{2n}}.$$

**Anderson [A69]** These intervals refer to the confidence intervals due to Anderson [9] which take a unique approach by considering the entire sample cumulative distribution function, rather than just the mean and variance. Consequently, however, Anderson's CIs require iid observations, rather than the more general setup we consider. We nevertheless find that even in the iid setting, our approach outperforms Anderson's.

Suppose  $X_1, \dots, X_n \stackrel{iid}{\sim} P$  are  $[0, 1]$ -bounded with mean  $\mathbb{E}_P(X_1) = \mu$ . Let  $X_{(1)}, \dots, X_{(n)}$  denote the order statistics of  $X_1^n$  with the convention that  $X_{(0)} := 0$  and  $X_{(n+1)} := 1$ . Following the notation of Learned-Miller and Thomas [170], Anderson's CI is given by

$$\left[ \sum_{i=1}^n u_i^{\text{DKW}} (-X_{(n-(i+1))} + X_{(n-i)}) , 1 - \sum_{i=1}^n u_i^{\text{DKW}} (X_{(i+1)} - X_{(i)}) \right],$$

where  $u_i^{\text{DKW}} = \left(i/n - \sqrt{\log(2/\alpha)/2n}\right) \vee 0$ . Learned-Miller and Thomas [170, Theorem 2] show that Anderson's CI is always tighter than Hoeffding's. The authors also introduce a bound which is strictly tighter than Anderson's which they conjecture has valid  $(1 - \alpha)$ -coverage, but we do not compare to this bound here.

**EB-CI [MP09]** The empirical Bernstein CI of Maurer and Pontil [187] is given by

$$\frac{1}{n} \sum_{i=1}^n X_i \pm \sqrt{\frac{2\hat{\sigma}^2 \log(4/\alpha)}{n}} + \frac{7 \log(4/\alpha)}{3(n-1)},$$

and  $\hat{\sigma}^2$  is the sample variance.

**Bentkus-CI [B04]** Bentkus' confidence interval requires an a-priori upper bound on  $\text{Var}(X_i)$  for each  $i$ . As alluded to in the introduction, we do not consider concentration bounds which require knowledge of the variance. However, since we assume  $X_i \in [0, 1]$ , we have the trivial upper bound,  $\text{Var}(X_i) \leq \frac{1}{4}$ , which we implicitly use throughout our computation of Bentkus' confidence interval.

Define the independent, mean-zero random variables  $(G_i)_{i=1}^n$  as

$$G_i := \begin{cases} -\frac{1}{4} & \text{w.p. } \frac{4}{5} \\ 1 & \text{w.p. } \frac{1}{5} \end{cases},$$

an important technical device which has appeared in seminal works by Hoeffding [120, Equation

(2.14)] and Bennett [27, Equation (10)]. Then the “Bentkus-CI” is

$$\frac{1}{n} \sum_{i=1}^n X_i \pm \frac{W_\alpha^*}{n},$$

where  $W_\alpha^* \in [0, n]$  is given by the value of  $W_\alpha$  such that

$$\inf_{y \in [0, n] : y \leq W_\alpha} \frac{\mathbb{E} \left[ \sum_{i=1}^n (G_i - y)_+^2 \right]}{(W_\alpha - y)_+^2} = \alpha.$$

Efficient algorithms have been developed to solve the above [29, Section 9], [164].

**PTL- $\ell_2$  [PTL21]** The work by Phan et al. [203] proposes an interesting but computationally intensive approach to constructing confidence intervals for means of iid bounded random variables. Specifically, we will focus on their tightest bound (according to [203, Figure 4]) which makes use of the  $\ell_2$  norm in its derivation (and which we thus refer to as PTL- $\ell_2$ ).

For example, computing PTL- $\ell_2$  confidence intervals<sup>3</sup> from a sample  $X_1, \dots, X_{300} \sim \text{Unif}[0, 1]$  of  $n = 300$  uniformly distributed random variables took upwards of 11 minutes while our betting confidence interval (Remark 3) took less than 0.5 seconds. For this reason, we conduct a small-scale simulation of sample sizes 5-200 (see Figure 3.16). We find that PTL- $\ell_2$  performs extremely well for the low-variance continuous distribution Beta(10, 30) but poorly for sample sizes closer to 200 for Bernoulli data. Nevertheless, PTL- $\ell_2$  requires i.i.d. data (while we only require boundedness and conditional mean  $\mu$ ) and PTL- $\ell_2$  does not have time-uniform or without-replacement analogues.

### 3.D.3 Time-uniform confidence sequences (without replacement)

The WoR CSs which were introduced in this chapter include

- **Hedged-WoR:** Without replacement hedged capital process (Theorem 3.5.1); and
- **hgKelly-WoR:** Without replacement analogue of hgKelly (Proposition 3.B.1).

The CSs labeled “H-WoR [WR20]” and “EB-WoR [WR20]” are the without-replacement Hoeffding- and empirical Bernstein-type CSs found in Chapter 2 which we recall now.

**H-WoR [WR20]** Define the weighted WoR mean estimator and the Hoeffding-type  $\lambda$ -sequence,

$$\hat{\mu}_t^{\text{WoR}}(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i (X_i + \frac{1}{N-i+1} \sum_{j=1}^{i-1} X_j)}{\sum_{i=1}^t \lambda_i (1 + \frac{i-1}{N-i+1})}, \quad \text{and} \quad \lambda_t := \sqrt{\frac{8 \log(2/\alpha)}{t \log(t+1)}} \wedge 1,$$

---

<sup>3</sup>We used code by Phan et al. [203] with their default tuning parameters, available at [github.com/myphan9/small\\_sample\\_mean\\_bounds](https://github.com/myphan9/small_sample_mean_bounds).

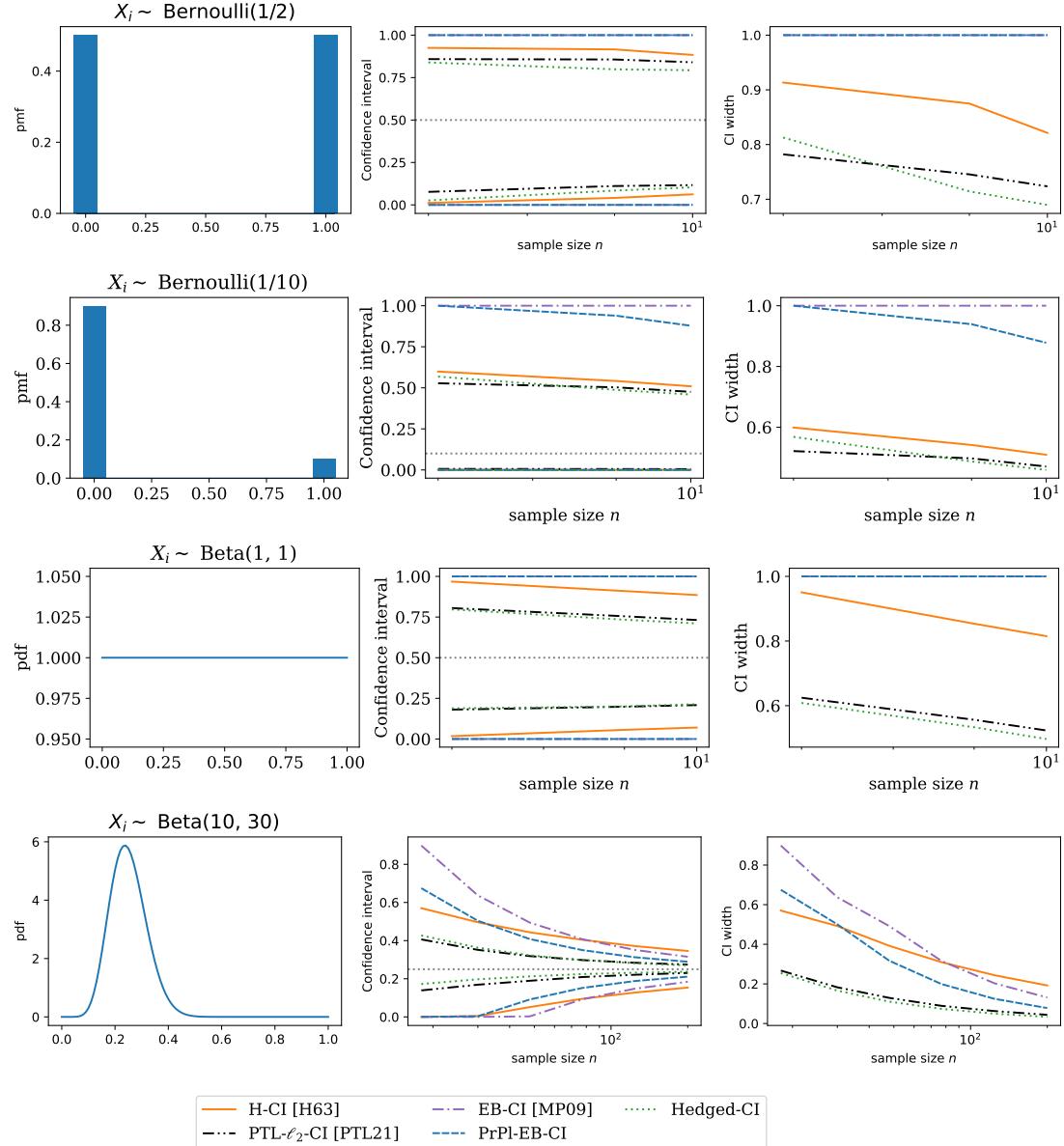


Figure 3.16: Various with-replacement fixed-time confidence intervals, including that of Phan et al. [203] (PTL- $\ell_2$ -CI). While PTL- $\ell_2$ -CI performs very well in the Beta(10, 30) regime, it appears to suffer for Bernoulli(1/2) with larger  $n$ . In any case, PTL- $\ell_2$ -CI relies on iid data, while the other four methods do not.

respectively. Then “H-CS [WR20]” refers to the WoR Hoeffding-type CS,

$$\hat{\mu}_t^{\text{WoR}}(\lambda_1^t) \pm \frac{\sum_{i=1}^t \psi_H(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)}.$$

**EB-WoR [WR20]** Analogously to the Hoeffding-type CSs, “EB-CS [WR20]” corresponds to the empirical Bernstein-type CSs for sampling WoR found in Chapter 2. These CSs take the form

$$\hat{\mu}_t^{\text{WoR}}(\lambda_1^t) \pm \frac{\sum_{i=1}^t 4(X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)},$$

where in this case, we have

$$\lambda_t := \sqrt{\frac{2 \log(2/\alpha)}{\hat{\sigma}_{t-1}^2 t \log(t+1)}} \wedge \frac{1}{2}, \quad \hat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \hat{\mu}_i)^2}{t+1}, \quad \text{and} \quad \hat{\mu}_t := \frac{1}{t} \sum_{i=1}^t X_i. \quad (3.52)$$

### 3.D.4 Fixed-time confidence intervals (without replacement)

The only fixed-time CI introduced in this chapter is **Hedged-WoR-CI**: the without-replacement hedged capital process CI described in Section 3.5. The other two are both found in Chapter 2 which we describe now.

**H-WoR-CI [WR20]** This corresponds to the CI described in Corollary 2.3.1 from Chapter 2. This has the form

$$\hat{\mu}_n^{\text{WoR}} \pm \frac{\sqrt{\frac{1}{2} \log(2/\alpha)}}{\sqrt{n} + \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{i-1}{N-i+1}}.$$

**EB-WoR-CI [WR20]** Similarly, this CI corresponds to that described in Corollary 2.3.2 from Chapter 2. Specifically, “EB-WoR-CI [WR20]” is defined as

$$\hat{\mu}_n^{\text{WoR}}(\lambda_1^n) \pm \frac{\sum_{i=1}^n 4(X_i - \hat{\mu}_{i-1})^2 \psi_E(\lambda_i) + \log(2/\alpha)}{\sum_{i=1}^n \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)},$$

where

$$\lambda_t := \sqrt{\frac{2 \log(2/\alpha)}{n \hat{\sigma}_{t-1}^2}} \wedge \frac{1}{2}, \quad \hat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \hat{\mu}_i)^2}{t+1}, \quad \text{and} \quad \hat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1}, \quad (3.53)$$

and  $\hat{\mu}_n^{\text{WoR}}$  is defined as

$$\hat{\mu}_t^{\text{WoR}}(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i (X_i + \frac{1}{N-i+1} \sum_{j=1}^{i-1} X_j)}{\sum_{i=1}^t \lambda_i \left(1 + \frac{i-1}{N-i+1}\right)}.$$

### 3.D.5 Betting “confidence distributions”: confidence sets at several resolutions

Figures 3.17 and 3.18 demonstrate two tools to visualize CSs at various  $\alpha$  and  $t$ .

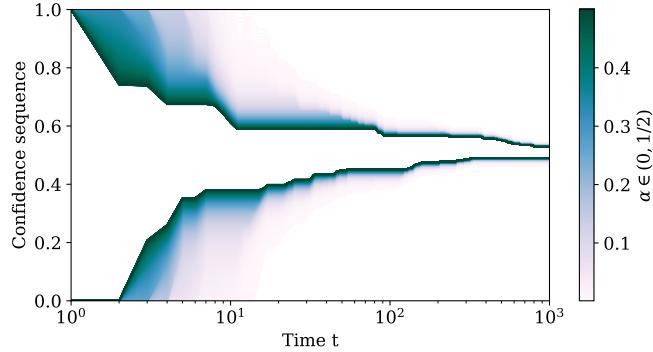


Figure 3.17: This plot shows the aGRAPA CS for all  $\alpha \in [0, 1/2]$  under  $\text{Unif}[0, 1]$  data.

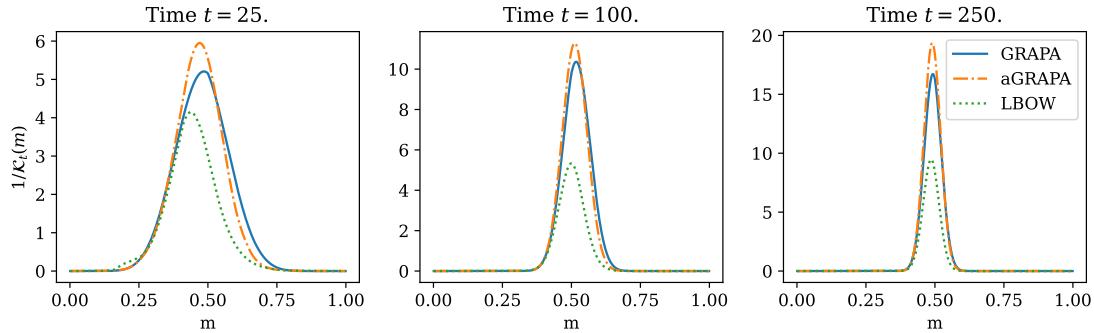


Figure 3.18: Here we plot the inverse wealth  $1/\mathcal{K}_t(m)$  in game  $m \in [0, 1]$ , at  $t = 25, 100, 250$  for three different betting strategies. Note the different  $y$ -axis scales. Despite not being normalized to yield a “confidence distribution”, this is a useful visual tool. For example, the mode in each plot signifies the  $m$  against which we have minimum wealth, which is a reasonable point estimator for  $\mu$ . Further, the superlevel set for any  $\alpha \in [0, 1]$  yields exactly the  $(1 - \alpha)$ -CS for  $\mu$  (for that corresponding time and strategy) since it yields all  $m$  with wealth less than  $1/\alpha$ . Last, for any  $m \in [0, 1]$ , the height (truncated at one) is anytime-valid  $p$ -value for the null hypothesis that the mean equals  $m$ .

## 3.E Additional theoretical results

### 3.E.1 Betting confidence sets are tighter than Hoeffding

In this section, we demonstrate that the betting approach can dominate Hoeffding for sufficiently large sample sizes. First, we show that for any  $x, m \in (0, 1)$  and any  $\lambda \in \mathbb{R}$ , then  $\gamma \equiv \gamma^m(\lambda)$  can be set as

$$\gamma^m(\lambda) := \exp\{-m\lambda - \lambda^2/8\} (\exp(\lambda) - 1),$$

so that

$$H^m(x) := \underbrace{\exp\{\lambda(x-m) - \lambda^2/8\}}_{\text{Hoeffding term}} \leq \underbrace{1 + \gamma(x-m)}_{\text{Capital process term}} =: \mathcal{K}^m(x)$$

for any  $x, m \in [0, 1]$ . In particular, the Hoeffding-type and capital process supermartingales are built from precisely the above terms, respectively, and so if  $H^m(x) \leq \mathcal{K}^m(x)$  for any  $x \in [0, 1]$ , then their respective supermartingales will satisfy the same inequality almost surely.

**Proposition 3.E.1** (Capital process dominates Hoeffding process). *Suppose  $x, m \in [0, 1]$  and  $\lambda \in \mathbb{R}$ . Then there exists  $\gamma^m(\lambda) \in \mathbb{R}$  such that*

$$H^m(x) := \exp(\lambda(x-m) - \lambda^2/8) \leq 1 + \gamma^m(\lambda)(x-m) =: \mathcal{K}^m(x).$$

Note that Proposition 3.E.1 alone does not confirm that the Hoeffding-based CIs will be dominated by capital process-based CIs since  $\gamma$  must be within  $[-1/(1-m), 1/m]$  for  $\mathcal{K}^m(x)$  to be nonnegative. However, it is easy to verify that for all  $\lambda \in [-0.45, 0.45]$ , we have that  $\gamma \in [-1, 1]$  and thus  $\mathcal{K}^m(x) \geq 0$ . When constructing a Hoeffding-type  $(1-\alpha)$ -confidence interval, for example, one would set  $\lambda_n^H := \sqrt{8 \log(2/\alpha)/n}$ , making  $\lambda_n^H \in [-0.45, 0.45]$  whenever  $n \geq 40 \log(2/\alpha)$ , in which case a capital process-based CI will dominate a Hoeffding-based CI almost surely.

*Proof of Proposition 3.E.1.* We prove the result for  $\lambda \geq 0$  and remark that this implies the result for the case when  $\lambda \leq 0$  by considering  $(1-x)$  and  $(1-m)$  instead of  $x$  and  $m$ , respectively.

The proof proceeds in 3 steps. First, we consider the line segment  $L^m(x)$  connecting  $H^m(0)$  and  $H^m(1)$  and note that by convexity of  $H^m(x)$ , we have that  $H^m(x) \leq L^m(x)$  for all  $x \in [0, 1]$ . We then find the slope of this line segment and set  $\gamma$  to this value so that the line  $\mathcal{K}^m(x) := 1 + \gamma(x-m)$  has the same slope as  $L^m(x)$ . Finally, we demonstrate that  $L^m(0) \leq \mathcal{K}^m(0)$ , and conclude that  $H^m(x) \leq L^m(x) \leq \mathcal{K}^m(x)$  for all  $x \in [0, 1]$ .

**Step 1.** Note that  $H^m(x)$  is a convex function in  $x \in [0, 1]$ , and thus

$$\forall x \in [0, 1], H^m(x) \leq H^m(0) + [H^m(1) - H^m(0)]x =: L^m(x).$$

**Step 2.** Observe that the slope of  $L^m(x)$  is  $H^m(1) - H^m(0)$ . Setting  $\gamma := H^m(1) - H^m(0)$  we have that  $\mathcal{K}^m(x)$  and  $L^m(x)$  are parallel.

**Step 3.** It remains to show that  $\mathcal{K}^m(0) \geq H^m(0) \equiv H^m(0)$  for every  $m \in [0, 1]$ . Consider the following equivalent statements:

$$\begin{aligned} & \mathcal{K}^m(0) \geq H^m(0) \\ \iff & 1 - m [H^m(1) - H^m(0)] \geq H^m(0) \\ \iff & 1 - m \exp(\lambda - \lambda m - \lambda^2/8) \geq (1 - m) \exp(-\lambda m - \lambda^2/8) \\ \iff & 1 \geq \exp(-\lambda m - \lambda^2/8) [1 - m + m \exp(\lambda)] \\ \iff & \exp(\lambda m + \lambda^2/8) \geq [1 - m + m \exp(\lambda)] \\ \iff & a(\lambda) := \exp(\lambda m + \lambda^2/8) - [1 - m + m \exp(\lambda)] \geq 0. \end{aligned}$$

Now, note that  $a$  is smooth and  $a(0) = 0$  and so it suffices to show that its derivative  $a'(\lambda) \geq 0$  for all  $\lambda \geq 0$ . To this end, consider the following equivalent statements.

$$\begin{aligned} & a'(\lambda) \equiv \left(m + \frac{\lambda}{4}\right) \exp(\lambda m + \lambda^2/8) - m \exp(\lambda) \geq 0 \\ \iff & \left(m + \frac{\lambda}{4}\right) \exp(\lambda m + \lambda^2/8) \geq m \exp(\lambda) \\ \iff & \ln\left(1 + \frac{\lambda}{4m}\right) + \lambda m + \lambda^2/8 \geq \lambda \\ \iff & b(\lambda) := \ln\left(1 + \frac{\lambda}{4m}\right) + \lambda m + \lambda^2/8 - \lambda \geq 0, \end{aligned}$$

and hence it suffices to show that  $b(\lambda) \geq 0$ . Similar to  $a(\lambda)$ , we have that  $b(0) = 0$  and so it suffices to show that its derivative,  $b'(\lambda) \geq 0$  for all  $\lambda \geq 0$ . Indeed,

$$\begin{aligned} & b'(\lambda) \equiv \frac{1}{4m + \lambda} + m + \frac{\lambda}{4} - 1 \geq 0 \\ \iff & c(\lambda) := 1 + m(4m + \lambda) + \frac{\lambda}{4}(4m + \lambda) - 4m - \lambda \geq 0 \end{aligned}$$

Since  $c(\lambda)$  is a convex quadratic, it is straightforward to check that

$$\operatorname{argmin}_{\lambda \in \mathbb{R}} c(\lambda) = 2 - 4m,$$

and that  $c(2 - 4m) = 0$ . In conclusion, if we set  $\gamma \equiv \gamma^m(\lambda)$  as

$$\gamma^m(\lambda) := H^m(1) - H^m(0) = \exp\{-m\lambda - \lambda^2/8\} (\exp(\lambda) - 1),$$

then  $H^m(x) \leq \mathcal{K}^m(x) := 1 + \gamma^m(\lambda)(x - m)$  for every  $m \in [0, 1]$ . This completes the proof.  $\square$

### 3.E.2 Optimal convergence of betting confidence sets

In Section 3.B, it was mentioned that for nonnegative martingales, Ville's inequality is nearly an equality and hence martingale-based CSs are nearly tight in a time-uniform sense. However, it is natural to wonder what other theoretical guarantees betting CSs/CIs can have in addition to their empirical performance. In the time-uniform setting, CSs for the mean cannot attain widths which scale faster than  $\asymp \sqrt{\log \log t/t}$ , due to the law of the iterated logarithm. Similarly, fixed-time CIs cannot scale faster than  $\asymp 1/\sqrt{n}$ . In this section, we show that it is possible to choose betting strategies such that the resulting CSs and CIs scale at the optimal rates of  $O(\sqrt{\log \log t/t})$  and  $O(1/\sqrt{n})$ , respectively.

#### 3.E.2.1 An iterated logarithm betting confidence sequence

We will establish the law of the iterated logarithm (LIL) convergence rate by carefully constructing a capital process martingale whose resulting CS is – for sufficiently large  $t$  – tighter than a larger CS which itself attains the required LIL rate.

Before stating the result in Proposition 3.E.2, let  $\zeta(s) := \sum_{k=1}^{\infty} \frac{1}{k^s}$  be the Riemann zeta function and for each  $k \in \{1, 2, \dots\}$ , define

$$\begin{aligned}\lambda_k &:= \sqrt{\frac{8 \log(k^s \zeta(s))}{\eta^{k+1/2}}}, \quad \text{and} \\ \gamma_k(m) &= \exp\{-m\lambda_k - \lambda_k^2/8\} (\exp(\lambda_k) - 1) \wedge 1,\end{aligned}$$

where  $\eta > 1$  is some user-chosen constant. Let  $k_t$  denote the (unique) integer such that  $\log_{\eta} t \leq k_t \leq \log_{\eta} t + 1$ . Define the process

$$\begin{aligned}\mathcal{K}_t^{\mathcal{L}} &:= \frac{1}{2} \mathcal{K}_t^{\mathcal{L}+}(m) + \frac{1}{2} \mathcal{K}_t^{\mathcal{L}-}(m) \\ \text{where } \mathcal{K}_t^{\mathcal{L}+}(m) &:= \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^t (1 + \gamma_{k_t}(X_i - m)) \quad \text{and} \\ \mathcal{K}_t^{\mathcal{L}-}(m) &:= \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^t (1 - \gamma_{k_t}(X_i - m)).\end{aligned}$$

Note that  $\mathcal{K}_t^{\mathcal{L}+}(m)$  and  $\mathcal{K}_t^{\mathcal{L}-}(m)$  are both upper-bounded by the infinite mixtures

$$\mathcal{K}_t^{\mathcal{L}+}(m) \leq \sum_{k=1}^{\infty} \frac{1}{k^s \zeta(s)} \prod_{i=1}^t (1 + \gamma_k(X_i - m)) \quad \text{and} \tag{3.54}$$

$$\mathcal{K}_t^{\mathcal{L}-}(m) \leq \sum_{k=1}^{\infty} \frac{1}{k^s \zeta(s)} \prod_{i=1}^t (1 - \gamma_k(X_i - m)), \tag{3.55}$$

which themselves form nonnegative martingales when  $m = \mu$  by Fubini's theorem. Conse-

quently,

$$C_t^{\mathcal{L}} := \left\{ m \in [0, 1] : \mathcal{K}_t^{\mathcal{L}}(m) < \frac{1}{\alpha} \right\}$$

forms a  $(1 - \alpha)$ -CS for  $\mu$ . The following proposition establishes the LIL rate of  $C_t^{\mathcal{L}}$ .

**Proposition 3.E.2.** *The CS  $(C_t^{\mathcal{L}})_{t=1}^{\infty}$  has a width of  $O(\sqrt{\log \log t/t})$ , meaning*

$$\nu(C_t^{\mathcal{L}}) = O\left(\sqrt{\frac{\log \log t}{t}}\right),$$

where  $\nu$  is the Lebesgue measure.

*Proof.* The proof proceeds in three steps. In Step 1, we construct a distinct but related CS (which we will denote by  $(C_t^{\times})_{t=1}^{\infty}$ ) via the stitching technique [125]. In Step 2, we demonstrate that this stitched CS achieves the desired rate by deriving an analytically tractible superset whose width scales as  $O(\sqrt{\log \log t/t})$ . Finally, in Step 3, we will show that the stitched CS  $C_t^{\times}$  is a superset of  $C_t^{\mathcal{L}}$  for all  $t$  sufficiently large, thus implying the final result.

**Step 1. Constructing the stitched CS  $C_t^{\times}$ :** In the language of betting, the idea behind stitching is to first divide one's capital up into infinitely many portions  $w_1, w_2, \dots$  such that  $\sum_{k=1}^{\infty} w_k = 1$ , and then place a constant bet  $\lambda_k$  using a capital of  $w_k$  on a designated epoch of time, which will be chosen to be geometrically spaced. In what follows, the portions  $w_k$  will be given by  $w_k = \frac{1}{\zeta(s)k^s}$ , and we will divide time  $\{1, 2, 3, \dots\}$  up into epochs demarcated by the endpoints  $\eta^{k-1}$  and  $\eta^k$  for each  $k \in \{1, 2, 3, \dots\}$  and for some user-specified  $\eta > 1$  (e.g.  $\eta = 1.1$ ). The constant bets  $\lambda_k$  will be chosen so that they are effective between  $\eta^{k-1}$  and  $\eta^k$  and lead to  $O(\sqrt{\log \log t/t})$  widths after being combined across epochs.

The construction of the stitched boundary essentially follows (a simplified version of) the proof of Theorem 1 in Howard et al. [125, Section A.1], but we present the derivation here for completeness. Consider the Hoeffding-type process for a fixed  $\lambda \in \mathbb{R}$ :

$$M_t^{\lambda}(m) := \exp\left\{\lambda S_t(m) - t\lambda^2/8\right\}, \quad (3.56)$$

where  $S_t(m) := \sum_{i=1}^t (X_i - m)$ . As discussed in Section 3.3,  $M_t(\mu)$  forms a test supermartingale, and hence by Ville's inequality we have

$$P\left(\exists t \geq 1 : S_t(\mu) \geq \underbrace{\frac{r + t\lambda^2/8}{\lambda}}_{g_{\lambda,r}(t)}\right) \leq e^{-r}.$$

We have typically used  $r = \log(1/\alpha)$  throughout the chapter, but the above alternative notation will help in the following discussion. Using the notation of Howard et al. [125, Section A.1],

define the boundary above as  $g_{\lambda,r}(t) := (r + t\lambda^2/8)/\lambda$ , and let

$$\lambda_k := \sqrt{\frac{8r_k}{\eta^{k-1/2}}},$$

where  $r_k := \log\left(\frac{k^s\zeta(s)}{\alpha/2}\right)$ .

Some algebra will reveal that plugging the above choices of  $\lambda_k$  and  $r_k$  into  $g_{\lambda,r}(t)$  yields

$$g_{\lambda_k, r_k}(t) := \sqrt{\frac{r_k t}{8}} \left( \sqrt{\frac{\eta^{k-1/2}}{t}} + \sqrt{\frac{t}{\eta^{k-1/2}}} \right),$$

resulting in the following concentration inequality for each  $k$ :

$$P(\exists t \geq 1 : S_t(\mu) \geq g_{\lambda_k, r_k}(t)) \leq \exp\{-r_k\}.$$

Let  $k_t$  denote the (unique) epoch number such that  $\eta^{k_t-1} \leq t \leq \eta^{k_t}$  (i.e. such that  $\log_\eta t \leq k_t \leq \log_\eta t + 1$ ). Now, we take a union bound over  $k = 1, 2, 3, \dots$  resulting in the following boundary,

$$P(\exists t \geq 1 : S_t(\mu) \geq g_{\lambda_{k_t}, r_{k_t}}(t)) \leq \sum_{k=1}^{\infty} \exp\{-r_k\} = \frac{\alpha/2}{\zeta(s)} \underbrace{\sum_{k=1}^{\infty} \frac{1}{k^s}}_{\zeta(s)} = \alpha/2.$$

Repeating all of the previous steps for  $-S(\mu)$  and taking a union bound, we arrive at the  $(1 - \alpha)$  stitched CS  $(C_t^\times)_{t=1}^\infty$  given by

$$C_t^\times := \left( \frac{1}{t} \sum_{i=1}^t X_i \pm \frac{g_{\lambda_{k_t}, r_{k_t}}(t)}{t} \right),$$

with the guarantee that  $P(\exists t \geq 1 : \mu \notin C_t^\times) \leq \alpha$ .

**Step 2. Demonstrating that  $C_t^\times$  achieves the desired LIL width:** Now, we will simply upper-bound  $g_{\lambda_{k_t}, r_{k_t}}(t)$  by an analytical boundary depending explicitly on  $t$  (rather than implicitly through  $k_t$ ) to see that it achieves the desired LIL width. First, notice that  $\sqrt{\eta^{k_t-1/2}/t} + \sqrt{t/\eta^{k_t-1/2}}$  is uniquely minimized when  $t = \eta^{k_t-1/2}$  and hence its maximum on the interval  $(\eta^{k_t-1}, \eta^{k_t})$  must be at the endpoints. Therefore,  $\sqrt{\eta^{k_t-1/2}/t} + \sqrt{t/\eta^{k_t-1/2}} \leq \eta^{1/4} + \eta^{-1/4}$  and thus for each  $k$ , we have

$$g_{\lambda_{k_t}, r_{k_t}}(t) \leq \sqrt{\frac{r_{k_t} t}{8}} \left( \eta^{1/4} + \eta^{-1/4} \right) \quad \text{for all } \eta^{k_t-1} \leq t \leq \eta^{k_t}.$$

Furthermore, for all  $\eta^{k_t-1} \leq t \leq \eta^{k_t}$ , we have that  $k_t \leq \log_\eta t + 1$ . Applying this inequality to the above, we obtain the final bound which does not depend on  $k$ ,

$$g_{\lambda_{k_t}, r_{k_t}}(t) \leq \sqrt{\frac{t \log(2(\log_\eta t + 1)^s \zeta(s)/\alpha)}{8}} (\eta^{1/4} + \eta^{-1/4}) \quad \text{for all } k.$$

In conclusion, we have that

$$C_t^\times \subseteq \left( \frac{1}{t} \sum_{i=1}^t X_i \pm \sqrt{\frac{\log(2(\log_\eta t + 1)^s \zeta(s)/\alpha)}{8t}} (\eta^{1/4} + \eta^{-1/4}) \right),$$

and thus  $C_t^\times = O(\sqrt{\log \log t / t})$ , as desired.

**Step 3. Showing that  $C_t^{\mathcal{L}} \subseteq C_t^\times$  for all  $t$  large enough:** This step in the proof essentially follows immediately from the discussion in Section 3.E.1. We justified that for  $\lambda \geq 0$ , setting  $\gamma$  as

$$\gamma = \exp\{-m\lambda - \lambda^2/8\} (\exp(\lambda) - 1) \wedge 1,$$

yields  $1 + \gamma(x - m) \geq \exp\{\lambda(x - m) - \lambda^2/8\}$  for all  $x, m \in [0, 1]$  if  $\lambda$  is sufficiently small (i.e. so that  $\gamma$  is not relying on truncation at 1). Since  $\lambda_k$  is decreasing in  $t$ , it follows that for  $t$  sufficiently large,

$$\prod_{i=1}^t (1 + \gamma_{k_t}(X_i - m)) \geq \exp\{\lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8\} \quad \text{almost surely.}$$

Therefore, for  $t$  sufficiently large,

$$\begin{aligned} \mathcal{K}_t^{\mathcal{L}+}(m) &:= \frac{1}{k_t^s \zeta(s)} \prod_{i=1}^t (1 + \gamma_{k_t}(X_i - m)) \\ &\geq \frac{1}{k_t^s \zeta(s)} \exp\{\lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8\} =: H_t^{\circ,+}(m) \end{aligned}$$

and similarly for  $K_t^{\mathcal{L}-}(m)$ ,

$$\mathcal{K}_t^{\mathcal{L}-}(m) \geq \frac{1}{k_t^s \zeta(s)} \exp\{-\lambda_{k_t} S_t(m) - \lambda_{k_t}^2/8\} =: H_t^{\circ,-}(m).$$

Therefore, for sufficiently large  $t$ , we have

$$C_t^{\mathcal{L}} := \left\{ m \in [0, 1] : \mathcal{K}_t^{\mathcal{L}}(m) < \frac{1}{\alpha} \right\}$$

$$\subseteq \underbrace{\left\{ m \in \mathbb{R} : \max \left\{ \frac{1}{2} H_t^{\infty+}(m), \frac{1}{2} H_t^{\infty-}(m) \right\} < \frac{1}{\alpha} \right\}}_{(\star)}$$

and it is straightforward to verify that  $(\star)$  is precisely  $C_t^\times$ .

In summary, we constructed a CS  $C_t^\times$  using the stitching technique in Step 1, and then showed that  $\nu(C_t^\times) = O(\sqrt{\log \log t/t})$  in Step 2. Finally in Step 3, we showed that our discrete mixture betting CS  $C_t^L$  is a subset of  $C_t^\times$  for  $t$  sufficiently large, and hence by subadditivity of measures,

$$\nu(C_t^L) = O\left(\sqrt{\frac{\log \log t}{t}}\right),$$

which completes the proof.  $\square$

*Remark 6.* Notice that  $\mathcal{K}_t^{L+}$  and  $\mathcal{K}_t^{L-}$  can be made strictly more powerful if they are replaced by adding additional terms, as long as the final sums are upper-bounded by (3.54) and (3.55), respectively. In particular, any finite sum analogue of (3.54) and (3.55) would have sufficed, as long as  $\mathcal{K}_t^{L+}$  and  $\mathcal{K}_t^{L-}$  form a term in each sum, respectively. We presented  $\mathcal{K}_t^{L+}$  and  $\mathcal{K}_t^{L-}$  in their current forms for the sake of notational (and computational) simplicity.

### 3.E.2.2 The $\sqrt{n}$ -convergence of betting CIs

**Proposition 3.E.3.** Suppose  $X_1^n \sim P$  are independent observations from a distribution  $P \in \mathcal{P}^\mu$  with mean  $\mu \in [0, 1]$ . Let  $\lambda_n \in (0, 1)$  such that  $\lambda_n \asymp 1/\sqrt{n}$ . Then the confidence interval,

$$C_n := \left\{ m \in [0, 1] : \mathcal{K}_n^\pm < \frac{1}{\alpha} \right\} \text{ has an asymptotic width of } O(1/\sqrt{n}).$$

*Proof.* Writing out the capital process with positive bets, we have by Lemma 3.B.1 that for any  $m \in [0, 1]$ ,

$$\begin{aligned} \mathcal{K}_n^+(m) &:= \prod_{i=1}^n (1 + \lambda_n(X_i - m)) \\ &\geq \exp \left( \lambda_n \sum_{i=1}^n (X_i - m) - \psi_E(\lambda_n) \sum_{i=1}^n 4(X_i - m)^2 \right) \\ &\geq \exp \left( \lambda_n \sum_{i=1}^n (X_i - m) - 4n\psi_E(\lambda_n) \right) =: B_t^+(m), \end{aligned}$$

and similarly for negative bets,

$$\mathcal{K}_n^-(m) := \prod_{i=1}^n (1 - \lambda_n(X_i - m))$$

$$\geq \exp \left( -\lambda_n \sum_{i=1}^t (X_i - m) - 4n\psi_E(\lambda_n) \right) =: B_t^-(m).$$

For any  $\theta \in (0, 1)$ , consider the set,

$$\mathcal{S}_n := \left\{ m : B_t^+(m) < \frac{1}{\theta\alpha} \right\} \cap \left\{ m : B_t^-(m) < \frac{1}{(1-\theta)\alpha} \right\}$$

Now notice that the  $1/\alpha$ -level set of  $\mathcal{K}_n^\pm(m) := \max \{\theta\mathcal{K}_n^+(m), (1-\theta)\mathcal{K}_n^-(m)\}$  is a subset of  $\mathcal{S}_n$ :

$$C_n = \left\{ m : \mathcal{K}_n^+(m) < \frac{1}{\theta\alpha} \right\} \cap \left\{ m : \mathcal{K}_n^-(m) < \frac{1}{(1-\theta)\alpha} \right\} \subseteq \mathcal{S}_n.$$

On the other hand, it is straightforward to derive a closed-form expression for  $\mathcal{S}_n$ :

$$\left( \frac{\sum_{i=1}^n X_i}{n} - \frac{\log(\frac{1}{\theta\alpha}) + 4n\psi_E(\lambda_n)}{n\lambda_n}, \frac{\sum_{i=1}^n X_i}{n} + \frac{\log(\frac{1}{(1-\theta)\alpha}) + 4n\psi_E(\lambda_n)}{n\lambda_n} \right),$$

which in the typical case of  $\theta = 1/2$  has the cleaner expression,

$$\frac{\sum_{i=1}^n X_i}{n} \pm \frac{\log(2/\alpha) + 4n\psi_E(\lambda_n)}{n\lambda_n}.$$

As discussed in Section 3.B, we have by two applications of L'Hôpital's rule that  $\frac{\psi_E(\lambda_n)}{\psi_H(\lambda_n)} \xrightarrow{n \rightarrow \infty} 1$ , where  $\psi_H(\lambda_n) := \lambda_n^2/8 \asymp 1/n$  and thus the width  $W_n$  of  $\mathcal{S}_n$  scales as

$$W_n := 2 \cdot \frac{\log(1/\alpha) + 4n\psi_E(\lambda_n)}{n\lambda_n} \asymp \frac{\log(1/\alpha)}{\sqrt{n}} + \frac{4n/n}{\sqrt{n}} \asymp \frac{1}{\sqrt{n}}.$$

Since  $C_n \subseteq \mathcal{S}_n$ , we have that  $C_n$  has a width of  $O(1/\sqrt{n})$ , which completes the proof.  $\square$

Despite these results, the hedged capital CI presented and recommended in Section 3.4.4 does not satisfy the assumptions of the above proof. In particular, we recommended using the variance-adaptive predictable plug-in,

$$\lambda_t^{\text{PrPl-EB}(n)} := \sqrt{\frac{2 \log(2/\alpha)}{n\hat{\sigma}_{t-1}^2}}, \quad \hat{\sigma}_t^2 := \frac{1/4 + \sum_{i=1}^t (X_i - \hat{\mu}_i)^2}{t+1}, \quad \text{and} \quad \hat{\mu}_t := \frac{1/2 + \sum_{i=1}^t X_i}{t+1}, \quad (3.57)$$

using a truncation which depends on  $m$ ,

$$\lambda_t^+(m) := \lambda_t^\pm \wedge \frac{c}{m}, \quad \lambda_t^-(m) := - \left( \lambda_t^\pm \wedge \frac{c}{1-m} \right), \quad (3.58)$$

and finally defining the hedged capital process for each  $t \in \{1, \dots, n\}$ :

$$\mathcal{K}_t^\pm(m) := \max \left\{ \theta \prod_{i=1}^t (1 + \lambda_i^+(m) \cdot (X_i - m)), (1 - \theta) \prod_{i=1}^t (1 - \lambda_i^-(m) \cdot (X_i - m)) \right\}.$$

Furthermore, the resulting CI is defined as an intersection,

$$\mathfrak{B}_n := \bigcap_{t=1}^n \left\{ m \in [0, 1] : \mathcal{K}_t^\pm(m) < \frac{1}{\alpha} \right\}. \quad (3.59)$$

All of these tweaks (i.e. making bets predictable, truncating beyond  $(0, 1)$ , and taking an intersection) do not in any way invalidate the type-I error, but we find (through simulations) that they tighten the CIs, especially in low-variance, asymmetric settings (see Figure 3.19).

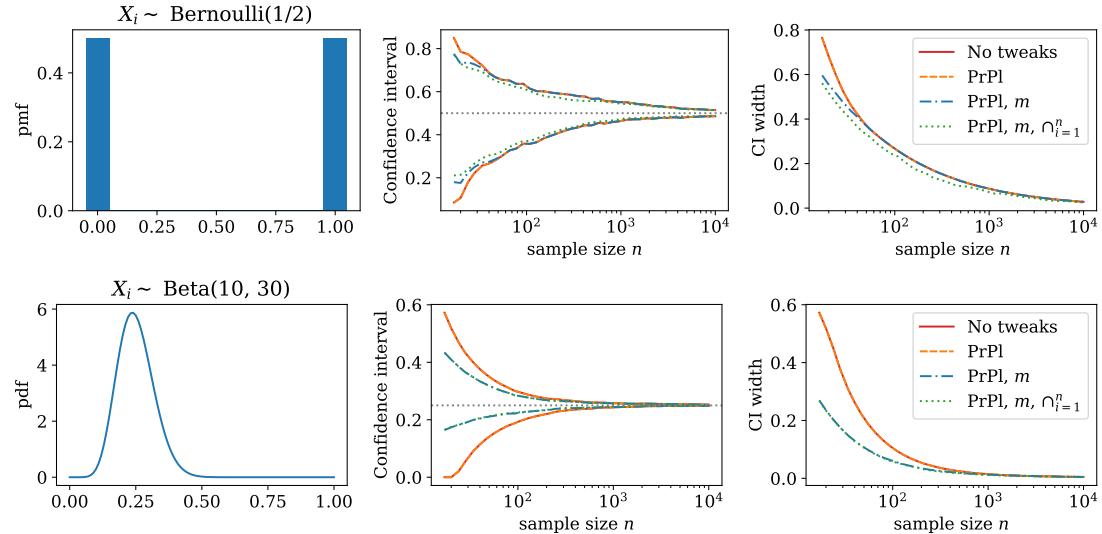


Figure 3.19: Hedged capital CIs with various added tweaks. The CIs labeled “No tweaks” refer to those which satisfy the conditions of Proposition 3.E.3. The other three plots differ in which “tweaks” have been added. Those with “PrPl” in the legend use the predictable plug-in approach defined in (3.57); those with  $m$  in the legend have been truncated using  $m$  as outlined in (3.58); finally, the plots with  $\cap_{i=1}^n$  in their legends had their running intersections taken as in (3.59).

### 3.E.3 On the width of empirical Bernstein confidence intervals

Recall the predictable plug-in empirical Bernstein confidence interval:

$$C_n^{\text{PrPl-EB}(n)} := \left( \frac{\sum_{i=1}^n \lambda_i X_i}{\sum_{i=1}^n \lambda_i} \pm \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i} \right),$$

where

$$\lambda_t := \sqrt{\frac{2 \log(2/\alpha)}{n \hat{\sigma}_{t-1}^2}}, \quad \hat{\sigma}_t^2 := \frac{\frac{1}{4} + \sum_{i=1}^t (X_i - \hat{\mu}_i)^2}{t+1}, \quad \text{and} \quad \hat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t X_i}{t+1}.$$

Below, we analyze the asymptotic behavior of the width of  $C_n^{\text{PrPl-EB}(n)}$  in the i.i.d. setting. In Proposition 3.E.4, we will show that if the data are drawn i.i.d. from a distribution  $Q \in \mathcal{Q}^\mu$  having variance  $\sigma^2$ , then the half-width  $W_n$  of  $C_n^{\text{PrPl-EB}(n)}$  scales as

$$\sqrt{n}W_n \equiv \sqrt{n} \left( \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i} \right) \xrightarrow{a.s.} \sigma \sqrt{2 \log(2/\alpha)}, \quad (3.60)$$

and hence the width is asymptotically proportional to the standard deviation.

First, let us prove a few lemmas about *nonrandom* sequences of numbers, which will be helpful in what follows. These are simple facts for which we could not find a proof to reference, so we prove them below for completeness.

**Lemma 3.E.1.** *Suppose  $(a_n)_{n=1}^\infty$  is a sequence of real numbers such that  $a_n \rightarrow a$ . Then their cumulative average also converges to  $a$ , meaning that  $\frac{1}{n} \sum_{i=1}^n a_i \rightarrow a$ .*

*Proof.* Let  $\epsilon > 0$  and choose  $N \equiv N_\epsilon \in \mathbb{N}$  such that whenever  $n \geq N$ , we have

$$|a_n - a| < \epsilon. \quad (3.61)$$

Moreover, choose

$$M \equiv M_N > \frac{\sum_{i=1}^N |a_i - a|}{\epsilon} \quad (3.62)$$

and note that

$$\frac{n - N - 1}{n} < 1. \quad (3.63)$$

Let  $n \geq \max\{N, M\}$ . Then we have by the triangle inequality,

$$\begin{aligned} \left| \frac{1}{n} \sum_{i=1}^n (a_i - a) \right| &\leq \frac{1}{n} \sum_{i=1}^N |a_i - a| + \frac{1}{n} \sum_{i=N+1}^n |a_i - a| \\ &\leq \frac{1}{n} \sum_{i=1}^N |a_i - a| + \frac{1}{n} (n - N - 1) \epsilon && \text{by (3.61)} \\ &\leq 2\epsilon && \text{by (3.62) and (3.63),} \end{aligned}$$

which can be made arbitrarily small. This completes the proof of Lemma 3.E.1.  $\square$

**Lemma 3.E.2.** *Let  $(a_n)_{n=1}^\infty$  and  $(b_n)_{n=1}^\infty$  be sequences of numbers such that*

$$a_n \rightarrow 0 \quad \text{and} \quad (3.64)$$

$$|b_n| \leq C \text{ for some } C \geq 0 \text{ and for all } n \geq 1. \quad (3.65)$$

Then  $a_n b_n \rightarrow 0$ . Further, if  $(A_n)$  is a sequence of random variables such that  $A_n \rightarrow 0$  almost surely, then  $A_n b_n \rightarrow 0$  almost surely.

The proof is trivial, since  $|A_n b_n| \leq C |A_n|$  which converges to zero almost surely.  $\square$

Now, we prove that a modified variance estimator is consistent.

**Lemma 3.E.3.** *Let  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} Q \in \mathcal{Q}^\mu$  with  $\text{Var}(X_i) = \sigma^2$ . Then the modified variance estimator*

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_{i-1})^2$$

converges to  $\sigma^2$ ,  $Q$ -almost surely.

*Proof.* By direct substitution,

$$\begin{aligned} \hat{\sigma}_n^2 &:= \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_{i-1})^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu + \mu - \hat{\mu}_{i-1})^2 \\ &= \underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2}_{\xrightarrow{a.s.} \sigma^2} - \underbrace{\frac{2}{n} \sum_{i=1}^n (X_i - \hat{\mu}_{i-1})(\hat{\mu}_{i-1} - \mu)}_{(*)} + \underbrace{\frac{1}{n} \sum_{i=1}^n (\mu - \hat{\mu}_{i-1})^2}_{(**)}. \end{aligned}$$

Now, note that  $\hat{\mu}_{i-1} - \mu \xrightarrow{a.s.} 0$  and  $|X_i - \hat{\mu}_{i-1}| \leq 1$  for each  $i$ . Therefore, by Lemma 3.E.2,  $(X_i - \hat{\mu}_{i-1})(\hat{\mu}_{i-1} - \mu) \xrightarrow{a.s.} 0$ , and by Lemma 3.E.1,  $(*) \xrightarrow{a.s.} 0$ . Furthermore, we have that  $(\mu - \hat{\mu}_{i-1})^2 \xrightarrow{a.s.} 0$  and so by another application of Lemma 3.E.1, we have  $(**) \xrightarrow{a.s.} 0$ . This completes the proof of Lemma 3.E.3.  $\square$

Next, let us analyze the second term in the numerator in the margin of  $C_n^{\text{PrPl-EB}(n)}$ ,

$$\frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i}. \quad (3.66)$$

**Lemma 3.E.4.** *Under the same assumptions as Lemma 3.E.3,*

$$\sum_{i=1}^n v_i \psi_E(\lambda_i) \xrightarrow{a.s.} \log(2/\alpha).$$

*Proof.* Recall that  $\frac{\psi_E(\lambda)}{\psi_H(\lambda)} \xrightarrow{\lambda \rightarrow 0} 1$ , and  $\hat{\sigma}_t^2 \xrightarrow{t \rightarrow \infty} \sigma^2$ . By definition of  $\lambda_i$ , we have that  $\lambda_i \xrightarrow{a.s.} 0$

and thus we may also write

$$\frac{\psi_E(\lambda_i)}{\psi_H(\lambda_i)} = 1 + R_i \text{ and} \quad (3.67)$$

$$\sqrt{\frac{\sigma^2}{\hat{\sigma}_t^2}} = 1 + R'_i \quad (3.68)$$

for some  $R_i, R'_i \xrightarrow{a.s.} 0$ . Thus, we rewrite the left hand side of the claim as

$$\begin{aligned} \sum_{i=1}^n v_i \psi_E(\lambda_i) &= \sum_{i=1}^n v_i \psi_H(\lambda_i) \frac{\psi_E(\lambda_i)}{\psi_H(\lambda_i)} = \sum_{i=1}^n v_i (\lambda_i^2/8)(1 + R_i) \\ &= \sum_{i=1}^n v_i \cdot \frac{2 \log(2/\alpha)}{8\hat{n}\sigma_{i-1}^2} \cdot (1 + R_i) \\ &= \sum_{i=1}^n v_i \cdot \frac{2 \log(2/\alpha)}{8n\sigma^2} \cdot (1 + R'_i) \cdot (1 + R_i) \\ &= \sum_{i=1}^n 4(X_i - \hat{\mu}_{i-1})^2 \cdot \frac{2 \log(2/\alpha)}{8n\sigma^2} \cdot (1 + R_i + R'_i + R_i R'_i). \end{aligned}$$

Defining  $R''_i = R_i + R'_i + R_i R'_i$  for brevity, and noting that  $R''_i \rightarrow 0$  almost surely, the above expression becomes

$$\begin{aligned} \sum_{i=1}^n v_i \psi_E(\lambda_i) &= \sum_{i=1}^n (X_i - \hat{\mu}_{i-1})^2 \cdot \frac{\log(2/\alpha)}{n\sigma^2} \cdot (1 + R''_i) \\ &= \frac{\log(2/\alpha)}{\sigma^2} \left[ \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_{i-1})^2 \cdot (1 + R''_i) \right] \\ &= \frac{\log(2/\alpha)}{\sigma^2} \left[ \underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_{i-1})^2}_{\xrightarrow{a.s.} \sigma^2 \text{ by Lemma 3.E.3}} + \underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_{i-1})^2 R''_i}_{\xrightarrow{a.s.} 0 \text{ by Lemma 3.E.2}} \right] \xrightarrow{a.s.} \log(2/\alpha), \end{aligned}$$

which completes the proof of Lemma 3.E.4.  $\square$

Now, consider the denominator in (3.66).

**Lemma 3.E.5.** *Continuing with the same notation,*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i \xrightarrow{a.s.} \sqrt{\frac{2 \log(2/\alpha)}{\sigma^2}}.$$

*Proof.* Let  $R'_i$  be as in (3.68). Then,

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \sqrt{\frac{2 \log(2/\alpha)}{n \hat{\sigma}_{i-1}^2}} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \sqrt{\frac{2 \log(2/\alpha)}{n \sigma^2}} \cdot (1 + R'_i) \\ &= \sqrt{\frac{2 \log(2/\alpha)}{\sigma^2}} \cdot \underbrace{\frac{1}{n} \sum_{i=1}^n (1 + R'_i)}_{\xrightarrow{\text{a.s.}} 1 \text{ by Lemma 3.E.1}} \xrightarrow{\text{a.s.}} \sqrt{\frac{2 \log(2/\alpha)}{\sigma^2}}, \end{aligned}$$

completing the proof of Lemma 3.E.5.  $\square$

We are now able to combine Lemmas 3.E.4 and 3.E.5 to prove the main result.

**Proposition 3.E.4.** *Denoting the half-width of  $C_n^{\text{PrPl-EB}(n)}$  as  $W_n$ , and assuming the data are drawn iid from a distribution  $Q \in \mathcal{Q}^\mu$  with variance  $\sigma^2$ , we have*

$$\sqrt{n} W_n \equiv \sqrt{n} \left( \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i} \right) \xrightarrow{\text{a.s.}} \sigma \sqrt{2 \log(2/\alpha)}. \quad (3.69)$$

Thus, the width is asymptotically proportional to the standard deviation.

*Proof.* By direct rearrangement of the left hand side, we see that

$$\begin{aligned} \sqrt{n} \left( \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\sum_{i=1}^n \lambda_i} \right) &= \frac{\log(2/\alpha) + \sum_{i=1}^n v_i \psi_E(\lambda_i)}{\frac{1}{\sqrt{n}} \sum_{i=1}^n \lambda_i} \\ &\xrightarrow{\text{a.s.}} \frac{\log(2/\alpha) + \log(2/\alpha)}{\sigma^{-1} \sqrt{2 \log(2/\alpha)}} = \sigma \sqrt{2 \log(2/\alpha)}, \end{aligned}$$

which completes the proof of Proposition 3.E.4.  $\square$

### 3.E.4 aGRAPA sublevel sets need not be intervals: a worst-case example

In the proof of Theorem 3.4.1, we demonstrated that the hedged capital process with predictable plug-in bets yielded convex confidence sets, making their construction more practical. However, this proof was made simple by taking advantage of the fact that the sequences before truncation  $(\dot{\lambda}_t^+)^{\infty}_{t=1}$  and  $(\dot{\lambda}_t^-)^{\infty}_{t=1}$  did not depend on  $m \in [0, 1]$ . This raises the natural question, of whether there are betting-based confidence sets which are nonconvex when these sequences depend on  $m$ . Here, we provide a (somewhat pathological) example of the aGRAPA process with nonconvex sublevel sets.

Consider the aGRAPA bets,

$$\lambda_t^{\text{aGRAPA}} := \frac{\hat{\mu}_{t-1} - m}{\hat{\sigma}_{t-1}^2 + (\hat{\mu}_{t-1} - m)^2} \text{ where } \hat{\mu}_t := \frac{1/2 + \sum_{i=1}^t X_i}{t+1}, \quad \hat{\sigma}_t^2 := \frac{1/20 + \sum_{i=1}^t (X_i - \hat{\mu}_i)^2}{t+1}. \quad (3.70)$$

Furthermore, suppose that the observed variables are  $X_1 = X_2 = 0$ . Then it can be verified that

$$\begin{aligned} \mathcal{K}_2^{\text{aGRAPA}}(m) &= (1 + \lambda_1^{\text{aGRAPA}}(X_1 - m)) (1 + \lambda_2^{\text{aGRAPA}}(X_2 - m)) \\ &= \left(1 + \frac{1/2 - m}{1/20 + (1/2 - m)^2}(-m)\right) \left(1 + \frac{1/4 - m}{0.05625 + (1/4 - m)^2}(-m)\right), \end{aligned}$$

which does not yield convex sublevel sets. For example,  $\mathcal{K}_2^{\text{aGRAPA}}(0.08) < 0.85$  and  $\mathcal{K}_2^{\text{aGRAPA}}(0.4) < 0.85$  but  $\mathcal{K}_2^{\text{aGRAPA}}(0.03) > 0.85$ . In particular, the sublevel set,

$$\{m \in [0, 1] : \mathcal{K}_2^{\text{aGRAPA}}(m) < 0.85\}$$

is not convex. In our experience, however, situations like the above do not arise frequently. In fact, we needed to actively search for these examples and use a rather small “prior” variance of  $1/20$  which we would not use in practice. Furthermore, the sublevel set given above is at the 0.85 level while confidence sets are compared against  $1/\alpha$  which is always larger than 1 and typically larger than 10. We believe that it may be possible to restrict  $(\lambda_t^{\text{aGRAPA}})_{t=1}^\infty$  and/or the confidence level,  $\alpha \in (0, 1)$  in some way so that the resulting confidence sets are convex. One reason to suspect that this may be possible is because of the intimate relationship between  $\lambda_t^{\text{aGRAPA}}$ ,  $\lambda_t^{\text{GRAPA}}$ , and the optimal hindsight bets,  $\lambda^{\text{HS}}$ . Specifically, we show in Section 3.E.6 that the optimal hindsight capital  $\mathcal{K}_t^{\text{HS}}$  is exactly the empirical likelihood ratio [200] which is known to generate convex confidence sets for the mean [116]. We leave this question as a direction for future work.

### 3.E.5 Betting confidence sequences for non-iid data

The CSs presented in this chapter are valid under the assumption that each observation is bounded in  $[0, 1]$  with conditional mean  $\mu$ . That is, we require that  $X_1, X_2, \dots$  are  $[0, 1]$ -valued with  $\mathbb{E}(X_t | \mathcal{F}_{t-1}) = \mu$  for each  $t$ , which includes familiar regimes such as independent and identically-distributed (iid) data from some common distribution  $P$  with mean  $\mu$ . Despite the generality of our results, we made matters simpler by focusing the simulations in Section 3.C on the iid setting. For the sake of completeness, we present a simulation to examine the behavior of our CSs in the presence of some non-iid data.

In this setup, we draw the first several hundred or thousand observations independently from a Beta(10, 10) – a distribution whose mean is  $1/2$  but whose variance is small ( $\approx 0.012$ ) – while the remaining observations are independently drawn from a Bernoulli( $1/2$ ) whose mean is also  $1/2$  but with a maximal variance of  $1/4$ . We chose to start the data off with low-variance observations in an attempt to “trick” our betting strategies into adapting to the wrong variance. Empirically, we find that the hedged capital (Theorem 3.4.1) and ConBo (Corollary 3.B.1) CSs start off strong, adapting to the small variance of a Beta(10, 10). After several Bernoulli( $1/2$ )

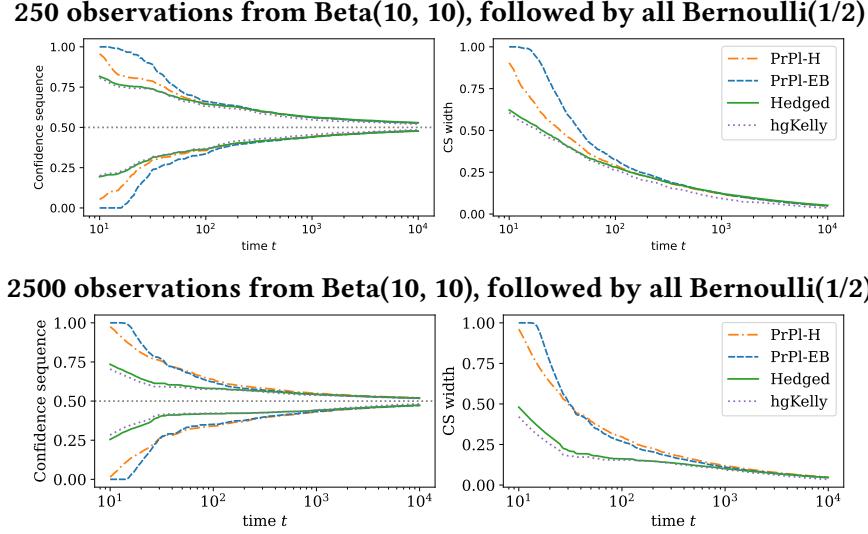


Figure 3.20: CSs for the true mean  $\mu = 1/2$  for non-iid data. In top pair of plots, the first 250 observations were independently drawn from a Beta(10, 10) while the subsequent observations are drawn from a Bernoulli(1/2). The bottom pair of plots is similar, but with 2500 initial draws from a Beta(10, 10) instead of 250. In both cases, the betting-based CSs (Hedged and ConBo) tend to outperform those based on supermartingales.

observations, the CSs remain tight, but seem to shrink less rapidly. Nevertheless, we find that the hedged capital and ConBo CSs greatly outperform the Hoeffding (Proposition 3.3.1) and empirical Bernstein (Theorem 3.3.1) predictable plug-in CSs (see Figure 3.20). Regardless of empirical performance, all methods considered produce *valid* CSs for  $\mu$ .

### 3.E.6 Owen's empirical likelihood ratio and Mykland's dual likelihood ratio

Let  $x_1, \dots, x_t \in [0, 1]$  and recall the optimal hindsight capital process  $\mathcal{K}_t^{\text{HS}}(m)$ ,

$$\mathcal{K}_t^{\text{HS}}(m) := \prod_{i=1}^t (1 + \lambda^{\text{HS}}(x_i - m)) \quad \text{where } \lambda^{\text{HS}} \text{ solves } \sum_{i=1}^t \frac{x_i - m}{1 + \lambda^{\text{HS}}(x_i - m)} = 0.$$

Now, let  $\mathcal{Q}^m \equiv \mathcal{Q}^m(x_1^t)$  be the collection of discrete probability measures with support  $\{x_1, \dots, x_t\}$  and mean  $m$ . Let  $\mathcal{Q} \equiv \mathcal{Q}(x_1^t) := \bigcup_{m \in [0, 1]} \mathcal{Q}^m$  and define the empirical likelihood ratio [200],

$$\text{EL}_t(m) := \frac{\sup_{Q \in \mathcal{Q}} \prod_{i=1}^t Q(x_i)}{\sup_{Q \in \mathcal{Q}^m} \prod_{i=1}^t Q(x_i)}.$$

Owen [200] showed that the numerator equals  $(1/t)^t$  and the denominator equals

$$\prod_{i=1}^t (1 + \lambda^{\text{EL}}(x_i - m))^{-1} \quad \text{where } \lambda^{\text{EL}} \text{ solves } \sum_{i=1}^t \frac{x_i - m}{1 + \lambda^{\text{EL}}(x_i - m)} = 0.$$

Notice that the above product is exactly the reciprocal of  $\mathcal{K}_t^{\text{HS}}$  and that  $\lambda^{\text{EL}} = \lambda^{\text{HS}}$ . Therefore for each  $m \in [0, 1]$ ,

$$\text{EL}_t(m) = (1/t)^t \mathcal{K}_t^{\text{HS}}(m).$$

Furthermore, given the connection between the empirical and dual likelihood ratios for independent data [192], the hindsight capital process is also proportional to the dual likelihood ratio in this case.

### 3.F An extended history of betting and its applications

(This is an expanded version of Section 3.6 and Figure 3.9.)

The use of betting-related ideas in probability, statistics, optimization, finance and machine learning has evolved in many different parallel threads, emanating from different influential early works and thus having different roots and evolutions. Since these threads have had little interaction for many decades now, we consider it worthwhile to mention them in some detail. Two notes of caution:

- We anticipate missing some authors and works in our broad strokes below, but a thorough coverage would be better suited to a longer survey paper on the topic. For example, we entirely skip the field of mathematical finance, since betting is literally a foundation of the entire field (and theoretical and applied progress on martingales, betting strategies, and related topics has been phenomenal).
- Many of the authors listed below have used the language of betting in their works explicitly, but others have not — and may even prefer (or have preferred) *not* to do so. Thus, our references should be treated with a pinch of salt, as some connections that we draw to betting may be more apparent in hindsight (to us) than foresight (to the authors).

If we had to pick the most critical early authors without whom our work would have been impossible, it would be Ville, Wald, Kelly and Robbins; later influences on us have been via Lai, Cover, Shafer, Vovk, Grunwald and the second author’s own earlier works [124, 125]. These authors stand out below.

**Probability.** Ville’s 1939 PhD thesis [264] contained an important and rather remarkable result of its time that connected measure-theoretic probability with betting, and indeed brought the very notion of a martingale into probability theory. In brief, Ville proved that for every event of measure zero, there exists a betting strategy for which a gambler’s wealth process (a nonnegative martingale) grows to infinity if that event occurs. For example, the strong law of large numbers (SLLN) and the law of the iterated logarithm (LIL) are two classic measure-theoretic statements that occur on all sequences of observations, except for a null set according to some underlying probability measure (where the two null sets for the two laws are different). Ville proved that it is possible to bet on the next outcome such that if the LIL were false for that particular sequence of observations, then the gambler’s wealth would grow in an unbounded fashion.

Doob’s monumental papers and book [86] in the following decades stripped martingales of their betting roots and presented them as some of the most powerful tools of measure-theoretic probability theory, with applications to many other branches of mathematics. (However, betting could be viewed as instances of “Doob’s martingale transform”.) These betting roots were revived in the 1960s with the renewed interest in algorithmic definitions of randomness, due to Kolmogorov, Martin-Löf [185] and many others.

More recently, Shafer and Vovk [235, 236] have produced two seminal books that aim bring

betting and martingales to the front and center of probability and finance, aiming to derive much (if not all) of probability theory from purely game-theoretic principles based on betting strategies. The product martingale wealth process that appears in our work also appears in theirs (indeed, it is a fundamental process), but Shafer and Vovk did not explore the topics in this chapter (confidence sequences, explicit computationally efficient betting strategies, sampling without replacement, thorough numerical simulations, and so on). Indeed, their book has a thorough treatment of probability and finance, but with respect to statistical inference, there is little explicit methodology for practice. Perhaps they were aware of such a statistical utility, but they did not explicitly recognize or demonstrate the excellent power of betting in practice (when properly developed) for problems such as ours.

**Statistical inference.** Using the power of hindsight, we now know that Wald’s influential work on the sequential probability ratio test was implicitly based on martingale techniques [271]. Wald derived many fundamental results that he required from scratch without having the general language that was being set up by Doob in parallel to his work. In the case of testing a simple null  $H_0 : \theta = \theta^*$  against a composite alternative  $H_0 : \theta \neq \theta^*$ , Wald [271, Eq (10:10)] suggests forming the likelihood ratio process  $\prod_{i=1}^n f_{\theta_{i-1}}(X_i) / \prod_{i=1}^n f_{\theta^*}(X_i)$ , where  $\theta_{i-1}$  is a mapping from  $X_1, \dots, X_{i-1}$  to  $\Theta$ ; in other words,  $\theta_{i-1}$  is predictable. In the language of this chapter, this is a predictable plug-in, and the first appearance of betting-like ideas in the statistical literature. However, beyond this passing equation in a parametric setup, the idea appears to have lain dormant.

Robbins (along with students and colleagues Siegmund, Darling, and Lai) quickly realized the power of Wald’s and Ville’s ideas as well as martingales more generally, and pursued a rather broad agenda around sequential testing and estimation, including the introduction and extensive study of confidence sequences and the method of mixtures [75, 73, 74, 218, 219, 220, 221, 222, 168]. Robbins and Siegmund also analyzed Wald’s “betting” test, and proved in some generality that its behavior is similar to a mixture likelihood ratio test [222, Section 6]. Most of Wald’s and Robbins’ work was parametric, but Robbins did explicitly study the sub-Gaussian setting in some detail [217]. Building on a vast literature of Chernoff-style concentration inequalities that exploded after Robbins’ time, Howard et al. [124, 125] recently extended mixture methods of Robbins to derive confidence sequences under a large class of nonparametric settings using exponential supermartingales. Howard et al. [124, 125] recognized Wald’s betting idea, but did not develop it nonparametrically beyond a brief mention in the paper as a direction for future work. The current work takes this natural next step in some thorough detail.

**Information and coding theory.** Soon after the seminal work of Shannon [240], another researcher at AT&T Bell Labs, John Larry Kelly Jr. wrote a paper titled “A New Interpretation of Information Rate” which explicitly connected betting with the new field of information theory, complementing the work of Shannon [152]. In short, he proved that it is possible to bet on the symbols in a communication channel at odds consistent with their probabilities in order to have a gambler’s wealth grow exponentially, with the exponent equaling the rate of transmission over the channel. More explicitly, given a sequence of Bernoulli random variables

with probability  $p > 1/2$ , Kelly proved that betting a  $(2p - 1)$  fraction of your current wealth on the next outcome being 1 is the unique strategy that maximizes the expected log wealth of the gambler.

When the probability  $p$  changes at each step in an unknown manner, the “universal coding” work of Krichevsky and Trofimov [163] showed that a mixture method involving the Jeffreys prior and maximum likelihood can achieve nearly the optimal wealth in hindsight, with the expected log wealth of their strategy only being worse than the optimal oracle log-wealth by a factor that is logarithmic in the number of rounds; these observations work for any discrete alphabet, not just a binary. Cover’s interest in these techniques spans several decades [67, 68, 69, 24, 23], culminating in his famous universal portfolio algorithm [70], that today forms a standard textbook topic in information theory.

There are other parts of information/coding theory that could be seen as related in some ways to betting through the use of (what are now called) e-variables: these include the topics of prequential model selection and minimum description length; see works by Rissanen [214, 215], Dawid [78, 79], Grünwald [111], Grünwald et al. [113], Li [173] and references therein.

**Online learning and sequential prediction under log loss.** In the 1990s, the problems studied by Krichevsky, Trofimov, and Cover continued to be extended — often dropping the information theoretic context — under the title of sequential prediction under the logarithmic loss. In the active subfield of online learning, the previous results were effectively “regret bounds” against potentially adversarial sequences of observations, with a chapter devoted to the problem in the book on prediction, learning and games by Cesa-Bianchi and Lugosi [50]. More recently, Orabona and colleagues such as Pal and Jun have found powerful implications of these ideas in deriving parameter-free algorithms for online convex optimization [196, 197, 140, 139].

Rakhlin and Sridharan [208] found that deterministic regret inequalities can be used to derive concentration inequalities for martingales, connecting the two rich fields. Later, Jun and Orabona [139] also derive concentration inequalities using their betting-based regret bounds, with explicit bounds derived in the sub-Gaussian and bounded settings. However, because regret bounds could be tight in rate but are typically loose in constants, the resulting concentration inequalities are not tight in practice. Thus, we view this line of work as important and complementary to our explorations, which are different in their motivation, derivation and practicality.

*Typically, none of these lines of literature have cited the others.* For example, the important paper of Rakhlin and Sridharan [208] does not mention the work of Ville, Wald or Robbins, or even of Vovk and Shafer. Similarly, despite the books of Shafer and Vovk having a wonderful coverage of the history of probability and martingales stemming back hundreds of years, even their recent 2019 book [236] does not cite the coding theory and online learning literature very much, including the works of Orabona and coauthors [196, 197, 140, 72, 139], Krichevsky and Trofimov [163], or Rakhlin and Sridharan [208]. Recent work of Orabona and colleagues also in turn has no mention of the books of Shafer and Vovk [235, 236], or works of Ville, Wald, Robbins, Howard, their coauthors and other recent authors. The work of Howard et al.

[124, 125] does cite the Wald and Robbins literatures, as well as the books of Shafer and Vovk and pioneering work of Ville, but does not form connections to information/coding theory nor to online learning. The excellent book of Cesa-Bianchi and Lugosi [50] does not cite Ville, the seminal martingale works of Robbins, or the 2001 book by Shafer and Vovk.<sup>4</sup>

The reason for the lack of intersection of these parallel threads is likely manifold, and definitely far from malicious: (a) these works were and continue to be published in different literatures, (b) these works had different goals in mind, meaning that they were addressing different problems and often using different techniques, (c) our understanding of these literatures and their relationships is constantly evolving and far from complete; it is likely that no author has a command over all these parallel literatures, and indeed this should not be expected.

In the preface of their 2006 book, Cesa-Bianchi and Lugosi write

Prediction of individual sequences, the main theme of this book, has been studied in various fields, such as statistical decision theory, information theory, game theory, machine learning, and mathematical finance. Early appearances of the problem go back as far as the 1950s, with the pioneering work of Blackwell, Hannan, and others. Even though the focus of investigation varied across these fields, some of the main principles have been discovered independently. Evolution of ideas remained parallel for quite some time. As each community developed its own vocabulary, communication became difficult. By the mid-1990s, however, it became clear that researchers of the different fields had a lot to teach each other. When we decided to write this book, in 2001, one of our main purposes was to investigate these connections and help ideas circulate more fluently. In retrospect, we now realize that the interplay among these many fields is far richer than we suspected. ... Today, several hundreds of pages later, we still feel there remains a lot to discover. This book just shows the first steps of some largely unexplored paths. We invite the reader to join us in finding out where these paths lead and where they connect.

Thus it is clear that Cesa-Bianchi and Lugosi already foresaw that there were many connections between the fields that have been unstated, underappreciated, undiscovered and underutilized. The connections we briefly point out above between these literatures, both historical and modern, are themselves new in their own right (not existing in any of the aforementioned books or papers) and may be considered a small contribution of this chapter. A more thorough investigation of these connections may be the topic of a future survey paper, or indeed, a book on these topics.

---

<sup>4</sup>Authors like Rissanen [214, 215] and Dawid [78, 79] are not cited in most of these works, perhaps because the connections of their works to betting are indirect.

# Chapter 4

## RiLACS: Risk-limiting audits via confidence sequences

### 4.1 Introduction

The reported outcome of an election may not match the validly cast votes for a variety of reasons, including software configuration errors, bugs, human error, and deliberate malfeasance. Trustworthy elections start with a trustworthy paper record of the validly cast votes. Given access to a trustworthy paper trail of votes, a risk-limiting audit (RLA) can provide a rigorous probabilistic guarantee:

1. If an initially announced assertion  $\mathcal{A}$  about an election is *false*, this will be corrected by the audit with high probability;
2. If the aforementioned assertion  $\mathcal{A}$  is *true*, then  $\mathcal{A}$  will be confirmed (with probability one).

Here, an electoral assertion  $\mathcal{A}$  is simply a claim about the aggregated votes cast (e.g. “Alice received more votes than Bob”). An auditor may wish to audit several claims: for example, whether the reported winner is correct or whether the margin of victory is as large as announced.

From a statistical point of view, efficient risk-limiting audits can be implemented as sequential hypothesis tests. Namely, one tests the null hypothesis  $H_0$ : “the assertion  $\mathcal{A}$  is false,” versus the alternative  $H_1$ : “the assertion  $\mathcal{A}$  is true”. Imagine then observing a random sequence of voter-cast ballots  $X_1, X_2, \dots, X_N$ , where  $N$  is the total number of ballots. A sequential hypothesis test is represented by a sequence  $(\phi_t)_{t=1}^N$  of binary-valued functions:

$$\phi_t := \phi(X_1, \dots, X_t) \mapsto \{0, 1\},$$

where  $\phi_t = 1$  represents rejecting  $H_0$  (typically in favor of  $H_1$ ), and  $\phi_t = 0$  means that  $H_0$  has not yet been rejected. The sequential test (and thus the RLA) stops as soon as  $\phi_t = 1$  or

once all  $N$  ballots are observed, whichever comes first. The “risk-limiting” property of RLAs states that if the assertion is false (in other words, if  $H_0$  holds), then

$$\mathbb{P}_{H_0} (\exists t \in \{1, \dots, N\} : \phi_t = 1) \leq \alpha,$$

which is equivalent to type-I error control of the sequential test. Another way of interpreting the above statement is as follows: if the assertion is incorrect, then with probability at least  $(1 - \alpha)$ ,  $\phi_t = 0$  for every  $t \in \{1, \dots, N\}$  and hence all  $N$  ballots will eventually be inspected, at which point the “true” outcome (which is the result of the full hand count) will be known with certainty.

#### 4.1.1 SHANGRLA Reduces Election Auditing to Sequential Testing

Designing the sequential hypothesis test  $(\phi_t)_{t=1}^N$  depends on the type of vote, the aggregation method, or the social choice function for the election, and thus past works have constructed a variety of tests. Some works have designed  $(\phi_t)_{t=1}^N$  in the context of a particular type of election [177, 199, 216]. On the other hand, the “SHANGRLA” (**S**ets of **H**alf-Average Nulls **G**enerate **R**LAs) framework unifies many common election types including plurality elections, approval voting, ranked-choice voting, and more by reducing each of these to a simple hypothesis test of whether a finite collection of finite lists of bounded numbers has mean  $\mu^*$  at most  $1/2$  [246, 37]. Let us give an illustrative example to show how SHANGRLA can be used in practice.

Suppose we have an election with two candidates, Alice and Bob. A ballot may contain a vote for Alice or for Bob, or it may contain no valid vote, e.g., because there was no selection or an overvote. It is reported that Alice and Bob received  $N_A$  and  $N_B$  votes respectively with  $N_A > N_B$  and that there were a total of  $N_I$  invalid ballots for a total of  $N = N_A + N_B + N_I$  voters. We encode votes for Alice as “1”, votes for Bob as “0” and invalid votes as “1/2”, to obtain a set of numbers  $\{x_1, x_2, \dots, x_N\}$ . Crucially, Alice indeed received more votes than Bob if and only if  $\mu^* := \frac{1}{N} \sum_{i=1}^N x_i > 1/2$ . In other words, *the report that Alice beat Bob can be translated into the assertion that  $\mu^* \in (1/2, 1]$* .

SHANGRLA proposes to audit an assertion by testing its complement: rejecting that “complementary null” is affirmative evidence that the assertion is indeed true. In other words, if one can ensure that  $X_1, X_2, \dots, X_N$  is a random permutation of  $\{x_1, \dots, x_N\}$  by sampling ballots without replacement (each ballot is chosen uniformly amongst remaining ballots), then we can concern ourselves with designing a hypothesis test  $(\phi_t)_{t=1}^N$  to test the null  $H_0 : \mu^* \leq 1/2$  against the alternative  $H_1 : \mu^* > 1/2$ .

One of the major benefits of SHANGRLA is the ability to reduce a wide range of election types to a testing problem of the above form. This permits the use of powerful statistical techniques which were designed specifically for such testing problems (but may not have been designed with RLAs in mind). Throughout this chapter, we adopt the SHANGRLA framework, and while we return to the example of plurality elections for illustrative purposes, all of our methods can be applied to any election audit which has a SHANGRLA-like testing reduction [246].

### 4.1.2 Confidence Sequences

In the fixed-time (i.e. non-sequential) hypothesis testing regime, there is a well-known duality between hypothesis tests and confidence intervals for a parameter  $\mu^*$  of interest. We describe this briefly for  $\mu^* \in [0, 1]$  for simplicity. For each  $\mu \in [0, 1]$ , suppose that  $\phi^\mu \equiv \phi^\mu(X_1, \dots, X_n) \mapsto \{0, 1\}$  is a level- $\alpha$  nonsequential, fixed-sample test for the hypothesis  $H_0 : \mu^* = \mu$  versus  $H_1 : \mu^* \neq \mu$ . Then, a nonsequential, fixed-sample  $(1 - \alpha)$  confidence interval for  $\mu^*$  is given by the set of all  $\mu \in [0, 1]$  for which  $\phi^\mu$  does not reject, that is  $\{\mu \in [0, 1] : \phi^\mu = 0\}$ .

As we discuss further in Section 4.2, an analogous duality holds for sequential hypothesis tests and time-uniform *confidence sequences* (here and throughout the chapter, “time” is used to refer to the number of samples so far, and need not correspond to any particular units such as hours or seconds). We first give a brief preview of the results to come. Consider a family of sequential hypothesis tests  $\{(\phi_t^\mu)_{t=1}^N\}_{\mu \in [0,1]}$ , meaning that for each  $\mu$ ,  $(\phi_t^\mu)_{t=1}^N$  is a sequential test for  $\mu$ . Then, the set of all  $\mu$  for which  $\phi_t^\mu = 0$ ,

$$C_t := \{\mu \in [0, 1] : \phi_t^\mu = 0\}$$

forms a  $(1 - \alpha)$  *confidence sequence* for  $\mu^*$ , meaning that

$$\mathbb{P}(\exists t \in [N] : \mu^* \notin C_t) \leq \alpha,$$

where  $[N]$  is used to denote the set  $\{1, 2, \dots, N\}$ . In other words,  $C_t$  will cover  $\mu^*$  at every single time  $t$ , except with some small probability  $\leq \alpha$ . Since  $C_t$  is typically an interval  $[L_t, U_t]$ , we call the lower endpoint  $(L_t)_{t=1}^N$  as a lower confidence sequence (and similarly for upper).

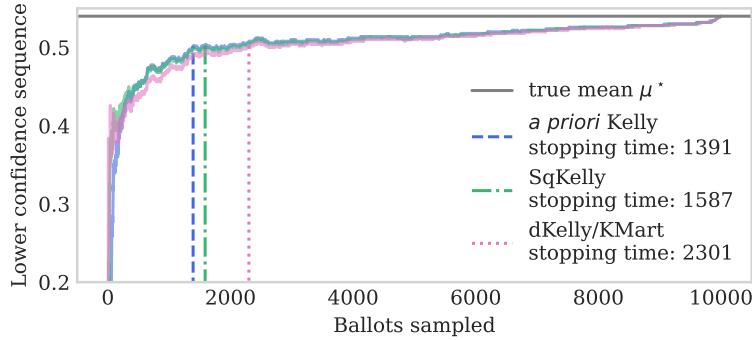


Figure 4.1: 95% Lower confidence sequences for the margin of a plurality election between Alice and Bob for three different auditing methods. Votes for Alice are encoded by “1” and those for Bob are encoded by “0”. The parameter of interest is then the average of these votes, which in this particular example is 54% (given by the horizontal grey line). The outcome is verified once the lower confidence sequence exceeds 1/2. The time at which this happens is given by the vertical blue, green, and pink lines.

In particular, given the sequential hypothesis testing problem that arises in SHANGRLA,

we can cast the RLA as a sequential estimation problem that can be solved by developing confidence sequences (see Figure 4.1).<sup>1</sup> As we will see in Section 4.2, our confidence sequences provide added flexibility and an intuitive visualizable interpretation for SHANGRLA-compatible election audits, without sacrificing any statistical efficiency.

### 4.1.3 Contributions and Outline

The contributions of this work are twofold. First, we introduce confidence sequences to the election auditing literature as intuitive and flexible ways of interpreting and visualizing risk-limiting audits. Second, we present algorithms for performing RLAs based on confidence sequences by deriving statistically and computationally efficient nonnegative martingales. At the risk of oversimplifying the issue, modern RLAs face a computational-statistical efficiency tradeoff. Methods such as BRAVO are easy to compute, but potentially less statistically efficient than the current state-of-the-art, KMart [246], but KMart can be prohibitively expensive to compute for large elections. The methods presented in this chapter resolve this tradeoff: they typically match or outperform both BRAVO and KMart, while remaining practical to compute in large elections.

In Section 4.2, we show how confidence sequences generate risk-limiting audits, how they relate to more familiar RLAs based on sequentially valid  $p$ -values, and how they can be used to audit multiple contests. Section 4.3 derives novel confidence sequence-based RLAs and compares them to past RLA methods via simulation. In Section 4.4, we illustrate how the previously derived techniques can be applied to an audit of Canada’s 43rd federal election. Finally, Section 4.5 discusses how all of the aforementioned results apply to risk-limiting tallies for coercion-resistant voting schemes.

## 4.2 Confidence sequences are risk-limiting

Consider an election consisting of  $N$  ballots. Following SHANGRLA [246], suppose that these can be transformed to a set of  $[0, u]$ -bounded real numbers  $x_1, \dots, x_N \in [0, u]$  with mean  $\mu^* := \frac{1}{N} \sum_{i=1}^N x_i$  for some known  $u > 0$ . Suppose that electoral assertions can be made purely in terms of  $\mu^*$ . A classical  $(1 - \alpha)$  confidence interval  $\text{CI}_n$  for  $\mu^*$  is an interval computed from data  $X_1, X_2, \dots, X_n$  with the guarantee that

$$\forall n \in [N], \mathbb{P}(\mu^* \in \text{CI}_n) \geq 1 - \alpha.$$

In contrast, a  $(1 - \alpha)$  confidence sequence for  $\mu^*$  is a sequence of confidence sets,  $C_1, C_2, \dots, C_N$  which all simultaneously capture  $\mu^*$  with probability at least  $(1 - \alpha)$ . That is,

$$\underbrace{\mathbb{P}(\forall t \in [N], \mu^* \in C_t) \geq 1 - \alpha}_{\text{simultaneous coverage probability}}, \quad \text{or equivalently} \quad \underbrace{\mathbb{P}(\exists t \in [N] : \mu^* \notin C_t) \leq \alpha}_{\text{error probability}}.$$

---

<sup>1</sup>Code to reproduce all plots can be found at [github.com/wannabesmith/RiLACS](https://github.com/wannabesmith/RiLACS).

The two probabilistic statements above are equivalent, but provide a different way of interpreting  $\alpha$  and the corresponding guarantee.

If we have access to a  $(1 - \alpha)$  confidence sequence for  $\mu^*$ , we can audit any assertion about the election outcome made in terms of  $\mu^*$  with risk limit  $\alpha$ . Here, we use  $\mathcal{A} \subseteq [0, u]$  to denote an assertion. For example, SHANGRLA typically uses assertions of the form “ $\mu^*$  is greater than  $1/2$ ”, in which case  $\mathcal{A} = (1/2, u]$ .

Algorithm 4.1: Risk limiting audits via confidence sequences (RiLACS)

**Require:** Assertion  $\mathcal{A} \subseteq [0, u]$ , risk limit  $\alpha \in (0, 1)$

```

1: for  $t \in [N]$  do
2:   Randomly sample and remove  $X_t$  from the remaining ballots.
3:   Compute  $C_t \equiv C(X_1, \dots, X_t)$  at level  $\alpha$ .
4:   if  $\mathcal{A} \subseteq C_t$  then
5:     Certify the assertion  $\mathcal{A}$  and stop if desired.
6:   end if
7: end for

```

If the goal is to finish the audit as soon as possible above all else, then one can ignore the “if desired” condition. However, continued sampling can provide added assurance in  $\mathcal{A}$ , and maintains the risk limit at  $\alpha$ . The following theorem summarizes the risk-limiting guarantee of the above algorithm.

**Theorem 4.2.1.** *Let  $(C_t)_{t=1}^N$  be a  $(1 - \alpha)$  confidence sequence for  $\mu^*$ . Let  $\mathcal{A} \subseteq [0, u]$  be an assertion about the electoral outcome (in terms of  $\mu^*$ ). The audit mechanism that certifies  $\mathcal{A}$  as soon as  $C_t \subseteq \mathcal{A}$  has risk limit  $\alpha$ .*

*Proof.* We need to prove that if  $\mu^* \notin \mathcal{A}$ , then  $\mathbb{P}(\exists t \in [N] : C_t \subseteq \mathcal{A}) \leq \alpha$ . First, notice that if  $C_t \subseteq \mathcal{A}$ , then we must have that  $\mu^* \notin C_t$  since  $\mu^* \notin \mathcal{A}$ . Then,

$$\begin{aligned} \mathbb{P}(\exists t \in [N] : C_t \subseteq \mathcal{A}) &\leq \mathbb{P}(\exists t \in [N] : \mu^* \notin C_t) \\ &\leq \alpha, \end{aligned}$$

where the second inequality follows from the definition of a confidence sequence. This completes the proof.  $\square$

Let us see how this theorem can be used in an example. Consider an election with two candidates, Alice and Bob, and a total of  $N$  cast ballots. Let  $\{x_1, \dots, x_N\}$  be the list of numbers that result from encoding votes for Alice as 1, votes for Bob as 0, and ballots that do not contain a valid vote as  $1/2$ . Let  $(C_t)_{t=1}^N$  be a  $(1 - \alpha)$  confidence sequence for  $\mu^* := \frac{1}{N} \sum_{i=1}^N x_i$ . If we wish to audit the assertion that “Alice beat Bob”, then  $u = 1$  and  $\mathcal{A} = (1/2, 1]$ . We can sequentially sample  $X_1, X_2, \dots, X_N$  without replacement, certifying the assertion once  $C_t \subseteq \mathcal{A}$ . By Theorem 4.2.1, this limits the risk to level  $\alpha$ .

### 4.2.1 Relationship to Sequential Hypothesis Testing

The earliest work on RLAs did not use anytime  $p$ -values [247, 244], but since about 2009, most RLA methods have used anytime  $p$ -values to conduct sequential hypothesis tests [245, 198, 199, 246, 129]. An anytime  $p$ -value is a sequence of  $p$ -values  $(p_t)_{t=1}^N$  with the property that under some null hypothesis  $H_0$ ,

$$\mathbb{P}_{H_0}(\exists t \in [N] : p_t \leq \alpha) \leq \alpha. \quad (4.1)$$

The anytime  $p$ -values  $p_t \equiv p_t(\mu)$  are typically defined implicitly for each null hypothesis  $H_0 : \mu^* = \mu$  and yield a sequential hypothesis test  $\phi_t^\mu := \mathbf{1}(p_t(\mu) \leq \alpha)$ . As alluded to in Section 4.1.2, this immediately recovers a confidence sequence:

$$C_t := \{\mu \in [0, u] : \phi_t^\mu = 0\}.$$

Notice in Figure 4.2 that the times at which nulls are rejected (or “stopping times”) are the same for both confidence sequences and the associated  $p$ -values. Thus, nothing is lost by basing the RLA on confidence sequences rather than anytime  $p$ -values. Confidence sequences benefit from being visually intuitive and are arguably easier to interpret than anytime  $p$ -values.

For example, consider conducting an RLA for a simple two-candidate election between Alice and Bob with no invalid votes. Suppose that it is reported that Alice won, i.e.,  $\mu^* := \frac{1}{N} \sum_{i=1}^N x_i > 1/2$  where  $x_i = 1$  if the  $i$ th ballot is for Alice, 0 if for Bob, and  $1/2$  if the ballot does not contain a valid vote for either candidate. A sequential RLA in the SHANGLA framework would posit a null hypothesis  $H_0 : \mu^* \leq 1/2$  (the complement of the announced result: Bob actually won or the outcome is a tie), sample random ballots sequentially, and stop the audit (confirming the announced result) if and when  $H_0$  is rejected at significance level  $\alpha$ . If  $H_0$  is not rejected before all ballots have been inspected, the true outcome is known.<sup>2</sup>

On the other hand, a ballot-polling RLA [177] based on confidence sequences proceeds by computing a lower  $1 - \alpha$  confidence bound for the fraction  $\mu^*$  of votes for Alice. The audit stops, confirming the outcome, if and when this lower bound is larger than  $1/2$ . If that does not occur before the last ballot has been examined, the true outcome is known. In this formulation, there is no need to define a null hypothesis as the complement of the announced result and interpret the resulting  $p$ -value, and so on. The approach also works for comparison audits using the “overstatement assorter” approach developed in [246], which transforms the problem into the same canonical form: testing whether the mean of any list in a collection of nonnegative, bounded lists is less than  $1/2$ .

### 4.2.2 Auditing Multiple Contests

It is known that RLAs of multi-candidate, multi-winner elections can be reduced to several pairwise contests without adjusting for multiplicity [177]. This is accomplished by testing whether every single reported winner beat every single reported loser, and stopping once

---

<sup>2</sup>At any point during the sampling, an election official can choose to abort the sampling and perform a full hand count for any reason. This cannot increase the risk limit: the chance of failing to correct an incorrect reported outcome does not increase.

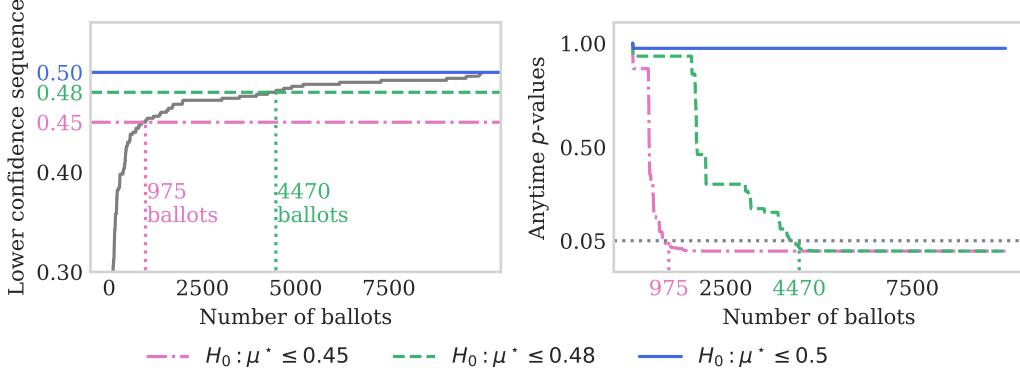


Figure 4.2: The duality between anytime  $p$ -values and confidence sequences for three nulls:  $H_0 : \mu^* \leq \mu_0$  for  $\mu_0 \in \{0.45, 0.48, 0.5\}$ . The  $p$ -value for  $H_0 : \mu^* \leq 0.45$  (pink dash-dotted line) drops below 5% after 975 samples, exactly when the 95% lower confidence sequence exceeds 0.45. However, the  $p$ -value for  $H_0 : \mu^* \leq 0.5$  never reaches 0.05 and the 95% confidence sequence never excludes 0.5, the true value of  $\mu^*$ .

each of these tests rejects their respective nulls at level  $\alpha \in (0, 1)$ . For example, suppose it is reported that a set of candidates  $\mathcal{W}$  beat a set of candidates  $\mathcal{L}$  in a  $k$ -winner plurality contest with  $K$  candidates in all (that is,  $|\mathcal{W}| = k$  and  $|\mathcal{L}| = K - k$ ). For each reported winner  $w \in \mathcal{W}$  and each reported loser  $\ell \in \mathcal{L}$ , encode votes for candidate  $w$  as “1”, votes for  $\ell$  as “0” and ballots with no valid vote in the contest or with a vote for any other candidate as “1/2” to obtain the population  $\{x_1^{w,\ell}, \dots, x_N^{w,\ell}\}$ . Then as before, candidate  $w$  beat candidate  $\ell$  if and only if  $\mu_{w,\ell}^* := \frac{1}{N} \sum_{i=1}^N x_i^{w,\ell} > 1/2$ . In a two-candidate plurality election we would have proceeded by testing the null  $H_0^{w,\ell} : \mu_{w,\ell}^* \leq 1/2$  against the alternative  $H_1^{w,\ell} : \mu_{w,\ell}^* > 1/2$ . To use the decomposition of a single winner or multi-winner plurality contest into a set of pairwise contests, we test each null  $H_0^{w,\ell} : \mu_{w,\ell}^* \leq 1/2$  for  $w \in \mathcal{W}$  and  $\ell \in \mathcal{L}$ . The audit stops if and when *all*  $k(K - k)$  null hypotheses are rejected. Crucially, if candidate  $w \in \mathcal{W}$  did not win (i.e.  $\mu_{w,\ell}^* \leq 1/2$  for some  $\ell \in \mathcal{L}$ ), then

$$\mathbb{P}(\text{reject all } H_{0,w,\ell} : w \in \mathcal{W}, \ell \in \mathcal{L}) \leq \min_{w \in \mathcal{W}, \ell \in \mathcal{L}} \mathbb{P}(\text{reject } H_{0,w,\ell}) \leq \alpha.$$

The same technique applies when auditing with confidence sequences. Let  $\{(C_t^{w,\ell})_{t=1}^N\}$  be  $(1 - \alpha)$  confidence sequences for  $\{\mu_{w,\ell}^*\}$ ,  $w \in \mathcal{W}$ ,  $\ell \in \mathcal{L}$ . We verify the electoral outcome of every contest once  $C_t^{w,\ell} \subseteq (1/2, u]$  for all  $w \in \mathcal{W}$ ,  $\ell \in \mathcal{L}$ . Again, if  $\mu_{w,\ell}^* \leq 1/2$  for some  $w \in \mathcal{W}$ , and  $\ell \in \mathcal{L}$ , then

$$\mathbb{P}(\forall w \in \mathcal{W}, \forall \ell \in \mathcal{L}, C_t^{w,\ell} \subseteq (1/2, u]) \leq \min_{w \in \mathcal{W}, \ell \in \mathcal{L}} \mathbb{P}(C_t^{w,\ell} \subseteq (1/2, u]) \leq \alpha.$$

This technique can be generalized to handle audits of any number of contests from the same audit sample, as explained in [246]. For the sake of brevity, we omit the derivation, but it is a

straightforward extension of the above.

### 4.3 Designing powerful confidence sequences for RLAs

So far we have discussed how to conduct RLAs from confidence sequences for the parameter  $\mu^*$ . In this section, we will discuss how to derive powerful confidence sequences for the purposes of conducting RLAs as efficiently as possible. For mathematical and notational convenience in the following derivations, we consider the case where  $u = 1$ . Note that nothing is lost in this setup since any population of  $[0, u]$ -bounded numbers can be scaled to the unit interval  $[0, 1]$  by dividing each element by  $u$  (thereby scaling the population's mean as well).

As discussed in Section 4.2.1, we can construct confidence sequences by “inverting” sequential hypothesis tests. In particular, given a sequential hypothesis test  $(\phi_t^\mu)_{t=1}^N$ , the sequence of sets,

$$C_t := \{\mu \in [0, 1] : \phi_t^\mu = 0\}$$

forms a  $(1 - \alpha)$  confidence sequence for  $\mu^*$ . Consequently, in order to develop powerful RLAs via confidence sequences, we can simply focus on carefully designing sequential tests  $(\phi_t^\mu)_{t=1}^N$ .<sup>3</sup>

To design sequential hypothesis tests, we start by finding *martingales* that translate to powerful tests. To this end, define  $M_0(\mu) := 1$  and consider the following process for  $t \in [N]$ :

$$M_t(\mu) := \prod_{i=1}^t (1 + \lambda_i(X_i - C_i(\mu))), \quad (4.2)$$

where  $\lambda_i \in \left[0, \frac{1}{C_i(\mu)}\right]$  is a tuning parameter depending only on  $X_1, \dots, X_{i-1}$ , and

$$C_i(\mu) := \frac{N\mu - \sum_{j=1}^{i-1} X_j}{N - i + 1}$$

is the conditional mean of  $X_i \mid X_1, \dots, X_{i-1}$  if the mean of  $\{x_1, \dots, x_N\}$  were  $\mu$ .

Following Chapter 3, the process  $(M_t(\mu^*))_{t=0}^N$  is a nonnegative martingale starting at one. Formally, this means that  $M_0(\mu^*) = 1$ ,  $M_t(\mu^*) \geq 0$ , and

$$\mathbb{E}(M_t(\mu^*) \mid X_1, \dots, X_{t-1}) = M_{t-1}(\mu^*)$$

for each  $t \in [N]$ . Importantly for our purposes, nonnegative martingales are unlikely to ever become very large. This fact is known as *Ville's inequality* [264, 124], which serves as a generalization of Markov's inequality to nonnegative (super)martingales, and can be stated formally as

$$\mathbb{P}(\exists t \in [N] : M_t(\mu^*) \geq 1/\alpha) \leq \alpha M_0(\mu^*) = \alpha, \quad (4.3)$$

---

<sup>3</sup>Notice that it is not always feasible to compute the set of all  $\mu \in [0, 1]$  such that  $\phi_t^\mu = 0$  since  $[0, 1]$  is uncountably infinite. However, all confidence sequences we will derive in this section are intervals (i.e. convex), and thus we can find the endpoints using a simple grid search or standard root-finding algorithms.

where  $\alpha \in (0, 1)$ , and the equality follows from the fact that  $M_0(\mu^*) = 1$ . As alluded to in Section 4.2,  $(M_t(\mu^*))_{t=0}^N$  can be interpreted as the reciprocal of an anytime  $p$ -value:

$$\mathbb{P}\left(\exists t \in [N] : \frac{1}{M_t(\mu^*)} \leq \alpha\right) \leq \alpha,$$

which matches the probabilistic guarantee in (4.1). As a direct consequence of Ville's inequality, if we define the test  $\phi_t^\mu := \mathbb{1}(M_t(\mu) \geq 1/\alpha)$ , then

$$\mathbb{P}(\exists t \in [N] : \phi_t^\mu = 1) \leq \alpha,$$

and thus  $(\phi_t^\mu)_{t=1}^N$  is a level- $\alpha$  sequential hypothesis test. We can then invert  $(\phi_t^\mu)_{t=1}^N$  and apply Theorem 4.2.1 to obtain confidence sequence-based RLAs with risk limit  $\alpha$ .

### 4.3.1 Designing Martingales and Tests from Reported Vote Totals

So far, we have found a process  $(M_t(\mu))_{t=0}^N$  that is a nonnegative martingale when  $\mu = \mu^*$ , but what happens when  $\mu \neq \mu^*$ ? This is where the tuning parameters  $(\lambda_t)_{t=1}^N$  come into the picture. Recall that an electoral assertion  $\mathcal{A}$  is certified once  $C_t \subseteq \mathcal{A}$ . Therefore, to audit assertions quickly, we want  $C_t$  to be as tight as possible. Since  $C_t$  is defined as the set of  $\mu \in [0, 1]$  such that  $M_t(\mu) < 1/\alpha$ , we can make  $C_t$  tight by making  $M_t(\mu)$  as *large* as possible. To do so, we must carefully choose  $(\lambda_t)_{t=1}^N$ . This choice will depend on the type of election as well as the amount of information provided prior to the audit. First consider the case where reported vote totals are given (in addition to the announced winner).

For example, recall the election between Alice and Bob of Section 4.2, and suppose that  $\{x_1, \dots, x_N\}$  is the list of numbers encoding votes for Alice as 1, votes for Bob as 0, and ballots with no valid vote for either candidate as 1/2. Recall that Alice beat Bob if and only if  $\mu^* := \frac{1}{N} \sum_{i=1}^N x_i > 1/2$ , so we are interested in testing the null hypothesis  $H_0 : \mu^* \leq 1/2$  against the alternative  $H_1 : \mu^* > 1/2$ . Suppose it is reported that Alice beat Bob with  $N'_A$  votes for Alice,  $N'_B$  for Bob, and  $N'_U$  nuisance votes (i.e. either invalid or for another party). If the reported outcome is *correct*, then for any fixed  $\lambda$ , we know the exact value of

$$\prod_{i=1}^N (1 + \lambda(x_i - 1/2)), \tag{4.4}$$

which is an inexact but reasonable proxy for  $M_N(1/2)$ , the final value of the process  $(M_t(1/2))_{t=0}^N$ . We can then choose the value of  $\lambda'$  that maximizes (4.4). Some algebra (which we defer to Section 4.A) reveals that the maximizer of (4.4) is given by

$$\lambda' := 2 \frac{N'_A - N'_B}{N'_A + N'_B}. \tag{4.5}$$

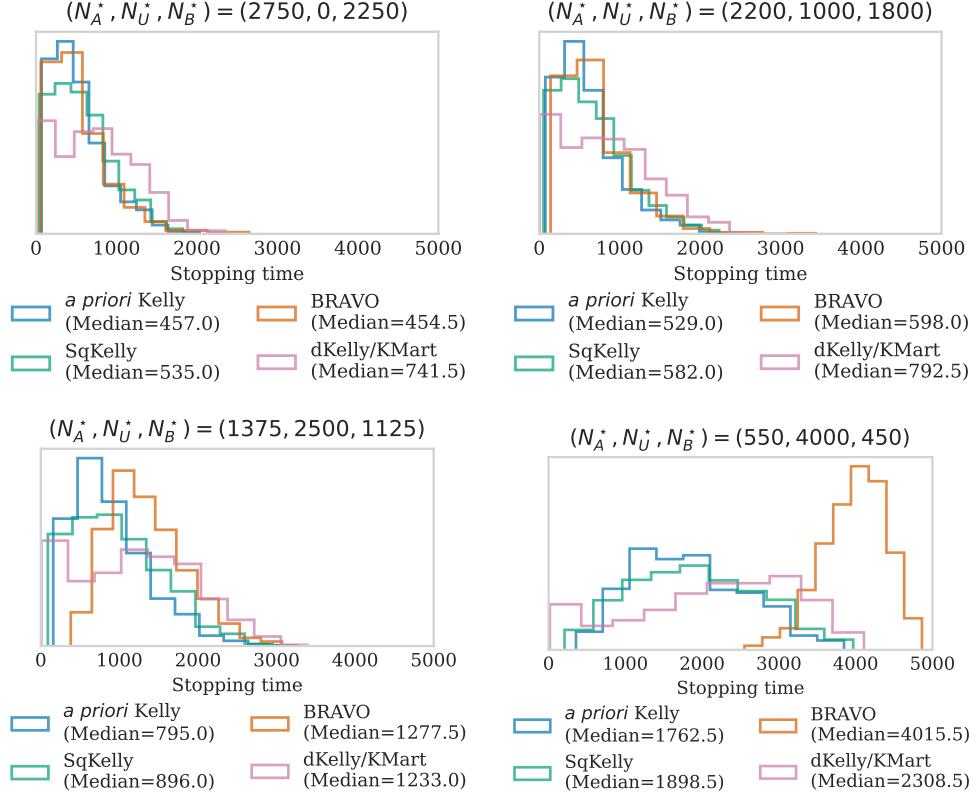


Figure 4.3: Ballot-polling audit workload distributions under four possible outcomes of a two-candidate plurality election. Workload is defined as the number of distinct ballots examined before completing the audit. The first example considers an outcome where Alice and Bob received 2750 and 2250 votes respectively, and no ballots were invalid, for a margin of 0.1. The second, third, and fourth examples have the same margin, but with increasing numbers of invalid or “nuisance” ballots represented by  $N_U^*$ . Notice that in the case with no nuisance ballots, *a priori* Kelly and BRAVO have an edge, while in the setting with many nuisance ballots, *a priori* Kelly vastly outperforms BRAVO. On the other hand, neither SqKelly nor dKelly require tuning based on the reported outcomes, but SqKelly outperforms dKelly in all four scenarios.

We then truncate  $\lambda'$  at each time step  $t$  to obtain

$$\lambda_t^{\text{apK}} := \min \left\{ \lambda', \frac{1}{C_t(\mu)} \right\}, \quad (4.6)$$

ensuring that it lies in the allowable range  $[0, 1/C_t(\mu)]$ . We call this choice of  $\lambda_t^{\text{apK}}$  ***a priori* Kelly** due to its connections to Kelly’s criterion [152] (see Chapter 3) for maximizing products of the form (4.4). This choice of  $\lambda_t^{\text{apK}}$  also has the desirable property of yielding convex confidence sequences, which we summarize below.

**Proposition 4.3.1.** Let  $X_1, \dots, X_N$  be a sequential random sample from  $\{x_1, \dots, x_N\}$  with  $\mu^* := \frac{1}{N} \sum_{i=1}^N x_i$ . Consider  $(\lambda_t^{\text{apK}})_{t=1}^N$  from (4.6) and define the process  $M_t(\mu) := \prod_{i=1}^t (1 + \lambda_i^{\text{apK}}(X_i - \mathcal{C}_i(\mu)))$  for any  $\mu \in [0, 1]$ . Then the confidence set

$$C_t^{\text{apK}} := \{\mu \in [0, 1] : M_t(\mu) < 1/\alpha\}$$

is an interval with probability one.

*Proof.* Notice that since  $\lambda' \geq 0$ ,  $\mathcal{C}_t(\mu) \geq 0$ , and  $X_i \geq 0$ , we have that

$$\lambda_t^{\text{apK}}(X_i - \mathcal{C}_t(\mu)) = \min\{\lambda' X_i, X_i/\mathcal{C}_t(\mu)\} - \min\{\lambda' \mathcal{C}_t(\mu), 1\}$$

is a nonincreasing function of  $\mu$  for each  $t \in [N]$ . Consequently,  $M_t(\mu)$  is a nonincreasing and quasiconvex function of  $\mu$ , so its sublevel sets are convex.  $\square$

Note that any sequence  $(\lambda_t)_{t=1}^N$  such that  $\lambda_t \in [0, 1/\mathcal{C}_t(\mu)]$  would have yielded a valid non-negative martingale, but we chose that which maximizes (4.4) so that the resulting hypothesis test  $\phi_t := \mathbb{1}(M_t(1/2) > 1/\alpha)$  is powerful. In situations more complex than two-candidate plurality contests, the maximizer of (4.4) can still be found efficiently via standard root-finding algorithms. All of these methods are implemented in our Python package.<sup>4</sup>

While audits based on *a priori* Kelly display excellent empirical performance (see Figure 4.3), their efficiency may be hurt when vote totals are erroneously reported. Small errors in reported vote totals seem to have minor adverse effects on stopping times (and in some cases can be slightly beneficial), but larger errors can significantly affect stopping time distributions (see Figure 4.4). If we wish to audit the reported winner of an election but prefer not to rely on (or do not have access to) exact reported vote totals, we need an alternative to *a priori* Kelly. In the following section, we describe a family of such alternatives.

### 4.3.2 Designing Martingales and Tests without Vote Totals

If the exact vote totals are not known, but we still wish to audit an assertion (e.g. that Alice beat Bob), we need to design a slightly different martingale that does not depend on maximizing (4.4) directly. Instead of finding an optimal  $\lambda'$ , we will take  $D \geq 2$  points evenly-spaced on the allowable range  $[0, 1/\mathcal{C}_t(\mu)]$  and “hedge our bets” among all of these. Making this more precise, note that a convex combination of martingales (with respect to the same filtration) is itself a martingale, and thus for any  $(\theta_1, \dots, \theta_D)$  such that  $\theta_d \geq 0$  and  $\sum_{d=1}^D \theta_d = 1$ , we have that

$$M_t^D(\mu^*) := \sum_{d=1}^D \theta_d \prod_{i=1}^t \left( 1 + \frac{d}{(D+1)\mathcal{C}_i(\mu^*)} (X_i - \mathcal{C}_i(\mu^*)) \right) \quad (4.7)$$

forms a nonnegative martingale starting at one. Notice that we no longer have to depend on the reported vote totals to begin an audit. Furthermore, confidence sequences generated using sublevel sets of  $M_t^D(\mu)$  are intervals with probability one (Theorem 3.5.1). Nevertheless,

---

<sup>4</sup>[github.com/wannabesmith/RiLACS](https://github.com/wannabesmith/RiLACS)

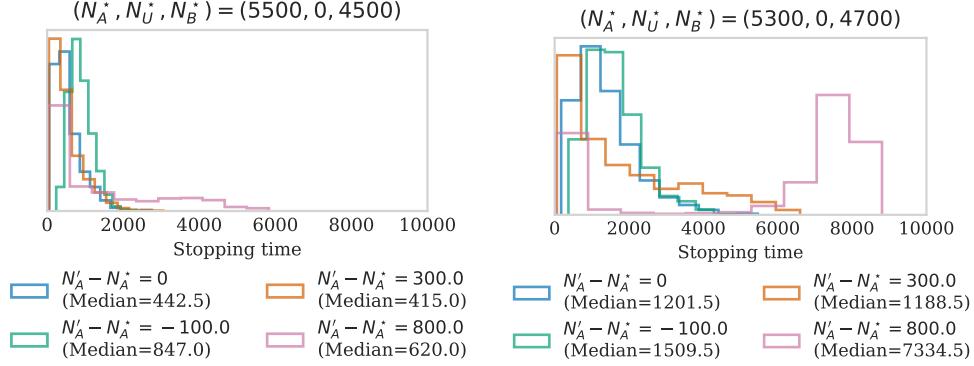


Figure 4.4: Stopping times for *a priori* Kelly under various degrees of error in reported outcomes. In the above legends,  $N_A^*$  refers to the *true* number of votes for Alice, while  $N'_A$  refers to the incorrectly reported number of votes. Notice that empirical performance is relatively strong for  $N'_A - N_A^* \in \{0, 300\}$  but is adversely affected when  $N'_A - N_A^* \in \{-100, 800\}$ , especially in the right-hand side plot with a narrower margin.

choosing  $(\theta_1, \dots, \theta_D)$  is a nontrivial task. A natural – but as we will see, suboptimal – choice is to set  $\theta_d = 1/D$  for each  $d \in [D]$ . In Chapter 3 we call this **dKelly** (for “diversified Kelly”), a name we further adopt here. In fact, this choice of  $(\theta_1, \dots, \theta_D)$  gives an arbitrarily close and computationally efficient approximation to the *Kaplan martingale* (**KMart**) [246] which can otherwise be prohibitively expensive to compute for large  $N$ .

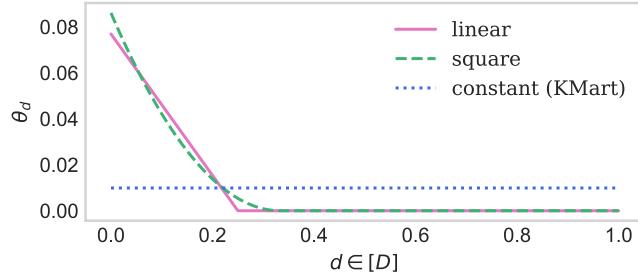


Figure 4.5: Various values of the convex weights  $(\theta_1, \dots, \theta_D)$ , which can be used in the construction of the diversified martingale (4.7). Notice that the linear and square weights are largest for  $d$  near 0, and decrease as  $d$  approaches  $1/4$ , finally remaining at 0 for all large  $d$ . Smaller values of  $d$  are upweighted since they correspond to those values of  $\lambda$  in  $M_t^D(\mu^*)$  that are optimal for smaller (i.e. interesting) electoral margins. This is in contrast to the constant weight function, which sets  $\theta_d = 1/D$  for each  $d \in [D]$ . We find that square weights perform well in practice (see Figure 4.3) but these can be tuned and tailored based on prior knowledge and the particular problem at hand.

Better choices of  $(\theta_d)_{d=1}^D$  exist for the types of elections one might encounter in practice. Recall that near-optimal values of  $\lambda$  are given by (4.5). However, setting  $\theta_d = 1/D$  for each

$d \in [D]$  implicitly treats all  $d/((D+1)\mathcal{C}_i(\mu^*))$  as equally reasonable values of  $\lambda$ . Elections with large values of  $\mu^*$  (e.g. closer to 1) are “easier” to audit, and the interesting or “difficult” regime is when  $\mu^*$  is close to (but strictly larger than) 1/2. Therefore, we recommend designing  $(\theta_1, \dots, \theta_D)$  so that  $(M_t^D(1/2))_{t=0}^N$  upweights optimal values of  $\lambda$  for margins close to 0, and downweights those for margins close to 1. Consider the following concrete examples. First, we have the truncated-square weights,

$$\theta_d^{\text{square}} := \frac{\gamma_d^{\text{square}}}{\sum_{d=1}^D \gamma_d^{\text{square}}}, \quad \text{where } \gamma_d^{\text{square}} := (1/3 - x)^2 \mathbb{1}_{d \leq 1/3}.$$

and we normalize by  $\sum_d \gamma_d^{\text{linear}}$  to ensure that  $\sum_d \theta_d = 1$ . Another sensible choice is given by the truncated-linear weights, where we simply replace  $\gamma_d^{\text{square}}$  by  $\gamma_d^{\text{linear}} := \max\{0, 1 - 2d\}$ . These values of  $\theta_d^{\text{linear}}$  and  $\theta_d^{\text{square}}$  are large for  $d \approx 0$  and small for  $d \gg 0$ , and hence the summands in the martingale given by (4.7) are upweighted for implicit values of  $\lambda$  which are optimal for “interesting” margins close to 0, and downweighted for simple margins much larger than 0 (see Figure 4.5).

When  $M_t^D$  is combined with  $\theta_d^{\text{square}}$ , we refer to the resulting martingales and confidence sequences as **SqKelly**. We compare their empirical workload against that of *a priori* Kelly, dKelly, and BRAVO in Figure 4.3. A hybrid approach is also possible: suppose we want to use reported outcomes or prior knowledge alongside these convex-weighted martingales. We can simply choose  $(\theta_1, \dots, \theta_D)$  so that  $M_t^D$  upweights values in a neighborhood of  $\lambda'$  (or some other value chosen based on prior knowledge<sup>5</sup>).

## 4.4 Illustration: auditing Canada’s 43rd federal election

We now apply the techniques derived in Section 4.3 to risk-limiting audits of the 2019 Canadian federal election, which is made up of many plurality contests between 6 major political parties.<sup>6</sup> These consisted of The Liberal Party of Canada, The Progressive Conservative Party of Canada (PC), The New Democratic Party (NDP), The Green Party, The Bloc Québécois (Bloc), and the People’s Party of Canada (PPC). Independent candidates were also included where appropriate. The country is made up of 338 so-called “ridings” (see Figure 4.6). These are geographic regions, each corresponding to one seat in the house of commons. For each riding, a multi-party, single-winner plurality contest takes place where the winner is awarded the respective seat. Generally speaking, the party with the greatest number of seats forms government (there are exceptions to this rule<sup>7</sup> but these will not be important for the purposes of auditing). In US elections, states and electoral college votes play similar roles to ridings and seats, respectively. Since each riding’s underlying contest takes the form of a multi-party, single-winner plurality election,

---

<sup>5</sup>The use of the word “prior” here should not be interpreted in a Bayesian sense. No matter what values of  $(\theta_1, \dots, \theta_D)$  are chosen, the resulting tests and confidence sequences have *frequentist* risk-limiting guarantees.

<sup>6</sup>While Canada has many registered political parties, only a handful have come close to winning seats in the house of commons, and hence should be considered in an audit. As a somewhat arbitrary rule, we considered those parties which satisfied the Leaders’ Debates Commission’s 2019 participation criteria.

<sup>7</sup>[www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=en](http://www.elections.ca/content.aspx?section=res&dir=ces&document=part1&lang=en)

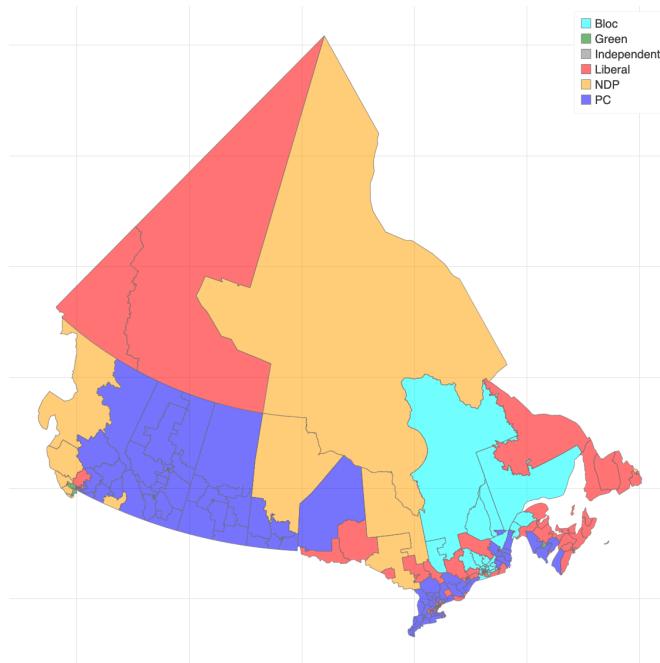


Figure 4.6: A map of Canada's 338 ridings, each representing one seat in the house of commons. Ridings are colored according to which party received the greatest number of votes in the 2019 federal election. The PPC is omitted from the legend here as they did not win any seats.

we can simply apply the techniques for auditing multiple contests outlined in Section 4.2.2 alongside the martingales and confidence sequences developed in Section 4.3.

**The data-driven web application** We designed and developed an interactive Python- and Bokeh-based [38] web application where users can display audits of any Canadian riding in a single click. This combined two data sources: one for electoral outcomes as recorded by hand-counted paper ballots in the 2019 federal election [40, 46], and one to draw the map of electoral districts [77]. After cleaning and merging, the data consisted of 347 records. Each record consists of a geographic information systems (GIS) polygon to draw the riding, vote totals for each party, and other information. The additional 9 records correspond to islands which are not separate ridings but require their own GIS polygon to be drawn on a map.

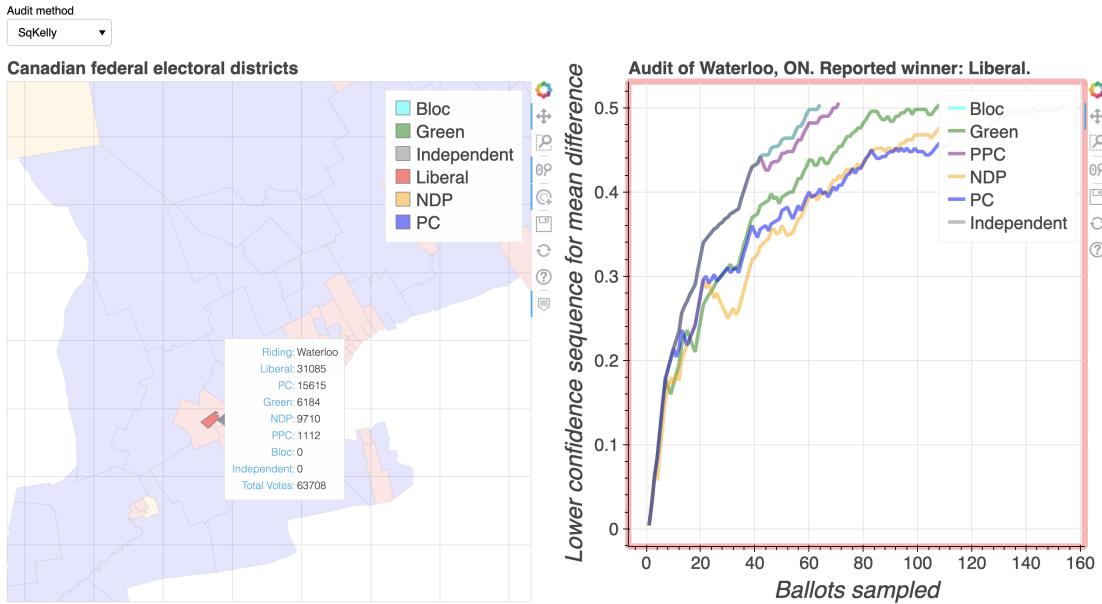


Figure 4.7: Example risk-limiting audit for the riding of Waterloo, Ontario using SqKelly. This screenshot was captured after zooming the map of Figure 4.6 in on southern Ontario. In this example, it was (correctly) reported that the Liberal party received 31,085 out of 63,708 total votes. Clicking on Waterloo’s polygon will begin the audit shown in the right-hand side, which displays six  $(1 - \alpha)$  lower confidence sequences for the pairwise contests between the Liberal party and each reportedly losing party. The Liberal party’s win is certified once each of these confidence sequences exceeds  $1/2$ , which in this case happened after sampling roughly 160 ballots.

Following the notation of Section 4.2.2, recall that the electoral parameter of interest  $\mu_{w,\ell}$  is defined as

$$\mu_{w,\ell}^* := \frac{1}{N} \sum_{i=1}^N x_i^{w,\ell},$$

where

- $x_i^{w,\ell} = 1$  if the  $i^{\text{th}}$  ballot shows a vote for  $w$ ,
- $x_i^{w,\ell} = 0$  if the  $i^{\text{th}}$  ballot shows a vote for  $\ell$ , and
- $x_i^{w,\ell} = 1/2$  if the  $i^{\text{th}}$  ballot shows a vote for any other party.

Also recall that the reported assertion – “ $w$  received more votes than  $\ell$  for each  $\ell \in \mathcal{L}$ ” – is certified once the  $(1 - \alpha)$  lower confidence sequences for  $\mu_{w,\ell}^*$  exceed  $1/2$  for each  $\ell \in \mathcal{L}$ . Furthermore, this yields an RLA with risk limit  $\alpha$ , without needing to perform any multiplicity adjustments for constructing several confidence sequences (see Section 4.2.2 for more details). For example, the right-hand side plot of Figure 4.7 displays an RLA with risk-limit  $\alpha$  for the assertion “the Liberal party received the largest number of votes” by computing six  $(1 - \alpha)$  lower confidence sequences for  $\mu_{w,\ell}^*$ , where  $w = \text{Liberal}$ , and  $\ell \in \{\text{Bloc}, \text{Green}, \text{PPC}, \text{NDP}, \text{PC}, \text{Independent}\}$ .

It is important to keep in mind that electoral outcomes in the underlying data sets correspond to hand-counted paper ballot vote totals [40, 46]. Therefore, the right-hand side plot in the web application (e.g. Figure 4.7) demonstrates the length of time that an audit would last, given correctly-reported outcomes, *and* assuming that the recorded data match the true votes cast. In practice, our confidence sequences would only rely on an assertion to audit (e.g. “The Liberal party received the most votes”) and a simple random sample without replacement from the physical stack of ballots cast. Moreover, the web application is easily adapted to this practical scenario, an extension we plan to pursue in future work.

A key feature of this app is its interactivity. Users can hover their cursors over ridings to see reported vote totals, click and drag the map around, zoom in on regions of interest, and so on. When the user has found a riding they wish to audit, they can simply click on that riding’s polygon to immediately compute lower confidence sequences and begin the RLA (see Figure 4.7). Server-side computation and client-side updates are fully asynchronous, meaning users can interact with the app while the audit is being conducted, and the audit will not “lock up”. A demo of these features can be found online<sup>8</sup> and the code is available on GitHub.<sup>9</sup>

## 4.5 Risk-limiting tallies via confidence sequences

Rather than audit an already-announced electoral outcome, it may be of interest to determine (for the purposes of making a first announcement) the election winner with high probability, without counting all  $N$  ballots. Such procedures are known as risk-limiting tallies (RLTs), which were developed for coercion-resistant, end-to-end verifiable voting schemes [133]. For example, suppose a voter is being coerced to vote for Bob. If the final vote tally reveals that Bob received few or no votes, then the coercer will suspect that the voter did not comply with instructions. RLTs provide a way to mitigate this issue by providing high-probability guarantees that the reported winner truly won, leaving a large proportion of votes shrouded. In such cases, the voter is guaranteed plausible deniability, as they can claim to the coercer that their ballot is simply among the unrevealed ones.

---

<sup>8</sup>[ianws.com/audit\\_demo.mov](http://ianws.com/audit_demo.mov)

<sup>9</sup>[github.com/WannabeSmith/RiLACS/tree/main/canada\\_audit](https://github.com/WannabeSmith/RiLACS/tree/main/canada_audit)

While the motivations for RLTs are quite different from those for RLAs, the underlying techniques are similar. The same is true for confidence sequence-based RLTs. All methods introduced in this chapter can be applied to RLTs (with the exception of “*a priori* Kelly” since it depends on the reported outcome) but with two-sided power. Consider the martingales we discussed in Section 4.3.2,

$$M_t^D(\mu^*) := \sum_{d=1}^D \theta_d \prod_{i=1}^t \left( 1 + \frac{d}{(D+1)\mathcal{C}_i(\mu^*)} (X_i - \mathcal{C}_i(\mu^*)) \right), \quad (4.8)$$

where  $(\theta_1, \dots, \theta_D)$  are convex weights. Recall that our confidence sequences at a given time  $t$  were defined as those  $\mu \in [0, 1]$  for which  $M_t^D(\mu) < 1/\alpha$ . In other words, a given value  $\mu$  is only excluded from the confidence set if  $M_t^D(\mu)$  is large. However, notice that  $M_t^D(\mu)$  will become large if the conditional mean  $\mathcal{C}_t(\mu^*) \equiv \mathbb{E}(X_t | X_1, \dots, X_{t-1})$  is larger than the null conditional mean  $\mathcal{C}_t(\mu)$ , but the same cannot be said if  $\mathcal{C}_t(\mu^*) < \mathcal{C}_t(\mu)$ . As a consequence, the resulting confidence sequences are all one-sided *lower* confidence sequences. To ensure that our bounds have non-trivial two-sided power, we can simply combine (4.8) with a martingale that also grows when  $\mathcal{C}_t(\mu^*) < \mathcal{C}_t(\mu)$ .

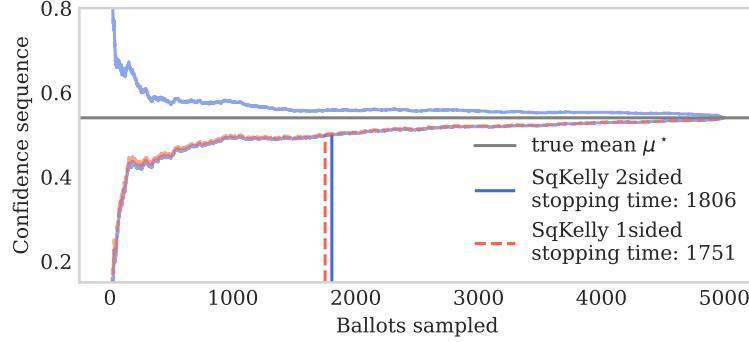


Figure 4.8: Confidence sequence-based risk-limiting tally for a two-candidate election. Unlike RLAs, RLTs require two-sided confidence sequences so that the true winner can be determined (with high probability) without access to an announced result. Notice that testing the same null  $H_0 : \mu^* \leq 0.5$  is less efficient in an RLT than in an RLA. This is a necessary sacrifice for having nontrivial power against other alternatives.

**Proposition 4.5.1.** *For nonnegative vectors  $(\theta_1^+, \dots, \theta_D^+)$  and  $(\theta_1^-, \dots, \theta_D^-)$  that each sum to one, define the processes*

$$\begin{aligned} M_t^{D+}(\mu) &:= \sum_{d=1}^D \theta_d^+ \prod_{i=1}^t \left( 1 + \frac{d}{(D+1)\mathcal{C}_i(\mu^*)} (X_i - \mathcal{C}_i(\mu^*)) \right), \\ M_t^{D-}(\mu) &:= \sum_{d=1}^D \theta_d^- \prod_{i=1}^t \left( 1 - \frac{d}{(D+1)(1-\mathcal{C}_i(\mu^*))} (X_i - \mathcal{C}_i(\mu^*)) \right). \end{aligned}$$

Next, for  $\beta \in [0, 1]$ , define their mixture

$$M_t^{D\pm}(\mu) := \beta M_t^{D+}(\mu) + (1 - \beta) M_t^{D-}(\mu).$$

Then,  $M_t^{D\pm}(\mu^*)$  is a nonnegative martingale starting at one. Consequently,

$$C_t^\pm := \{\mu \in [0, 1] : M_t^{D\pm}(\mu) < 1/\alpha\}$$

forms a  $(1 - \alpha)$  confidence sequence for  $\mu^*$ .

*Proof.* This follows immediately from the fact that both  $M_t^{D+}(\mu^*)$  and  $M_t^{D-}(\mu^*)$  are martingales with respect to the same filtration, and that convex combinations of such martingales are also martingales.  $\square$

With this setup and notation in mind,  $M_t^D$  as defined in Section 4.3.2 is a special case of  $M_t^{D\pm}$  with  $\beta = 1$ . As noted by [133], RLTs involving multiple assertions *do* require correction for multiple testing, unlike RLAs. The same is true for confidence sequence-based RLTs (and hence the tricks of Section 4.2.2 do not apply). It suffices to perform a simple Bonferroni correction by constructing  $(1 - \alpha/K)$  confidence sequences to establish  $K$  simultaneous assertions.

## 4.6 Summary

This chapter presented a general framework for conducting risk-limiting audits based on confidence sequences, and derived computationally and statistically efficient martingales for computing them. We showed how *a priori* Kelly takes advantage of the reported vote totals (if available) to stop ballot-polling audits significantly earlier than extant ballot-polling methods, and how alternative martingales such as SqKelly also provide strong empirical performance in the absence of reported outcomes. Finally, we demonstrated how a simple tweak to the aforementioned algorithms provides two-sided confidence sequences, which can be used to perform risk-limiting tallies. Confidence sequences and these martingales can be applied to ballot-level comparison audits and batch-level comparison audits as well, using “overstatement assorters” [246], which reduce comparison audits to the same canonical statistical problem: testing whether the mean of any list in a collection of non-negative bounded lists is at most  $1/2$ . We hope that this new perspective on RLAs and its associated software will aid in making election audits simpler, faster, and more transparent.

## 4.A Maximizing a proxy for $M_N(1/2)$

In Section 4.3, equation (4.4), we considered the product

$$\widetilde{M}_N^\lambda := \prod_{i=1}^N (1 + \lambda(x_i - 1/2)), \quad (4.9)$$

as an (inexact) proxy for  $M_N(1/2)$ , the final value of the process  $(M_t(1/2))_{t=0}^N$ . Let us now show that the maximizer of  $\widetilde{M}_N^\lambda$  is given by

$$\lambda' := 2 \frac{N'_A - N'_B}{N'_A + N'_B}. \quad (4.10)$$

*Proof.* To begin, note that the maximizer of  $\widetilde{M}_N^\lambda$  is exactly the maximizer of  $\log(\widetilde{M}_N^\lambda)$  due to the monotonicity of  $\log(\cdot)$ . Taking the derivative of  $\log(\widetilde{M}_N^\lambda)$  and setting it to zero, we find that  $\widetilde{M}_N^\lambda$  is maximized by the value of  $\lambda'$  that solves

$$\sum_{i=1}^N \frac{x_i - 1/2}{1 + \lambda'(x_i - 1/2)} = 0. \quad (4.11)$$

Breaking above sum up into terms for which the ballots are ones, zeros, and halves, respectively, we have that (4.11) reduces to

$$\begin{aligned} 0 &= \sum_{i:x_i=1} \frac{x_i - 1/2}{1 + \lambda'(x_i - 1/2)} + \sum_{i:x_i=0} \frac{x_i - 1/2}{1 + \lambda'(x_i - 1/2)} + \sum_{i:x_i=1/2} \frac{x_i - 1/2}{1 + \lambda'(x_i - 1/2)} \\ &= \sum_{i:x_i=1} \frac{1/2}{1 + \lambda'/2} + \sum_{i:x_i=0} \frac{-1/2}{1 + -\lambda'/2} + \sum_{i:x_i=1/2} \frac{0}{1 + \lambda' \cdot 0} \\ &= N'_A \frac{1/2}{1 + \lambda'/2} - N'_B \frac{1/2}{1 - \lambda'/2}. \end{aligned}$$

Solving the above equation for  $\lambda'$  yields the desired result given in (4.10). This completes the proof.  $\square$

# Chapter 5

## Nonparametric extensions of randomized response for private confidence sets

### 5.1 Introduction

It is easier than ever for mobile apps and web browsers to collect massive amounts of sensitive data about individuals. *Differential privacy* (DP) provides a framework that leverages statistical noise to limit the risk of sensitive information disclosure [97]. The goal of private data analysis is to extract meaningful population-level information from the data (whether in the form of machine learning model training, statistical inference, etc.) while preserving the privacy of individuals via DP. In particular, this chapter will focus on statistical inference (e.g. confidence intervals and  $p$ -values) for population means under DP constraints.

As motivating examples, suppose a city wishes to survey households to calculate the approval rating of their mayor, or an IT company aims to understand whether a redesigned homepage will lead to the average user spending more time on it. Both problems can be framed as estimating the mean of some (potentially large) population, but it may be infeasible to query every single household or all possible website users. Fortunately, a *sample mean* can still be used to estimate the population mean with some degree of precision. For example, a city may randomly choose households to query, or the technology company may show 10% of users the redesigned webpage at random. This is often referred to as “A/B testing”, and we expand on this application under privacy constraints in Section 5.4. When making decisions, however, it is crucial to both calculate sample means *and* quantify the uncertainty in those estimates. However, calculating confidence intervals under local differential privacy constraints (defined in Section 5.1.1) poses a unique statistical challenge, because these intervals must incorporate both the uncertainty introduced from random sampling *and* from the privacy mechanism. This chapter studies and provides a nonparametric solution to precisely this challenge.

### 5.1.1 Background: Local Differential Privacy

There are two main models of privacy within the DP framework: *central* and *local* DP (LDP) [97, 146, 96]. The former involves a centralized data aggregator that is trusted with constructing privatized output from raw data, while the latter performs privatization at the “local” or “personal” level (e.g. on an individual’s smartphone before leaving the device) so that trust need not be placed in any data collector. Both models have their advantages and disadvantages: LDP is a more restrictive model of privacy and thus in general requires more noise to be added. On the other hand, the stronger privacy guarantees that do not require a trusted central aggregator make LDP an attractive framework in practice. This chapter deals exclusively with LDP.

Making our setup more precise, suppose  $X_1, X_2, \dots$  is a (potentially infinite) sequence of  $[0, 1]$ -valued random variables. We could instead have assumed boundedness on any known interval  $[a, b]$  since we can always translate and scale the interval to  $[0, 1]$  via the transformation  $x \mapsto (x - a)/(b - a)$ . We will refer to  $(X)_{t=1}^{\infty}$  as the “raw” or “sensitive” data that are yet to be privatized. Following the notation of Duchi et al. [90] the privatized views  $Z_1, Z_2, \dots$  of  $X_1, X_2, \dots$ , respectively are generated by a sequence of conditional distributions  $Q_1, Q_2, \dots$  which we refer to as the *privacy mechanism*. Throughout this chapter, we will allow this privacy mechanism to be *sequentially interactive*, meaning that the distribution  $Q_i$  of  $Z_i$  may depend on the past privatized observations  $Z_1^{i-1} := (Z_1, \dots, Z_{i-1})$  [90]. In other words, the privatized view  $Z_i$  of  $X_i$  has a conditional distribution  $Q_i(\cdot | X_i = x, Z_1^{i-1} = z_1^{i-1})$ . Following Duchi et al. [90, 92] we say that  $Q_i$  satisfies  $\varepsilon$ -local differential privacy if for all  $z_1, \dots, z_{i-1} \in [0, 1]$  and  $x, \tilde{x} \in [0, 1]$ , the following likelihood ratio is uniformly bounded:

$$\sup_{z \in [0, 1]} \frac{q_i(z | X_i = x, Z_1^{i-1} = z_1^{i-1})}{q_i(z | X_i = \tilde{x}, Z_1^{i-1} = z_1^{i-1})} \leq \exp\{\varepsilon\}, \quad (5.1)$$

where  $q_i$  is the density (or Radon-Nikodym derivative) of  $Q_i$  with respect to some dominating measure. In the non-interactive case where the dependence on  $Z_1^{i-1}$  is dropped, (5.1) simplifies to the usual  $\varepsilon$ -LDP definition [96]. To put  $\varepsilon > 0$  in a real-world context, Apple uses privacy levels in the range of  $\varepsilon \in \{2, 4, 8\}$  on macOS and iOS devices for various  $\varepsilon$ -LDP data collection tasks, including health data type usage, emoji suggestions, and lookup hints [10]. See Figure 5.1 to intuit how  $\varepsilon$  affects the widths of confidence intervals that we develop.

### 5.1.2 Contributions and Outline

Our primary contributions are threefold: (a) privacy mechanisms, (b) CIs, and (c) time-uniform CSs.

- (a) We prove local DP guarantees of “Nonparametric randomized response” (NPRR) — a sequentially interactive, nonparametric generalization of Warner’s randomized response [278] for bounded data (Section 5.2).
- (b) We derive several CIs for the mean of bounded random variables that are privatized by NPRR (Section 5.3). We believe Section 5.3 introduces the first private nonparametric and nonasymptotic CIs for means of bounded random variables.

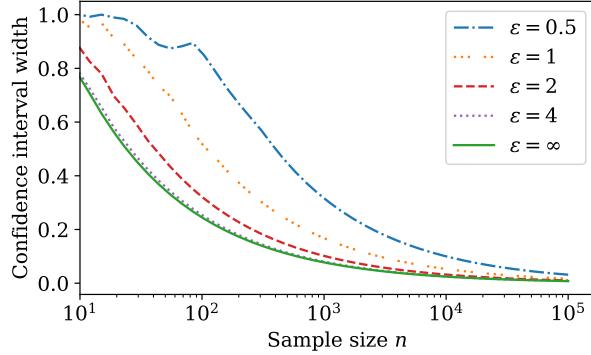


Figure 5.1: Widths of private 90%-CIs for the mean of a uniform distribution using our private Hoeffding CI given in (5.9) for various levels of  $\varepsilon$  ranging from  $\varepsilon = 0.5$  (very private) to  $\varepsilon = \infty$  (no privacy). Unsurprisingly, less privacy leads to sharper inference, but notice that inference is still practical, especially for  $\varepsilon \geq 2$ . For context, Apple uses  $\varepsilon \in \{2, 4, 8\}$  for various data collection tasks on iPhones [10]. At these levels of privacy, our CIs perform nearly as well as — and are in some cases indistinguishable from — the non-private Hoeffding CI.

- (c) We derive time-uniform CSs for the mean of bounded random variables that are privatized by NPPR, enabling private nonparametric sequential inference (Section 5.3.3). We also introduce a CS that is able to capture means that change over time under no stationarity conditions on the time-varying means (Section 5.3.4). We believe Sections 5.3.3 and 5.3.4 are the first private nonparametric CSs in the DP literature.

Furthermore, we show how all of the aforementioned techniques can be used to conduct private online A/B tests (Section 5.4). Finally, Section 5.5 summarizes our findings and discusses some additional results whose details can be found in the appendix. A Python package implementing our methods as well as code to reproduce the figures can be found on GitHub at [github.com/WannabeSmith/nppr](https://github.com/WannabeSmith/nppr).

### 5.1.3 Related Work

The literature on differentially private statistical inference is rich, including nonparametric estimation rates [279, 90, 91, 92, 143, 44, 6], parametric hypothesis testing and confidence intervals [269, 277, 108, 13, 145, 48, 138, 103, 71], median estimation [87], independence testing [66], online convex optimization [139], and parametric sequential hypothesis testing [276]. A more detailed summary can be found in Section 5.D.

The aforementioned works do not study the problem of private nonparametric confidence sets for population means. Prior work does exist on confidence intervals for the *sample mean of the data* [85, 274]. The most closely related work is that of Ding et al. [85, Section 2.1] who introduce the “1BitMean” mechanism which can be viewed as a special case of NPPR (Algorithm 5.2). They derive a private Hoeffding-type confidence interval for the *sample* mean of the data, but it is important to distinguish this from the more classical statistical task of *population* mean estimation. For example, if  $X_1, \dots, X_n$  are random variables drawn

from a distribution with mean  $\mu^*$ , then the *population mean* is  $\mu^*$ , while the *sample mean* is  $\hat{\mu}_n := \frac{1}{n} \sum_{i=1}^n X_i$ . A private CI for  $\mu^*$  incorporates randomness from both the mechanism *and* the data, while a CI for  $\hat{\mu}_n$  incorporates randomness from the mechanism *only*. Neither is a special case of the other, and some of our techniques allow for the (sequential) estimation of sample means (see Section 5.B.2 for details and explicit bounds) but this chapter is primarily focused on the problem of private *population* mean estimation.

## 5.2 Extending Warner's randomized response

Before discussing a nonparametric extension of randomized response, let us briefly review Warner's classical randomized response mechanism as well as the Laplace mechanism, discuss their shortcomings, and present a different mechanism that remedies them.

**Warner's randomized response.** When the raw data  $(X)_{t=1}^\infty$  are binary, one of the oldest and simplest privacy mechanisms is Warner's *randomized response* (RR) [278]. Warner's RR was introduced decades before the very definition of DP, but was later shown to satisfy LDP by Dwork and Roth [96]. RR was introduced as a method to provide plausible deniability to subjects when answering sensitive survey questions [278], and proceeds as follows: when presented with a sensitive Yes/No question (e.g. “have you ever used illicit drugs?”), the subject flips a biased coin with  $\mathbb{P}(\text{Heads}) = r \in (0, 1]$ . If the coin comes up heads, then the subject answers truthfully; if tails, the subject answers “Yes” or “No” (encoded as 1 and 0, respectively) with equal probability 1/2. It is easy to see that this mechanism satisfies  $\varepsilon$ -LDP with  $\varepsilon = \log(1 + \frac{2r}{1-r})$  by bounding the likelihood ratio of the privatized response distributions: for any true response  $x \in \{0, 1\}$ , let  $q(z | X = x) = r\mathbf{1}(z = x) + (1 - r)/2$  denote the conditional probability mass function of its privatized view. Then for any  $x, \tilde{x} \in \{0, 1\}$ ,

$$\sup_{z \in \{0, 1\}} \frac{q(z | X = x)}{q(z | X = \tilde{x})} \leq 1 + \frac{2r}{1-r}, \quad (5.2)$$

and hence RR satisfies  $\varepsilon$ -LDP with  $\varepsilon = \log(1 + \frac{2r}{1-r})$ . In Section 5.B.1, we show how one can derive a CI for the mean of Bernoulli random variables when they are privatized via RR, but as we will see in Section 5.3, this will be an immediate corollary of a more general result for bounded random variables (Theorem 5.3.2).

One downside of RR, however, is that it takes binary data as input. On the other hand, the famous Laplace mechanism satisfies  $\varepsilon$ -LDP for *bounded* data, including binary ones.

**The Laplace mechanism.** The Laplace mechanism appeared in the very same paper that introduced DP [97]. Algorithm 5.1 recalls the (sequentially interactive) Laplace mechanism [90].

Algorithm 5.1: Sequentially interactive Laplace mechanism

```

1: for  $t = 1, 2, \dots$  do
2:   Choose  $\varepsilon_t$  based on  $Z_1^{t-1}$ 
3:   Generate  $\mathcal{L}_t \sim \text{Laplace}(1/\varepsilon_t)$ 
4:    $Z_t \leftarrow X_t + \mathcal{L}_t$ 
5: end for

```

It is well-known that  $Z_t$  is (conditionally)  $\varepsilon_t$ -LDP (given  $Z_1^{t-1}$ ) for each  $t$  [97]. Section 5.B.4 derives novel CIs and CSs for population means under the Laplace mechanism, but we omit them here for brevity as a different mechanism (to be described shortly) will yield better bounds.

**Nonparametric randomized response (NPRR).** The mechanism we use, which we call “Nonparametric randomized response” (NPRR) serves as a sequentially interactive generalization of RR for arbitrary bounded data by simply combining stochastic rounding [21, 107, 130] with  $k$ -RR – a categorical but non-interactive generalization of Warner’s RR introduced by Kairouz et al. [141, 142], and also considered by Li et al. [174] under the name “Generalized Randomized Response”. Note that Kairouz et al. [141, 142] use  $k$  to refer to the number of unique values that the input and output data can take on, which is  $k = G + 1$  in the case of Algorithm 5.2. NPRR is explicitly described in Algorithm 5.2, and we summarize its LDP guarantees in Theorem 5.2.1.

Algorithm 5.2: Nonparametric randomized response (NPRR)

```

1: Algorithm: Nonparametric randomized response (NPRR)
2: for  $t = 1, 2, \dots$  do
3:   // Step 1: Discretize  $X_t$  into  $Y_t$  via stochastic rounding.
4:   Choose integer  $G_t \geq 1$  based on  $Z_1^{t-1}$ 
5:    $X_t^{\text{ceil}} \leftarrow \lceil G_t X_t \rceil / G_t$ ,  $X_t^{\text{floor}} \leftarrow \lfloor G_t X_t \rfloor / G_t$ 
6:   if  $X_t^{\text{ceil}} == X_t^{\text{floor}}$  then
7:      $Y_t \leftarrow X_t$ 
8:   else
9:     Generate  $Y_t \sim \begin{cases} X_t^{\text{ceil}} & \text{w.p. } G_t \cdot (X_t - X_t^{\text{floor}}) \\ X_t^{\text{floor}} & \text{w.p. } G_t \cdot (X_t^{\text{ceil}} - X_t) \end{cases}$ 
10:  end if
11:  // Step 2: Privatize  $Y_t$  into  $Z_t$  via  $k$ -RR.
12:  Choose  $r_t \in (0, 1]$  based on  $Z_1^{t-1}$ 
13:  Generate  $\mathcal{U}_t \sim \text{Unif} \left\{ 0, \frac{1}{G_t}, \frac{2}{G_t}, \dots, \frac{G_t}{G_t} \right\}$ 
14:  Generate  $Z_t \sim \begin{cases} Y_t & \text{w.p. } r_t \\ \mathcal{U}_t & \text{w.p. } 1 - r_t \end{cases}$ 
15: end for

```

Up to rescaling of the outputs, NPRR can be viewed as a sequentially interactive analogue of the local randomizer given in Balle et al. [16, Algorithm 2] in the context of the shuffle model. We nevertheless refer to this mechanism as “NPRR” to emphasize its connection to

Warner's RR. Indeed, if  $(X)_{t=1}^{\infty} 1$  are  $\{0, 1\}$ -valued, and if we set  $G_1 = G_2 = \dots = 1$  and  $r_1 = r_2 = \dots = r \in (0, 1]$ , then no stochastic rounding occurs and NPPR recovers RR exactly, making NPPR a sequentially interactive and nonparametric generalization for bounded data. Finally, if we let NPPR be non-interactive and set  $G_1 = \dots = G_n = 1$ , then NPPR recovers the "1BitMean" mechanism of Ding et al. [85]. As such, the form of NPPR given in Algorithm 5.2 makes transparent generalizations of and connections between Warner [278], Kairouz et al. [142], Li et al. [174], Ding et al. [85], and Balle et al. [16].<sup>1</sup> Let us now formalize NPPR's LDP guarantees.

**Theorem 5.2.1** (NPPR satisfies LDP). *Suppose  $(Z)_{t=1}^{\infty} 1$  are generated according to NPPR. Then for each  $t \in \{1, 2, \dots\}$ ,  $Z_t$  is a conditionally  $\varepsilon_t$ -LDP view of  $X_t$  with*

$$\varepsilon_t := \log \left( 1 + \frac{(G_t + 1)r_t}{1 - r_t} \right). \quad (5.3)$$

The proof in Section 5.A.2 proceeds by bounding the conditional likelihood ratio for any two data points  $x, \tilde{x} \in [0, 1]$  similar to (5.1). In all of the results that follow in the following sections, we will write expressions in terms of  $(r)_{t=1}^n$ , but these can always be chosen given desired  $(\varepsilon)_{t=1}^n$  levels via the relationship

$$r_t = \frac{\exp\{\varepsilon_t\} - 1}{\exp\{\varepsilon_t\} + G_t}. \quad (5.4)$$

In the familiar special case of  $r_t = r \in (0, 1]$  and  $G_t = G \in \{1, 2, \dots\}$  for each  $t$ , we have that  $(Z)_{t=1}^{\infty} 1$  satisfy  $\varepsilon$ -LDP with  $\varepsilon := \log(1 + (G + 1)r/(1 - r))$ . Notice that when  $G_t = 1$  for each  $t$ , we have that NPPR satisfies  $\varepsilon$ -LDP with the same value of  $\varepsilon$  as Warner's RR. Consequently, there is no privacy lost from instantiating the more general NPPR to the binary case.

*Remark 7* (Who chooses  $\varepsilon_t$ ,  $r_t$ , or  $G_t$ , and how?). Due to the sequential interactivity of NPPR, individuals can specify their own levels of privacy, or the parameters  $(r_t, G_t)_{t=1}^{\infty}$  can be adjusted over time (e.g. if the data collector chooses to decrease  $\varepsilon_t$  for regulatory reasons, or increase  $\varepsilon_t$  to obtain sharper inference). Formally,  $(r_t, G_t)$  can be chosen in any way as long as they are *predictable*, meaning that they can depend on  $Z_1^{t-1}$ . Nevertheless, sequential interactivity is completely optional, and the data collector is free to set  $(r_t, G_t) = (r, G)$  for every  $t$  to recover the familiar notion of  $\varepsilon$ -LDP.

**Why use NPPR instead of Laplace?** While RR is limited to privatizing binary data, the Laplace mechanism can handle bounded data, so why use NPPR as an alternative to the two? The reason stems from our original motivation: to derive locally private nonparametric, nonasymptotic confidence sets for means of bounded random variables. To achieve this, we will ultimately use modern concentration techniques from the literature on (non-private) confidence sets, many of which exploit boundedness in clever ways to yield clean, closed-form expressions and/or empirically tight confidence intervals. Since the Laplace mechanism does not preserve the boundedness of its input, it is not clear how those techniques can be used for Laplace-privatized data (though we do derive novel Laplace-based solutions using a different approach

---

<sup>1</sup>Notice that Ding et al. [85]'s  $\alpha$ -point rounding mechanism is different from NPPR as NPPR shifts the mean of the inputs but alpha-point rounding leaves the mean unchanged.

in Section 5.B.4, but they are ultimately outperformed by those that we derive based on NPPR). NPPR on the other hand, preserves the input's boundedness, making it possible to apply analogues of these modern concentration techniques for NPPR-privatized data. The efficiency gains that result from this approach are illustrated in Figures 5.4 and 5.5.

In addition to being useful for deriving simple and efficient confidence sets, NPPR has some other orthogonal advantages over the Laplace mechanism. First, NPPR has reduced storage requirements: Once a  $[0, 1]$ -bounded random variable has been privatized via Laplace, the output is a floating-point number, requiring 64 bits to store as a double-precision float. In contrast, NPPR outputs one of  $(G + 1)$  different values, hence requiring only  $\lceil \log_2(G + 1) \rceil$  bits to store. Moreover, storing the NPPR-privatized view of  $x$  will never require more memory than storing  $x$  itself (unless  $G$  is set to nonsensical values larger than  $2^{64}$ ), while Laplace-privatized views will always require at least enough memory to represent floating point numbers.

Second, NPPR is automatically resistant to the floating-point attacks that the Laplace mechanism suffers from. Mironov [189] showed that storing Laplace output as a floating-point number can leak information about the input  $x$ , thereby compromising its LDP guarantees. While Mironov [189] discusses remedies to this issue, practitioners may still naively apply the Laplace mechanism using common software packages and remain vulnerable to these so-called “floating-point attacks”. In contrast, the discrete representation of NPPR’s output is not vulnerable to such attacks, without the need for remedies at all. Note that while NPPR may have to deal with floating point numbers as input, they are transformed into discrete random variables *before* any  $\epsilon$ -LDP guarantees are added. The privatization step (transforming  $Y_t$  into  $Z_t$  in Algorithm 5.2) takes one of  $G_t + 1$  values as input and produces one of  $G_t + 1$  values as output, thereby sidestepping any need to handle floating point numbers.

The remainder of this chapter will focus solely on constructing efficient locally private confidence sets, but the above benefits can be seen as “free” byproducts of NPPR’s design.

### 5.3 Private confidence intervals for bounded data

Making matters formal, let  $\mathcal{P}_\mu$  be the set of distributions on  $[0, 1]$  with population mean  $\mu \in [0, 1]$ .  $\mathcal{P}_\mu$  is a convex set of distributions with no common dominating measure, since it consists of discrete and continuous distributions, as well as their mixtures. We will consider sequences of random variables  $(X_i)_{i=1}^n$  drawn from the product distribution  $\prod_{i=1}^n P_i$  where  $n \in \{1, 2, \dots, \infty\}$  and each  $P_i \in \mathcal{P}_\mu$ . For succinctness, define the following set of distributions,

$$\mathcal{P}_\mu^n := \left\{ \prod_{i=1}^n P_i \text{ such that each } P_i \in \mathcal{P}_\mu \right\}, \quad (5.5)$$

for  $n \in \{1, 2, \dots, \infty\}$ . In words,  $\mathcal{P}_\mu^n$  contains distributions for which the random variables are independent and  $[0, 1]$ -bounded with mean  $\mu$  but need not be identically distributed. We use the notation  $(X)_{t=1}^n \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^n$  to indicate that  $(X)_{t=1}^n$  are independent with mean  $\mu^*$ . The goal is now to derive sharp CIs and time-uniform CSs for  $\mu^*$  given NPPR-privatized views of  $(X)_{t=1}^n$ .

Let us write  $\mathcal{Q}_{\mu^*}^n$  to denote the set of joint distributions on NPPR's output, where we have left the dependence on each  $G_t$  and  $r_t$  implicit. In other words, given  $(X)_{t=1}^n \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^n$ , their NPPR-induced privatized views  $(Z)_{t=1}^n$  have a joint distribution from  $Q$  for some  $Q \in \mathcal{Q}_{\mu^*}^n$ .

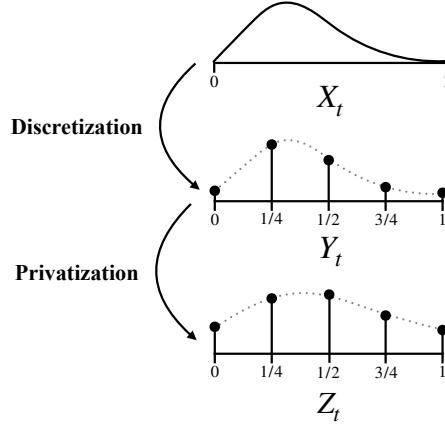


Figure 5.2: An illustration of how a distribution  $Q \in \mathcal{Q}_{\mu^*}^n$  can arise from applying NPPR with  $G_t = 4$  to draws from the input distribution  $P \in \mathcal{P}_{\mu^*}^n$ . Raw data  $X_t$  are discretized into  $Y_t$  so that  $Y_t$  has finite support but so that  $\mu^* = \mathbb{E}(X_t) = \mathbb{E}(Y_t)$ . The discrete  $Y_t$  are then privatized into  $Z_t$  with conditional mean  $\mathbb{E}(Z_t | Z_1^{t-1}) = \zeta_t(\mu^*) = r_t \mu^* + (1 - r_t)/2$  by being mixed with independent uniform noise  $U_t \sim \text{Unif}\{0, 1/4, 1/2, 3/4, 1\}$ .

### 5.3.1 What is a Locally Private Confidence Set?

Let first define what we mean by locally private confidence intervals (LPCI) and sequences (LPCS), and subsequently derive them for means of bounded random variables.

*Definition 5.3.1* (Locally private confidence sets). Let  $\varepsilon \equiv (\varepsilon)_{t=1}^n \equiv (\varepsilon_t)_t$ . We say that  $L_n$  is a lower  $(1 - \alpha, \varepsilon)$ -LPCI for a parameter  $\theta^*$ , and with respect to the raw data  $(X)_{t=1}^n$  if  $L_n$  is a lower  $(1 - \alpha)$ -CI for  $\theta^*$ , meaning

$$\mathbb{P}(\theta^* \geq L_n) \geq 1 - \alpha, \quad (5.6)$$

and if  $L_n \equiv L(Z_1, \dots, Z_n)$  is only a function of the  $\varepsilon_t$ -LDP view  $Z_t$  of  $X_t$  for each  $t$ , but not of  $(X)_{t=1}^n$  directly.

Similarly, we say that  $(L_t)_{t=1}^\infty$  is a lower  $(1 - \alpha, \varepsilon)$ -LPCS for  $\theta^*$  if (5.6) is replaced with the time-uniform guarantee

$$\mathbb{P}(\forall t, \theta^* \geq L_t) \geq 1 - \alpha. \quad (5.7)$$

Upper CIs and CSs are defined analogously.

Note that LPCIs and LPCSs also satisfy  $\varepsilon$ -LDP, since DP is closed under post-processing [96].

### 5.3.2 A Locally Private Hoeffding CI via NPPR

First, we present a private generalization of Hoeffding's inequality under NPPR.

**Theorem 5.3.2 (NPPR-H).** Suppose  $(X)_{t=1}^n \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^n$ , and let  $(Z)_{t=1}^n \sim Q \in \mathcal{Q}_{\mu^*}^n$  be their privatized views via NPPR. Define the NPPR-adjusted sample mean

$$\hat{\mu}_n := \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - (1 - r_i)/2)}{\frac{1}{n} \sum_{i=1}^n r_i}. \quad (5.8)$$

Then,

$$\dot{L}_n^H := \hat{\mu}_n - \sqrt{\frac{\log(1/\alpha)}{2n(\frac{1}{n} \sum_{i=1}^n r_i)^2}} \quad (5.9)$$

is a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCI for  $\mu^*$ .

The proof in Section 5.A.3 uses a locally private supermartingale variant of the Cramér-Chernoff bound. We recommend setting  $r_t$  for the desired  $\varepsilon_t$ -LDP level via the relationship in (5.4) and  $G_t := 1$  for all  $t$  (the reason behind which we will discuss in Remark 10). Notice that in the non-private setting where we set  $r_i = 1$  for all  $i$ , then  $\dot{L}_n^H$  recovers the classical Hoeffding inequality exactly [120]. Moreover, notice that if  $(X)_{t=1}^n$  took values in  $[a, b]$  instead of  $[0, 1]$ , then (5.9) would simply scale with  $(b - a)$  in the same manner as Hoeffding [120]. Recall as discussed in Remark 7 that  $(r)_{t=1}^n$  could be chosen either by the data collector or by the subject whose data are being collected, but that sequential interactivity is optional.

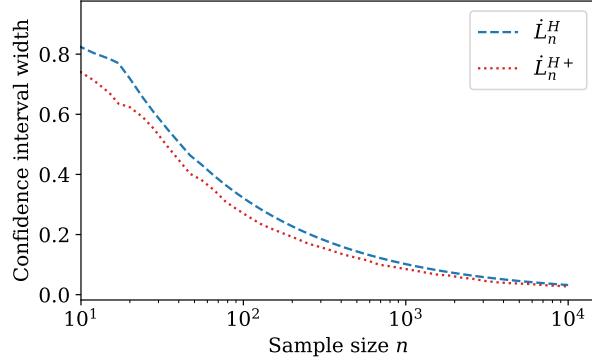


Figure 5.3: Two  $(90\%, 2)$ -LPCIs:  $\dot{L}_n^H$  given in (5.9) and  $\dot{L}_n^{H+}$  given in (5.10) – i.e. these are  $(1 - \alpha, \varepsilon)$ -LPCIs with  $\alpha = 0.1$  and  $\varepsilon = 2$ . Notice that the latter can be tighter than the former. Indeed this is because  $L_n^{H+}$  is never looser than  $L_n^H$  (by definition) but strictly tighter with positive probability.

In fact, we can strictly improve on (5.9) by exploiting the martingale dependence of this problem. Indeed, under the same assumptions as Theorem 5.3.2, we have that

$$\dot{L}_n^{H+} := \max_{1 \leq t \leq n} \left\{ \hat{\mu}_t - \frac{\log(1/\alpha) + t\lambda_n^2/8}{\lambda_n \sum_{i=1}^t r_i} \right\} \quad (5.10)$$

is also a lower  $(1 - \alpha, (\varepsilon)_{t=1}^n)$ -LPCI for  $\mu^*$ , where  $\lambda_n := \sqrt{8 \log(1/\alpha)/n}$ . Notice that  $\dot{L}_n^{H+}$  is at least as tight as  $\dot{L}_n^H$  since the  $n^{\text{th}}$  term in the above  $\max_{1 \leq t \leq n}$  recovers  $\dot{L}_n^H$  exactly. Moreover,  $\dot{L}_n^{H+}$  is *strictly* tighter than  $\dot{L}_n^H$  with positive probability, and hence strictly tighter in expectation:  $\mathbb{E}(\dot{L}_n^{H+}) > \mathbb{E}(\dot{L}_n^H)$ .

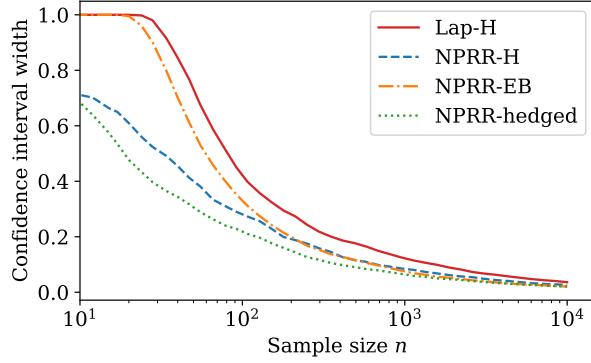


Figure 5.4: Widths of  $(90\%, 2)$ -LPCIs for the mean of a  $\text{Beta}(50, 50)$  distribution. Hoeffding-based methods (Lap-H and NPPR-H found in Corollary 5.B.3 and Theorem 5.3.2) do slightly worse than the variance-adaptive ones (NPPR-EB and NPPR-hedged in Proposition 5.B.2 and Theorem 5.B.1), but in all cases, CIs that rely on NPPR seem to outperform Lap-H in both small and large  $n$  regimes.

*Remark 8* (Minimax rate optimality of (5.9)). In the case of  $\varepsilon_1 = \dots = \varepsilon_n = \varepsilon \in (0, 1]$ , Duchi et al. [90, Proposition 1] give minimax estimation rates for the problem of nonparametric mean estimation. Their lower bounds say that for any  $\varepsilon$ -LDP mechanism and estimator  $\hat{\mu}_n$  for  $\mu^*$ , the root mean squared error  $\sqrt{\mathbb{E}(\hat{\mu}_n - \mu^*)^2}$  cannot scale faster than  $O(1/\sqrt{n\varepsilon^2})$ . Since NPPR is  $\varepsilon$ -LDP with  $\varepsilon = \log(1 + 2r/(1 - r))$ , we have that  $r \asymp \varepsilon$  up to constants on  $\varepsilon \in (0, 1]$ . It follows that  $\dot{L}_n^H \asymp 1/\sqrt{n\varepsilon^2}$ , matching the minimax estimation rate. Of course, the midpoint of a CI for  $\mu^*$  can always be used as an estimator for  $\mu^*$ , and hence we cannot expect the width of the CI to shrink faster than the minimax estimation rate. While explicit minimax lower bounds do not exist for the setting where  $\varepsilon_i \neq \varepsilon_j$  for some  $i, j$ , notice that instead of scaling with  $r^{-1}$  (which we would have if  $r_i = r_j$  for  $i \neq j$ ),  $\dot{L}_n^H$  scales with  $(\frac{1}{n} \sum_{i=1}^n r_i)^{-1}$ , and hence our bounds seem to be of the right order when  $\varepsilon$  is permitted to change.

*Remark 9* (The relationship between  $\varepsilon$  and (5.9) for practical levels of privacy). As mentioned in the introduction and in Figure 5.1, Apple uses values of  $\varepsilon \in \{2, 4, 8\}$  for various  $\varepsilon$ -LDP data collection tasks on iPhones [10]. Note that for  $G = 1$ , having  $\varepsilon$  take values of 2, 4, and 8 corresponds to  $r$  being roughly 0.762, 0.964, and 0.999, respectively, via the relationship  $r = (\exp(\varepsilon) - 1)/(\exp(\varepsilon) + 1)$ . As such, (5.9) simply inflates the width of the non-private Hoeffding CI by  $0.762^{-1}$ ,  $0.964^{-1}$ , and  $0.999^{-1}$ , respectively. Hence larger  $\varepsilon$  (e.g.  $\varepsilon \geq 4$ ) leads to CIs that are nearly indistinguishable from the non-private case (Figure 5.1).

*Remark 10.* Since Hoeffding-type bounds are not variance-adaptive (meaning they use a worst-case upper-bound on the variance of bounded random variables as in Hoeffding [120]), they

do not benefit from the additional granularity when setting  $G_t \geq 2$  (see Section 5.B.3 for a detailed mathematical explanation). As such, we set  $G_t = 1$  for each  $t$  when running NPPR-H. Nevertheless, other CIs are capable of adapting to the variance with  $G_t \geq 2$ , and these are discussed in Section 5.B.5, with some suggestions for how to choose  $G_t \geq 2$  in Section 5.B.6. Nevertheless, the empirical performance of our variance-adaptive CIs is illustrated in Figure 5.4.

### 5.3.3 Time-uniform Confidence Sequences for $\mu^*$

Previously, we focused on constructing a (lower) CI  $L_n$  for  $\mu^*$ , meaning that  $L_n$  satisfies the high-probability guarantee  $\mathbb{P}(\mu^* \geq L_n) \geq 1 - \alpha$  for the prespecified sample size  $n$ . We will now derive CSs – i.e. entire sequences of CIs  $(L_t)_{t=1}^\infty$  – which have the stronger *time-uniform* coverage guarantee  $\mathbb{P}(\forall t, \mu^* \geq L_t) \geq 1 - \alpha$ , enabling anytime-valid inference in sequential regimes. In summary, if  $(L_t)_{t=1}^\infty$  is a lower  $(1 - \alpha)$ -CS, then  $L_\tau$  forms a valid  $(1 - \alpha)$ -CI at arbitrary stopping times  $\tau$  (including random and data-dependent times) and hence a practitioner can continuously update inferences as new data are collected, without any penalties for “peeking” at the data early. Let us now present a Hoeffding-type CS for  $\mu^*$ , serving as a time-uniform analogue of Theorem 5.3.2.

**Theorem 5.3.3 (NPPR-H-CS).** *Let  $(Z_t)_{t=1}^\infty \sim Q$  for some  $Q \in \mathcal{Q}_{\mu^*}^\infty$ . Define the modified mean estimator under NPPR:*

$$\hat{\mu}_t(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i \cdot (Z_i - (1 - r_i)/2)}{\sum_{i=1}^t r_i \lambda_i}, \quad (5.11)$$

and let  $(\lambda_t)_{t=1}^\infty$  be a real-valued sequence of tuning parameters (discussed in (5.32)). Then,

$$\bar{L}_t^H := \hat{\mu}_t(\lambda_1^t) - \frac{\log(1/\alpha) + \sum_{i=1}^t \lambda_i^2 / 8}{\sum_{i=1}^t r_i \lambda_i} \quad (5.12)$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCS for  $\mu^*$ .

The proof can be found in Section 5.A.4. Unlike Theorem 5.3.2, we suggest setting

$$\lambda_t := \sqrt{\frac{8 \log(1/\alpha)}{t \log(t+1)}} \wedge 1, \quad (5.13)$$

to ensure that  $\bar{L}_t^H \asymp O(\sqrt{\log t/t})$  up to  $\log \log t$  factors. Section 3.3.3 from Chapter 3 gives a derivation and discussion of  $\lambda_t$  and the  $O(\sqrt{\log t/t})$  rate. Similar to Theorem 5.3.2, we recommend setting  $r_t$  for the desired  $\varepsilon_t$ -LDP level via (5.4) and  $G_t := 1$  for all  $t$ .

The similarity between Theorem 5.3.3 and Theorem 5.3.2 is no coincidence: indeed, Theorem 5.3.2 is a corollary of Theorem 5.3.3 where we instantiated a CS at a fixed sample size  $n$  and set  $\lambda_1 = \dots = \lambda_n = \sqrt{8 \log(1/\alpha)/n}$ . In fact, every Cramér-Chernoff bound (even in the non-private regime) has an underlying supermartingale and CS that are rarely exploited [124], but setting  $\lambda$ 's as in Theorem 5.3.2 tightens these CSs for the fixed time  $n$  – yielding  $O(1/\sqrt{n})$  rates but only for a fixed  $n$  – while tuning  $\lambda_t$  as in (5.13) allows them to spread their efficiency over all  $t$  – yielding  $O(\sqrt{\log t/t})$  rates but for all  $t$  simultaneously. Notice that both the time-uniform and fixed-time bounds in Theorems 5.3.2 and 5.3.3 cover an unchanging

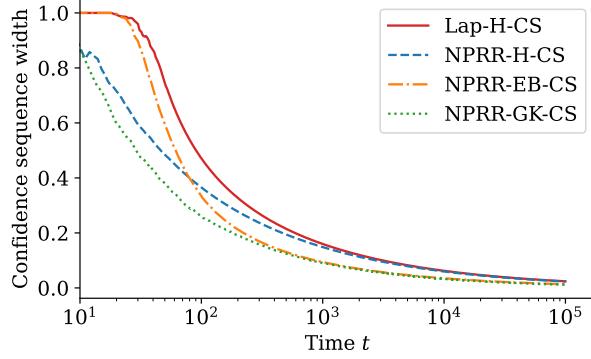


Figure 5.5: Widths of (90%, 2)-LPCSSs for the mean of a Beta(50, 50) distribution. Like Figure 5.4, Hoeffding-based methods (Lap-H-CS and NPPR-H-CS found in Proposition 5.B.1 and Theorem 5.3.3) do worse than the variance-adaptive ones (NPPR-EB-CS and NPPR-GK-CS in Proposition 5.B.3 and Theorem 5.B.2) for large  $t$ , though NPPR-H-CS does outperform NPPR-EB-CS for small  $t$ . Nevertheless, in all cases, we find that NPPR-based CSs outperform Lap-H-CS in both small and large  $t$  regimes.

real-valued mean  $\mu^* \in \mathbb{R}$  – in the following section, we will relax this assumption and allow for the mean of each  $X_i$  to change over time in an arbitrary manner, but still derive CSs for sensible parameters.

### 5.3.4 Confidence Sequences for Time-varying Means

All of the bounds derived thus far have been concerned with estimating some common  $\mu^*$  under the nonparametric assumption  $(X)_{t=1}^\infty \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^\infty$  and hence  $(Z)_{t=1}^\infty \sim Q$  for some  $Q \in \mathcal{Q}_{\mu^*}^\infty$ . Let us now consider the more general (and challenging) task of constructing CSs for the average mean so far  $\tilde{\mu}_t^* := \frac{1}{t} \sum_{i=1}^t \mu_i^*$  under the assumption that each  $X_t$  has a different mean  $\mu_t^*$ . In what follows, we require that NPPR is non-interactive, i.e.  $r_t = r \in (0, 1]$  and  $G_t = G \in \{1, 2, \dots\}$  for each  $t$ .

**Theorem 5.3.4** (Confidence sequences for time-varying means). *Suppose  $X_1, X_2, \dots$  are independent  $[0, 1]$ -bounded random variables with individual means  $\mathbb{E}X_t = \mu_t^*$  for each  $t$ , and let  $Z_1, Z_2, \dots$  be their privatized views according to NPPR without sequential interactivity. Define*

$$\hat{\mu}_t := \frac{\sum_{i=1}^t (Z_i - (1-r)/2)}{tr}, \quad (5.14)$$

$$\text{and } \tilde{B}_t^\pm := \sqrt{\frac{t\beta^2 + 1}{2(tr\beta)^2} \log \left( \frac{\sqrt{t\beta^2 + 1}}{\alpha} \right)}, \quad (5.15)$$

for any  $\beta > 0$ . Then,  $\tilde{C}_t^\pm := (\hat{\mu}_t \pm \tilde{B}_t^\pm)$  forms a two-sided  $(1 - \alpha, \varepsilon)$ -LPCS for  $\tilde{\mu}_t^*$ , where  $\varepsilon = \log(1 + \frac{2r}{1-r})$ .

The proof in Section 5.A.5 uses a sub-Gaussian mixture supermartingale technique similar

to Robbins [217] and Howard et al. [124, 125]. The parameter  $\beta > 0$  is a tuning parameter dictating a time for which the CS boundary is optimized. Regardless of how  $\beta$  is chosen,  $\tilde{C}_t^\pm$  has the time-uniform coverage guarantee given in Theorem 5.3.4 but finite-sample performance can be improved near a particular time  $t_0$  by selecting

$$\beta_\alpha(t_0) := \sqrt{\frac{-2 \log \alpha + \log(-2 \log \alpha + 1)}{t_0}}, \quad (5.16)$$

which approximately minimizes  $\tilde{B}_{t_0}$ ; see Howard et al. [125, Section 3.5] for details.

Notice that in the non-private case where  $r = 1$ , we have that  $\tilde{C}_t^\pm$  recovers Robbins' sub-Gaussian mixture CS [217, 125]. Notice that while Theorem 5.3.4 handles a strictly more general and challenging problem than the previous sections (by tracking a time-varying mean  $(\tilde{\mu}_t)_{t=1}^\infty$ ), it has the restriction that NPPR must be non-interactive. There is a technical reason for this that boils down to it being difficult to combine time-varying *tuning parameters* (such as those in Theorem 5.3.3) with time-varying *estimands* in the same CS. This challenge has appeared in other (non-private) works on CSs [282, 125]. In short, this chapter has methods for tracking a time-varying mean under non-interactive NPPR or a fixed mean under sequentially interactive NPPR, but not both simultaneously — this would be an interesting direction to explore in future work.

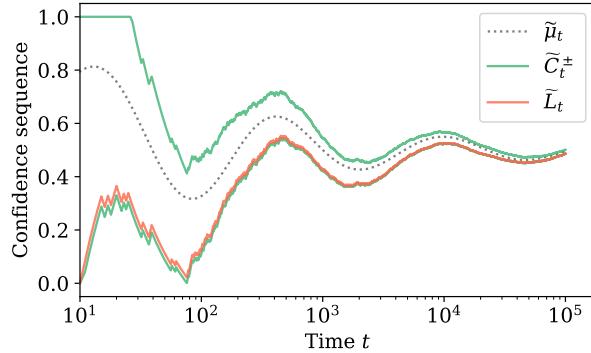


Figure 5.6: (90%, 2)-LPCSSs for the average time-varying mean so far  $\tilde{\mu}_t^*$  with the boundary optimized for  $t_0 = 100$ . In this example, we set  $\mu_t^* = \frac{1}{2} [1 - \sin(2 \log(e + t)) / \log(e + 0.01t)]$  to produce the displayed sinusoidal behavior. Notice that  $\tilde{L}_t$  is tighter at the expense of only being one-sided. In either case, however, the CSs adapt to non-stationarity and capture  $\tilde{\mu}_t^*$  uniformly over time.

A one-sided analogue of Theorem 5.3.4 is presented in Section 5.B.7 via slightly different techniques.

## 5.4 Illustration: Private online A/B testing

Our methods can be used to conduct locally private *online A/B tests* (sequential randomized experiments). Broadly, an A/B test is a statistically principled way of comparing two different

*treatments* – e.g. administering drug A versus drug B in a clinical trial. In its simplest form, A/B testing proceeds by (i) randomly assigning subjects to receive treatment A with some probability  $\pi \in (0, 1)$  and treatment B with probability  $1 - \pi$ , (ii) collecting some outcome measurement  $Y_t$  for each subject  $t \in \{1, 2, \dots\}$  – e.g. severity of headache after taking drug A or B – and (iii) measuring the difference in that outcome between the two groups. An *online* A/B test is one that is conducted sequentially over time – e.g. a sequential clinical trial where patients are recruited one after the other or in batches.

We now illustrate how to sequentially test for the mean difference in outcomes between groups A and B when only given access to locally private data. To set the stage, suppose that  $(A_1, Y_1), (A_2, Y_2), \dots$  are random variables such that  $A_t \sim \text{Bernoulli}(\pi)$  is 1 if subject  $t$  received treatment A and 0 if they received treatment B, and  $Y_t$  is a  $[0, 1]$ -bounded outcome of interest after being assigned treatment  $A_t$ .

Using the techniques of Section 5.3.4, we will construct  $(1 - \alpha)$ -CSs for the *time-varying mean*  $\tilde{\Delta}_t := \frac{1}{t} \sum_{i=1}^t \Delta_i$  where  $\Delta_i := \mathbb{E}(Y_i | A_i = 1) - \mathbb{E}(Y_i | A_i = 0)$  is the mean difference in the outcomes at time  $i$ . In words,  $\tilde{\Delta}_t$  is the mean difference in outcomes *among the subjects so far*.

Unlike Section 5.3.4, however, we will not directly privatize  $(Y)_{t=1}^\infty 1$ , but instead will apply NPPR to some “pseudo-outcomes”  $\varphi_t \equiv \varphi_t(Y_t, A_t)$  – functions of  $Y_t$  and  $A_t$ ,

$$\varphi_t := \frac{f_t + \frac{1}{1-\pi}}{\frac{1}{\pi} + \frac{1}{1-\pi}}, \quad \text{where } f_t := \left[ \frac{Y_t A_t}{\pi} - \frac{Y_t(1-A_t)}{1-\pi} \right].$$

Notice that due to the fact that  $Y_t, A_t \in [0, 1]$ , we have  $f_t \in [-1/(1-\pi), 1/\pi]$ , and hence  $\varphi_t \in [0, 1]$ . Now that we have  $[0, 1]$ -bounded random variables  $(\varphi)_{t=1}^\infty 1$ , we can obtain their NPPR-induced  $\varepsilon$ -LDP views  $(\psi)_{t=1}^\infty 1$  by setting  $G_t = 1$  and  $r_t = \exp\{\varepsilon - 1\}/\exp\{\varepsilon + 1\}$  for each  $t$ . Notice that we are privatizing  $\varphi_t$  which is a function of both  $Y_t$  and  $A_t$ , so both the outcome *and* the treatment are protected with  $\varepsilon$ -LDP.

**Corollary 5.4.1** (Locally private online A/B estimation). *Following the setup above, let  $(\psi)_{t=1}^\infty 1$  be the NPPR-induced privatized views of  $(\varphi)_{t=1}^\infty 1$ . Define the estimator*

$$\hat{\varphi}_t := \frac{\sum_{i=1}^t (\psi_i - (1-r)/2)}{tr}, \tag{5.17}$$

and set  $\tilde{B}_t$  as in (5.44). Then,

$$\tilde{L}_t^\Delta := -\frac{1}{1-\pi} + \left( \frac{1}{\pi} + \frac{1}{1-\pi} \right) (\hat{\varphi}_t - \tilde{B}_t) \tag{5.18}$$

is a lower  $(1 - \alpha, \varepsilon)$ -LPCS for  $\tilde{\Delta}_t$ .

The proof is an immediate consequence of the well-known fact about “inverse-probability-weighted” estimators that  $\mathbb{E}f_t = \Delta_t$  for every  $t$  [122, 225], combined with Proposition 5.B.4. Similarly, a two-sided CS can be obtained by replacing  $\hat{\varphi}_t - \tilde{B}_t$  in (5.18) with  $\hat{\varphi}_t \pm \tilde{B}_t^\pm$ , where  $\tilde{B}_t^\pm$  is given in (5.15).

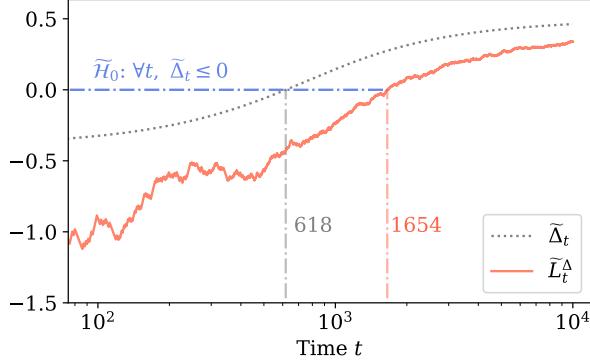


Figure 5.7: An example of Corollary 5.4.1 applied to the time-varying mean given by  $\Delta_t := 1.8(\exp\{t/300\}/(1 + \exp\{t/300\}) - 1/2)$ . In this particular example, we have that  $\tilde{\Delta}_t := \frac{1}{t} \sum_{i=1}^t \Delta_i$  changes from negative to positive at time 618, and yet our lower CS  $\tilde{L}_t^\Delta$  later detects this change at time 1654, at which point the weak null  $\tilde{\mathcal{H}}_0: \forall t, \tilde{\Delta}_t \leq 0$  can be rejected (see Section 5.B.9 for details regarding the composite hypothesis  $\tilde{\mathcal{H}}_0$  and how to test it).

**Practical implications.** The implications of Corollary 5.4.1 for the practitioner are threefold:

1. The CSs can be continuously monitored from the start of the A/B test and for an indefinite amount of time;
2. Inferences made from  $\tilde{L}_\tau^\Delta$  are valid at any stopping time  $\tau$ , regardless of why the test is stopped; and
3.  $\tilde{L}_t^\Delta$  adapts to non-stationarity: if the treatment differences  $\Delta_t$  drift over time,  $\tilde{L}_t^\Delta$  still forms an LPCS for  $\tilde{\Delta}_t$ . But if  $\Delta_1 = \Delta_2 = \dots = \Delta^*$  is constant, then  $\tilde{L}_t^\Delta$  forms an LPCS for  $\Delta^*$ .

## 5.5 Additional results & summary

Both NPPR and our proof techniques are general-purpose tools with several other implications for locally private statistical inference, including confidence sets via the Laplace mechanism, variance-adaptive inference, and sequential hypothesis testing. We briefly expand on these implications here, and leave their details to the appendix.

- §5.B.4: **Confidence sets via the Laplace mechanism.** We used NPPR as an extension of randomized response for arbitrary bounded data (rather than just binary), but of course the Laplace mechanism also handles bounded data. While NPPR enjoys advantages over Laplace as discussed in Section 5.2, it may still be of interest to derive confidence sets from data that are privatized via Laplace, given its ubiquity and simplicity. Section 5.B.4 presents new nonparametric CIs and CSs for population means under the Laplace mechanism.
- §5.B.5: **Variance-adaptive inference.** Notice that the CIs and CSs presented in Sec-

tion 5.3 were not variance-adaptive due to the fact that they relied on sub-Gaussianity of bounded random variables. However, this is not necessary, and we present other locally private *variance-adaptive* CIs and CSs in Section 5.B.5.

- §5.B.8: **Sequential hypothesis testing.** While the statistical procedures of this chapter have taken the form of CIs and CSs rather than hypothesis tests, there is a deep relationship between the two, and our results have analogues that could have been presented in the language of the latter. Section 5.B.8 articulates this relationship and presents explicit (sequential) tests.
- §5.B.10: **Adaptive online A/B testing.** Corollary 5.4.1 assumes a common propensity score  $\pi$  among all subjects for simplicity of exposition, but it is also possible to derive CSs for  $\tilde{\Delta}_t$  under an adaptive framework where propensity scores  $(\pi_t(X_t))_{t=1}^\infty$  can change over time in a data-dependent fashion, and be functions of some measured covariates  $(X_t)_{t=1}^\infty$ . The details of this more complex setup are left to Section 5.B.10.

Another followup problem that we do not explicitly address here but that can be solved using our techniques is locally private *variance* estimation. Notice that the variance  $\text{Var}(X) := \mathbb{E}(X^2) - (\mathbb{E}(X))^2$  is a function of two expectations,  $\mathbb{E}(X^2)$  and  $\mathbb{E}(X)$ . Since  $X^2$  is also  $[0, 1]$ -bounded if  $X$  is, we can use all of the techniques in this chapter to derive two separate  $(1 - \alpha/2, \varepsilon/2)$ -LPCIs (or LPCSs) to derive a  $(1 - \alpha, \varepsilon)$ -LPCI for  $\text{Var}(X)$ . Of course this requires collecting privatized views of both  $X^2$  and  $X$  separately. As a further generalization, a similar argument can be made for the construction of LPCIs for the covariance of  $X$  and  $Y$  since  $\text{cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$  (though here we would need to construct  $(1 - \alpha/3, \varepsilon/3)$ -LPCIs, etc.).

A limitation of the present chapter is that we have only discussed confidence sets for univariate parameters. Indeed, it is not immediately clear to us what is the right way to generalize NRRR to the multivariate case, or how to derive LPCIs and LPCSs for means of random *vectors* given such a generalization. This is an open direction for future work.

With the growing interest in protecting user privacy, an increasingly important addition to the statistician's toolbox are methods that can extract population information from privatized data. In this chapter, we derived nonparametric confidence intervals and time-uniform confidence sequences for population means from locally private data. We studied and used NRRR a nonparametric and sequentially interactive extension of Warner's randomized response for bounded data. The privatized output from NRRR can then be harnessed to produce confidence sets for the mean of the raw data distribution. Importantly, our confidence sets are sharp, some attaining optimal theoretical convergence rates and others simply having excellent empirical performance, not only making private nonparametric (sequential) inference possible, but practical. In future work, we aim to apply these general-purpose tools to changepoint detection, two-sample testing, and (conditional) independence testing.

## 5.A Proofs of main results

### 5.A.1 Prelude: filtrations, supermartingales, and Ville's inequality

By far the most common way to derive a CS is by constructing a nonnegative supermartingales and then applying Ville's maximal inequality to it. Indeed, all of the proofs for our CS and CI results employ this technique. However, in order to discuss supermartingales we must first review *filtrations*. A filtration  $\mathcal{F} \equiv (\mathcal{F})_{t=1}^\infty$  is a nondecreasing sequence of sigma-algebras  $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots$ , and a stochastic process  $(M)_{t=1}^\infty$  is said to be *adapted* to  $\mathcal{F}$  if  $M_t$  is  $\mathcal{F}_t$ -measurable for all  $t \in \mathbb{N}$ . On the other hand,  $(M)_{t=1}^\infty$  is said to be  $\mathcal{F}$ -predictable if each  $M_t$  is  $\mathcal{F}_{t-1}$ -measurable — informally “ $M_t$  depends on the past”.

For example, the canonical filtration  $\mathcal{X}$  generated by a sequence of random variables  $(X)_{t=1}^\infty$  is given by the sigma-algebra generated by  $X_1^t$ , i.e.  $\mathcal{X}_t := \sigma(X_1^t)$  for each  $t \in \{1, 2, \dots\}$ , and  $\mathcal{X}_0$  is the trivial sigma-algebra. A function  $M_t \equiv M(X_1^t)$  depending only on  $X_1^t$  forms a  $\mathcal{X}$ -adapted process, while  $(M_{t-1})_{t=1}^\infty$  is  $\mathcal{X}$ -predictable. Likewise, if we obtain a privatized view  $(Z)_{t=1}^\infty$  of  $(X)_{t=1}^\infty$  using some locally private mechanism, a different filtration  $\mathcal{Z}$  emerges, given by  $\mathcal{Z}_t := \sigma(Z_1^t)$ . Throughout our proofs,  $\mathcal{Z}$ -adapted and  $\mathcal{Z}$ -predictable processes will be central mathematical objects.

A process  $(M)_{t=1}^\infty$  adapted to  $\mathcal{F}$  is a *supermartingale* if

$$\mathbb{E}(M_t | \mathcal{F}_{t-1}) \leq M_{t-1} \text{ for each } t \geq 1. \quad (5.19)$$

If the above inequality is replaced by an equality, then  $(M)_{t=1}^\infty$  is a *martingale*. The methods in this chapter will involve derivations of (super)martingales which are nonnegative and begin at one — often referred to as “test (super)martingales” [237] or simply “nonnegative (super)martingales” (NMs or NSMs for martingales and supermartingales, respectively) [217, 124]. NSMs  $(M)_{t=1}^\infty$  satisfy the following powerful concentration inequality due to Ville [264]:

$$\mathbb{P}(\exists t \in \mathbb{N} : M_t \geq 1/\alpha) \leq \alpha. \quad (5.20)$$

In other words, they are unlikely to ever grow too large.

In the CS proofs that follow, we will focus on deriving processes  $(M_t(\mu))_{t=1}^\infty$  for any  $\mu \in [0, 1]$  such that when  $\mu$  is equal to the true mean of interest  $\mu^*$ , we have that  $M_t(\mu^*)$  forms a NSM. In this case, it turns out that the set of  $\mu$  such that  $M_t(\mu)$  is less than  $1/\alpha$  — i.e.  $C_t := \{\mu \in [0, 1] : M_t(\mu) < 1/\alpha\}$  — forms a  $(1 - \alpha)$ -CS for  $\mu^*$ . This is easy to see since  $\mu^* \notin C_t$  if and only if  $M_t(\mu^*) \geq 1/\alpha$ , and thus

$$\mathbb{P}(\exists t \in \mathbb{N} : \mu^* \notin C_t) = \mathbb{P}(\exists t \in \mathbb{N} : M_t(\mu^*) \geq 1/\alpha) \leq \alpha, \quad (5.21)$$

where the last inequality is precisely (5.20). The CS proofs that follow will make the exact processes  $(M_t(\mu))_{t=1}^\infty$  explicit.

### 5.A.2 Proof of Theorem 5.2.1

**Theorem 5.2.1** (NP RR satisfies LDP). *Suppose  $(Z)_{t=1}^{\infty}$  are generated according to NP RR. Then for each  $t \in \{1, 2, \dots\}$ ,  $Z_t$  is a conditionally  $\varepsilon_t$ -LDP view of  $X_t$  with*

$$\varepsilon_t := \log \left( 1 + \frac{(G_t + 1)r_t}{1 - r_t} \right). \quad (5.3)$$

*Proof.* We will prove the result for fixed  $r \in (0, 1)$ ,  $G \geq 1$  but it is straightforward to generalize the proof for  $r_t$  depending on  $Z_1^{t-1}$ . It suffices to verify that the likelihood ratio  $L(x, \tilde{x})$  is bounded above by  $\exp(\varepsilon)$  for any  $x, \tilde{x} \in [0, 1]$ . Writing out the likelihood ratio  $L(x, \tilde{x})$ , we have

$$L(x, \tilde{x}) := \frac{\frac{1-r}{G+1} + rG \cdot \{\mathbb{1}(Z = x^{\text{ceil}})(x - x^{\text{floor}}) + \mathbb{1}(Z = x^{\text{floor}})[1/G - (x - x^{\text{floor}})]\}}{\frac{1-r}{G+1} + rG \cdot \{\mathbb{1}(Z = \tilde{x}^{\text{ceil}})(\tilde{x} - \tilde{x}^{\text{floor}}) + \mathbb{1}(Z = \tilde{x}^{\text{floor}})[1/G - (\tilde{x} - \tilde{x}^{\text{floor}})]\}},$$

which is dominated by the counting measure. Notice that the numerator of  $L$  is maximized when  $x$  already lies in the discretized range, i.e.  $Z = x = x^{\text{ceil}} = x^{\text{floor}}$  so that the numerator becomes  $\frac{1-r}{G+1} + r$ , while the denominator is minimized when  $Z \neq \tilde{x}^{\text{ceil}}$  and  $Z \neq \tilde{x}^{\text{floor}}$  so that the denominator becomes  $\frac{1-r}{G+1}$ . Therefore, we have that with probability one,

$$L(x, \tilde{x}) \leq \frac{\frac{1-r}{G+1} + r}{\frac{1-r}{G+1}} = 1 + \frac{(G+1)r}{1-r},$$

and thus NP RR is  $\varepsilon$ -locally DP with  $\varepsilon := \log(1 + (G+1)r/(1-r))$ .  $\square$

### 5.A.3 Proof of Theorem 5.3.2

**Theorem 5.3.2** (NP RR-H). *Suppose  $(X)_{t=1}^n \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^n$ , and let  $(Z)_{t=1}^n \sim Q \in \mathcal{Q}_{\mu^*}^n$  be their privatized views via NP RR. Define the NP RR-adjusted sample mean*

$$\hat{\mu}_n := \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - (1 - r_i)/2)}{\frac{1}{n} \sum_{i=1}^n r_i}. \quad (5.8)$$

Then,

$$\bar{L}_n^H := \hat{\mu}_n - \sqrt{\frac{\log(1/\alpha)}{2n(\frac{1}{n} \sum_{i=1}^n r_i)^2}} \quad (5.9)$$

is a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCI for  $\mu^*$ .

*Proof.* The proof proceeds in two steps. First we note that  $\bar{L}_t^H$  forms a  $(1 - \alpha)$ -lower confidence sequence, and then instantiate this fact at the sample size  $n$ .

**Step 1.  $\bar{L}_t^H$  forms a  $(1 - \alpha)$ -lower CS.** This is exactly the statement of Theorem 5.3.3.

**Step 2.  $\dot{L}_n^H$  is a lower-**CI**.** By Step 1, we have that  $\bar{L}_t^H$  forms a  $(1 - \alpha)$ -lower CS, meaning

$$\mathbb{P}(\forall t \in \{1, \dots, n\}, \mu^* \geq \bar{L}_t^H) \geq 1 - \alpha.$$

Therefore,

$$\mathbb{P}\left(\mu^* \geq \max_{1 \leq t \leq n} \bar{L}_t^H\right) = \mathbb{P}(\mu^* \geq \dot{L}_n^H) \geq 1 - \alpha,$$

which completes the proof.  $\square$

#### 5.A.4 Proof of Theorem 5.3.3

**Theorem 5.3.3 (NPRR-H-CS).** Let  $(Z)_{t=1}^\infty 1 \sim Q$  for some  $Q \in \mathcal{Q}_{\mu^*}^\infty$ . Define the modified mean estimator under NRR:

$$\hat{\mu}_t(\lambda_1^t) := \frac{\sum_{i=1}^t \lambda_i \cdot (Z_i - (1 - r_i)/2)}{\sum_{i=1}^t r_i \lambda_i}, \quad (5.11)$$

and let  $(\lambda)_{t=1}^\infty 1$  be a real-valued sequence of tuning parameters (discussed in (5.32)). Then,

$$\bar{L}_t^H := \hat{\mu}_t(\lambda_1^t) - \frac{\log(1/\alpha) + \sum_{i=1}^t \lambda_i^2/8}{\sum_{i=1}^t r_i \lambda_i} \quad (5.12)$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCS for  $\mu^*$ .

*Proof.* The proof proceeds in two steps. First, we construct an NSM adapted to the private filtration  $\mathcal{Z} \equiv (\mathcal{Z})_{t=1}^\infty 0$ . Second and finally, we apply Ville's inequality to obtain a high-probability upper bound on the NSM, and show that this inequality results in the CS given in Theorem 5.3.3.

**Step 1.** Consider the nonnegative process starting at one given by

$$M_t(\mu^*) := \prod_{i=1}^t \exp\left\{\lambda_i(Z_i - \zeta_i(\mu^*)) - \lambda_i^2/8\right\}, \quad (5.22)$$

where  $(\lambda)_{t=1}^\infty 1$  is a real-valued sequence<sup>2</sup> and  $\zeta_t(\mu^*) := r_t \mu^* + (1 - r_t)/2$  as usual. We claim that  $(M_t(\mu^*))_{t=0}^\infty$  is a supermartingale, meaning  $\mathbb{E}(M_t(\mu^*) | \mathcal{Z}_{t-1}) \leq M_{t-1}(\mu^*)$ . Writing out the conditional expectation of  $M_t(\mu^*)$ , we have

$$\begin{aligned} & \mathbb{E}(M_t(\mu^*) | \mathcal{Z}_{t-1}) \\ &= \mathbb{E}\left(\prod_{i=1}^t \exp\left\{\lambda_i(Z_i - \zeta_i(\mu^*)) - \lambda_i^2/8\right\} \mid \mathcal{Z}_{t-1}\right) \end{aligned}$$

---

<sup>2</sup>The proof also works if  $(\lambda)_{t=1}^\infty 1$  is  $\mathcal{Z}$ -predictable but we omit this detail since we typically recommend using real-valued sequences anyway.

$$= \underbrace{\prod_{i=1}^{t-1} \exp \{ \lambda_i(Z_i - \zeta_i(\mu^*)) - \lambda_i^2/8 \}}_{M_{t-1}(\mu^*)} \cdot \underbrace{\mathbb{E} \left( \exp \{ \lambda_t(Z_t - \zeta_t(\mu^*)) - \lambda_t^2/8 \} \mid \mathcal{Z}_{t-1} \right)}_{(\dagger)},$$

since  $M_{t-1}(\mu^*)$  is  $\mathcal{Z}_{t-1}$ -measurable, and thus it can be written outside of the conditional expectation. It now suffices to show that  $(\dagger) \leq 1$ . To this end, note that  $Z_t$  is a  $[0, 1]$ -bounded random variable with conditional mean  $\mathbb{E}(Z_t \mid \mathcal{Z}_{t-1}) = \zeta_t(\mu^*)$  by design of NPPR (Algorithm 5.2). Since bounded random variables are sub-Gaussian [120], we have that

$$\mathbb{E}(\lambda_t(Z_t - \zeta_t(\mu^*)) \mid \mathcal{Z}_{t-1}) \leq \exp \{ \lambda_t^2/8 \},$$

and hence  $(\dagger) \leq 1$ . Therefore,  $(M_t(\mu^*))_{t=0}^\infty$  is a  $\mathcal{Q}_{\mu^*}^\infty$ -NSM.

**Step 2.** By Ville's inequality for NSMs [264], we have that

$$\mathbb{P}(\exists t : M_t(\mu^*) \geq 1/\alpha) \leq \alpha.$$

In other words, we have that  $M_t(\mu^*) < 1/\alpha$  for all  $t \in \mathbb{N}$  with probability at least  $1 - \alpha$ . Using some algebra to rewrite the inequality  $M_t(\mu^*) < 1/\alpha$ , we have

$$\begin{aligned} M_t(\mu^*) < 1/\alpha &\iff \prod_{i=1}^t \exp \{ \lambda_i(Z_i - \zeta_i(\mu^*)) - \lambda_i^2/8 \} < \frac{1}{\alpha} \\ &\iff \sum_{i=1}^t [\lambda_i(Z_i - \zeta_i(\mu^*)) - \lambda_i^2/8] < \log(1/\alpha) \\ &\iff \sum_{i=1}^t \lambda_i Z_i - \mu^* \sum_{i=1}^t \lambda_i r_i - \sum_{i=1}^t \lambda_i \cdot (1 - r_i)/2 - \sum_{i=1}^t \lambda_i^2/8 < \log(1/\alpha) \\ &\iff \mu^* > \underbrace{\frac{\sum_{i=1}^t \lambda_i \cdot (Z_i - (1 - r_i)/2)}{\sum_{i=1}^t r_i \lambda_i}}_{\hat{\mu}_t(\lambda_1^t)} - \underbrace{\frac{\log(1/\alpha) + \sum_{i=1}^t \lambda_i^2/8}{\sum_{i=1}^t r_i \lambda_i}}_{\bar{B}_t(\lambda_1^t)} \end{aligned}$$

Therefore,  $\bar{L}_t := \hat{\mu}_t(\lambda_1^t) - \bar{B}_t(\lambda_1^t)$  forms a lower  $(1 - \alpha)$ -CS for  $\mu^*$ . The upper CS  $\bar{U}_t := \hat{\mu}_t(\lambda_1^t) + \bar{B}_t(\lambda_1^t)$  can be derived by applying the above proof to  $(-Z)_{t=1}^\infty$  and their conditional means  $(-\zeta_i(\mu^*))_{t=1}^\infty$ . This completes the proof  $\square$

### 5.A.5 Proof of Theorem 5.3.4

**Theorem 5.3.4** (Confidence sequences for time-varying means). *Suppose  $X_1, X_2, \dots$  are independent  $[0, 1]$ -bounded random variables with individual means  $\mathbb{E}X_t = \mu_t^*$  for each  $t$ , and*

let  $Z_1, Z_2 \dots$  be their privatized views according to NPPR without sequential interactivity. Define

$$\hat{\mu}_t := \frac{\sum_{i=1}^t (Z_i - (1-r)/2)}{tr}, \quad (5.14)$$

$$\text{and } \tilde{B}_t^\pm := \sqrt{\frac{t\beta^2 + 1}{2(tr\beta)^2} \log \left( \frac{\sqrt{t\beta^2 + 1}}{\alpha} \right)}, \quad (5.15)$$

for any  $\beta > 0$ . Then,  $\tilde{C}_t^\pm := (\hat{\mu}_t \pm \tilde{B}_t^\pm)$  forms a two-sided  $(1 - \alpha, \varepsilon)$ -LPCS for  $\tilde{\mu}_t^*$ , where  $\varepsilon = \log(1 + \frac{2r}{1-r})$ .

*Proof.* The proof proceeds in three steps. First, we derive a sub-Gaussian NSM indexed by a parameter  $\lambda \in \mathbb{R}$ . Second, we mix this NSM over  $\lambda$  using the density of a Gaussian distribution, and justify why the resulting process is also an NSM. Third and finally, we apply Ville's inequality and invert the NSM to obtain  $(\tilde{C}_t^\pm)_{t=1}^\infty 1$ .

**Step 1: Constructing the  $\lambda$ -indexed NSM.** Let  $(X)_{t=1}^\infty 1$  be independent  $[0, 1]$ -bounded random variables with individual means given by  $\mathbb{E}X_t = \mu_t^*$ , and let  $(Z)_{t=1}^\infty 1$  be the NPPR-induced private views of  $(X)_{t=1}^\infty 1$ . Define  $\zeta(\mu) := r\mu + (1-r)/2$  for any  $\mu \in [0, 1]$ , and  $r \in (0, 1]$ . Let  $\lambda \in \mathbb{R}$  and consider the process,

$$M_t(\lambda) := \prod_{i=1}^t \exp \left\{ \lambda(Z_i - \zeta(\mu_i^*)) - \lambda^2/8 \right\}, \quad (5.23)$$

with  $M_0(\lambda) \equiv 0$ . We claim that (5.23) forms an NSM with respect to the private filtration  $\mathcal{Z}$ . The proof technique is nearly identical to that of Theorem 5.3.3 but with changing means and  $\lambda = \lambda_1 = \lambda_2 = \dots \in \mathbb{R}$ . Indeed,  $M_t(\lambda)$  is nonnegative with initial value one by construction, so it remains to show that  $(M_t(\lambda))_{t=0}^\infty$  is a supermartingale. That is, we need to show that for every  $t$ , we have  $\mathbb{E}(M_t(\lambda) \mid \mathcal{Z}_{t-1}) \leq M_{t-1}(\lambda)$ . Writing out the conditional expectation of  $M_t(\lambda)$ , we have

$$\begin{aligned} \mathbb{E}(M_t(\lambda) \mid \mathcal{Z}_{t-1}) &= \mathbb{E} \left( \prod_{i=1}^t \exp \left\{ \lambda(Z_i - \zeta(\mu_i^*)) - \lambda^2/8 \right\} \mid Z_1^{t-1} \right) \\ &= \underbrace{\prod_{i=1}^{t-1} \exp \left\{ \lambda(Z_i - \zeta(\mu_i^*)) - \lambda^2/8 \right\}}_{M_{t-1}(\lambda)} \cdot \mathbb{E} \left( \exp \left\{ \lambda(Z_t - \zeta(\mu_t^*)) - \lambda^2/8 \right\} \mid Z_1^{t-1} \right) \\ &= M_{t-1}(\lambda) \cdot \underbrace{\mathbb{E} \left( \exp \left\{ \lambda(Z_t - \zeta(\mu_t^*)) - \lambda^2/8 \right\} \right)}_{(\dagger)}, \end{aligned}$$

where the last inequality follows by independence of  $(Z)_{t=1}^\infty 1$ , and hence the conditional expectation becomes a marginal expectation. Therefore, it now suffices to show that  $(\dagger) \leq 1$ . Indeed,  $Z_t$  is a  $[0, 1]$ -bounded, mean- $\zeta(\mu_t^*)$  random variable. By Hoeffding's sub-Gaussian inequality

for bounded random variables [120], we have that  $\mathbb{E}[\exp\{\lambda(Z_t - \zeta(\mu_t^*))\}] \leq \exp\{\lambda^2/8\}$ , and thus

$$(\dagger) = \mathbb{E}[\exp\{\lambda(Z_t - \zeta(\mu_t^*))\}] \cdot \exp\{-\lambda^2/8\} \leq 1.$$

It follows that  $(M_t(\lambda))_{t=0}^\infty$  is an NSM.

**Step 2.** Let us now construct a sub-Gaussian mixture NSM. Note that the mixture of an NSM with respect to a probability distribution is itself an NSM [217, 124] – a straightforward consequence of Fubini’s theorem. Concretely, let  $f_{\rho^2}(\lambda)$  be the probability density function of a mean-zero Gaussian random variable with variance  $\rho^2$ ,

$$f_{\rho^2}(\lambda) := \frac{1}{\sqrt{2\pi\rho^2}} \exp\left\{\frac{-\lambda^2}{2\rho^2}\right\}.$$

Then, since mixtures of NSMs are themselves NSMs, the process  $(M)_{t=1}^\infty$  given by

$$M_t := \int_{\lambda \in \mathbb{R}} M_t(\lambda) f_{\rho^2}(\lambda) d\lambda \tag{5.24}$$

is an NSM. We will now find a closed-form expression for  $M_t$ . To ease notation, define the partial sum  $S_t^* := \sum_{i=1}^t (Z_i - \zeta(\mu_i^*))$ . Writing out the definition of  $M_t$ , we have

$$\begin{aligned} M_t &:= \int_{\lambda \in \mathbb{R}} \prod_{i=1}^t \exp\{\lambda(Z_i - \zeta(\mu_i^*)) - \lambda^2/8\} f_{\rho^2}(\lambda) d\lambda \\ &= \int_{\lambda} \exp\left\{\lambda \underbrace{\sum_{i=1}^t (Z_i - \zeta(\mu_i^*))}_{S_t^*} - t\lambda^2/8\right\} f_{\rho^2}(\lambda) d\lambda \\ &= \int_{\lambda} \exp\{\lambda S_t^* - t\lambda^2/8\} \frac{1}{\sqrt{2\pi\rho^2}} \exp\left\{\frac{-\lambda^2}{2\rho^2}\right\} d\lambda \\ &= \frac{1}{\sqrt{2\pi\rho^2}} \int_{\lambda} \exp\{\lambda S_t^* - t\lambda^2/8\} \exp\left\{\frac{-\lambda^2}{2\rho^2}\right\} d\lambda \\ &= \frac{1}{\sqrt{2\pi\rho^2}} \int_{\lambda} \exp\left\{\lambda S_t^* - \frac{\lambda^2(t\rho^2/4 + 1)}{2\rho^2}\right\} d\lambda \\ &= \frac{1}{\sqrt{2\pi\rho^2}} \int_{\lambda} \exp\left\{\frac{-\lambda^2(t\rho^2/4 + 1) + 2\lambda\rho^2 S_t^*}{2\rho^2}\right\} d\lambda \\ &= \frac{1}{\sqrt{2\pi\rho^2}} \int_{\lambda} \exp\left\{\frac{-a(\lambda^2 - \frac{b}{a}2\lambda)}{2\rho^2}\right\} d\lambda, \end{aligned}$$

where we have set  $a := t\rho^2/4 + 1$  and  $b := \rho^2 S_t^*$ . Completing the square in the exponent, we have that

$$\begin{aligned} \exp \left\{ \frac{-\lambda^2 - 2\lambda \frac{b}{a} + \left(\frac{b}{a}\right)^2 - \left(\frac{b}{a}\right)^2}{2\rho^2/a} \right\} &= \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} + \frac{a(b/a)^2}{2\rho^2} \right\} \\ &= \underbrace{\exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} \right\}}_{(*)} \exp \left\{ \frac{b^2}{2a\rho^2} \right\}. \end{aligned}$$

Now notice that  $(*)$  is proportional to the density of a Gaussian random variable with mean  $b/a$  and variance  $\rho^2/a$ . Plugging the above back into the integral and multiplying the entire quantity by  $a^{-1/2}/a^{-1/2}$ , we obtain the closed-form expression of the mixture NSM,

$$\begin{aligned} M_t &:= \underbrace{\frac{1}{\sqrt{2\pi\rho^2/a}} \int_{\lambda \in \mathbb{R}} \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} \right\} d\lambda}_{=1} \frac{\exp \left\{ \frac{b^2}{2a\rho^2} \right\}}{\sqrt{a}} \\ &= \frac{1}{\sqrt{t\rho^2/4 + 1}} \exp \left\{ \frac{\rho^2(S_t^*)^2}{2(t\rho^2/4 + 1)} \right\}. \end{aligned} \tag{5.25}$$

**Step 3.** Now that we have computed the mixture NSM  $(M)_{t=1}^\infty 0$ , we are ready to apply Ville's inequality and invert the process. Since  $(M)_{t=1}^\infty 0$  is an NSM, we have by Ville's inequality [264],

$$\mathbb{P}(\exists t : M_t \geq 1/\alpha) \leq \alpha \quad \text{or equivalently, } \mathbb{P}(\forall t, M_t < 1/\alpha) \geq 1 - \alpha.$$

Therefore, with probability at least  $(1 - \alpha)$ , we have that for all  $t \in \{1, 2, \dots\}$ ,

$$\begin{aligned} M_t < 1/\alpha &\iff \frac{1}{\sqrt{t\rho^2/4 + 1}} \exp \left\{ \frac{\rho^2(S_t^*)^2}{2(t\rho^2/4 + 1)} \right\} < 1/\alpha \\ &\iff \frac{\rho^2(S_t^*)^2}{2(t\rho^2/4 + 1)} - \log \left( \sqrt{t\rho^2/4 + 1} \right) < \log(1/\alpha) \\ &\iff \frac{\rho^2(S_t^*)^2}{2(t\rho^2/4 + 1)} < \log \left( \frac{\sqrt{t\rho^2/4 + 1}}{\alpha} \right) \\ &\iff (S_t^*)^2 < \frac{2(t\rho^2/4 + 1)}{\rho^2} \log \left( \frac{\sqrt{t\rho^2/4 + 1}}{\alpha} \right) \\ &\iff \frac{(S_t^*)^2}{t^2r^2} < \underbrace{\frac{2(t(\rho/2)^2 + 1)}{(tr\rho)^2} \log \left( \frac{\sqrt{t(\rho/2)^2 + 1}}{\alpha} \right)}_{(**)}. \end{aligned}$$

Set  $\beta := \rho/2$  and notice that  $(\star\star) = (\tilde{B}_t^\pm)^2$  where  $\tilde{B}_t^\pm$  is the boundary given by (5.15) in the statement of Theorem 5.3.4. Also recall from Theorem 5.3.4 the private estimator  $\hat{\mu}_t := \frac{1}{tr} \sum_{i=1}^t [Z_i - (1-r)/2]$  and the quantity we wish to capture – the moving average of *population means*  $\tilde{\mu}_t^\star := \frac{1}{t} \sum_{i=1}^t \mu_i^\star$ , where  $\mu_i^\star = \mathbb{E}X_i$ . Putting these together with the above high-probability bound, we have that with probability  $\geq (1-\alpha)$ , for all  $t$ ,

$$\begin{aligned}
M_t < 1/\alpha &\iff \frac{(S_t^\star)^2}{t^2 r^2} < (\tilde{B}_t^\pm)^2 \\
&\iff -\tilde{B}_t^\pm < \frac{S_t^\star}{tr} < \tilde{B}_t^\pm \\
&\iff -\tilde{B}_t^\pm < \frac{\sum_{i=1}^t [Z_i - \zeta(\mu_i^\star)]}{tr} < \tilde{B}_t^\pm \\
&\iff -\tilde{B}_t^\pm < \frac{\sum_{i=1}^t [Z_i - (r\mu_i^\star + (1-r)/2)]}{tr} < \tilde{B}_t^\pm \\
&\iff -\tilde{B}_t^\pm < \frac{\sum_{i=1}^t [Z_i - (1-r)/2]}{tr} - \frac{\sum_{i=1}^t \mu_i^\star}{tr} < \tilde{B}_t^\pm \\
&\iff -\frac{\sum_{i=1}^t [Z_i - (1-r)/2]}{tr} - \tilde{B}_t^\pm < -\frac{\sum_{i=1}^t \mu_i^\star}{t} < -\frac{\sum_{i=1}^t [Z_i - (1-r)/2]}{tr} + \tilde{B}_t^\pm \\
&\iff -\hat{\mu}_t - \tilde{B}_t^\pm < -\tilde{\mu}_t^\star < -\hat{\mu}_t + \tilde{B}_t^\pm \\
&\iff \hat{\mu}_t - \tilde{B}_t^\pm < \tilde{\mu}_t^\star < \hat{\mu}_t + \tilde{B}_t^\pm.
\end{aligned}$$

In summary, we have that  $\tilde{C}_t^\pm := (\hat{\mu}_t \pm \tilde{B}_t^\pm)$  forms a  $(1-\alpha)$ -CS for the time-varying parameter  $\tilde{\mu}_t^\star$ , meaning

$$\mathbb{P}\left(\forall t, \tilde{\mu}_t^\star \in \tilde{C}_t^\pm\right) \geq 1 - \alpha.$$

This completes the proof. □

## 5.B Additional results

### 5.B.1 Confidence sets under randomized response

Since NPPR is a strict generalization for bounded random variables, it can be used to construct confidence sets for the mean of Bernoulli random variables which are privatized via randomized response (RR). The following corollary provides a Hoeffding-type CI for the mean under RR.

**Corollary 5.B.1** (Locally private Hoeffding inequality under RR). *Let  $X_1, \dots, X_n \sim \text{Bernoulli}(p^\star)$ , and let  $Z_1, \dots, Z_n$  be their privatized views according to RR for some fixed  $r \in (0, 1]$ . Then,*

$$\dot{L}_n^H := \frac{\sum_{i=1}^n (Z_i - (1-r)/2)}{nr} - \sqrt{\frac{\log(1/\alpha)}{2nr^2}} \tag{5.26}$$

*is a  $(1-\alpha, \varepsilon)$ -lower LPCI for  $p^\star$ , where  $\varepsilon = \log(1 + 2r/(1-r))$ .*

Corollary 5.B.1 is a special case of Theorem 5.3.2. Notice that in the non-private setting when  $r = 1$ , Corollary 5.B.1 recovers Hoeffding's inequality exactly [120].

## 5.B.2 Confidence sets for sample means

While we primarily focused on deriving CIs and CSs for population means, our techniques can also be applied to the construction of CIs and CSs for the *sample mean*. Indeed, in the non-interactive case, the proof of Theorem 5.3.2 can be modified so that the bound (5.9) is a lower  $(1 - \alpha)$ -CI for the sample mean  $\mu^* := \frac{1}{n} \sum_{i=1}^n x_i$ , recovering essentially the same result as Ding et al. [85, Theorem 1].<sup>3</sup> However, implicit in our results are also time-uniform CSs for the *running sample mean so far*. Concretely, we have the following corollary.

**Corollary 5.B.2** (A confidence sequence for the running sample mean). *Let  $(x_t)_{t=1}^\infty$  be a sequence of  $[0, 1]$ -bounded numbers and let  $(Z_t)_{t=1}^\infty$  be their privatized views according to NPPR without sequential interactivity. Then, the same bound as given in Theorem 5.3.4, i.e.*

$$\tilde{C}_t := \left( \frac{\sum_{i=1}^t (Z_i - (1-r)/2)}{tr} \pm \sqrt{\frac{t\beta^2 + 1}{2(tr\beta)^2} \log \left( \frac{\sqrt{t\beta^2 + 1}}{\alpha} \right)} \right) \quad (5.27)$$

forms a  $(1 - \alpha, \varepsilon)$ -LPCS for the running sample mean  $\tilde{\mu}_t^* := \frac{1}{t} \sum_{i=1}^t x_i$ , i.e.

$$\mathbb{P} \left( \forall t, \tilde{\mu}_t^* \in \tilde{C}_t \right) \geq 1 - \alpha. \quad (5.28)$$

The above corollary is an immediate consequence of Theorem 5.3.4 instantiated for random variables  $(X_t)_{t=1}^\infty$  with degenerate distributions. (and hence  $\mathbb{E}X_t = X_t = \mu^*$ ).

Corollary 5.B.2 also sheds some light on how the two estimands (population vs sample means) are related but fundamentally different. Both the (a) stochastic setting with data  $X_1, X_2, \dots$  that have a constant mean  $\mathbb{E}X_1 = \mu^* \in [0, 1]$  and (b) nonstochastic setting with deterministic data  $x_1, x_2, \dots$  are special cases of the stochastic setting with data that have time-varying means  $\mathbb{E}X_t = \mu_t$  for  $t \geq 1$ . Setting (a) is recovered by assuming that  $\mu_1 = \mu_2 = \dots = \mu^*$ , while setting (b) is recovered by assuming  $(X_t)_{t=1}^\infty$  have degenerate distributions (or by conditioning on them). Clearly, neither is a special case of the other, and hence we cannot expect CIs/CSs for one to work for the other in general (though in this case,  $(\tilde{C}_t)_{t=1}^\infty$  works for both).

## 5.B.3 Why one should set $G = 1$ for Hoeffding-type methods

In Section 5.3, we recommended setting  $G$  to the smallest possible value of 1 because Hoeffding-type bounds cannot benefit from larger values. We will now justify mathematically where this recommendation came from.

---

<sup>3</sup>Technically, a one-sided CI is more general than Ding et al. [85]'s since theirs is a two-sided CI that we recover after taking a union bound over lower and upper CIs, but the lower CI is also implicit in their proof.

Suppose  $(X)_{t=1}^n \sim P$  for some  $\mathcal{P}_{\mu^*}^n$  where we have chosen  $r \in (0, 1]$  and an integer  $G \geq 1$  to satisfy  $\varepsilon$ -LDP with

$$\varepsilon := \log \left( 1 + \frac{(G+1)r}{1-r} \right). \quad (5.29)$$

Recall the NPPR-Hoeffding lower LPCI given (5.9),

$$\dot{L}_n^H := \frac{\sum_{i=1}^n (Z_i - (1-r)/2)}{nr} - \underbrace{\sqrt{\frac{\log(1/\alpha)}{2nr^2}}}_{\dot{B}_n^H}, \quad (5.30)$$

and take particular notice of  $\dot{B}_n^H$ , the ‘‘boundary’’. Making this bound as sharp as possible amounts to minimizing  $\dot{B}_n^H$ , which is clearly when  $r = 1$  – the non-private case – but what if we want to minimize  $\dot{B}_n^H$  subject to  $\varepsilon$ -LDP? Given the relationship between  $\varepsilon$ ,  $r$ , and  $G$ , we have that  $r$  can be written as

$$r := \frac{\exp\{\varepsilon\} - 1}{\exp\{\varepsilon\} + G}.$$

Plugging this into  $\dot{B}_n^H$ , we have

$$\dot{B}_n^H := \sqrt{\frac{\log(1/\alpha)}{2n \left( \frac{\exp\{\varepsilon\} - 1}{\exp\{\varepsilon\} + G} \right)^2}} = \left( \frac{\exp\{\varepsilon\} + G}{\exp\{\varepsilon\} - 1} \right) \cdot \sqrt{\frac{\log(1/\alpha)}{2n}},$$

which is a strictly increasing function of  $G$ . It follows that  $G$  should be set to the minimal value of 1 to make  $\dot{L}_n^H$  as sharp as possible.

#### 5.B.4 Confidence sets under the sequentially interactive Laplace mechanism

**Proposition 5.B.1** (Lap-H-CS). *Suppose  $(X)_{t=1}^\infty \sim P$  for some  $\mathcal{P}_{\mu^*}^\infty$  and let  $(Z)_{t=1}^\infty$  be their privatized views according to Algorithm 5.1. Let  $\psi_t^L(\lambda) := -\log(1 - \lambda^2/\varepsilon_t^2)$  be the (conditional) cumulant generating function of a mean-zero Laplace random variable with scale  $1/\varepsilon_t$ . Let  $(\lambda)_{t=1}^\infty$  be a sequence of random variables such that  $\lambda_t$  depends on  $Z_1^{t-1}$  – formally  $\sigma(Z_1^{t-1})$ -measurable – and  $[0, \varepsilon_t]$ -valued. Then,*

$$\bar{L}_t^L := \frac{\sum_{i=1}^t \lambda_i Z_i}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\alpha) + \sum_{i=1}^t (\lambda_i^2/8 + \psi_i^L(\lambda_i))}{\sum_{i=1}^t \lambda_i} \quad (5.31)$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_{t=1}^\infty)$ -LPCS for  $\mu^*$ .

To obtain sharp CSs for  $\mu^*$ , we recommend setting

$$\lambda_t := \sqrt{\frac{\log(1/\alpha)}{\sum_{i=1}^t (1/8 + 1/\varepsilon_i^2) \log(t+1)}} \wedge c \cdot \varepsilon_t, \quad (5.32)$$

for some prespecified truncation scale  $c \in (0, 1)$ . We choose  $\lambda_t$  as scaling like  $1/\sqrt{t \log t}$  so that the CS  $\bar{L}_t^L$  is  $O(\sqrt{\log t/t})$  up to log log factors (see Table 3.1 from Chapter 3 for more

details).<sup>4</sup> The constants provided in (5.32) arise from approximating  $\psi^{\mathcal{L}}(\lambda)$  by  $\lambda^2/\varepsilon^2$  for  $\lambda$  near 0 — an approximation that can be justified by a simple application of L'Hopital's rule — and attempting to minimize the CI width.

Similar to Section 5.3, we can choose  $(\lambda)_{t=1}^\infty 1$  so that  $\bar{L}_t^{\mathcal{L}}$  is tight for a fixed sample size  $n$ . Indeed, we have the following Laplace-Hoeffding CIs for  $\mu^*$ .

**Corollary 5.B.3 (Lap-H).** *Given the same assumptions as Proposition 5.B.1 for a fixed sample size  $n$ , define*

$$\lambda_{t,n} := \sqrt{\frac{\log(1/\alpha)}{\frac{n}{t} \sum_{i=1}^t (1/8 + 1/\varepsilon_i^2)}} \wedge c \cdot \varepsilon_t, \quad (5.33)$$

and plug it into  $\bar{L}_t^{\mathcal{L}}$  as given above. Then,

$$\dot{L}_n^{\mathcal{L}} := \max_{1 \leq t \leq n} \bar{L}_t$$

is a  $(1 - \alpha, (\varepsilon_t)_t)$ -lower LPCI for  $\mu^*$ .

The proof of Proposition 5.B.1 (and hence Corollary 5.B.3) can be found in Section 5.C.1. Note that any prespecified value of  $c \in (0, 1)$  yields valid CSs and CIs, we find that smaller values (e.g. near 0.1) yield tighter intervals, and we set  $c = 0.1$  in our simulations (Figures 5.4 and 5.5).

## 5.B.5 Variance-adaptive confidence intervals and sequences

### 5.B.5.1 Variance-adaptive confidence intervals

Notice that if  $G_t = 1$  for each  $t$ , then regardless of how low-variance  $(X)_{t=1}^n$  are, the observations that are ultimately used for confidence set construction are still Bernoulli. In other words, it does not matter whether  $(X)_{t=1}^n$  are Bernoulli(1/2), Uniform[0, 1], or Beta(100, 100) — with variances of roughly 0.25, 0.083, and 0.0012, respectively — the privatized observations  $(Z)_{t=1}^n$  are all Bernoulli(1/2) with a maximal variance 0.25. Unfortunately, this means that variance-adaptive techniques cannot be used to derive tighter CIs from  $(Z)_{t=1}^n$  directly. The story changes, however, when  $G_t \geq 2$ . Concretely, for the same value of  $r_t$ , setting  $G_t$  to be very large does not change the conditional mean of  $Z_t$  but it can substantially lower its conditional variance (e.g. if  $X_t$  has a continuous distribution, such as Beta( $\alpha, \beta$ )). Of course, given the fact that NPPR satisfies  $\varepsilon_t$ -LDP with  $\varepsilon_t = \log\left(1 + \frac{(G_t+1)r_t}{1-r_t}\right)$ , there are privacy implications to increasing  $G_t$ , and hence there is a tradeoff that must be carefully navigated when choosing  $(r_t, G_t)$  to satisfy  $\varepsilon_t$  when attempting to derive variance-adaptive CIs. We will leave that delicate discussion for later — for now, it is just important to keep in mind that larger  $G_t$  can lower the variance of  $(Z)_{t=1}^n$ , and our goal will be to exploit this fact for the sake of tighter CIs.

We will proceed by turning to the literature on nonasymptotic CIs for bounded random vari-

---

<sup>4</sup>This specific rate assumes  $\varepsilon_t = \varepsilon \in (0, 1)$  for each  $t$ .

ables, focusing on the (super)martingale-based CIs of Chapter 3 and adapting those techniques to the locally private setting.

**Product “betting” martingales.** Beginning with the former, we follow the discussions in Section 3.3.3 and set

$$\begin{aligned}\lambda_{t,n}(\mu) &:= \sqrt{\frac{2 \log(1/\alpha)}{\hat{\gamma}_{t-1}^2 n}} \wedge \frac{c}{\zeta_t(\mu)}, \text{ where} \\ \hat{\gamma}_t^2 &:= \frac{1/4 + \sum_{i=1}^t (Z_i - \hat{\zeta}_i)^2}{t+1}, \quad \hat{\zeta}_t := \frac{1/2 + \sum_{i=1}^t Z_i}{t+1},\end{aligned}\tag{5.34}$$

and  $c \in (0, 1)$  is some prespecified truncation scale (e.g. 1/2 or 3/4). Given the above, we have the following variance-adaptive CI for  $\mu^*$  under NPPR.

**Theorem 5.B.1** (NPPR-hedged). *Suppose  $(X)_{t=1}^n \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^n$  and let  $(Z)_{t=1}^n \sim Q$  be their NPPR-privatized views where  $Q \in \mathcal{Q}_{\mu^*}^n$ . Define*

$$\mathcal{K}_{t,n}(\mu) := \prod_{i=1}^t [1 + \lambda_{i,n}(\mu) \cdot (Z_i - \zeta_i(\mu))] \tag{5.35}$$

with  $\lambda_{t,n}(\mu)$  given by (5.34). Then,  $\mathcal{K}_{t,n}(\mu)$  is a nonincreasing function of  $\mu \in [0, 1]$ , and  $\mathcal{K}_{t,n}(\mu^*)$  forms a  $\mathcal{Q}_{\mu^*}^n$ -NM. Consequently,

$$\dot{L}_n := \max_{1 \leq t \leq n} \inf \{\mu \in [0, 1] : \mathcal{K}_{t,n}(\mu) < 1/\alpha\} \tag{5.36}$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCI for  $\mu^*$ , meaning  $\mathbb{P}(\mu^* \geq \dot{L}_n) \geq 1 - \alpha$ .

The proof in Section 5.C.3 follows a similar technique to that of Theorem 5.B.2. As is apparent in the proof,  $\mathcal{K}_{t,n}(\mu^*)$  forms a  $\mathcal{Q}_{\mu^*}^n$ -NM regardless of how  $\lambda_{t,n}(\mu)$  is chosen, in which case the resulting  $\dot{L}_n$  would still be a bona fide lower confidence bound. However, the choice of  $\lambda_{t,n}(\mu)$  given in (5.34) provides excellent empirical performance for the reasons discussed in Chapter 3 and guarantees that  $\dot{L}_n$  is an interval (rather than a union of disjoint sets, for example). We find that Theorem 5.B.1 has the best empirical performance out of the private CIs in this chapter (see Figure 5.4). In our simulations (Figure 5.4), we set  $c = 0.8$ , and  $(r, G)$  were chosen using the technique outlined in Section 5.B.6.

**Empirical Bernstein supermartingales.** While Theorem 5.B.1 improves on Theorem 5.3.2 in terms of variance-adaptivity, the resulting bounds given in (5.36) are *implicit*, and hence require numerical methods (e.g. root-finding algorithms) to compute the downstream CI. The numerical operations required are both computationally efficient and straightforward to implement in code, but closed-form bounds may nevertheless be preferable for the sake of simplicity. Empirical Bernstein CIs occupy a middle ground between the Hoeffding-style CIs of Theorem 5.3.3 and the implicit CIs of Theorem 5.B.1 by being both closed-form and

variance-adaptive. To this end, consider the following tuning parameters which are similar (but not identical) to (5.34):

$$\begin{aligned}\lambda_{t,n}^{\text{EB}}(\mu) &:= \sqrt{\frac{2 \log(1/\alpha)}{\hat{\gamma}_{t-1}^2 n}} \wedge c, \text{ where} \\ \hat{\gamma}_t^2 &:= \frac{1/4 + \sum_{i=1}^t (Z_i - \hat{\zeta}_i)^2}{t+1}, \quad \hat{\zeta}_t := \frac{1/2 + \sum_{i=1}^t Z_i}{t+1},\end{aligned}\tag{5.37}$$

and  $c \in (0, 1)$ . Then, we have the following variance-adaptive empirical Bernstein CIs under NPPR.

**Proposition 5.B.2 (NPPR-EB).** *Under the same assumptions as Theorem 5.B.1, let  $(\lambda_{t,n}^{\text{EB}})_{t=1}^n$  be the  $[0, 1]$ -valued  $\mathcal{Z}$ -predictable sequence given in (5.37) and define*

$$\begin{aligned}\hat{\mu}_t(\lambda_1^t) &:= \frac{\sum_{i=1}^t \lambda_i \cdot (Z_i - (1 - r_i)/2)}{\sum_{i=1}^t r_i \lambda_i}, \\ \bar{B}_t^{\text{EB}}(\lambda_1^t) &:= \frac{\log(1/\alpha) + \sum_{i=1}^t 4(Z_i - \hat{\zeta}_{i-1})^2 \psi_E(\lambda_i)}{\sum_{i=1}^t r_i \lambda_i}.\end{aligned}$$

where  $\psi_E(\lambda) := (-\log(1 - \lambda) - \lambda)/4$ . Then,

$$\dot{L}_t^{\text{EB}} := \max_{1 \leq t \leq n} \{\hat{\mu}_t - \bar{B}_t^{\text{EB}}\}\tag{5.38}$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCI for  $\mu^*$ , meaning  $\mathbb{P}(\mu^* \geq \dot{L}_t^{\text{EB}}) \geq 1 - \alpha$ .

Proposition 5.B.2 is a corollary of Proposition 5.B.3 whose proof can be found in Section 5.C.5. Similar to Theorem 5.B.1, one can use any  $(\lambda_{t,n}^{\text{EB}})_{t=1}^n$  as long as they are predictable and  $[0, 1]$ -valued, but we presented (5.37) as it tends to exhibit good empirical performance for the reasons discussed in Chapter 3. As previously alluded to, the essential difference between Theorem 5.B.1 and Proposition 5.B.2 is that the former tends to be tighter in practice, while the latter has the advantage of having a computationally and analytically simple closed-form expression. In principle, the proof and techniques of Theorem 5.B.1 and Proposition 5.B.2 may be adapted to many other variance-adaptive CIs for bounded random variables, including Bentkus [28], Audibert et al. [12], Maurer and Pontil [187], Orabona and Jun [195], or other bounds in Chapter 3, but we presented the aforementioned two for simplicity and illustration. Let us now turn our attention to a more challenging but related problem of constructing time-uniform *confidence sequences* instead of fixed-time *confidence intervals*.

### 5.B.5.2 Variance-adaptive time-uniform confidence sequences

In Section 5.3, we presented Hoeffding-type CSs for  $\mu^*$  under NPPR. As discussed in Section 5.B.5.1, Hoeffding-type inequalities are not variance-adaptive. In this section, we will derive a simple-to-compute, variance-adaptive CS at the expense of a closed-form expression. Adapting the so-called “grid Kelly capital process” (GridKelly) of Chapter 3 to the locally private

setting, consider the family of processes for each  $\mu \in [0, 1]$ , and for any user-chosen integer  $D \geq 2$ ,

$$\begin{aligned}\mathcal{K}_t^+(\mu) &:= \sum_{d=1}^D \prod_{i=1}^t \left[ 1 + \lambda_{i,d}^+ \cdot (Z_i - \zeta_i(\mu)) \right], \\ \text{and } \mathcal{K}_t^-(\mu) &:= \sum_{d=1}^D \prod_{i=1}^t \left[ 1 - \lambda_{i,d}^- \cdot (Z_i - \zeta_i(\mu)) \right],\end{aligned}$$

where  $\lambda_{i,d}^+ := \frac{d}{(D+1)\zeta_i(\mu)}$  and  $\lambda_{i,d}^- := \frac{d}{(D+1)(1-\zeta_i(\mu))}$  for each  $i$ . Then we have the following locally private CSs for  $\mu^*$ .

**Theorem 5.B.2 (NPRR-GK-CS).** *Let  $(Z_t)_{t=1}^\infty \sim Q$  for some  $Q \in \mathcal{Q}_{\mu^*}^\infty$  be the output of NPRR as described in Section 5.2. For any prespecified  $\theta \in [0, 1]$ , define the process  $(\mathcal{K}_t^{\text{GK}}(\mu))_{t=0}^\infty$  given by*

$$\mathcal{K}_t^{\text{GK}}(\mu) := \theta \mathcal{K}_t^+(\mu) + (1 - \theta) \mathcal{K}_t^-(\mu),$$

with  $\mathcal{K}_0^{\text{GK}}(\mu) \equiv 1$ . Then,  $\mathcal{K}_t^{\text{GK}}(\mu^*)$  forms a  $\mathcal{Q}_{\mu^*}^\infty$ -NM, and

$$\bar{C}_t^{\text{GK}} := \left\{ \mu \in [0, 1] : \mathcal{K}_t^{\text{GK}}(\mu) < \frac{1}{\alpha} \right\}$$

forms a  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCS for  $\mu^*$ , meaning  $\mathbb{P}(\forall t, \mu^* \in \bar{C}_t^{\text{GK}}) \geq 1 - \alpha$ . Moreover,  $\bar{C}_t^{\text{GK}}$  forms an interval almost surely.

The proof of Theorem 5.B.2 is given in Section 5.C.4 and follows from Ville's inequality for nonnegative supermartingales [264, 124]. If a lower or upper CS is desired, one can set  $\theta = 1$  or  $\theta = 0$ , respectively, with  $\theta = 1/2$  yielding a two-sided CS. In our simulations (Figure 5.5), we set  $D = 30$ , and  $(r, G)$  were chosen using the technique outlined in Section 5.B.6.

In Proposition 5.B.2, we presented a closed-form empirical Bernstein CI for  $\mu^*$  under NPRR. Similar to the relationship between the fixed-time NPRR-Hoeffding CI (Theorem 5.3.3) and the time-uniform NPRR-Hoeffding CS (Theorem 5.3.2), Proposition 5.B.2 is a corollary of a more general closed-form empirical Bernstein CS instantiated at a fixed sample size. We omitted this CS from the main discussion for brevity, but provide its details here.

**Proposition 5.B.3 (NPRR-EB-CS).** *Given  $(Z_t)_{t=1}^\infty \sim \mathcal{Q}_{\mu^*}^\infty$  and let  $\hat{\mu}_t(\lambda_1^t)$  and  $\bar{B}_t^{\text{EB}}(\lambda_1^t)$  be as in Proposition 5.B.2:*

$$\begin{aligned}\hat{\mu}_t(\lambda_1^t) &:= \frac{\sum_{i=1}^t \lambda_i \cdot (Z_i - (1 - r_i)/2)}{\sum_{i=1}^t r_i \lambda_i}, \text{ and} \\ \bar{B}_t^{\text{EB}}(\lambda_1^t) &:= \frac{\log(1/\alpha) + \sum_{i=1}^t 4(Z_i - \hat{\zeta}_{i-1})^2 \psi_E(\lambda_i)}{\sum_{i=1}^t r_i \lambda_i}.\end{aligned}$$

where  $\psi_E(\lambda) := (-\log(1 - \lambda) - \lambda)/4$ . Then,

$$\bar{L}_t^{\text{EB}} := \hat{\mu}_t(\lambda_1^t) - \bar{B}_t^{\text{EB}}(\lambda_1^t) \quad (5.39)$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCS for  $\mu^\star$ , meaning  $\mathbb{P}(\forall t \geq 1, \mu^\star \geq \bar{L}_t^{\text{EB}}) \geq 1 - \alpha$ .

The proof can be found in Section 5.C.5, and combines the techniques for deriving private concentration inequalities (such as in Theorem 5.3.3) with those for deriving predictable plug-in empirical Bernstein inequalities (such as in Chapter 3).

Similar to Theorem 5.B.1 and Proposition 5.B.2, the proofs and techniques of Theorem 5.B.2 and Proposition 5.B.3 could potentially be adapted to many other variance-adaptive CSs for bounded random variables, including other bounds contained in Chapter 3, Kuchibhotla and Zheng [164], or Orabona and Jun [195].

## 5.B.6 Choosing $(r, G)$ for variance-adaptive confidence sets

Unlike Hoeffding-type bounds, it is not immediately clear how we should choose  $(r, G)$  to satisfy  $\varepsilon$ -LDP and obtain sharp confidence sets using Theorems 5.B.2 and 5.B.1, since there is no closed form bound to optimize. Nevertheless, certain heuristic calculations can be performed to choose  $(r, G)$  in a principled way.<sup>5</sup>

One approach is to view the raw-to-private data mapping  $X \mapsto Z$  as a channel over which information is lost, and we would like to choose the mapping so that as much information is preserved as possible. We will aim to measure ‘‘information lost’’ by the conditional entropy  $H(Z | X)$  and minimize a surrogate of this value.

For the sake of illustration, suppose that  $X$  has a continuous uniform distribution. This is a reasonable starting point because it captures the essence of preserving information about a continuous random variable  $X$  on a discretely supported output space  $\mathcal{G} := \{0, 1/G, \dots, G/G\}$ . Then, the entropy  $H(Z | X = x)$  conditioned on  $X = x$  is given by

$$H(Z | X = x) := \sum_{z \in \mathcal{G}} \mathbb{P}(Z = z | X = x) \log_2 \mathbb{P}(Z = z | X = x), \quad (5.40)$$

and we know that by definition of NPPR, the conditional probability mass function of  $(Z | X)$  is

$$\mathbb{P}(Z = z | X = x) = \frac{1 - r}{G + 1} + rG \cdot \left[ \mathbb{1}(z = x^{\text{ceil}})(x - x^{\text{floor}}) + \mathbb{1}(z = x^{\text{floor}})(x^{\text{ceil}} - x) \right].$$

We will use the heuristic approximation  $x - x^{\text{floor}} \approx x^{\text{ceil}} - x \approx 1/(2G)$ , which would hold with equality if  $x$  were at the midpoint between  $x^{\text{floor}}$  and  $x^{\text{ceil}}$ . With this approximation in

---

<sup>5</sup>Note that ‘‘heuristics’’ do not invalidate the method – no matter what  $(r, G)$  are chosen to be,  $\varepsilon$ -LDP and confidence set coverage are preserved. We are just using heuristic to choose  $(r, G)$  in a smart way for the sake of gaining power.

mind, we can write

$$\begin{aligned}\mathbb{P}(Z = z \mid X = x) &\approx \frac{1-r}{G+1} + rG \cdot \left[ \frac{1}{2G} \mathbb{1}(z = x^{\text{ceil}} \text{ or } z = x^{\text{floor}}) \right] \\ &= \frac{1-r}{G+1} + \frac{r}{2} \mathbb{1}(z = x^{\text{ceil}} \text{ or } z = x^{\text{floor}})\end{aligned}\quad (5.41)$$

Given (5.41), we can heuristically compute  $H(Z \mid X = x)$  because for exactly two terms in the sum  $\sum_{z \in \mathcal{G}} \mathbb{P}(Z = z \mid X = x) \log_2 \mathbb{P}(Z = z \mid X = x)$ , we will have  $\mathbb{1}(z = x^{\text{ceil}} \text{ or } z = x^{\text{floor}}) = 1$  and the other  $G - 1$  terms will have the indicator set to 0. Simplifying notation slightly, let  $p_1(r, G) := (1-r)/(G+1) + r/2$  be (5.41) for those whose indicator is 1, and  $p_0(r, G) := (1-r)/(G+1)$  for those whose indicator is 0. Therefore, we can write

$$H(Z \mid X = x) \approx (G-1)p_0(r, G) \log_2 p_0(r, G) + 2p_1(r, G) \log_2 p_1(r, G). \quad (5.42)$$

Finally, the conditional entropy  $H(Z \mid X)$  can be approximated by

$$H(Z \mid X) = \int_0^1 H(Z \mid X = x) dx \approx (G-1)p_0(r, G) \log_2 p_0(r, G) + 2p_1(r, G) \log_2 p_1(r, G), \quad (5.43)$$

since we assumed that  $X$  was uniform on  $[0, 1]$ .

Given a fixed privacy level  $\varepsilon \in (0, \infty)$ , the approximation (5.43) gives us an objective function to minimize with respect to  $r$  (since  $G$  is completely determined by  $r$  once  $\varepsilon$  is fixed). This can be done using standard numerical minimization solvers. Once an optimal  $(r_{\text{opt}}, \tilde{G}_{\text{opt}})$  pair is determined numerically,  $\tilde{G}_{\text{opt}}$  may not be an integer (but we require  $G \geq 1$  to be an integer for NPPR). As such, one can then choose the final  $G_{\text{opt}}$  to be  $\lfloor \tilde{G}_{\text{opt}} \rfloor$  or  $\lceil \tilde{G}_{\text{opt}} \rceil$ , depending on which one minimizes  $H(Z \mid X)$  while keeping  $\varepsilon$  fixed. If the numerically determined  $\tilde{G}_{\text{opt}}$  is  $\leq 1$ , then one can simply set  $G_{\text{opt}} := 1$  and adjust  $r_{\text{opt}}$  accordingly.

### 5.B.7 One-sided time-varying

The following one-sided analogue of Theorem 5.3.4 can be derived via slightly different techniques; the details can be found in its proof.

**Proposition 5.B.4.** *Given the same setup as Theorem 5.3.4, define*

$$\tilde{B}_t := \sqrt{\frac{t\beta^2 + 1}{2(tr\beta)^2} \log \left( 1 + \frac{\sqrt{t\beta^2 + 1}}{2\alpha} \right)}. \quad (5.44)$$

*Then,  $\tilde{L}_t := \hat{\mu}_t - \tilde{B}_t$  forms a lower  $(1 - \alpha, \varepsilon)$ -LPCS for  $\tilde{\mu}_t^* := \frac{1}{t} \sum_{i=1}^t \mu_i^*$ , meaning*

$$\mathbb{P}(\forall t, \tilde{\mu}_t^* \geq \tilde{L}_t) \geq 1 - \alpha. \quad (5.45)$$

The proof is provided in Section 5.C.7 and uses a one-sided sub-Gaussian mixture supermartingale technique similar to Howard et al. [125, Proposition 6]. Since  $\tilde{B}_t$  resembles  $\tilde{B}_t^\pm$  but

with  $\alpha$  doubled, we suggest choosing  $\beta$  using (5.16) but with  $\beta_{2\alpha}(t_0)$ . We display  $\tilde{L}_t$  alongside the two-sided bound  $\bar{C}_t^\pm$  of Theorem 5.3.4 in Figure 5.6.

### 5.B.8 Private hypothesis testing and $p$ -values

So far, we have focused on the use of *confidence sets* for statistical inference, but another closely related perspective is through the lens of hypothesis testing and  $p$ -values (and their sequential counterparts). Fortunately, we do not need any additional techniques to derive methods for testing, since they are byproducts of our previous results.

Following the nonparametric conditions<sup>6</sup> outlined in Section 5.3, suppose that  $(X)_{t=1}^\infty \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^\infty$  which are then privatized into  $(Z)_{t=1}^\infty \sim Q \in \mathcal{Q}_{\mu^*}^\infty$  via NPPR. The goal now — “locally private sequential testing” — is to use the private data  $(Z)_{t=1}^\infty$  to test some null hypothesis  $\mathcal{H}_0$ . For example, to test  $\mu^* = \mu_0$ , we set  $\mathcal{H}_0 = \mathcal{Q}_{\mu_0}^\infty$  or to test  $\mu^* \leq \mu_0$ , we set  $\mathcal{H}_0 = \{Q \in \mathcal{Q}_\mu^\infty : \mu \leq \mu_0\}$ .

Concretely, we are tasked with designing a binary-valued function  $\bar{\phi}_t \equiv \bar{\phi}(Z_1, \dots, Z_t) \rightarrow \{0, 1\}$  with outputs of 1 and 0 being interpreted as “reject  $\mathcal{H}_0$ ” and “fail to reject  $\mathcal{H}_0$ ”, respectively, so that

$$\sup_{Q \in \mathcal{H}_0} Q(\exists t : \bar{\phi}_t = 1) \leq \alpha. \quad (5.46)$$

A sequence of functions  $(\bar{\phi})_{t=1}^\infty 1$  satisfying (5.46) is known as a *level- $\alpha$  sequential test*. Another common tool in hypothesis testing is the  $p$ -value, which also has a sequential counterpart, known as the *anytime p-value* [135, 125]. We say that a sequence of  $p$ -values  $(\bar{p})_{t=1}^\infty 1$  is an *anytime p-value* if

$$\sup_{Q \in \mathcal{H}_0} Q(\exists t : \bar{p}_t \leq \alpha) \leq \alpha. \quad (5.47)$$

There are at least two ways to achieve (5.46) and (5.47): (a) by using CSs to reject non-intersecting null hypotheses, and (b) by explicitly deriving  $e$ -processes. We will first discuss (a) and leave (b) to Section 5.B.8.2 as the discussion is more involved.

#### 5.B.8.1 Private hypothesis testing using confidence sets.

The simplest and most direct way to test hypotheses using the results of this chapter is to exploit the duality between CSs and sequential tests (or CIs and fixed-time tests). Suppose  $(\bar{C}_t(\alpha))_{t=1}^\infty$  is an LDP  $(1 - \alpha)$ -CS for  $\mu^*$ , and let  $\mathcal{H}_0 : \{Q \in \mathcal{Q}_\mu^\infty : \mu \in \Theta_0\}$  be a null hypothesis that we wish to test. Then, for any  $\alpha \in (0, 1)$ ,

$$\bar{\phi}_t := \mathbb{1}(\bar{C}_t(\alpha) \cap \Theta_0 = \emptyset) \quad (5.48)$$

forms an LDP level- $\alpha$  sequential test for  $\mathcal{H}_0$ , meaning it satisfies (5.46). In particular, if  $\bar{C}_t(\alpha)$  shrinks to a single point as  $t \rightarrow \infty$ , then  $(\bar{\phi})_{t=1}^\infty 1$  has asymptotic power one. Furthermore,  $\inf\{\alpha : \bar{C}_t(\alpha) \cap \Theta_0 = \emptyset\}$  forms an anytime  $p$ -value for  $\mathcal{H}_0$ , meaning it satisfies (5.47).

---

<sup>6</sup>The discussion that follows also applies to the parametric case.

Similarly, if  $\dot{C}_n(\alpha)$  is a  $(1 - \alpha)$  CI for  $\mu^*$ , then  $\dot{\phi}_n := \mathbb{1}(\dot{C}_n(\alpha) \cap \Theta_0 = \emptyset)$  is a level- $\alpha$  test:  $\sup_{Q \in \mathcal{H}_0} Q(\dot{\phi}_n = 1) \leq \alpha$ , and  $\dot{p}_n := \inf\{\alpha : \dot{C}_n(\alpha) \cap \Theta_0 = \emptyset\}$  is a  $p$ -value for  $\mathcal{H}_0$ :  $\sup_{Q \in \mathcal{H}} Q(\dot{p}_n \leq \alpha) \leq \alpha$ .

One can also derive sequential tests using so-called *e-processes* – processes that are upper-bounded by nonnegative supermartingales under a given null hypothesis. In fact, every single one of our CSs is derived by first deriving an explicit *e*-process. Let us now discuss how one can derive sequential tests and CSs using *e*-processes.

### 5.B.8.2 Testing via *e*-processes

To achieve (5.46) and (5.47), it is also sufficient to derive an *e*-process  $(\bar{E})_{t=1}^\infty 1$  – a  $\mathcal{Z}$ -adapted process that is upper bounded by an NSM for every element of  $\mathcal{H}_0$ . Formally,  $(\bar{E})_{t=1}^\infty 1$  is an *e*-process for  $\mathcal{H}_0$  if for every  $Q \in \mathcal{H}_0$ , there exists a  $Q$ -NSM  $(M_t^Q)_{t=1}^\infty 1$  such that

$$\forall t, \bar{E}_t \leq M_t^Q, \quad Q\text{-almost surely.} \quad (5.49)$$

Here,  $(M_t^Q)_{t=1}^\infty 1$  being a  $Q$ -NSM means that  $\mathbb{E}_Q M_t^Q \leq M_{t-1}^Q$ , and  $M_0^Q \equiv 1$ , and  $M_t^Q \geq 0$ ,  $Q$ -almost surely. Note that these upper-bounding NSMs need not be the same, i.e.  $(\bar{E})_{t=1}^\infty 1$  can be upper bounded by a different  $Q$ -NSM for each  $Q \in \mathcal{H}_0$ .

Importantly, if  $(\bar{E})_{t=1}^\infty 1$  is an *e*-process under  $\mathcal{H}_0$ , then  $\phi_t := \mathbb{1}(\bar{E}_t \geq 1/\alpha)$  forms a level- $\alpha$  sequential test satisfying (5.46) by applying Ville's inequality to the NSM that upper bounds  $(\bar{E})_{t=1}^\infty 1$ :

$$\sup_{Q \in \mathcal{H}_0} Q(\exists t : \bar{E}_t \geq 1/\alpha) \leq \alpha. \quad (5.50)$$

Using the same technique it is easy to see that,  $\bar{p}_t := 1/\bar{E}_t$  forms an anytime  $p$ -value satisfying (5.47). Similarly to Section 5.3, if we are only interested in inference at a fixed sample size  $n$ , we can still leverage *e*-processes to obtain sharp finite-sample  $p$ -values from private data by simply taking

$$\dot{p}_n := \min_{1 \leq t \leq n} 1/\bar{E}_t. \quad (5.51)$$

As an immediate consequence of (5.50), we have  $\sup_{Q \in \mathcal{H}_0} Q(\dot{p}_n \leq \alpha) \leq \alpha$ .

With all of this in mind, the question becomes: where can we find *e*-processes? The answer is simple: every single CS and CI in this chapter was derived by first constructing an *e*-process under a point null, and Table 5.1 explicitly links all of these CSs to their corresponding *e*-processes.<sup>7</sup> For more complex composite nulls however, there may exist *e*-processes that are not NSMs [210], and we touch on one such example in Proposition 5.B.5.

**A note on locally private *e*-values.** Similar to how the  $p$ -value is the fixed-time version of an anytime  $p$ -value, the so-called *e-value* is the fixed-time version of an *e*-process. An *e*-value for a null  $\mathcal{H}_0$  is a nonnegative random variable  $\dot{E}$  with  $Q$ -expectation at most one, meaning  $\mathbb{E}_Q(\dot{E}) \leq 1$  for any  $Q \in \mathcal{H}_0$  [113, 266], and clearly by Markov's inequality,  $1/\dot{E}$  is a  $p$ -value

<sup>7</sup>We do not link CIs to *e*-processes since all of our CIs are built using the aforementioned CSs.

Table 5.1: A mapping between theorems containing confidence sequences and equations containing the explicit  $e$ -processes that underlie them.

Confidence sequence	$e$ -process
Theorem 5.3.3	(5.22)
Theorem 5.3.4	(5.25)
Proposition 5.B.4	(5.73)

for  $\mathcal{H}_0$ . Indeed, the time-uniform property (5.50) for the  $e$ -process  $(E)_{t=1}^\infty 1$  is *equivalent* to saying  $E_\tau$  is an  $e$ -value for any stopping time  $\tau$  [125, Lemma 3]; [295, Proposition 1].

Given the shared goals between  $e$ - and  $p$ -values, a natural question arises: “Should one use  $e$ -values or  $p$ -values for inference?”. While  $p$ -values are the canonical measure of evidence in hypothesis testing, there are several reasons why one may prefer to work with  $e$ -values directly; some practical, and others philosophical. From a purely practical perspective,  $e$ -values make it simple to combine evidence across several studies [113, 251, 266] or to control the false discovery rate under arbitrary dependence [275]. They have also received considerable attention for philosophical reasons including how they relate testing to betting [234] and connect frequentist and Bayesian notions of uncertainty [113, 281]. While the details of these advantages are well outside the scope of this chapter, they are advantages that can now be enjoyed in locally private inference using our methods.

### 5.B.9 A/B testing the weak null

As described in Section 5.B.8, there is a close connection between CSs and sequential hypothesis tests. The lower CS  $(\tilde{L}^\Delta)_{t=1}^\infty 1$  presented in Proposition 5.B.5 is no exception, and can be used to test the weak null hypothesis,  $\tilde{\mathcal{H}}_0: \forall t, \tilde{\Delta}_t \leq 0$  (see Figure 5.7). In words,  $\tilde{\mathcal{H}}_0$  is testing “is the new treatment as bad or worse than placebo *among the patients so far?*”. Indeed, adapting (5.74) from the proof of Proposition 5.B.4 to the current setting, we have the following anytime  $p$ -value for the weak null under locally private online A/B tests.

**Proposition 5.B.5.** *Consider the same setup as Corollary 5.4.1, and let  $\Phi(\cdot)$  be the cumulative distribution function of a standard Gaussian. Define for any  $\beta > 0$ ,*

$$\tilde{E}_t^\Delta := \frac{2}{\sqrt{t\beta^2 + 1}} \exp \left\{ \frac{2\beta^2(S_{t,0}^\Delta)^2}{t\beta^2 + 1} \right\} \Phi \left( \frac{2\beta S_{t,0}^\Delta}{\sqrt{t\beta^2 + 1}} \right),$$

where  $S_{t,0}^\Delta := \sum_{i=1}^t (\psi_i - (1-r)/2) - tr \frac{1/(1-\pi)}{1/\pi + 1/(1-\pi)}$  and  $\beta > 0$ . Then,  $\tilde{E}_t^\Delta$  forms an  $e$ -process and hence  $\tilde{p}_t^\Delta := 1/\tilde{E}_t^\Delta$  forms an anytime  $p$ -value, and  $\tilde{\phi}_t^\Delta := \mathbb{1}(\tilde{p}_t^\Delta \leq \alpha)$  forms a level- $\alpha$  sequential test for the weak null  $\tilde{\mathcal{H}}_0$ .

The proof provided in Section 5.C.6 relies on the simple observation that under  $\tilde{\mathcal{H}}_0$ ,

$(\tilde{E}^\Delta)_{t=1}^\infty \mathbf{1}$  is upper bounded by a nonnegative supermartingale, and is hence an “ $e$ -process”. We suggest choosing  $\beta > 0$  in a similar manner to Proposition 5.B.4.

### 5.B.10 Locally private adaptive online A/B testing

In Section 5.4, we demonstrated how our techniques can be used to conduct online A/B tests. However, those A/B tests were non-adaptive, in the sense that the propensity score  $\pi \in (0, 1)$  was required to be the same constant for all individuals (e.g. in a Bernoulli experiment). In this section, we briefly describe an alternative CS that can be used to conduct *adaptive* online A/B tests, where the propensity scores  $(\pi_t(X_t))_{t=1}^\infty$  can change over time in a data-dependent fashion and be a function of some measured baseline covariates  $(X_t)_{t=1}^\infty$ . Note that while we will still consider private tests in the sense of the outcomes  $(Y_t)_{t=1}^\infty$  being privatized, we will not be privatizing the covariates  $(X_t)_{t=1}^\infty$  (though this is an interesting direction for future work).

To set the stage, suppose that  $(X_1, A_1, Y_1), (X_2, A_2, Y_2), \dots$  are joint random variables such that covariates  $X_t \sim p_X(\cdot)$ , are drawn according to some common distribution, treatments  $A_t \sim \text{Bernoulli}(\pi_t(X_t))$  are drawn from a conditional distribution  $\pi_t$  (called the propensity score) which can be chosen based on  $(X_i, A_i, Y_i)_{i=1}^{t-1}$ , and  $Y_t \sim p_Y(\cdot | A_t, X_t)$  is drawn from a common conditional distribution.<sup>8</sup> In words, we have that for each subject  $t$ , covariates  $X_t$  are observed, a propensity score  $\pi_t$  is chosen based on all previous subjects, a binary treatment  $A_t$  is drawn with probability  $\pi_t(X_t)$ , and a  $[0, 1]$ -bounded outcome  $Y_t$  is observed based on subject  $t$ 's covariates and their treatment  $A_t$ . Of course, if  $\pi(X_t) \equiv \pi$  for each  $t$ , then the above setup recovers the classical (non-adaptive) A/B testing setup considered in Section 5.4.

Similarly to Section 5.4, we will construct  $(1 - \alpha)$ -CSs for the *time-varying mean*  $\tilde{\Delta}_t := \frac{1}{t} \sum_{i=1}^t \Delta_i$  where

$$\Delta_i := \mathbb{E}\{\mathbb{E}(Y_i | X_i, A_i = 1) - \mathbb{E}(Y_i | X_i, A_i = 0)\} \quad (5.52)$$

is the individual treatment effect for subject  $i$ . To state our main result, we need to prepare some notation. Let  $w_t^{(1)} := \frac{\mathbb{1}(A_t=1)}{\pi_t(X_t)}$  and  $w_t^{(0)} := \frac{\mathbb{1}(A_t=0)}{1-\pi_t(X_t)}$  denote the inverse propensity score weights for treatment and control groups, respectively, and define the following pseudo-outcomes  $\theta_t := [w_t^{(1)} Z_t - (1 - w_t^{(0)})(1 - Z_t)]/r$ , and the resulting variance process

$$V_t := \frac{1}{t} \sum_{i=1}^t \left( \theta_i - \hat{\theta}_{i-1} \right)^2, \text{ where } \hat{\theta}_t := \left( \frac{1}{t} \sum_{i=1}^t \theta_i \right) \wedge 1 \quad (5.53)$$

We are now ready to state the main result of this section.

**Theorem 5.B.3** (Locally private adaptive A/B estimation). *Let  $S_t(\tilde{\Delta}'_t) := (\sum_{i=1}^t \theta_i - t\tilde{\Delta}'_t)/2$  for any  $\tilde{\Delta}'_t \in [0, 1]$  and define for any  $\rho > 0$ ,*

$$\widetilde{M}_t^{\text{EB}}(\tilde{\Delta}'_t) := \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{V_t + \rho} \right) F_t(\tilde{\Delta}'_t), \quad (5.54)$$

---

<sup>8</sup>These distributional assumptions can be substantially weakened as in Chapter 6, but we present this simplified setting for the sake of exposition.

where  $F_t(\tilde{\Delta}'_t) := {}_1F_1(1, V_t + \rho + 1, S_t(\tilde{\Delta}'_t) + V_t + \rho)$ , and  ${}_1F_1$  is Kummer's confluent hypergeometric function, and  $\Gamma(\cdot, \cdot)$  is the upper incomplete gamma function. Then, when evaluated at the true  $\tilde{\Delta}_t$ , we have that  $\tilde{M}_t^{\text{EB}}(\tilde{\Delta}_t)$  forms a nonnegative supermartingale. Consequently,

$$\tilde{L}_t^\Delta := \inf \left\{ \tilde{\Delta}_t \in [0, 1] : \tilde{M}_t^{\text{EB}}(\tilde{\Delta}_t) < 1/\alpha \right\} \quad (5.55)$$

forms a lower  $(1 - \alpha)$ -CS for the running ATE  $\tilde{\Delta}_t$ .

The proof can be found in Section 5.C.8. Readers familiar with the semiparametric causal inference literature will notice that  $\theta_t$  takes the form of a modified inverse-probability-weighted (IPW) influence function, and that doubly robust (also known as “augmented IPW”) approaches are often superior both theoretically and empirically. In principle, the above discussion can be modified to handle doubly robust pseudo-outcomes and CSs using the ideas contained in Chapter 6, but we presented the IPW-based approach instead for the sake of simplicity.

## 5.C Proofs of additional results

### 5.C.1 Proof of Proposition 5.B.1

**Proposition 5.B.1** (Lap-H-CS). Suppose  $(X)_{t=1}^\infty 1 \sim P$  for some  $\mathcal{P}_{\mu^*}^\infty$  and let  $(Z)_{t=1}^\infty 1$  be their privatized views according to Algorithm 5.1. Let  $\psi_t^{\mathcal{L}}(\lambda) := -\log(1 - \lambda^2/\varepsilon_t^2)$  be the (conditional) cumulant generating function of a mean-zero Laplace random variable with scale  $1/\varepsilon_t$ . Let  $(\lambda)_{t=1}^\infty 1$  be a sequence of random variables such that  $\lambda_t$  depends on  $Z_1^{t-1}$  — formally  $\sigma(Z_1^{t-1})$ -measurable — and  $[0, \varepsilon_t]$ -valued. Then,

$$\bar{L}_t^{\mathcal{L}} := \frac{\sum_{i=1}^t \lambda_i Z_i}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\alpha) + \sum_{i=1}^t (\lambda_i^2/8 + \psi_i^{\mathcal{L}}(\lambda_i))}{\sum_{i=1}^t \lambda_i} \quad (5.31)$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_{t=1}^\infty)$ -LPCS for  $\mu^*$ .

*Proof.* The proof proceeds in two steps. First, we construct an exponential NSM using the cumulant generating function of a Laplace distribution. Second and finally, we apply Ville’s inequality to the NSM and invert it to obtain the lower CS.

**Step 1.** Consider the following process for any  $\mu \in [0, 1]$ ,

$$M_t^{\mathcal{L}}(\mu) := \prod_{i=1}^t \exp \left\{ \lambda_i (Z_i - \mu) - \lambda_i^2/8 - \psi_i^{\mathcal{L}}(\lambda_i) \right\},$$

with  $M_t^{\mathcal{L}}(\mu) \equiv 1$ . We claim that  $(M_t^{\mathcal{L}}(\mu^*))_{t=0}^\infty$  forms an NSM with respect to the private filtration  $\mathcal{Z}$ . Indeed,  $(M_t^{\mathcal{L}}(\mu^*))_{t=0}^\infty$  is nonnegative and starts at one by construction. It remains to prove that  $(M_t^{\mathcal{L}}(\mu^*))$  is a supermartingale, meaning  $\mathbb{E}(M_t^{\mathcal{L}}(\mu^*) \mid \mathcal{Z}_{t-1}) \leq M_{t-1}^{\mathcal{L}}(\mu^*)$ . Writing out the conditional expectation of  $M_t^{\mathcal{L}}(\mu^*)$ , we have

$$\mathbb{E}(M_t^{\mathcal{L}}(\mu^*) \mid \mathcal{Z}_{t-1})$$

$$\begin{aligned}
&= \mathbb{E} \left( \prod_{i=1}^t \exp \{ \lambda_i(Z_i - \mu^*) - \lambda_i^2/8 - \psi_i^{\mathcal{L}}(\lambda_i) \} \mid \mathcal{Z}_{t-1} \right) \\
&= \underbrace{\prod_{i=1}^{t-1} \exp \{ \lambda_i(Z_i - \mu^*) - \lambda_i^2/8 - \psi_i^{\mathcal{L}}(\lambda_i) \}}_{M_{t-1}(\mu^*)} \cdot \underbrace{\mathbb{E} \left( \exp \{ \lambda_t(Z_t - \mu^*) - \lambda_t^2/8 - \psi_t^{\mathcal{L}}(\lambda_t) \} \mid \mathcal{Z}_{t-1} \right)}_{(\dagger)},
\end{aligned}$$

since  $M_{t-1}^{\mathcal{L}}(\mu^*)$  is  $\mathcal{Z}_{t-1}$ -measurable, and thus it can be written outside of the conditional expectation. It now suffices to show that  $(\dagger) \leq 1$ . To this end, note that  $Z_t = X_t + \mathcal{L}_t$  where  $X_t$  is a  $[0, 1]$ -bounded, mean- $\mu^*$  random variable, and  $\mathcal{L}_t$  is a mean-zero Laplace random variable (conditional on  $\mathcal{Z}_{t-1}$ ). Consequently,  $\mathbb{E}(\exp \{ \lambda_t X_t \} \mid \mathcal{Z}_{t-1}) \leq \exp \{ \lambda_t^2/8 \}$  by Hoeffding's inequality [120], and  $\mathbb{E}(\exp \{ \lambda_t \mathcal{L}_t \} \mid \mathcal{Z}_{t-1}) = \exp \{ \lambda_t^{\mathcal{L}}(\lambda_t) \}$  by definition of a Laplace random variable. Moreover, note that by design of Algorithm 5.1,  $X_t$  and  $\mathcal{L}_t$  are conditionally independent. It follows that

$$\begin{aligned}
(\dagger) &= \mathbb{E} \left( \exp \{ \lambda_t(Z_t - \mu^*) - \lambda_t^2/8 - \psi_t^{\mathcal{L}}(\lambda_t) \} \mid \mathcal{Z}_{t-1} \right) \\
&= \mathbb{E} \left( \exp \{ \lambda_t(X_t - \mu^*) + \lambda_t \mathcal{L}_t - \lambda_t^2/8 - \psi_t^{\mathcal{L}}(\lambda_t) \} \mid \mathcal{Z}_{t-1} \right) \\
&= \underbrace{\mathbb{E} \left( \exp \{ \lambda_t(X_t - \mu^*) - \lambda_t^2/8 \} \mid \mathcal{Z}_{t-1} \right)}_{\leq 1} \cdot \underbrace{\mathbb{E} \left( \exp \{ \lambda_t \mathcal{L}_t - \psi_t^{\mathcal{L}}(\lambda_t) \} \mid \mathcal{Z}_{t-1} \right)}_{=1} \leq 1,
\end{aligned}$$

where the third equality follows from the conditional independence of  $X_t$  and  $\mathcal{L}_t$ . Therefore,  $(M_t^{\mathcal{L}}(\mu^*))_{t=0}^{\infty}$  is an NSM.

**Step 2.** By Ville's inequality, we have that

$$\mathbb{P}(\forall t, M_t^{\mathcal{L}}(\mu^*) < 1/\alpha) \geq 1 - \alpha.$$

Let us rewrite the inequality  $M_t^{\mathcal{L}}(\mu^*) < 1/\alpha$  so that we obtain the desired lower CS.

$$\begin{aligned}
M_t^{\mathcal{L}}(\mu^*) < 1/\alpha &\iff \prod_{i=1}^t \exp \{ \lambda_t(Z_t - \mu^*) - \lambda_i^2/8 - \psi_i^{\mathcal{L}}(\lambda_i) \} < 1/\alpha \\
&\iff \sum_{i=1}^t [\lambda_t(Z_t - \mu^*) - \lambda_i^2/8 - \psi_i^{\mathcal{L}}(\lambda_i)] < \log(1/\alpha) \\
&\iff \sum_{i=1}^t \lambda_t Z_t - \sum_{i=1}^t [\lambda_i^2/8 + \psi_i^{\mathcal{L}}(\lambda_i)] - \mu^* \sum_{i=1}^t \lambda_i < \log(1/\alpha) \\
&\iff \mu^* > \frac{\sum_{i=1}^t \lambda_i Z_t}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\alpha) + \sum_{i=1}^t (\lambda_i^2/8 + \psi_i^{\mathcal{L}}(\lambda_i))}{\sum_{i=1}^t \lambda_i}.
\end{aligned}$$

In summary, the above inequality holds uniformly for all  $t \in \{1, 2, \dots\}$  with probability at least  $(1 - \alpha)$ . In other words,

$$\frac{\sum_{i=1}^t \lambda_i Z_t}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\alpha) + \sum_{i=1}^t (\lambda_i^2/8 + \psi_i^L(\lambda_i))}{\sum_{i=1}^t \lambda_i}$$

forms a  $(1 - \alpha)$ -lower CS for  $\mu^*$ . An analogous upper-CS can be derived by applying the same technique to  $-Z_1, -Z_2, \dots$  and their mean  $-\mu^*$ . This completes the proof.  $\square$

### 5.C.2 A lemma for Theorems 5.B.1 and 5.B.2

To prove Theorems 5.B.2 and 5.B.1, we will prove a more general result (Lemma 5.C.1), and use it to instantiate both theorems as immediate consequences. The proof follows a similar technique to Chapter 3 but adapted to the locally private setting.

**Lemma 5.C.1.** *Suppose  $(X)_{t=1}^\infty \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^\infty$  and let  $(Z)_{t=1}^\infty$  be their NPPR-induced privatized views. Let  $\theta_1, \dots, \theta_D \in [0, 1]$  be convex weights satisfying  $\sum_{d=1}^D \theta_d = 1$  and let  $(\lambda^{(d)})_{t=1}^\infty$  be a  $\mathcal{Z}$ -predictable sequence for each  $d \in \{1, \dots, D\}$  such that  $\lambda_t \in (-(1 - \zeta_t(\mu^*))^{-1}, \zeta_t(\mu^*)^{-1})$ . Then the process formed by*

$$M_t := \sum_{d=1}^D \theta_d \prod_{i=1}^t (1 + \lambda_i^{(d)} \cdot (Z_i - \zeta_i(\mu^*))) \quad (5.56)$$

*is a nonnegative martingale starting at one. Further suppose that  $(\check{M}_t(\mu))_{t=0}^\infty$  is a process for any  $\mu \in (0, 1)$  that when evaluated at  $\mu^*$ , satisfies  $\check{M}_t(\mu^*) \leq M_t$  almost surely for each  $t$ . Then*

$$\check{C}_t := \left\{ \mu \in (0, 1) : \check{M}_t(\mu) < 1/\alpha \right\} \quad (5.57)$$

*forms a  $(1 - \alpha)$ -CS for  $\mu^*$ .*

*Proof.* The proof proceeds in three steps. First, we will show that the product processes given by  $\prod_{i=1}^t (1 + \lambda_i \cdot (Z_i - \zeta_i(\mu^*)))$  form nonnegative martingales with respect to  $\mathcal{Z}$ . Second, we argue that  $\sum_{d=1}^D \theta_d M_t^{(d)}$  forms a martingale for any  $\mathcal{Z}$ -adapted martingales. Third and finally, we argue that  $\check{C}_t$  forms a  $(1 - \alpha)$ -CS despite not being constructed from a martingale directly.

**Step 1.** We wish to show that  $M_t^{(d)} := \prod_{i=1}^t (1 + \lambda_i^{(d)} \cdot (Z_i - \zeta_i(\mu^*)))$  forms a nonnegative martingale starting at one given a fixed  $d \in \{1, \dots, D\}$ . Nonnegativity follows immediately from the fact that  $\lambda_t \in (-(1 - \zeta_t(\mu^*))^{-1}, \zeta_t(\mu^*)^{-1})$ , and  $M_t^{(d)}$  begins at one by design. It remains to show that  $M_t^{(d)}$  forms a martingale. To this end, consider the conditional expectation of  $M_t^{(d)}$  for any  $t \in \{1, 2, \dots\}$ ,

$$\mathbb{E} \left( M_t^{(d)} \mid \mathcal{Z}_{t-1} \right) = \mathbb{E} \left( \prod_{i=1}^t (1 + \lambda_i^{(d)} \cdot (Z_i - \zeta_i(\mu^*))) \mid \mathcal{Z}_{t-1} \right)$$

$$\begin{aligned}
&= \underbrace{\prod_{i=1}^{t-1} (1 + \lambda_i^{(d)} \cdot (Z_i - \zeta_i(\mu^*)))}_{M_{t-1}^{(d)}} \cdot \mathbb{E} \left( 1 + \lambda_t^{(d)} \cdot (Z_t - \zeta_t(\mu^*)) \mid \mathcal{Z}_{t-1} \right) \\
&= M_{t-1}^{(d)} \cdot \left( 1 + \lambda_t^{(d)} \cdot \underbrace{[\mathbb{E}(Z_t \mid \mathcal{Z}_{t-1}) - \zeta_t(\mu^*)]}_{=0} \right) \\
&= M_{t-1}^{(d)}.
\end{aligned}$$

Therefore,  $(M^{(d)})_{t=1}^\infty 0$  forms a martingale.

**Step 2.** Now, suppose that  $M_t^{(1)}, \dots, M_t^{(D)}$  are test martingales with respect to the private filtration  $\mathcal{Z}$ , and let  $\theta_1, \dots, \theta_D \in [0, 1]$  be convex weights, i.e. satisfying  $\sum_{d=1}^D \theta_d = 1$ . Then  $M_t := \sum_{d=1}^D \theta_d M_t^{(d)}$  also forms a martingale since

$$\begin{aligned}
\mathbb{E}(M_t \mid \mathcal{Z}_{t-1}) &= \mathbb{E} \left( \sum_{d=1}^D \theta_d M_t^{(d)} \mid \mathcal{Z}_{t-1} \right) \\
&= \sum_{d=1}^D \theta_d \mathbb{E} \left( M_t^{(d)} \mid \mathcal{Z}_{t-1} \right) \\
&= \sum_{d=1}^D \theta_d M_{t-1}^{(d)} \\
&= M_{t-1}.
\end{aligned}$$

Moreover,  $(M)_{t=1}^\infty 0$  starts at one since  $M_0 := \sum_{d=1}^D \theta_d M_0^{(d)} = \sum_{d=1}^D \theta_d = 1$ . Finally, nonnegativity follows from the fact that  $\theta_1, \dots, \theta_D$  are convex and each  $(M^{(d)})_{t=1}^\infty 0$  is almost-surely nonnegative. Therefore,  $(M)_{t=1}^\infty 0$  is a test martingale.

**Step 3.** Now, suppose  $(\check{M}_t(\mu))_{t=0}^\infty$  is a process that is almost-surely upper-bounded by  $(M)_{t=1}^\infty 0$ . Define  $\check{C}_t := \{\mu \in (0, 1) : \check{M}_t(\mu) < 1/\alpha\}$ . Writing out the probability of  $\check{C}_t$  miscovering  $\mu^*$  for any  $t$ , we have

$$\begin{aligned}
\mathbb{P}(\exists t : \mu^* \notin \check{C}_t) &= \mathbb{P}(\exists t : \check{M}_t(\mu^*) \geq 1/\alpha) \\
&\leq \mathbb{P}(\exists t : M_t \geq 1/\alpha) \\
&\leq \alpha,
\end{aligned}$$

where the first inequality follows from the fact that  $\check{M}_t(\mu^*) \leq M_t$  almost surely for each  $t$ , and the second follows from Ville's inequality [264]. This completes the proof of Lemma 5.C.1.  $\square$

In fact, a more general “meta-algorithm” extension of Lemma 5.C.1 holds, following the

derivation of the “Sequentially Rebalanced Portfolio” in Chapter 3 but we omit these details for the sake of simplicity.

### 5.C.3 Proof of Theorem 5.B.1

**Theorem 5.B.1 (NPRR-hedged).** Suppose  $(X)_{t=1}^n \sim P$  for some  $P \in \mathcal{P}_{\mu^*}^n$  and let  $(Z)_{t=1}^n \sim Q$  be their NRR-privatized views where  $Q \in \mathcal{Q}_{\mu^*}^n$ . Define

$$\mathcal{K}_{t,n}(\mu) := \prod_{i=1}^t [1 + \lambda_{i,n}(\mu) \cdot (Z_i - \zeta_i(\mu))] \quad (5.35)$$

with  $\lambda_{t,n}(\mu)$  given by (5.34). Then,  $\mathcal{K}_{t,n}(\mu)$  is a nonincreasing function of  $\mu \in [0, 1]$ , and  $\mathcal{K}_{t,n}(\mu^*)$  forms a  $\mathcal{Q}_{\mu^*}^n$ -NM. Consequently,

$$\dot{L}_n := \max_{1 \leq t \leq n} \inf \{\mu \in [0, 1] : \mathcal{K}_{t,n}(\mu) < 1/\alpha\} \quad (5.36)$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCI for  $\mu^*$ , meaning  $\mathbb{P}(\mu^* \geq \dot{L}_n) \geq 1 - \alpha$ .

*Proof.* The proof of Theorem 5.B.1 proceeds in three steps. First, we show that  $\mathcal{K}_{t,n}$  is non-increasing and continuous in  $\mu \in [0, 1]$ , making  $\dot{L}_n$  simple to compute via line/grid search. Second, we show that  $\mathcal{K}_{t,n}(\mu^*)$  forms a  $\mathcal{Q}_{\mu^*}^\infty$ -NM. Third and finally, we show that  $\dot{L}_n$  is a lower CI by constructing a lower CS that yields  $\dot{L}_n$  when instantiated at  $n$ .

**Step 1.  $\mathcal{K}_{t,n}(\mu)$  is nonincreasing and continuous.** To simplify the notation that follows, write  $g_{i,n}(\mu) := 1 + \lambda_{i,n}(\mu) \cdot (Z_i - \zeta_i(\mu))$  so that

$$\mathcal{K}_{t,n}(\mu) = \prod_{i=1}^t g_{i,n}(\mu).$$

Now, recall the definition of  $\lambda_{i,n}(\mu)$ ,

$$\begin{aligned} \lambda_{t,n}(\mu) &:= \underbrace{\sqrt{\frac{2 \log(1/\alpha)}{\hat{\gamma}_{t-1}^2 n}}}_{\eta} \wedge \frac{c}{\zeta_t(\mu)}, \text{ where} \\ \hat{\gamma}_t^2 &:= \frac{1/4 + \sum_{i=1}^t (Z_i - \hat{\zeta}_i)^2}{t+1}, \quad \hat{\zeta}_t := \frac{1/2 + \sum_{i=1}^t Z_i}{t+1}. \end{aligned}$$

Notice that  $\lambda_{t,n}(\mu) \equiv \eta \wedge c/\zeta_t(\mu)$  is nonnegative and does not depend on  $\mu$  except through the truncation with  $c/\zeta_t(\mu)$ . In particular we can write  $g_{i,n}(\mu)$  as

$$g_{i,n}(\mu) \equiv 1 + \left( \eta \wedge \frac{c}{\zeta_i(\mu)} \right) (Z_i - \zeta_i(\mu))$$

$$= 1 + (\eta Z_i) \wedge \frac{cZ_i}{\zeta_i(\mu)} - \eta \zeta_i(\mu) \wedge c,$$

which is a nonincreasing (and continuous) function of  $\zeta_i(\mu)$ . Since  $\zeta_i(\mu) := r_i\mu + (1 - r_i)/2$  is an increasing (and continuous) function of  $\mu$ , we have that  $g_{i,n}(\mu)$  is nonincreasing and continuous in  $\mu$ .

Moreover, we have that  $g_{i,n}(\mu) \geq 0$  by design, and the product of nonnegative nonincreasing functions is also nonnegative and nonincreasing, so  $\mathcal{K}_{t,n} = \prod_{i=1}^t g_{i,n}(\mu)$  is nonincreasing.

**Step 2.  $\mathcal{K}_{t,n}(\mu^*)$  is a  $\mathcal{Q}_{\mu^*}^\infty$ -NM.** Recall the definition of  $\mathcal{K}_{t,n}(\mu^*)$

$$\mathcal{K}_{t,n}(\mu^*) := \prod_{i=1}^t [1 + \lambda_{i,n}(\mu^*) \cdot (Z_i - \zeta_i(\mu^*))]$$

Then by Lemma 5.C.1 with  $D = 1$  and  $\theta_1 = 1$ ,  $\mathcal{K}_{t,n}(\mu^*)$  is a  $\mathcal{Q}_{\mu^*}^n$ -NM.

**Step 3.  $\dot{L}_n$  is a lower CI.** First, note that by Lemma 5.C.1, we have that

$$C_t := \{\mu \in [0, 1] : \mathcal{K}_{t,n}(\mu) < 1/\alpha\}$$

forms a  $(1 - \alpha)$ -CS for  $\mu^*$ . In particular, define

$$\bar{L}_{t,n} := \inf\{\mu \in [0, 1] : \mathcal{K}_{t,n}(\mu) < 1/\alpha\}.$$

Then,  $[\bar{L}_{t,n}, 1]$  forms a  $(1 - \alpha)$ -CS for  $\mu^*$ , meaning  $\mathbb{P}(\forall t, \mu^* \geq \bar{L}_{t,n}) \geq 1 - \alpha$ , and hence

$$\mathbb{P}\left(\mu^* \geq \max_{1 \leq t \leq n} L_{t,n}\right) = \mathbb{P}\left(\mu^* \geq \dot{L}_n\right) \geq 1 - \alpha.$$

This completes the proof. □

#### 5.C.4 Proof of Theorem 5.B.2

**Theorem 5.B.2 (NPRR-GK-CS).** Let  $(Z)_{t=1}^\infty \sim Q$  for some  $Q \in \mathcal{Q}_{\mu^*}^\infty$  be the output of NRR as described in Section 5.2. For any prespecified  $\theta \in [0, 1]$ , define the process  $(\mathcal{K}_t^{\text{GK}}(\mu))_{t=0}^\infty$  given by

$$\mathcal{K}_t^{\text{GK}}(\mu) := \theta \mathcal{K}_t^+(\mu) + (1 - \theta) \mathcal{K}_t^-(\mu),$$

with  $\mathcal{K}_0^{\text{GK}}(\mu) \equiv 1$ . Then,  $\mathcal{K}_t^{\text{GK}}(\mu^*)$  forms a  $\mathcal{Q}_{\mu^*}^\infty$ -NM, and

$$\bar{C}_t^{\text{GK}} := \left\{ \mu \in [0, 1] : \mathcal{K}_t^{\text{GK}}(\mu) < \frac{1}{\alpha} \right\}$$

forms a  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCS for  $\mu^*$ , meaning  $\mathbb{P}(\forall t, \mu^* \in \bar{C}_t^{\text{GK}}) \geq 1 - \alpha$ . Moreover,  $\bar{C}_t^{\text{GK}}$  forms an interval almost surely.

*Proof.* The proof will proceed in two steps. First, we will invoke Lemma 5.C.1 to justify that  $\bar{C}_t^{\text{GK}}$  indeed forms a CS. Second and finally, we prove that  $\bar{C}_t^{\text{GK}}$  forms an interval almost surely for each  $t \in \{1, 2, \dots\}$  by showing that  $\mathcal{K}_t^{\text{GK}}(\mu)$  is a convex function.

**Step 1.  $\bar{C}_t^{\text{GK}}$  forms a CS.** Notice that by Lemma 5.C.1, we have that  $\mathcal{K}_t^+(\mu^*)$  and  $\mathcal{K}_t^-(\mu^*)$  defined in Theorem 5.B.2 are both test martingales. Consequently, their convex combination

$$\mathcal{K}_t^{\text{GK}}(\mu^*) := \theta \mathcal{K}_t^+(\mu^*) + (1 - \theta) \mathcal{K}_t^-(\mu^*)$$

is also a test martingale. Therefore,  $\bar{C}_t^{\text{GK}} := \{\mu \in [0, 1] : \mathcal{K}_t^{\text{GK}}(\mu) < 1/\alpha\}$  indeed forms a  $(1 - \alpha)$ -CS.

**Step 2.  $\bar{C}_t^{\text{GK}}$  is an interval almost surely.** We will now justify that  $\bar{C}_t^{\text{GK}}$  forms an interval by proving that  $\mathcal{K}_t^{\text{GK}}(\mu)$  is a convex function of  $\mu \in [0, 1]$  and noting that the sublevel sets of convex functions are themselves convex.

To ease notation, define the multiplicands  $g_i^+(\mu) := 1 + \lambda_{i,d}^+ \cdot (Z_i - \zeta_i(\mu))$  so that

$$\mathcal{K}_t^+(\mu) \equiv \prod_{i=1}^t g_i(\mu).$$

Rewriting  $g_i(\mu)$ , we have that

$$1 + \lambda_{i,d}^+ \cdot (Z_i - \zeta_i(\mu)) = 1 + \frac{d}{D+1} \cdot \left( \frac{Z_i}{r_i\mu + (1-r_i)/2} - 1 \right),$$

from which it is clear that each  $g_i(\mu)$  is (a) nonnegative, (b) nonincreasing, and (c) convex in  $\mu \in [0, 1]$ . Now, note that properties (a)–(c) are preserved under products (see Chapter 3), meaning

$$\mathcal{K}_t^+(\mu) \equiv \prod_{i=1}^t g_i(\mu)$$

also satisfies (a)–(c).

A similar argument goes through for  $\mathcal{K}_t^-(\mu)$ , except that this function is nonincreasing rather than nondecreasing, but it is nevertheless nonnegative and convex. Since convexity of functions is preserved under convex combinations, we have that

$$\mathcal{K}_t^{\text{GK}}(\mu) := \theta \mathcal{K}_t^+(\mu) + (1 - \theta) \mathcal{K}_t^-(\mu)$$

is a convex function of  $\mu \in [0, 1]$ .

Finally, observe that  $\bar{C}_t^{\text{GK}}$  is the  $(1/\alpha)$ -sublevel set of  $\mathcal{K}_t^{\text{GK}}(\mu)$  by definition, and the sublevel sets of convex functions are convex. Therefore,  $\bar{C}_t^{\text{GK}}$  is an interval almost surely. This completes the proof of Theorem 5.B.2.

□

### 5.C.5 Proof of Proposition 5.B.3

**Proposition 5.B.3 (NPRR-EB-CS).** *Given  $(Z_t)_{t=1}^\infty \sim \mathcal{Q}_{\mu^*}^\infty$  and let  $\hat{\mu}_t(\lambda_1^t)$  and  $\bar{B}_t^{\text{EB}}(\lambda_1^t)$  be as in Proposition 5.B.2:*

$$\begin{aligned}\hat{\mu}_t(\lambda_1^t) &:= \frac{\sum_{i=1}^t \lambda_i \cdot (Z_i - (1 - r_i)/2)}{\sum_{i=1}^t r_i \lambda_i}, \quad \text{and} \\ \bar{B}_t^{\text{EB}}(\lambda_1^t) &:= \frac{\log(1/\alpha) + \sum_{i=1}^t 4(Z_i - \hat{\zeta}_{i-1})^2 \psi_E(\lambda_i)}{\sum_{i=1}^t r_i \lambda_i}.\end{aligned}$$

where  $\psi_E(\lambda) := (-\log(1 - \lambda) - \lambda)/4$ . Then,

$$\bar{L}_t^{\text{EB}} := \hat{\mu}_t(\lambda_1^t) - \bar{B}_t^{\text{EB}}(\lambda_1^t) \tag{5.39}$$

forms a lower  $(1 - \alpha, (\varepsilon_t)_t)$ -LPCS for  $\mu^*$ , meaning  $\mathbb{P}(\forall t \geq 1, \mu^* \geq \bar{L}_t^{\text{EB}}) \geq 1 - \alpha$ .

*Proof.* The proof proceeds in two steps. First, we derive a sub-exponential NSM. Second and finally, we apply Ville's inequality to the NSM and invert it to obtain  $(\dot{L}_t^{\text{EB}})_{t=1}^\infty$ .

**Step 1: Deriving a sub-exponential nonnegative supermartingale.** Consider the process  $(M_t^{\text{EB}}(\mu^*))_{t=1}^\infty$  given by

$$M_t^{\text{EB}}(\mu^*) := \prod_{i=1}^t \exp \left\{ \lambda_i \cdot (Z_i - \zeta_i(\mu^*)) - 4(Z_i - \hat{\zeta}_{i-1}(\mu^*))^2 \psi_E(\lambda_i) \right\}, \tag{5.58}$$

and defined as  $M_0^{\text{EB}}(\mu^*) \equiv 1$ . Clearly,  $M_t^{\text{EB}} > 0$ , and hence in order to show that  $(M_t^{\text{EB}}(\mu^*))_{t=1}^\infty$  is an NSM, it suffices to show that  $\mathbb{E}(M_t^{\text{EB}}(\mu^*) \mid \mathcal{Z}_{t-1}) = M_{t-1}^{\text{EB}}(\mu^*)$  for each  $t \geq 1$ . To this end, we have that

$$\begin{aligned}\mathbb{E}(M_t^{\text{EB}}(\mu^*) \mid \mathcal{Z}_{t-1}) &= \mathbb{E} \left( \prod_{i=1}^t \exp \left\{ \lambda_i \cdot (Z_i - \zeta_i(\mu^*)) - 4(Z_i - \hat{\zeta}_{i-1}(\mu^*))^2 \psi_E(\lambda_i) \right\} \mid \mathcal{Z}_{t-1} \right) \\ &= M_{t-1}^{\text{EB}}(\mu^*) \underbrace{\mathbb{E} \left( \exp \left\{ \lambda_t \cdot (Z_t - \zeta_t(\mu^*)) - 4(Z_t - \hat{\zeta}_{t-1}(\mu^*))^2 \psi_E(\lambda_t) \right\} \mid \mathcal{Z}_{t-1} \right)}_{(*)},\end{aligned} \tag{5.59}$$

$$= M_{t-1}^{\text{EB}}(\mu^*) \underbrace{\mathbb{E} \left( \exp \left\{ \lambda_t \cdot (Z_t - \zeta_t(\mu^*)) - 4(Z_t - \hat{\zeta}_{t-1}(\mu^*))^2 \psi_E(\lambda_t) \right\} \mid \mathcal{Z}_{t-1} \right)}_{(*)}, \tag{5.60}$$

and hence it suffices to show that  $(*) \leq 1$ . Following the proof of Theorem 3.3.1 from Chapter 3, denote

$$Y_t := Z_t - \zeta_t(\mu^*) \quad \text{and} \quad \delta_t := \hat{\zeta}_t(\mu^*) - \zeta_t(\mu^*). \tag{5.61}$$

Note that  $\mathbb{E}(Y_t | \mathcal{Z}_{t-1}) = 0$ , and thus it suffices to prove that for any  $[0, 1]$ -bounded,  $\mathcal{Z}_{t-1}$ -measurable  $\lambda_t$ ,

$$\mathbb{E} \left( \exp \left\{ \lambda_t Y_t - 4(Y_t - \delta_{t-1})^2 \psi_E(\lambda_t) \right\} \mid \mathcal{F}_{t-1} \right) \leq 1.$$

Indeed, in the proof of Fan et al. [102, Proposition 4.1],  $\exp\{\xi\lambda - 4\xi^2\psi_E(\lambda)\} \leq 1 + \xi\lambda$  for any  $\lambda \in [0, 1]$  and  $\xi \geq -1$ . Setting  $\xi := Y_t - \delta_{t-1} = Z_t - \hat{\zeta}_{t-1}(\mu^*)$ ,

$$\begin{aligned} & \mathbb{E} \left( \exp \left\{ \lambda_t Y_t - 4(Y_t - \delta_{t-1})^2 \psi_E(\lambda_t) \right\} \mid \mathcal{Z}_{t-1} \right) \\ &= \mathbb{E} \left( \exp \left\{ \lambda_t(Y_t - \delta_{t-1}) - 4(Y_t - \delta_{t-1})^2 \psi_E(\lambda_t) \right\} \mid \mathcal{Z}_{t-1} \right) \exp(\lambda_t \delta_{t-1}) \\ &\leq \mathbb{E} (1 + (Y_t - \delta_{t-1})\lambda_t \mid \mathcal{Z}_{t-1}) \exp(\lambda_t \delta_{t-1}) \stackrel{(i)}{=} \mathbb{E} (1 - \delta_{t-1}\lambda_t \mid \mathcal{Z}_{t-1}) \exp(\lambda_t \delta_{t-1}) \stackrel{(ii)}{\leq} 1, \end{aligned}$$

where (i) follows from the fact that  $Y_t$  is conditionally mean zero, and (ii) follows from the inequality  $1 - x \leq \exp(-x)$  for all  $x \in \mathbb{R}$ . This completes the proof of Step 1.

**Step 2: Applying Ville's inequality and inverting.** Now that we have established that  $(M_t^{\text{EB}}(\mu^*))_{t=1}^\infty$  is an NSM, we have by Ville's inequality [264] that

$$\mathbb{P}(\exists t \geq 1 : M_t^{\text{EB}}(\mu^*) \geq 1/\alpha) \leq \alpha, \quad (5.62)$$

or equivalently,  $\mathbb{P}(\forall t \geq 1, M_t^{\text{EB}}(\mu^*) < 1/\alpha) \geq 1 - \alpha$ . Consequently, we have that with probability at least  $(1 - \alpha)$ ,

$$\begin{aligned} M_t^{\text{EB}}(\mu^*) &> 1/\alpha \prod_{i=1}^t \exp \left\{ \lambda_i \cdot (Z_i - \zeta_i(\mu^*)) - 4(Z_i - \hat{\zeta}_{i-1}(\mu^*))^2 \psi_E(\lambda_i) \right\} < 1/\alpha \\ & \quad (5.63) \end{aligned}$$

$$\iff \sum_{i=1}^t \lambda_i (Z_i - \zeta_i(\mu^*)) - \sum_{i=1}^t 4(Z_i - \hat{\zeta}_{i-1}(\mu^*))^2 \psi_E(\lambda_i) < \log(1/\alpha) \quad (5.64)$$

$$\iff \sum_{i=1}^t \lambda_i \zeta_i(\mu^*) > \sum_{i=1}^t \lambda_i Z_i - \sum_{i=1}^t 4(Z_i - \hat{\zeta}_{i-1}(\mu^*))^2 \psi_E(\lambda_i) - \log(1/\alpha) \quad (5.65)$$

$$\iff \sum_{i=1}^t \lambda_i \left( r_i \mu^* + \frac{1 - r_i}{2} \right) > \sum_{i=1}^t \lambda_i Z_i - \sum_{i=1}^t 4(Z_i - \hat{\zeta}_{i-1}(\mu^*))^2 \psi_E(\lambda_i) - \log(1/\alpha) \quad (5.66)$$

$$\iff \mu^* \sum_{i=1}^t \lambda_i r_i + \sum_{i=1}^t \lambda_i \frac{1 - r_i}{2} > \sum_{i=1}^t \lambda_i Z_i - \sum_{i=1}^t 4(Z_i - \hat{\zeta}_{i-1}(\mu^*))^2 \psi_E(\lambda_i) - \log(1/\alpha) \quad (5.67)$$

$$\iff \mu^* > \underbrace{\frac{\sum_{i=1}^t \lambda_i(Z_i - (1 - r_i)/2)}{\sum_{i=1}^t \lambda_i r_i}}_{\hat{\mu}_t(\lambda_1^t)} - \underbrace{\frac{\sum_{i=1}^t 4(Z_i - \hat{\zeta}_{i-1}(\mu^*))^2 \psi_E(\lambda_i) + \log(1/\alpha)}{\sum_{i=1}^t \lambda_i r_i}}_{\bar{B}_t(\lambda_1^t)}, \quad (5.68)$$

(5.69)

and hence  $\hat{\mu}_t(\lambda_1^t) - \bar{B}_t(\lambda_1^t)$  forms a lower  $(1 - \alpha)$ -CS. This completes the proof of Proposition 5.B.3.

□

### 5.C.6 Proof of Proposition 5.B.5

**Proposition 5.B.5.** Consider the same setup as Corollary 5.4.1, and let  $\Phi(\cdot)$  be the cumulative distribution function of a standard Gaussian. Define for any  $\beta > 0$ ,

$$\tilde{E}_t^\Delta := \frac{2}{\sqrt{t\beta^2 + 1}} \exp \left\{ \frac{2\beta^2(S_{t,0}^\Delta)^2}{t\beta^2 + 1} \right\} \Phi \left( \frac{2\beta S_{t,0}^\Delta}{\sqrt{t\beta^2 + 1}} \right),$$

where  $S_{t,0}^\Delta := \sum_{i=1}^t (\psi_i - (1 - r)/2) - tr \frac{1/(1-\pi)}{1/\pi+1/(1-\pi)}$  and  $\beta > 0$ . Then,  $\tilde{E}_t^\Delta$  forms an  $e$ -process and hence  $\tilde{p}_t^\Delta := 1/\tilde{E}_t^\Delta$  forms an anytime  $p$ -value, and  $\tilde{\phi}_t^\Delta := \mathbb{1}(\tilde{p}_t^\Delta \leq \alpha)$  forms a level- $\alpha$  sequential test for the weak null  $\tilde{\mathcal{H}}_0$ .

*Proof.* In order to show that  $\tilde{E}_t^\Delta$  is an  $e$ -process, it suffices to find an NSM that almost surely upper bounds  $\tilde{E}_t^\Delta$  for each  $t$  under the weak null  $\tilde{\mathcal{H}}_0$ :  $\tilde{\Delta}_t \leq 0$ . As such, the proof proceeds in three steps. First, we justify why the one-sided NSM (5.73) given by Proposition 5.B.4 is a nonincreasing function of  $\tilde{\mu}_t$ . Second, we adapt the aforementioned NSM to the A/B testing setup to obtain  $M_t^\Delta(\tilde{\Delta}_t)$  and note that it is a nonincreasing function of  $\tilde{\Delta}_t$ . Third and finally, we observe that  $E_t^\Delta := M_t^\Delta(0)$  is upper bounded by  $M_t^\Delta(\tilde{\Delta}_t)$  under the weak null, thus proving the desired result.

**Step 1: The one-sided NSM (5.73) is nonincreasing in  $\tilde{\mu}_t$ .** Recall the  $\lambda$ -indexed process from Step 1 of the proof of Proposition 5.B.4 given by

$$M_t(\lambda) := \prod_{i=1}^t \exp \left\{ \lambda(Z_i - \zeta(\mu_i)) - \lambda^2/8 \right\},$$

which can be rewritten as

$$M_t(\lambda) := \exp \left\{ S_t(\tilde{\mu}_t) - \lambda^2/8 \right\},$$

where  $S_t(\tilde{\mu}_t) := \sum_{i=1}^t (Z_i - (1-r)/2) - tr\tilde{\mu}_t$  and  $\tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \mu_t$ . In particular, notice that  $M_t(\lambda)$  is a nonincreasing function of  $\tilde{\mu}_t$  for any  $\lambda \geq 0$ , and hence we also have that

$$M_t(\lambda) f_{\rho^2}^+(\lambda)$$

is a nonincreasing function of  $\tilde{\mu}_t$  where  $f_{\rho^2}^+(\lambda)$  is the density of a folded Gaussian distribution given in (5.71), by virtue of  $f_{\rho^2}^+(\lambda)$  being everywhere nonnegative, and 0 for all  $\lambda < 0$ . Finally, by Step 2 of the proof of Proposition 5.B.4, we have that

$$\int_{\lambda} M_t(\lambda) f_{\rho^2}^+(\lambda) d\lambda \equiv \frac{2}{\sqrt{t\rho^2/4 + 1}} \exp \left\{ \frac{\rho^2 S_t(\tilde{\mu}_t)^2}{2(t\rho^2/4 + 1)} \right\} \Phi \left( \frac{\rho S_t(\tilde{\mu}_t)}{\sqrt{t\rho^2/4 + 1}} \right)$$

is nonincreasing in  $\tilde{\mu}_t$ , and forms an NSM when evaluated at the true means  $(\tilde{\mu}^*)_{t=1}^\infty 1$ .

**Step 2: Applying Step 1 to the A/B testing setup to yield  $M_t^\Delta(\tilde{\delta}_t)$ .** Adapting Step 1 to the setup described in Proposition 5.B.5, let  $\delta_1, \delta_2, \dots \in \mathbb{R}$  and let  $\tilde{\delta}_t := \sum_{i=1}^t \delta_i$ . Define the partial sum process,

$$S_t^\Delta(\tilde{\delta}_t) := \sum_{i=1}^t (\psi_i - (1-r)/2) - rt \frac{\tilde{\delta}_t + \frac{1}{1-\pi}}{\frac{1}{\pi} + \frac{1}{1-\pi}}$$

and the associated process,

$$M_t^\Delta(\tilde{\delta}_t) := \frac{2}{\sqrt{t\beta^2 + 1}} \exp \left\{ \frac{2\beta^2 S_t^\Delta(\tilde{\delta}_t)^2}{t\beta^2 + 1} \right\} \Phi \left( \frac{2\beta S_t^\Delta(\tilde{\delta}_t)}{\sqrt{t\beta^2 + 1}} \right),$$

where we have substituted  $\rho := 2\beta > 0$ . Notice that by construction,  $\psi_t$  is a  $[0, 1]$ -bounded random variable with mean  $r \frac{\tilde{\Delta}_t + 1/(1-\pi)}{1/\pi + 1/(1-\pi)} + (1-r)/2$ , so  $M_t^\Delta(\tilde{\Delta}_t)$  forms an NSM. We are now ready to invoke the main part of the proof.

**Step 3: The process  $\tilde{E}_t^\Delta$  is upper-bounded by the NSM  $M_t^\Delta(\tilde{\Delta}_t)$ .** Define the nonnegative process  $(\tilde{E}^\Delta)_{t=1}^\infty 0$  starting at one given by

$$\tilde{E}_t^\Delta := M_t^\Delta(0) \equiv \frac{2}{\sqrt{t\beta^2 + 1}} \exp \left\{ \frac{2\beta^2 S_t^\Delta(0)^2}{t\beta^2 + 1} \right\} \Phi \left( \frac{2\beta S_t^\Delta(0)}{\sqrt{t\beta^2 + 1}} \right).$$

By Steps 1 and 2, we have that  $\tilde{E}_t^\Delta \leq M_t^\Delta(\tilde{\Delta}_t)$  for any  $\tilde{\Delta}_t \leq 0$ , and since  $M_t^\Delta(\tilde{\Delta}_t)$  is an NSM, we have that  $(\tilde{E}^\Delta)_{t=1}^\infty 0$  forms an  $e$ -process for  $\mathcal{H}_0: \tilde{\Delta}_t \leq 0$ . This completes the proof.  $\square$

### 5.C.7 Proof of Proposition 5.B.4

**Proposition 5.B.4.** *Given the same setup as Theorem 5.3.4, define*

$$\tilde{B}_t := \sqrt{\frac{t\beta^2 + 1}{2(tr\beta)^2} \log \left( 1 + \frac{\sqrt{t\beta^2 + 1}}{2\alpha} \right)}. \quad (5.44)$$

Then,  $\tilde{L}_t := \hat{\mu}_t - \tilde{B}_t$  forms a lower  $(1 - \alpha, \varepsilon)$ -LPCS for  $\tilde{\mu}_t^* := \frac{1}{t} \sum_{i=1}^t \mu_i^*$ , meaning

$$\mathbb{P} \left( \forall t, \tilde{\mu}_t^* \geq \tilde{L}_t \right) \geq 1 - \alpha. \quad (5.45)$$

*Proof.* The proof begins similar to that of Theorem 5.3.4 but with a slightly modified mixing distribution, and proceeds in four steps. First, we derive a sub-Gaussian NSM indexed by a parameter  $\lambda \in \mathbb{R}$  identical to that of Theorem 5.3.4. Second, we mix this NSM over  $\lambda$  using a folded Gaussian density, and justify why the resulting process is also an NSM. Third, we derive an implicit lower CS for  $(\tilde{\mu}^*)_{t=1}^\infty$ . Fourth and finally, we compute a closed-form lower bound for the implicit CS.

**Step 1: Constructing the  $\lambda$ -indexed NSM.** This is exactly the same step as Step 1 in the proof of Theorem 5.3.4, found in Section 5.A.5. In summary, we have that for any  $\lambda \in \mathbb{R}$ ,

$$M_t(\lambda) := \prod_{i=1}^t \exp \left\{ \lambda(Z_i - \zeta(\mu_i^*)) - \lambda^2/8 \right\}, \quad (5.70)$$

with  $M_0(\lambda) \equiv 0$  forms an NSM with respect to the private filtration  $\mathcal{Z}$ .

**Step 2: Mixing over  $\lambda \in (0, \infty)$  to obtain a mixture NSM.** Let us now construct a one-sided sub-Gaussian mixture NSM. First, note that the mixture of an NSM with respect to a probability density is itself an NSM [217, 124] and is a simple consequence of Fubini's theorem. For our purposes, we will consider the density of a *folded Gaussian* distribution with location zero and scale  $\rho^2$ . In particular, if  $\Lambda \sim N(0, \rho^2)$ , let  $\Lambda_+ := |\Lambda|$  be the folded Gaussian. Then  $\Lambda_+$  has a probability density function  $f_{\rho^2}^+(\lambda)$  given by

$$f_{\rho^2}^+(\lambda) := \mathbb{1}(\lambda > 0) \frac{2}{\sqrt{2\pi\rho^2}} \exp \left\{ \frac{-\lambda^2}{2\rho^2} \right\}. \quad (5.71)$$

Note that  $f_{\rho^2}^+$  is simply the density of a mean-zero Gaussian with variance  $\rho^2$ , but truncated from below by zero, and multiplied by two to ensure that  $f_{\rho^2}^+(\lambda)$  integrates to one.

Then, since mixtures of NSMs are themselves NSMs, the process  $(M)_{t=1}^\infty$  given by

$$M_t := \int_{-\infty}^t M_s(\lambda) f_{\rho^2}^+(\lambda) d\lambda \quad (5.72)$$

is an NSM. We will now find a closed-form expression for  $M_t$ . Many of the techniques used to

derive the expression for  $M_t$  are identical to Step 2 of the proof of Theorem 5.3.4, but we repeat them here for completeness. To ease notation, define the partial sum  $S_t^* := \sum_{i=1}^t (Z_i - \zeta(\mu_i^*))$ . Writing out the definition of  $M_t$ , we have

$$\begin{aligned}
M_t &:= \int_{\lambda} \prod_{i=1}^t \exp \left\{ \lambda(Z_i - \zeta(\mu_i^*)) - \lambda^2/8 \right\} f_{\rho^2}^+(\lambda) d\lambda \\
&= \int_{\lambda} \exp \left\{ \lambda \underbrace{\sum_{i=1}^t (Z_i - \zeta(\mu_i^*))}_{S_t^*} - t\lambda^2/8 \right\} f_{\rho^2}^+(\lambda) d\lambda \\
&= \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \lambda S_t^* - t\lambda^2/8 \right\} \frac{2}{\sqrt{2\pi\rho^2}} \exp \left\{ \frac{-\lambda^2}{2\rho^2} \right\} d\lambda \\
&= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \lambda S_t^* - t\lambda^2/8 \right\} \exp \left\{ \frac{-\lambda^2}{2\rho^2} \right\} d\lambda \\
&= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \lambda S_t^* - \frac{\lambda^2(t\rho^2/4 + 1)}{2\rho^2} \right\} d\lambda \\
&= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \frac{-\lambda^2(t\rho^2/4 + 1) + 2\lambda\rho^2 S_t^*}{2\rho^2} \right\} d\lambda \\
&= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \underbrace{\exp \left\{ \frac{-a(\lambda^2 - \frac{b}{a}2\lambda)}{2\rho^2} \right\}}_{(*)} d\lambda,
\end{aligned}$$

where we have set  $a := t\rho^2/4 + 1$  and  $b := \rho^2 S_t^*$ . Completing the square in  $(*)$ , we have that

$$\begin{aligned}
\exp \left\{ \frac{-a(\lambda^2 - \frac{b}{a}2\lambda)}{2\rho^2} \right\} &= \exp \left\{ \frac{-\lambda^2 + 2\lambda \frac{b}{a} + (\frac{b}{a})^2 - (\frac{b}{a})^2}{2\rho^2/a} \right\} \\
&= \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} + \frac{a(b/a)^2}{2\rho^2} \right\} \\
&= \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} \right\} \exp \left\{ \frac{b^2}{2a\rho^2} \right\}.
\end{aligned}$$

Plugging this back into our derivation of  $M_t$  and multiplying the entire quantity by  $a^{-1/2}/a^{-1/2}$ ,

we have

$$\begin{aligned}
M_t &= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \underbrace{\exp \left\{ \frac{-a(\lambda^2 + \frac{b}{a}2\lambda)}{2\rho^2} \right\}}_{(*)} d\lambda \\
&= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} \right\} \exp \left\{ \frac{b^2}{2a\rho^2} \right\} d\lambda \\
&= \frac{2}{\sqrt{a}} \exp \left\{ \frac{b^2}{2a\rho^2} \right\} \underbrace{\int_{\lambda} \mathbb{1}(\lambda > 0) \frac{1}{\sqrt{2\pi\rho^2/a}} \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} \right\} d\lambda}_{(**)}
\end{aligned}$$

Now, notice that  $(**)$  =  $\mathbb{P}(N(b/a, \rho^2/a) \geq 0)$ , which can be rewritten as  $\Phi(b/\rho\sqrt{a})$ , where  $\Phi$  is the CDF of a standard Gaussian. Putting this all together and plugging in  $a = t\rho^2/4 + 1$  and  $b = \rho^2 S_t^*$ , we have the following expression for  $M_t$ ,

$$\begin{aligned}
M_t &= \frac{2}{\sqrt{a}} \exp \left\{ \frac{b^2}{2a\rho^2} \right\} \Phi \left( \frac{b}{\rho\sqrt{a}} \right) \\
&= \frac{2}{\sqrt{t\rho^2/4 + 1}} \exp \left\{ \frac{\rho^4(S_t^*)^2}{2(t\rho^2/4 + 1)\rho^2} \right\} \Phi \left( \frac{\rho^2 S_t^*}{\rho\sqrt{t\rho^2/4 + 1}} \right) \\
&= \frac{2}{\sqrt{t\rho^2/4 + 1}} \exp \left\{ \frac{\rho^2(S_t^*)^2}{2(t\rho^2/4 + 1)} \right\} \Phi \left( \frac{\rho S_t^*}{\sqrt{t\rho^2/4 + 1}} \right). \tag{5.73}
\end{aligned}$$

**Step 3: Deriving a  $(1 - \alpha)$ -lower CS  $(L'_t)_{t=1}^\infty$  for  $(\tilde{\mu}_t^*)_{t=1}^\infty$ .** Now that we have computed the mixture NSM  $(M)_{t=1}^\infty 0$ , we apply Ville's inequality to it and "invert" a family of processes — one of which is  $M_t$  — to obtain an *implicit* lower CS (we will further derive an *explicit* lower CS in Step 4).

First, let  $(\mu)_{t=1}^\infty 1$  be an arbitrary real-valued process — i.e. not necessarily equal to  $(\mu^*)_{t=1}^\infty 1$  — and define their running average  $\tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \mu_i$ . Define the partial sum process in terms of  $(\tilde{\mu})_{t=1}^\infty 1$ ,

$$S_t(\tilde{\mu}_t) := \sum_{i=1}^t Z_i - tr\tilde{\mu}_t - t(1-r)/2,$$

and the resulting nonnegative process,

$$M_t(\tilde{\mu}_t) := \frac{2}{\sqrt{t\rho^2/4 + 1}} \exp \left\{ \frac{\rho^2 S_t(\tilde{\mu}_t)^2}{2(t\rho^2/4 + 1)} \right\} \Phi \left( \frac{\rho S_t(\tilde{\mu}_t)}{\sqrt{t\rho^2/4 + 1}} \right). \tag{5.74}$$

Notice that if  $(\tilde{\mu})_{t=1}^\infty 1 = (\tilde{\mu}^*)_{t=1}^\infty 1$ , then  $S_t(\tilde{\mu}_t^*) = S_t^*$  and  $M_t(\tilde{\mu}_t^*) = M_t$  from Step 2. Impor-

tantly,  $(M_t(\tilde{\mu}_t^*))_{t=0}^\infty$  is an NSM. Indeed, by Ville's inequality, we have

$$\mathbb{P}(\exists t : M_t(\tilde{\mu}_t^*) \geq 1/\alpha) \leq \alpha. \quad (5.75)$$

We will now “invert” this family of processes to obtain an implicit lower boundary given by

$$L'_t := \inf\{\tilde{\mu}_t : M_t(\tilde{\mu}_t) < 1/\alpha\}, \quad (5.76)$$

and justify that  $(L'_t)_{t=1}^\infty$  is indeed a  $(1 - \alpha)$ -lower CS for  $\tilde{\mu}_t^*$ . Writing out the probability of miscoverage at any time  $t$ , we have

$$\begin{aligned} \mathbb{P}(\exists t : \tilde{\mu}_t^* < L'_t) &\equiv \mathbb{P}\left(\exists t : \tilde{\mu}_t^* < \inf_{\tilde{\mu}_t} \{M_t(\tilde{\mu}_t) < 1/\alpha\}\right) \\ &= \mathbb{P}(\exists t : M_t(\tilde{\mu}_t^*) \geq 1/\alpha) \\ &\leq \alpha, \end{aligned}$$

where the last line follows from Ville's inequality applied to  $(M_t(\tilde{\mu}_t^*))_{t=0}^\infty$ . In particular,  $L'_t$  forms a  $(1 - \alpha)$ -lower CS, meaning

$$\mathbb{P}(\forall t, \tilde{\mu}_t \geq L'_t) \geq 1 - \alpha.$$

**Step 4: Obtaining a closed-form lower bound  $(\tilde{L}_t)_{t=1}^\infty$  for  $(L'_t)_{t=1}^\infty$ .** The lower CS of Step 3 is simple to evaluate via line- or grid-searching, but a closed-form expression may be desirable in practice, and for this we can compute a sharp lower bound on  $L'_t$ .

First, take notice of two key facts:

- (a) When  $\tilde{\mu}_t = \frac{1}{tr} \sum_{i=1}^t Z_i - (1-r)/2r$ , we have that  $S_t(\tilde{\mu}_t) = 0$  and hence  $M_t(\tilde{\mu}_t) < 1$ , and
- (b)  $S_t(\tilde{\mu}_t)$  is a strictly decreasing function of  $\tilde{\mu}_t \leq \frac{1}{tr} \sum_{i=1}^t Z_i - (1-r)/2r$ , and hence so is  $M_t(\tilde{\mu}_t)$ .

Property (a) follows from the fact that  $\Phi(0) = 1/2$ , and that  $\sqrt{t\rho^2/4 + 1} > 1$  for any  $\rho > 0$ . Property (b) follows from property (a) combined with the definitions of  $S_t(\cdot)$ ,

$$S_t(\tilde{\mu}_t) := \sum_{i=1}^t Z_i - tr\tilde{\mu}_t - t(1-r)/2,$$

and of  $M_t(\cdot)$ ,

$$M_t(\tilde{\mu}_t) := \frac{2}{\sqrt{t\rho^2/4 + 1}} \exp\left\{\frac{\rho^2 S_t(\tilde{\mu}_t)^2}{2(t\rho^2/4 + 1)}\right\} \Phi\left(\frac{\rho S_t(\tilde{\mu}_t)}{\sqrt{t\rho^2/4 + 1}}\right),$$

In particular, by facts (a) and (b), the infimum in (5.76) must be attained when  $S_t(\cdot) \geq 0$ . That

is,

$$S_t(L'_t) \geq 0. \quad (5.77)$$

Using (5.77) combined with the inequality  $1 - \Phi(x) \leq \exp\{-x^2/2\}$  (a straightforward consequence of the Cramér-Chernoff technique), we have the following lower bound on  $M_t(L'_t)$ :

$$\begin{aligned} M_t(L'_t) &= \frac{2}{\sqrt{t\rho^2/4 + 1}} \exp\left\{\frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2/4 + 1)}\right\} \Phi\left(\frac{\rho S_t(L'_t)}{\sqrt{t\rho^2/4 + 1}}\right) \\ &\geq \frac{2}{\sqrt{t\rho^2/4 + 1}} \exp\left\{\frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2/4 + 1)}\right\} \left(1 - \exp\left\{-\frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2/4 + 1)}\right\}\right) \\ &= \frac{2}{\sqrt{t\rho^2/4 + 1}} \left(\exp\left\{\frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2/4 + 1)}\right\} - 1\right) \\ &=: \tilde{M}_t(L'_t). \end{aligned}$$

Finally, the above lower bound on  $M_t(L'_t)$  implies that  $1/\alpha \geq M_t(L'_t) \geq \tilde{M}_t(L'_t)$  which yields the following lower bound on  $L'_t$ :

$$\begin{aligned} \tilde{M}_t(L'_t) \leq 1/\alpha &\iff \frac{2}{\sqrt{t\rho^2/4 + 1}} \left(\exp\left\{\frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2/4 + 1)}\right\} - 1\right) \leq 1/\alpha \\ &\iff \exp\left\{\frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2/4 + 1)}\right\} \leq 1 + \frac{\sqrt{t\rho^2/4 + 1}}{2\alpha} \\ &\iff \frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2/4 + 1)} \leq \log\left(1 + \frac{\sqrt{t\rho^2/4 + 1}}{2\alpha}\right) \\ &\iff S_t(L'_t) \leq \sqrt{\frac{2(t\rho^2/4 + 1)}{\rho^2} \log\left(1 + \frac{\sqrt{t\rho^2/4 + 1}}{2\alpha}\right)} \\ &\iff \sum_{i=1}^t Z_i - trL'_t - t(1-r)/2 \leq \sqrt{\frac{2(t\rho^2/4 + 1)}{\rho^2} \log\left(1 + \frac{\sqrt{t\rho^2/4 + 1}}{2\alpha}\right)} \\ &\iff trL'_t \geq \sum_{i=1}^t Z_i - t(1-r)/2 - \sqrt{\frac{2(t\rho^2/4 + 1)}{\rho^2} \log\left(1 + \frac{\sqrt{t\rho^2/4 + 1}}{2\alpha}\right)} \\ &\iff L'_t \geq \frac{\sum_{i=1}^t (Z_i - (1-r)/2)}{tr} - \sqrt{\frac{2(t\rho^2/4 + 1)}{(tr\rho)^2} \log\left(1 + \frac{\sqrt{t\rho^2/4 + 1}}{2\alpha}\right)} \\ &\iff L'_t \geq \underbrace{\frac{\sum_{i=1}^t (Z_i - (1-r)/2)}{tr} - \sqrt{\frac{t\beta^2 + 1}{2(tr\beta)^2} \log\left(1 + \frac{\sqrt{t\beta^2 + 1}}{2\alpha}\right)}}_{\tilde{L}_t}, \end{aligned}$$

where we set  $\rho = 2\beta$  in the right-hand side of the final inequality. This precisely yields  $\tilde{L}_t$  as given in Proposition 5.B.4, completing the proof.  $\square$

### 5.C.8 Proof of Theorem 5.B.3

**Theorem 5.B.3** (Locally private adaptive A/B estimation). *Let  $S_t(\tilde{\Delta}'_t) := (\sum_{i=1}^t \theta_i - t\tilde{\Delta}'_t)/2$  for any  $\tilde{\Delta}'_t \in [0, 1]$  and define for any  $\rho > 0$ ,*

$$\tilde{M}_t^{\text{EB}}(\tilde{\Delta}'_t) := \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{V_t + \rho} \right) F_t(\tilde{\Delta}'_t), \quad (5.54)$$

where  $F_t(\tilde{\Delta}'_t) := {}_1F_1(1, V_t + \rho + 1, S_t(\tilde{\Delta}'_t) + V_t + \rho)$ , and  ${}_1F_1$  is Kummer's confluent hypergeometric function, and  $\Gamma(\cdot, \cdot)$  is the upper incomplete gamma function. Then, when evaluated at the true  $\tilde{\Delta}_t$ , we have that  $\tilde{M}_t^{\text{EB}}(\tilde{\Delta}_t)$  forms a nonnegative supermartingale. Consequently,

$$\tilde{L}_t^\Delta := \inf \left\{ \tilde{\Delta}_t \in [0, 1] : \tilde{M}_t^{\text{EB}}(\tilde{\Delta}_t) < 1/\alpha \right\} \quad (5.55)$$

forms a lower  $(1 - \alpha)$ -CS for the running ATE  $\tilde{\Delta}_t$ .

*Proof.* The proof proceeds in three steps and follows a similar form to the proof of Theorem 6.3.1 from Chapter 6. First, we show that a collection of processes (indexed by  $\lambda \in (0, 1)$ ) each form  $\mathcal{Q}_{\mu^*}^\infty$ -NSMs with respect to the private filtration  $\mathcal{Z}$ . Second, we mix over  $\lambda \in (0, 1)$  using the truncated gamma density to obtain the NSM obtained in Theorem 5.B.3. Third and finally, we “invert” the aforementioned NSM to obtain the LPCS of Theorem 5.B.3.

**Step 1: Showing that  $M_t^\lambda$  forms an NSM.** Consider the process  $(M_t^\lambda)_{t=1}^\infty$  given by

$$M_t^\lambda := \prod_{i=1}^t \exp \left\{ \lambda(\theta_i - \Delta_i) - \psi_E(\lambda)(\theta_i - \hat{\theta}_{i-1})^2/4 \right\}. \quad (5.78)$$

We will show that  $(M_t^\lambda)_{t=1}^\infty$  forms an NSM. First, note that  $M_0^\lambda \equiv 1$  by construction, and  $M_t^\lambda$  is always positive. It remains to show that  $M_t^\lambda$  forms a supermartingale. Writing out the conditional expectation of  $M_t^\lambda$  given  $\mathcal{Z}_{t-1}$ , we have that

$$\mathbb{E}(M_t^\lambda | \mathcal{Z}_{t-1}) = M_{t-1}^\lambda \underbrace{\mathbb{E} \left[ \exp \left\{ \lambda(\theta_t - \Delta_t) - \psi_E(\lambda)(\theta_t - \hat{\theta}_{t-1})^2/4 \right\} | \mathcal{Z}_{t-1} \right]}_{(\dagger)}, \quad (5.79)$$

and hence it suffices to prove that  $(\dagger) \leq 1$ . Denote for the sake of succinctness,

$$\xi_t := \theta_t - \Delta_t \quad \text{and} \quad \eta_t := \hat{\theta}_{t-1} - \Delta_t,$$

and note that  $\mathbb{E}(\xi_t | \mathcal{Z}_{t-1}) = 0$ . Using the proof of Fan et al. [102, Proposition 4.1], we have that  $\exp\{b\lambda - b^2\psi_E(\lambda)\} \leq 1 + b\lambda$  for any  $\lambda \in [0, 1]$  and  $b \geq -1$ . Noticing that  $(\theta_t - \hat{\theta}_{t-1})/2 \geq -1$

and setting  $b := (\xi_t - \eta_t)/2 = (\theta_t - \hat{\theta}_{t-1})/2$ , we have that

$$\begin{aligned} & \mathbb{E} \left[ \exp \left\{ \lambda \xi_t - \psi_E(\lambda) (\xi_t - \eta_t)^2/4 \right\} \mid \mathcal{Z}_{t-1} \right] \\ &= \mathbb{E} \left[ \exp \left\{ \lambda (\xi_t - \eta_t) - \psi_E(\lambda) (\xi_t - \eta_t)^2/4 \right\} \mid \mathcal{Z}_{t-1} \right] \exp(\lambda \eta_t) \\ &\leq \mathbb{E} [1 + (\xi_t - \eta_t)\lambda \mid \mathcal{Z}_{t-1}] \exp(\lambda \eta_t) \\ &= \mathbb{E} [1 - \eta_t \lambda \mid \mathcal{Z}_{t-1}] \exp(\lambda \eta_t) \leq 1, \end{aligned}$$

where the last line follows from the fact that  $\xi_t$  is conditionally mean zero and the inequality  $1 - x \leq \exp(-x)$  for all  $x \in \mathbb{R}$ . This completes Step 1 of the proof.

**Step 2: Mixing over  $\lambda$  using the truncated gamma density.** For any distribution  $F$  on  $(0, 1)$ ,

$$\widetilde{M}_t^{\text{EB}} := \int_{\lambda \in (0, 1)} M_t^\lambda dF(\lambda) \quad (5.80)$$

forms a test supermartingale by Fubini's theorem. In particular, we will use the truncated gamma density  $f(\lambda)$  given by

$$f(\lambda) = \frac{\rho^\rho e^{-\rho(1-\lambda)} (1-\lambda)^{\rho-1}}{\Gamma(\rho) - \Gamma(\rho, \rho)}, \quad (5.81)$$

as the mixing density. Writing out  $\widetilde{M}_t^{\text{EB}} \equiv \widetilde{M}_t^{\text{EB}}(\tilde{\Delta}_t)$  using  $dF(\lambda) := f(\lambda)d\lambda$ , and using the shorthand  $S_t \equiv S_t(\tilde{\Delta}_t)$ , we have

$$\begin{aligned} \widetilde{M}_t^{\text{EB}} &:= \int_0^1 \exp \{ \lambda S_t - V_t \psi_E(\lambda) \} f(\lambda) d\lambda \\ &= \int_0^1 \exp \{ \lambda S_t - V_t \psi_E(\lambda) \} \frac{\rho^\rho e^{-\rho(1-\lambda)} (1-\lambda)^{\rho-1}}{\Gamma(\rho) - \Gamma(\rho, \rho)} d\lambda \\ &= \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \int_0^1 \exp \{ \lambda (\rho + S_t + V_t) \} (1-\lambda)^{\rho-1} d\lambda \\ &= \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{V_t + \rho} \right) \left( \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 e^{zu} u^{a-1} (1-u)^{b-a-1} du \right) \Big|_{\substack{a=1 \\ b=V_t+\rho+1 \\ z=S_t+V_t+\rho}} \\ &= \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{V_t + \rho} \right) {}_1F_1(1, V_t + \rho + 1, S_t + V_t + \rho), \end{aligned}$$

which completes this step.

**Step 3: Applying Ville's inequality and inverting the mixture NSM.** Notice that  $\tilde{\Delta}_t < \tilde{L}_t^\Delta$  if and only if  $\widetilde{M}_t(\tilde{\Delta}_t) \geq 1/\alpha$ , and hence by Ville's inequality for nonnegative

supermartingales [264], we have that

$$\mathbb{P}(\exists t : \tilde{\Delta}_t < \tilde{L}_t^\Delta) = \mathbb{P}(\exists t : \tilde{M}_t^{\text{EB}} \geq 1/\alpha) \leq \alpha,$$

and hence  $\tilde{L}_t^\Delta$  forms a lower  $(1 - \alpha, \varepsilon)$ -LPCS for  $\tilde{\Delta}_t$ . This completes the proof.  $\square$

## 5.D A more detailed survey of related work

There is a rich literature exploring the intersection of statistics and differential privacy. Wasserman and Zhou [279] studied DP estimation rates under various metrics for several privacy mechanisms. Duchi et al. [90, 91, 92] articulated a new “locally private minimax rate” — the fastest worst-case estimation rate with respect to any estimator *and* LDP mechanism together — and studied them in several estimation problems. To accomplish this they provide locally private analogues of the famous Le Cam, Fano, and Assouad bounds that are common in the nonparametric minimax estimation literature. As an example application, Duchi et al. [90, 91, 92] derived minimax rates for nonparametric density estimation in Sobolev spaces, and showed that a naive application of the Laplace mechanism cannot achieve said rates, but a different carefully designed mechanism can. This study of density estimation was extended to Besov spaces by Butucea et al. [44]. Butucea and Issartel [43] employed this minimax framework to study the fundamental limits of private estimation of nonlinear functionals. Acharya et al. [6] extended the locally private Le Cam, Fano, and Assouad bounds to central DP setting. Duchi and Ruan [89] developed a framework akin to Duchi et al. [90, 91, 92] but from the *local* minimax point of view (here, “local” refers to the type of minimax rate considered, not “local DP”). Barnes et al. [20] studied the locally private Fisher information for parametric models. All of the aforementioned works are focused on estimation rates, rather than inference — i.e. confidence sets,  $p$ -values, and so on (though some asymptotic inference is implicitly possible in [89, 20]).

The first work to explicitly study inference under DP constraints was Vu and Slavkovic [269], who developed asymptotically valid private hypothesis tests for some parametric problems, including Bernoulli proportion estimation, and independence testing. Several works have furthered the study of differentially private goodness-of-fit and independence testing [277, 108, 31, 8, 3, 4, 5, 7]. Couch et al. [66] develop nonparametric tests of independence between categorical and continuous variables. Awan and Slavković [13] derive private uniformly most powerful nonasymptotic hypothesis tests in the binomial case. Karwa and Vadhan [145], Gaboardi et al. [109], and Joseph et al. [138] study nonasymptotic CIs for the mean of Gaussian random variables. Canonne et al. [48] study optimal private tests for simple nulls against simple alternatives. Covington et al. [71] derive nonasymptotic CIs for parameters of location-scale families. Ferrando et al. [103] introduces a parametric bootstrap method for deriving asymptotically valid CIs.

All of the previously mentioned works either consider goodness-of-fit testing, independence testing, or parametric problems where distributions are known up to some finite-dimensional parameter. Drechsler et al. [87] study nonparametric CIs for medians. To the best of our

knowledge, no prior work derives private nonasymptotic CIs (nor CSs) for means of bounded random variables.

Moreover, like most of the statistics literature, the prior work on private statistical inference is non-sequential, with the exception of Wang et al. [276] who study private analogues of Wald’s sequential probability ratio test [271] for simple hypotheses, and Berrett and Yu [32] who study locally private online changepoint detection. Another interesting paper is that of Jun and Orabona [139, Sections 7.1 & 7.2] – the authors study online convex optimization with sub-exponential noise, but also consider applications to martingale concentration inequalities (and thus CSs) as well as locally private stochastic subgradient descent.

# Chapter 6

## Anytime-valid off-policy inference for contextual bandits

### 6.1 Introduction

The so-called “contextual bandit” problem is an abstraction that can be used to describe several problem setups in statistics and machine learning [169, 172]. For example, it generalizes the multi-armed bandit problem by allowing for “contextual” side information, and it can be used to describe many adaptive sequential experiments. The general contextual bandit problem can be described informally as follows: an agent (such as a medical scientist in a clinical trial) views contextual information  $X_t \in \mathcal{X}$  for subject  $t$  (such as the clinical patient’s demographics, medical history, etc.), takes an action  $A_t \in \mathcal{A}$  (such as whether to administer a placebo, a low dose, or a high dose), and observes some reward  $R_t$  (such as whether their adverse symptoms have subsided). This description is made formal in the protocol for the generation of contextual bandit data in Algorithm 6.1. In the present chapter, no restrictions are placed on the dimensionality or structure of the context and action spaces  $\mathcal{X}$  and  $\mathcal{A}$  beyond them being measurable, but it is often helpful to think about  $\mathcal{X}$  as a  $d$ -dimensional Euclidean space, and  $\mathcal{A}$  as  $\{0, 1\}$  for binary treatments, or  $\mathbb{R}$  for different dosages, and so on. Indeed, while high-dimensional settings often pose certain challenges in contextual bandits (such as computational ones, or inflated variances), none of these issues will affect the *validity* of our statistical inference methods. Throughout, we will require that the rewards are real-valued and bounded in  $[0, 1]$  — a common assumption in contextual bandits [252, 144] — except for Section 6.4 where we relax the boundedness constraint.

There are two main objectives that one can study in the contextual bandit setup: (1) policy optimization, and (2) off-policy evaluation (OPE) [172, 169, 93, 94]. Here, a “policy”  $\pi(a | x)$  is simply a conditional distribution over actions, such as the probability that patient  $t$  should receive various treatments given their context  $X_t$ . Policy *optimization* is concerned with finding policies that achieve high cumulative rewards (typically measured through regret), while off-policy *evaluation* is concerned with asking the counterfactual question: “how would

we have done if we used some policy  $\pi$  instead of the policy that is currently collecting data?”. In this chapter, we study the latter with a particular focus on statistical inference in adaptive, sequential environments under nonparametric assumptions.

### 6.1.1 Off-policy inference, confidence intervals, and confidence sequences

By far the most common parameter of interest in the OPE problem is the expected reward  $\nu := \mathbb{E}_\pi(R)$  that would result from taking an action from the policy  $\pi$ . This expectation  $\nu$  is called the “value” of the policy  $\pi$ . While several estimators for  $\nu$  have been derived and refined over the years, many practical problems call for more than just a point estimate: we may also wish to quantify the uncertainty surrounding our estimates via statistical inference tools such as confidence intervals (CI). However, a major drawback of CIs is the fact that they are only valid at *fixed and prespecified* sample sizes, while contextual bandit data are collected in a sequential and adaptive fashion over time.

We lay out the assumed protocol for the generation of contextual bandit data in Algorithm 6.1, and in particular, all of our results will assume access to the output of this algorithm, namely the (potentially infinite) sequence of tuples  $(X_t, A_t, R_t)_{t=1}^T$  for  $T \in \mathbb{N} \cup \{\infty\}$ . As is standard in OPE, we will always assume that the policy  $\pi$  is (almost surely) absolutely continuous with respect to  $h_t$  so that  $\pi/h_t$  is almost surely finite (without which, estimation and inference are not possible in general). Indeed, this permits many bandit techniques and in principle allows for Thompson sampling since it always assigns positive probability to an action (note that it may not always be easy to compute the probability of taking that action via Thompson sampling, but if those probabilities can be computed, they can be used directly within our framework). However,  $(h_t)_{t=1}^\infty$  cannot be the result of Upper Confidence Bound (UCB)-style algorithms since they take conditionally deterministic actions given the past, violating the absolute continuity of  $\pi$  with respect to  $h_t$ .

In Algorithm 6.1, the term “exogenously time-varying” simply means that the context and reward distributions at time  $t$  can only depend on the past through  $X_1^{t-1} \equiv (X_1, \dots, X_{t-1})$ , and not on the actions taken (or rewards received). Formally, we allow for any joint distribution over  $(X_t, A_t, R_t)_{t=1}^\infty$  as long as

$$p_{R_t}(r | x, a, \mathcal{H}_{t-1}) = p_{R_t}(r | x, a, X_1^{t-1}) \quad \text{and} \quad p_{X_t}(x | \mathcal{H}_{t-1}) = p_{X_t}(x | X_1^{t-1}), \quad (6.1)$$

where  $\mathcal{H}_t$  is all of the history  $\sigma((X_i, A_i, R_i)_{i=1}^t)$  up until time  $t$ . This conditional independence requirement (6.1) includes as a special case more classical setups where  $X_t$  is independent of all else given  $A_t$ , such as those considered in Bibaut et al. [33] or iid scenarios [144], but is strictly more general, since, for example,  $(X_t)_{t=1}^\infty$  can be a highly dependent time-series. However, we do not go as far as to consider the adversarial setting that is sometimes studied in the context of regret minimization. We impose this conditional independence requirement since otherwise, the interpretation of  $\mathbb{E}_\pi(R_t | \mathcal{H}_{t-1})$  changes depending on which sequence of actions were played by the logging policy. Making matters more concrete, the conditional off-policy value

$\mathbb{E}_\pi(R_t | \mathcal{H}_{t-1})$  at time  $t$  is given by

$$\nu_t := \mathbb{E}_\pi(R_t | \mathcal{H}_{t-1}) \equiv \int_{\mathcal{X} \times \mathcal{A} \times \mathbb{R}} r \cdot p_{R_t}(r | a, x, \mathcal{H}_{t-1}) \pi(a | x) p_{X_t}(x | \mathcal{H}_{t-1}) dx da dr \quad (6.2)$$

$$= \int_{\mathcal{X} \times \mathcal{A} \times \mathbb{R}} r \cdot p_{R_t}(r | a, x, X_1^{t-1}) \pi(a | x) p_{X_t}(x | X_1^{t-1}) dx da dr, \quad (6.3)$$

and the equality (6.3) follows from (6.1). Notice that (6.2) could in principle depend on the logging policies and actions played, but (6.3) does not. Despite imposing the conditional independence assumption (6.1), the integral (6.2) is still a perfectly well-defined functional, and if (6.1) is not satisfied, then our CSs will still cover a quantity in terms of this functional. However, its interpretation would no longer be counterfactual with respect to the entire sequence of actions (only conditional on the past).

While most prior work on OPE in contextual bandits is not written *causally* in terms of potential outcomes (e.g. [252, 144, 33, 51, 128]), it is nevertheless possible to write down a causal target  $\nu_t^*$  (i.e. a functional of the potential outcome distribution) and show that it is equal to  $\nu_t$  under certain causal identification assumptions. These assumptions resemble the familiar *consistency*, *exchangeability*, and *positivity* conditions that are ubiquitous in the treatment effect estimation literature. Moreover, there is a close relationship between OPE and the estimation of so-called *stochastic interventions* in causal inference; indeed, they can essentially be seen as equivalent but with slightly different emphases and setups. However, given that neither the potential outcomes view nor the stochastic intervention interpretation of OPE are typically emphasized in the contextual bandit literature (with the exception of Zhan et al. [292], who use potential outcomes throughout), we leave this discussion to Section 6.B.

Algorithm 6.1: Protocol for the generation of contextual bandit data

```

1: // Here,  $T \in \mathbb{N} \cup \{\infty\}$ .
2: for  $t = 1, 2, \dots, T$  do
3:   // The agent selects a policy  $h_t$  based on the history  $\mathcal{H}_{t-1} \equiv \sigma((X_i, A_i, R_i)_{i=1}^{t-1})$ .
4:    $h_t \in \mathcal{H}_{t-1}$ 
5:   // The environment draws a context from an (exogenously time-varying) distribution.
6:    $X_t \sim p_{X_t}(\cdot)$ 
7:   // The agent plays a random action drawn from the selected policy.
8:    $A_t \sim h_t(\cdot | X_t)$ 
9:   // The environment draws a reward from an (exogenously time-varying) distribution
    // based on the action and context.
10:   $R_t \sim p_{R_t}(\cdot | A_t, X_t)$ 
11: end for
12: // Return a (potentially infinite) sequence of contextual bandit data.
13: return  $(X_t, A_t, R_t)_{t=1}^T$ 
```

To illustrate the shortcomings of CIs for OPE, suppose we run a contextual bandit algorithm

and want to see whether  $\pi$  is better than the current state-of-the-art policy  $h$  – e.g. whether  $\mathbb{E}_\pi(R) > \mathbb{E}_h(R)$ . (Here, we are implicitly assuming that  $\mathbb{E}_{\pi'}(R) = \mathbb{E}_{\pi'}(R_t | \mathcal{H}_{t-1})$  for any policy  $\pi'$  for the sake of illustration.) Suppose we compute a CI for the value of  $\pi$  based on  $n$  samples (for some prespecified  $n$ ), and while  $\pi$  seems promising, the CI turns out to be inconclusive (the CI for  $\mathbb{E}_\pi(R)$  includes  $\mathbb{E}_h(R)$  if the latter is known, or the two CIs overlap if the latter is unknown). It is tempting to collect more data, for a total of  $n'$  points, to see if the result is now conclusive; however the resulting sample size  $n'$  is now a *data-dependent* quantity, rendering the CI invalid. (This could happen more than once.)

Fortunately, there exist statistical tools that permit adaptive stopping in these types of sequential data collection problems: *confidence sequences* (CSs [76, 168]). A CS is a sequence of confidence intervals, valid at all sample sizes uniformly (and hence at arbitrary stopping times). Importantly for the aforementioned OPE setup, CSs allow practitioners to collect additional data and continuously monitor it, so that the resulting CI is indeed valid at the data-dependent stopped sample size  $n'$ . More formally, we say that a sequence of intervals  $[L_t, U_t]_{t=1}^\infty$  is a CS for the parameter  $\theta \in \mathbb{R}$  if

$$\mathbb{P}(\forall t \in \mathbb{N}, \theta \in [L_t, U_t]) \geq 1 - \alpha, \text{ or equivalently, } \mathbb{P}(\exists t \in \mathbb{N} : \theta \notin [L_t, U_t]) \leq \alpha. \quad (6.4)$$

Contrast (6.4) above with the definition of a CI which states that  $\forall n \in \mathbb{N}$ ,  $\mathbb{P}(\theta \in [L_n, U_n]) \geq 1 - \alpha$ , so that the “ $\forall n$ ” is outside the probability  $\mathbb{P}(\cdot)$  rather than inside. A powerful consequence of (6.4) is that  $[L_\tau, U_\tau]$  is a valid CI for *any stopping time*  $\tau$ . In fact,  $[L_\tau, U_\tau]$  being a valid CI is not just an implication of (6.4) but the two statements are actually equivalent; see Howard et al. [125, Lemma 3].

The consequence for the OPE practitioner is that they can continuously update and monitor a CS for the value of  $\pi$  *while* the contextual bandit algorithm is running, and deploy  $\pi$  as soon as they are confident that it is better than the current state-of-the-art  $h$ . Karampatziakis et al. [144] refer to this adaptive policy switching as “gated deployment”, and we will return to this motivating example through the chapter. Let us now lay out five desiderata that we want all of our CSs to satisfy.

### 6.1.2 Desiderata for anytime-valid off-policy inference

Throughout this chapter, we will derive methods for off-policy evaluation and inference in a variety of settings – including fixed policies (Section 6.2), time-varying policies (Section 6.3), and for entire cumulative distribution functions (Section 6.4). However, what all of these approaches will have in common is that they will satisfy five desirable properties which we enumerate here.

1. **Nonasymptotic:** Our confidence sets will satisfy *exact* coverage guarantees for *any* sample size, unlike procedures based on the central limit theorem which only satisfy *approximate* guarantees for large samples.<sup>1</sup>

---

<sup>1</sup>Note that nonasymptotic procedures may be more conservative than asymptotic ones as they satisfy more rigorous coverage (similarly, type-I error) guarantees. Whether one should sacrifice stronger guarantees for tightness

2. **Nonparametric:** We will not make any parametric assumptions on the distribution of the contexts, policies, or rewards.
3. **Time-uniform / anytime-valid:** Our confidence sets will be *uniformly* valid for all sample sizes, and permit off-policy inference at arbitrary data-dependent stopping times.
4. **Adaptive data collection (via online learning):** All of our off-policy inference methods will allow for the sequence of logging policies  $(h_t)_{t=1}^{\infty}$  to be predictable (i.e.  $h_t$  can depend on  $\mathcal{H}_{t-1}$ ). In particular  $(h_t)_{t=1}^{\infty}$  can be the result of an online learning algorithm.
5. **Unknown and unbounded  $w_{\max}$ :** In all of our algorithms, the maximal importance weight

$$w_{\max} := \underset{t \in \mathbb{N}, a \in \mathcal{A}, x \in \mathcal{X}}{\text{ess sup}} \frac{\pi(a \mid x)}{h_t(a \mid x)}$$

can be unknown, and need not be uniformly bounded (i.e.  $w_{\max}$  can be infinite). Note that we do require that importance weights  $\pi(a \mid x)/h_t(a \mid x)$  themselves are finite for each  $(t, a, x)$ , but their essential *supremum* need not be. Perhaps surprisingly, even if  $w_{\max}$  is infinite, it is still possible for our CSs to shrink to zero-width since they depend only on *empirical variances*. As an illustrative example, see Proposition 6.3.1 for a closed-form CS whose width can shrink at a rate of  $\sqrt{\log \log t/t}$  as long as the importance-weighted rewards are well-behaved (e.g. in the iid setting, if they have finite second moments).

In addition to the above, we will design procedures so that they have strong empirical performance and are straightforward to compute. While some of these desiderata are quite intuitive and common in statistical inference (such as nonasymptotic, nonparametric, and time-uniform validity, so as to avoid relying on large sample theory, unrealistic parametric assumptions, or prespecified sample sizes), desiderata 4 and 5 are more specific to OPE and have not been satisfied in several prior works as we outline in Sections 6.2, 6.3, and 6.4. Given their central importance to our work, let us elaborate on them here.

**Why allow for logging policies to be predictable?** For the purpose of policy optimization, contextual bandit algorithms are tasked with balancing exploration and exploitation: simultaneously figuring out which policies will yield high rewards (at the expense of trying out suboptimal policies) and playing actions from the policies that have proven effective so far. On the other hand, in adaptive sequential trials, an experimenter might aim to balance context distributions between treatment arms (such as via Efron’s biased coin design [98]) or to adaptively find the treatment policy that yields the most efficient treatment effect estimators [148]. In both cases, the logging policies  $(h_t)_{t=1}^{\infty}$  are not only changing over time, but adaptively so based on previous observations. We strive to design procedures that permit inference in precisely these types of adaptive data collection regimes, despite most prior works on off-policy inference for contextual bandits having assumed that there is a fixed, prespecified logging policy [252, 144, 51, 128]. Of course, if a CS or CI is valid under adaptive data collection, they are also valid when fixed logging policies are used instead.

---

comes down to philosophical preference. For the purposes of this chapter, we focus solely on nonasymptotics.

**Why not rely on knowledge of  $w_{\max}$ ?** Related to the previous discussion, it may not be known *a priori* how the range of the predictable logging policies will evolve over time. Moreover, one can imagine a situation where  $\sup_{a,x} \pi(a | x)/h_t(a | x) \rightarrow \infty$ , even if every individual importance weight  $\pi/h_t$  is finite. In such cases, having CSs be agnostic to the value of  $w_{\max}$  is essential. However, even if  $w_{\max}$  is known, it may be preferable to design CSs that do not depend on this worst-case value. Suppose for the sake of illustration that a logging policy  $h$  assigns a novel treatment (denoted by  $a = 1$ ) with probability 1/5 and a placebo (denoted by  $a = 0$ ) with probability 4/5 for most subjects, except for a small but high-risk subpopulation, who receive the novel treatment with probability 1/1000. To estimate the expected reward of the novel treatment, note that the importance weight for subject  $t$  will take on the value  $w_t := 1/h(A_t | X_t)$  for treatment  $A_t \in \{0, 1\}$  and context  $X_t \in \mathcal{X}$ . Despite the fact that most of the importance weights are only 5, and hence most importance-weighted pseudo-outcomes  $w_t R_t$  will take values in  $[0, 5]$ , the worst-case  $w_{\max}$  is much larger at 1000. Consequently, we should expect a CS that scales with  $w_{\max} = 1000$  to be much wider than one that only scales with a quantity like an empirical variance. For these reasons we prefer procedures that depend on an empirical variance term (defined later) rather than the worst-case importance weight  $w_{\max}$ .

### 6.1.3 Outline and contributions

Our fundamental contribution is in the derivation of CSs for various off-policy parameters, including fixed policy values, time-varying policy values, and quantiles of the off-policy reward distribution. We begin in Section 6.2 with the most common formulation of the OPE problem: estimating the value  $\nu$  of a target policy  $\pi$ . Theorem 6.2.1 presents time-uniform CSs for  $\nu$ , a result that generalizes and improves upon the current state-of-the-art CSs for  $\nu$  by Karampatziakis et al. [144]. In Section 6.3, we consider the more challenging problem of estimating a time-varying average policy value  $\nu_t$ , where the distribution of off-policy rewards can change over time in an arbitrary and unknown fashion. In Section 6.4, we derive CSs for quantiles of the off-policy reward distribution, and in particular, Theorem 6.4.1 presents a confidence band for the entire cumulative distribution function (CDF) that is both uniformly valid in time *and* in the quantiles. For the results of Sections 6.3 and 6.4, no other solutions to this problem exist in the literature, to the best of our knowledge. Finally, in Section 6.5, we summarize our results and describe some natural extensions and implications of them, namely false discovery rate control under arbitrary dependence when evaluating several policies, and differentially private off-policy inference.

### 6.1.4 Related work

Throughout the chapter, we will draw detailed comparisons to work that is most closely related to ours, i.e. papers that are broadly concerned with estimating policy values and/or the off-policy CDF from contextual bandit data in a model-free setting — here we are using the term “model-free” to mean that no restrictions are placed on the functional form between the rewards  $R$  and the actions  $A$  nor on the covariates  $X$ . Specifically, Table 6.1 and the preceding text provides a (selective) property-by-property comparison to the directly related

works of Karampatziakis et al. [144], Bibaut et al. [33], Zhan et al. [292], and Howard et al. [125], and Table 6.2 provides a similar comparison to the works of Howard and Ramdas [123], Chandak et al. [51], and Huang et al. [128]. However, there are several other works that focus on *fixed-n* (i.e. not time-uniform) and *asymptotic* statistical inference from adaptively collected data (e.g. in the form of multi-armed bandits, contextual bandits, or more general reinforcement learning). For example, Ramprasad et al. [212] develop a bootstrap procedure for estimating policy values under Markov noise with temporal difference learning algorithms, Dimakopoulou et al. [84] perform adaptive inference in the multi-armed bandit setting, Hadad et al. [115] provide asymptotic confidence intervals for treatment effects in adaptive experiments, and Zhang et al. [294] provide distribution-uniform asymptotic procedures for M-estimation from contextual bandit data. Other works that consider more model-based approaches include Zhang et al. [293], Khamaru et al. [157], Shen et al. [241], and Chen et al. [55].

### 6.1.5 Notation: supermartingales, filtrations, and stopping times

Since all of our results will rely on the analysis of nonnegative (super)martingales, predictable processes, stopping times, and so on, it is worth defining some of these terms before proceeding. Consider a universe of distributions  $\Pi$  on a filtered probability space  $(\Omega, \mathcal{F})$ . A single draw from any distribution  $P \in \Pi$  results in a sequence  $Z_1, Z_2, \dots$  of potentially dependent observations. (In the context of this chapter,  $Z_t$  may represent  $(X_t, A_t, R_t)$ , for example, and the distribution  $P$  may be induced by the policy, and not specified in advance.) If  $Z_1, Z_2, \dots$  are independent and identically distributed (iid), we will explicitly say so, but in general we eschew iid assumptions in this chapter.

As is common in the statistics literature, we will use upper-case letters like  $Z$  to refer to random variables and lower-case letters  $z$  to refer to non-stochastic values in the same space that  $Z$  takes values. Let  $Z_1^t$  denote the tuple  $(Z_1, \dots, Z_t)$  and let  $\mathcal{H} \equiv (\mathcal{H}_t)_{t=1}^\infty$  by default represent the data (or “canonical”) filtration, meaning that  $\mathcal{H}_t = \sigma(Z_1^t)$ .

A sequence of random variables  $Y \equiv (Y_t)_{t=1}^\infty$  is called a *process* if it is adapted to  $\mathcal{H}$ , that is if  $Y_t$  is measurable with respect to  $\mathcal{H}_t$  for every  $t$ . A process  $Y$  is *predictable* if  $Y_t$  is measurable with respect to  $\mathcal{H}_{t-1}$  — informally “ $Y_t$  only depends on the past”. A process  $M$  is a *martingale* for  $P$  with respect to  $\mathcal{H}$  if

$$\mathbb{E}_P[M_t | \mathcal{H}_{t-1}] = M_{t-1} \tag{6.5}$$

for all  $t \geq 1$ .  $M$  is a *supermartingale* for  $P$  if it satisfies (6.5) with “=” relaxed to “ $\leq$ ”. A (super)martingale is called a *test (super)martingale* if it is nonnegative and  $M_0 = 1$ . A process  $M$  is called a test (super)martingale for  $\mathcal{P} \subset \Pi$  if it is a test (super)martingale for every  $P \in \mathcal{P}$ .

Throughout, if an expectation  $\mathbb{E}$  operator is used without a subscript  $P$ , or if a **boldface**  $\mathbb{P}$  is used to denote a probability, these are always referring to the distribution of  $(X_t, A_t, R_t)_{t \geq 1}$  induced by Algorithm 6.1 and the logging policies  $(h_t)_{t=1}^\infty$ .

An  $\mathcal{H}$ -stopping time  $\tau$  is a  $\mathbb{N} \cup \{\infty\}$ -valued random variable such that  $\{\tau \leq t\} \in \mathcal{H}_t$  for each  $t \geq 0$ . Informally and in the context of this chapter, a stopping time can be thought of as a sample size that was chosen based on all of the information  $\mathcal{H}_t$  up until time  $t$ .

## 6.2 Warmup: Off-policy inference for constant policy values

This section deals with the case where  $\nu_t$  from (6.2) does not depend on  $t$ , meaning that it is constant as a function of time. We handle the time-varying case in the next section.

We begin by extending a result of Karampatziakis et al. [144, Section 5.2] which applied in the iid setting, meaning that the logging policy  $h$  is fixed and the contexts and rewards are assumed to be iid. Their paper derives several CSs for the value  $\nu := \mathbb{E}_{A \sim \pi}(R)$  of the policy  $\pi$  for  $[0, 1]$ -bounded rewards  $R$ , but some of their CSs require knowledge of  $w_{\max}$ , which we would like to avoid as per our desiderata in Section 6.1.2. However, their so-called “scalar betting” approach in [144, Section 5.2] makes use of importance-weighted random variables and does not depend on knowing  $w_{\max}$ . To elaborate, let  $w_t$  be the importance weight for the target policy  $\pi$  versus the logging policy  $h$  given by

$$w_t := \frac{\pi(A_t | X_t)}{h(A_t | X_t)}, \quad (6.6)$$

and let  $\phi_t^{(\text{IW}-\ell)} := w_t R_t$  and  $\phi_t^{(\text{IW}-u)} := w_t(1 - R_t)$  be importance-weighted rewards that will be used to construct lower and upper bounds respectively. Note that  $\phi_t^{(\text{IW}-\ell)}$  is ubiquitous in the bandit and causal inference literatures, and the authors were not concerned with deriving *new estimators*, but rather *new confidence sequences* using existing estimators. While  $w_t \equiv w_t(X_t, A_t)$  does depend on both  $A_t$  and  $X_t$ , we leave the dependence on them implicit going forward to reduce notational clutter.

**Proposition 6.2.1** (Scalar betting off-policy CS [144]). *Suppose  $(X_t, A_t, R_t)_{t=1}^\infty$  are iid with  $[0, 1]$ -valued rewards  $(R_t)_{t=1}^\infty$ , and the logging policy  $h$  is fixed. For each  $\nu' \in [0, 1]$ , let  $(\lambda_t^L(\nu'))_{t=1}^\infty$  be any  $[0, 1/\nu']$ -valued predictable sequence. Then,*

$$L_t^{\text{IW}} := \inf \left\{ \nu' \in [0, 1] : \prod_{i=1}^t \left( 1 + \lambda_i^L(\nu') \cdot (\phi_i^{(\text{IW}-\ell)} - \nu') \right) < \frac{1}{\alpha} \right\} \quad (6.7)$$

forms a lower  $(1 - \alpha)$ -CS for  $\nu$ , meaning  $\mathbb{P}(\forall t \in \mathbb{N}, \nu \geq L_t^{\text{IW}}) \geq 1 - \alpha$ . Similarly, for any  $[0, 1/(1 - \nu')]$ -valued predictable sequence  $(\lambda_t^U(\nu'))_{t=1}^\infty$ .

$$U_t^{\text{IW}} := 1 - \inf \left\{ 1 - \nu' \in [0, 1] : \prod_{i=1}^t \left[ 1 + \lambda_i^U(\nu') \cdot (\phi_i^{(\text{IW}-u)} - (1 - \nu')) \right] < \frac{1}{\alpha} \right\} \quad (6.8)$$

forms an upper  $(1 - \alpha)$ -CS for  $\nu$ , meaning  $\mathbb{P}(\forall t \in \mathbb{N}, \nu \leq U_t^{\text{IW}}) \geq 1 - \alpha$ . A two-sided CS can be formed using  $[L_t^{\text{IW}}, U_t^{\text{IW}}]_{t=1}^\infty$  combined with a union bound.

The above CS  $[L_t^{\text{IW}}, U_t^{\text{IW}}]_{t=1}^\infty$  due to Karampatziakis et al. [144] has a number of desirable properties. Namely, it satisfies the first four of five desiderata in Section 6.1.2, meaning it is a nonasymptotic, nonparametric, time-uniform confidence sequence that does not require knowledge of  $w_{\max}$ . Note that while infima appear in the definitions of  $L_t^{\text{IW}}$  and  $U_t^{\text{IW}}$  they are straightforward to compute (e.g. via line or grid search) when the product is quasiconvex in  $\nu' \in$

$[0, 1]$  which is often the case as we discuss in Section 6.2.2. The idea behind Proposition 6.2.1 is to show that the product inside the above infima are nonnegative martingales when  $\nu' = \nu$  and then apply Ville's inequality to it [264]. Our main results in the coming sections use similar techniques albeit with very different (super)martingales tailored to different problem settings.

We also wish to highlight that indeed,  $[L_t^{\text{IW}}, U_t^{\text{IW}}]_{t=1}^{\infty}$  forms a valid  $(1 - \alpha)$ -CS *regardless* of how the sequences  $(\lambda_t^L(\nu'))_t$  and  $(\lambda_t^U(\nu'))_t$  are chosen. Such phenomena are common in martingale-based statistical procedures such as in Chapter 3 (see also the review paper of Ramdas et al. [211]) and will be seen in several of the results to follow. We will discuss some guiding principles for how to choose these sequences in Remark 11.

### 6.2.1 Tighter confidence sequences via doubly robust pseudo-outcomes

Here, we generalize and improve upon Proposition 6.2.1 in three ways. First, we show that the logging policy  $h$  can be replaced by a *sequence* of predictable logging policies  $(h_t)_{t=1}^{\infty}$  so that  $h_t$  can be built from the entire history  $\mathcal{H}_{t-1}$  up until time  $t - 1$  (and in particular,  $(h_t)_{t=1}^{\infty}$  can be the result of an online learning algorithm), so that the importance weight  $w_t$  at time  $t$  is given by

$$w_t := \frac{\pi(A_t | X_t)}{h_t(A_t | X_t)}. \quad (6.9)$$

Second, we show how the importance-weighted pseudo-outcomes  $(\phi_t)_{t=1}^{\infty}$  can be made doubly robust in the sense of Dudík et al. [93, 94]. Indeed, define the lower and upper doubly robust pseudo-outcomes  $(\phi_t^{(\text{DR}-\ell)})_{t=1}^{\infty}$  and  $(\phi_t^{(\text{DR}-u)})_{t=1}^{\infty}$  given by

$$\phi_t^{(\text{DR}-\ell)} := w_t \cdot \left( R_t - \left[ \hat{r}_t(X_t; A_t) \wedge \frac{k_t}{w_t} \right] \right) + \mathbb{E}_{a \sim \pi(\cdot | X_t)} \left( \hat{r}_t(X_t; a) \wedge \frac{k_t}{w_t} \right), \quad (6.10)$$

$$\phi_t^{(\text{DR}-u)} := w_t \cdot \left( 1 - R_t - \left[ (1 - \hat{r}_t(X_t; A_t)) \wedge \frac{k_t}{w_t} \right] \right) + \mathbb{E}_{a \sim \pi(\cdot | X_t)} \left( [1 - \hat{r}_t(X_t; a)] \wedge \frac{k_t}{w_t} \right), \quad (6.11)$$

where  $\hat{r}_t(X_t; A_t)$  is any  $[0, 1]$ -valued predictor of  $R_t$  built from  $\mathcal{H}_{t-1}$  and  $k_t$  is a  $\mathbb{R}_{\geq 0} \cup \{\infty\}$ -valued tuning parameter built from  $\mathcal{H}_{t-1}$  that determines how “doubly robust”  $\phi_t^{\text{DR}}$  should be. Note that  $\phi_t^{(\text{DR}-\ell)}$  and  $\phi_t^{(\text{DR}-u)}$  are both at least  $-k_t$  by construction, and have conditional means of  $\nu$  and  $1 - \nu$  respectively. Note that the phrase “doubly robust” is sometimes used to refer to *properties* of estimators (e.g. that their bias is second order and depends only on products of nuisance errors in observational studies where importance weights are unknown [155]) and sometimes to refer to *types of* estimators that enjoy variance-reduction without compromising validity in experiments where importance weights are known. We are using this phrase in the second sense following the conventions of Dudík et al. [93, 94].

Similar to the discussion surrounding Proposition 6.2.1, the doubly robust pseudo-outcomes in (6.10) are ubiquitous in the causal inference and bandit literatures [225, 93, 94, 256] with the minor tweak that we are truncating the reward predictor. Note that we are *not* doing this for the purposes of deriving better estimators — instead, we are doing so for the purposes of sharp

concentration of measure in the pursuit of tighter CSs.

Setting  $k_1 = k_2 = \dots = 0$  recovers the IW outcomes exactly, while setting  $k_1 = k_2 = \dots = \infty$  recovers the classic doubly robust outcomes [93, 94] (this could also be achieved by setting  $k_1 = k_2 = \dots = w_{\max}$ , provided  $w_{\max}$  is finite and known). We discuss the need to truncate  $\hat{r}_t$  in Remark 12, but the motivation to include a reward predictor at all is to reduce the variance of  $\phi_t^{(\text{DR}-\ell)}$  and  $\phi_t^{(\text{DR}-u)}$  if  $R_t$  can be well-predicted by  $\hat{r}_t$ , a well-known phenomenon in doubly robust estimation [225, 258, 56]. Indeed, we find that the resulting CSs are able to adapt to this reduced variance accordingly for large  $t$  (see Figure 6.1 for an illustration).

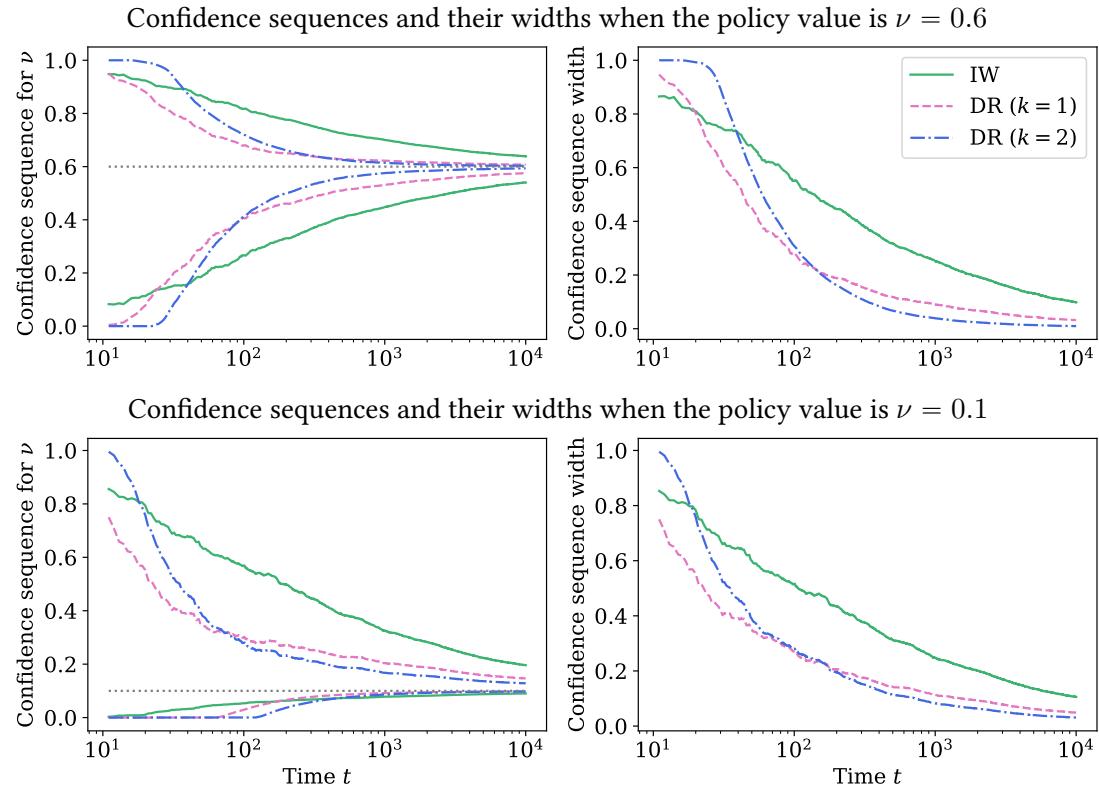


Figure 6.1: Three confidence sequences for a policies with values  $\nu = 0.6$  and  $\nu = 0.1$ . The first CS is built from importance-weighted pseudo-outcomes (“IW”), and the other two are built from doubly robust pseudo-outcomes (“DR”) with  $k$  taking values 1 and 2, respectively. In these examples, the reward  $R_t$  can be predicted easily, a property that only the doubly robust CSs can exploit. Notice that a larger value of  $k$  allows the doubly robust CS to become narrower for large  $t$ , but it pays for this adaptivity with wider bounds at small  $t$ . Nevertheless, all three CSs are time-uniform, and nonasymptotically valid in both simulation scenarios.

Third and finally, we relax the iid assumption, and only require that  $\mathbb{E}_{\pi}(R_t | \mathcal{H}_{t-1}) = \nu \equiv \mathbb{E}_{\pi}(R_t)$  and  $R_t \in [0, 1]$  almost surely. This relaxation of assumptions can be obtained for free, without any change to the resulting CSs whatsoever, and with only a slight modification to the

proof. We summarize our extensions in the following theorem.

**Theorem 6.2.1** (Doubly robust betting off-policy CS). *Suppose  $(X_t, A_t, R_t)_{t=1}^\infty$  is an infinite sequence of contextual bandit data generated by the predictable policies  $(h_t)_{t=1}^\infty$  whose  $[0, 1]$ -valued reward  $R_t$  at time  $t$  has conditional mean  $\mathbb{E}_\pi(R_t \mid \mathcal{H}_{t-1}) = \nu \equiv \mathbb{E}_\pi(R_t)$ . For any predictable sequence  $(\lambda_t^L(\nu'))_{t=1}^\infty$  such that  $\lambda_t^L(\nu') \in [0, (\nu' + k_t)^{-1}]$ , we have*

$$L_t^{\text{DR}} := \inf \left\{ \nu' \in [0, 1] : \prod_{i=1}^t \left[ 1 + \lambda_i^L(\nu') \cdot (\phi_i^{(\text{DR}-\ell)} - \nu') \right] < \frac{1}{\alpha} \right\} \quad (6.12)$$

forms a lower  $(1 - \alpha)$ -CS for  $\nu$ . Similarly, if  $\lambda_t^U(\nu') \in [0, (1 - \nu' + k_t)^{-1}]$  is predictable, then

$$U_t^{\text{DR}} := 1 - \inf \left\{ 1 - \nu' \in [0, 1] : \prod_{i=1}^t \left[ 1 + \lambda_i^U(\nu') \cdot (\phi_i^{(\text{DR}-u)} - (1 - \nu')) \right] < \frac{1}{\alpha} \right\} \quad (6.13)$$

forms an upper  $(1 - \alpha)$ -CS for  $\nu$ .

The proof of Theorem 6.2.1 can be found in Section 6.A.2 and relies on applying Ville's inequality [264] to the products in (6.12) and (6.13). Note that the dimensionality of  $\mathcal{X}$  does not in any way affect the validity of Theorem 6.2.1. Moreover,  $\phi_t^{(\text{DR}-\ell)}$  and  $\phi_t^{(\text{DR}-u)}$  are always unbiased and yield valid CSs regardless of how  $\hat{r}_t$  is chosen. The reason to introduce these doubly robust pseudo-outcomes is to obtain lower-variance CSs (as illustrated in Figure 6.1) since doubly robust estimators can be semiparametric efficient thereby attaining the optimal asymptotic mean squared error in a local minimax sense. These details are outside the scope of the present chapter, but we direct the interested reader to Kennedy [155] and Uehara et al. [256] for modern reviews discussing this subject.

Notice that Theorem 6.2.1 is a generalization of Proposition 6.2.1. Indeed, if the logging policies do not change (i.e.  $h_1 = h_2 = \dots = h$ ), and if the observations  $(X_t, A_t, R_t)_{t=1}^\infty$  are iid, and if  $k_t = 0$  for each  $t$ , then Theorem 6.2.1 recovers Proposition 6.2.1 exactly. For this reason, we do not elaborate on empirical comparisons between Proposition 6.2.1 and Theorem 6.2.1 – any CS that can be derived using the former is a special case of the latter. Moreover, in the *on-policy* setting with all importance weights set to 1 and without a reward predictor, Theorem 6.2.1 recovers the betting-style CSs of Theorem 3.4.1 from Chapter 3. As alluded to in the discussion following Proposition 6.2.1, the infima above are straightforward to compute for many choices of predictable sequences  $(\lambda_t^L(\nu'))_{t=1}^\infty$  and  $(\lambda_t^U(\nu'))_{t=1}^\infty$  including all of those discussed in the following section.

## 6.2.2 Tuning, truncating, and mirroring

We make three remarks below, that are important on both theoretical and practical fronts.

*Remark 11* (Tuning  $(\lambda_t^L)_{t=1}^\infty$  and  $(k_t)_{t=1}^\infty$ ). As stated, Theorem 6.2.1 yields a valid lower CS for  $\nu$  using any predictable sequence of  $[0, (\nu' + k_t)^{-1}]$ -valued tuning parameters  $(\lambda_t^L(\nu'))$  – referred to as “bets” by Karampatziakis et al. [144] and Chapter 3, but how should these

bets be chosen? Section 3.B from Chapter 3 discuss several possible options, but in practice none of them uniformly dominate the others. (This should not be surprising, since there is a certain formal sense in which different nontrivial nonnegative martingales cannot uniformly dominate each other for a given composite sequential testing problem; see Ramdas et al. [209] for a precise statement.) For a simple-to-implement and empirically compelling option, we suggest scaling  $\phi_t^{\text{DR}}$  as  $\xi_t := \phi_t^{\text{DR}}/(k_t + 1)$  and setting  $\lambda_t^L(\nu')$  as

$$\lambda_t^L(\nu') := \sqrt{\frac{2 \log(1/\alpha)}{\hat{\sigma}_{t-1}^2 t \log(1+t)}} \wedge \frac{c}{k_t + \nu'}, \quad \text{where} \quad (6.14)$$

$$\hat{\sigma}_t^2 := \frac{\sigma_0^2 + \sum_{i=1}^t (\xi_i - \bar{\xi}_i)^2}{t+1}, \quad \text{and} \quad \bar{\xi}_t := \left( \frac{1}{t} \sum_{i=1}^t \xi_i \right) \wedge \frac{1}{k_t + 1}, \quad (6.15)$$

with a similar definition for  $\lambda_t^U(\nu')$  but with  $c/(k_t + \nu')$  replaced by  $c/(k_t + (1 - \nu'))$ . Here,  $c \in (0, 1)$  is some truncation scale, reasonable values of which may lie between  $1/4$  and  $3/4$ , but it is of relatively minor practical importance, and for sufficiently large  $t$ , the choice of  $c$  will be inconsequential. A justification for why  $\lambda_t^L(\nu')$  is a sensible choice can be found in Section 3.B from Chapter 3, but the practitioner is nevertheless free to use any other sequence of bets, as long as they are predictable and satisfy the aforementioned boundedness constraints. Furthermore, as in Chapter 3, when  $\lambda_t^L(\nu')$  is chosen as above, the product in (6.12) is quasiconvex in  $\nu' \in [0, 1]$  and hence the infima in Theorem 6.2.1 (and Proposition 6.2.1) can be computed straightforwardly via line or grid search as in Chapter 3.

The sequence of nonnegative  $(k_t)_{t=1}^\infty$  that truncate the reward predictors can also be chosen in any way as long as they are predictable. There are several heuristics that one might use, with increasing levels of complexity. One option is to have a prior guess for  $w_{\max}$  (or some value  $w'_{\max}$  that the practitioner believes will upper-bound most importance weights) and set  $k_t = w'_{\max}/C$  for some  $C \geq 1$ , e.g.  $C = 2$ . For a more adaptive option, one could set  $k_t := \text{median}(w_1, \dots, w_{t-1})$ , or even try out a grid of values  $\{k^{(1)}, \dots, k^{(J)}\}$  and choose the  $k^{(j)}$  that would have yielded the tightest CSs in hindsight. Nevertheless, all three of these options yield nonasymptotically valid  $(1 - \alpha)$ -CSs for  $\nu$ .

*Remark 12* (Why truncate the reward predictor  $\hat{r}_t$ ?). Readers familiar with doubly robust estimation in causal inference or contextual bandits will notice that if  $k_1 = k_2 = \dots = \infty$ , then  $\phi_t^{(\text{DR-}\ell)}$  takes the form of a classical doubly robust estimator of the policy value, and that such estimators often vastly outperform those based on importance weighting alone (and in many cases, are provably more efficient, at least in an asymptotic sense), so why would we want to truncate  $\hat{r}_t$  at all?

The reason has to do with the fact that for nonasymptotic inference, we exploit the lower-boundedness of  $\phi_t^{(\text{DR-}\ell)}$  in order to show that the product in (6.12) is a nonnegative martingale. However, if we introduce a non-truncated reward predictor, we can only say that  $\phi_t^{(\text{DR-}\ell)}$  is lower-bounded by  $-w_{\max}$ , which we do not want to assume knowledge of (or that it is finite at all). Truncation of  $\hat{r}_t$  allows us to occupy a middle ground, so that many of the efficiency gains from doubly robust estimation can be realized, without entirely losing the lower-boundedness

structure of  $\phi_t^{(\text{DR}-\ell)}$ .

This same line of thought helps to illustrate why including a reward predictor tightens our CSs for large  $t$ , potentially at the expense of tightness for smaller  $t$ . Notice that truncating the reward predictor at  $k_t/w_t$  simply restricts the tuning parameter  $\lambda_t^L(\nu')$  to lie in  $[0, (\nu' + k_t)^{-1}]$  instead of  $[0, \nu'^{-1}]$  for importance weighting. Without dwelling on the details too much, it is known that smaller values of  $\lambda_t$  correspond to CSs and CIs being tighter for larger  $t$  – e.g. the role that  $\lambda$  plays in Hoeffding’s CIs looks like  $\sqrt{\log(2/\alpha)/2n}$  [120] – but we refer the reader to Howard et al. [125] or Chapter 3 for more in-depth discussions. The important takeaway for our purposes here, is that larger  $k_t$  corresponds to more variance adaptivity via double robustness, but does more to restrict the CSs tightness at small  $t$ . Nevertheless, this tradeoff is clearly worth it in some cases (see Figure 6.1).

We note that the idea to truncate  $\hat{r}_t$  based on  $k_t/w_t$  was inspired by the so-called “reduced-variance” estimators of Zimmert and Lattimore [297], Zimmert and Seldin [298]. However, their reduced-variance estimators are slightly different since they multiply by an indicator  $\mathbb{1}(w_i \leq \eta)$  for some  $\eta \geq 0$ , which sends the reward predictor to zero for large importance weights, whereas ours only truncates the reward predictor.

*Remark 13* (Mirroring trick for upper CSs). Notice that in Proposition 6.2.1 and Theorem 6.2.1, upper CSs for  $\nu$  were obtained by importance weighting  $1 - R_t$  rather than  $R_t$  to obtain a *lower* CS for  $1 - \nu$ , which was then translated into an *upper* CS for  $\nu$ . This “mirroring trick” – first used in the OPE setting by Thomas et al. [252] to the best of our knowledge – applies to all of the results that follow, but for the sake of succinctness, we will only explicitly write the lower CSs.

### 6.2.3 Closed-form confidence sequences

In Theorem 6.2.1, we derived CSs for the policy value that generalize and improve on prior work [144]. These bounds are empirically tight and can be computed efficiently, but are not closed-form, which may be desirable in practice. In this section, we derive a simple, closed-form, variance-adaptive CS for the fixed policy value  $\nu := \mathbb{E}_\pi(R_t | \mathcal{H}_{t-1})$ . Let

$$\xi_t := \frac{\phi_t^{(\text{DR}-\ell)}}{k_t + 1}, \quad \text{and} \quad \hat{\xi}_{t-1} := \left( \frac{1}{t-1} \sum_{i=1}^{t-1} \xi_i \right) \wedge \frac{1}{k_t + 1}. \quad (6.16)$$

With the above notation in mind, we are ready to state the main result of this section.

**Proposition 6.2.2** (Closed-form predictable plug-in CS for  $\nu$ ). *Given contextual bandit data  $(X_t, A_t, R_t)_{t=1}^\infty$  with  $[0, 1]$ -valued rewards, choose nonnegative and predictable tuning parameters  $(k_t)_{t=1}^\infty$ , and define  $(\lambda_t)_{t=1}^\infty$  as*

$$\lambda_t := \sqrt{\frac{2 \log(1/\alpha)}{\hat{\sigma}_{t-1}^2 t \log(1+t)}} \wedge c, \quad \hat{\sigma}_t^2 := \frac{\sigma_0^2 + \sum_{i=1}^t (\xi_i - \bar{\xi}_t)^2}{t+1}, \quad \bar{\xi}_t := \left( \frac{1}{t} \sum_{i=1}^t \xi_i \right) \wedge \frac{1}{k_t + 1}, \quad (6.17)$$

where  $c \in (0, 1)$  is some truncation parameter (reasonable values of which may include  $1/2$  or  $3/4$ ); and  $\xi_0 \in (0, 1)$  and  $\sigma_0^2 > 0$  are some user-chosen parameters that can be thought of as prior guesses for the mean and variance of  $\xi$ , respectively. Then,

$$L_t^{\text{PrPl}} := \left( \frac{\sum_{i=1}^t \lambda_i \xi_i}{\sum_{i=1}^t \lambda_i / (k_i + 1)} - \frac{\log(1/\alpha) + \sum_{i=1}^t (\xi_i - \hat{\xi}_{i-1})^2 \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i / (k_i + 1)} \right) \vee 0 \quad (6.18)$$

forms a lower  $(1 - \alpha)$ -CS for  $\nu$ . An analogous upper CS  $(U_t^{\text{PrPl}})_{t=1}^\infty$  follows by mirroring (Remark 13).

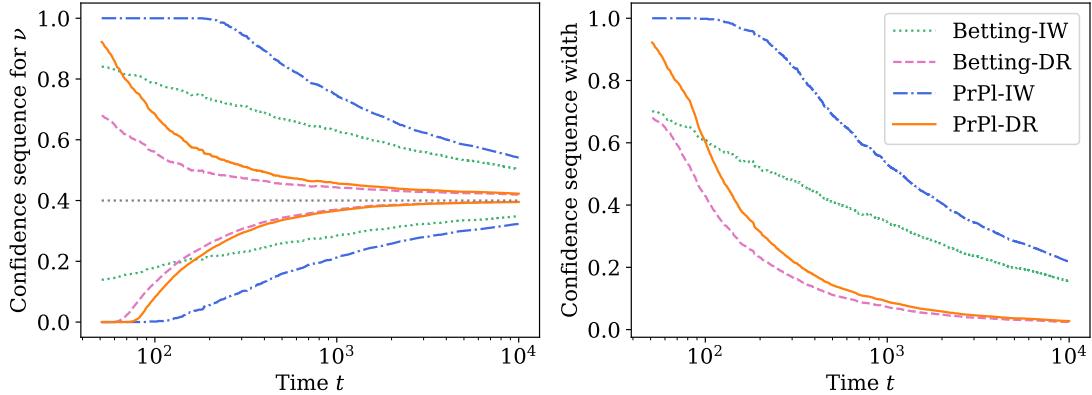


Figure 6.2: Betting-based (Theorem 6.2.1) and predictable plug-in (PrPl) (Proposition 6.2.2) CSs for  $\nu$  with both importance-weighted (IW) and doubly robust (DR) variants. Notice that for both IW and DR CSs, the betting-based approach of Theorem 6.2.1 outperforms the PrPl CSs. Nevertheless, the closed-form PrPl CSs are simpler to implement, and can still benefit from doubly robust variance adaptation.

The proof can be found in Section 6.A.4 and relies on Ville’s inequality [264] applied to a predictable plug-in empirical Bernstein supermartingale similar to that appearing in Chapter 3 but with a variant of Fan’s inequality [102] for lower-bounded random variables with upper-bounded means. As seen in Figure 6.2, the betting CSs of Theorem 6.2.1 still have better empirical performance than the closed-form predictable plug-in CSs of Proposition 6.2.2, but the latter are more computationally and analytically convenient, and are valid under the same set of assumptions. A similar phenomenon was observed in Chapter 3 for bounded random variables (outside the context of OPE). In the on-policy setting with all importance weights set to 1 and without a reward predictor, Proposition 6.2.2 recovers Theorem 3.3.1 from Chapter 3.

It is important not to confuse  $\hat{\xi}_t$  with  $\bar{\xi}_t$ . The difference between them may seem minor since the former simply has access to one less data point than the latter, but they play two very different roles in Proposition 6.2.2:  $\hat{\xi}_{t-1}$  is a *predictable* sample mean that shows up in the width of  $L_t^{\text{PrPl}}$  explicitly, and its predictability is fundamental to the proof technique, while  $\bar{\xi}_t$  is just used as a tool to obtain better estimates of  $\hat{\sigma}_t^2$  so that they can be plugged in to the tuning

parameters  $\lambda_t$ . Consequently,  $\hat{\xi}_t$  can be found in CSs that rely on a similar proof technique (such as Theorem 6.3.1), while  $\bar{\xi}_t$  can be found in other CSs that make use of predictable tuning parameters (such as Theorem 6.2.1).

### 6.2.4 Fixed-time confidence intervals

While this chapter is focused on time-uniform CSs for OPE, our methods also naturally give rise to fixed-time CIs that are *not* anytime-valid but can still benefit from our general techniques. In this section, we will briefly discuss what minor modifications are needed to derive sharp fixed-time instantiations of our otherwise time-uniform bounds. We will also compare our fixed-time CIs to the CIs of Thomas et al. [252], but this comparison is by no means comprehensive. Indeed, our goal is not to show that our methods are “better” than prior work, even if some simulations may suggest this — instead, we aim to provide the reader with some context as to how our fixed-time instantiations fit within the broader literature on CIs for OPE.

**Confidence intervals for policy values.** We begin by deriving fixed-time analogues of the CSs for  $\nu$  presented in Theorem 6.2.1 and Proposition 6.2.2 — the former being a “betting-style” CS that is very tight in practice, and the latter being a closed-form predictable plug-in (PrPl) CS that is slightly more analytically and computationally convenient. In both cases, our suggested modification is essentially the same: choose a predictable sequence  $(\lambda_t)_{t=1}^n$  that is tuned for the desired sample size  $n$  — to be elaborated on shortly — and take the intersection of the implicit CS  $(C_t)_{t=1}^n$  that is formed from times 1 through  $n$ . Concretely, define the predictable sequence  $(\lambda_t)_{t=1}^n$  given by

$$\dot{\lambda}_{t,n} := \sqrt{\frac{2 \log(1/\alpha)}{n \hat{\sigma}_{t-1}^2}}, \quad (6.19)$$

where  $\hat{\sigma}_t^2$  is given as in (6.15). Then, we have the following corollary for betting-style CIs for  $\nu$ .

**Corollary 6.2.1.** *Let  $(X_t, A_t, R_t)_{t=1}^n$  be a finite sequence of contextual bandit data with  $[0, 1]$ -valued rewards and define  $(\lambda_t^L(\nu'))_{t=1}^n$  and  $(\lambda_t^U(\nu'))_{t=1}^n$  by*

$$\lambda_t^L(\nu') := \dot{\lambda}_{t,n} \wedge \frac{c}{k_t + \nu'}, \quad \text{and} \quad \lambda_t^U(\nu') := \dot{\lambda}_{t,n} \wedge \frac{c}{k_t + (1 - \nu')}, \quad (6.20)$$

where  $c \in (0, 1)$  is some truncation scale as in Theorem 6.2.1. Let  $L_t^{\text{DR}}$  and  $U_t^{\text{DR}}$  be as in (6.12) and (6.13). Then,

$$\dot{L}_n := \max_{1 \leq t \leq n} L_t^{\text{DR}} \quad \text{and} \quad \dot{U}_n := \min_{1 \leq t \leq n} U_t^{\text{DR}} \quad (6.21)$$

form lower and upper  $(1 - \alpha)$ -CIs for  $\nu$ , respectively, meaning  $\mathbb{P}(\nu \in [\dot{L}_n, \dot{U}_n]) \geq 1 - \alpha$ .

Corollary 6.2.1 is an immediate consequence of Theorem 6.2.1 where the sequence of tuning parameters was chosen to tighten the CI for the sample size  $n$ . This particular choice of  $(\lambda_t^L(\nu'))_{t=1}^n$  is inspired by the fact that the product in (6.12) resembles an exponential supermartingale whose resulting CI can be tightened using tuning parameters that are well-

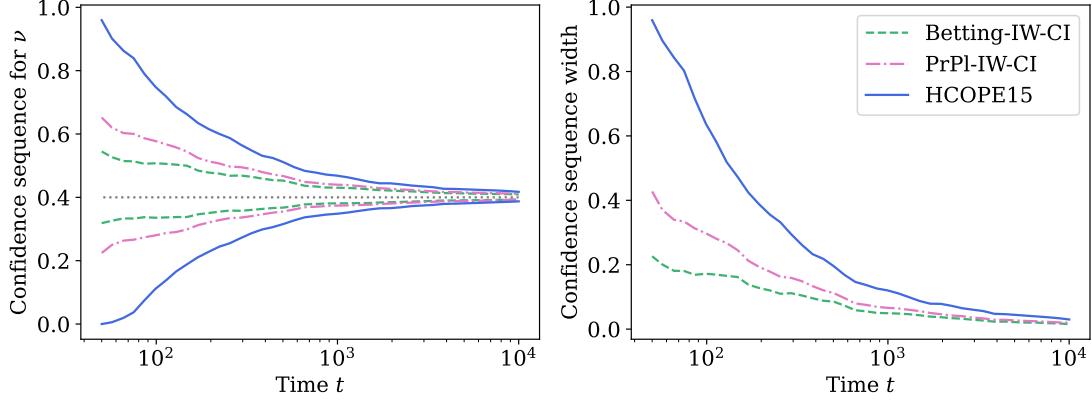


Figure 6.3: Fixed-time 90% confidence intervals for  $\nu$  using three different methods: a betting-based CI (Corollary 6.2.1), a predictable plug-in (PrPl) CI (Corollary 6.2.2), and those presented in a paper entitled “High-confidence off-policy evaluation” (HCOPE15) by Thomas et al. [252]. Notice that the betting-based CI outperforms the closed-form PrPl CI, which itself significantly outperforms the bounds in Thomas et al. [252].

estimated by (6.20). For more details, we refer the reader to Section 3.3 from Chapter 3. The fact that the maximum and minimum can be taken in (6.21) follows from the fact that  $(L_t^{\text{DR}})_{t=1}^n$  satisfies  $\mathbb{P}(\forall t \in \{1, \dots, n\}, \nu \geq L_t^{\text{DR}}) \geq 1 - \alpha$ , and similarly for the upper CI  $\dot{U}_n$ . Figure 6.3 demonstrates what these CIs may look like in practice.

Similar to how Corollary 6.2.1 is a fixed-time instantiation of Theorem 6.2.1, the following corollary is a fixed-time instantiation of the closed-form PrPl CSs of Proposition 6.2.2.

**Corollary 6.2.2.** *Let  $(X_t, A_t, R_t)_{t=1}^n$  be a finite sequence of contextual bandit data with  $[0, 1]$ -valued rewards and define  $(\lambda_t)_{t=1}^n$  by*

$$\lambda_t := \dot{\lambda}_{t,n} \wedge c, \quad (6.22)$$

*with  $c$  chosen as in Proposition 6.2.2. Then, with  $L_t^{\text{PrPl}}$  and  $U_t^{\text{PrPl}}$  defined in Proposition 6.2.2, we have that*

$$\dot{L}_n^{\text{PrPl}} := \max_{1 \leq t \leq n} L_t^{\text{PrPl}} \quad \text{and} \quad \dot{U}_n^{\text{PrPl}} := \min_{1 \leq t \leq n} U_t^{\text{PrPl}} \quad (6.23)$$

*form lower and upper  $(1 - \alpha)$ -CIs for  $\nu$ , respectively, meaning  $\mathbb{P}(\nu \in [\dot{L}_n^{\text{PrPl}}, \dot{U}_n^{\text{PrPl}}]) \geq 1 - \alpha$ .*

Corollary 6.2.2 is an immediate consequence of Proposition 6.2.2 instantiated for a different choice of  $\lambda_t$  and with an intersection being taken over the implicit CS from times 1 through  $n$ .

While the methods of this section improve on past work both theoretically and empirically, each of our results thus far have assumed that  $\nu \equiv \mathbb{E}_\pi(R_t)$  is fixed and does not change over time, an assumption that we may not always wish to make in practice (e.g. if the environment is nonstationary). Fortunately, it is still possible to design CSs that capture an interpretable parameter: the *time-varying average policy value thus far*. However, we will need completely

different supermartingales to achieve this, which we outline in the following section.

### 6.3 Inference for time-varying policy values

Let us now consider the more challenging task of performing anytime-valid off-policy inference for a time-varying average policy value. Concretely, suppose that the value of the  $[0, 1]$ -bounded reward  $R_t$  under policy  $\pi$  is given by  $\nu_t := \mathbb{E}_\pi(R_t | \mathcal{H}_{t-1}) \in [0, 1]$ , and hence  $(\nu_t)_{t=1}^\infty$  is now a *sequence* of conditional policy values. Our goal is to derive CSs for  $(\tilde{\nu}_t)_{t=1}^\infty$  where  $\tilde{\nu}_t := \frac{1}{t} \sum_{i=1}^t \nu_i$  is the *average conditional policy value so far*. In addition to satisfying desiderata 1–5 in Section 6.1.2 our CSs will impose no restrictions on how  $(\nu_t)_{t=1}^\infty$  changes over time. Unfortunately, the techniques of Proposition 6.2.1 and Theorem 6.2.1 will not work here because their underlying test statistics cannot be written explicitly as functions of a candidate average policy value  $\tilde{\nu}'_t := \frac{1}{t} \sum_{i=1}^t \nu_i$ , but only of the entire candidate tuple  $(\nu_1, \dots, \nu_t)$ . To remedy this, we will rely on test statistics that are functions of a candidate average  $\tilde{\nu}'_t$  to derive CSs for  $\tilde{\nu}_t$  in Theorems 6.3.1 and 6.3.1. An empirical demonstration of the failure of Theorem 6.2.1 juxtaposed with the remedy provided by Theorem 6.3.1 can be seen in the left-hand side of Figure 6.4 and a comparison between Theorem 6.3.1 and Proposition 6.3.1 can be seen in the right-hand side of the same figure.

We will present two CSs for  $\tilde{\nu}_t$ : (1) the “empirical Bernstein” CS in Theorem 6.3.1 whose underlying supermartingale is constructed using Robbins’ method of mixtures [217], and (2) the “iterated logarithm” CS in Proposition 6.3.1 which uses the stitching technique. Both yield time-uniform, nonasymptotically valid bounds, and are easy to compute. However, the former tends to have better empirical performance in finite samples, while the latter achieves the (optimal) rate of convergence, matching the law of the iterated logarithm. Nevertheless, both boundaries shrink to zero-width at a rate of  $\tilde{O}(\sqrt{V_t}/t)$  [125]. Here,  $\tilde{O}(\cdot)$  means  $O(\cdot)$  up to logarithmic factors.

In order to write down the empirical Bernstein CS, we first define the scaled doubly robust pseudo-outcomes  $\xi_t := \phi_t/(1+k)$  where  $\phi_t \equiv \phi_t^{(\text{DR-}\ell)}$  is given in (6.10) and with all  $k_t$  equal to a fixed  $k$ . Then, define the corresponding centered sum process  $(S_t(\tilde{\nu}'_t))_{t=1}^\infty$  and variance process  $(V_t)_{t=1}^\infty$  given by

$$S_t(\tilde{\nu}'_t) := \sum_{i=1}^t \xi_i - \frac{t\tilde{\nu}'_t}{1+k}, \quad \text{and} \tag{6.24}$$

$$V_t := \sum_{i=1}^t (\xi_i - \hat{\xi}_{i-1})^2, \quad \text{where } \hat{\xi}_t := \left( \frac{1}{t} \sum_{i=1}^t \xi_i \right) \wedge \frac{1}{1+k}, \tag{6.25}$$

and  $\xi_0 \in [0, (1+k)^{-1}]$  is chosen by the user. With this setup and notation in mind, we are ready to state the empirical Bernstein CS for  $\tilde{\nu}_t$ .

**Theorem 6.3.1** (Empirical Bernstein confidence sequence for  $\tilde{\nu}_t$ ). *Let  $(X_t, A_t, R_t)_{t=1}^\infty$  be an infinite sequence of contextual bandit data with  $[0, 1]$ -valued rewards generated by the sequence*

of policies  $(h_t)_{t=1}^\infty$ , and let  $(S_t(\tilde{\nu}'_t))_{t=1}^\infty$  and  $(V_t)_{t=1}^\infty$  be the centered sum and variance processes as in (6.24) and (6.25). Then for any  $\rho > 0$ ,

$$M_t^{\text{EB}}(\tilde{\nu}_t) := \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{V_t + \rho} \right) {}_1F_1(1, V_t + \rho + 1, S_t(\tilde{\nu}_t) + V_t + \rho) \quad (6.26)$$

forms a nonnegative supermartingale starting at one, where  ${}_1F_1(\cdot, \cdot, \cdot)$  is Kummer's confluent hypergeometric function and  $\Gamma(\cdot, \cdot)$  is the upper incomplete gamma function. Consequently,

$$L_t^{\text{EB}} := \inf \{ \tilde{\nu}'_t \in [0, 1] : M_t^{\text{EB}}(\tilde{\nu}'_t) < 1/\alpha \} \quad (6.27)$$

forms a lower  $(1 - \alpha)$ -CS for  $\tilde{\nu}_t$ , meaning  $\mathbb{P}(\forall t \in \mathbb{N}, \tilde{\nu}_t \geq L_t^{\text{EB}}) \geq 1 - \alpha$ . Similarly, an upper CS can be derived by using  $\phi_t^{(\text{DR-u})}$  defined in (6.11) and employing the mirroring trick described in Remark 13.

The proof of Theorem 6.3.1 can be found in Section 6.A.3 and relies on an inequality due to Fan et al. [102] along with a mixture supermartingale analogous to that of Howard et al. [125, Proposition 9]. In the on-policy setting with all importance weights set to 1 and with no reward predictor (and hence  $k = 0$ ), we have that Theorem 6.3.1 recovers the gamma-exponential mixture CS of Howard et al. [125, Proposition 9].

The tuning parameter  $\rho > 0$  effectively dictates the neighborhood of intrinsic time – i.e. the value of  $V_t$  – at which  $L_t^{\text{EB}}$  is tightest, and it is rather straightforward to choose  $\rho > 0$  given this interpretation. Following Howard et al. [125],  $\rho > 0$  can be chosen to (approximately) tighten  $L_t^{\text{EB}}$  at  $V_t = V^*$  by setting

$$\rho(V^*) := \sqrt{\frac{-2 \log \alpha + \log(-2 \log \alpha + 1)}{V^*}}. \quad (6.28)$$

Nevertheless,  $L_t^{\text{EB}}$  forms a valid lower  $(1 - \alpha)$ -CS for  $\tilde{\nu}_t$  regardless of how  $\rho > 0$  is chosen, as long as this is done data-independently.

Readers familiar with gamma-exponential mixture supermartingales such as those in Howard et al. [124, 125] and Choe and Ramdas [57] may have expected to see a lower incomplete gamma function  $\gamma(a, b)$  instead of  ${}_1F_1(1, a, b)$ . Indeed,  ${}_1F_1(1, a, b)$  reduces to  $\gamma(a, b)$  when  $b \equiv S_t(\tilde{\nu}_t) + V_t + \rho$  is nonnegative, but unlike the lower incomplete gamma,  ${}_1F_1(1, a, b)$  is well-defined when  $b < 0$ . Writing (6.26) in terms of a lower incomplete gamma would have required lower-bounding this term by a piece-wise function when  $b \equiv S_t + V_t + \rho < 0$  as in Choe and Ramdas [57, Appendix C]; see Remark 15 for details.

While  $L_t^{\text{EB}}$  is not a closed-form bound, it can be computed efficiently using root-finding algorithms, and it can also be shown to achieve an asymptotic width of  $O(\sqrt{V_t \log V_t}/t)$  (the justifications in Howard et al. [125] carry over to this scenario). While this rate is sufficient for deriving CSs with strong empirical performance that shrink to zero-width, one can derive CSs that achieve an improved rate of  $O(\sqrt{V_t \log \log V_t}/t)$  using a different technique known as “stitching”.

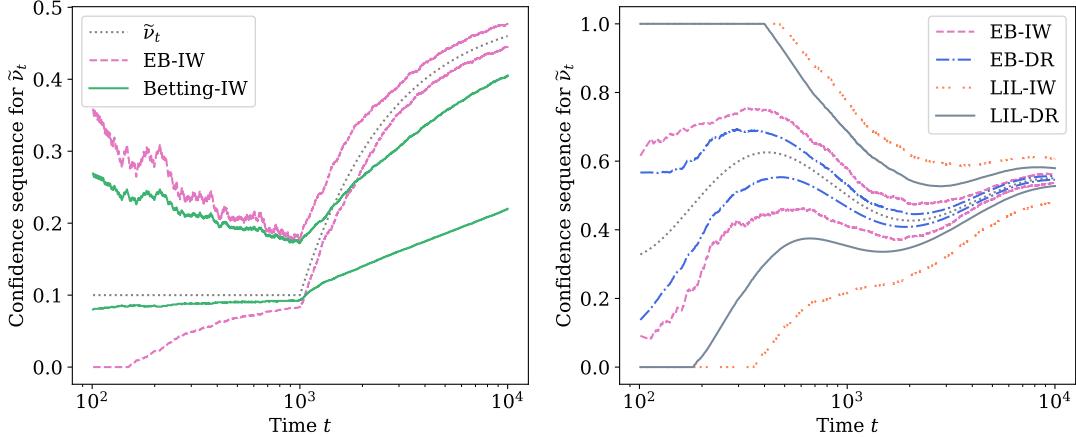


Figure 6.4: Various CSs for the time-varying policy value  $\tilde{\nu}_t$ . The left-hand side plot illustrates that while the betting-style CS of Theorem 6.2.1 is tight when  $\tilde{\nu}_t$  remains fixed, it fails to cover when  $\tilde{\nu}_t$  changes (in this case, there is an abrupt change at  $t = 1000$ ). The right-hand side plot illustrates how Theorem 6.3.1 and Proposition 6.3.1 compare, both using their importance-weighted (IW) and doubly robust (DR) variants. Notice that while LIL-IW and LIL-DR attain optimal rates of convergence, the empirical Bernstein CSs (EB-IW and EB-DR) are much tighter in practice. In both cases, the DR variant outperforms the IW variant due to the reward being easy to predict in this particular example.

**Proposition 6.3.1** (Variance-adaptive iterated logarithm confidence sequence for  $\tilde{\nu}_t$ ). *Let  $\bar{V}_t := V_t \vee 1$  where  $V_t$  is given as in (6.25). Define the function  $\ell_t(\alpha)$*

$$\ell_t(\alpha) := 2 \log (\log \bar{V}_t + 1) + \log \left( \frac{1.65}{\alpha} \right). \quad (6.29)$$

*Then we have that under the same conditions as Theorem 6.3.1,*

$$L_t^{\text{LIL}} := (k+1) \left( \frac{1}{t} \sum_{i=1}^t \xi_i - \frac{\sqrt{2.13\ell_t(\alpha)\bar{V}_t + 1.76\ell_t(\alpha)^2}}{t} - \frac{1.33\ell_t(\alpha)^2}{t} \right) \vee 0 \quad (6.30)$$

*forms a lower  $(1-\alpha)$ -CS for  $\tilde{\nu}_t := \frac{1}{t} \sum_{i=1}^t \nu_i$ , meaning  $\mathbb{P}(\forall t, \tilde{\nu}_t \geq L_t^{\text{LIL}}) \geq 1-\alpha$ . An analogous upper CS  $(U_t^{\text{LIL}})_{t=1}^\infty$  can be derived using the mirroring trick of Remark 13.*

The proof in Section 6.A.5 uses the ‘‘stitching’’ (sometimes called ‘‘peeling’’) technique that is common in the derivation of LIL-type bounds [76, 132, 150] applied to linear sub-exponential boundaries. Importantly,  $L_t^{\text{LIL}} \asymp \mathcal{O}(t^{-1}\sqrt{V_t \log \log V_t})$  which matches the unimprovable rate implied by the law of the iterated logarithm. We take a maximum with 0 in  $L_t^{\text{LIL}}$  (and hence an implicit minimum with 1 in  $U_t^{\text{LIL}}$ ) since  $\tilde{\nu}_t \in [0, 1]$  for every  $t \in \mathbb{N}$  by assumption. Of course, if one is in a stationary environment so that  $\nu_1 = \nu_2 = \dots = \nu$ , then  $\tilde{\nu}_t = \nu$ , and hence  $(L_t^{\text{LIL}}, U_t^{\text{LIL}})$  forms a  $(1-\alpha)$ -CS for  $\nu$ .

**Comparison of Theorems 6.2.1 and 6.3.1 with prior work.** To the best of our knowledge, Thomas et al. [252] were the first to derive nonasymptotic confidence intervals for policy values in contextual bandits, and they did so without knowledge of  $w_{\max}$ . However, their bounds are not time-uniform, and the authors do not consider time-varying policy values nor data-dependent logging policies  $(h_t)_{t=1}^{\infty}$ . Four other prior works stand out as being related to the results of this section, namely Karampatziakis et al. [144], Howard et al. [125, Section 4.2], Bibaut et al. [33], and Zhan et al. [292] and we discuss each of them in some detail below. Note that in the last row labeled “Doubly robust” of Table 6.1, we are referring to the property of confidence sets to be potentially sharpened in the presence of regression estimators without compromising validity (as discussed in the paragraphs surrounding Theorem 6.2.1).

- **KMR21:** The off-policy CSs of Karampatziakis et al. [144] – reviewed in Proposition 6.2.1 – can in several ways be seen as an improvement of Thomas et al. [252] since they are time-uniform, in addition to being empirically tight. As discussed in Section 6.2, their importance-weighted off-policy CSs are strictly generalized and extended in our Theorem 6.2.1, but their result nevertheless yields the current state-of-the-art CSs for  $\nu$ .
- **BDKCvdL21 & ZHHA21:** Bibaut et al. [33] and Zhan et al. [292] both study the off-policy inference problem in contextual bandits from an asymptotic point of view. Their off-policy estimators take the form of sample averages of influence functions – what Bibaut et al. [33] refer to as the canonical gradient – to which martingale central limit theorems may be applied to obtain asymptotically valid inference.

In contrast to our work, the confidence intervals of Bibaut et al. [33] and Zhan et al. [292] (a) are asymptotic, and hence do not have finite-sample guarantees, (b) are not time-uniform, and hence cannot be used at stopping times, and (c) do not track time-varying policy values.

- **HRMS21:** Howard et al. [125, Section 4.2] derive time-uniform, nonasymptotic CSs for the average treatment effect (ATE) in randomized experiments. The main difference between our results and those of [125, Section 4.2] is that they do not study the contextual bandit off-policy evaluation problem. However, since estimating the ATE in randomized experiments can be seen as a special case of the contextual bandit problem, it is natural to wonder how our approach differs in this special case. The main difference here is that Howard et al. [125] require knowledge of  $w_{\max}$  – or equivalently in their setup, the maximal and minimal propensity scores – while ours do not. Moreover, our results allow  $w_{\max}$  to be infinite and nevertheless enjoy variance-adaptivity. See Section 6.3.2 for a more detailed discussion of the implications of our bounds for ATE estimation in randomized experiments.

### 6.3.1 A remark on policy value differences

The results in this chapter have taken the form of CSs for policy values, e.g. a sequence of sets  $[L_t, U_t]_{t=1}^{\infty}$  such that  $\mathbb{P}(\forall t \in \mathbb{N}, \nu \in [L_t, U_t]) \geq 1 - \alpha$ , but it may be of interest to directly estimate policy value *differences* – e.g.  $\delta \equiv \nu_1 - \nu_2 \equiv \nu(\pi_1) - \nu(\pi_2)$ , where  $\nu(\pi)$  is the value

Table 6.1: Comparison of various CSs and CIs for mean off-policy values.

	KMR21	BDKCvdL21 & ZHHA21	HRMS21	Thm. 6.2.1	Thm. 6.3.1
Contextual bandits	✓	✓		✓	✓
Time-varying rewards			✓		✓
Nonasymptotic	✓		✓	✓	✓
Time-uniform	✓		✓	✓	✓
Predictable $(h_t)_{t=1}^\infty$		✓	✓	✓	✓
$w_{\max}$ -free	✓	✓		✓	✓
Doubly robust		✓	✓	✓	✓

of some policy  $\pi$ , and  $\pi_1$  and  $\pi_2$  are two policies we would like to compare. In many cases including the gated deployment problem studied by Karampatziakis et al. [144],  $\pi_1$  is some target policy of interest and  $\pi_2 = h_1 = h_2 = \dots = h$  is the logging policy so that  $\nu(\pi_1) - \nu(h)$  can be interpreted as the additional value (or “lift”) in the target policy  $\pi_1$  over the logging policy. However, our setup allows for  $\pi_1$  and  $\pi_2$  to be any two policies that are absolutely continuous with respect to the logging policies.

Of course, one can always solve this problem by union bounding: construct  $(1 - \alpha/2)$ -CSs for  $\nu_1$  and  $\nu_2$  separately to yield a  $(1 - \alpha)$ -CS for their difference. However, it is possible to remove this small amount of slack introduced by union bounding and instead derive a CS for the difference directly.

The idea is simple: rather than only leverage lower-boundedness of importance-weighted rewards  $w_t R_t$ , we construct a new random variable  $\theta_t := w_t^{(1)} R_t - [1 - w_t^{(2)}(1 - R_t)]$  and leverage its lower-boundedness directly — here,  $w_t^{(1)} = \pi_1/h_t$  and  $w_t^{(2)} = \pi_2/h_t$  are the importance weights for policies  $\pi_1$  and  $\pi_2$ , respectively. In particular, notice that

$$\mathbb{E}[\theta_t] = \delta, \quad \text{and } \theta_t \geq -1, \quad \text{and hence} \tag{6.31}$$

$$\frac{1}{2} [\theta_t - \delta] \geq -1 \quad \text{almost surely.} \tag{6.32}$$

Consequently, (6.32) can be used in the proofs of our theorems to derive a CS for  $\delta$  directly, since those proofs fundamentally rely on the centered (i.e. with their mean subtracted) random variables being almost surely lower-bounded by  $-1$ . For instance, we have that for any  $(0, 1)$ -

valued predictable sequence  $(\lambda_t(\delta))_{t=1}^\infty$ ,

$$M_t(\delta) := \prod_{i=1}^t (1 + \lambda_i(\delta) \cdot (\theta_t - \delta)/2) \quad (6.33)$$

forms a test martingale and hence  $L_t := \inf \{\delta' \in [-1, 1] : M_t(\delta') < 1/\alpha\}$  forms a lower  $(1 - \alpha)$ -CS for  $\delta$ . As usual, the mirroring trick can be used to obtain an upper CS for this policy value difference. Moreover, the above discussion can be extended to time-varying policy value differences and doubly robust pseudo-outcomes (rather than just their importance-weighted counterparts), as well as sequences of policies — i.e. analyzing the sequences  $(\pi_1^{(t)})_{t=1}^\infty$  and  $(\pi_2^{(t)})_{t=1}^\infty$  — but we omit these derivations for the sake of brevity.

### 6.3.2 Time-varying treatment effects in adaptive experiments

While this chapter is focused on anytime-valid *contextual bandit* inference — i.e. inference for policy values or their CDFs from contextual bandit data — one can nevertheless view off-policy evaluation as a generalization of treatment effect estimation from adaptive experiments. Consequently, every single result in this chapter also has powerful implications for nonasymptotic inference for treatment effects from such experiments. In this section, we will focus on adaptive experiments with binary treatments for simplicity, but the analogy extends to more general settings.

**From contextual bandits to adaptive experiments with binary treatments.** The contextual bandit problem can be seen as a generalization of adaptive experiments since the latter has three key notational differences.

1. The “context”  $X_t$  is typically referred to as a “covariate” or a “feature”, and may be used to represent baseline demographics and medical history in a clinical trial, for example.
2. The “action”  $A_t$  (which is binary in this case) is referred to as a “treatment”, and the policy  $h_t$  is called the “propensity score”, and is simply the probability of a subject with covariates  $X_t$  receiving treatment  $A_t = 1$  at time  $t$ .
3. The “reward”  $R_t$  is often referred to as the “outcome” for subject  $t$ .

There are many reasons why one may wish to run an adaptive sequential experiment rather than a simple Bernoulli( $h$ ) experiment with a constant pre-specified propensity score  $h$ . Two simple examples include: (a) balancing designs such as Efron’s biased coin [98] which vary the propensity scores  $(h_t)_{t=1}^\infty$  over time to ensure that treatment groups are “balanced” within certain levels of the covariates, and (b) the experimental designs of Kato et al. [148] which adaptively choose propensity scores to minimize the variance of the resulting doubly robust and inverse propensity-weighted (IPW) estimators, yielding sharper confidence sets. (In the language of contextual bandits and off-policy evaluation, IPW and importance weighting are equivalent.) Both (a) and (b) — or any other design that varies propensity scores adaptively over time — can be paired with the CSs of the current chapter.

**Implications for causal inference in adaptive experiments.** From the perspective of treatment effect estimation, the current chapter provides nonasymptotic, nonparametric, time-uniform inference for treatment effects, all without knowledge of the minimal propensity score  $h_{\min} := \text{ess inf}_{t,a,x} h_t(a | x)$  and this essential infimum can even be 0 (as long as it is not attained, meaning each  $h_t(a | x)$  is itself positive). Contrast this with prior work on nonasymptotic, nonparametric, time-uniform inference for treatment effects such as Howard et al. [125, Section 4.2], which require *a priori* knowledge of  $h_{\min} > 0$ . Their bounds necessarily scale with an implied upper-bound on the variance of  $h_t^{-1} R_t$  implied by  $h_{\min}^{-1}$  while ours only scale with the *empirical variance* of  $(h_t^{-1} R_t)_{t=1}^{\infty}$  — the latter always being smaller.

Concretely, Theorem 6.2.1 can be used to derive CSs for the average treatment effect from adaptively collected data in experiments with binary treatments and bounded outcomes. Theorem 6.3.1 goes further, enabling the construction of CSs for *time-varying average treatment effects* similar to Howard et al. [125, Section 4.2]. Finally, Theorem 6.4.1 — to be presented in Section 6.4 — allows for the construction of time-uniform confidence bands for the CDF of the outcome distribution under a given treatment. Moreover, all of this is possible in a nonparametric, nonasymptotic framework, without knowledge (or strict positivity) of  $h_{\min}$ . To the best of our knowledge, all three of these implied results are new in the literature for treatment effect estimation.

### 6.3.3 Sequential testing and anytime $p$ -values for off-policy inference

While we have thus far taken an *estimation* perspective (i.e. deriving CSs and CIs rather than  $p$ -values), all of our results have hypothesis testing analogues. In particular, the CSs and CIs developed in this chapter have all been built by first deriving implicit *e-processes*. Formally, given a set of distributions  $\mathcal{P}_0$  (referred to as “the null hypothesis”), an *e*-process  $E \equiv (E_t)_{t=1}^{\infty}$  for  $\mathcal{P}_0$  is a nonnegative process such that  $\mathbb{E}_P[E_{\tau}] \leq 1$  for any  $P \in \mathcal{P}_0$  and any stopping time  $\tau$ . (In particular, all test supermartingales for  $\mathcal{P}_0$  are *e*-processes by the optional stopping theorem, but not vice versa.)

While *e*-processes can serve as tools to derive CSs, they can also be used as interpretable testing tools in their own right, or as a way to derive anytime  $p$ -values —  $p$ -values that are uniformly valid over time in the same sense as CSs. Formally, an anytime  $p$ -value for  $\mathcal{P}_0$  is an  $\mathcal{H}$ -adapted process  $(p_t)_{t=1}^{\infty}$  such that

$$\sup_{P \in \mathcal{P}_0} P(\exists t \in \mathbb{N} : p_t \leq \alpha) \leq \alpha. \quad (6.34)$$

Compare (6.34) with a traditional fixed-time  $p$ -value  $p_n$  that satisfies  $\forall n \in \mathbb{N}, \sup_{P \in \mathcal{P}_0} P(p_n \leq \alpha) \leq \alpha$ . On the other hand, an *e*-process  $(E_t)_{t=1}^{\infty}$  for  $\mathcal{P}_0$  also satisfies Ville’s inequality:

$$\sup_{P \in \mathcal{P}_0} P(\exists t \in \mathbb{N} : E_t \geq 1/\alpha) \leq \alpha. \quad (6.35)$$

As a direct consequence of (6.35), notice that *e*-processes yield anytime  $p$ -values via the transformation  $p_t := (1/E_t) \wedge 1$ .

*Remark 14* (Which should you choose:  $e$  or  $p$ ?). There are several philosophical and practical reasons why one may wish to use  $e$ -processes over anytime  $p$ -values, despite the fact that they can both be used for sequential hypothesis testing. Philosophically,  $e$ -processes (and hence test supermartingales) have game-theoretic interpretations and connections to Bayesian statistics [237, 113, 281], and have been argued to serve as a better foundation for statistical communication [234]. Practically, stopped  $e$ -processes form  $e$ -values — nonnegative random variables with expectation at most one [266] — which have several attractive properties over  $p$ -values, including the fact that they are very straightforward to combine for the sake of testing a global null [266], to perform meta-analyses [251], or to control the false discovery rate under arbitrary dependence [275]. In this section, we remain agnostic as to which of the two one should use: we will simply derive  $e$ -processes and note that their philosophical and practical properties can be enjoyed within OPE, and that if anytime  $p$ -values are preferred, they are always available via the transformation  $p_t := (1/E_t) \wedge 1$ .

Following Section 6.3.1, let us now derive sequential tests for whether a policy  $\pi_1$  has a higher average value than some other policy  $\pi_2$ . Technically, one could also replace  $\pi_1$  or  $\pi_2$  with a *sequence* of predictable policies, but for simplicity we will only discuss fixed policies. Concretely, let  $\Delta_t$  denote the difference in the values of policies  $\pi_1$  and  $\pi_2$  at time  $t$ ,

$$\delta_t := \nu_t(\pi_1) - \nu_t(\pi_2) \equiv \mathbb{E}_{A_t \sim \pi_1}(R_t) - \mathbb{E}_{A_t \sim \pi_2}(R_t), \quad (6.36)$$

and let  $\Delta_t := \frac{1}{t} \sum_{i=1}^t \delta_i$  denote the running average difference. We are interested in testing the *weak* null hypothesis  $H_0$ ,

$$H_0 : \forall t, \Delta_t \leq 0, \quad \text{vs} \quad H_1 : \exists t : \Delta_t > 0. \quad (6.37)$$

In words,  $H_0$  says that “ $\pi_1$  is no better than  $\pi_2$  *on average thus far*” and was used to compare sequential forecasters in Choe and Ramdas [57]. This “weak null” should be contrasted with the “strong null” that would posit  $H_0^* : \forall t, \delta_t \leq 0$  — clearly, the latter implies the former, and hence any  $e$ -process (or anytime  $p$ -value) for  $H_0$  can also be used for  $H_0^*$ . Mathematically,  $H_0$  is a composite superset of the point null  $H_0^*$ . From a practical perspective,  $H_0$  may be a favorable null to test since it allows for  $\nu_t(\pi_1) > \nu_t(\pi_2)$  at various  $t$  as long as  $\frac{1}{t} \sum_{i=1}^t \nu_t(\pi_1) \leq \frac{1}{t} \sum_{i=1}^t \nu_t(\pi_2)$  for all  $t$ , whereas  $H_0^*$  requires  $\pi_1$  to be uniformly dominated by  $\pi_2$ .

Let us now derive an explicit  $e$ -process for  $H_0$ . Using the techniques of Section 6.3.1, define

$$\theta_t := w_t^{(1)} R_t - (1 - w_t^{(2)})(1 - R_t), \quad (6.38)$$

where  $w_t^{(1)} := \pi_1(A_t | X_t)/h_t(A_t | X_t)$  and  $w_t^{(2)} := \pi_2(A_t | X_t)/h_t(A_t | X_t)$  are the importance weights for policies  $\pi_1$  and  $\pi_2$ . As before, we note that  $\theta_t \geq -1$  and  $\mathbb{E}(\theta_t | \mathcal{H}_{t-1}) = \delta_t$ , and hence

$$\mathbb{E}\left(\frac{1}{t} \sum_{i=1}^t \theta_i \mid \mathcal{H}_{t-1}\right) = \Delta_t. \quad (6.39)$$

Given the above setup, we are ready to derive an  $e$ -process (and hence an anytime  $p$ -value) for

the weak null  $H_0$  (an illustration is provided in Figure 6.5).

**Proposition 6.3.2.** *Given contextual bandit data  $(X_t, A_t, R_t)_{t=1}^{\infty}$  and two target policies  $\pi_1$  and  $\pi_2$  that we would like to compare, define  $S_t(\Delta'_t)$  and  $V_t$  by*

$$S_t(\Delta'_t) := \frac{1}{2} \left( \sum_{i=1}^t \theta_i - t\Delta'_t \right) \quad \text{and} \quad (6.40)$$

$$V_t := \frac{1}{2} \sum_{i=1}^t (\theta_i - \hat{\theta}_{i-1})^2, \quad \text{where } \hat{\theta}_t := \frac{1}{2} \left[ \left( \frac{1}{t} \sum_{i=1}^t \theta_i \right) \wedge 1 \right] \quad (6.41)$$

Then, given any  $\rho > 0$ , we have that

$$M_t^{\text{EB}}(0) := \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{V_t + \rho} \right) {}_1F_1(1, V_t + \rho + 1, S_t(0) + V_t + \rho) \quad (6.42)$$

forms an  $e$ -process for  $H_0$ . Consequently,  $p_t := (1/M_t^{\text{EB}}) \wedge 1$  forms an anytime  $p$ -value for the weak null  $H_0 : \forall t, \Delta_t \leq 0$ , meaning  $\sup_{P \in H_0} P(\exists t : p_t \leq \alpha) \leq \alpha$ .

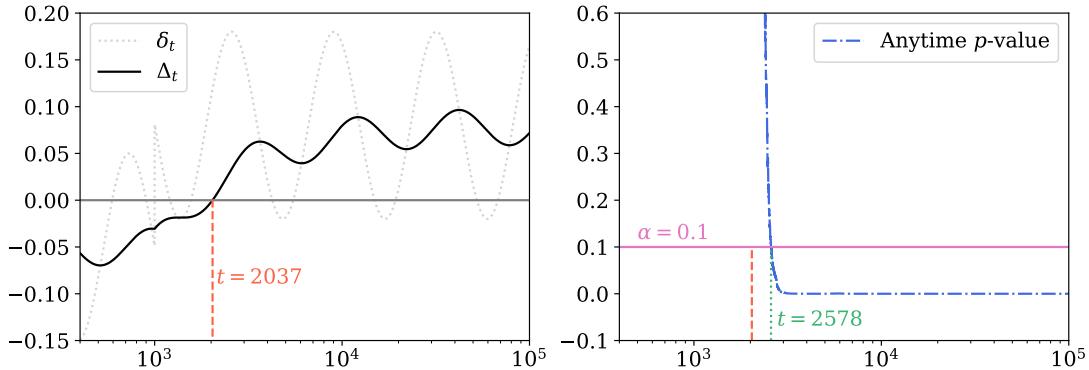


Figure 6.5: An illustration of how the anytime  $p$ -value derived in Proposition 6.3.2 can be used to test the weak null  $H_0 : \forall t, \Delta_t \leq 0$ . In the left-hand side plot, notice that  $\delta_t$  ventures above 0 at several points prior to  $t = 2037$ , but the average policy value difference is positive for the first time at  $t = 2037$ . In the right-hand side plot, we see that the anytime  $p$ -value dips below  $\alpha$  shortly after  $\Delta_t > 0$ , at which point the weak null can be safely rejected, with no penalties for the  $p$ -value having been continuously monitored.

The fact that  $M_t^{\text{EB}}(0)$  forms an  $e$ -process is easy to see: under  $H_0$ , we have that  $\Delta_t \leq 0$  and notice that  $M_t^{\text{EB}}$  with  $S_t(0)$  replaced with  $S_t(\Delta_t)$  forms a test supermartingale using the same techniques as Section 6.3. Since  $S_t(\cdot)$  is nonincreasing and  ${}_1F_1$  is nondecreasing in its third argument, we have that  $M_t^{\text{EB}}$  is upper-bounded by the aforementioned test supermartingale whenever  $\Delta_t \leq 0$ . The claimed  $e$ -process property is then an immediate consequence of the optional stopping theorem applied to the above test supermartingale.

## 6.4 Time-uniform inference for the off-policy CDF

Thus far we have focused on off-policy inference for *mean policy values*, i.e. functionals of the form  $\nu := \mathbb{E}_\pi(R)$ . In some cases, however, it may be of interest to study quantiles (e.g. median or 75<sup>th</sup> percentile) or perhaps the entire cumulative distribution function (CDF) of the reward distribution under policy  $\pi$ . In this section, we focus on the latter, deriving confidence bands for the CDF  $\mathbb{P}_\pi(R \leq r)$  of the reward  $R$  under policy  $\pi$ . Our confidence bands will be uniform in two senses: in time, and in the quantiles. Concretely, if  $Q(p)$  and  $Q^-(p)$  are the right (standard) and left quantiles, respectively – meaning  $Q(p) := \sup \{x \in \mathbb{R} : \mathbb{P}_\pi(R \leq x)\}$  and  $Q^-(p) := \sup \{x \in \mathbb{R} : \mathbb{P}_\pi(R < x)\}$  – then we will derive a sequence of confidence bands  $[L_t(p), U_t(p)]_{t \in \mathbb{N}}$  such that

$$\mathbb{P}(\forall t \in \mathbb{N}, p \in (0, 1), L_t(p) < Q^-(p) \text{ and } Q(p) < U_t(p)) \geq 1 - \alpha. \quad (6.43)$$

Such a guarantee enables anytime-valid inference at arbitrary stopping times for *all quantiles* simultaneously, (as well as any functional thereof). In addition, all our confidence bands will satisfy all five desiderata laid out in Section 6.1.2, and they will consistently shrink to the true quantile  $Q(p)$  for all  $p$ .

In order to state our main result, we first need to define a few terms. Define  $W_t, \bar{W}_t, \bar{q}_t(p)$ , and  $\ell_t(p; \alpha)$  given by

$$W_t := \sum_{i=1}^t w_i^2, \quad \bar{W}_t := W_t \vee 1, \quad (6.44)$$

$$\bar{q}_t(p) := \text{logit}^{-1} \left( \text{logit}(p) + 4\sqrt{\frac{e}{\bar{W}_t}} \right), \quad (6.45)$$

$$\ell_t(p; \alpha) := 2 \log (\log \bar{W}_t + 1) + 2 \log \left( \left| \left| \frac{\sqrt{\bar{W}_t} \text{logit}(p)}{4} \right| \right| \vee 1 \right) + \log \left( \frac{7.06}{\alpha} \right), \quad (6.46)$$

$$\text{and } \mathfrak{B}_t(p; \alpha) := \frac{\sqrt{2.13\ell_t(p; \alpha)\bar{W}_t + 1.76\bar{q}_t(p)^2\ell_t(p; \alpha)^2}}{t} \quad (6.47)$$

$$+ \frac{1.33\bar{q}_t(p)\ell_t(p; \alpha) + t(\bar{q}_t(p) - p)}{t}. \quad (6.48)$$

While some of the above may seem complicated, they arise naturally from the proof technique discussed below, and it is straightforward to implement them into code. Given the above setup, we are ready to state the main result of this section.

**Theorem 6.4.1** (Time-uniform confidence band for the off-policy CDF). *Consider a sequence of contextual bandit data  $(X_t, A_t, R_t)_{t=1}^\infty$  with real-valued (i.e. not necessarily  $[0, 1]$ -bounded) rewards. Let  $\hat{F}_t^\pi(x) := \frac{1}{t} \sum_{i=1}^t w_i \mathbb{1}(R_i \leq x)$  be the importance-weighted empirical CDF, and let*

$\hat{Q}_t(p)$  and  $\hat{Q}_t^-(p)$  be the upper and lower empirical quantiles, meaning

$$\hat{Q}_t(p) := \sup \left\{ x \in \mathbb{R} : \hat{F}_t^\pi(x) \leq p \right\}, \quad (6.49)$$

and similarly for  $\hat{Q}_t^-(p)$  with  $\leq$  in the above supremum replaced by a strict inequality  $<$ . Then,

$$\mathbb{P} \left( \forall t \in \mathbb{N}, p \in (0, 1), Q(p) < \hat{Q}_t^-([p + \mathfrak{B}_t(p; \alpha)] \wedge 1) \right) \geq 1 - \alpha. \quad (6.50)$$

Similarly, after applying the mirroring trick of Remark 13, we have that

$$\mathbb{P} \left( \forall t \in \mathbb{N}, p \in (0, 1), \hat{Q}_t \left( \left[ p + \frac{1}{t} \sum_{i=1}^t w_i - 1 - \mathfrak{B}_t(1-p; \alpha) \right] \vee 0 \right) < Q^-(p) \right) \geq 1 - \alpha. \quad (6.51)$$

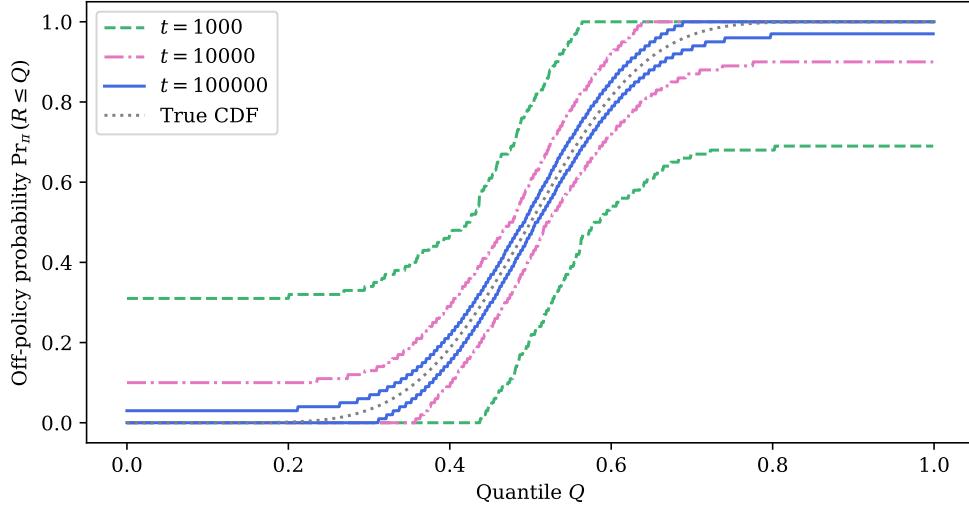


Figure 6.6: Time- and quantile-uniform 90% confidence band for the off-policy CDF  $\mathbb{P}_\pi(R \leq Q)$  in a Bernoulli(1/2) experiment with the target policy set to  $\pi(a | x) = \mathbb{1}(a = 1)$  – i.e. “always play action 1”. Here, the off-policy distribution of  $R$  is a beta distribution with  $\alpha = \beta = 10$ . These confidence bands are simultaneously valid for all  $Q \in \mathbb{R}$  and all  $t \in \mathbb{N}$  (though we only display them at  $t \in \{10^3, 10^4, 10^5\}$  above). In particular, notice that as  $t$  gets larger, the confidence bands shrink towards the true CDF (and will continue to do so in the limit).

The proof in Section 6.A.6 modifies the “double stitching” technique of Howard and Ramdas [123, Theorem 5] to handle importance-weighted observations, and relies on a sub-exponential concentration inequality rather than a sub-Bernoulli one. Notice that (6.50) and (6.51) could be written without  $\hat{Q}_t^-$  and  $Q^-$  – i.e. replacing  $\hat{Q}_t^-$  and  $Q^-$  with  $\hat{Q}_t$  and  $Q$ , respectively – but this would never result in a tighter bound. Illustrations of the time-uniform confidence bands derived in Theorem 6.4.1 are can be found in Figure 6.6.

Many of the CSs throughout this chapter have recovered prior CSs in the literature when specialized to the on-policy regime (that is, when all importance weights are set to 1 and reward predictors are set to 0). Examples include Theorem 6.2.1 recovering Theorem 3.4.1 from Chapter 3 from Chapter 3 or Theorem 6.3.1 recovering Howard et al. [125, Proposition 9]. However, Theorem 6.4.1 does not recover the on-policy bound it most resembles (Howard and Ramdas [123, Theorem 5]). The reason for this is subtle, and has to do with the fact that in the on-policy regime,  $\mathbb{1}(R_t \leq Q(p)) - p$  is a Bernoulli( $p$ ) random variable, hence their partial sums  $\frac{1}{t} \sum_{i=1}^t [\mathbb{1}(R_i \leq Q(p)) - p]$  form sub-Bernoulli processes [124, 125, 123] with variance process  $tp(1-p)$  regardless of the value of  $Q(p)$ . On the other hand, in the off-policy setting, we use importance-weighted indicators  $w_t \mathbb{1}(R_t \leq Q(p)) - p$  whose partial sums are not sub-Bernoulli, but are instead sub-exponential with a variance process that depends on  $Q(p)$ . This fundamental difference changes the test supermartingales that we have access to, and consequently alters the downstream CSs.

**Comparison with prior work.** There are three prior works that are related to the results of this section, namely those of Howard and Ramdas [123], Chandak et al. [51], and Huang et al. [128], but it is important to note that none of them solve the problem that we are studying – time-uniform confidence bands for off-policy CDFs – and hence we focus on a theoretical comparison with these prior works, rather than an empirical one. We discuss each of them in detail below and summarize how they compare with the present chapter in Table 6.2.

- **HR22:** Howard and Ramdas [123] derive time- and quantile-uniform confidence bands for the CDF of iid random variables in the *on-policy* setting, and in particular, Theorem 5 in their paper satisfies a guarantee of the form (6.43). However, Howard and Ramdas [123] do not consider the off-policy inference problem that we do here, and hence the “Predictable  $(h_t)_{t=1}^\infty$ ” and “ $w_{\max}$ -free” rows are not applicable (N/A). In addition, our setup can be seen as a generalization of theirs if all importance weights are set to 1.
- **UnO21:** In the paper entitled “Universal off-policy evaluation” (UnO), Chandak et al. [51] derive fixed-time quantile-uniform confidence bands for the off-policy CDF. Cleverly exploiting monotonicity of CDFs, they reduce the problem to computing finitely many confidence intervals for means of importance-weighted bounded random variables, and taking a union bound over them. Notably, their bounds do not require knowledge of  $w_{\max}$ .

The main difference between our bounds and those of Chandak et al. [51] is that ours are both time- and quantile-uniform, while theirs are only quantile-uniform. Note that Chandak et al. [51] do consider the more general setup of reinforcement learning in Markov Decision Processes (MDP) – an area to which we intend to extend all of our CSs in the future – and MDPs include the contextual bandit setting as a special case. When focusing on contextual bandits specifically, however, and even when ignoring time-uniformity, Theorem 6.4.1 improves on the fixed-time results of Chandak et al. [51] in the two following ways.

First, the confidence bands of Chandak et al. [51] are not guaranteed to shrink to the true CDF as the sample size grows to infinity while ours are (and at an explicit rate of

$O(\sqrt{\log t/t})$ , which we refer to as “consistency” in Table 6.2.

Second, their bounds assume that the logging policy  $h$  is fixed, while ours can take the form of a sequence of data-dependent logging policies  $(h_t)_{t=1}^\infty$ , such as those that result from online learning algorithms. However, since Chandak et al. [51, Theorem 2] simply requires taking a union bound over several CIs for importance-weighted bounded random variables, their Theorem 2 can presumably be extended to handle predictable logging policy sequences by employing the CIs provided in Corollaries 6.2.1 or 6.2.2.

- **HLLA21:** Huang et al. [128] derive impressive quantile-uniform confidence bands for the off-policy CDF. Their bounds are elegant and simple to state, resembling the famous Dvoretzky-Kiefer-Wolfowitz (DKW) inequalities and their sharpened forms [95, 186]. Notably, their bounds are consistent for the true CDF at  $O(1/\sqrt{n})$  rates. Similar to Chandak et al. [51], however, their results are not time-uniform, and hence do not permit valid inference at stopping times, unlike ours. Moreover, all of their bounds require knowledge of  $w_{\max}$  (and for this value to be finite), while our bounds do not. Finally, similar to Chandak et al. [51], their bounds assume that the logging policy  $h$  is fixed.

Table 6.2: Comparison of various uniform confidence bands for the CDF

	HR22	UnO21	HLLA21	Thm. 6.4.1
Off-policy		✓	✓	✓
Time-uniform	✓			✓
Consistency	✓		✓	✓
Predictable $(h_t)_{t=1}^\infty$	N/A			✓
$w_{\max}$ -free	N/A	✓		✓

## 6.5 Summary & extensions

This chapter derived time-uniform confidence sequences for various parameters in off-policy evaluation which remain valid even in contextual bandit setups where data are collected adaptively and sequentially over time. We began in Section 6.1.2 by laying out our desiderata for off-policy inference: we sought methods that (1) are exact and nonasymptotically valid, (2) only make nonparametric assumptions such as boundedness, (3) are time-uniform, and hence valid at arbitrary stopping times, (4) do not require knowledge of extreme values of importance weights, and (5) allow data to be collected by data-dependent logging policies.

In Section 6.2, we began by studying the most classical off-policy parameter – a fixed policy value  $\nu$  – and we derived CSs that strictly generalize prior state-of-the-art CSs by weakening the required assumptions and allowing for variance reduction via double robustness. In the same section, we also develop the first closed-form confidence sequences for policy values, as

well as some tight fixed-time confidence intervals that are instantiations of our time-uniform bounds. Section 6.3 then developed CSs for a more general parameter: the time-varying average policy value  $\tilde{\nu}_t := \frac{1}{t} \sum_{i=1}^t \nu_i$ , and we discussed what implications these bounds have for adaptive sequential experiments, such as online A/B tests.

Finally, in Section 6.4, we derived simultaneously valid CSs for every quantile of the off-policy reward distribution. Said differently, these bounds form time-uniform confidence bands for the CDF of the off-policy reward distribution.

There are a few other works that consider CSs and test supermartingales without reference to OPE or contextual bandits, but that nevertheless now have interesting consequences in the OPE problem once paired with the present chapter. In particular, we want to highlight the implications that Wang and Ramdas [275] and Xu et al. [290] have on false discovery rate control in OPE, and how Chapter 5 immediately yields algorithms for differentially private OPE. We briefly discuss these implications below, but omit their full derivations since these extensions are rather simple and not central to the current chapter.

**False discovery/coverage rate control under arbitrary dependence.** Suppose that rather than estimate a single policy value  $\nu$ , we are interested in a *collection*  $(\nu_1, \dots, \nu_J)$  containing the values of the policies  $(\pi_1, \dots, \pi_J)$ . When testing several hypothesis or constructing several CIs, etc., it is often of interest to control some multiple testing metric, such as the false discovery rate (FDR), or the false coverage rate (FCR), respectively [25, 26]. Rather surprisingly, Wang and Ramdas [275] and Xu et al. [290] show that for tests and CIs built from *e-values* – nonnegative test statistics with expectation at most one – the FDR and FCR can be controlled under *arbitrary* dependence with virtually no modification to the famous Benjamini-Hochberg [25] and Benjamini-Yekutieli [26] procedures, while this fact is not true for generic tests and CIs. Relevant to the current chapter, *all* of our CSs are fundamentally built from test supermartingales which form *e*-values at arbitrary stopping times. As a concrete consequence, we can take a collection of stopped CSs  $C_\tau^{(1)}, \dots, C_\tau^{(J)}$  for  $(\pi_j)_{j \in [J]}$ , adjust them via the *e*-BY procedure of Xu et al. [290] to produce  $\tilde{C}_\tau^{(1)}, \dots, \tilde{C}_\tau^{(J)}$ , so that the FCR is controlled at some desired level  $\delta \in (0, 1)$ . The ability control the FCR under arbitrary dependence is crucial for our setting since the CSs  $(C_\tau^{(j)})_{j \in [J]}$  are highly dependent and constructed from the same data, but with different importance weights. Similar implications hold for sequential tests and control of the FDR via the *e*-BH procedure of Wang and Ramdas [275].

**Locally differentially private off-policy evaluation in contextual bandits.** Chapter 5 developed nonparametric CSs and CIs for means of bounded random variables under privacy constraints. The authors developed a so-called “Nonparametric randomized response” (NPRR) mechanism that serves as a nonparametric generalization of Warner’s randomized response [278], mapping a  $[0, 1]$ -bounded random variable  $Y_t$  to a new random variable  $Z_t$  so that each  $Z_t$  is an  $\varepsilon$ -locally differentially private view of  $Y_t$  with mean  $r\mathbb{E}(Y_t) + (1 - r)/2$ , where  $r$  is a known quantity that depends on  $\varepsilon$  (and hence it is possible to work out what  $\mathbb{E}(Y_t)$  is). While Chapter 5 did not explicitly consider the contextual bandit setup, it did develop CSs for time-varying treatment effects in sequential experiments, similar to the discussion

in Section 6.3.2. However, like other prior work, those CSs require *a priori* knowledge of the minimal propensity score (in the language of this chapter: they require knowledge of  $w_{\max}$ ). Nevertheless, it is possible to derive locally private CSs for (time-varying) policy values without knowledge of  $w_{\max}$  using the techniques of the current chapter. Moreover, several policies can be evaluated from a *single* application of NPPR, thereby avoiding inflation of the privacy parameter  $\varepsilon$  from evaluating multiple policies. That is, given  $[0, 1]$ -bounded rewards  $(R_t)_{t=1}^{\infty}$ , we can use NPPR to generate private views  $(Z_t)_{t=1}^{\infty}$  of these rewards, and notice that  $\mathbb{E}(w_t Z_t) = \mathbb{E}_{A_t \sim \pi}(Z_t) = r\mathbb{E}_{A_t \sim \pi}(R_t) + (1 - r)/2$ , and hence a CS for  $\mathbb{E}(w_t Z_t)$  can be translated into a CS for  $\mathbb{E}_{A_t \sim \pi}(R_t)$  even though we only see a privatized version of  $R_t$ . In particular, practitioners can derive locally private CSs for time-varying policy values using Theorem 6.3.1 for several policies  $\pi_1, \dots, \pi_J$ , with only a single application of NPPR.

We believe that this chapter presents a comprehensive treatment of OPE inference, yielding procedures that are theoretically valid under more general settings and yet deliver state-of-the-art practical performance. A challenging open problem is to extend these techniques to the off-policy MDP (Markov Decision Process) setting, where the actions at each step affect subsequent covariate and reward distributions, as captured by state variables. Another important open problem is to design practical OPE inference methods not just for one policy, but uniformly over an entire family of policies.

## 6.A Proofs of the main results

### 6.A.1 A technical lemma

**Lemma 6.A.1.** *Let  $Z$  and  $\hat{Z}$  be  $\mathcal{H}$ -adapted processes such that  $Z_t - \hat{Z}_{t-1} \geq -1$  almost surely for all  $t$ . Denoting  $\mu_t := \mathbb{E}(Z_t | \mathcal{H}_{t-1})$ , we have that for any  $(0, 1)$ -valued predictable process  $(\lambda_t)_{t=1}^{\infty}$ ,*

$$M_t := \exp \left( \sum_{i=1}^t \lambda_i (Z_t - \mu_t) - \sum_{i=1}^t \psi_E(\lambda_i) (Z_t - \hat{Z}_{t-1})^2 \right), \quad (6.52)$$

*forms a test supermartingale, where  $\psi_E(\lambda) := -\log(1 - \lambda) - \lambda$ .*

Above,  $\hat{Z}_{t-1}$  is to be interpreted as an estimator of  $Z_t$  using the first  $t - 1$  samples. Closely related lemmas have appeared in [102, 124, 125, 282], but those papers assumed  $Z_t - \mu_t \geq -1$ , which does not suffice for our purposes. What is somewhat surprising above is that we do not require a particular lower bound on  $Z_t$  or an upper bound on  $\mu_t$ , as long as  $Z_t - \hat{Z}_{t-1} \geq -1$ .

*Proof.* First, note that  $M_0 \equiv 1$  by construction, and  $M_t$  is always positive. It remains to show that  $M_t$  forms a supermartingale. Writing out the conditional expectation of  $M_t$  given  $\mathcal{H}_{t-1}$ , we have that

$$\mathbb{E}(M_t | \mathcal{H}_{t-1}) = M_{t-1} \underbrace{\mathbb{E} \left( \exp \left\{ \lambda_t (Z_t - \mu_t) - \psi_E(\lambda_t) (Z_t - \hat{Z}_{t-1})^2 \right\} | \mathcal{H}_{t-1} \right)}_{(\dagger)}, \quad (6.53)$$

and hence it suffices to prove that  $(\dagger) \leq 1$ . Denote for the sake of succinctness,

$$Y_t := Z_t - \mu_t \quad \text{and} \quad \delta_t := \hat{Z}_{t-1} - \mu_t,$$

and note that  $\mathbb{E}(Y_t | \mathcal{H}_{t-1}) = 0$ . Using the proof of Fan et al. [102, Proposition 4.1], we have that  $\exp\{b\lambda - b^2\psi_E(\lambda)\} \leq 1 + b\lambda$  for any  $\lambda \in [0, 1)$  and  $b \geq -1$ . Setting  $b := Y_t - \delta_t = Z_t - \hat{Z}_{t-1}$ ,

$$\begin{aligned} & \mathbb{E} \left[ \exp \left\{ \lambda_t Y_t - (Y_t - \delta_t)^2 \psi_E(\lambda_t) \right\} \mid \mathcal{H}_{t-1} \right] \\ &= \mathbb{E} \left[ \exp \left\{ \lambda_t (Y_t - \delta_t) - (Y_t - \delta_t)^2 \psi_E(\lambda_t) \right\} \mid \mathcal{H}_{t-1} \right] \exp(\lambda_t \delta_t) \\ &\leq \mathbb{E} [1 + (Y_t - \delta_t) \lambda_t \mid \mathcal{H}_{t-1}] \exp(\lambda_t \delta_t) \\ &= \mathbb{E} [1 - \delta_t \lambda_t \mid \mathcal{H}_{t-1}] \exp(\lambda_t \delta_t) \leq 1, \end{aligned}$$

where the last line follows from the fact that  $Y_t$  is conditionally mean zero and the inequality  $1 - x \leq \exp(-x)$  for all  $x \in \mathbb{R}$ . This completes the proof.  $\square$

### 6.A.2 Proof of Theorem 6.2.1

We will only derive the lower CS for  $\nu$ , since the upper CS follows analogously. Consider the process  $(M_t(\nu))_{t=1}^\infty$  given by

$$M_t(\nu) := \prod_{i=1}^t \left[ 1 + \lambda_i^L(\nu) \cdot (\phi_i^{(\text{DR}-\ell)} - \nu) \right]. \quad (6.54)$$

The proof proceeds in three steps, following the strategy of [125], Chapter 3, and [144]. In Step 1, we show that the pseudo-outcomes have conditional mean  $\nu$ , i.e.  $\mathbb{E}(\phi_t^{(\text{DR}-\ell)} \mid \mathcal{H}_{t-1}) = \mathbb{E}_\pi(R_t \mid \mathcal{H}_{t-1}) = \nu$ . In Step 2, we use Step 1 to show that  $M_t(\nu)$  forms a test martingale and apply Ville's inequality to it. In Step 3, we “invert” this test martingale to obtain the lower CS found in Theorem 6.2.1.

**Step 1: Computing the conditional mean of the doubly robust pseudo-outcomes.** Writing out the conditional expectation of  $\phi_t^{(\text{DR}-\ell)}$ , we have

$$\begin{aligned} & \mathbb{E}[\phi_t^{(\text{DR}-\ell)} \mid \mathcal{H}_{t-1}] \\ &= \mathbb{E}(w_t R_t \mid \mathcal{H}_{t-1}) - \mathbb{E} \left\{ w_t \cdot \left( \hat{r}_t(X_t; A_t) \wedge \frac{k_t}{w_t} \right) - \mathbb{E}_{a \sim \pi(\cdot \mid X_t)} \left( \hat{r}_t(X_t; a) \wedge \frac{k_t}{w_t} \right) \mid \mathcal{H}_{t-1} \right\} \\ &= \mathbb{E}(w_t R_t \mid \mathcal{H}_{t-1}) \\ &= \int_{(x,a,r)} \frac{\pi(a \mid x)}{h_t(a \mid x)} r \cdot p_{R_t}(r \mid a, x, \mathcal{H}_{t-1}) h_t(a \mid x) p_{X_t}(x \mid \mathcal{H}_{t-1}) dx da dr \\ &= \int_{(x,a,r)} r \cdot p_{R_t}(r \mid a, x, \mathcal{H}_{t-1}) \pi(a \mid x) p_{X_t}(x \mid \mathcal{H}_{t-1}) dx da dr \end{aligned}$$

$$= \mathbb{E}_\pi(R_t \mid \mathcal{H}_{t-1}) = \nu.$$

**Step 2: Showing that  $M_t(\nu)$  forms a test martingale.** First, note that  $M_0 \equiv 1$  by construction. To show that  $M_t$  is nonnegative, notice that since  $R_t, \hat{r}_t \in [0, 1]$  almost surely, we have that  $\phi_t^{(\text{DR}-\ell)} \geq -k_t$ . Therefore, for any  $\nu \in [0, 1]$ ,

$$\begin{aligned} 1 + \lambda_t^L(\nu) \cdot (\phi_t^{(\text{DR}-\ell)} - \nu) &\geq 1 + \lambda_t^L(\nu) \cdot (-k_t - \nu) \\ &> 1 + \frac{-k_t - \nu}{k_t + \nu} \quad (\text{since } \lambda_t^L(\nu) \in [0, (\nu + k_t)^{-1}]) \\ &= 0. \end{aligned}$$

Lastly, it remains to show that  $\mathbb{E}[M_t(\nu) \mid \mathcal{H}_{t-1}] = M_{t-1}(\nu)$ . Writing out the conditional expectation of  $M_t(\nu)$ , we have

$$\begin{aligned} \mathbb{E}[M_t(\nu) \mid \mathcal{H}_{t-1}] &= \mathbb{E}\left[M_{t-1}(\nu) \left\{1 + \lambda_t^L(\nu) \cdot (\phi_t^{(\text{DR}-\ell)} - \nu)\right\} \mid \mathcal{H}_{t-1}\right] \\ &= M_{t-1}(\nu) \cdot \left[1 + \lambda_t^L(\nu) \cdot \mathbb{E}\left\{(\phi_t^{(\text{DR}-\ell)} - \nu) \mid \mathcal{H}_{t-1}\right\}\right] \\ &= M_{t-1}(\nu) \cdot (1 + \lambda_t^L(\nu) \cdot 0) = M_{t-1}(\nu), \end{aligned}$$

where the second line follows from the fact that  $M_{t-1}, \lambda_t^L$  are predictable, and the third line follows from Step 1. Therefore, by Ville's inequality for nonnegative supermartingales [264], we have

$$\mathbb{P}\left(\exists t \in \mathbb{N}, M_t(\nu) \geq \frac{1}{\alpha}\right) \leq \alpha. \quad (6.55)$$

**Step 3: Inverting Ville's inequality to obtain a lower CS.** Recall the lower boundary given by (6.12),

$$L_t^{\text{DR}} := \inf \left\{ \nu' \in [0, 1] : \prod_{i=1}^t \left[1 + \lambda_i^L(\nu') \cdot (\phi_i^{(\text{DR}-\ell)} - \nu')\right] < \frac{1}{\alpha} \right\}$$

and notice that if  $\nu < L_t^{\text{DR}}$ , then  $M_t(\nu) \geq 1/\alpha$  by definition of  $L_t^{\text{DR}}$ . Consequently,

$$\mathbb{P}(\exists t \in \mathbb{N}, \nu < L_t^{\text{DR}}) \leq \mathbb{P}\left(\exists t \in \mathbb{N}, M_t(\nu) \geq \frac{1}{\alpha}\right) \leq \alpha.$$

Therefore, we have  $\mathbb{P}(\forall t \in \mathbb{N}, \nu \geq L_t^{\text{DR}}) \geq 1 - \alpha$ , so  $L_t^{\text{DR}}$  forms a lower  $(1 - \alpha)$ -CS for  $\nu$ , which completes the proof.  $\square$

### 6.A.3 Proof of Theorem 6.3.1

*Proof of Theorem 6.3.1.* The proof proceeds in three steps, following the high level outline of the conjugate mixture method in [125]. First, we invoke Lemma 6.A.1 to derive a test supermartingale for each  $\lambda \in (0, 1)$ . Second, we mix over  $\lambda \in (0, 1)$  using the truncated gamma density to obtain (6.26). Third and finally, we invert this test supermartingale to obtain a lower CS for  $\tilde{\nu}_t$ .

**Step 1: Deriving a test supermartingale indexed by  $\lambda \in (0, 1)$ .** Let  $Z_t := \xi_t$  and  $\hat{Z}_{t-1} := \hat{\xi}_{t-1}$  as in the setup of Theorem 6.3.1. First, notice that  $\mathbb{E}(\xi_t | \mathcal{H}_{t-1}) = \nu_t$ :

$$\mathbb{E}(\xi_t | \mathcal{H}_{t-1}) = \mathbb{E}(w_t R_t | \mathcal{H}_{t-1}) \quad (6.56)$$

$$= \int_{x,a,r} \frac{\pi(a|x)}{h_t(a|x)} r p_{R_t}(r|a,x,\mathcal{H}_{t-1}) h_t(a|x) p_{X_t}(x|\mathcal{H}_{t-1}) dx da dr. \quad (6.57)$$

Notice that  $\xi_t - \hat{\xi}_{t-1} \geq -1$ , and hence by Lemma 6.A.1, we have that for any  $\lambda \in (0, 1)$

$$M_t(\tilde{\nu}_t; \lambda) := \exp\{\lambda S_t(\tilde{\nu}_t) - V_t \psi_E(\lambda)\}$$

forms a test supermartingale.

**Step 2: Mixing over  $\lambda$  using the truncated gamma density.** For any distribution  $F$  on  $(0, 1)$ ,

$$M_t^{\text{EB}}(\tilde{\nu}_t) := \int_{\lambda \in (0,1)} M_t(\tilde{\nu}_t; \lambda) dF(\lambda) \quad (6.58)$$

forms a test supermartingale by Fubini's theorem. In particular, we will use the truncated gamma density  $f(\lambda)$  given by

$$f(\lambda) = \frac{\rho^\rho e^{-\rho(1-\lambda)} (1-\lambda)^{\rho-1}}{\Gamma(\rho) - \Gamma(\rho, \rho)}, \quad (6.59)$$

as the mixing density. Writing out  $M_t(\nu)$  using  $dF(\lambda) := f(\lambda)d\lambda$ , we have

$$\begin{aligned} M_t^{\text{EB}}(\tilde{\nu}_t) &:= \int_0^1 \exp\{\lambda S_t(\tilde{\nu}_t) - V_t \psi_E(\lambda)\} f(\lambda) d\lambda \\ &= \int_0^1 \exp\{\lambda S_t(\tilde{\nu}_t) - V_t \psi_E(\lambda)\} \frac{\rho^\rho e^{-\rho(1-\lambda)} (1-\lambda)^{\rho-1}}{\Gamma(\rho) - \Gamma(\rho, \rho)} d\lambda \\ &= \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \int_0^1 \exp\{\lambda(\rho + S_t + V_t)\} (1-\lambda)^{V_t+\rho-1} d\lambda \\ &= \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{V_t + \rho} \right) \left( \frac{\Gamma(b)}{\Gamma(a)\Gamma(b-a)} \int_0^1 e^{zu} u^{a-1} (1-u)^{b-a-1} du \right) \Big|_{\substack{b=V_t+\rho+1 \\ z=S_t+V_t+\rho}} \end{aligned}$$

$$= \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{V_t + \rho} \right) {}_1F_1(1, V_t + \rho + 1, S_t + V_t + \rho),$$

which completes this step.

**Step 3: Inverting the mixture test supermartingale to obtain (6.26).** Similar to Step 3 of the proof of Theorem 6.2.1, we have that  $\tilde{\nu}_t < L_t^{\text{EB}}$  if and only if  $M_t(\tilde{\nu}_t) \geq 1/\alpha$ , and hence by Ville's inequality for nonnegative supermartingales, we have that

$$\mathbb{P}(\exists t : \tilde{\nu}_t < L_t^{\text{EB}}) = \mathbb{P}(\exists t : M_t^{\text{EB}}(\tilde{\nu}_t) \geq 1/\alpha) \leq \alpha,$$

and hence  $L_t^{\text{EB}}$  forms a lower  $(1 - \alpha)$ -CS for  $\tilde{\nu}_t$ . This completes the proof.  $\square$

*Remark 15* (Writing (6.26) in terms of the lower incomplete gamma function). For readers familiar with Howard et al. [125, Proposition 9], we can rewrite (6.26) in terms of the lower incomplete gamma function via the identity  ${}_1F_1(1, b, z) = (b-1)e^z z^{1-b}(\Gamma(b-1) - \Gamma(b-1, z))$ , resulting in

$$\begin{aligned} & \left( \frac{\rho^\rho e^{-\rho}}{\Gamma(\rho) - \Gamma(\rho, \rho)} \right) \left( \frac{1}{v + \rho} \right) {}_1F_1(1, v + \rho + 1, s + v + \rho) \\ &= \left( \frac{\rho^\rho}{\Gamma(\rho)\gamma(\rho, \rho)} \right) \frac{\Gamma(v + \rho)\gamma(v + \rho, s + v + \rho)}{(s + v + \rho)^{v+\rho}} \exp(s + v), \end{aligned}$$

where  $\gamma(\cdot, \cdot)$  is the lower regularized incomplete gamma function and  $v = V_t$  and  $s = S_t(\tilde{\nu}_t)$ . This matches Howard et al. [125, Eq. (66)] when setting  $c = 1$ . The final representation above is real-valued after some complex terms are cancelled (in the case where  $(s + v + \rho)$  is negative), but the representation in terms of  ${}_1F_1(1, \cdot, \cdot)$  sidesteps this subtlety altogether, which is why we prefer to use it in Theorem 6.3.1.

#### 6.A.4 Proof of Proposition 6.2.2

*Proof.* Consider the process  $M \equiv (M_t)_{t=1}^\infty$  given by

$$M_t := \exp \left\{ \sum_{i=1}^t \lambda_i \left( \xi_i - \frac{\nu}{k_i + 1} \right) - \sum_{i=1}^t \left( \xi_i - \hat{\xi}_{i-1} \right)^2 \psi_E(\lambda_i) \right\}. \quad (6.60)$$

Then by Lemma 6.A.1, we have that  $M$  is a test supermartingale, and hence by Ville's inequality,  $\mathbb{P}(\exists t : M_t \geq 1/\alpha) \leq \alpha$ . Inverting this time-uniform concentration inequality, we have that with probability at least  $(1 - \alpha)$  and for all  $t \in \mathbb{N}$ ,

$$\begin{aligned} M_t < 1/\alpha &\iff \exp \left\{ \sum_{i=1}^t \lambda_i \left( \xi_i - \frac{\nu}{k_i + 1} \right) - \sum_{i=1}^t \left( \xi_i - \hat{\xi}_{i-1} \right)^2 \psi_E(\lambda_i) \right\} < \frac{1}{\alpha} \\ &\iff \sum_{i=1}^t \lambda_i \left( \xi_i - \frac{\nu}{k_i + 1} \right) - \sum_{i=1}^t \left( \xi_i - \hat{\xi}_{i-1} \right)^2 \psi_E(\lambda_i) < \log(1/\alpha) \end{aligned}$$

$$\begin{aligned} &\iff \sum_{i=1}^t \lambda_i \xi_i - \nu \sum_{i=1}^t \frac{\lambda_i}{k_i + 1} - \sum_{i=1}^t (\xi_i - \hat{\xi}_{i-1})^2 \psi_E(\lambda_i) < \log(1/\alpha) \\ &\iff \nu > \frac{\sum_{i=1}^t \lambda_i \xi_i}{\sum_{i=1}^t \lambda_i / (k_i + 1)} - \frac{\log(1/\alpha) + \sum_{i=1}^t (\xi_i - \hat{\xi}_{i-1})^2 \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i / (k_i + 1)}, \end{aligned}$$

which completes the proof.  $\square$

### 6.A.5 Proof of Proposition 6.3.1

We will prove a more general result below for arbitrary  $\eta, s > 1$ , but the exact constants in Proposition 6.3.1 can be obtained by setting  $\eta = e, s = 2$ . By Lemma 6.A.1 combined with Howard et al. [124, Table 5, row 7], we have that  $S_t(\tilde{\nu}_t)$  is a sub-gamma process with scale parameter  $c = 1$ , meaning for any  $\lambda \in [0, 1]$

$$M_t^G(\lambda) := \exp \{ \lambda S_t(\tilde{\nu}_t) - V_t \psi_G(\lambda) \}, \quad (6.61)$$

where  $\psi_G(\lambda) \equiv \psi_{G,1}(\lambda) = \frac{\lambda^2}{2(1-\lambda)}$ . Define the following parameters:

$$\begin{aligned} \lambda_k &:= \psi^{-1}(\log(1/\alpha)/\eta^{k+1/2}), \text{ where } \psi_G^{-1}(a) := \frac{2}{1 + \sqrt{1 + 2/a}}, \\ \alpha_k &:= \frac{\alpha}{(k+1)^s \zeta(s)}, \text{ and} \\ b_{t,k} &:= \frac{V_t \psi_G(\lambda_k) + \log(1/\alpha_k)}{\lambda_k}. \end{aligned}$$

Taking a union bound over  $k \in \mathbb{N}$ , we have that

$$\mathbb{P}(\forall t \in \mathbb{N}, k \in \mathbb{N}, S_t(\tilde{\nu}_t) \leq b_{t,k}) \geq 1 - \alpha.$$

It remains to find a deterministic upper bound on  $b_{t,k}$  that does not depend on  $k$ . Indeed, similar to Howard et al. [125, Eq. (39)], we have that

$$b_{t,k} = A \left( \frac{\log(1/\alpha_k)}{\eta^{k+1/2}} \right) \underbrace{\left[ \sqrt{\frac{\eta^{k+1/2}}{V_t}} + \sqrt{\frac{V_t}{\eta^{k+1/2}}} \right]}_{(*)} \sqrt{\frac{\log(1/\alpha_k)V_t}{2}}, \quad (6.62)$$

where  $A(a) := \sqrt{2a}/\psi_G^{-1}(a) = \sqrt{1+a/2} + \sqrt{a/2}$ . Now, notice that  $(*)$  is convex in  $V_t$  on  $V_t \in [\eta^k, \eta^{k+1}]$ , and hence  $(*)$  is maximized at the endpoints  $\eta^k$  and  $\eta^{k+1}$ . Consequently, on the  $k^{\text{th}}$  epoch – i.e. when  $\eta^k \leq V_t \leq \eta^{k+1}$  – we have that

$$b_{t,k} \leq A \left( \frac{\log(1/\alpha_k)}{\eta^{k+1/2}} \right) \left[ \eta^{-1/4} + \eta^{1/4} \right] \sqrt{\frac{\log(1/\alpha_k)V_t}{2}}$$

$$\begin{aligned}
&\leq A \left( \frac{\sqrt{\eta} \log(1/\alpha_k)}{V_t} \right) \left[ \eta^{-1/4} + \eta^{1/4} \right] \sqrt{\frac{\log(1/\alpha_k)V_t}{2}} \\
&= \left[ \sqrt{1 + \frac{\sqrt{\eta} \log(1/\alpha_k)}{2V_t}} + \sqrt{\frac{\sqrt{\eta} \log(1/\alpha_k)}{2V_t}} \right] \cdot \left[ \eta^{-1/4} + \eta^{1/4} \right] \cdot \sqrt{\frac{\log(1/\alpha_k)V_t}{2}} \quad (6.63)
\end{aligned}$$

where the first inequality follows from our analysis of  $(\star)$ , the second follows from monotonicity of  $A(\cdot)$  and the fact that  $V_t \leq \eta^{k+1}$  on the  $k^{\text{th}}$  epoch, the third follows from the definition of  $A(\cdot)$ . Rewriting the final line (6.63) more succinctly, we have the following upper bound on  $b_{t,k}$  for every  $k \in \mathbb{N}$ ,

$$b_{t,k} \leq \sqrt{\gamma_1^2 \log(1/\alpha_k)V_t + \gamma_2^2 \log^2(1/\alpha_k)} + \gamma_2 \log(1/\alpha_k), \quad (6.64)$$

$$\text{where } \gamma_1 := \frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}}, \text{ and } \gamma_2 := \frac{\sqrt{\eta} + 1}{2}. \quad (6.65)$$

Now, notice that the above bound only depends on  $k$  through  $\log(1/\alpha_k)$ . As such, we will upper bound  $\log(1/\alpha_k)$  solely in terms of  $V_t$  and other constants. Indeed, on the  $k^{\text{th}}$  epoch, we have

$$\begin{aligned}
\log(1/\alpha_k) &\equiv \log \left( \frac{(k+1)^s \zeta(s)}{\alpha} \right) = s \log(k+1) + \log \left( \frac{\zeta(s)}{\alpha} \right) \\
&\leq s \log(\log_\eta V_t + 1) + \log \left( \frac{\zeta(s)}{\alpha} \right) \equiv \ell_t
\end{aligned} \quad (6.66)$$

where the final line used the upper bound  $k \leq \log_\eta V_t$  which follows because  $\eta^k \leq V_t$  on the  $k^{\text{th}}$  epoch. Combining (6.64) and (6.66), we have that

$$b_{t,k} \leq \sqrt{\gamma_1^2 \ell_t V_t + \gamma_2^2 \ell_t^2} + \gamma_2 \ell_t, \text{ where } \ell_t := s \log(\log_\eta V_t + 1) + \log \left( \frac{\zeta(s)}{\alpha} \right), \quad (6.67)$$

which no longer depends on  $k$ . Consequently, we have that

$$\begin{aligned}
1 - \alpha &\leq \mathbb{P} \left( \forall t \in \mathbb{N}, \sum_{i=1}^t \xi_i - \sum_{i=1}^t \nu_i \leq \sqrt{\gamma_1^2 \ell_t V_t + \gamma_2^2 \ell_t^2} + \gamma_2 \ell_t \right) \\
&= \mathbb{P} \left( \forall t \in \mathbb{N}, \tilde{\nu}_t \geq \underbrace{\frac{1}{t} \sum_{i=1}^t \xi_i - \frac{\sqrt{\gamma_1^2 \ell_t V_t + \gamma_2^2 \ell_t^2}}{t} - \frac{\gamma_2 \ell_t}{t}}_{(\dagger)} \right),
\end{aligned}$$

and hence  $(\dagger)$  forms a lower  $(1 - \alpha)$ -CS for  $\tilde{\nu}_t$ .

### 6.A.6 Proof of Theorem 6.4.1

We will prove a more general result below for arbitrary  $\eta, s, \delta > 1$ , but the exact constants in Proposition 6.3.1 can be obtained by setting  $\eta = e$ , and  $s = \delta = 2$ . The proof will proceed in five steps. First, we derive an exponential  $e$ -process — i.e. an adapted process upper-bounded by a test supermartingale — from  $S_t(p) := \sum_{i=1}^t w_i \mathbf{1}(R_i \leq Q(p)) - tp$ . Second, we apply Ville's inequality to the aforementioned  $e$ -process to obtain a level- $\alpha$  linear boundary  $b_t(p)$  on  $S_t(p)$ , meaning  $\mathbb{P}(\exists t \in \mathbb{N} : S_t(p) \geq b_t(p)) \leq \alpha$ . Third, we derive one level- $\alpha_{k,j}$  linear boundary for each  $k \in \mathbb{N}, j \in \mathbb{Z}$  using the techniques of Step 2 so that  $\sum_{k \in \mathbb{N}} \sum_{j \in \mathbb{Z}} \alpha_{k,j} \leq \alpha$  and take a union bound over all of them. Here,  $k \in \mathbb{N}$  will index exponentially spaced epochs of time  $t \in \mathbb{N}$ , while  $j \in \mathbb{Z}$  will index evenly-spaced log-odds of  $p \in (0, 1)$ . Fourth, we modify the boundaries derived in Step 3 to obtain a boundary that is uniform in both  $t \in \mathbb{N}$  and in  $p \in (0, 1)$ . Fifth and finally, we obtain an analytic upper bound on the boundary derived in Step 4.

At several points throughout the proof, we will make use of various functions that depend on  $k$  and  $j$ . While we will define them as they are needed, we also list them here for reference.

$$W_t := \sum_{i=1}^t w_i^2, \quad (6.68a)$$

$$\alpha_{k,j} := \frac{\alpha}{(k+1)^s (|j| \vee 1)^s \zeta(s) (2\zeta(s) + 1)}, \quad (6.68b)$$

$$q(k, j) := \frac{1}{1 + \exp\{-2j\delta/\eta^{k/2}\}}, \quad (6.68c)$$

$$j(k, p) := \left\lceil \frac{\eta^{k/2} \text{logit}(p)}{2\delta} \right\rceil, \quad (6.68d)$$

$$\lambda(k, j) := \psi_{G,q(k,j)}^{-1}(\log(1/\alpha_{k,j})/\eta^{k+1/2}), \quad \text{where } \psi_{G,c}^{-1}(a) := \frac{2}{c + \sqrt{c^2 + 2/a}}, \quad \text{and} \quad (6.68e)$$

$$b_{t,k}(p) := \frac{W_t \psi_{G,p}(\lambda_{k,j}) + \log(1/\alpha_{k,j})}{\lambda_{k,j}}. \quad (6.68f)$$

**Step 1: Deriving an  $e$ -process.** Invoking Lemma 6.A.1 combined with Howard et al. [124, Table 5, row 7] we have that for any  $p \in (0, 1)$ ,  $S_t(p)$  is sub-gamma [124, 125] with variance process  $V_t(p) := \sum_{i=1}^t (w_i \mathbf{1}\{R_i \leq Q(p)\})^2$  and scale  $c = p$ , meaning we have that for any  $\lambda \in [0, 1/c]$ ,

$$M_t^G(\lambda; p) := \exp\{\lambda S_t(p) - V_t(p) \psi_{G,p}(\lambda)\} \quad (6.69)$$

forms a test supermartingale. Now, since  $V_t(p) \leq \sum_{i=1}^t w_i^2 \equiv W_t$  almost surely, we have that

$$E_t^G(\lambda; p) := \exp\{\lambda S_t(p) - W_t \psi_{G,p}(\lambda)\} \leq M_t^G(\lambda; p) \quad (6.70)$$

forms an  $e$ -process — i.e. it is upper-bounded by a test supermartingale. This completes the first step of the proof.

**Step 2: Applying Ville's inequality to  $E_t^G(\lambda; p)$ , yielding a time-uniform linear boundary.** In Step 1, we showed that  $E_t^G(\lambda; p)$  forms an  $e$ -process. By Ville's maximal inequality for nonnegative supermartingales [264], we have that

$$\mathbb{P}(\exists t \in \mathbb{N} : E_t^G(\lambda; p) \geq 1/\alpha) \leq \mathbb{P}(\exists t \in \mathbb{N} : M_t^G(\lambda; p) \geq 1/\alpha) \leq \alpha. \quad (6.71)$$

Now, we will rewrite the inequality  $E_t^G(\lambda; p) \geq 1/\alpha$  slightly more conveniently so that we can derive a time-uniform concentration inequality for  $S_t(p)$ . Indeed,

$$\begin{aligned} E_t^G(\lambda; p) \geq 1/\alpha &\iff \lambda S_t(p) - W_t \psi_{G,p}(\lambda) \geq \log(1/\alpha) \\ &\iff S_t(p) \geq \underbrace{\frac{W_t \psi_{G,p}(\lambda) + \log(1/\alpha)}{\lambda}}_{b_t(p)}. \end{aligned}$$

In summary, we have the following time-uniform concentration inequality on  $S_t(p)$  for any  $p \in (0, 1)$ ,  $\alpha \in (0, 1)$  and  $\lambda \in [0, 1/p]$ ,

$$\mathbb{P}(\exists t \in \mathbb{N} : S_t(p) \geq b_t(p)) \leq \alpha, \quad \text{where } b_t(p) := \frac{W_t \psi_{G,p}(\lambda) + \log(1/\alpha)}{\lambda}, \quad (6.72)$$

which could also be written as a time-uniform high-probability upper bound on  $S_t(p)$ :

$$\mathbb{P}(\forall t \in \mathbb{N}, S_t(p) < b_t(p)) \geq 1 - \alpha. \quad (6.73)$$

**Step 3: Union-bounding over infinitely many choices of  $\lambda$ ,  $\alpha$ , and  $p$ .** In Step 2, we showed that  $b_t(p)$  forms a time-uniform high-probability upper bound for  $S_t(p)$ . We will now take a union bound over a countably infinite two-dimensional grid of  $t$  and  $p$ . Concretely, for each  $k \in \mathbb{N}$  and  $j \in \mathbb{Z}$ , recall  $\alpha_{k,j}$ ,  $q(k, j)$ , and  $\lambda(k, j)$  as in (6.68b), (6.68c), and (6.68e). The exact choices of  $q(k, j)$  and  $\lambda(k, j)$  will become relevant later. For now, note that by (6.72) from Step 2 combined with a union bound, we have that

$$\mathbb{P}(\exists t \in \mathbb{N}, k \in \mathbb{N}, j \in \mathbb{Z} : S_t(q(k, j)) \geq b_{t,k}(q(k, j))) \leq \sum_{k \in \mathbb{N}} \sum_{j \in \mathbb{Z}} \alpha_{k,j}, \quad (6.74)$$

$$\text{where } b_{t,k}(q(k, j)) := \frac{W_t \psi_{G,q(k,j)}(\lambda_{k,j}) + \log(1/\alpha_{k,j})}{\lambda_{k,j}}. \quad (6.75)$$

We will now show that  $\sum_{k \in \mathbb{N}} \sum_{j \in \mathbb{Z}} \alpha_{k,j} = \alpha$  so that (6.74) holds with probability at most  $\alpha$ . Indeed,

$$\begin{aligned} \sum_{k \in \mathbb{N}} \sum_{j \in \mathbb{Z}} \alpha_{k,j} &= \frac{\alpha}{\zeta(s)(2\zeta(s) + 1)} \sum_{k \in \mathbb{N}} \frac{1}{(k+1)^s} \sum_{j \in \mathbb{Z}} \frac{1}{(|j| \vee 1)^s} \\ &= \frac{\alpha}{\zeta(s)(2\zeta(s) + 1)} \underbrace{\sum_{k=0}^{\infty} \frac{1}{(k+1)^s}}_{=\zeta(s)} \underbrace{\left(1 + 2 \sum_{m=1}^{\infty} \frac{1}{m^s}\right)}_{=2\zeta(s)+1} \end{aligned}$$

$$= \alpha.$$

Therefore, in summary, we have that

$$\mathbb{P}(\exists t \in \mathbb{N}, k \in \mathbb{N}, j \in \mathbb{Z} : S_t(q(k, j)) \geq b_{t,k}(q(k, j))) \leq \alpha. \quad (6.76)$$

**Step 4: Removing dependence on  $j \in \mathbb{Z}$  and obtaining  $p$ -uniformity.** We will obtain a bound that is uniform in  $p \in (0, 1)$  by replacing  $j$  with  $j(k, p)$  — a function of both  $k \in \mathbb{N}$  and  $p \in (0, 1)$ . For any  $k \in \mathbb{N}$  and any  $p \in (0, 1)$ , define  $j(k, p)$  as

$$j(k, p) := \left\lceil \frac{\eta^{k/2} \text{logit}(p/(1-p))}{2\delta} \right\rceil. \quad (6.77)$$

Of course,  $j(k, p) \in \mathbb{Z}$  is not unique. It is easy to check that  $p \leq q(k, j(k, p))$ , a fact that we will use shortly. Abusing notation slightly, let  $j_1, j_2, \dots$  denote the integers generated by  $j(k, p)$  for every  $k \in \mathbb{N}$  and  $p \in (0, 1)$ , and let  $\mathcal{J} := \{j_1, j_2, \dots\} \subseteq \mathbb{Z}$  denote their image. Given this setup and applying (6.76) from Step 3, we have that

$$\begin{aligned} & \mathbb{P}(\exists t \in \mathbb{N}, k \in \mathbb{N}, p \in (0, 1) : S_t(q(k, j(k, p))) \geq b_{t,k}(q(k, j(k, p)))) \\ &= \mathbb{P}(\exists t \in \mathbb{N}, k \in \mathbb{N}, j \in \mathcal{J} : S_t(q(k, j)) \geq b_{t,k}(q(k, j))) \\ &\leq \mathbb{P}(\exists t \in \mathbb{N}, k \in \mathbb{N}, j \in \mathbb{Z} : S_t(q(k, j)) \geq b_{t,k}(q(k, j))) \\ &\leq \alpha, \end{aligned}$$

where the second line follows from the definition of  $\mathcal{J}$ , the third follows from the fact that  $\mathcal{J} \subseteq \mathbb{Z}$ , and the last follows from (6.76). In summary, we have the time- and  $p$ -uniform concentration inequality given by

$$\mathbb{P}(\exists t \in \mathbb{N}, k \in \mathbb{N}, p \in (0, 1) : S_t(q(k, j(k, p))) \geq b_{t,k}(q(k, j(k, p)))) \leq \alpha, \text{ or equivalently,} \quad (6.78)$$

$$\mathbb{P}(\forall t \in \mathbb{N}, k \in \mathbb{N}, p \in (0, 1), S_t(q(k, j(k, p))) < b_{t,k}(q(k, j(k, p)))) \geq 1 - \alpha \quad (6.79)$$

**Step 5: Obtaining a time- and  $p$ -uniform upper bound on  $S_t(p)$ .** While (6.79) is now written to be  $p$ -uniform, the quantity  $b_{t,k}(q(k, j(k, p)))$  is only a high-probability upper bound on  $S_t(q(k, j(k, p)))$ , but what we need is a high-probability upper bound on  $S_t(p)$ . To this end, we use a similar technique to Howard and Ramdas [123] to bound the distance between  $S_t(q(k, j(k, p)))$  and  $S_t(p)$  for any  $p \in (0, 1)$ . Indeed, consider the following representation of  $S_t(p)$  in terms of  $S_t(q(k, j(k, p)))$ :

$$\begin{aligned} S_t(p) &:= \sum_{i=1}^t w_i \mathbb{1}(R_i \leq Q(p)) - tp \\ &\leq \sum_{i=1}^t w_i \mathbb{1}(R_i \leq Q(q(k, j(k, p)))) - tp \end{aligned}$$

$$= S_t(q(k, j(k, p))) + t(q(k, j(k, p)) - p),$$

where the first line follows by definition of  $S_t(p)$ , the second by monotonicity of  $Q \mapsto \mathbb{1}(R_t \leq Q)$  and the fact that  $p \leq p_{k,j(k,p)}$ , and the third follows from the definition of  $S_t(q(k, j(k, p)))$ . Combining (6.79) with the above representation of  $S_t(p)$ , we have that

$$\mathbb{P} \left( \forall t \in \mathbb{N}, k \in \mathbb{N}, p \in (0, 1), S_t(p) < \underbrace{b_{t,k}(q(k, j(k, p)))}_{\text{(i)}} + \underbrace{t(q(k, j(k, p)) - p)}_{\text{(ii)}} \right) \geq 1 - \alpha, \quad (6.80)$$

where (i)  $\equiv b_{t,k}(q(k, j(k, p)))$  is given by

$$b_{t,k}(q(k, j(k, p))) := \frac{W_t \psi_{G,q(k,j(k,p))}(\lambda_{k,j(k,p)}) + \log(1/\alpha_{k,j(k,p)})}{\lambda_{k,j(k,p)}}. \quad (6.81)$$

**Step 5(i): Upper-bounding (i) without dependence on  $k$ .** Applying Lemma 6.A.2 but with  $j(k, p)$  in place of  $j$ , we have that for every  $\eta^k \leq W_t \leq \eta^{k+1}$ ,

$$\begin{aligned} b_{t,k}(q(k, j(k, p))) &\leq \sqrt{\gamma_1^2 \log(1/\alpha_{k,j(k,p)}) W_t + \gamma_2^2 q(k, j(k, p))^2 \log^2(1/\alpha_{k,j(k,p)})} \\ &\quad + \gamma_2 q(k, j(k, p)) \log(1/\alpha_{k,j(k,p)}). \end{aligned}$$

Now notice that the above upper bound depends on  $k$  solely through  $q(k, j(k, p))$  and  $\log(1/\alpha_{k,j(k,p)})$ , each of which we will upper-bound independently of  $k$ . By Lemma 6.A.3, we have that

$$q(k, j(k, p)) \leq \bar{q}_t(p) \equiv \text{logit}^{-1} \left( \text{logit}(p) + 2\delta \sqrt{\frac{\eta}{W_t}} \right) \quad \text{for all } \eta^k \leq W_t \leq \eta^{k+1}, \quad (6.82)$$

so it remains to upper-bound  $\log(1/\alpha_{k,j(k,p)})$ . Recall the definition of  $\alpha_{k,j}$  for any  $k \in \mathbb{N}, j \in \mathbb{Z}$  given in (6.68b). Then we can write  $\log(1/\alpha_{k,j(k,p)})$  as

$$\log(1/\alpha_{k,j(k,p)}) = \underbrace{s \log(k+1)}_{(\star k)} + \underbrace{2 \log(|j(k, p)| \vee 1)}_{(\star j)} + \log \zeta(s) + \log(2\zeta(s)+1) + \log(1/\alpha), \quad (6.83)$$

and we observe that  $(\star k)$  and  $(\star j)$  are the only terms depending on  $k$ . Firstly, notice that  $(\star j)$  can be upper bounded for every  $\eta^k \leq W_t \leq \eta^{k+1}$  as

$$\begin{aligned} (\star j) &\equiv 2 \log(|j(k, p)| \vee 1) = 2 \log \left( \left\| \frac{\eta^{k/2} \text{logit}(p)}{2\delta} \right\| \vee 1 \right) \\ &\leq 2 \log \left( \left\| \frac{\sqrt{W_t} \text{logit}(p)}{2\delta} \right\| \vee 1 \right). \end{aligned}$$

Second, notice that we can easily upper-bound  $(\star k)$  on epoch  $\eta^k \leq W_t \leq \eta^{k+1}$  as

$$s \log(k+1) \leq s \log(\log_\eta W_t + 1).$$

Therefore, we have the following upper-bound on  $\log(1/\alpha_{k,j(k,p)})$ :

$$\begin{aligned} \log(1/\alpha_{k,j}) &\leq s \log(\log_\eta W_t + 1) + 2 \log\left(\left|\left|\frac{\sqrt{W_t} \text{logit}(p)}{2\delta}\right|\right| \vee 1\right) + \log\zeta(s) + \log(2\zeta(s) + 1) + \log(1/\alpha), \\ &\equiv \ell_t(p), \end{aligned}$$

which no longer depends on  $k$ . In summary, we have that

$$b_{t,k}(q(k, j(k, p))) \leq \sqrt{\gamma_1^2 \ell_t(p) W_t + \gamma_2^2 \bar{q}_t(p)^2 \ell_t(p)^2} + \gamma_2 \bar{q}_t(p) \ell_t(p).$$

**Step 5(ii): Upper-bounding (ii) without dependence on  $k$ .** By Lemma 6.A.3, we have that  $q(k, j(k, p)) \leq \bar{q}_t(p) \equiv \text{logit}^{-1}\left(\text{logit}(p) + 2\delta\sqrt{\frac{\eta}{W_t}}\right)$ . Therefore, we can upper bound (ii) as

$$(ii) \equiv t(q(k, j(k, p)) - p) \leq t(\bar{q}_t(p) - p) \equiv t\left[\text{logit}^{-1}\left(\text{logit}(p) + 2\delta\sqrt{\frac{\eta}{W_t}}\right) - p\right], \quad (6.84)$$

where the final inequality no longer depends on  $k$ . In sum, with probability at least  $1 - \alpha$ ,

$$\forall t \in \mathbb{N}, p \in (0, 1), S_t(p) < \sqrt{\gamma_1^2 \ell_t(p) W_t + \gamma_2^2 \bar{q}_t(p)^2 \ell_t(p)^2} + \gamma_2 \bar{q}_t(p) \ell_t(p) + t(\bar{q}_t(p) - p). \quad (6.85)$$

**Lemma 6.A.2.** For any  $k \in \mathbb{N}$  and any  $j \in \mathbb{Z}$ , we have that for all  $\eta^k \leq W_t \leq \eta^{k+1}$ ,

$$b_{t,k}(q(k, j)) \leq \sqrt{\gamma_1^2 \log(1/\alpha_{k,j}) W_t + \gamma_2^2 q(k, j)^2 \log^2(1/\alpha_{k,j})} + q(k, j) \gamma_2 \log(1/\alpha_{k,j}), \quad (6.86)$$

where  $\gamma_1, \gamma_2$  are constants defined as

$$\gamma_1 := \frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}} \text{ and } \gamma_2 := \frac{\sqrt{\eta} + 1}{2}. \quad (6.87)$$

*Proof.* Recall the chosen value of  $\lambda_{k,j}$  given in (6.68e),

$$\lambda(k, j) := \psi_{G,q(k,j)}^{-1}(\log(1/\alpha_{k,j})/\eta^{k+1/2}), \text{ where } \psi_{G,c}^{-1}(a) := \frac{2}{c + \sqrt{c^2 + 2/a}} \quad (6.88)$$

Similar to Howard et al. [125, Eq. (39)], some algebra will reveal that for any  $t, k \in \mathbb{N}, j \in \mathbb{Z}$ ,

we have that

$$b_{t,k}(q(k, j)) = A_{q(k,j)} \left( \frac{\log(1/\alpha_{k,j})}{\eta^{k+1/2}} \right) \underbrace{\left[ \sqrt{\frac{\eta^{k+1/2}}{W_t}} + \sqrt{\frac{W_t}{\eta^{k+1/2}}} \right]}_{(\star)} \sqrt{\frac{\log(1/\alpha_{k,j}) W_t}{2}},$$

where  $A_c(a) := \sqrt{2a}/\psi_{G,c}^{-1}(a) = \sqrt{1+c^2a/2} + c\sqrt{a/2}$ . Now, notice that the second derivative of  $(\star)$  with respect to  $W_t$  is positive on  $W_t \in [\eta^k, \eta^{k+1}]$ , and hence  $(\star)$  is convex in  $W_t$ . As such, for every  $W_t \in [\eta^k, \eta^{k+1}]$  — i.e. the  $k^{\text{th}}$  epoch — we have that  $(\star)$  is maximized at the endpoints  $W_t = \eta^k$  and  $W_t = \eta^{k+1}$ , and we thus have the following upper bound on  $b_{t,k}(q(k, j))$  on the  $k^{\text{th}}$  epoch:

$$b_{t,k}(q(k, j)) \leq A_{q(k,j)} \left( \frac{\log(1/\alpha_{k,j})}{\eta^{k+1/2}} \right) \left[ \eta^{1/4} + \eta^{-1/4} \right] \sqrt{\frac{\log(1/\alpha_{k,j}) W_t}{2}}. \quad (6.89)$$

Furthermore, since  $W_t/\sqrt{\eta} \leq \eta^{k+1/2}$  on the  $k^{\text{th}}$  epoch, we also have that

$$A_{q(k,j)} \left( \frac{\log(1/\alpha_{k,j})}{\eta^{k+1/2}} \right) \leq A_{q(k,j)} \left( \frac{\sqrt{\eta} \log(1/\alpha_{k,j})}{W_t} \right) \quad \text{for all } \eta^k \leq W_t \leq \eta^{k+1}. \quad (6.90)$$

Putting (6.89) and (6.90) together, we have that for all  $\eta^k \leq W_t \leq \eta^{k+1}$ ,

$$\begin{aligned} b_{t,k}(q(k, j)) &\leq \frac{\eta^{1/4} + \eta^{-1/4}}{\sqrt{2}} \left( \sqrt{\log(1/\alpha_{k,j}) W_t + \frac{\sqrt{\eta} q(k, j)^2 \log^2(1/\alpha_{k,j})}{2}} + q(k, j) \frac{\eta^{1/4} \log(1/\alpha_{k,j})}{\sqrt{2}} \right) \\ &= \sqrt{\gamma_1^2 \log(1/\alpha_{k,j}) W_t + \gamma_2^2 q(k, j)^2 \log^2(1/\alpha_{k,j})} + q(k, j) \gamma_2 \log(1/\alpha_{k,j}), \end{aligned} \quad (6.91)$$

where  $\gamma_1, \gamma_2$  are constants defined in (6.87). This completes the proof of Lemma 6.A.2.  $\square$

**Lemma 6.A.3.** Define  $\bar{q}_t(p)$  as

$$\bar{q}_t(p) := \text{logit}^{-1} \left( \text{logit}(p) + 2\delta \sqrt{\frac{\eta}{W_t}} \right). \quad (6.92)$$

For all  $\eta^k \leq W_t \leq \eta^{k+1}$ , we have that  $q(k, j(k, p)) \leq \bar{q}_t(p)$ .

*Proof.* The result follows by definition of  $q(k, j(k, p))$ . Indeed, we have that for all  $\eta^k \leq W_t \leq \eta^{k+1}$ ,

$$q(k, j(k, p)) := \frac{1}{1 + \exp \left\{ -2j(k, p)\delta/\eta^{k/2} \right\}}$$

$$\begin{aligned}
&= \left( 1 + \exp \left\{ -2 \left[ \frac{\eta^{k/2} \text{logit}(p)}{2\delta} \right] \delta / \eta^{k/2} \right\} \right)^{-1} \\
&\leq \left( 1 + \exp \left\{ -2 \left( \frac{\eta^{k/2} \text{logit}(p)}{2\delta} + 1 \right) \delta / \eta^{k/2} \right\} \right)^{-1} \\
&= \left( 1 + \exp \left\{ -(\text{logit}(p) - 2\delta/\eta^{k/2}) \right\} \right)^{-1} \\
&= \text{logit}^{-1}(\text{logit}(p) + 2\delta/\eta^{k/2}) \\
&\leq \text{logit}^{-1} \left( \text{logit}(p) + 2\delta \sqrt{\frac{\eta}{W_t}} \right),
\end{aligned}$$

which completes the proof.

## 6.B A causal view of contextual bandits via potential outcomes

In Section 6.1, we discussed how the OPE problem can be interpreted as asking a *counterfactual* question, such as “how would the rewards have been, had we used a different policy  $\pi$  than the logging policy  $h$  that collected the data?”. While it is somewhat reasonable to think about the functional  $\nu_t = \mathbb{E}_\pi(R_t \mid X_1^{t-1})$  in a counterfactual sense, the Neyman-Rubin potential outcomes framework was designed for the rigorous study of precisely these types of causal questions [193, 228]. In this section, we will define a target *causal* functional  $\nu_t^*$  in terms of potential outcomes, and outline the identification assumptions under which  $\nu_t^*$  is equal to  $\nu_t$  (and hence, the conditions under which our CSs can be interpreted as covering the causal quantity  $\tilde{\nu}_t^* := \frac{1}{t} \sum_{i=1}^t \nu_i^*$ ). We emphasize that these identification assumptions are not required for the CSs to be useful or sensible — indeed,  $\tilde{\nu}_t$  is still an interpretable statistical quantity that we may wish to estimate — but they cannot otherwise be said to cover a causal functional defined in terms of potential outcomes.

Making our setup precise, we posit that for each time  $t$ , there is one potential outcome  $R_t(a)$  for every action  $a \in \mathcal{A}$ . The functional we are ultimately interested in estimating is the *conditional mean potential outcome reward under the policy  $\pi$* , i.e.

$$\nu_t^\star \equiv \mathbb{E}_\pi(R_t(G) \mid X_1^{t-1}) := \mathbb{E} \left\{ \mathbb{E}_{G \sim \pi(\cdot \mid X_t)}(R_t(G) \mid X_t, X_1^{t-1}) \mid X_1^{t-1} \right\} \quad (6.93)$$

$$= \int_{\mathcal{A} \times \mathcal{X}} \mathbb{E}(R_t(g) \mid G = g, X_t = x, X_1^{t-1}) \pi(g \mid x) p_{X_t}(x \mid X_1^{t-1}) dg dx. \quad (6.94)$$

In words,  $\nu_t$  is the average of the potential outcomes  $\{R_t(g)\}_{g \in \mathcal{A}}$  conditional on  $X_1^{t-1}$  with respect to the distribution  $\pi(\cdot | X_t)$ . We use  $g$  and  $G$  in place of  $a$  and  $A$  to avoid confusion between the actual (random) action  $A_t$  played according to the logging policy  $h_t(\cdot | X_t)$  and the hypothetical (random) action  $G$ . Without further assumptions, however, the counterfactual quantity  $\nu_t^*$  is not necessarily identified, meaning it cannot necessarily be written as a functional of the distribution of the observed data  $(X_t, A_t, R_t)_{t=1}^\infty$ . This is simply due to the fact that

the potential outcome  $R_t(g)$  is not directly observable from  $(X_t, A_t, R_t)_{t=1}^\infty$ . To remedy this, consider the following causal identification assumptions for every subject  $t$ ,

(IA1): **Consistency:**  $A_t = a \implies R_t(a) = R_t$  for every  $a \in \mathcal{A}$  with positive  $\pi$ -density,

(IA2): **Sequential exchangeability:**  $A_t \perp\!\!\!\perp R_t(a) \mid X_1^t$ , and

(IA3): **Positivity:**  $\pi \ll h_t$ , meaning  $h_t(A_t \mid X_t) = 0 \implies \pi(A_t \mid X_t) = 0$  almost surely.

Notice that in the contextual bandit setup, IA2 and IA3 are known to hold *by design*, while IA1 is more subtle (e.g. IA1 may not hold even in a randomized experiment due to interference between subjects, such as in a vaccine trial). Nevertheless, with IA1, IA2, and IA3 in mind, we are ready to state the main identification result of this section.

**Lemma 6.B.1.** *Under causal assumptions IA1, IA2, and IA3, we have that*

$$\nu_t^* = \nu_t, \text{ and hence } \tilde{\nu}_t^* := \frac{1}{t} \sum_{i=1}^t \nu_i^* = \frac{1}{t} \sum_{i=1}^t \nu_i =: \tilde{\nu}_t. \quad (6.95)$$

In other words, the counterfactual conditional mean  $\nu_t^*$  can be represented as a function of the distribution of observed data  $(X_t, A_t, R_t)_{t=1}^\infty$ , and that representation is given by  $\nu_t$ .

*Proof.* The proof is an exercise in causal identification and essentially follows that of Kennedy [154, Theorem 1] and Robins'  $g$ -formula [223], but we nevertheless provide a derivation here for completeness.

Writing out the definition of  $\nu_t^*$ , we have

$$\nu_t^* := \int_{\mathcal{A} \times \mathcal{X}} \mathbb{E}(R_t(g) \mid G = g, X_t = x, X_1^{t-1}) \pi(g \mid x_t) p_{X_t}(x \mid X_1^{t-1}) dg dx \quad (6.96)$$

$$= \int_{\mathcal{A} \times \mathcal{X}} \mathbb{E}(R_t(g) \mid G = g, A_t = g, X_t = x, X_1^{t-1}) \pi(g \mid x) p_{X_t}(x \mid X_1^{t-1}) dg dx \quad (6.97)$$

$$= \int_{\mathcal{A} \times \mathcal{X}} \mathbb{E}(R_t(g) \mid A_t = g, X_t = x, X_1^{t-1}) \pi(g \mid x) p_{X_t}(x \mid X_1^{t-1}) dg dx \quad (6.98)$$

$$= \int_{\mathcal{A} \times \mathcal{X}} \mathbb{E}(R_t \mid A_t = g, X_t = x, X_1^{t-1}) \pi(g \mid x) p_{X_t}(x \mid X_1^{t-1}) dg dx \quad (6.99)$$

$$= \mathbb{E}\{\mathbb{E}_{A_t \sim \pi(\cdot \mid X_t)}(R_t \mid X_t, X_1^{t-1}) \mid X_1^{t-1}\} \quad (6.100)$$

$$\equiv \mathbb{E}_\pi(R_t \mid X_1^{t-1}) =: \nu_t, \quad (6.101)$$

where (6.97) follows from IA2 (sequential exchangeability), (6.98) follows from the fact that  $R_t(G) \perp\!\!\!\perp G \mid X_1^t$  (by definition), and (6.99) follows from IA1 (consistency). Throughout, we implicitly used IA3 (positivity) so that the outer integral is well-defined. That is, we conditioned on  $A_t = g$  at several points, which implicitly leaves us with a factor of  $\pi(g \mid x)/h_t(g \mid x)$  — positivity ensures that this quantity is well-defined with probability one. This completes the proof of Lemma 6.B.1.

□

*Remark 16* (On the relationship between OPE and stochastic intervention effect estimation). It is no surprise that the proof of Lemma 6.B.1 follows Robins [223] and Kennedy [154] who study *stochastic intervention effects* in causal inference. Indeed, OPE and estimation of stochastic interventions are two different framings of essentially the same problem, and use the same importance-weighted and doubly robust estimators. The main differences between these fields lie in their emphases: the former is focused on adaptive experiments where logging policies are data-adaptive and *known*, whereas the latter typically places more emphasis on potential outcomes and causal identification, observational studies (i.e. where logging policies must be *estimated*), and more complex causal functionals, such as those of Haneuse and Rotnitzky [118] and Kennedy [154]. Of course, these are incomplete characterizations made with broad strokes; for a more detailed summary of prior work in stochastic interventions, see Kennedy [154, Section 1].

*Remark 17* (Implications for design-based confidence sequences). As an alternative to estimating treatment effects in superpopulations, one can opt to consider a so-called “design-based” approach to causal inference where the potential outcomes of all individuals are conditioned on, and confidence intervals are constructed for the *sample* average treatment effect (SATE) given by  $\text{SATE}_t := \frac{1}{t} \sum_{i=1}^t (R_i(1) - R_i(0))$  where  $R_i(a)$  is subject  $i$ ’s potential outcome under treatment  $a \in \{0, 1\}$ . Here, the resulting confidence intervals cover the SATE with high probability, where the probability is taken with respect to the randomness in the treatment assignment mechanism only. The design-based approach goes back to Fisher and has a deep and extensive literature [105, 193, 131], and more recent work has constructed nonasymptotic CSs for the time-varying effect  $(\text{SATE}_t)_{t=1}^\infty$  in Howard et al. [125, Section 4.2] and asymptotic ones in Ham et al. [117]. For a more comprehensive literature review, we direct readers to Abadie et al. [1] and Ham et al. [117] as well as the references therein.

We simply remark here that the results of Section 6.3 simultaneously apply to the design-based *and* superpopulation settings as immediate corollaries. Indeed, in the stochastic (non-design-based) setting for binary experiments and under the causal identification assumptions IA1–IA3, we have that Lemma 6.B.1 yields

$$\Delta_t^* := \frac{1}{t} \sum_{i=1}^t \mathbb{E}[R_i(1) - R_i(0)] = \frac{1}{t} \sum_{i=1}^t [\mathbb{E}(R_i | A_i = 1) - \mathbb{E}(R_i | A_i = 0)] =: \Delta_t. \quad (6.102)$$

Now, to recover CSs for  $(\text{SATE}_t)_{t=1}^\infty$  we simply condition on  $(R_t(1), R_t(0), X_t)_{t=1}^\infty$  so that  $(A_t)_{t=1}^\infty$  are the only non-degenerate random variables here. The techniques for time-varying treatment effects described in Section 6.3.1 and Section 6.3.2, yield a  $(1 - \alpha)$ -CS  $[L_t, U_t]_{t=1}^\infty$  for  $(\Delta_t)_{t=1}^\infty \equiv (\Delta_t^*)_{t=1}^\infty$  and hence for  $(\text{SATE}_t)_{t=1}^\infty$ . Going further, when instantiated for the design-based setting, our CSs substantially improve on Howard et al. [125, Section 4.2], both practically and theoretically. Indeed, as discussed in Ham et al. [117, Section 3.2], one of the drawbacks of existing nonasymptotic CSs in the literature is that the minimal propensity score – i.e.  $p_{\min} := \text{ess inf}_{t,a,x} \mathbb{P}(A_t = a | x | \mathcal{H}_{t-1})$  – must be specified in advance, and

the downstream CSs always scale with  $p_{\min}^{-1}$ . However, as we have emphasized throughout this chapter, beginning with desideratum 5 in Section 6.1.2, *none* of our CSs suffer from this limitation.

Simultaneously, if we consider the superpopulation setting where  $\mathbb{E}(R_t(1)) - \mathbb{E}(R_t(0)) = \delta$  for all  $t \geq 1$  and for some  $\delta \in [-1, 1]$ , then under identification assumptions IA1–IA3, the same CS  $[L_t, U_t]_{t=1}^{\infty}$  also covers  $\delta$  by Lemma 6.B.1. In this way, our time-varying CSs simultaneously handle the stationary superpopulation setting where treatment effects do not change over time, as well as the design-based setting where all potential outcomes are conditioned on, since these are both special cases of the time-varying stochastic setting considered in Section 6.3.

## **Part II**

# **Asymptotic inference, strong laws, and Gaussian approximation**

# Chapter 7

## Time-uniform central limit theory and asymptotic confidence sequences

### 7.1 Introduction

The central limit theorem (CLT) is arguably the most widely used result in applied statistical inference, due to its ability to provide large-sample confidence intervals (CI) and  $p$ -values in a broad range of problems under weak assumptions. Examples include (a) nonparametric estimation of means, such as population proportions, (b) maximum likelihood and other M-estimation problems [260], and (c) modern semiparametric causal inference methodology involving (augmented) inverse propensity score weighting [225, 258, 153, 56]. Crucially, in some of these problems such as doubly robust estimation in observational studies, nonasymptotic inference is typically not possible, and hence the CLT yields asymptotic CIs for an otherwise unsolvable inference problem.

While the CLT makes efficient statistical inference possible in a broad array of problems, the resulting CIs are only valid at a prespecified sample size  $n$ , invalidating any inference that occurs at data-dependent stopping times, for example under continuous monitoring. CIs that retain validity in sequential environments are known as *confidence sequences* (CS) [76, 217] and can be used to make decisions at arbitrary stopping times (e.g. while adaptively sampling, continuously peeking at the data, etc.). CSs are an inherently nonasymptotic notion, and thus essentially every published CS is nonasymptotic, including various recent state-of-the-art constructions in different settings [125, 123, 282, 273].

This chapter presents a new notion: an “asymptotic confidence sequence”. For the familiar reader, this might at first sound like an oxymoron. Further, it is not obvious how to posit a definition that is simultaneously sensible and tractable, meaning whether it is possible to develop such asymptotic CSs (whatever it may mean). We believe that we have formulated the

“right” definition, because we accompany it with a universality result that parallels the CLT — a universal asymptotic CS that is valid under the exact same moment assumptions required by the CLT, and exploits certain time-uniform central limit theory to arrive at boundaries that one would use if the data were Gaussian. This enables the construction of asymptotic CSs in a myriad of new situations where the distributional assumptions are weak enough to remain out of the reach of nonasymptotic techniques even in fixed-time settings. The width of this universal asymptotic CS scales with the variance of the data, just like the empirical variance used in the CLT — such variance-adaptivity is only achievable for nonasymptotic methods in very specialized settings (e.g. Chapter 3).

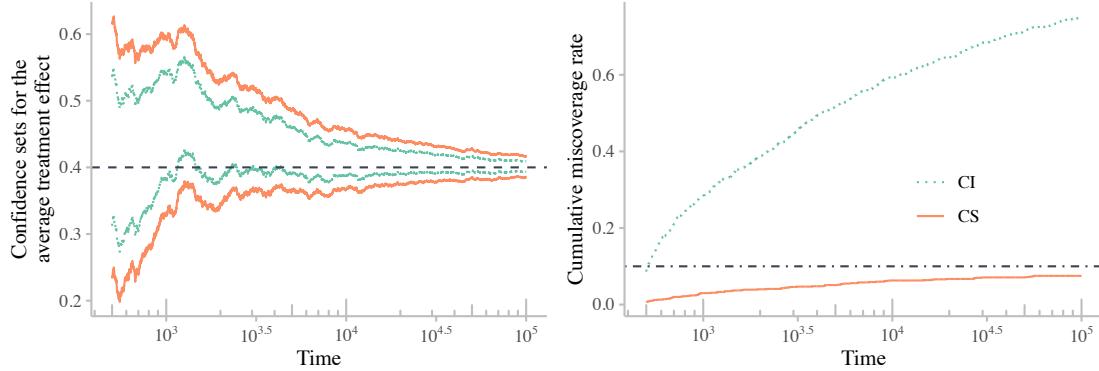


Figure 7.1: The left plot shows one run of a single experiment: an asymptotic CS alongside an asymptotic CI for a parameter of interest (in this case, the average treatment effect (ATE) of 0.4, an example we expand on in Section 7.3). The true value of the ATE is covered by the CS simultaneously from time 30 to 10000. On the other hand, the CI fails to cover the true ATE at several points in time. By repeating such an experiment hundreds of times, one obtains the right plot which displays the cumulative probability of miscoverage — i.e. the probability of the CS or CI failing to capture the true ATE at any time up to  $t$ . Notice that the CI error rate begins at  $\alpha = 0.1$  and quickly grows, while the CS error rate never exceeds  $\alpha = 0.1$ .

As mentioned in Chapter 1, while nonparametric CSs have been developed for several problems, they have thus far been *nonasymptotic*. Nonasymptotic inference for means of random variables *requires* strong assumptions on the distribution of the data [14]. These assumptions often take the form of a parametric likelihood [280, 123], known bounds on the random variables themselves [125, 282], on their moments [273], or on their moment generating functions [125].

These added distributional assumptions make existing CSs quite unlike CLT-based CIs which (a) are universal, meaning they take the same form — up to a change in influence functions — and are computed in the same way for most problems, and (b) are often applicable even when no nonasymptotic CI is known, such as in doubly robust inference of causal effects in observational studies. Our work bridges this gap, bringing properties (a) and (b) to the anytime-valid sequential regime by making one simple modification to the usual CIs. Just as CLT-based CIs yield approximate inference for a wide variety of problems in fixed- $n$  settings,

this chapter yields the same for *sequential* settings.

### 7.1.1 Contributions and outline

We begin by rigorously defining “asymptotic confidence sequences” (AsympCSs) in Definition 7.2.1 and providing a general recipe to derive explicit AsympCSs that are as easy to implement and apply as CLT-based CIs in Section 7.2.3. Using this recipe, we develop a Lindeberg-type AsympCS that is able to capture time-varying means under martingale dependence (Section 7.2.4). Furthermore, in Section 7.2.5, we give a definition of asymptotic time-uniform coverage (akin to coverage of asymptotic CIs) and show how *sequences* of our AsympCSs enjoy this property. In Section 7.3 we illustrate how the AsympCSs of Section 7.2.1 enable asymptotically anytime-valid semiparametric inference for causal effects in both randomized experiments and observational studies (Section 7.3). To be clear, we are not focused on deriving new semiparametric estimators; we simply demonstrate how semiparametric causal inference — a problem for which no known CSs exist in the observational setting — can now be tackled in fully sequential environments using the existing state-of-the-art estimators combined with our AsympCSs (Theorems 7.3.1 and 7.3.2). In Section 7.4, we provide a simulation study to illustrate empirical widths and miscoverage rates of AsympCSs and compare them to some existing (nonasymptotic) CSs in the literature. Finally, we apply the AsympCSs of Section 7.3 to a real observational data set by sequentially estimating the effects of fluid intake on 30-day mortality in sepsis patients. In sum, this work expands the scope of anytime-valid inference by tackling sequential estimation problems under CLT-like moment assumptions and guarantees.

## 7.2 Asymptotic confidence sequences

We first define what it means for a sequence of intervals to form an asymptotic confidence sequence (AsympCS). Then, we derive a “universal” AsympCS in the sense that the AsympCS does not depend on any features of the distribution beyond its mean and variance.<sup>1</sup> Much like classical asymptotic confidence intervals based on the CLT, this universal AsympCS fundamentally relies on Gaussian approximation. However, in this setting, the particular type of central limit theory being invoked is that of strong invariance principles, where an implicit Gaussian process is coupled with a partial sum with probability one (more details are provided in Section 7.2.2). Finally, similar to CIs based on martingale CLTs, we derive a Lindeberg-type martingale AsympCS that can track a moving average of conditional means.

### 7.2.1 Defining asymptotic confidence sequences

Here, we define and present “asymptotic confidence sequences” as time-uniform analogues of CLT-based asymptotic CIs, making similarly weak moment assumptions and providing a universal closed-form boundary.

---

<sup>1</sup>We use “universal” in the same way that the CLT and law of large numbers are considered universal [250], as they describe macroscopic behaviors that are independent of most microscopic details of the system.

The term “asymptotic confidence sequence” may at first seem paradoxical. Indeed, ever since their introduction by Robbins and collaborators [76, 167, 168], CSs have been defined nonasymptotically. So how could a bound be both time-uniform and asymptotically valid? We clarify this critical point soon, with an analogy to classical asymptotic CIs. Similar to asymptotic CIs, AsympCSs trade nonasymptotic guarantees for (a) simplicity and universality, and (b) the ability to tackle a much wider variety of problems, especially those for which there is no known nonasymptotic CS. Said differently, AsympCSs trade finite sample validity for versatility (exemplified in Section 7.3 with a particular emphasis on modern causal inference).

Indeed, there is a clear desire for (asymptotically) time-uniform methods with CLT-like simplicity and versatility, especially in the context of causal inference. For example, Johari et al. [135, Section 4.3] use a Gaussian mixture sequential probability ratio test (SPRT) to conduct A/B tests (i.e. randomized experiments) for data coming from (non-Gaussian) exponential families and mentions that CLT approximations hold at large sample sizes. Similarly, Yu et al. [291] develop a mixture SPRT for causal effects in generalized linear models, where they say that their likelihood ratio forms an “approximate martingale”, meaning its conditional expectation is constant up to a factor of  $\exp\{o_{\mathbb{P}}(1)\}$ . Moreover, Pace and Salvan [201] suggest using Robbins’ Gaussian mixture CS as a closed-form “approximate CS” and they demonstrate through simulations that the time-uniform coverage guarantee tends to hold in the asymptotic regime. However, all of these approaches justify time-uniform inference with  $o_{\mathbb{P}}(\cdot)$  approximations that only hold at a *fixed, pre-specified sample size*, and yet inferences are being carried out at *data-dependent sample sizes*. This section remedies the tension between fixed- $n$  approximations and time-uniform inference by defining AsympCSs such that Gaussian approximations must hold almost surely for *all sample sizes simultaneously*. The AsympCSs we define will also be valid in a wide range of nonparametric scenarios (beyond exponential families, parametric models, and so on).

To motivate the definition of an AsympCS that follows, let us briefly review the CLT in the batch (non-sequential) setting. Suppose  $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} \mathbb{P}$  with mean  $\mathbb{E}(Y_1) = \mu$  and variance  $\text{Var}(Y_1) = \sigma^2$ . Then the standard CLT-based CI for  $\mu$  (with known variance  $\sigma$ ) takes the form

$$\dot{C}_n := [\hat{\mu}_n \pm \dot{\mathfrak{B}}_n] \equiv \left[ \hat{\mu}_n \pm \sigma \cdot \frac{\Phi^{-1}(1 - \alpha/2)}{\sqrt{n}} \right], \quad (7.1)$$

where  $\hat{\mu}_n$  is the sample mean and  $\Phi^{-1}(1 - \alpha/2)$  is the  $(1 - \alpha/2)$ -quantile of a standard Gaussian  $N(0, 1)$  (e.g. for  $\alpha = 0.05$ , we have  $\Phi^{-1}(0.975) \approx 1.96$ ). The classical notion of “asymptotic validity” is

$$\liminf_{n \rightarrow \infty} \mathbb{P}(\mu \in \dot{C}_n) \geq 1 - \alpha. \quad (7.2)$$

While the above is the standard definition of an asymptotic CI, one could have arrived at an alternative definition by noting the following rather strong statement that can be made under

the same conditions: there exist<sup>2</sup> i.i.d. standard Gaussians  $Z_1, \dots, Z_n \sim N(0, 1)$  such that

$$\frac{1}{n} \sum_{i=1}^n (Y_i - \mu)/\sigma = \frac{1}{n} \sum_{i=1}^n Z_i + o_{\mathbb{P}}(1/\sqrt{n}). \quad (7.3)$$

From this vantage point, we note that the CI in (7.1) has the additional guarantee that there in fact exists an (unknown) *nonasymptotic*  $(1 - \alpha)$ -CI  $[\hat{\mu}_n \pm \dot{\mathfrak{B}}_n^*]$  such that

$$\dot{\mathfrak{B}}_n^*/\dot{\mathfrak{B}}_n \xrightarrow{\mathbb{P}} 1. \quad (7.4)$$

We deliberately highlight the above property of asymptotic CIs because it ends up serving as a natural starting point for defining asymptotic confidence *sequences*. In particular, we will define AsympCSs so that an analogous approximation to (7.4) holds uniformly over time, almost surely. Statements like (7.3) are known as “couplings” and appear in the literature on strong approximations and invariance principles where similar guarantees can indeed be shown to hold almost surely and at faster rates under additional moment assumptions [101, 160, 161].

*Definition 7.2.1* (Asymptotic confidence sequences). Let  $\mathcal{T}$  be a totally ordered infinite set (denoting time) that has a minimum value  $t_0 \in \mathcal{T}$ . We say that the intervals  $(\hat{\theta}_t - L_t, \hat{\theta}_t + U_t)_{t \in \mathcal{T}}$  centered at the estimators  $(\hat{\theta}_t)_{t \in \mathcal{T}}$  with non-zero bounds  $L_t, U_t > 0, \forall t \in \mathcal{T}$  form a  $(1 - \alpha)$ -asymptotic confidence sequence (AsympCS) for a sequence of real parameters  $(\theta_t)_{t \in \mathcal{T}}$  if there exists a (typically unknown) nonasymptotic  $(1 - \alpha)$ -CS  $(\hat{\theta}_t - L_t^*, \hat{\theta}_t + U_t^*)_{t \in \mathcal{T}}$  for  $(\theta_t)_{t \in \mathcal{T}}$  – i.e. satisfying

$$\mathbb{P} \left( \forall t \in \mathcal{T}, \theta_t \in [\hat{\theta}_t - L_t^*, \hat{\theta}_t + U_t^*] \right) \geq 1 - \alpha, \quad (7.5)$$

and such that  $L_t, U_t$  become arbitrarily precise almost-sure approximations to  $L_t^*$  and  $U_t^*$ :

$$L_t^*/L_t \xrightarrow{\text{a.s.}} 1 \quad \text{and} \quad U_t^*/U_t \xrightarrow{\text{a.s.}} 1. \quad (7.6)$$

In words, Definition 7.2.1 says that an AsympCS  $(C_t)_{t \in \mathcal{T}}$  centered at  $(\hat{\theta}_t)_{t \in \mathcal{T}}$  is an arbitrarily precise approximation of some nonasymptotic CS  $(C_t^*)_{t \in \mathcal{T}}$  centered at  $(\hat{\theta}_t)_{t \in \mathcal{T}}$  as  $t \rightarrow \infty$ . Throughout this chapter, we will mostly focus on the case where  $\mathcal{T} = \mathbb{N}_0$  with  $t_0 = 0$ .

It is important to note that alternate definitions fail to be coherent in different ways. As one example, we could have hypothetically defined a sequence of intervals  $(C_t(\alpha))_{t \in \mathcal{T}}$  to be a  $(1 - \alpha)$ -AsympCS if  $\limsup_{m \rightarrow \infty} \mathbb{P}(\exists t \geq m : \mu \notin C_t(\alpha)) \leq \alpha$ , analogously to asymptotic CIs which satisfy  $\limsup_{n \rightarrow \infty} \mathbb{P}(\mu \notin C_n(\alpha)) \leq \alpha$ . In words, we could have posited that if we just start peeking late enough, then the probability of eventual miscoverage would indeed be below  $\alpha$ . Unfortunately, even for nonasymptotic CSs constructed at any level  $\alpha' \in (0, 1)$ , the former limit is *zero*, so this inequality would be vacuously true, regardless of  $\alpha'$ , even if  $\alpha' \gg \alpha$ . However, we do show that *sequences* of our AsympCSs have a guarantee of this type if peeking starts late enough (see Section 7.2.5), but we delay these definitions until later as they

---

<sup>2</sup>Technically, writing (7.3) may require enriching the probability space so that both  $Y$  and  $Z$  can be defined (but without changing their laws). See Einmahl [101, Equation (1.2)] for a precise statement.

are slightly more involved.

By virtue of being defined in terms of their limiting behavior, one can obviously construct AsympCSs (as well as asymptotic confidence intervals) with nonsensical finite-sample behavior. It is thus imperative that if a practitioner decides to employ an asymptotic method, they do so with the understanding that its effectiveness relies on it exploiting some well-approximated nonasymptotic phenomenon, and that its limiting behavior should be viewed as a rigorous manifestation of a guiding principle rather than a panacea.

*Remark 18* (Why almost surely?). One may wonder why it is necessary to define AsympCSs so that  $L_t^*/L_t \rightarrow 1$  *almost surely* (rather than in probability, for example). Since CSs are bounds that hold uniformly over time with high probability, convergence in probability  $L_t^*/L_t = 1 + o_{\mathbb{P}}(1)$  is not the right notion of convergence, as it only requires that the approximation term  $o_{\mathbb{P}}(1)$  be small with high probability for sufficiently large *fixed*  $t$ , but not for all  $t$  uniformly. It is natural to try to extend convergence in probability to *time-uniform convergence with high probability* – i.e.  $\sup_{k \geq t} (L_k^*/L_k) = 1 + o_{\mathbb{P}}(1)$  – but it turns out that this is in fact *equivalent* to almost-sure convergence  $L_t^*/L_t = 1 + o_{\text{a.s.}}(1)$ ; see Section 7.B.3.

Going forward, we may omit “a.s.” from  $o_{\text{a.s.}}(\cdot)$  and  $O_{\text{a.s.}}(\cdot)$  and instead simply write  $o(\cdot)$  and  $O(\cdot)$ , respectively to simplify notation. Now that we have defined AsympCSs as time-uniform analogues of asymptotic CIs, we will explicitly derive AsympCSs for the mean of i.i.d. random variables with finite variances (i.e. under the same assumptions as the CLT).

## 7.2.2 Warmup: AsympCSs for the mean of i.i.d. random variables

We now construct an explicit AsympCS for the mean of i.i.d. random variables by combining a variant of Robbins’ (nonasymptotic) Gaussian mixture boundary [217] with Strassen’s strong approximation theorem [248]. Before presenting the result, let us review Robbins’ boundary and Strassen’s result, and discuss how they can be used in conjunction to arrive at the AsympCS in Theorem 7.2.2.

### 7.2.2.1 Robbins’ Gaussian mixture boundary

The study of CSs began with Herbert Robbins and colleagues [76, 217, 220, 167, 168], leading to several fundamental results and techniques including the famous Gaussian mixture boundary for partial sums of i.i.d. Gaussian random variables [217] (see also Howard et al. [125, §3.2]) which we recall here. Suppose  $(Z_t)_{t=1}^\infty$  are i.i.d. standard Gaussian random variables. Then for any  $\rho > 0$ ,

$$\mathbb{P} \left( \exists t \geq 1 : \left| \frac{1}{t} \sum_{i=1}^t Z_i \right| \geq \sqrt{\frac{2(t\rho^2 + 1)}{t^2 \rho^2} \log \left( \frac{\sqrt{t\rho^2 + 1}}{\alpha} \right)} \right) \leq \alpha. \quad (7.7)$$

Notice that the above boundary scales as  $O(\sqrt{\log t/t})$  for any  $\rho > 0$ . In fact, Robbins [217, Eq. 11] noted that (7.7) holds not only for Gaussian random variables, but for those that are 1-sub-Gaussian, and hence pre-multiplying the boundary by  $\sigma$  yields a  $\sigma$ -sub-Gaussian time-

uniform concentration inequality, serving as a time-uniform analogue of Chernoff or Hoeffding inequalities. The connections between these fixed-time and time-uniform concentration inequalities are made explicit in Howard et al. [125]. Nevertheless, Equation (7.7) requires *a priori* knowledge of  $\sigma > 0$  unlike CLT-based CIs which we aim to emulate in the (asymptotically) time-uniform regime. The following strong Gaussian approximation due to Strassen [248] will serve as a technical tool allowing the nonasymptotic sub-Gaussian bound in (7.7) to be applied to partial sums of arbitrary random variables with finite variances.

### 7.2.2.2 Strassen's strong approximation

Strassen [248] initiated the study of “strong approximation” (also called strong invariance principles or strong embeddings) which blossomed into an active and impactful corner of probability theory research over the subsequent years, culminating in now-classical results such as the Komlós-Major-Tusnády embeddings [160, 161, 181] and other related works [249, 100, 191, 54].

In his 1964 paper, Strassen [248, §2] used the Skorokhod embedding [243] (see also [36, p. 513]) to obtain a strong invariance principle which connects asymptotic Gaussian behavior with the law of the iterated logarithm. Concretely, let  $(Y_t)_{t=1}^\infty$  be an infinite sequence of i.i.d. random variables from a distribution  $\mathbb{P}$  with mean  $\mu$  and variance  $\sigma^2$ . Then, on a potentially richer probability space,<sup>3</sup> there exist standard Gaussian random variables  $(Z_t)_{t=1}^\infty$  whose partial sums are almost-surely coupled with those of  $(Y_t)_{t=1}^\infty$  up to iterated logarithm rates, i.e.

$$\left| \sum_{i=1}^t (Y_i - \mu)/\sigma - \sum_{i=1}^t Z_i \right| = o\left(\sqrt{t \log \log t}\right) \quad \text{almost surely.} \quad (7.8)$$

Notice that the law of the iterated logarithm states that  $|\sum_{i=1}^t (Y_i - \mu)/\sigma| = O(\sqrt{t \log \log t})$  while (7.8) states that the same partial sum is almost-surely coupled with an implicit Gaussian process – i.e. replacing  $O(\cdot)$  with  $o(\cdot)$ . It may be convenient to divide by  $t$  and interpret (7.8) on the level of sample averages rather than partial sums, in which case the right-hand side becomes  $o(\sqrt{\log \log t/t})$ . Let us now describe how Strassen's strong approximation can be combined with Robbins' Gaussian mixture boundary to derive an AsympCS under finite moment assumptions akin to the CLT.

### 7.2.2.3 The Gaussian mixture asymptotic confidence sequence

Given the juxtaposition of (7.7) and (7.8), the high-level approach to the derivation of AsympCSs becomes clearer. Indeed, the essential idea behind Theorem 7.2.2 is as follows. By Strassen's strong approximation, we couple the partial sums  $S_t := \sum_{i=1}^t (Y_i - \mu)/\sigma$  with implicit partial sums  $G_t := \sum_{i=1}^t Z_i$  of Gaussians  $(Z_t)_{t=1}^\infty$ , and then use Robbins' mixture boundary to obtain

---

<sup>3</sup>A richer probability space may be needed to describe Gaussian random variables, if for example,  $(Y_t)_{t=1}^\infty$  are  $\{0, 1\}$ -valued on a probability space whose probability measure is dominated by the counting measure. This construction of a richer probability space imposes no additional assumptions on  $(Y_t)_{t=1}^\infty$ , and is only a technical device used to rigorously couple two sequences of random variables, and appears in essentially all papers on strong invariance principles, not just Strassen [248].

a time-uniform high-probability bound on the deviations of  $|G_t|$ , noting that the coupling rate  $o(\sqrt{t \log \log t})$  is asymptotically dominated by the concentration rate  $O(\sqrt{t \log t})$ , leading to asymptotic validity in the formal sense of Definition 7.2.1.

**Theorem 7.2.2** (Gaussian mixture asymptotic confidence sequence). *Suppose  $(Y_t)_{t=1}^{\infty} \stackrel{iid}{\sim} \mathbb{P}$  is an infinite sequence of i.i.d. observations from a distribution  $\mathbb{P}$  with mean  $\mu$  and finite variance. Let  $\hat{\mu}_t := \frac{1}{t} \sum_{i=1}^t Y_i$  be the sample mean, and  $\hat{\sigma}_t^2 := \frac{1}{t} \sum_{i=1}^t Y_i^2 - (\hat{\mu}_t)^2$  the sample variance based on the first  $t$  observations. Then, for any prespecified constant  $\rho > 0$ ,*

$$\bar{C}_t^{\mathcal{G}} \equiv (\hat{\mu}_t \pm \bar{\mathfrak{B}}_t^{\mathcal{G}}) := \left( \hat{\mu}_t \pm \hat{\sigma}_t \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log \left( \frac{t\rho^2 + 1}{\alpha^2} \right)} \right) \quad (7.9)$$

forms a  $(1 - \alpha)$ -AsympCS for  $\mu$ .

The proof of Theorem 7.2.2 is in Appendix 7.A.1. We can think of  $\rho > 0$  as a user-chosen tuning parameter which dictates the time at which (7.9) is tightest, and we discuss how to easily tune this value in Section 7.B.2. A one-sided analogue of (7.9) can be found in Section 7.B.1.

While (7.9) may look visually similar to Robbins' (sub)-Gaussian mixture CS [217] – written explicitly in Howard et al. [125, Eq. (14)] – it is worth pausing to reflect on how they are markedly different. Firstly, Robbins' CS is a nonasymptotic bound that is only valid for  $\sigma$ -sub-Gaussian random variables, meaning  $\mathbb{E} \exp\{\lambda(Y_1 - \mathbb{E}Y_1)\} \leq \exp\{\sigma^2 \lambda^2 / 2\}$  for some *a priori* known  $\sigma > 0$ , while Theorem 7.2.2 does not require the existence of a finite MGF at all (much less a known upper bound on it). Secondly, Robbins' CS uses this known (possibly conservative)  $\sigma$  in place of  $\hat{\sigma}_t$  in (7.9), and thus it cannot adapt to an unknown variance, while (7.9) always scales with  $\sqrt{\text{Var}(Y_1)}$ . In simpler terms, Theorem 7.2.2 is an asymptotically time-uniform analogue of the CLT in the same way that Robbins' CS is a time-uniform analogue of a sub-Gaussian concentration inequality (e.g. Hoeffding's or Chernoff's inequality [120, 124]).

It is important not to confuse Theorem 7.2.2 with a martingale CLT as the latter still gives fixed-time CIs in the spirit of the usual CLT but under different assumptions on the martingale difference sequence (however, we do present an analogue of Theorem 7.2.2 under martingale dependence in Proposition 7.2.2).

#### 7.2.2.4 An asymptotic confidence sequence with iterated logarithm rates

As a consequence of the law of the iterated logarithm, a confidence sequence for  $\mu$  centered at  $\hat{\mu}_t$  cannot have an asymptotic width smaller than  $O(\sqrt{\log \log t / t})$ . This is easy to see since

$$\limsup_{t \rightarrow \infty} \frac{\sqrt{t} |\hat{\mu}_t - \mu|}{\sigma \sqrt{2 \log \log t}} = 1.$$

This raises the question as to whether  $\bar{C}_t^{\mathcal{G}}$  can be improved so that the optimal asymptotic width of  $O(\sqrt{\log \log t / t})$  is achieved. Indeed, we can replace Robbins' Gaussian mixture boundary with Howard et al. [125, Eq. (2)] (or virtually any other Gaussian boundary for that matter) in

the proof of Theorem 7.2.2 to derive such an AsympCS, but as the authors discuss, mixture boundaries such as the one in Theorem 7.2.2 may be preferable in practice, because any bound that is tighter “later on” (asymptotically) must be looser “early on” (at practical sample sizes) due to the fact that all such bounds have a cumulative miscoverage probability  $\leq \alpha$ . This is formally a concern for nonasymptotic CSs, but only applies to AsympCSs insofar as they are asymptotic approximations of nonasymptotic bounds. Nevertheless, we present an AsympCS with an iterated logarithm rate here for completeness.

**Proposition 7.2.1** (Iterated logarithm asymptotic confidence sequences). *Under the same conditions as Theorem 7.2.2,*

$$\bar{C}_t^{\mathcal{L}} \equiv (\hat{\mu}_t \pm \bar{\mathfrak{B}}_t^{\mathcal{L}}) := \left( \hat{\mu}_t \pm \hat{\sigma}_t \cdot 1.7 \sqrt{\frac{\log \log(2t) + 0.72 \log(10.4/\alpha)}{t}} \right)$$

forms a  $(1 - \alpha)$ -AsympCS for  $\mu$ .

We omit the proof of Proposition 7.2.1 as it proceeds in a similar fashion to that of Theorem 7.2.2. In fact, both of these AsympCSs are simply instantiations of a more general recipe for deriving AsympCSs by combining strong approximations with time-uniform boundaries for the approximating process, an approach that we discuss in the following section.

### 7.2.3 A general recipe for deriving asymptotic confidence sequences

The proofs of both Theorem 7.2.2 and Proposition 7.2.1 follow the same general structure, combining strong approximations with time-uniform boundaries along with some other almost-sure asymptotic behavior. Abstracting away the details specific to these particular results, we provide the following four general conditions under which many AsympCSs can be derived, including those from the previous section but also Lyapunov- and Lindeberg-type AsympCS that we will state in Section 7.2.4).

In what follows, let  $\mathcal{T}$  be a totally ordered infinite set that includes a minimum value  $t_0 \in \mathcal{T}$  (for example, one may think about  $\mathcal{T}$  as  $\mathbb{R}^{>0}$  or  $\mathbb{N}_0$  with  $t_0 = 0$ ) and let  $(\hat{\theta}_t)_{t \in \mathcal{T}}$  be a sequence of estimators for the real-valued parameters  $(\theta_t)_{t \in \mathcal{T}}$ . Then, consider the following four conditions where we use the “Condition G-X” enumeration as a mnemonic for the X<sup>th</sup> condition in the section on General recipes for AsympCSs).

**Condition G-1** (Strong approximation). *On a potentially enriched probability space, there exists a process  $(Z_t)_{t \in \mathcal{T}}$  starting at  $Z_{t_0} \equiv 0$  that strongly approximates  $(\theta_t - \hat{\theta}_t)_{t \in \mathcal{T}}$  up to a rate of  $(r_t)_{t \in \mathcal{T}}$ , i.e.*

$$(\theta_t - \hat{\theta}_t) - Z_t = O(r_t) \quad \text{almost surely.} \tag{7.10}$$

**Condition G-2** (Boundary for the approximating process). *There exist  $\hat{L}_t > 0$  and  $\hat{U}_t > 0$  for each  $t \in \mathcal{T}$  so that  $[-\hat{L}_t, \hat{U}_t]_{t \in \mathcal{T}}$  forms a  $(1 - \alpha)$ -boundary for the process  $(Z_t)_{t \in \mathcal{T}}$  given in (7.10):*

$$\mathbb{P}\left(\forall t \in \mathcal{T}, Z_t \in [-\hat{L}_t, \hat{U}_t]\right) \geq 1 - \alpha. \tag{7.11}$$

**Condition G-3** (Strong approximation rate). *The approximation rate  $(r_t)_{t \in \mathcal{T}}$  in (7.10) is faster than both  $(\hat{L}_t)_{t \in \mathcal{T}}$  and  $(\hat{U}_t)_{t \in \mathcal{T}}$  in (7.11), i.e.*

$$r_t = o(\hat{L}_t \wedge \hat{U}_t) \quad \text{almost surely.} \quad (7.12)$$

**Condition G-4** (Almost-sure approximate boundary). *The  $(1 - \alpha)$ -boundary  $[-\hat{L}_t, \hat{U}_t]_{t \in \mathcal{T}}$  for  $(Z_t)_{t \in \mathcal{T}}$  is almost-surely approximated by the sequence  $[-L_t, U_t]_{t \in \mathcal{T}}$ , i.e.*

$$L_t/\hat{L}_t \xrightarrow{\text{a.s.}} 1 \quad \text{and} \quad U_t/\hat{U}_t \xrightarrow{\text{a.s.}} 1. \quad (7.13)$$

Deriving new AsympCSs then reduces to the conceptually simpler but nevertheless non-trivial task of satisfying the requisite conditions above. For example, in the previous section, we satisfied Condition G-1 via Strassen [248], Condition G-2 via Robbins [217], Condition G-3 via the combination of Strassen [248] and Robbins [217], and Condition G-4 via the strong law of large numbers (SLLN). The only difference between Theorem 7.2.2 and Proposition 7.2.1 was in what boundaries were being used for  $[L_t, U_t]_{t \in \mathcal{T}}$  and  $[\hat{L}_t, \hat{U}_t]_{t \in \mathcal{T}}$ . More generally under Conditions G-1–G-4, we have the following abstract Theorem for AsympCSs.

**Theorem 7.2.3** (An abstract AsympCS for well-approximated processes). *Let  $\mathcal{T}$  be a totally ordered infinite set containing a minimal element  $t_0 \in \mathcal{T}$  and let  $(\hat{\theta}_t)_{t \in \mathcal{T}}$  be a real-valued process. Under Conditions G-1–G-4,*

$$\left[ \hat{\theta}_t - L_t, \hat{\theta}_t + U_t \right] \quad (7.14)$$

*forms a  $(1 - \alpha)$ -AsympCS for  $\theta_t$  meaning there exists (on a potentially enriched probability space) some nonasymptotic  $(1 - \alpha)$ -CS  $[\hat{\theta}_t - L_t^*, \hat{\theta}_t + U_t^*]_{t \in \mathcal{T}}$  for  $(\theta_t)_{t \in \mathcal{T}}$ , i.e.*

$$\mathbb{P}\left(\forall t \in \mathcal{T}, \theta_t \in \left[ \hat{\theta}_t - L_t^*, \hat{\theta}_t + U_t^* \right]\right) \geq 1 - \alpha \quad (7.15)$$

such that

$$L_t^*/L_t \xrightarrow{\text{a.s.}} 1 \quad \text{and} \quad U_t^*/U_t \xrightarrow{\text{a.s.}} 1. \quad (7.16)$$

We provide a short proof of Theorem 7.2.3 in Section 7.A.2. Note that the lower boundaries given by  $L_t^*$  and  $\hat{L}_t$  are not the same, but rather  $L_t^*$  is constructed from  $\hat{L}_t$  (and similarly for  $U_t^*$  and  $\hat{U}_t$ ). In the following section, we will use the general recipe of Theorem 7.2.3 to obtain AsympCSs for time-varying means from non-i.i.d. random variables under martingale dependence akin to the Lindeberg CLT [176, 36].

#### 7.2.4 Lindeberg- and Lyapunov-type AsympCSs for time-varying means

The results in Theorem 7.2.2 and Proposition 7.2.1 focused on the situation where the observed random variables are independent and identically distributed, as this is one of the most commonly studied regimes in statistical inference. One may also be interested in the case where means and variances do not remain constant over time, or where observations are

dependent. We will now show that an analogue of Theorem 7.2.2 holds for random variables with time-varying *means and variances* under *martingale dependence*. In this case, rather than the AsympCS covering some fixed  $\mu$ , it covers the average conditional mean thus far:  $\tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \mu_i$  – to be made precise shortly.<sup>4</sup>

Given the additional complexity introduced by considering time-varying conditional distributions, we will first explicitly spell out the conditions required to achieve a time-varying analogue of Theorem 7.2.2. Suppose  $(Y_t)_{t=1}^\infty$  is a sequence of random variables with conditional means and variances given by  $\mu_t := \mathbb{E}(Y_t | Y_1^{t-1})$  and  $\sigma_t^2 := \text{Var}(Y_t | Y_1^{t-1})$ , respectively where we use the shorthand  $Y_1^{t-1}$  for  $\{Y_1, \dots, Y_{t-1}\}$ . First, we require that the average conditional variance  $\tilde{\sigma}_t^2 := \frac{1}{t} \sum_{i=1}^t \sigma_i^2$  either does not vanish, or does so superlinearly; equivalently, we require that the cumulative conditional variance diverges almost surely.

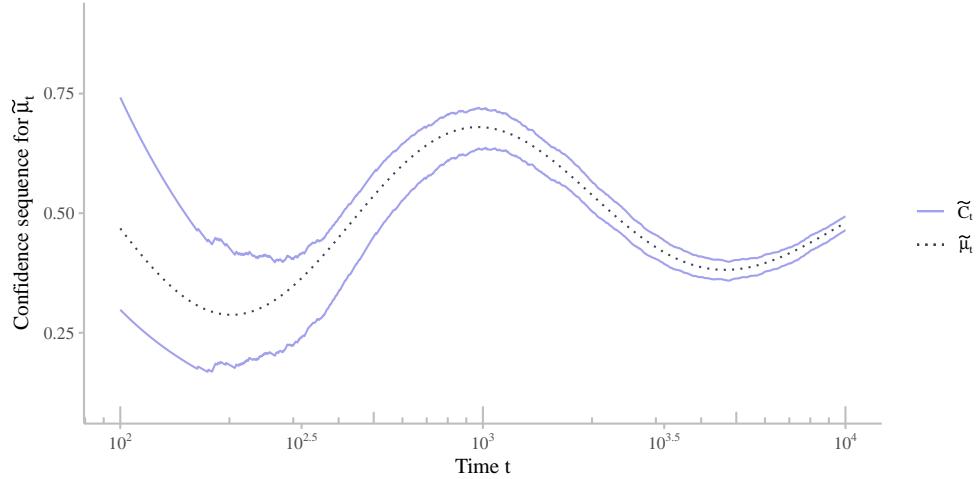


Figure 7.2: A 90%-AsympCS for the time-varying mean  $\tilde{\mu}_t$  using Proposition 7.2.2 with  $\rho$  optimized for  $t^* = 500$  based on the exact solution of Section 7.B.2. Here, we have set  $\mu_t := \frac{1}{2}(1 - \sin(2 \log(e + 10t)) / \log(e + 0.01t))$  to produce the sinusoidal behavior of  $\tilde{\mu}_t$ . Notice that  $\tilde{C}_t$  uniformly captures  $\tilde{\mu}_t$ , adapting to its non-stationarity.

**Condition L-1** (Cumulative variance diverges almost surely). *For each  $t \geq 1$ , let  $\sigma_t^2 := \text{Var}(Y_t | Y_1^{t-1})$  be the conditional variance of  $Y_t$ . Then,*

$$V_t := \sum_{i=1}^t \sigma_i^2 \rightarrow \infty \text{ almost surely.} \quad (7.17)$$

Equation (7.17) can also be interpreted as saying that the average conditional variance

---

<sup>4</sup>Throughout this section and the remainder of the chapter, we use the overhead tilde (e.g.  $\tilde{\mu}_t$ ,  $\tilde{\sigma}_t$ , and  $\tilde{C}_t$ ) to emphasize that these quantities can change over time. For example, Figure 7.2 explicitly displays means and CSs with sinusoidal behaviors resembling a tilde.

$\tilde{\sigma}_t^2 := \frac{1}{t} \sum_{i=1}^t \sigma_i^2$  does not vanish faster than  $1/t$  (if at all), meaning  $\tilde{\sigma}_t^2 = \omega_{\text{a.s.}}(1/t)$ . For example, Condition L-1 would hold if  $\tilde{\sigma}_t^2 \xrightarrow{\text{a.s.}} \sigma_*^2$  for some  $\sigma_*^2 > 0$  or in the i.i.d. case where  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_*^2$ . Second, we require a Lindeberg-type uniform integrability condition on the tail behavior of  $(Y_t)_{t=1}^\infty$ .

**Condition L-2** (Lindeberg-type uniform integrability). *There exists some  $0 < \kappa < 1$  such that*

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[(Y_t - \mu_t)^2 \mathbb{1}((Y_t - \mu_t)^2 > V_t^\kappa) | Y_1^{t-1}]}{V_t^\kappa} < \infty \text{ almost surely.} \quad (7.18)$$

Notice that Equation (7.18) is satisfied if all conditional  $q^{\text{th}}$  moments are almost surely uniformly bounded for some  $q > 2$ , meaning  $1/K \leq \mathbb{E}(|Y_t - \mu_t|^q | Y_1^{t-1}) < K$  a.s. for all  $t \geq 1$  and for some constant  $K > 0$ , or more generally under a Lyapunov-type condition that states  $\sum_{t=1}^{\infty} [\mathbb{E}(|Y_t - \mu_t|^{2+\delta} | Y_1^{t-1}) / \sqrt{V_t}^{2+\delta}] < \infty$  a.s. for some  $\delta > 0$ .<sup>5</sup> Third and finally, we require a consistent variance estimator.

**Condition L-3** (Consistent variance estimation). *Let  $\hat{\sigma}_t^2$  be an estimator of  $\tilde{\sigma}_t^2$  constructed using  $Y_1, \dots, Y_t$  such that*

$$\hat{\sigma}_t^2 / \tilde{\sigma}_t^2 \xrightarrow{\text{a.s.}} 1. \quad (7.19)$$

Note that in the i.i.d. case, (7.19) would hold using the sample variance by the SLLN. More generally for independent but non-identically distributed data, Condition L-3 holds as long as the variation in means vanishes — i.e.  $\frac{1}{t} \sum_{i=1}^t (\mu_i - \tilde{\mu}_t)^2 = o(1)$  — but we will expand on this later in Corollary 7.2.1. Given Conditions L-1, L-2, and L-3, we have the following AsympCS for the time-varying conditional mean  $\tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \mu_i$ .

**Proposition 7.2.2** (Lindeberg-type Gaussian mixture martingale AsympCS). *Let  $(Y_t)_{t=1}^\infty$  be a sequence of random variables with conditional mean  $\mu_t := \mathbb{E}(Y_t | Y_1^{t-1})$  and conditional variance  $\sigma_t^2 := \text{Var}(Y_t | Y_1^{t-1})$ . Then under Assumptions L-1, L-2, and L-3, we have that*

$$\tilde{C}_t \equiv (\hat{\mu}_t \pm \tilde{\mathfrak{B}}_t) := \left( \hat{\mu}_t \pm \sqrt{\frac{t\hat{\sigma}_t^2\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\hat{\sigma}_t^2\rho^2 + 1}{\alpha^2}\right)} \right) \quad (7.20)$$

forms a  $(1 - \alpha)$ -AsympCS for the running average conditional mean  $\tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \mu_i$ .

At a high level, the proof of Proposition 7.2.2 (found in Section 7.A.3) follows from the general AsympCS procedure of Theorem 7.2.3 by using Condition L-2 and Strassen's 1967 strong approximation [249] (not to be confused with his 1964 result that we used in Theorem 7.2.2) to satisfy Conditions G-1 and G-3, and a variant of Robbins' mixture martingale for non-i.i.d. random variables along with Conditions L-1 and L-3 to satisfy Conditions G-2 and G-4.

Notice that if the data happen to be i.i.d., then  $(\tilde{C}_t)_{t=1}^\infty$  is asymptotically equivalent to  $(\bar{C}_t)_{t=1}^\infty$  given in Theorem 7.2.2 (here, “asymptotic equivalence” simply means that the ratio of

---

<sup>5</sup>We show that the Lyapunov-type condition implies Condition L-2 in Section 7.B.5.

the two boundaries converges a.s. to 1). In other words, Proposition 7.2.2 is valid in a more general (non-i.i.d.) setting, but will essentially recover Theorem 7.2.2 in the i.i.d. case. Figure 7.2 illustrates what  $\tilde{C}_t$  may look like in practice. Note that when  $(Y_t)_{t=1}^\infty$  are independent with  $\mu_1 = \mu_2 = \dots = \mu_\star$ , and  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_\star^2$ , it is nevertheless the case that  $\tilde{C}_t$  forms a  $(1 - \alpha)$ -AsympCS for  $\mu_\star$  under the same assumptions as Theorem 7.2.2. In this sense, we can view  $(\tilde{C}_t)_{t=1}^\infty$  as “robust” to deviations from independence and stationarity.<sup>6</sup> A one-sided analogue of Proposition 7.2.2 is presented in Proposition 7.B.2 within Section 7.B.1.

As a near-immediate corollary of Proposition 7.2.2, we have the following Lyapunov-type AsympCS under independent but non-identically distributed random variables.

**Corollary 7.2.1** (Lyapunov-type AsympCS). *Suppose  $(Y_t)_{t=1}^\infty$  is a sequence of independent random variables with individual means and variances given by  $\mu_t := \mathbb{E}(Y_t)$  and  $\sigma_t^2 := \text{Var}(Y_t)$ , respectively. Suppose that in addition to Condition L-1 and the Lyapunov-type condition  $\sum_{i=1}^\infty [\mathbb{E}|Y_i - \mu_i|^{2+\delta}/\sqrt{V_i}^{2+\delta}] < \infty$ , we have the following regularity conditions:*

$$\sum_{i=1}^\infty \frac{\mathbb{E}|Y_i^2 - \mathbb{E}Y_i^2|^{1+\beta}}{V_i^{1+\beta}} < \infty, \quad \tilde{\mu}_t^2 = o(V_t) \text{ a.s.,} \quad \text{and} \quad \frac{1}{t} \sum_{i=1}^t (\mu_i - \tilde{\mu}_t)^2 = o(\tilde{\sigma}_t^2) \quad (7.21)$$

for some  $\beta \in (0, 1)$ . In other words, the higher moments of  $(Y_t)_{t=1}^\infty$ , the running mean  $\tilde{\mu}_t$ , and the cumulative “variation in means”  $\sum_{i=1}^t (\mu_i - \tilde{\mu}_t)^2$  all cannot diverge too quickly relative to  $(V_t)_{t=1}^\infty$ . Then using the sample variance for  $\tilde{\sigma}_t^2$ ,  $\tilde{C}_t$  forms a  $(1 - \alpha)$ -AsympCS for the running average mean  $\tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \mu_i$ .

Clearly, the conditions of (7.21) are trivially satisfied if the means and the  $(2 + 2\beta)^{\text{th}}$  absolute central moments are uniformly bounded over time. Since Conditions L-1 and L-2 hold by the assumptions of Corollary 7.2.1, the proof in Section 7.A.4 simply shows how the conditions in (7.21) imply Condition L-3.

As suggested by Section 7.2.3, we can combine Theorem 7.2.3 with essentially any other Gaussian boundary, and indeed there are others that can yield Lindeberg- and Lyapunov-type AsympCSs but we do not enumerate any more here, though we do mention one inspired by Robbins [217, Eq. (20)] in passing in Section 7.2.6. The next section discusses how all of the aforementioned AsympCSs satisfy a certain formal asymptotic coverage guarantee.

## 7.2.5 Asymptotic coverage and type-I error control

While the AsympCSs derived thus far serve as sequential analogues of CLT-based CIs, it is not immediately obvious whether the bounds introduced in the previous section enjoy similar asymptotic coverage (equivalently, type-I error) guarantees. We will now give a positive answer to this question by showing that after appropriate tuning, our AsympCSs have asymptotic  $(1 - \alpha)$ -coverage uniformly for all  $t \geq m$  as  $m \rightarrow \infty$  (to be formalized in Definition 7.2.4).

---

<sup>6</sup>Here, the term “robust” should not be interpreted in the same spirit as “doubly robust”, where the latter is specific to the discussions surrounding functional estimation and causal inference in Section 7.3.

Recall that the coverage of CLT-based CIs is at least  $(1 - \alpha)$  in the limit:

$$\liminf_{n \rightarrow \infty} \mathbb{P}(\mu \in \dot{C}_n) \geq 1 - \alpha, \quad (7.22)$$

but what is the right time-uniform analogue of (7.22)? Since any single AsympCS will simply have *some* coverage, we provide the following definition as a time-uniform analogue of (7.22) for *sequences* of sets that start later and later. In the definition that follows,  $m \geq 1$  will play the role of an “initial peeking time” and time-uniformity will be provided with respect to all  $t \geq m$ .

**Definition 7.2.4** (Asymptotic time-uniform coverage). For each  $m \in \mathbb{N}$ , let  $(C_t(m))_{t=m}^{\infty}$  be a sequence of sets, and let  $\alpha \in (0, 1)$  be the desired miscoverage level. We say that  $(C_t(m))_{t=m}^{\infty}$  has *asymptotic time-uniform*  $(1 - \alpha)$ -coverage for  $(\mu_t)_{t=1}^{\infty}$  if

$$\liminf_{m \rightarrow \infty} \mathbb{P}(\forall t \geq m, \mu_t \in C_t(m)) \geq 1 - \alpha, \quad (7.23)$$

and we say that this coverage is *sharp* if the above inequality holds with equality and the limit infimum is replaced by a limit.

To the best of our knowledge, the existing literature lacks a concrete definition of asymptotic time-uniform coverage (or type-I error control) like Definition 7.2.4, but sequences of AsympCSs satisfying (7.23) have been implicit in Robbins [217] and Robbins and Siegmund [220], and the followup work of Bibaut et al. [34]. In what follows, we provide (sharp) coverage guarantees for our AsympCSs. Furthermore, in Section 7.2.6 we strictly improve on aforementioned bounds by Robbins [217] and Robbins and Siegmund [220]. Furthermore, we note that a bound Bibaut et al. [34] is in a certain sense equivalent to one that we provide here.

In order to obtain asymptotic time-uniform coverage, we need a stronger variant of Condition L-3 so that variances are estimated at polynomial rates (rather than at arbitrary rates).

**Condition L-3- $\eta$**  (Polynomial rate variance estimation). *There exists some  $0 < \eta < 1$  such that*

$$\hat{\sigma}_t^2 - \tilde{\sigma}_t^2 = o\left(\frac{(t\tilde{\sigma}_t^2)^{\eta}}{t}\right) \quad \text{almost surely.} \quad (7.24)$$

Note that while Condition L-3- $\eta$  is stronger than Condition L-3, it is still quite mild. For instance, if  $\tilde{\sigma}_t^2$  is uniformly bounded, then (7.24) simply requires that  $\hat{\sigma}_t^2 - \tilde{\sigma}_t^2 = o(t^{\eta-1})$  (i.e. strong consistency at *any* polynomial rate, potentially much slower than  $t^{-1/2}$ ). Moreover, in the i.i.d. case with at least  $(2 + \delta)$  finite absolute moments, Condition L-3- $\eta$  always holds by the SLLNs of Marcinkiewicz and Zygmund [184].

Our goal now is to show that sequences of AsympCSs given in Proposition 7.2.2 have asymptotic time-uniform coverage, and we will achieve this by effectively tuning them for later and later start times. Recall that Section 7.B.2 allows us to choose the parameter  $\rho > 0$  so that the AsympCS is tightest at some particular time — we will now choose  $\rho_m$  based on the

first peeking time  $m$  as

$$\rho_m := \rho(\hat{\sigma}_m^2 m \log(m \vee e)) \equiv \sqrt{\frac{-2 \log \alpha + \log(-2 \log \alpha) + 1}{\hat{\sigma}_m^2 m \log(m \vee e)}}.$$
 <sup>7</sup> (7.25)

Then, let  $(\tilde{C}_t(m))_{t=m}^\infty$  be the Gaussian mixture AsympCS with  $\rho_m$  plugged into the expression of the boundary for all  $t \geq m$ :

$$\tilde{C}_t(m) := \left( \hat{\mu}_t \pm \sqrt{\frac{t \hat{\sigma}_t^2 \rho_m^2 + 1}{t^2 \rho_m^2} \log \left( \frac{t \hat{\sigma}_t^2 \rho_m^2 + 1}{\alpha^2} \right)} \right).$$
 (7.26)

In other words,  $(\tilde{C}_t(m))_{t=m}^\infty$  should be thought of as an AsympCS that only starts after time  $m$ , and is vacuous beforehand. The following theorem formalizes the coverage guarantees satisfied by this *sequence* of AsympCSs as  $m \rightarrow \infty$ .

**Theorem 7.2.5** (Asymptotic  $(1 - \alpha)$ -coverage for Gaussian mixture AsympCSs). *Given the same setup as Proposition 7.2.2 and Conditions L-1, L-2, and L-3- $\eta$ , the AsympCSs  $(\tilde{C}_t(m))_{t=m}^\infty$  given in (7.26) have sharp asymptotic  $(1 - \alpha)$ -coverage for  $\tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \mu_i$  as  $m \rightarrow \infty$ , meaning*

$$\lim_{m \rightarrow \infty} \mathbb{P} \left( \forall t \geq m, \tilde{\mu}_t \in \tilde{C}_t(m) \right) = 1 - \alpha.$$
 (7.27)

The proof can be found in Section 7.A.7. Clearly, when the mean is constant – i.e.  $\mu_1 = \mu_2 = \dots = \mu_\star$  – the above also holds for the running intersection of intervals  $\bigcap_{m \leq s \leq t} \tilde{C}_s(m)$ . Notice that in the i.i.d. setting, as  $m \rightarrow \infty$ ,  $(\tilde{C}_t(m))_{t=m}^\infty$  is asymptotically equivalent to  $(\bar{C}_t^G(m))_{t=m}^\infty$  given by

$$\bar{C}_t^G(m) := \left( \hat{\mu}_t \pm \hat{\sigma}_t \sqrt{\frac{t \bar{\rho}_m^2 + 1}{t^2 \bar{\rho}_m^2} \log \left( \frac{t \bar{\rho}_m^2 + 1}{\alpha^2} \right)} \right),$$
 (7.28)

where  $\bar{\rho}_m := \rho(m \log(m \vee e))$ . A quick inspection of the proof will reveal that (7.28) also satisfies the coverage guarantee provided in Theorem 7.2.5 under the condition that  $\hat{\sigma}_t^2 \rightarrow \sigma_\star^2 > 0$  almost surely. In summary, (7.28) can be thought of as an analogue of (7.26) for the AsympCSs that were derived in Theorem 7.2.2.

## 7.2.6 Asymptotic confidence sequences using Robbins' delayed start

As is clear from Theorem 7.2.3, virtually any boundary for Gaussian observations can be used to derive an AsympCS as long as an appropriate strong invariance principle can be applied under the given assumptions – indeed, Theorem 7.2.2, Proposition 7.2.1, Proposition 7.2.2, and Corollary 7.2.1 are all instantiations of the general phenomenon outlined in Theorem 7.2.3.

---

<sup>7</sup>In fact,  $\rho_m$  can be replaced by  $\rho(\hat{\sigma}_m^2 m d_m)$  where  $(d_m)_{m=1}^\infty$  is any positive increasing sequence diverging to  $\infty$ .

Another AsympCS that may be of interest to practitioners is one that leverages Robbins' CS for means of Gaussian random variables with a delayed start time [217, Eq. (20)]. In a nutshell, Robbins calculated a lower bound on the probability that a centered Gaussian random walk would remain within a particular two-sided boundary for all times  $t \geq m$  given some starting time  $m \geq 1$ . That is, he showed that for i.i.d. Gaussians  $(Z_t)_{t=1}^{\infty}$  with mean zero and variance  $\sigma^2$ , letting  $G_t := \sum_{i=1}^t Z_i/\sigma$ , and for any  $a > 0$ ,

$$\mathbb{P}\left(\forall t \geq m : |G_t| < \sqrt{t(a^2 + \log(t/m))}\right) \geq 1 - \Lambda(a), \quad (7.29)$$

where  $\Lambda(a) := 2(1 - \Phi(a) + a\phi(a))$  and  $\Phi$  and  $\phi$  are the CDF and PDF of a standard Gaussian, respectively. In particular, setting  $a \in (0, \infty)$  so that  $\Lambda(a) = \alpha$  yields a two-sided  $(1 - \alpha)$ -boundary for the Gaussian random walk  $(G_t)_{t=1}^{\infty}$ , and indeed, a solution to  $\Lambda(a) = \alpha$  always exists and is trivial to compute due to the fact that  $\Lambda$  is strictly decreasing, starting at  $\Lambda(0) = 1$  and  $\lim_{a \rightarrow \infty} \Lambda(a) = 0$ . In a followup paper, Robbins and Siegmund [220] extended the ideas of Robbins [217] to a large class of boundaries for Wiener processes so that the probabilistic inequality in (7.29) can be shown to be an equality when  $|G_t|$  is replaced by the absolute value of a Wiener process (which would imply the inequality in (7.29) for i.i.d. standard Gaussians as a corollary). Using this fact within the general framework of Theorem 7.2.3 combined with the strong invariance principle of Strassen [249] and the techniques found in the proof of Theorem 7.2.5 yields the following result.

**Proposition 7.2.3** (Delayed-start AsympCS). *Consider the same setup as Theorem 7.2.5 so that  $(Y_t)_{t=1}^{\infty}$  have conditional means and variances given by  $\mu_t := \mathbb{E}(Y_t | Y_1^{t-1})$  and  $\sigma_t^2 := \text{Var}(Y_t | Y_1^{t-1})$ . Then under Conditions L-1, L-2, and L-3, we have that for any  $m \geq 1$ ,*

$$\tilde{C}_t^{\text{DS}}(m) := \left( \hat{\mu}_t \pm \hat{\sigma}_t \sqrt{\frac{1}{t} \left[ a^2 + \log \left( \frac{t\hat{\sigma}_t^2}{m\hat{\sigma}_m^2} \right) \right]} \right) \quad \text{if } t \geq m \quad (7.30)$$

(and all of  $\mathbb{R}$  otherwise) forms a  $(1 - \alpha)$ -AsympCS for  $\tilde{\mu}_t$ , where  $a$  is chosen so that  $\Lambda(a) = \alpha$ . Furthermore, under Condition L-3- $\eta$ ,  $\tilde{C}_t^{\text{DS}}(m)$  has sharp asymptotic time-uniform  $(1 - \alpha)$ -coverage in the sense of Definition 7.2.4.

The proof is provided in Section 7.A.9. Similar to the relationship between Theorem 7.2.5 and (7.28), we have that if variances happen to converge  $\tilde{\sigma}_t^2 \rightarrow \sigma_*^2 > 0$  almost surely, then as a corollary of Proposition 7.2.3, the sequence  $(\tilde{C}_t^{\text{DS}*}(m))_{t=1}^{\infty}$  given by

$$\tilde{C}_t^{\text{DS}*}(m) := \left( \hat{\mu}_t \pm \hat{\sigma}_t \sqrt{\frac{1}{t} \left( a^2 + \log \left( \frac{t}{m} \right) \right)} \right) \quad \text{for } t \geq m \quad (7.31)$$

has asymptotic time-uniform coverage as in Definition 7.2.4. This can be seen as a generalization and improvement of results implied by Robbins [217] and Robbins and Siegmund [220], and has connections to Bibaut et al. [34]. We elaborate on these below.

### 7.2.6.1 Relationship to Robbins and Siegmund

Informally, Robbins and Siegmund [220, Theorem 2(i)] show that for independent and identically distributed random variables  $(Y_t)_{t=1}^{\infty} \sim \mathbb{P}$  with mean zero and unit variance (without loss of generality), the probability of their scaled partial sums  $S_t := \sum_{i=1}^t Y_i$  exceeding a particular boundary behaves like a rescaled Wiener process exceeding that boundary. Consequently, for i.i.d. data with known variance, their result combined with Robbins' delayed start [217, Eq. (20)] implies an asymptotic coverage guarantee (in the sense of Definition 7.2.4) for  $\tilde{C}_t^{\text{DS}*}$  given in (7.31) but with  $\hat{\sigma}_t$  replaced by the true  $\sigma$ . As such, (7.31) should be thought of as a generalization of their boundary when variances are unknown under martingale dependence. Nevertheless, Proposition 7.2.3 is strictly more general, allowing for variances to never converge.

### 7.2.6.2 Relationship to Bibaut et al. [34]

In version 1 of Bibaut et al. [34], the authors derive a particular AsympCS and show that sequences thereof satisfy an asymptotic coverage guarantee in the same sense as Definition 7.2.4. That bound resembles — but is always looser than — a corollary of Proposition 7.2.3 in (7.31) for a fixed  $m \geq 1$ ; see Claim 7.B.1. Nevertheless, it was their asymptotic type-I error results that inspired us to show that the same guarantees hold for the bounds in (7.26), (7.30), and (7.31), and with our explicit conditions on how variances are allowed to diverge in the former two, rather than their implicit conditions (or sufficient conditions in special cases) through almost-surely convergent variance-stabilized pseudo-outcomes. Nevertheless, one can always apply our bounds to these same variance-stabilized pseudo-outcomes to weaken these implicit assumptions. After our advances, the second version of their paper introduced a bound called the “running maximum likelihood SPRT” (rmlSPRT) which is identical to the corollary of Proposition 7.2.3 found in (7.31) (modulo differences in variance estimation techniques); see Claim 7.B.2. Since the exact connections may not be obvious to the reader, we derive them explicitly in Section 7.B.7. Despite their rmlSPRT being identical to (7.31), we remark that their paper focuses on testing guarantees and contains several additional interesting investigations including a sophisticated analysis of the expected rejection time, enriching the landscape of asymptotic anytime-valid methodology.

## 7.3 Illustration: Causal effects and semiparametric estimation

Given the groundwork laid in Section 7.2, we now demonstrate the use of AsympCSs for conducting anytime-valid causal inference. Since it is an important and thoroughly studied functional, we place a particular emphasis on the average treatment effect (ATE) for illustrative purposes but we discuss how these techniques apply to semiparametric functional estimation more generally in Section 7.3.5. The literature on semiparametric functional inference often falls within the asymptotic regime and hence AsympCSs form a natural time-uniform extension thereof.

It is important to note that obtaining AsympCSs for the ATE is not as simple as directly applying the theorems of Section 7.2.1 to some appropriately chosen augmented inverse-

probability-weighted (AIPW) influence functions (otherwise the illustration of this section would have been trivial). Indeed, satisfying the conditions of the aforementioned theorems – Theorem 7.2.3 in particular – in the presence of infinite-dimensional nuisance parameters is nontrivial and the analysis proceeds rather differently from the fixed- $n$  setting. Nevertheless, after introducing and carefully analyzing sequential sample splitting and cross-fitting (Section 7.3.1), we will see that asymptotic time-uniform inference for the ATE is possible.

To solidify the notation and problem setup, suppose that we observe a (potentially infinite) sequence of i.i.d. variables  $Z_1, Z_2, \dots$  from a distribution  $\mathbb{P}$  where  $Z_t := (X_t, A_t, Y_t)$  denotes the  $t^{\text{th}}$  subject's triplet and  $X_t \in \mathbb{R}^d$  are their measured baseline covariates,  $A_t \in \{0, 1\}$  is the treatment that they receive, and  $Y_t \in \mathbb{R}$  is their measured outcome after treatment. Our target estimand is the average treatment effect (ATE)  $\psi$  defined as

$$\psi := \mathbb{E}(Y^1 - Y^0), \quad (7.32)$$

where  $Y^a$  is the counterfactual outcome for a randomly selected subject had they received treatment  $a \in \{0, 1\}$ . The ATE  $\psi$  can be interpreted as the average population outcome if everyone were treated  $\mathbb{E}(Y^1)$  versus if no one were treated  $\mathbb{E}(Y^0)$ . Under standard causal identification assumptions – typically referred to as consistency, positivity, and exchangeability (see e.g. Kennedy [153, §2.2]) – we have that  $\psi$  can be written as a (non-counterfactual) functional of the distribution  $\mathbb{P}$ :

$$\psi \equiv \psi(\mathbb{P}) = \mathbb{E}\{\mathbb{E}(Y | X, A = 1) - \mathbb{E}(Y | X, A = 0)\}. \quad (7.33)$$

Throughout the remainder of this section, we will operate under these identification assumptions and aim to derive efficient AsympCSs for  $\psi$  using tools from semiparametric theory. At a high level, we will construct AsympCSs for  $\psi$  by combining the results of Section 7.2 with sample averages of *influence functions* for  $\psi$  and in the ideal case, these influence functions will be *efficient* (in the semiparametric sense).

### 7.3.1 Sequential sample splitting and cross fitting

Following Robins et al. [226], Zheng and van der Laan [296], and Chernozhukov et al. [56], we employ sample splitting to derive an estimate  $\hat{f}$  of the influence function  $f$  on a “training” sample, and evaluate  $\hat{f}$  on values of  $Z_t$  in an independent “evaluation” sample. Sample splitting sidesteps complications introduced from “double-dipping” (i.e. using  $Z_t$  to both construct  $f$  and evaluate  $\hat{f}(Z_t)$ ) and greatly simplifies the analysis of the downstream estimator. Since the aforementioned authors employed sample splitting in the *batch* (non-sequential) regime while we are concerned with settings where data are continually observed in an online stream over time, we modify the sample splitting procedure as follows. We will denote  $\mathcal{D}_\infty^{\text{trn}}$  and  $\mathcal{D}_\infty^{\text{eval}}$  as the “training” and “evaluation” sets, respectively. At time  $t$ , we assign  $Z_t$  to either group with equal probability:

$$Z_t \in \begin{cases} \mathcal{D}_\infty^{\text{trn}} & \text{with probability } 1/2, \\ \mathcal{D}_\infty^{\text{eval}} & \text{otherwise.} \end{cases}$$

Note that at time  $t + 1$ ,  $Z_t$  is *not* re-randomized into either split — once  $Z_t$  is randomly assigned to one of  $\mathcal{D}_\infty^{\text{trn}}$  or  $\mathcal{D}_\infty^{\text{eval}}$ , they remain in that split for the remainder of the study. In this way, we can write  $\mathcal{D}_\infty^{\text{trn}} = (Z_1^{\text{trn}}, Z_2^{\text{trn}}, \dots)$  and  $\mathcal{D}_\infty^{\text{eval}} = (Z_1^{\text{eval}}, Z_2^{\text{eval}}, \dots)$  and think of these as independent, sequential observations from a common distribution  $\mathbb{P}$ . To keep track of how many subjects have been randomized to  $\mathcal{D}_\infty^{\text{trn}}$  and  $\mathcal{D}_\infty^{\text{eval}}$  at time  $t$ , define

$$T := |\mathcal{D}_\infty^{\text{eval}}| \quad \text{and} \quad T' := |\mathcal{D}_\infty^{\text{trn}}| \equiv t - T, \quad (7.34)$$

where we have left the dependence on  $t$  implicit.

*Remark 19.* Strictly speaking, under the i.i.d. assumption, we do not need to randomly assign subjects to training and evaluation groups for the forthcoming results to hold (e.g. we could simply assign even-numbered subjects to  $\mathcal{D}_\infty^{\text{trn}}$  and odd-numbered subjects to  $\mathcal{D}_\infty^{\text{eval}}$ ). However, the analysis is not further complicated by this randomization, and it can be used to combat bias in treatment assignments when the i.i.d. assumption is violated [98].

### 7.3.1.1 The sequential sample-split estimators $(\hat{\psi}_t^{\text{split}})_{t=1}^\infty$

After employing sequential sample splitting, the sequence of sample-split estimators  $(\hat{\psi}_t^{\text{split}})_{t=1}^\infty$  for  $\psi$  are given by

$$\hat{\psi}_t^{\text{split}} := \frac{1}{T} \sum_{i=1}^T \hat{f}_{T'}(Z_i^{\text{eval}}), \quad (7.35)$$

where  $\hat{f}_{T'}$  is given by the so-called *efficient influence function* (a brief review of semiparametric efficient estimators can be found in Section 7.B.8),

$$f(z) \equiv f(x, a, y) := \{\mu^1(x) - \mu^0(x)\} + \left( \frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)} \right) \{y - \mu^a(x)\}, \quad (7.36)$$

with  $\eta \equiv (\mu^1, \mu^0, \pi)$  replaced by  $\hat{\eta}_{T'} \equiv (\hat{\mu}_{T'}^1, \hat{\mu}_{T'}^0, \bar{\pi}_{T'})$  — where  $\bar{\pi}_{T'}$  may be an estimator  $\hat{\pi}_{T'}$  of the propensity score  $\pi$ , or the propensity score itself, depending on whether one is considering an observational study or randomized experiment — so that  $\hat{\eta}_{T'}$  is built solely from  $\mathcal{D}_\infty^{\text{trn}}$ . The sample splitting procedure for constructing  $\hat{\psi}_t^{\text{split}}$  is summarized pictorially in Figure 7.3. In the batch setting for a fixed sample size, (7.35) is often referred to as the *augmented inverse probability weighted* (AIPW) estimator [227, 231] (an instantiation of so-called “one-step correction” in the semiparametrics literature) and we adopt similar nomenclature here.

### 7.3.1.2 The sequential cross-fit estimators $(\hat{\psi}_t^\times)_{t=1}^\infty$

A commonly cited downside of sample splitting is the loss in efficiency by using  $T \approx t/2$  subjects instead of  $t$  when evaluating the sample mean  $\frac{1}{T} \sum_{i=1}^T \hat{f}_{T'}(Z_i^{\text{eval}})$ . An easy fix is to *cross-fit*: swap the two samples, using the evaluation set  $\mathcal{D}_\infty^{\text{eval}}$  for training and the training set  $\mathcal{D}_\infty^{\text{trn}}$  for evaluation to recover the full sample size of  $t \equiv T + T'$  [226, 296, 56]. That is,

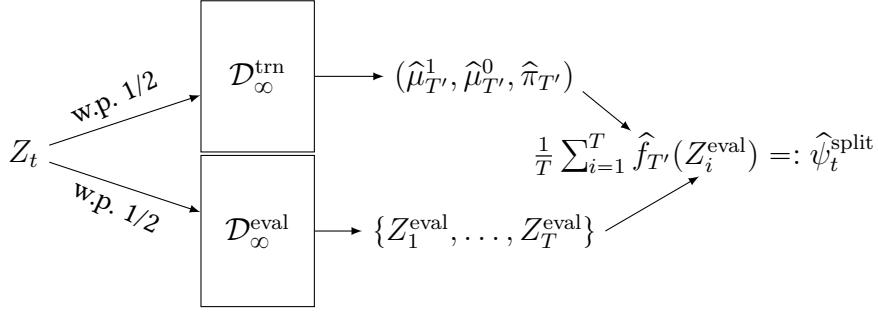


Figure 7.3: A schematic illustrating sequential sample splitting. At each time step  $t$ , the new observation  $Z_t$  is randomly assigned to  $\mathcal{D}_\infty^{\text{trn}}$  or  $\mathcal{D}_\infty^{\text{eval}}$  with equal probability (1/2). Nuisance function estimators  $(\hat{\mu}_{T'}^1, \hat{\mu}_{T'}^0, \hat{\pi}_{T'})$  are constructed using  $\mathcal{D}_\infty^{\text{trn}}$  which then yield  $\hat{f}_{T'}$ . The sample-split estimator  $\hat{\psi}_t^{\text{split}}$  is defined as the sample average  $\frac{1}{T} \sum_{i=1}^T \hat{f}_{T'}(Z_i^{\text{eval}})$  where each  $Z_i^{\text{eval}} \in \mathcal{D}_\infty^{\text{eval}}$ .

construct  $\hat{f}_T$  solely from  $\mathcal{D}_\infty^{\text{eval}}$  and define the cross-fit estimator  $\hat{\psi}_t^\times$  as

$$\hat{\psi}_t^\times := \frac{\sum_{i=1}^T \hat{f}_{T'}(Z_i^{\text{eval}}) + \sum_{i=1}^{T'} \hat{f}_T(Z_i^{\text{trn}})}{t}, \quad (7.37)$$

and the associated cross-fit variance estimate

$$\widehat{\text{Var}}_t(\hat{f}) := \frac{\widehat{\text{Var}}_T(\hat{f}_{T'}) + \widehat{\text{Var}}_{T'}(\hat{f}_T)}{2}, \quad (7.38)$$

where  $\widehat{\text{Var}}_T(\hat{f}_{T'})$  is the  $\mathcal{D}_\infty^{\text{eval}}$ -sample variance of the pseudo-outcomes  $(\hat{f}_{T'}(Z_i^{\text{eval}}))_{i=1}^T$  and similarly for  $\widehat{\text{Var}}_{T'}$  (we deliberately omit the subscript on  $\hat{f}$  in the left-hand side of (7.38)). All of the results that follow are stated in terms of the cross-fit estimators  $(\hat{\psi}_t^\times)_{t=1}^\infty$  but they also hold using  $(\hat{\psi}_t^{\text{split}})_{t=1}^\infty$  instead. With the setup of Section 7.B.8 and Section 7.3.1 in mind, we are ready to derive AsympCSs for  $\psi$ , first in randomized experiments.

### 7.3.2 Asymptotic confidence sequences in randomized experiments

Consider a sequential randomized experiment so that a subject with covariates  $x$  has a known propensity score

$$\pi(x) := \mathbb{P}(A = 1 \mid X = x). \quad (7.39)$$

Consider the cross-fit AIPW estimator  $\hat{\psi}_t^\times$  as given in (7.37) but with estimated propensity scores —  $\hat{\pi}_{T'}(x)$  and  $\hat{\pi}_T(x)$  — replaced by their true values  $\pi(x)$ , and with  $\hat{\mu}_T^a$  and  $\hat{\mu}_T^a$  being possibly misspecified estimators for  $\mu^a$ . We will assume that  $\hat{\mu}_t^a$  converges to some function  $\bar{\mu}^a$ , which need not coincide with  $\mu^a$ . In what follows, when we use  $\hat{\mu}_t^a$  or  $\hat{f}_t$  in writing  $\|\hat{\mu}_t^a - \bar{\mu}^a\|_{L_2(\mathbb{P})}$  or  $\|\hat{f}_t - \bar{f}\|_{L_2(\mathbb{P})}$ , we are referring to large-sample properties of the estimator (and hence  $\hat{f}_t$  could be replaced by  $\hat{f}_{T'}$  or  $\hat{f}_T$  without loss of generality).

**Theorem 7.3.1** (AsympCSs for the ATE in randomized experiments). *Let  $\hat{\psi}_t^\times$  be the cross-fit AIPW estimator as in (7.37). Suppose  $\|\hat{\mu}_t^a(X) - \bar{\mu}^a(X)\|_{L_2(\mathbb{P})} = o(1)$  for each  $a \in \{0, 1\}$  where  $\bar{\mu}^a$  is some function (but need not be  $\mu^a$ ), and hence  $\|\hat{f}_t - \bar{f}\|_{L_2(\mathbb{P})} = o(1)$  for some influence function  $\bar{f}$ . Suppose that propensity scores are bounded away from 0 and 1, i.e.  $\pi(X) \in [\delta, 1 - \delta]$  almost surely for some  $\delta > 0$ , and suppose that  $\text{Var}(\bar{f}(Z)) < \infty$ . Then for any constant  $\rho > 0$ ,*

$$\hat{\psi}_t^\times \pm \sqrt{\widehat{\text{Var}}_t(\hat{f})} \cdot \sqrt{\frac{2(t\rho^2 + 1)}{t^2\rho^2} \log\left(\frac{\sqrt{t\rho^2 + 1}}{\alpha}\right)} \quad (7.40)$$

forms a  $(1 - \alpha)$ -AsympCS for  $\psi$ .

The proof in Appendix 7.A.5 combines an analysis of the almost-sure convergence of  $(\hat{\psi}_t^\times - \psi)$  with the AsympCS of Theorem 7.2.2. Notice that since  $\hat{\mu}_t^a$  is consistent for a function  $\bar{\mu}^a$ , we have that  $\hat{f}_t$  is converging to some influence function  $\bar{f}$  of the form

$$\bar{f}(z) \equiv \bar{f}(x, a, y) := \{\bar{\mu}^1(x) - \bar{\mu}^0(x)\} + \left(\frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)}\right)\{y - \bar{\mu}^a(x)\}. \quad (7.41)$$

In practice, however, one must choose  $\hat{\mu}_t^a$ . As alluded to at the beginning of Section 7.3, the best possible influence function is the EIF  $f(z)$  defined in (7.36), and thus it is natural to attempt to construct  $\hat{\mu}_t^a$  so that  $\|\hat{f}_t - f\|_{L_2(\mathbb{P})} = o(1)$ . The resulting confidence sequences would inherit such optimality properties, a point which we discuss further in Section 7.B.9.

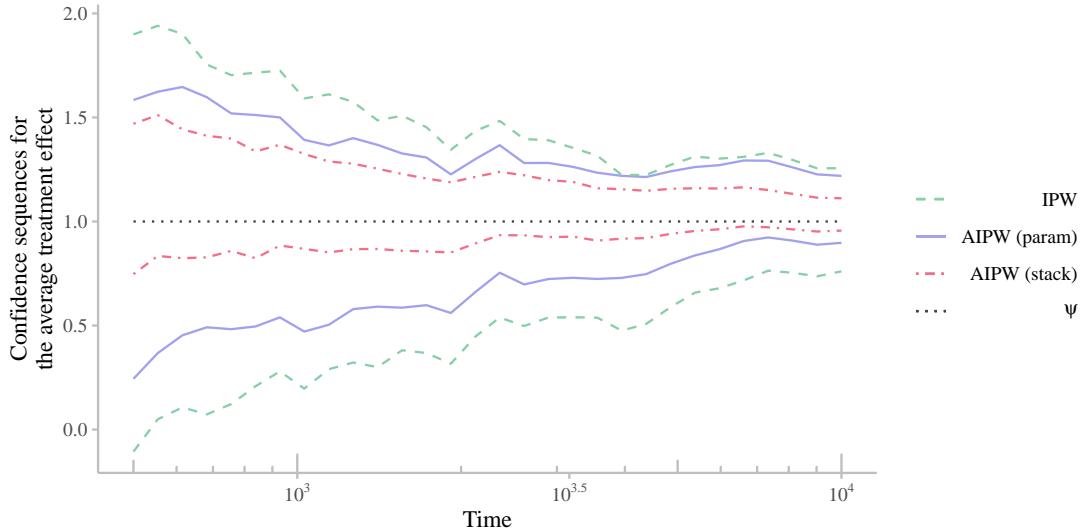


Figure 7.4: Three 90%-AsympCSs for the average treatment effect in a simulated randomized experiment using different regression estimators. Notice that all three confidence sequences uniformly capture the average treatment effect  $\psi$ , but more sophisticated models do so more efficiently, with AIPW+stacking greatly outperforming IPW.

Since  $\mu^a$  is simply a conditional mean function, we can use virtually any regression techniques to estimate it. Here we will consider the general approach of *stacking* introduced by Breiman [42] and further studied by Tsybakov [255] and van der Laan et al. [259] (see also [206, 207]) under the names of “aggregation” and “Super Learning” respectively. In short, stacking uses cross-validation to choose a weighted combination of  $K$  candidate predictors where the weights are chosen based on data in held-out samples. Importantly (and under certain conditions), the stacked predictor will have a mean squared error that scales with that of the best of the  $K$  predictors up to an additive  $\log K$  term [255, 262]. This advantage can be seen empirically in Figure 7.4 where the true regression functions  $\mu^0$  and  $\mu^1$  are nonsmooth and nonlinear in  $x$ . Such advantages via stacking are not new – we are only highlighting the observation that similar phenomena carry over to AsympCSs.

So far, the use of flexible regression techniques like stacking were used only for the purposes of deriving sharper AsympCSs in sequential randomized experiments. In observational studies, however, consistent estimation of nuisance functions at fast rates is essential to the construction of *valid* fixed- $n$  CIs, and indeed the same is true for AsympCSs.

### 7.3.3 Asymptotic confidence sequences in observational studies

Consider now a sequential observational study (e.g. we are able to continuously monitor  $(X_t, A_t, Y_t)_{t=1}^\infty$  but do not know  $\pi(x)$  exactly, or we are in a sequentially randomized experiment with noncompliance, etc.). The only difference in this setting with respect to setup is the fact that  $\pi(x)$  is no longer known and must be estimated. As in the fixed- $n$  setting, this complicates estimation and inference. The following theorem provides the conditions under which we can construct AsympCSs for  $\psi$  using the cross-fit AIPW estimator in observational studies.

**Theorem 7.3.2** (Confidence sequence for the ATE in observational studies). *Consider the same setup as Theorem 7.3.1 but with  $\pi(x)$  no longer known. Suppose that regression functions and propensity scores are consistently estimated in  $L_2(\mathbb{P})$  at a product rate of  $o(\sqrt{\log t/t})$ , meaning that*

$$\|\hat{\pi}_t - \pi\|_{L_2(\mathbb{P})} \sum_{a=0}^1 \|\hat{\mu}_t^a - \mu^a\|_{L_2(\mathbb{P})} = o\left(\sqrt{\log t/t}\right).$$

Moreover, suppose that  $\|\hat{f}_t - f\|_{L_2(\mathbb{P})} = o(1)$  where  $f$  is the efficient influence function (7.36) and that  $\text{Var}(f(Z)) < \infty$ . Then for any constant  $\rho > 0$ ,

$$\hat{\psi}_t^\times \pm \sqrt{\widehat{\text{Var}}_t(\hat{f})} \cdot \sqrt{\frac{2(t\rho^2 + 1)}{t^2\rho^2} \log\left(\frac{\sqrt{t\rho^2 + 1}}{\alpha}\right)}$$

forms a  $(1 - \alpha)$ -AsympCS for  $\psi$ .

The proof in Appendix 7.A.5.2 proceeds similarly to the proof of Theorem 7.3.1 by combining Theorem 7.2.2 with an analysis of the almost-sure behavior of  $(\hat{\psi}_t^\times - \psi)$ . Notice that the nuisance estimation rate of  $\sqrt{\log t/t}$  is slower than  $1/\sqrt{t}$  which is usually required in the fixed- $n$  regime,

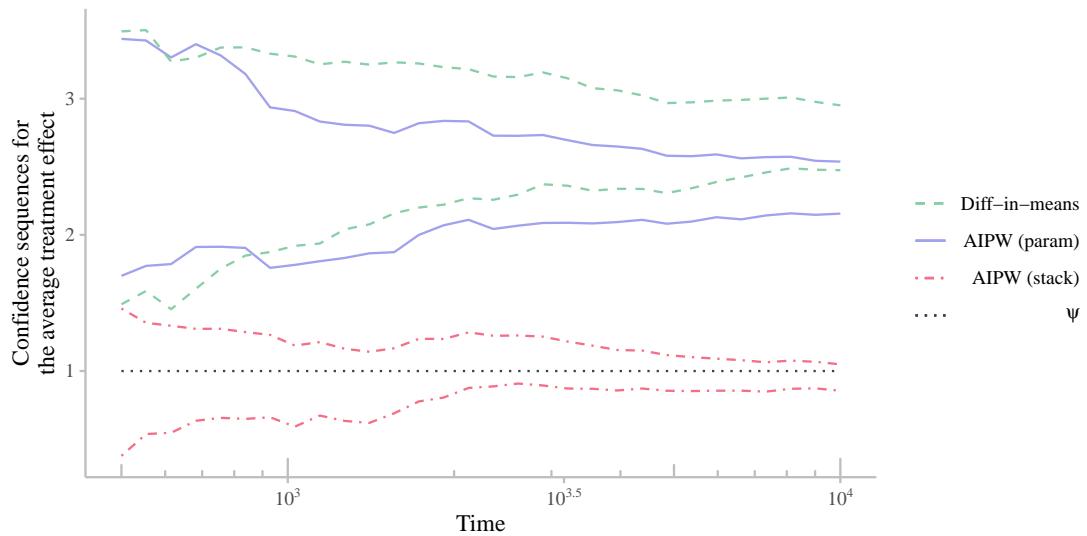


Figure 7.5: Three 90%-AsympCSs for the ATE in an observational study using three different estimators — a difference-in-means estimator, AIPW with parametric models, and AIPW with an ensemble of predictors combined via stacking. Unlike the randomized setup, only the stacking ensemble is consistent, since the other two are misspecified. Not only is the stacking-based AsympCS converging to  $\psi$ , but it is also the tightest of the three models at each time step.

but we do require almost-sure convergence rather than convergence in probability.

Unlike the experimental setting of Section 7.3.2, Theorem 7.3.2 requires that  $\hat{\mu}_t^a$  and  $\hat{\pi}_t$  consistently estimate  $\mu^a$  and  $\pi$ , respectively. As a consequence, the stacking-based AIPW AsympCS is both the tightest of the three *and* is uniquely consistent for  $\psi$  (see Figure 7.5).

### 7.3.4 The running average of individual treatment effects

The results in Sections 7.3.2 and 7.3.3 considered the classical regime where the ATE  $\psi$  is a fixed functional that does not change over time. Consider a strict generalization where distributions – and hence individual treatment effects in particular – may change over time. In other words,

$$\psi_t := \mathbb{E} \{Y_t^1 - Y_t^0\} \stackrel{(*)}{=} \mathbb{E} \{\mathbb{E}(Y_t | X_t, A_t = 1) - \mathbb{E}(Y_t | X_t, A_t = 0)\}, \quad (7.42)$$

where the equality  $(*)$  holds under the familiar causal identification assumptions discussed earlier. Despite the non-stationary and non-i.i.d. structure, it is nevertheless possible to derive AsympCSs for the *running average* of individual treatment effects  $\tilde{\psi}_t := \frac{1}{t} \sum_{i=1}^t \psi_i$  – or simply, the running average treatment effect – using the Lyapunov-type bounds of Corollary 7.2.1. However, given this more general and complex setup, the assumptions required are more subtle (but no more restrictive) than those for Theorems 7.3.1 and 7.3.2; as such, we explicitly describe their details here but handle the randomized and observational settings simultaneously for brevity.

**Condition  $\widetilde{\text{ATE-1}}$**  (Regression estimator is uniformly well-behaved in  $L_2(\mathbb{P})$ ). We assume that regression estimators  $\mu_t^a(X_i)$  converge in  $L_2(\mathbb{P})$  to any function  $\bar{\mu}^a(X_i)$  uniformly for  $i \in \{1, 2, \dots\}$  i.e.

$$\sup_{1 \leq i < \infty} \|\hat{\mu}_t^a(X_i) - \bar{\mu}^a(X_i)\|_{L_2(\mathbb{P})} = o(1) \quad (7.43)$$

for each  $a \in \{0, 1\}$ .

Condition  $\widetilde{\text{ATE-1}}$  simply requires that the regression estimator  $\hat{\mu}_t^a$  must converge to some function  $\bar{\mu}^a$ , which need not coincide with true regression function  $\mu^a$ . In the i.i.d. setting where  $X_1, X_2, \dots$  all have the same distribution, we would simply drop the  $\sup_{1 \leq i < \infty}$ , recovering the conditions for Theorems 7.3.1 and 7.3.2.

**Condition  $\widetilde{\text{ATE-2}}$**  (Convergence of average nuisance errors). Let  $\hat{\mu}_t^a$  be an estimator of the regression function  $\mu^a$ ,  $a \in \{0, 1\}$  and  $\hat{\pi}_t$  an estimator of the propensity score  $\pi$ . We assume that the average bias shrinks at a  $\sqrt{\log t/t}$  rate, i.e.

$$\frac{1}{t} \sum_{i=1}^t \left\{ \|\hat{\pi}_t(X_i) - \pi(X_i)\|_{L_2(\mathbb{P})} \sum_{a=0}^1 \|\hat{\mu}_t^a(X_i) - \mu^a(X_i)\|_{L_2(\mathbb{P})} \right\} = o\left(\sqrt{\frac{\log t}{t}}\right). \quad (7.44)$$

Note that Condition  $\widetilde{\text{ATE-2}}$  would hold in two familiar scenarios. Firstly, in a randomized experiment (Theorem 7.3.3) where  $\hat{\pi}_t = \pi$  is known by design, we have that (7.44) is always zero, satisfying Condition  $\widetilde{\text{ATE-2}}$  trivially. Second, in an observational study where the product

of errors  $\|\widehat{\pi}_t(X_i) - \pi(X_i)\|_{L_2(\mathbb{P})} \|\widehat{\mu}_t^a(X_i) - \mu^a(X_i)\|_{L_2(\mathbb{P})}$  vanishes at a rate faster than  $\sqrt{\log t/t}$ , for each  $i$  and for both  $a \in \{0, 1\}$ , we also have that their average product errors vanish at the same rate (7.44). With these assumptions in mind, let us summarize how running average treatment effects can be captured in randomized experiments.

**Theorem 7.3.3** (AsympCSs for the running average treatment effect). *Suppose  $Z_1, Z_2, \dots$  are independent triples  $Z_t := (X_t, A_t, Y_t)$  and that Conditions  $\widetilde{\text{ATE-1}}$  and  $\widetilde{\text{ATE-2}}$  hold. Finally, suppose that the conditions of Corollary 7.2.1 hold, but with  $(Y_t)_{t=1}^\infty$  replaced by the influence functions  $(\bar{f}(Z_t))_{t=1}^\infty$ . Then,*

$$\widehat{\psi}_t^\times \pm \sqrt{\frac{2(t\rho^2 \widehat{\text{Var}}_t(\bar{f}) + 1)}{t^2 \rho^2} \log \left( \frac{\sqrt{t\rho^2 \widehat{\text{Var}}_t(\bar{f}) + 1}}{\alpha} \right)} \quad (7.45)$$

forms a  $(1 - \alpha)$ -AsympCS for the running average treatment effect  $\tilde{\psi}_t := \frac{1}{t} \sum_{i=1}^t \psi_i$ .

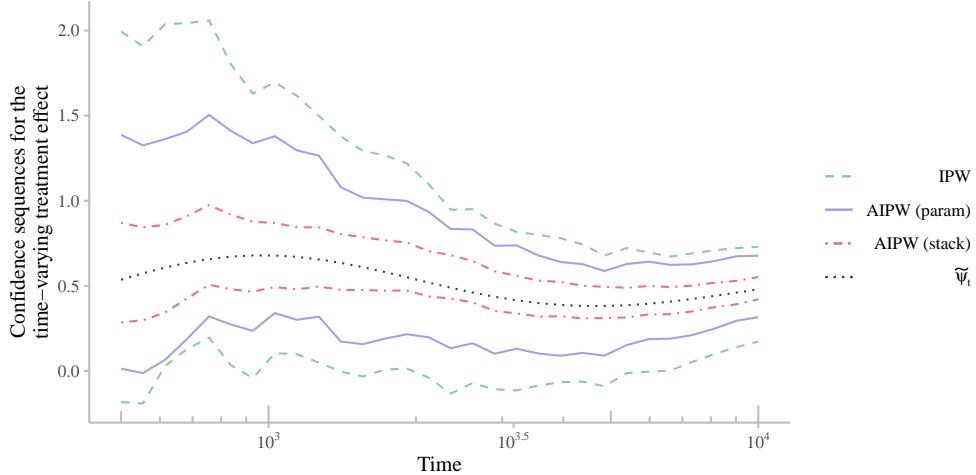


Figure 7.6: Three 90% AsympCSs for  $\tilde{\psi}_t$  constructed using various estimators via Theorem 7.3.3. Since this is a randomized experiment, all three CSs capture  $\tilde{\psi}_t$  uniformly over time with high probability. Similar to Figure 7.4, however, the stacking-based AIPW estimator greatly outperforms those based on parametric models or IPW.

The proof can be found in Section 7.A.6, and it is not hard to see that both Theorems 7.3.1 and 7.3.2 are particular instantiations of Theorem 7.3.3. The important takeaway from Theorem 7.3.3 is that under some rather mild conditions on the moments of  $(\bar{f}(Z_t))_{t=1}^\infty$ , it is possible to derive an AsympCS for a running average treatment effect  $\tilde{\psi}_t$  (see Figure 7.6 for what these look like in practice). Nevertheless, under the commonly considered regime where the treatment effect is constant  $\psi_1 = \psi_2 = \dots = \psi$ , we have that (7.45) forms a  $(1 - \alpha)$ -AsympCS for  $\psi$ . Note that unlike Theorems 7.3.1 and 7.3.2, Theorem 7.3.3 actually does require the use of the cross-fit AIPW estimator  $\widehat{\psi}_t^\times$  and would not capture  $\tilde{\psi}_t$  if the sample-split version were

used in its place.

*Remark 20* (Avoiding sample splitting via martingale AsympCSs). The reader may wonder whether it is possible to simply plug in a *predictable* estimate of  $\hat{\mu}_t^a$  in a randomized experiment – i.e. so that  $\hat{\mu}_t^a$  only depends on  $Z_1^{t-1}$  – and employ the Lindeberg-type martingale AsympCS of Proposition 7.2.2 in place of Corollary 7.2.1, thereby sidestepping the need for sequential sample splitting and cross fitting altogether. Indeed, such an analogue of Theorem 7.3.3 is possible to derive, but the conditions required are less transparent than those we have provided above so we defer it to Section 7.B.6.

### 7.3.5 Extensions to general semiparametric estimation and the delta method

The discussion thus far has been focused on deriving confidence sequences for the ATE in the context of causal inference. However, the tools presented in this chapter are more generally applicable to any pathwise differentiable functional with positive (and finite) semiparametric information bound. Some prominent examples in causal inference include modified interventions, complier-average effects, time-varying effects, and controlled mediation effects, among others. Examples outside causal inference include the expected density, entropy, the expected conditional variance, and the expected conditional covariance, to list a few.

All of the aforementioned problems, including estimation of the ATE in Section 7.3 can be written in the following general form. Suppose  $Z_1, Z_2, \dots \stackrel{\text{iid}}{\sim} \mathbb{Q}$  and let  $\theta(\mathbb{Q})$  be some functional (such as those listed above) of the distribution  $\mathbb{Q}$ . In the case of a finite sample size  $n$ ,  $\hat{\theta}_n$  is said to be an asymptotically linear estimator [254] for  $\theta$  if

$$\hat{\theta}_n - \theta = \frac{1}{n} \sum_{i=1}^n \phi(Z_i) + o_{\mathbb{Q}}\left(\frac{1}{\sqrt{n}}\right), \quad (7.46)$$

where  $\phi$  is the influence function of  $\hat{\theta}_n$ . When the sample size is not fixed in advance, we may analogously say that  $\hat{\theta}_t$  is an *asymptotically linear time-uniform estimator* if instead,

$$\hat{\theta}_t - \theta = \frac{1}{t} \sum_{i=1}^t \phi(Z_i) + o\left(\sqrt{\log t/t}\right), \quad (7.47)$$

with  $\phi$  being the same influence function as before. For example, in the case of the ATE with  $(Z_t)_{t=1}^\infty \stackrel{\text{iid}}{\sim} \mathbb{P}$ , we presented an efficient estimator  $\hat{\psi}_t$  which took the form,

$$\hat{\psi}_t - \psi = \frac{1}{t} \sum_{i=1}^t (f(Z_i) - \psi) + o\left(\sqrt{\log t/t}\right), \quad (7.48)$$

where  $f$  is the uncentered efficient influence function (EIF) defined in (7.36). In order to justify that the remainder term is indeed  $o(\sqrt{\log t/t})$ , we used sequential sample splitting and additional analysis in the randomized and observational settings (see the proofs in Sections 7.A.5 and 7.A.5.2 for more details). In general, as long as an estimator  $\hat{\theta}_t$  for  $\theta$  has the form (7.47), we

may derive AsympCSs for  $\theta$  as a simple corollary of Theorem 7.2.2.

**Corollary 7.3.1** (AsympCSs for general functional estimation). *Suppose  $\hat{\theta}_t$  is an asymptotically linear time-uniform estimator of  $\theta$  with influence function  $\phi$ , that is, satisfying (7.47). Additionally, suppose that  $\text{Var}(\phi) < \infty$ . Then,*

$$\hat{\theta}_t \pm \sqrt{\text{Var}(\phi)} \cdot \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)} \quad (7.49)$$

forms a  $(1 - \alpha)$ -AsympCS for  $\theta$ .

Clearly, the boundaries of Sections 7.2.2.4 or 7.2.6 could be used here in place of Theorem 7.2.2, though for the boundary in Proposition 7.2.1, we would need to strengthen the  $o(\sqrt{\log t/t})$  rate in (7.47) to  $o(\sqrt{\log \log t/t})$ . If computing  $\hat{\theta}_t$  additionally involves the estimation of a nuisance parameter  $\eta$  such as in Theorems 7.3.1 and 7.3.2, this must be handled carefully on a case-by-case basis where sequential sample splitting and cross fitting (Section 7.3.1) may be helpful, and higher moments on  $\phi(Z_i)$  may be needed. We now derive an analogue of the delta method for asymptotically linear time-uniform estimators.

**Proposition 7.3.1** (The delta method for AsympCSs). *Consider the same setup as Corollary 7.3.1 and let  $g : \mathbb{R} \rightarrow \mathbb{R}$  be a continuously differentiable function with first derivative  $g'$ . Then,  $g(\hat{\theta}_t)$  is an asymptotically linear time-uniform estimator for  $g(\theta)$  with influence function given by  $g'(\theta)\phi(\cdot)$ , i.e.*

$$g(\hat{\theta}_t) - g(\theta) = \frac{1}{t} \sum_{i=1}^t g'(\theta)\phi(Z_i) + o\left(\sqrt{\log t/t}\right). \quad (7.50)$$

The short proof in Section 7.A.8 is similar to the proof of the classical delta method but with the almost-sure continuous mapping theorem used in place of the in-probability one, and with the law of the iterated logarithm used in place of the central limit theorem.

## 7.4 Simulation studies: Widths and empirical coverage

We now perform simulations focusing on the setting where  $(Y_t)_{t=1}^\infty$  are bounded random variables (with known bounds) and the parameters of interest may include means or treatment effects. We consider this setting as it is well-studied, allowing us to draw on a rich literature containing several nonasymptotic CSs to which we will compare AsympCSs (though it is important to keep in mind that there are many unbounded problems for which nonasymptotic CSs do not exist, and we discuss these at the end of the section). In particular, we will compare the AsympCSs of Theorems 7.2.2 and 7.3.1 to Robbins' sub-Gaussian mixture CS [217] (see also [125, §3.2]), the empirical Bernstein CSs of Howard et al. [125, Thm 4 and Cor 2], and CLT-based CIs. Of course, CLT-based CIs are not time-uniform and are only included for reference. Figure 7.7 considers three cases of parameter estimation for bounded random variables:

- (a) The first simulation focuses on estimating the mean  $\mu$  of i.i.d. Uniform(0, 1) random

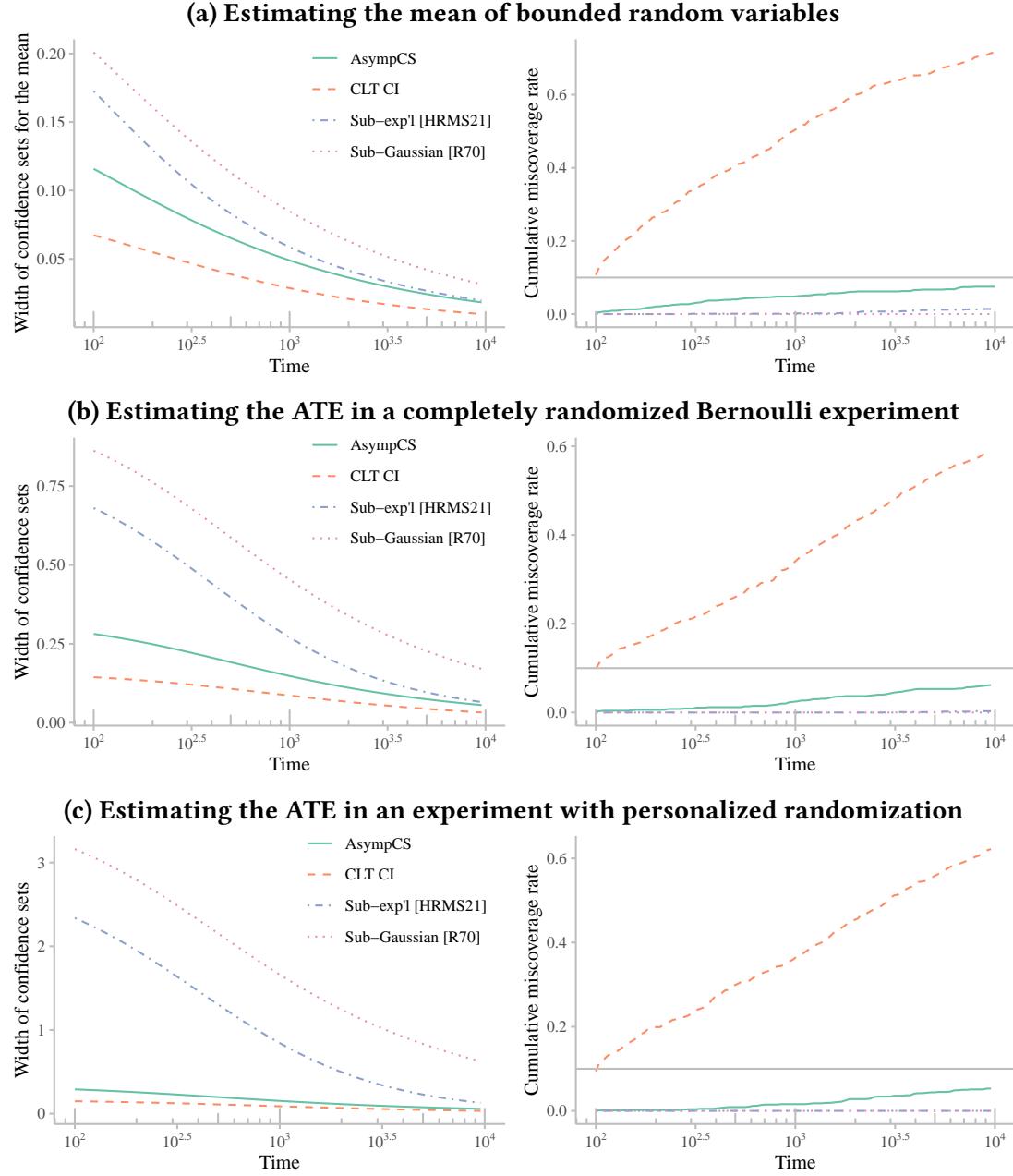


Figure 7.7: A comparison of  $(1 - \alpha) \equiv 90\%$  confidence sets for parameters in three simulation scenarios: (a) mean estimation of bounded random variables, (b) ATE estimation from a completely randomized Bernoulli experiment, and (c) ATE estimation from an experiment with covariate-dependent (“personalized”) randomization. Empirical widths and miscoverage rates were computed with 1000 replications beginning at time 100 for mean estimation (Simulation (a)) and time 500 for ATE estimation (Simulations (b) and (c)). Notice that in all three scenarios, only CSs have miscoverage rates below  $\alpha$ , but AsympCSs are the only ones that appear to sharply approach this level. The tuning parameter  $\bar{\rho}_m$  was chosen for a start time of  $m$  as  $\bar{\rho}_m := \rho(m \log(m \vee e))$  following (7.28).

variables only using knowledge of  $[0, 1]$ -boundedness. For this setting, boundedness allows the CSs of Robbins [217] and Howard et al. [125] to be immediately applicable.

- (b) The second considers average treatment effect estimation in a randomized experiment with  $\{0, 1\}$ -valued outcomes where everyone is randomly assigned to treatment or control with probability  $1/2$ . Since all propensity scores are equal to  $1/2$ , we have that estimates of the influence functions in (7.41) from Section 7.3.2 are bounded in  $[-2, 2]$ , and hence the techniques of Robbins [217] and Howard et al. [125] are applicable (as suggested by Howard et al. [125, §4.2]).
- (c) The third and final simulation considers a similar setup to (b) with the key difference that propensity scores are now covariate-dependent (“personalized”). In this case, as suggested by Howard et al. [125, §4.2], note that estimates of the influence functions (7.41) lie in  $[-1/p_{\min} - 1, 1/p_{\min} + 1]$  where  $p_{\min} := \min\{\text{ess inf}_x \pi(x), \text{ess inf}_x [1 - \pi(x)]\}$ , permitting the application of Robbins [217] and Howard et al. [125] as before.<sup>8</sup>

Notice that in all three scenarios, CLT-based CIs have cumulative miscoverage rates that quickly diverge beyond  $\alpha = 0.1$  while those of CSs – both asymptotic and nonasymptotic – never exceed  $\alpha$  before time  $10^4$ . (Note that a longer time horizon of  $10^5$  is considered only for AsympCSs and CLT-based CIs in Figure 7.1, but is omitted here due to the computational expense of Howard et al. [125, Thm 2] at large  $t$ .) Moreover, notice that nonasymptotic CSs appear to be conservative, while our AsympCSs are much tighter and have miscoverage rates approaching  $\alpha$  (as expected in light of Theorem 7.2.5). Of course, Theorem 7.2.5 states that miscoverage will only be close to  $\alpha$  for large values of the first peeking time  $m$ . In Figure 7.8, we illustrate this phenomenon by using the bound in (7.28) to estimate the means of two distributions (uniform and Student  $t$ -distributed in the left-hand and right-hand side plots, respectively) given several initial peeking times  $m \in \{2, 5, 10, 50, 100\}$ .

These particular simulation scenarios illustrate situations where AsympCSs are increasingly beneficial over nonasymptotic bounds. First, estimating means of bounded random variables (Figure 7.7(a)) is a problem for which several nonasymptotic CIs and CSs exist, and indeed both the bounds of Robbins [217] and especially Howard et al. [125] fare well in this setting. The relatively small variance of Uniform( $0, 1$ ) random variables means that asymptotic methods can quickly tighten in this setting while the empirical Bernstein CSs of Howard et al. [125] take a while to adapt to the variance (while those of Robbins [217] never will).

AsympCSs are particularly well-suited to the settings of ATE estimation in Figs (b) and (c) where almost-sure bounds on observed random variables can be quite large in comparison to their variances. Figure 7.7(b) considers the setting where all propensity scores are equal to  $1/2$  which is the “easiest” regime for nonasymptotic methods, meaning that a.s. bounds on the influence functions are as tight as possible. Even here, AsympCSs are tighter than the nonasymptotic bounds. On the other hand, Figure 7.7(c) considers a setting where propensity scores are highly covariate-dependent so that bounds on propensity scores  $\pi(x) \in [p_{\min}, 1 - p_{\max}]$

---

<sup>8</sup>Note that Howard et al. [125, §4.2] use a more conservative bound of  $[-2/p_{\min}, 2/p_{\min}]$  but it is not hard to see that this can be improved to  $[-1/p_{\min} - 1, 1/p_{\min} + 1]$ . Consequently, we are ultimately comparing our AsympCSs to a *strictly tighter* nonasymptotic CS than the one proposed by Howard et al. [125, §4.2].

$p_{\min}$ ] must hold for almost all  $x$  (in this simulation,  $p_{\min} = 0.2$ ). Here we see that AsympCSs drastically outperform nonasymptotic CSs without inflating miscoverage rates. Taking this to a logical extreme, it is possible to construct scenarios where  $p_{\min}$  is closer and closer to 0, but influence function variances remain bounded. In other words, it is possible to consider scenarios where AsympCSs are arbitrarily tighter than nonasymptotic CSs, without inflating miscoverage rates.

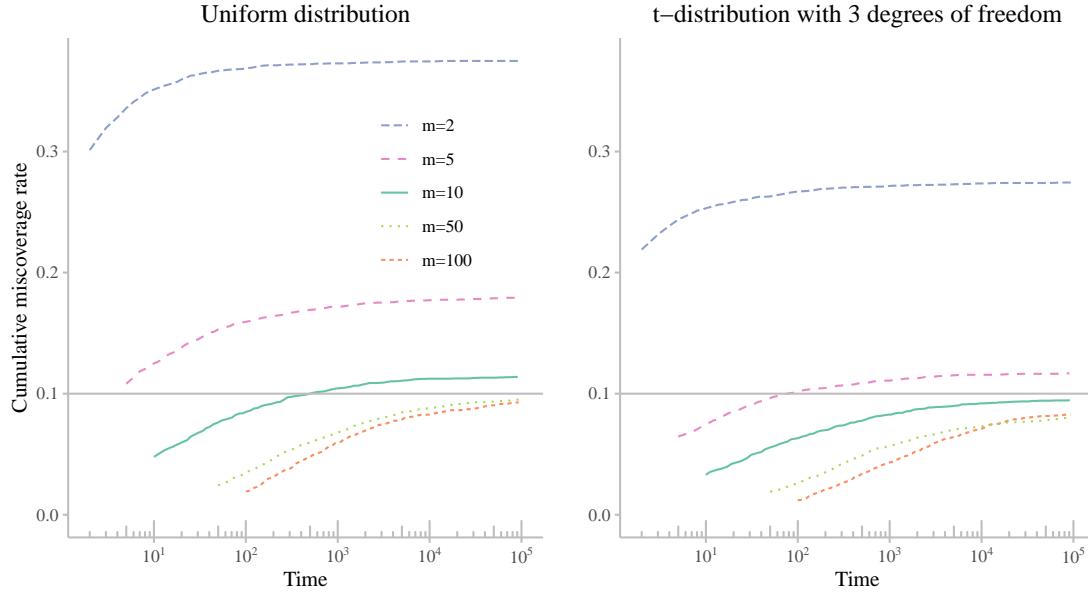


Figure 7.8: Cumulative miscoverage rates using (7.28) at level  $\alpha = 0.1$  to estimate the mean of i.i.d.  $\text{Uniform}[0, 1]$  and  $t$ -distributed random variables in the left-hand and right-hand side plots, respectively. Notice that in both cases including the heavy-tailed setting of a  $t$ -distribution with 3 degrees of freedom (so that the variance is finite but third and higher absolute moments are all infinite), cumulative miscoverage rates do not exceed  $\alpha = 0.1$  even after  $10^5$  observations as long as the first peeking time  $m$  is at least 50. It is worth remarking that asymptotic approximations appear to “kick in” earlier for the heavy-tailed  $t$ -distribution.

Finally, we remark that while AsympCSs demonstrate substantial benefits over nonasymptotic CSs in terms of *tightness*, we wish to also highlight their benefits of *versatility*, and in particular, note that there are many settings for which no simulations could be run since AsympCSs provide the first (asymptotically) time-uniform solution in the literature. For example, as a consequence of Bahadur and Savage [14], it is impossible to derive nonasymptotic CSs (or CIs) for the mean of random variables without prior bounds on their moments. By contrast, AsympCSs can handle mean estimation under finite (but unknown) moment assumptions. It is also impossible to derive nonasymptotic CS (and CIs) for the ATE from observational studies (without substantial and unrealistic knowledge of nuisance function estimation errors) but Section 7.3.3 outlines an asymptotically time-uniform solution. In both settings, we do not run simulations akin to Figure 7.7 since there are no prior CSs to compare to.

## 7.5 Real data application: effects of IV fluid caps in sepsis patients

Let us now illustrate the use of Theorem 7.3.2 by sequentially estimating the effect of fluid-restrictive strategies on mortality in an observational study of real sepsis patients. We will use data from the Medical Information Mart for Intensive Care III (MIMIC-III), a freely available database consisting of health records associated with more than 45,000 critical care patients at the Beth Israel Deaconess Medical Center [137, 205]. The data are rich, containing demographics, vital signs, medications, and mortality, among other information collected over the span of 11 years.

Following Shahn et al. [239], we aim to estimate the effect of restricting intravenous (IV) fluids within 24 hours of intensive care unit (ICU) admission on 30-day mortality in sepsis patients. In particular, we considered patients at least 16 years of age satisfying the Sepsis-3 definition – i.e. those with a suspected infection and a Sequential Organ Failure Assessment (SOFA) score of at least 2 [242]. Sepsis-3 patients can be obtained from MIMIC-III using SQL scripts provided by Johnson and Pollard [136], but we provide detailed instructions for reproducing our data collection and analysis process on GitHub.<sup>9</sup> This resulted in a total of 5231 sepsis patients, each of whom received out-of-hospital followup of at least 90 days.

Consider IV fluid intake within 24 hours of ICU admission  $\mathcal{L}^{24h}$ . To construct a binary treatment  $A \in \{0, 1\}$ , we dichotomize  $\mathcal{L}^{24h}$  so that  $A_i = \mathbb{1}(\mathcal{L}_i^{24h} \leq 6L)$ . 30-day mortality  $Y$  is defined as 1 if the patient died within 30 days of hospital admission, and 0 otherwise. We will consider baseline covariates  $X$  including a patient's age and sex, whether they are diabetic, modified Elixhauser scores [263], and SOFA scores. We are interested in the causal estimand

$$\psi := \mathbb{P}(Y^{\mathcal{L}^{24h} \leq 6L} = 1) - \mathbb{P}(Y^{\mathcal{L}^{24h} > 6L} = 1), \quad (7.51)$$

i.e. the difference in average 30-day mortality that would be observed if all sepsis patients were randomly assigned an IV fluid level according to the lower truncated distribution  $\mathbb{P}(\mathcal{L}^{24h} \leq l | \mathcal{L}^{24h} \leq 6L)$  versus the upper truncated distribution  $\mathbb{P}(\mathcal{L}^{24h} \leq l | \mathcal{L}^{24h} > 6L)$  [83]. While this is technically a stochastic intervention effect, we have that under the same causal identification assumptions discussed in Section 7.3,  $\psi$  is identified as

$$\psi = \mathbb{E}\{\mathbb{E}(Y | X, A = 1) - \mathbb{E}(Y | X, A = 0)\}, \quad (7.52)$$

the same functional considered in the previous sections. Therefore, we can estimate  $\psi$  under the same assumptions and with the same techniques as Section 7.3.3. Figure 7.9 contains AsympCSs for  $\psi$  using difference-in-means, parametric AIPW, and stacking-based AIPW estimators to demonstrate the impacts of different modeling choices on AsympCS width.

*Remark 21.* These simple binary treatment and outcome variables were used for simplicity so that the methods outlined in Section 7.3.3 are immediately applicable, but Section 7.3.5 points out that our AsympCSs may be used to sequentially estimate other causal functionals.

---

<sup>9</sup> [github.com/WannabeSmith/drconfseq/tree/main/paper\\_plots/sepsis](https://github.com/WannabeSmith/drconfseq/tree/main/paper_plots/sepsis)

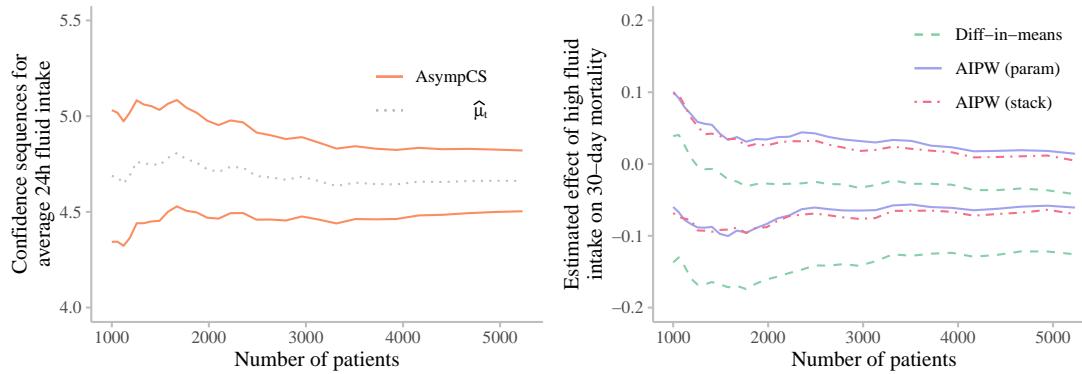


Figure 7.9: Left-hand side: a 90% AsympCS used to track the average 24h fluid intake over time. Right-hand side: Three 90%-AsympCSs for the causal effect of capped IV fluid intake (defined as  $\leq 6$  litres) on 30-day mortality using the same three estimators as those outlined in Figure 7.5. Notice that an analysis using a difference-in-means estimator would conclude that the treatment effect is negative after observing fewer than 1500 patients.

The stacking-based AIPW AsympCSs cover the null treatment effect of 0 from the 1000<sup>th</sup> to the 5231<sup>st</sup> observed patient, and thus we cannot conclude whether 6L IV fluid caps have an effect on 30-day mortality in sepsis patients. Interestingly, the width and center of AsympCSs based on both stacking and parametric estimators are roughly equal, which is in contrast to the simulations provided in the previous sections. This could be due to the fact that the true regression or propensity score functions lie in the parametric models considered by “AIPW (param)” or because they are good approximations. Since this analysis is not a simulation, we cannot know with certainty one way or another.

Note that these stacking-based AsympCSs nearly drop below 0 after observing the 5231<sup>st</sup> patient’s outcome. If we were using fixed-time confidence intervals, the analyst would need to resist the temptation to resume data collection (e.g. to see whether the null hypothesis  $H_0 : \psi = 0$  could be rejected with a slightly larger sample size) as this would inflate type-I error rates (as seen in Figure 7.1). On the other hand, AsympCSs flexibly permit precisely this form of continued sampling.

## 7.6 Conclusion

This chapter introduced the notion of an “asymptotic confidence sequence” as the time-uniform analogue of an asymptotic confidence interval based on the central limit theorem. We derived an explicit universal asymptotic confidence sequence for the mean from i.i.d. observations under weak moment assumptions by appealing to the strong invariance principle of Strassen [248]. These results were extended to the setting where observations’ distributions (including means and variances) can vary over time under martingale dependence, such that our confidence sequences capture a moving parameter — the running average of the conditional means so far. We then applied the aforementioned results to the problem of doubly robust sequential inference

for the average treatment effect in both randomized experiments and observational studies under i.i.d. sampling. Finally, we showed how these causal applications remain valid in the non-i.i.d. setting where distributions change over time, in which case our confidence sequences capture a running average of individual treatment effects. The aforementioned results will enable researchers to continuously monitor sequential experiments – such as clinical trials and online A/B tests – as well as sequential observational studies even if treatment effects do not remain stationary over time.

## 7.A Proofs of the main results

### 7.A.1 Proof of Theorem 7.2.2

We first introduce a lemma that will later be used in the main proof.

**Lemma 7.A.1** (Strong Gaussian approximation of the sample average). *Let  $(Y_t)_{t=1}^{\infty}$  be an i.i.d. sequence of random variables with mean  $\mu$ , variance  $\sigma^2$ , and let  $\hat{\sigma}_t^2$  be the sample variance. Then (after sufficiently enriching the probability space), there exist i.i.d. Gaussian random variables  $(G_t)_{t=1}^{\infty}$  such that*

$$\frac{1}{t} \sum_{i=1}^t (Y_i - \mu) = \frac{\hat{\sigma}_t}{t} \sum_{i=1}^t G_i + \varepsilon_t$$

where  $\varepsilon_t = o(\sqrt{\log \log t / t})$ . One could also replace  $\sum_{i=1}^t G_i$  with a Wiener process  $(W_t)_{t \geq 0}$ .

*Proof.* By the strong approximation theorem of Strassen [248], we have that (after sufficiently enriching the probability space) there exist i.i.d. Gaussian random variables  $(G_t)_{t=1}^{\infty}$  such that

$$\frac{1}{t} \sum_{i=1}^t (Y_i - \mu) = \frac{\sigma}{t} \sum_{i=1}^t G_i + \kappa_t \tag{7.53}$$

with  $\kappa_t = o(\sqrt{\log \log t / t})$ . Since  $(G_t)_{t=1}^{\infty}$  have mean zero and unit variance, we have by the law of the iterated logarithm (LIL) that

$$\frac{1}{t} \sum_{i=1}^t G_i = O\left(\sqrt{\frac{\log \log t}{t}}\right). \tag{7.54}$$

Now, by the strong law of large numbers,  $\hat{\sigma}_t \xrightarrow{a.s.} \sigma$ . Combining this fact with with (7.53), we have

$$\frac{1}{t} \sum_{i=1}^t (Y_i - \mu) = \frac{\hat{\sigma}_t + o(1)}{t} \sum_{i=1}^t G_i + \kappa_t \tag{7.55}$$

$$= \frac{\hat{\sigma}_t}{t} \sum_{i=1}^t G_i + \kappa_t + o\left(\sqrt{\frac{\log \log t}{t}}\right) \tag{7.56}$$

$$= \frac{\hat{\sigma}_t}{t} \sum_{i=1}^t G_i + o\left(\sqrt{\frac{\log \log t}{t}}\right) \quad (7.57)$$

almost surely, which completes the proof.  $\square$

*Proof of Theorem 7.2.2.* The proof proceeds in 3 steps. First, we use the fact that for any martingale  $M_t(\lambda)$ , we have that the mixture  $\int_{\mathbb{R}} M_t(\lambda) dF(\lambda)$  is also a martingale where  $F$  is any probability distribution on  $\mathbb{R}$  [124, 125]. We apply this fact to an exponential Gaussian martingale and use a Gaussian density  $f(\lambda; 0, \rho^2)$  as the mixing distribution. Second, we apply Ville's inequality [264] to this mixture exponential Gaussian martingale to obtain Robbins' normal mixture confidence sequence [217]. Third, we use Lemma 7.A.1 to approximate  $\sum_{i=1}^t (Y_i - \mu)$  by a cumulative sum of Gaussian random variables and apply the results from steps 1 and 2.

**Step 1** Let  $(W_t)_{t \geq 0}$  be a Wiener process and write the exponential process for any  $\lambda \in \mathbb{R}$ ,

$$M_t(\lambda) := \exp \left\{ \lambda W_t - t\lambda^2/2 \right\}.$$

It is well-known that  $M_t(\lambda)$  is a *nonnegative martingale starting at  $M_0 \equiv 1$*  with respect to the Brownian filtration. By Fubini's theorem, for any probability distribution  $F(\lambda)$  on  $\mathbb{R}$ , we also have that the mixture,

$$\int_{\lambda \in \mathbb{R}} M_t(\lambda) dF(\lambda)$$

is a nonnegative martingale with initial value one with respect to the same filtration [217]. In particular, consider the Gaussian probability distribution function  $f(\lambda; 0, \rho^2)$  with mean zero and variance  $\rho^2 > 0$  as the mixing distribution. The resulting martingale can be written as

$$M_t := \int_{\lambda \in \mathbb{R}} \exp \left\{ \lambda W_t - \frac{t\lambda^2}{2} \right\} f(\lambda; 0, \rho^2) d\lambda \quad (7.58)$$

$$= \frac{1}{\sqrt{2\pi\rho^2}} \int_{\lambda} \exp \left\{ \lambda W_t - \frac{t\lambda^2}{2} \right\} \exp \left\{ \frac{-\lambda^2}{2\rho^2} \right\} d\lambda \quad (7.59)$$

$$= \frac{1}{\sqrt{2\pi\rho^2}} \int_{\lambda} \exp \left\{ \lambda W_t - \frac{\lambda^2(t\rho^2 + 1)}{2\rho^2} \right\} d\lambda \quad (7.60)$$

$$= \frac{1}{\sqrt{2\pi\rho^2}} \int_{\lambda} \exp \left\{ \frac{-\lambda^2(t\rho^2 + 1) + 2\lambda\rho^2 W_t}{2\rho^2} \right\} d\lambda \quad (7.61)$$

$$= \frac{1}{\sqrt{2\pi\rho^2}} \int_{\lambda} \exp \left\{ \frac{-a(\lambda^2 - \frac{b}{a}2\lambda)}{2\rho^2} \right\} d\lambda \quad (7.62)$$

by setting  $a := t\rho^2 + 1$  and  $b := \rho^2 W_t$ . Focusing on the integrand and completing the square, we have

$$\exp \left\{ \frac{-\lambda^2 + 2\lambda \frac{b}{a} + (\frac{b}{a})^2 - (\frac{b}{a})^2}{2\rho^2/a} \right\} = \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} + \frac{a(b/a)^2}{2\rho^2} \right\} \quad (7.63)$$

$$= \underbrace{\exp\left\{\frac{-(\lambda - b/a)^2}{2\rho^2/a}\right\}}_{\propto f(\lambda, b/a, \rho^2/a)} \exp\left\{\frac{b^2}{2a\rho^2}\right\}. \quad (7.64)$$

Plugging this back into the integral and multiplying the entire quantity by  $\frac{a^{-1/2}}{a^{-1/2}}$ , we finally get the closed-form expression of the mixture exponential Wiener process,

$$\begin{aligned} M_t &:= \underbrace{\frac{1}{\sqrt{2\pi\rho^2/a}} \int_{\lambda \in \mathbb{R}} \exp\left\{\frac{-(\lambda - b/a)^2}{2\rho^2/a}\right\} d\lambda}_{=1} \frac{\exp\left\{\frac{b^2}{2a\rho^2}\right\}}{\sqrt{a}} \\ &= \frac{\exp\left\{\frac{\rho^2 W_t^2}{2(t\rho^2+1)}\right\}}{\sqrt{t\rho^2 + 1}}. \end{aligned} \quad (7.65)$$

**Step 2** Since  $M_t$  is a nonnegative martingale with initial value one, we have by Ville's inequality [264] that

$$\mathbb{P}(\forall t \geq 1, M_t < 1/\alpha) \geq 1 - \alpha.$$

Writing this out explicitly for  $M_t$  and solving for  $W_t$  algebraically, we have that

$$\mathbb{P}\left(\forall t \geq 1, \frac{\rho^2 W_t^2}{t\rho^2 + 1} < \log(1/\alpha^2) + \log(t\rho^2 + 1)\right) \quad (7.66)$$

$$= \mathbb{P}\left(\forall t \geq 1, |W_t/t| < \sqrt{\underbrace{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)}_{\mathfrak{B}_t^G}}\right) \geq 1 - \alpha. \quad (7.67)$$

**Step 3** First, note that by the triangle inequality,

$$\left|\frac{1}{t} \sum_{i=1}^t Y_i - \mu\right| \leq \left|\frac{1}{t} \sum_{i=1}^t (Y_i - \mu) - \frac{\hat{\sigma}_t}{t} W_t\right| + \frac{\hat{\sigma}_t}{t} |W_t|,$$

and thus by Lemma 7.A.1 and Step 2, we have with probability at least  $(1 - \alpha)$ ,

$$\forall t \geq 1, \left|\frac{1}{t} \sum_{i=1}^t Y_i - \mu\right| < \underbrace{\hat{\sigma}_t \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)}}_{\mathfrak{B}_t^G} + |\varepsilon_t| \quad (7.68)$$

where  $\varepsilon_t$  is defined as in Lemma 7.A.1. Finally, note that

$$\frac{\mathfrak{B}_t^G + |\varepsilon_t|}{\bar{\mathfrak{B}}_t^G} \xrightarrow{\text{a.s.}} 1, \quad (7.69)$$

completing the proof.  $\square$

### 7.A.2 Proof of Theorem 7.2.3

*Proof.* Using Conditions G-1 and G-2, take  $L_t^* := \hat{L}_t + (\hat{\theta}_t - \theta_t + Z_t)$  and  $U_t^* := \hat{U}_t - (\hat{\theta}_t - \theta_t + Z_t)$ . Then we have that

$$\mathbb{P} \left( \forall t \in \mathcal{T}, \theta_t \in [\hat{\theta}_t - L_t^*, \hat{\theta}_t + U_t^*] \right) \quad (7.70)$$

$$= \mathbb{P} \left( \forall t \in \mathcal{T}, \theta_t \in [\hat{\theta}_t - \hat{L}_t - (\hat{\theta}_t - \theta_t + Z_t), \hat{\theta}_t + \hat{U}_t - (\hat{\theta}_t - \theta_t + Z_t)] \right) \quad (7.71)$$

$$= \mathbb{P} \left( \forall t \in \mathcal{T}, Z_t \in [-\hat{L}_t, \hat{U}_t] \right) \geq 1 - \alpha. \quad (7.72)$$

It remains to show that

$$L_t^*/L_t \xrightarrow{\text{a.s.}} 1 \quad \text{and} \quad U_t^*/U_t \xrightarrow{\text{a.s.}} 1. \quad (7.73)$$

Indeed, we have that

$$L_t^*/L_t = (L_t^*/\hat{L}_t) \cdot (\hat{L}_t/L_t) \quad (7.74)$$

$$= \left( \frac{\hat{L}_t + (\hat{\theta}_t - \theta_t + Z_t)}{\hat{L}_t} \right) \cdot (\hat{L}_t/L_t) \quad (7.75)$$

$$= (1 + o(1)) \cdot (1 + o(1)) \quad (7.76)$$

$$= 1 + o(1) \quad (7.77)$$

almost surely, where the second-last line follows from the combination of Conditions G-3 and G-4. A similar calculation goes through for  $U_t^*/U_t$ , completing the proof.  $\square$

### 7.A.3 Proof of Proposition 7.2.2

First, we present a lemma that is implicit in the proof of Strassen [249, Theorem 4.4] and which will be central to the proof of Proposition 7.2.2.

**Lemma 7.A.2** (Strong approximation under martingale dependence). *Let  $(Y_t)_{t=1}^\infty$  be a sequence of random variables with conditional means and variances given by  $\mu_t = \mathbb{E}(Y_t | Y_1^{t-1})$  and  $\sigma_t^2 = \text{Var}(Y_t | Y_1^{t-1})$ , respectively. Let  $V_t = \sum_{i=1}^t \sigma_i^2$  be the cumulative conditional variance process, and suppose that  $V_t \rightarrow \infty$  almost surely. Furthermore, assume that the Lindeberg-type condition given in Condition L-2 holds. Then, on a potentially enriched probability space, there exist i.i.d. standard Gaussians  $(G_t)_{t=1}^\infty$  such that*

$$\frac{1}{t} \sum_{i=1}^t (Y_i - \mu_i) - \frac{1}{t} \sum_{i=1}^t \sigma_i G_i = o \left( \frac{V_t^{3/8} \log V_t}{t} \right). \quad (7.78)$$

*Proof of Lemma 7.A.2.* The proof centrally relies on Strassen [249, Eq. 159] which states that on a potentially enriched probability space,

$$\sum_{i=1}^t (Y_i - \mu_i) = \xi(V_t) + o(h(V_t)) \quad (7.79)$$

where  $\xi$  is a standard Brownian motion, and  $h(v) = (vf(v))^{1/4} \log v$  is any function so that  $f(v)$  is increasing in  $v$ , but  $f(v)/v$  is decreasing. For the purposes of this proof, we set  $f(v) = v^{1/2}$ , and hence  $h(v) = v^{3/8} \log v$ . Since a standard Brownian motion evaluated at  $V_t = \sum_{i=1}^t \sigma_i^2$  is equal in distribution to the discrete time process  $\sum_{i=1}^t \sigma_i G_i$  at each  $t$ , we have that

$$\sum_{i=1}^t (Y_i - \mu_i) = \sum_{i=1}^t \sigma_i G_i + o\left(V_t^{3/8} \log V_t\right), \quad (7.80)$$

which completes the proof after dividing both sides by  $t$ .  $\square$

With Lemma 7.A.2 in mind, we can now prove the main result (Proposition 7.2.2).

*Proof of Proposition 7.2.2.* The proof proceeds in four steps, each step being dedicated to satisfying one of Theorem 7.2.3's conditions (G-1–G-4). Step 1 follows quickly from Lemma 7.A.2, while Step 2 requires us to derive a (sub)-Gaussian boundary for non-i.i.d. observations under martingale dependence. Step 3 follows from the arguments in Steps 1 and 2, and Step 4 follows from a simple argument that uses Condition L-3.

**Step 1: Satisfying Condition G-1 via Strassen [249] and Conditions L-1 and L-2** Notice that the assumptions of Lemma 7.A.2 are satisfied by Conditions L-1 and L-2, and hence Condition G-1 follows using the approximating process formed by  $\frac{1}{t} \sum_{i=1}^t \sigma_i G_i$  as in Lemma 7.A.2 with a rate of  $r_t := t^{-1} V_t^{3/8} \log V_t$ . That is,

$$(\hat{\mu}_t - \tilde{\mu}_t) - \frac{1}{t} \sum_{i=1}^t \sigma_i G_i = o\left(\frac{V_t^{3/8} \log V_t}{t}\right). \quad (7.81)$$

In Step 2, we provide a nonasymptotic boundary for the process given by  $\frac{1}{t} \sum_{i=1}^t \sigma_i G_i$ .

**Step 2: Satisfying Condition G-2 using a nonasymptotic sub-Gaussian boundary** Let  $\sigma_i G_i$  be as in Lemma 7.A.2. Define the conditional variance  $\sigma_t^2 := \text{Var}(Y_t \mid Y_1^{t-1})$  and note that

$$\widetilde{M}_t(\lambda) := \exp \left\{ \sum_{i=1}^t (\lambda \sigma_i G_i - \lambda^2 \sigma_i^2 / 2) \right\}$$

is a nonnegative martingale starting at one (with respect to the filtration generated by  $(G_t, \sigma_t)_{t=1}^\infty$ ). Mixing over  $\lambda$  with the probability density  $dF(\lambda)$  of a mean-zero Gaussian with variance  $\rho^2$  as in the proof of Theorem 7.2.2, we have that

$$\widetilde{M}_t := \int_{\lambda \in \mathbb{R}} \widetilde{M}_t(\lambda) dF(\lambda) = \exp \left\{ \frac{\rho^2 (\sum_{i=1}^t \sigma_i G_i)^2}{2(V_t \rho^2 + 1)} \right\} \cdot (V_t \rho^2 + 1)^{-1/2}$$

is also a martingale. By Ville's inequality for nonnegative (super)martingales, we have that  $\mathbb{P}(\exists t : \tilde{M}_t \geq 1/\alpha) \leq \alpha$  and hence with probability at least  $(1 - \alpha)$ ,

$$\forall t \geq 1, \left| \frac{1}{t} \sum_{i=1}^t \sigma_i G_i \right| < \sqrt{\frac{V_t \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{V_t \rho^2 + 1}{\alpha^2} \right)}. \quad (7.82)$$

In particular, combined with Step 1, we have that

$$\left( \hat{\mu}_t \pm \sqrt{\frac{V_t \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{V_t \rho^2 + 1}{\alpha^2} \right)} \right) \quad (7.83)$$

forms a  $(1 - \alpha)$ -AsympCS for  $\tilde{\mu}_t$ .

**Step 3: Satisfying Condition G-3 as a consequence of Conditions G-1 and G-2** Inspecting the boundary in (7.82), we notice that  $V_t^{3/8} \log V_t/t = o(\sqrt{V_t \log V_t}/t)$  and hence Condition G-3 is satisfied.

**Step 4: Satisfying Condition G-4 via Condition L-3** Writing out the margin of (7.83) combined with Condition L-3:  $\hat{\sigma}_t^2 - \tilde{\sigma}_t^2 = o(\tilde{\sigma}_t^2)$  and recalling that  $V_t := t\tilde{\sigma}_t$ , we have

$$\begin{aligned} \sqrt{\frac{V_t \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{V_t \rho^2 + 1}{\alpha^2} \right)} &= \sqrt{\frac{t(\hat{\sigma}_t^2 + o(\tilde{\sigma}_t^2)\rho^2 + 1)}{t^2 \rho^2} \log \left( \frac{t(\hat{\sigma}_t^2 + o(\tilde{\sigma}_t^2)\rho^2 + 1)}{\alpha^2} \right)} \\ &= \sqrt{\frac{t\hat{\sigma}_t^2 \rho^2 + o(t\tilde{\sigma}_t^2) + 1}{t^2 \rho^2} \log \left( \frac{t\hat{\sigma}_t^2 \rho^2 + o(t\tilde{\sigma}_t^2) + 1}{\alpha^2} \right)} \\ &= \sqrt{\left( \frac{t\hat{\sigma}_t^2 \rho^2 + 1}{t^2 \rho^2} + o(\tilde{\sigma}_t^2/t) \right) \log \left( \frac{t\hat{\sigma}_t^2 \rho^2 + o(t\tilde{\sigma}_t^2) + 1}{\alpha^2} \right)}. \end{aligned} \quad (7.84)$$

Focusing on the logarithmic factor, we have

$$\begin{aligned} \log \left( \frac{t\hat{\sigma}_t^2 \rho^2 + o(t\tilde{\sigma}_t^2) + 1}{\alpha^2} \right) &= \log \left( \frac{1 + t\hat{\sigma}_t^2 \rho^2}{\alpha^2} + o(t\tilde{\sigma}_t^2) \right) \\ &= \log \left( \frac{1 + t\hat{\sigma}_t^2 \rho^2}{\alpha^2} [1 + o(1)] \right) \\ &= \log \left( \frac{1 + t\hat{\sigma}_t^2 \rho^2}{\alpha^2} \right) + \log(1 + o(1)) \\ &= \log \left( \frac{1 + t\hat{\sigma}_t^2 \rho^2}{\alpha^2} \right) + o(1) \end{aligned} \quad (7.85)$$

where the last line follows from the Taylor expansion  $\log(1 + x) = x + o(1)$  for  $|x| < 1$ . Combining (7.84) and (7.85), we have that the margin of (7.83) can be written as

$$\begin{aligned} \sqrt{\frac{V_t \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{V_t \rho^2 + 1}{\alpha^2} \right)} &= \sqrt{\frac{t \hat{\sigma}_t^2 \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{1 + t \hat{\sigma}_t^2 \rho^2}{\alpha^2} \right) + o(V_t/t^2) + o(V_t \log V_t/t^2)} \\ &= \sqrt{\frac{t \hat{\sigma}_t^2 \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{1 + t \hat{\sigma}_t^2 \rho^2}{\alpha^2} \right) + o(V_t \log V_t/t^2)} \\ &\leq \sqrt{\frac{t \hat{\sigma}_t^2 \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{1 + t \hat{\sigma}_t^2 \rho^2}{\alpha^2} \right) + o\left(\frac{\sqrt{V_t \log V_t}}{t}\right)}, \end{aligned} \quad (7.86)$$

where the last inequality follows from  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$  for  $a, b \geq 0$ . In particular, letting

$$[L_t, U_t] := \left[ \hat{\mu}_t \pm \sqrt{\frac{V_t \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{V_t \rho^2 + 1}{\alpha^2} \right)} \right] \quad (7.87)$$

$$\text{and } [\hat{L}_t, \hat{U}_t] := \left[ \hat{\mu}_t \pm \sqrt{\frac{t \hat{\sigma}_t^2 \rho^2 + 1}{t^2 \rho^2} \log \left( \frac{t \hat{\sigma}_t^2 \rho^2 + 1}{\alpha^2} \right)} \right], \quad (7.88)$$

we have that  $\hat{L}_t/L_t \xrightarrow{\text{a.s.}} 1$  and  $\hat{U}_t/U_t \xrightarrow{\text{a.s.}} 1$ , satisfying Condition G-4 and completing the proof of Proposition 7.2.2.

□

#### 7.A.4 Proof of Corollary 7.2.1

*Proof.* As alluded to in the paragraph following Corollary 7.2.1, it suffices to show that if the following regularity conditions hold, i.e.

$$\sum_{i=1}^{\infty} \frac{\mathbb{E}|Y_i|^2 - \mathbb{E}Y_i^2|^{1+\beta}}{V_i^{1+\beta}} < \infty, \quad \tilde{\mu}_t^2 = o(V_t), \quad \text{and} \quad \frac{1}{t} \sum_{i=1}^t (\mu_i - \tilde{\mu}_t)^2 = o(\tilde{\sigma}_t^2), \quad (7.89)$$

then Condition L-3 is satisfied, meaning  $\hat{\sigma}_t^2/\tilde{\sigma}_t^2 \xrightarrow{\text{a.s.}} 1$  where  $\hat{\sigma}_t^2 := \frac{1}{t} \sum_{i=1}^t Y_i^2 - \left(\frac{1}{t} \sum_{i=1}^t Y_i\right)^2$ . To this end, notice that by the SLLN for independent random variables (see Petrov [202, Theorem 12]), we have that the left-most inequality of (7.89) implies that

$$\frac{1}{t} \sum_{i=1}^t [Y_i^2 - \mathbb{E}(Y_i^2)] = o(\tilde{\sigma}_t^2). \quad (7.90)$$

Writing out the difference  $\hat{\sigma}_t^2 - \tilde{\sigma}_t^2$ , we see that

$$\hat{\sigma}_t^2 - \tilde{\sigma}_t^2 \equiv \frac{1}{t} \sum_{i=1}^t Y_i^2 - \left( \frac{1}{t} \sum_{i=1}^t Y_i \right)^2 - \frac{1}{t} \sum_{i=1}^t [\mathbb{E}(Y_i^2) - (\mathbb{E}Y_i)^2] \quad (7.91)$$

$$= \underbrace{\frac{1}{t} \sum_{i=1}^t [Y_i^2 - \mathbb{E}(Y_i^2)]}_{=o(\tilde{\sigma}_t^2) \text{ by (7.90)}} - \underbrace{\left[ \left( \frac{1}{t} \sum_{i=1}^t Y_i \right)^2 - \frac{1}{t} \sum_{i=1}^t (\mathbb{E}Y_i)^2 \right]}_{(*)}. \quad (7.92)$$

Focusing now only  $(*)$ , we use the Lyapunov-type condition assumption  $\sum_{i=1}^{\infty} \frac{\mathbb{E}|Y_i - \mu_i|^{2+\delta}}{\sqrt{V_i^{2+\delta}}} < \infty$  combined with Petrov [202, Theorem 12] to note that  $\frac{1}{t} \sum_{i=1}^t (Y_i - \mu_i) = o(\sqrt{V_t}/t)$  and hence

$$(*) \equiv \left( \tilde{\mu}_t + o(\sqrt{V_t}/t) \right)^2 - \frac{1}{t} \sum_{i=1}^t (\mathbb{E}Y_i)^2 \quad (7.93)$$

$$= \tilde{\mu}_t^2 + o(\tilde{\mu}_t \sqrt{V_t}/t) + o(V_t/t^2) - \frac{1}{t} \sum_{i=1}^t (\mathbb{E}Y_i)^2 \quad (7.94)$$

$$= \left( \frac{1}{t} \sum_{i=1}^t \mu_i \right)^2 + o(\tilde{\mu}_t \sqrt{V_t}/t) + o(V_t/t^2) - \frac{1}{t} \sum_{i=1}^t (\mathbb{E}Y_i)^2 \quad (7.95)$$

$$= -\frac{1}{t} \sum_{i=1}^t (\mu_i - \tilde{\mu}_t)^2 + o(\tilde{\mu}_t \sqrt{V_t}/t) + o(V_t/t^2) \quad (7.96)$$

$$= o(\tilde{\sigma}_t^2), \quad (7.97)$$

where the final line follows from the assumptions of (7.89). This completes the proof of Corollary 7.2.1.

□

### 7.A.5 Proof of Theorems 7.3.1 and 7.3.2

In the proofs that follow, we will make extensive use of some convenient notation, namely the sample average operator  $\mathbb{P}_t f(Z) \equiv \frac{1}{t} \sum_{i=1}^t f(Z_i)$  and the conditional expectation operator  $\mathbb{P}\hat{f}(Z) \equiv \mathbb{P}(\hat{f}(Z_i) | Z'_1, \dots, Z'_n)$  where  $Z'_1, \dots, Z'_n$  are the data used to construct  $\hat{f}$ .

First, let us analyze the almost-sure behavior of the AIPW estimator  $\hat{\psi}_t$  for the average treatment effect  $\psi$ .

**Lemma 7.A.3** (Decomposition of  $\hat{\psi}_t - \psi$ ). *Let  $\hat{\psi}_t := \mathbb{P}_T(\hat{f}_{T'}) = \frac{1}{T} \sum_{i=1}^T \hat{f}_{T'}(Z_i^{\text{eval}})$  be a (possibly misspecified) estimator of  $\psi := \mathbb{P}(f) = \mathbb{E}\{f(Z^{\text{eval}})\}$  based on  $(Z_1^{\text{eval}}, \dots, Z_T^{\text{eval}})$  where  $\hat{f}_{T'}$  can be any estimator built from  $(Z_1^{\text{trn}}, \dots, Z_{T'}^{\text{trn}})$  and  $f : \mathcal{Z} \rightarrow \mathbb{R}$  any function. Furthermore, assume that there exists  $\bar{f}$  such that  $\|\hat{f}_{T'} - \bar{f}\|_{L_2(\mathbb{P})} \rightarrow 0$ . In other words,  $\hat{f}_{T'}$  is an estimator of  $f$  but may instead converge to  $\bar{f}$ . Then we have the decomposition,*

$$\hat{\psi}_t - \psi = \Gamma_t^{\text{SA}} + \Gamma_t^{\text{EP}} + \Gamma_t^{\text{B}}$$

where

$$\Gamma_t^{\text{SA}} := (\mathbb{P}_T - \mathbb{P})\bar{f} \quad \text{is the centered sample average term,} \quad (7.98)$$

$$\Gamma_t^{\text{EP}} := (\mathbb{P}_T - \mathbb{P})(\hat{f} - \bar{f}) \quad \text{is the empirical process term, and} \quad (7.99)$$

$$\Gamma_t^{\text{B}} := \mathbb{P}(\hat{f} - f) \quad \text{is the bias term.} \quad (7.100)$$

*Proof.* By definition of the quantities involved, we decompose

$$\hat{\psi}_t - \psi = \mathbb{P}_T(\hat{f}_{T'}) - \mathbb{P}(f) \quad (7.101)$$

$$= (\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'}) + \mathbb{P}(\hat{f}_{T'} - f) \quad (7.102)$$

$$= \underbrace{(\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f})}_{\Gamma_t^{\text{EP}}} + \underbrace{(\mathbb{P}_T - \mathbb{P})\bar{f}}_{\Gamma_t^{\text{SA}}} + \underbrace{\mathbb{P}(\hat{f}_{T'} - f)}_{\Gamma_t^{\text{B}}}, \quad (7.103)$$

which completes the proof.  $\square$

Now, let us analyze the almost-sure behaviour of the empirical process term  $\Gamma_t^{\text{EP}}$  and the bias term  $\Gamma_t^{\text{B}}$  to show that they vanish asymptotically at sufficiently fast rates. First, let us examine  $\Gamma_t^{\text{EP}}$ .

**Lemma 7.A.4** (Almost sure convergence of  $\Gamma_t^{\text{EP}}$ ). *Let  $\mathbb{P}_T$  denote the empirical measure over  $\mathbf{Z}_T^{\text{eval}} := (Z_1^{\text{eval}}, \dots, Z_T^{\text{eval}})$  and let  $\hat{f}_{T'}(z)$  be any function estimated from a sample  $\mathcal{D}_{T'}^{\text{trn}} = (Z_1^{\text{trn}}, Z_2^{\text{trn}}, \dots, Z_{T'}^{\text{trn}})$  which is independent of  $\mathcal{D}_T^{\text{eval}}$ . If  $\hat{\pi}_t \in [\delta, 1 - \delta]$  almost surely, then,*

$$\Gamma_t^{\text{EP}} := (\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f}) = O\left(\left\{\sum_{a=0}^1 \|\hat{\mu}_t^a - \bar{\mu}^a\|_{L_2(\mathbb{P})}\right\} \sqrt{\frac{\log \log t}{t}}\right).$$

In particular, if  $\|\hat{\mu}_t^a - \bar{\mu}^a\|_{L_2(\mathbb{P})} = o(1)$  for each  $a$ , then we have that  $\Gamma_t^{\text{EP}}$  almost-surely converges to 0 at a rate of  $o(\sqrt{\log \log t/t})$ , but possibly faster.

The proof proceeds in two steps. First, we use an argument from Kennedy et al. [156] and the law of the iterated logarithm to show  $\Gamma_t^{\text{EP}} = O\left(\|\hat{f}_t - \bar{f}\| \sqrt{\log \log t/t}\right)$ . Second and finally, we upper bound  $\|\hat{f}_t - \bar{f}\|$  by  $O\left(\sum_{a=0}^1 \|\hat{\mu}_t^a - \bar{\mu}^a\|\right)$ .

*Proof.* **Step 1.** Following the proof of Kennedy et al. [156, Lemma 2], we have that conditional on  $\mathcal{D}_\infty^{\text{trn}} := (Z_t^{\text{trn}})_{t=1}^\infty$  and  $\mathcal{S}_\infty^{\text{trn}} := (\mathbb{1}(Z_t \in \mathcal{D}_\infty^{\text{trn}}))_{t=1}^\infty$  the term of interest has mean zero:

$$\mathbb{E}\left\{\mathbb{P}_T(\hat{f}_{T'} - \bar{f}) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right\} = \mathbb{E}(\hat{f}_{T'} - \bar{f} \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}) = \mathbb{P}(\hat{f}_{T'} - \bar{f}).$$

Now, we upper bound the conditional variance of a single summand,

$$\text{Var}\left\{(1 - \mathbb{P})(\hat{f}_{T'} - \bar{f}) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right\} = \text{Var}\left\{(\hat{f}_{T'} - \bar{f}) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right\} \quad (7.104)$$

$$\leq \|\hat{f}_{T'} - \bar{f}\|^2. \quad (7.105)$$

In particular, this means that

$$\left( \frac{T(\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f})}{\|\hat{f}_{T'} - \bar{f}\|} \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}} \right)$$

is a sum of i.i.d. random variables with conditional mean zero and conditional variance at most 1, and thus by the law of the iterated logarithm,

$$\mathbb{P} \left( \limsup_{t \rightarrow \infty} \frac{\pm \sqrt{T}(\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f})}{\sqrt{2 \log \log T} \|\hat{f}_{T'} - \bar{f}\|} \leq 1 \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}} \right) = 1.$$

Therefore, we have that

$$\mathbb{P} \left( \frac{(\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f})}{\|\hat{f}_t - \bar{f}\| \sqrt{\log \log t/t}} = O(1) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}} \right) \quad (7.106)$$

$$= \mathbb{P} \left( \frac{\|\hat{f}_{T'} - \bar{f}\|}{\|\hat{f}_t - \bar{f}\|} \cdot \frac{\sqrt{t}(\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f})}{\sqrt{\log \log t} \|\hat{f}_{T'} - \bar{f}\|} = O(1) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}} \right) \quad (7.107)$$

$$= \mathbb{P} \left( \underbrace{\lim_{t \rightarrow \infty} \frac{\|\hat{f}_{T'} - \bar{f}\|}{\|\hat{f}_t - \bar{f}\|}}_{O(1)} \cdot \underbrace{\frac{\sqrt{t} \log \log T}{\sqrt{T} \log \log t}}_{O(1)} \cdot \underbrace{\frac{\sqrt{T}(\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f})}{\sqrt{\log \log T} \|\hat{f}_{T'} - \bar{f}\|}}_{O(1)} = O(1) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}} \right) = 1. \quad (7.108)$$

$$(7.109)$$

Finally, by iterated expectation,

$$\mathbb{P} \left( \frac{(\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f})}{\|\hat{f}_t - \bar{f}\| \sqrt{\log \log t/t}} = O(1) \right) = \mathbb{E} \left[ \mathbb{P} \left( \frac{(\mathbb{P}_T - \mathbb{P})(\hat{f}_{T'} - \bar{f})}{\|\hat{f}_t - \bar{f}\| \sqrt{\log \log t/t}} = O(1) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}} \right) \right] \quad (7.110)$$

$$= \mathbb{E} 1 = 1, \quad (7.111)$$

which completes Step 1.

**Step 2.** Now, let us upper bound  $\|\hat{f}_t - \bar{f}\|$  by  $O\left(\sum_{a=0}^1 \|\hat{\mu}_t^a - \bar{\mu}^a\|\right)$ . To simplify the calculations which follow, define

$$\hat{f}^1(Z_i) := \hat{\mu}^1(X_i) + \frac{A_i}{\hat{\pi}(X_i)} \{Y_i - \hat{\mu}^{A_i}(X_i)\} \text{ and } \bar{f}^1(Z_i) := \bar{\mu}^1(X_i) + \frac{A_i}{\pi(X_i)} \{Y_i - \bar{\mu}^1(X_i)\}.$$

Analogously define  $\hat{f}^0$  and  $\bar{f}^0$  so that  $\hat{f} = \hat{f}^1 - \hat{f}^0$  and  $\bar{f} = \bar{f}^1 - \bar{f}^0$ . Writing out  $\|\hat{f}_t^1 - \bar{f}^1\|$ ,

$$\|\hat{f}_t^1 - \bar{f}^1\| = \left\| \hat{\mu}_t^1 + \frac{A}{\hat{\pi}_t} \{Y - \hat{\mu}_t^1\} - \bar{\mu}_t^1 - \frac{A}{\hat{\pi}_t} \{Y - \bar{\mu}^1\} \right\| \quad (7.112)$$

$$= \left\| \{\hat{\mu}_t^1 - \bar{\mu}^1\} \left\{ 1 - \frac{A}{\hat{\pi}_t} \right\} \right\| \quad (7.113)$$

$$\leq \frac{1}{\delta} \cdot \|\hat{\mu}_t^1 - \bar{\mu}^1\| = O(\|\hat{\mu}_t^1 - \bar{\mu}^1\|), \quad (7.114)$$

where the last inequality follows from the assumed bounds on  $\hat{\pi}(X)$ . A similar story holds for  $\|\hat{f}_t^0 - \bar{f}^0\|$ , and hence by the triangle inequality,

$$\|\hat{f}_t - \bar{f}\| = O\left(\sum_{a=0}^1 \|\hat{\mu}_t^a - \bar{\mu}^a\|\right),$$

which completes the proof.  $\square$

Now, we examine the asymptotic almost-sure behaviour of the bias term,  $\Gamma_t^B$  by upper-bounding this term by a product of  $L_2(\mathbb{P})$  estimation errors of nuisance functions.

**Lemma 7.A.5** (Almost-surely bounding  $\Gamma_T^B$  by  $L_2(\mathbb{P})$  errors of nuisance functions). *Suppose  $\hat{\pi}_t \in [\delta, 1 - \delta]$  almost surely for some  $\delta > 0$ . Then*

$$\Gamma_T^B = O\left(\|\hat{\pi}_t - \pi\|_{L_2(\mathbb{P})} \left\{ \|\hat{\mu}_t^1 - \mu^1\|_{L_2(\mathbb{P})} + \|\hat{\mu}_t^0 - \mu^0\|_{L_2(\mathbb{P})} \right\}\right)$$

This is an immediate consequence of the usual proof for  $O_{\mathbb{P}}$  combined with the fact that expectations are real numbers, and thus stochastic boundedness is equivalent to almost-sure boundedness. For completeness, we recall this proof here as it is short and illustrative.

*Proof.* To simplify the calculations which follow, define

$$\hat{f}^1(Z_i) := \hat{\mu}^1(X_i) + \frac{A_i}{\hat{\pi}(X_i)} \{Y_i - \hat{\mu}^{A_i}(X_i)\} \text{ and } f^1(Z_i) := \mu^1(X_i) + \frac{A_i}{\pi(X_i)} \{Y_i - \mu^1(X_i)\}.$$

Analogously define  $\hat{f}^0$  and  $f^0$  so that  $\hat{f} = \hat{f}^1 - \hat{f}^0$  and  $f = f^1 - f^0$ . Therefore,

$$\mathbb{P}(\hat{f}^1 - f^1) \stackrel{(i)}{=} \mathbb{P}\left(\frac{A}{\hat{\pi}}(Y - \hat{\mu}^A) + \hat{\mu}^1 - \mu^1\right) \quad (7.115)$$

$$\stackrel{(ii)}{=} \mathbb{P}\left[\left(\frac{\pi}{\hat{\pi}} - 1\right)(\hat{\mu}^1 - \mu^1)\right] \quad (7.116)$$

$$\stackrel{(iii)}{\leq} \frac{1}{\delta} \mathbb{P}(|(\hat{\pi} - \pi)(\hat{\mu}^1 - \mu^1)|) \quad (7.117)$$

$$\stackrel{(iv)}{\leq} \frac{1}{\delta} \|\hat{\pi} - \pi\|_{L_2(\mathbb{P})} \|\hat{\mu}^1 - \mu^1\|_{L_2(\mathbb{P})}, \quad (7.118)$$

where (i) and (ii) follow by iterated expectation, (iii) follows from the assumed bounds on  $\hat{\pi}$ , and (iv) by Hölder's inequality. Similarly, we have

$$\mathbb{P}(\hat{f}^0 - f^0) \leq \frac{1}{1-\delta} \|\hat{\pi} - \pi\| \|\hat{\mu}^0 - \mu^0\|.$$

Finally by the triangle inequality,

$$\mathbb{P}(\hat{f} - f) = O\left(\|\hat{\pi} - \pi\| \sum_{a=0}^1 \|\hat{\mu}^a - \mu^a\|\right),$$

which completes the proof.  $\square$

**Lemma 7.A.6** (Almost-sure consistency of the influence function variance estimator). *Suppose that  $\|\hat{f}_{T'} - \bar{f}\|_2 = o(1)$  and that  $\bar{f}(Z)$  has a finite second moment. Then,*

$$\widehat{\text{Var}}_T(\hat{f}_{T'}) = \text{Var}(\bar{f}) + o(1).$$

*Proof.* By direct calculation, we see that

$$\mathbb{P}_T \left( \hat{f}_{T'} - \mathbb{P}_T \hat{f}_{T'} \right)^2 = \underbrace{\mathbb{P}_T \left( \hat{f}_{T'} - \mathbb{P} \bar{f} \right)^2}_{(i)} + \underbrace{\left[ \mathbb{P}_T (\hat{f}_{T'} - \mathbb{P} \bar{f}) \right]^2}_{(ii)}, \quad (7.119)$$

and we have by the preceding lemmas that (ii)  $\xrightarrow{\text{a.s.}} 0$  since  $a_t \xrightarrow{\text{a.s.}} 0 \implies a_t^2 \xrightarrow{\text{a.s.}} 0$  and thus it suffices to show that (i)  $\xrightarrow{\text{a.s.}} \text{Var}(\bar{f})$ . Indeed, we use the inequality  $2ab \leq |ab| \leq a^2 + b^2$  and notice that

$$(i) = \mathbb{P}_T \left( \hat{f}_{T'} - \bar{f} + \bar{f} - \mathbb{P} \bar{f} \right)^2 \quad (7.120)$$

$$= \mathbb{P}_T \left( \hat{f}_{T'} - \bar{f} \right)^2 + 2\mathbb{P}_T \left( \hat{f}_{T'} - \bar{f} \right) (\bar{f} - \mathbb{P} \bar{f}) + \mathbb{P}_T (\bar{f} - \mathbb{P} \bar{f})^2 \quad (7.121)$$

$$\leq \mathbb{P}_T \left( \hat{f}_{T'} - \bar{f} \right)^2 + \mathbb{P}_T \left( \hat{f}_{T'} - \bar{f} \right)^2 + \mathbb{P}_T (\bar{f} - \mathbb{P} \bar{f})^2 + \mathbb{P}_T (\bar{f} - \mathbb{P} \bar{f})^2 \quad (7.122)$$

$$= 2 \left[ \underbrace{\mathbb{P}_T \left( \hat{f}_{T'} - \bar{f} \right)^2}_{(i.i)} + \underbrace{\mathbb{P}_T (\bar{f} - \mathbb{P} \bar{f})^2}_{(i.ii)} \right] \quad (7.123)$$

Let us now study (i.i) and (i.ii) separately. Note that  $\left\{ (\hat{f}_{T'}(Z_i) - \bar{f}(Z_i))^2 \right\}_{i \in \mathbb{N}}$  are conditionally i.i.d. and hence  $\mathbb{P}_T \left( \hat{f}_{T'} - \bar{f} \right)^2$  is a (conditional) reverse submartingale. By Manole and Ramdas [183, Theorem 2], we have

$$\mathbb{P} \left( \exists k \geq T' : \mathbb{P}_T (\hat{f}_{T'} - \bar{f})^2 \geq \varepsilon \right) \leq \mathbb{P}(\hat{f}_{T'} - \bar{f})^2 = \|\hat{f}_{T'} - \bar{f}\|^2 \xrightarrow{\text{a.s.}} 0, \quad (7.124)$$

so that  $(i.i) \rightarrow 0$   $\mathbb{P}$ -almost surely. Furthermore, we have by the finite second moment assumption  $\mathbb{P}|\bar{f} - \mathbb{P}\bar{f}|^2 < \infty$  that  $(i.ii) \rightarrow \text{Var}(\bar{f})$   $\mathbb{P}$ -almost surely by the strong law of large numbers. Putting these together, we have that  $(i) \rightarrow \text{Var}(\bar{f})$   $\mathbb{P}$ -almost surely, completing the proof.  $\square$

**Proposition 7.A.1** (General AsympCSs under sequential cross fitting). *Consider the cross-fit estimator as defined in (7.37):*

$$\hat{\psi}_t^\times := \frac{\sum_{i=1}^T f_{T'}(Z_i^{\text{eval}}) + \sum_{i=1}^{T'} f_T(Z_i^{\text{trn}})}{t}, \quad (7.125)$$

and the cross-fit variance estimator as defined in (7.38):

$$\widehat{\text{Var}}_t(f) := \frac{\widehat{\text{Var}}_T(\hat{f}_{T'}) + \widehat{\text{Var}}_{T'}(\hat{f}_T)}{2}. \quad (7.126)$$

Suppose that  $\Gamma_t^B$  and  $\Gamma_t^{\text{EP}}$  are both  $o(\sqrt{\log \log t/t})$ . Then,

$$\hat{\psi}_t^\times \pm \sqrt{\widehat{\text{Var}}_t(\hat{f})} \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)}$$

forms a  $(1 - \alpha)$ -AsympCS for  $\psi$ .

*Proof.* Writing out the centered cross-fit estimator  $\hat{\psi}_t^\times - \psi$  using the decomposition of Lemma 7.A.3, we have

$$\begin{aligned} \hat{\psi}_t^\times - \psi &= \frac{\sum_{i=1}^T \hat{f}_{T'}(Z_i^{\text{eval}}) + \sum_{i=1}^{T'} \hat{f}_T(Z_i^{\text{trn}}) - t\psi}{t} \\ &= \frac{\sum_{i=1}^T (f_{T'}(Z_i^{\text{eval}}) - \psi) + \sum_{i=1}^{T'} (\hat{f}_T(Z_i^{\text{trn}}) - \psi)}{t} \\ &= \frac{(T\Gamma_{t,\text{eval}}^{\text{SA}} + T'\Gamma_{t,\text{trn}}^{\text{SA}}) + T\Gamma_{t,\text{eval}}^{\text{EP}} + T'\Gamma_{t,\text{trn}}^{\text{EP}} + T\Gamma_{t,\text{eval}}^B + T'\Gamma_{t,\text{trn}}^B}{t} \\ &= (\mathbb{P}_t - \mathbb{P})\bar{f}(Z) + \frac{T\Gamma_{t,\text{eval}}^{\text{EP}} + T'\Gamma_{t,\text{trn}}^{\text{EP}} + T\Gamma_{t,\text{eval}}^B + T'\Gamma_{t,\text{trn}}^B}{t} \\ &= (\mathbb{P}_t - \mathbb{P})\bar{f}(Z) + \underbrace{O(\Gamma_t^B + \Gamma_t^{\text{EP}})}_{o(\sqrt{\log \log t/t})} \end{aligned} \quad (7.127)$$

where  $\Gamma_{t,\text{eval}}^{\text{EP}} := \frac{1}{T} \sum_{i=1}^T \hat{f}_{T'}(Z_i^{\text{eval}})$  and  $\Gamma_{t,\text{trn}}^{\text{EP}} := \frac{1}{T'} \sum_{i=1}^{T'} \hat{f}_T(Z_i^{\text{trn}})$ , and similarly for  $\Gamma_{t,\text{eval}}^{\text{SA}}$ ,  $\Gamma_{t,\text{trn}}^{\text{SA}}$ ,  $\Gamma_{t,\text{eval}}^B$ , and  $\Gamma_{t,\text{trn}}^B$ . Applying the proof of Theorem 7.2.2 (but with variance consistency  $\widehat{\text{Var}}_t(\hat{f}) \xrightarrow{\text{a.s.}} \text{Var}(\bar{f})$  obtained via Lemma 7.A.6), we have that

$$\hat{\psi}_t^\times \pm \sqrt{\widehat{\text{Var}}_t(\hat{f})} \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)} + o(\sqrt{\log \log t/t})$$

forms a nonasymptotic  $(1 - \alpha)$ -CS for  $\psi$ . Consequently,

$$\hat{\psi}_t^\times \pm \sqrt{\widehat{\text{Var}}_t(\hat{f})} \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\rho^2 + 1}{\alpha^2}\right)}$$

forms a  $(1 - \alpha)$ -AsympCS for  $\psi$  with rate  $o(\sqrt{\log \log t/t})$  which completes the proof.  $\square$

#### 7.A.5.1 Proof of Theorem 7.3.1

*Proof.* When propensity scores are known, we have that  $\Gamma_t^B = 0$  by Lemma 7.A.5. By assumption,  $\mathbb{E}\|\hat{\mu}_{T'}^a(X) - \bar{\mu}^a(X)\|_2 = o(1)$ , and thus by Lemma 7.A.4,  $\Gamma_t^{\text{EP}} = o(\sqrt{\log \log t/t})$ . Combining these conditions on  $\Gamma_t^B$  and  $\Gamma_t^{\text{EP}}$  with Proposition 7.A.1, we obtain the desired result. This completes the proof of Theorem 7.3.1.  $\square$

#### 7.A.5.2 Proof of Theorem 7.3.2

*Proof.* By Lemmas 7.A.4 and 7.A.5 we have that both  $\Gamma_t^B$  and  $\Gamma_t^{\text{EP}}$  are  $o(\sqrt{\log \log t/t})$ . Applying Proposition 7.A.1, we obtain the desired result, completing the proof of Theorem 7.3.2.  $\square$

#### 7.A.6 Proof of Theorem 7.3.3

**Lemma 7.A.7** (Decomposition of  $t\hat{\psi}_t^\times - t\tilde{\psi}_t$ ). *Let  $\hat{\psi}_t^\times$  be as in (7.37). Furthermore, assume that there exists  $\bar{f}$  such that  $\|\hat{f}_t - \bar{f}\|_{L_2(\mathbb{P})} \rightarrow 0$ . In other words,  $\hat{f}_t$  is an estimator of  $f$  but may instead converge to  $\bar{f}$ . Then we have the decomposition,*

$$t\hat{\psi}_t^\times - t\tilde{\psi}_t = \tilde{S}_t^{\text{SA}} + \underbrace{\tilde{S}_{t,\text{eval}}^{\text{EP}} + \tilde{S}_{t,\text{trn}}^{\text{EP}}}_{\tilde{S}_t^{\text{EP}}} + \underbrace{\tilde{S}_{t,\text{eval}}^B + \tilde{S}_{t,\text{trn}}^B}_{\tilde{S}_t^B} \quad (7.128)$$

where

$$\tilde{S}_t^{\text{SA}} := \sum_{i=1}^t [\bar{f}(Z_i) - \mathbb{P}(\bar{f}(Z_i))], \quad (7.129)$$

$$\tilde{S}_{t,\text{eval}}^{\text{EP}} := \sum_{i=1}^T \left\{ [\hat{f}_{T'}(Z_i^{\text{eval}}) - \mathbb{P}(\hat{f}_{T'}(Z_i^{\text{eval}}))] - [\bar{f}(Z_i^{\text{eval}}) - \mathbb{P}(\bar{f}(Z_i^{\text{eval}}))] \right\}, \quad (7.130)$$

$$\tilde{S}_{t,\text{trn}}^{\text{EP}} := \sum_{i=1}^{T'} \left\{ [\hat{f}_T(Z_i^{\text{trn}}) - \mathbb{P}(\hat{f}_T(Z_i^{\text{trn}}))] - [\bar{f}(Z_i^{\text{trn}}) - \mathbb{P}(\bar{f}(Z_i^{\text{trn}}))] \right\}, \quad (7.131)$$

$$\tilde{S}_{t,\text{eval}}^B := \sum_{i=1}^T \mathbb{P}(\hat{f}_{T'}(Z_i^{\text{eval}}) - f(Z_i^{\text{eval}})), \quad \text{and} \quad (7.132)$$

$$\tilde{S}_{t,\text{trn}}^B := \sum_{i=1}^{T'} \mathbb{P}(\hat{f}_T(Z_i^{\text{trn}}) - f(Z_i^{\text{trn}})). \quad (7.133)$$

*Proof.* First, note that  $t\hat{\psi}_t^\times - t\tilde{\psi}_t$  can be written as

$$t\hat{\psi}_t^\times - t\tilde{\psi}_t = \sum_{i=1}^T \hat{f}_{T'}(Z_i^{\text{eval}}) + \sum_{i=1}^{T'} \hat{f}_T(Z_i^{\text{trn}}) - \sum_{i=1}^T \mathbb{P}f(Z_i^{\text{eval}}) - \sum_{i=1}^{T'} \mathbb{P}f(Z_i^{\text{trn}}) \quad (7.134)$$

$$= \underbrace{\sum_{i=1}^T [\hat{f}_{T'}(Z_i^{\text{eval}}) - \mathbb{P}f(Z_i^{\text{eval}})]}_{(i)} + \underbrace{\sum_{i=1}^{T'} [\hat{f}_T(Z_i^{\text{trn}}) - \mathbb{P}f(Z_i^{\text{trn}})]}_{(ii)}. \quad (7.135)$$

We will handle each sum separately and then combine them to arrive at the final decomposition (7.128). Taking a closer look at (i) first, we have

$$(i) = \sum_{i=1}^T \left\{ [\hat{f}_{T'}(Z_i^{\text{eval}}) - \mathbb{P}(\hat{f}_{T'}(Z_i^{\text{eval}}))] - [\bar{f}(Z_i^{\text{eval}}) - \mathbb{P}(\bar{f}(Z_i^{\text{eval}}))] \right\} \quad (7.136)$$

$$+ \sum_{i=1}^T \mathbb{P} \left\{ \hat{f}_{T'}(Z_i^{\text{eval}}) - f(Z_i^{\text{eval}}) \right\} + \sum_{i=1}^T \left\{ \bar{f}(Z_i^{\text{eval}}) - \mathbb{P}(\bar{f}(Z_i^{\text{eval}})) \right\}. \quad (7.137)$$

Similarly for (ii), we have

$$(ii) = \sum_{i=1}^{T'} \left\{ [\hat{f}_T(Z_i^{\text{trn}}) - \mathbb{P}(\hat{f}_T(Z_i^{\text{trn}}))] - [\bar{f}(Z_i^{\text{trn}}) - \mathbb{P}(\bar{f}(Z_i^{\text{trn}}))] \right\} \quad (7.138)$$

$$+ \sum_{i=1}^{T'} \mathbb{P} \left\{ \hat{f}_T(Z_i^{\text{trn}}) - f(Z_i^{\text{trn}}) \right\} + \sum_{i=1}^{T'} \left\{ \bar{f}(Z_i^{\text{trn}}) - \mathbb{P}(\bar{f}(Z_i^{\text{trn}})) \right\}. \quad (7.139)$$

Putting (i) and (ii) together, we have

$$t\hat{\psi}_t^\times - t\tilde{\psi}_t = \sum_{i=1}^{T'} \left\{ \bar{f}(Z_i^{\text{eval}}) - \mathbb{P}(\bar{f}(Z_i^{\text{eval}})) \right\} + \sum_{i=1}^T \left\{ \bar{f}(Z_i^{\text{trn}}) - \mathbb{P}(\bar{f}(Z_i^{\text{trn}})) \right\} + \tilde{S}_t^{\text{EP}} + \tilde{S}_t^{\text{B}} \quad (7.140)$$

$$= \underbrace{\sum_{i=1}^t \left\{ \bar{f}(Z_i) - \mathbb{P}(\bar{f}(Z_i)) \right\}}_{\tilde{S}_t^{\text{SA}}} + \tilde{S}_t^{\text{EP}} + \tilde{S}_t^{\text{B}}, \quad (7.141)$$

which completes the proof.  $\square$

**Lemma 7.A.8** (Almost sure behavior of  $\tilde{S}_t^{\text{EP}}$ ). *Suppose that there exists  $\delta > 0$  such that*

$\hat{\pi}_t \in [\delta, 1 - \delta]$  almost surely for all  $t$ . Then,

$$\tilde{S}_t^{\text{EP}} = o\left(\left\{\sum_{a=0}^1 \sup_i \|\hat{\mu}_t^a(X_i) - \bar{\mu}^a(X_i)\|_{L_2(\mathbb{P})}\right\} \sqrt{t \log \log t}\right). \quad (7.142)$$

*Proof.* We will show that the result (7.142) holds for each of  $\tilde{S}_{t,\text{eval}}^{\text{EP}}$  and  $\tilde{S}_{t,\text{trn}}^{\text{EP}}$ , thereby yielding the same result for their sum  $\tilde{S}_t^{\text{EP}}$ . The proof proceeds in two steps. First, we use an argument from Kennedy et al. [156] and the law of the iterated logarithm to bound  $\tilde{S}_t^{\text{SA}}$  in terms of  $\sup_i \|\hat{f}_T(Z_i) - \bar{f}(Z_i)\|$ . Second and finally, we upper bound  $\sup_i \|\hat{f}_t(Z_i) - \bar{f}(Z_i)\|$  by  $O\left(\sum_{a=0}^1 \sup_i \|\hat{\mu}_t^a(Z_i) - \bar{\mu}(Z_i)\|\right)$ .

**Step 1** Let us first consider  $\tilde{S}_{t,\text{eval}}^{\text{EP}}$ . Following the proof of Kennedy et al. [156, Lemma 2] and of Lemma 7.A.4, note that conditional on  $\mathcal{D}_\infty^{\text{trn}} := (Z_t^{\text{trn}})_{t=1}^\infty$  and  $\mathcal{S}_\infty^{\text{trn}} := (\mathbb{1}(Z_t \in \mathcal{D}_\infty^{\text{trn}}))_{t=1}^\infty$  the summands of  $\tilde{S}_{t,\text{eval}}^{\text{EP}}$  have mean zero:

$$\mathbb{P}\left\{\left[\hat{f}_{T'}(Z_i^{\text{eval}}) - \bar{f}(Z_i^{\text{eval}})\right] - \left[\bar{f}(Z_i^{\text{eval}}) - \mathbb{P}(\bar{f}(Z_i^{\text{eval}}))\right] \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right\} = 0.$$

Similar to the proof of Lemma 7.A.4, we upper bound the conditional variance of a single summand,

$$\text{Var}\left\{(1 - \mathbb{P})(\hat{f}_{T'}(Z_i^{\text{eval}}) - \bar{f}(Z_i^{\text{eval}})) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right\} = \text{Var}\left\{(\hat{f}_{T'}(Z_i^{\text{eval}}) - \bar{f}(Z_i^{\text{eval}})) \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right\} \quad (7.143)$$

$$\leq \|\hat{f}_{T'}(Z_i^{\text{eval}}) - \bar{f}(Z_i^{\text{eval}})\|_{L_2(\mathbb{P})}^2. \quad (7.144)$$

Denote the following process  $\nu_{t,\text{eval}}$  as the supremum of the above with respect to  $i \in \{1, 2, \dots\}$ :

$$\nu_{t,\text{eval}} := \sup_{1 \leq i \leq \infty} \|\hat{f}_{T'}(Z_i^{\text{eval}}) - \bar{f}(Z_i^{\text{eval}})\|_{L_2(\mathbb{P})}.$$

Then we can lower bound the following conditional probability

$$\begin{aligned} & \mathbb{P}\left(\limsup_{t \rightarrow \infty} \frac{\pm \tilde{S}_{t,\text{eval}}^{\text{EP}}}{\nu_{t,\text{eval}} \sqrt{2t \log \log t}} \leq 1 \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right) \\ &= \mathbb{P}\left(\limsup_{t \rightarrow \infty} \sum_{i=1}^T \frac{\pm \left\{\left[\hat{f}_{T'}(Z_i^{\text{eval}}) - \mathbb{P}(\hat{f}_{T'}(Z_i^{\text{eval}}))\right] - \left[\bar{f}(Z_i^{\text{eval}}) - \mathbb{P}(\bar{f}(Z_i^{\text{eval}}))\right]\right\}}{\nu_{t,\text{eval}} \sqrt{2t \log \log t}} \leq 1 \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right) \\ &\geq \mathbb{P}\left(\limsup_{t \rightarrow \infty} \sum_{i=1}^T \frac{\pm \left\{\left[\hat{f}_{T'}(Z_i^{\text{eval}}) - \mathbb{P}(\hat{f}_{T'}(Z_i^{\text{eval}}))\right] - \left[\bar{f}(Z_i^{\text{eval}}) - \mathbb{P}(\bar{f}(Z_i^{\text{eval}}))\right]\right\}}{\|\hat{f}_{T'}(Z_i^{\text{eval}}) - \bar{f}(Z_i^{\text{eval}})\|_{L_2(\mathbb{P})} \sqrt{2t \log \log t}} \leq 1 \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right) \\ &= \mathbb{P}\left(\limsup_{t \rightarrow \infty} \sum_{i=1}^T \frac{\pm \zeta_i}{\sqrt{2t \log \log t}} \leq 1 \mid \mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}\right), \end{aligned} \quad (7.145)$$

where  $\zeta_i$  are independent mean-zero random variables with variance at most one (conditional on  $\mathcal{D}_\infty^{\text{trn}}, \mathcal{S}_\infty^{\text{trn}}$ ). By the law of the iterated logarithm, we have that (7.145) = 1. In particular, since this event happens with probability one conditionally, it also happens with probability one marginally. It follows that

$$\tilde{S}_{t,\text{eval}}^{\text{EP}} = O \left( \sup_i \|\hat{f}_t(Z_i) - \bar{f}(Z_i)\|_{L_2(\mathbb{P})} \sqrt{t \log \log t} \right). \quad (7.146)$$

Applying the same technique to  $\tilde{S}_{t,\text{trn}}^{\text{EP}}$ , we have that  $\tilde{S}_{t,\text{eval}}^{\text{EP}} = O \left( \sup_i \|\hat{f}_t(Z_i) - \bar{f}(Z_i)\|_{L_2(\mathbb{P})} \sqrt{t \log \log t} \right)$ , and hence

$$\tilde{S}_t^{\text{EP}} = O \left( \sup_i \|\hat{f}_t(Z_i) - \bar{f}(Z_i)\|_{L_2(\mathbb{P})} \sqrt{t \log \log t} \right). \quad (7.147)$$

**Step 2** Now, following the same technique as Step 2 in the proof of Lemma 7.A.4, we have that

$$\|\hat{f}_{T'}(Z_i) - \bar{f}(Z_i)\| = O \left( \sum_{a=0}^1 \|\hat{\mu}_t^a(X_i) - \bar{\mu}_i^a(X_i)\| \right). \quad (7.148)$$

Combining (7.147) and (7.148), we have the desired result,

$$\tilde{S}_t^{\text{EP}} = O \left( \left\{ \sum_{a=0}^1 \sup_{1 \leq i \leq \infty} \|\hat{\mu}_t^a(X_i) - \bar{\mu}_i^a(X_i)\| \right\} \sqrt{t \log \log t} \right), \quad (7.149)$$

which completes the proof.  $\square$

Now, we examine the asymptotic almost-sure behaviour of the bias term,  $\Gamma_t^B$  by upper-bounding this term by a product of  $L_2(\mathbb{P})$  estimation errors of nuisance functions.

**Lemma 7.A.9** (Almost-sure behavior of  $\tilde{S}_t^B$ ). *Suppose  $\hat{\pi}_t \in [\delta, 1 - \delta]$  for every  $t$  almost surely for some  $\delta > 0$ . Then,*

$$\tilde{S}_t^B = O \left( \sum_{i=1}^t \|\hat{\pi}_t(X_i) - \pi(X_i)\|_{L_2(\mathbb{P})} \sum_{a=0}^1 \|\hat{\mu}_t^a(X_i) - \mu^a(X_i)\|_{L_2(\mathbb{P})} \right) \quad (7.150)$$

The proof proceeds similarly to that of Lemma 7.A.5 but with additional care given to the fact that observations are no longer identically distributed.

*Proof.* Similar to the proof of Lemma 7.A.8, we will first prove the result for  $\tilde{S}_{t,\text{eval}}^B$ , and the proof proceeds similarly for  $\tilde{S}_{t,\text{trn}}^B$ , thereby yielding the desired result for  $\tilde{S}_t^B \equiv \tilde{S}_{t,\text{eval}}^B + \tilde{S}_{t,\text{trn}}^B$ . Following the same technique as Lemma 7.A.5, we have that

$$\mathbb{P}(\hat{f}_{T'}(Z_i^{\text{eval}}) - f(Z_i^{\text{eval}})) = O \left( \|\hat{\pi}_{T'}(X_i^{\text{eval}}) - \pi(X_i^{\text{eval}})\| \sum_{a=0}^1 \|\hat{\mu}_{T'}^a(X_i^{\text{eval}}) - \mu^a(X_i^{\text{eval}})\| \right).$$

Putting the above term back into the sum  $\tilde{S}_{t,\text{eval}}^B$ , we have

$$\tilde{S}_{t,\text{eval}}^B := \sum_{i=1}^T \mathbb{P}(\hat{f}_{T'}(Z_i^{\text{eval}}) - f(Z_i^{\text{eval}})) \quad (7.151)$$

$$= O\left(\sum_{i=1}^T \|\hat{\pi}_{T'}(X_i^{\text{eval}}) - \pi(X_i^{\text{eval}})\| \sum_{a=0}^1 \|\hat{\mu}_{T'}^a(X_i^{\text{eval}}) - \mu^a(X_i^{\text{eval}})\|\right). \quad (7.152)$$

Using a similar argument to bound  $\tilde{S}_{t,\text{trn}}^B$  and putting these together, we have the following bound for  $\tilde{S}_t^B = \tilde{S}_{t,\text{eval}}^B + \tilde{S}_{t,\text{trn}}^B$ ,

$$\tilde{S}_t^B = O\left(\sum_{i=1}^t \|\hat{\pi}_t(X_i) - \pi(X_i)\| \sum_{a=0}^1 \|\hat{\mu}_t^a(X_i) - \mu^a(X_i)\|\right), \quad (7.153)$$

which completes the proof.  $\square$

**Proposition 7.A.2** (General AsympCSs for time-varying causal effects under sequential cross fitting). *Consider the cross-fit estimator as defined in (7.37):*

$$\hat{\psi}_t^\times := \frac{\sum_{i=1}^T f_{T'}(Z_i^{\text{eval}}) + \sum_{i=1}^{T'} f_T(Z_i^{\text{trn}})}{t}, \quad (7.154)$$

and suppose we have access to a variance estimator  $\widehat{\text{Var}}_t(\hat{f})$  such that

$$\widehat{\text{Var}}_t(\hat{f}) - \widetilde{\text{Var}}(\bar{f}) = o(1).$$

Suppose that  $\tilde{S}_t^B$  and  $\tilde{S}_t^{\text{EP}}$  are both  $o(\sqrt{t \log \log t})$ , and that the conditions of Corollary 7.2.1 hold but with  $(Y)_{t=1}^\infty 1$  replaced by  $(\bar{f}(Z_t))_{t=1}^\infty$ . Then,

$$\hat{\psi}_t^\times \pm \sqrt{\frac{t\rho^2 \widehat{\text{Var}}_t(\bar{f}) + 1}{t^2 \rho^2} \log\left(\frac{t\rho^2 \widehat{\text{Var}}_t(\bar{f}) + 1}{\alpha^2}\right)}$$

forms a  $(1 - \alpha)$ -AsympCS for  $\tilde{\psi}_t := \frac{1}{t} \sum_{i=1}^t \psi_i$ .

*Proof.* Writing out the centered cross-fit estimator on the “sum scale”  $t(\hat{\psi}_t^\times - \tilde{\psi})$  using the decomposition of Lemma 7.A.7, we have

$$t(\hat{\psi}_t^\times - \tilde{\psi}_t) = \tilde{S}_t^{\text{SA}} + \underbrace{\tilde{S}_{t,\text{eval}}^{\text{EP}} + \tilde{S}_{t,\text{trn}}^{\text{EP}}}_{\tilde{S}_t^{\text{EP}} = o(\sqrt{t \log \log t})} + \underbrace{\tilde{S}_{t,\text{eval}}^B + \tilde{S}_{t,\text{trn}}^B}_{\tilde{S}_t^B = o(\sqrt{t \log \log t})}.$$

Therefore, we have that

$$\hat{\psi}_t - \tilde{\psi}_t = \frac{1}{t} \sum_{i=1}^t (\bar{f}(Z_i) - \psi_i) + o(\sqrt{\log \log t/t}).$$

Applying Corollary 7.2.1 to  $(\bar{f}(Z_t))_{t=1}^\infty$  above, we have that

$$\tilde{C}_t^{(\times\star)} := \hat{\psi}_t^\times \pm \sqrt{\frac{t\rho^2 \widehat{\text{Var}}_t(\bar{f}) + 1}{t^2 \rho^2} \log \left( \frac{t\rho^2 \widehat{\text{Var}}_t(\bar{f}) + 1}{\alpha^2} \right)} + o(\sqrt{\log \log t/t})$$

forms a nonasymptotic  $(1 - \alpha)$ -CS for  $\tilde{\psi}_t := \frac{1}{t} \sum_{i=1}^t \psi_i$ , meaning  $\mathbb{P}(\exists t : \tilde{\psi}_t \notin \tilde{C}_t^{(\times\star)}) \leq \alpha$ . Consequently,

$$\hat{\psi}_t^\times \pm \sqrt{\frac{t\rho^2 \widehat{\text{Var}}_t(\bar{f}) + 1}{t^2 \rho^2} \log \left( \frac{t\rho^2 \widehat{\text{Var}}_t(\bar{f}) + 1}{\alpha^2} \right)}$$

forms a  $(1 - \alpha)$ -AsympCS for  $\tilde{\psi}_t$ , which completes the proof.  $\square$

#### 7.A.6.1 Proof of Theorem 7.3.3

*Proof.* By Lemma 7.A.8 combined with Assumption  $\widetilde{\text{ATE-1}}$ , we have that  $\tilde{S}_t^{\text{EP}} = o(\sqrt{t \log \log t})$ . In a randomized experiment, Assumption  $\widetilde{\text{ATE-2}}$  holds by design, and thus by Lemma 7.A.9, we have that  $\tilde{S}_t^{\text{B}} = o(\sqrt{t \log \log t})$ . Invoking Proposition 7.A.2, we obtain the desired result.  $\square$

#### 7.A.7 Proof of Theorem 7.2.5

Before diving into the proof of Theorem 7.2.5, we will introduce some notation. Let  $(S(v))_{v \in [0, \infty)}$  be the (continuous-time) process obtained via  $S(V_t) = S_t$  for each  $t \in \mathbb{N}$  (with  $S(0) = 0$ ) and remaining piecewise constant in between the integers. Let  $\hat{S}(v)$  be defined analogously but with  $\hat{V}_t$  instead of  $V_t$ . Our first lemma will allow  $\hat{S}(v)$  to be written in terms of  $S(v)$  up to a smaller term in the argument.

**Lemma 7.A.10.** *Suppose  $\hat{V}_t - V_t = o(V_t^\eta)$  for some  $0 < \eta < 1$ , or in other words,  $\hat{\sigma}_t^2$  is a consistent estimator for  $\tilde{\sigma}_t^2$  at any polynomial rate in  $V_t$ . Then,*

$$\hat{S}(v) = S(v + o(v^\eta)) \text{ as } v \rightarrow \infty. \quad (7.155)$$

*Proof.* The result follows from the fact that  $\hat{V}_t - V_t = o(V_t^\eta)$  and that  $V_t \rightarrow \infty$  as  $t \rightarrow \infty$ .  $\square$

The next lemma approximates  $\hat{S}(v)$  by a Wiener process using the strong approximation results of Strassen [249] combined with Lemma 7.A.10.

**Lemma 7.A.11.** *After potentially enlarging the probability space, there exists a Wiener process  $(W(v))_{v \in \mathbb{R}^+}$  such that*

$$|\hat{S}(v) - W(v)| = o(v^{\kappa/2}) \quad (7.156)$$

for some  $0 < \kappa < 1$ .

*Proof.* By direct calculation, we have

$$|\widehat{S}(v) - W(v)| \stackrel{(i)}{=} |S(v + o(v^\eta)) - W(v)| \quad (7.157)$$

$$= \underbrace{|S(v + o(v^\eta)) - W(v + o(v^\eta))|}_{(*)} + \underbrace{|W(v + o(v^\eta)) - W(v)|}_{(\dagger)}, \quad (7.158)$$

where (i) follows from Lemma 7.A.10. We will subsequently bound (\*) using Strassen [249, Theorem 4.4], and (†) via the Borel-Cantelli lemma combined with elementary properties of the Wiener process.

**Bounding (★)** By Strassen [249, Theorem 4.4], we have that on a sufficiently rich probability space,  $S(v) - W(v) = o(v^{3/8} \log v)$ , and hence

$$(*) := |S(v + o(v^\eta)) - W(v + o(v^\eta))| \quad (7.159)$$

$$= o\left((v + o(v^\eta))^{3/8} \log(v + o(v^\eta))\right) \quad (7.160)$$

$$= o\left(v^{3/8} \log v\right), \quad (7.161)$$

where (7.161) follows from the assumption that  $\eta < 1$ .

**Bounding (†)** By elementary properties of the Wiener process, we have  $W(v + o(v^\eta)) - W(v) = \sqrt{o(v^\eta)}Z$  where  $Z \sim N(0, 1)$  is an independent standard Gaussian random variable. It then remains to show that  $\sqrt{o(v^\eta)}Z = o(v^{\kappa/2})$ . Indeed, notice that

$$\mathbb{P}\left(\frac{v^{\eta/2}}{v^{\kappa/2}}Z \geq \varepsilon\right) \leq \exp\left\{-\frac{\varepsilon^2 v^\kappa}{2v^\eta}\right\} = \exp\left\{-\frac{\varepsilon^2 v^{\kappa-\eta}}{2}\right\}, \quad (7.162)$$

and hence by the Borel-Cantelli lemma,

$$\frac{v^{1/2q}}{v^{\kappa/2}}Z \rightarrow 0 \text{ almost surely.} \quad (7.163)$$

for any  $\eta < \kappa \leq 1$ . □

*Proof of Theorem 7.2.5.* Recall that  $\rho_m = \sqrt{c/\widehat{V}_m d_m}$  where  $d_m$  is any increasing sequence  $\rightarrow \infty$  as  $m \rightarrow \infty$ . Define  $(\widehat{V}'(t))_{t \in [0, \infty)}$  as the continuous-time process obtained by setting  $\widehat{V}'(t) = \widehat{V}_t$  for each  $t \in \mathbb{N}$  and piecewise-constant between the integers. Writing down the

limit supremum in Theorem 7.2.5,

$$\limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists t \in \mathbb{N}^{\geq m} : |S_t| \geq \sqrt{\frac{t\hat{\sigma}_t^2 \rho_m^2 + 1}{\rho_m^2} \log \left( \frac{t\hat{\sigma}_t^2 \rho_m^2 + 1}{\alpha^2} \right)} \right) \quad (7.164)$$

$$= \limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists t \in \mathbb{R}^{\geq m} : |S'(t)| \geq \sqrt{\frac{\hat{V}'(t)\rho_m^2 + 1}{\rho_m^2} \log \left( \frac{\hat{V}'(t)\rho_m^2 + 1}{\alpha^2} \right)} \right), \quad (7.165)$$

where the equality follows from the fact that  $S'(t)$  and  $\hat{V}'(t)$  are piecewise-constant between integers. Note that since  $V_t \rightarrow \infty$  as  $t \rightarrow \infty$ , we have that for any real sequence  $(v_m)_{m=1}^\infty$  such that  $v_m \rightarrow \infty$ ,

$$\limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists t \in \mathbb{R}^{\geq m} : |S'(t)| \geq \sqrt{\frac{\hat{V}'(t)\rho_m^2 + 1}{\rho_m^2} \log \left( \frac{\hat{V}'(t)\rho_m^2 + 1}{\alpha^2} \right)} \right) \quad (7.166)$$

$$= \limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists v \geq v_m : |\hat{S}(v)| \geq \sqrt{\frac{v\rho_m^2 + 1}{\rho_m^2} \log \left( \frac{v\rho_m^2 + 1}{\alpha^2} \right)} \right). \quad (7.167)$$

Re-writing  $v$  as  $sv_md_m$  for some  $s \in \mathbb{R}^+$ , the above probability can be written as

$$\mathbb{P} \left( \exists v \geq v_m : |\hat{S}(v)| \geq \sqrt{\frac{v\rho_m^2 + 1}{\rho_m^2} \log \left( \frac{v\rho_m^2 + 1}{\alpha^2} \right)} \right) \quad (7.168)$$

$$= \mathbb{P} \left( \exists sv_md_m \geq v_m : |\hat{S}(sv_md_m)| \geq \sqrt{\frac{sv_md_m\rho_m^2 + 1}{\rho_m^2} \log \left( \frac{sv_md_m\rho_m^2 + 1}{\alpha^2} \right)} \right) \quad (7.169)$$

$$= \mathbb{P} \left( \exists sv_md_m \geq v_m : |\hat{S}(sv_md_m)| \geq \sqrt{\frac{sv_md_m c/v_md_m + 1}{c/v_md_m} \log \left( \frac{sv_md_m c/v_md_m + 1}{\alpha^2} \right)} \right). \quad (7.170)$$

Applying Lemma 7.A.11, we can approximate  $\hat{S}(sv_md_m)$  by  $W(sv_md_m)$  up to a  $o((sv_md_m)^{\kappa/2})$ , yielding

$$\mathbb{P} \left( \exists sv_md_m \geq v_m : |\hat{S}(sv_md_m)| \geq \sqrt{\frac{sc + 1}{c/v_md_m} \log \left( \frac{sc + 1}{\alpha^2} \right)} \right) \quad (7.171)$$

$$= \mathbb{P} \left( \exists s \geq \frac{1}{d_m} : \left| W(sv_md_m) + o((sv_md_m)^{\kappa/2}) \right| \geq \sqrt{\frac{sc + 1}{c/v_md_m} \log \left( \frac{sc + 1}{\alpha^2} \right)} \right) \quad (7.172)$$

$$= \mathbb{P} \left( \exists s \geq \frac{1}{d_m} : \left| \sqrt{v_md_m} W(s) + o((sv_md_m)^{\kappa/2}) \right| \geq \sqrt{\frac{sc + 1}{c/v_md_m} \log \left( \frac{sc + 1}{\alpha^2} \right)} \right) \quad (7.173)$$

$$= \mathbb{P} \left( \exists s \geq \frac{1}{d_m} : \left| W(s) + o \left( \frac{(sv_m d_m)^{\kappa/2}}{(v_m d_m)^{1/2}} \right) \right| \geq \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right), \quad (7.174)$$

where (7.173) follows from the fact that  $(W(cs))_{s \in \mathbb{R}^+} \stackrel{d}{=} (\sqrt{c}W(s))_{s \in \mathbb{R}^+}$  for any  $c > 0$ . Recalling the limit supremum of Theorem 7.2.5 and taking the limsup of the above, we have

$$\limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists t \in \mathbb{N}^{\geq m} : |S_t| \geq \sqrt{\frac{t\hat{\sigma}_t^2 \rho_m^2 + 1}{\rho_m^2} \log \left( \frac{t\hat{\sigma}_t^2 \rho_m^2 + 1}{\alpha^2} \right)} \right) \quad (7.175)$$

$$= \limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists s \geq \frac{1}{d_m} : \left| W(s) + o \left( s^{\kappa/2} (v_m d_m)^{\frac{\kappa-1}{2}} \right) \right| \geq \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right) \quad (7.176)$$

It remains to show that (7.176) converges to  $\alpha$ . We will do so by separately showing that (7.176) is upper- and lower-bounded by  $\alpha$ . Indeed for the upper bound, we have that

$$\sup_{s \geq 0} \left\{ \left| W(s) + o \left( s^{\kappa/2} (v_m d_m)^{\frac{\kappa-1}{2}} \right) \right| - \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right\} \quad (7.177)$$

$$\stackrel{d}{\rightarrow} \sup_{s \geq 0} \left\{ |W(s)| - \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right\} \text{ as } m \rightarrow \infty, \quad (7.178)$$

and hence by classical results for boundary-crossing of Wiener processes, we have that

$$(7.176) \equiv \limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists s \geq \frac{1}{d_m} : \left| W(s) + o \left( s^{\kappa/2} (v_m d_m)^{\frac{\kappa-1}{2}} \right) \right| \geq \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right) \quad (7.179)$$

$$\leq \limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists s \geq 0 : \left| W(s) + o \left( s^{\kappa/2} (v_m d_m)^{\frac{\kappa-1}{2}} \right) \right| \geq \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right) \quad (7.180)$$

$$= \mathbb{P} \left( \exists s \geq 0 : |W(s)| \geq \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right) \quad (7.181)$$

$$= \alpha. \quad (7.182)$$

Now, for the lower bound on (7.176), we have that

$$(7.176) \equiv \limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists s \geq \frac{1}{d_m} : \left| W(s) + o \left( s^{\kappa/2} (v_m d_m)^{\frac{\kappa-1}{2}} \right) \right| \geq \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right)$$

$$(7.183)$$

$$\geq \underbrace{\limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists s \geq 0 : \left| W(s) + o \left( s^{\kappa/2} (v_m d_m)^{\frac{\kappa-1}{2}} \right) \right| \geq \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \right)}_{(i)} \quad (7.184)$$

$$- \underbrace{\limsup_{m \rightarrow \infty} \mathbb{P} \left( \sup_{s \in [0, 1/d_m]} \left| W(s) + o \left( s^{\kappa/2} (v_m d_m)^{\frac{\kappa-1}{2}} \right) \right| - \sqrt{\frac{sc+1}{c} \log \left( \frac{sc+1}{\alpha^2} \right)} \geq 0 \right)}_{(ii)} \quad (7.185)$$

where the inequality (7.184) results from breaking the boundary-crossing event into the cases  $s \in [0, 1/d_m]$  and  $s \in [1/d_m, \infty)$ . From the same argument as the upper bound, we have that (i) =  $\alpha$ , and hence it remains to show that (ii)  $\rightarrow 0$ . Indeed, the local modulus of continuity of the Wiener process states that

$$\limsup_{s \rightarrow 0} \frac{|W(s)|}{\sqrt{2s \log \log(1/s)}} = 1 \text{ almost surely,} \quad (7.186)$$

and thus (ii)  $\rightarrow 0$ , so (7.176)  $\geq \alpha$ . This completes the proof of Theorem 7.2.5.  $\square$

### 7.A.8 Proof of Proposition 7.3.1

*Proof.* By Taylor's theorem, for each  $t$ , let  $\tilde{\theta}_t$  lie between  $\hat{\theta}_t$  and  $\theta$  so that with probability one,

$$g(\hat{\theta}_t) - g(\theta) = g'(\tilde{\theta}_t)(\hat{\theta}_t - \theta) \quad (7.187)$$

$$= (\hat{\theta}_t - \theta)g'(\theta) + \underbrace{(\hat{\theta}_t - \theta)}_{(*)} \underbrace{(g'(\tilde{\theta}_t) - g'(\theta))}_{(\dagger)} \quad (7.188)$$

$$= \frac{1}{t} \sum_{i=1}^t g'(\theta)\phi(Z_i) + o\left(\sqrt{\log t/t}\right) \quad (7.189)$$

where we have used the fact that  $(*) = O(\sqrt{\log \log t/t}) + o(\sqrt{\log t/t})$  by the law of the iterated logarithm and  $(\dagger) = o(1)$  by the almost-sure continuous mapping theorem, completing the proof.  $\square$

### 7.A.9 Proof of Proposition 7.2.3

First, we need the following lemma that calculates the probability of a mixture exponential Brownian motion exceeding  $\varepsilon > 0$  for any  $t \geq 1$ .

**Lemma 7.A.12** (A maximal (in)equality for mixture exponential Brownian motion with a

delayed start). Define the continuous-time process  $(M^1(t))_{t \geq 0}$  given by

$$M^1(t) := \frac{\exp\left\{\frac{1}{2t}W(t)^2\right\}}{\sqrt{t}} \quad (7.190)$$

where  $(W(t))_{t \geq 0}$  is a Wiener process. Then,

$$\mathbb{P}\left(\exists t \geq 1 : M^1(t) > \varepsilon\right) = 2[1 - \Phi(a) + a\phi(a)], \quad (7.191)$$

where  $a := \sqrt{2 \log \varepsilon}$ .

*Proof.* The proof proceeds in 3 steps. First, we construct a conditional mixture martingale based on geometric Brownian motion akin to the one found in the proof of Proposition 7.2.2 but with the Wiener process replaced by itself plus independent standard Gaussian noise. Second, we show via iterated expectation and Step 1 that

$$\mathbb{E}\left[\mathbb{P}(\exists t \geq 1 : M^1(t) \geq \varepsilon \mid W(1))\right] = \mathbb{E}\left[(\varepsilon^{-1} \exp\{W(1)^2/2\}) \wedge 1\right]. \quad (7.192)$$

Finally, we evaluate the expectation on the right-hand side by noting that  $W(1) \sim N(0, 1)$ .

**Step 1: Showing that  $(M^{1-}(s))_{s \geq 0}$  forms a conditional nonnegative martingale given  $Z$**  Let  $Z$  be a standard Gaussian random variable independent of  $(W(s))_{s \geq 0}$ . Then since geometric Brownian motion is a martingale, it is obvious that the process  $(M^{1-}(s; \lambda))_{s \geq 0}$  given by

$$M^{1-}(s; \lambda) := \exp\{\lambda[W(s) + Z] - s\lambda^2/2\} \quad (7.193)$$

forms a conditional nonnegative martingale given  $Z$  with  $M^{1-}(0; \lambda) \equiv \exp\{\lambda Z\}$ . Formally,  $M^{1-}(s; \lambda)$  forms a nonnegative martingale with respect to the filtration  $(\mathcal{F}_s)_{s \geq 0}$  given by  $\mathcal{F}_s := \sigma(W(s), Z)$ . By Fubini's theorem, we have that for any probability distribution  $F(\lambda)$ , the mixture

$$\int_{\lambda \in \mathbb{R}} M^{1-}(s; \lambda) dF(\lambda) \quad (7.194)$$

also forms a conditional nonnegative martingale given  $Z$  but now starting at  $\int_{\lambda} M^{1-}(0; \lambda) dF(\lambda)$ . In particular, using the techniques found in the proof of Proposition 7.2.2, we have that for  $dF(\lambda) := \frac{1}{\sqrt{2\pi}} \exp\{\lambda^2/2\} d\lambda$ ,

$$M^{1-}(s) := \int_{\lambda \in \mathbb{R}} M^{1-}(s; \lambda) dF(\lambda) = (s+1)^{-1/2} \exp\left\{\frac{(W(s) + Z)^2}{2(s+1)}\right\} \quad (7.195)$$

forms a conditional nonnegative martingale starting at  $M^{1-}(0) \equiv \exp\{Z^2/2\}$  conditional on  $Z$ .

**Step 2: Showing that**  $\mathbb{E}[\mathbb{P}(\exists t \geq 1 : M^1(t) \geq \varepsilon \mid W(1))] = \mathbb{E}[(\varepsilon^{-1} \exp \{W(1)^2/2\}) \wedge 1]$  Notice that if we let  $s := t - 1$ , then  $M^1(t)$  can be written as

$$M^1(t) := \frac{\exp \left\{ \frac{1}{2t} W(t)^2 \right\}}{\sqrt{t}} \quad (7.196)$$

$$= (s+1)^{-1/2} \exp \left\{ \frac{W(s+1)^2}{2(s+1)} \right\}, \quad (7.197)$$

and notice that by definition of the Wiener process, we have that  $(W(s+1))_{s \geq 0} = (W'(s) + W(1))_{s \geq 0}$  where  $(W'(s))_{s \geq 0}$  is a Wiener process independent of  $W(1)$ . Since  $W(1) \sim N(0, 1)$ , we have that

$$(M^1(t))_{t \geq 1} \mid W(1) \stackrel{d}{=} (M^{1-}(t-1))_{t \geq 1} \mid Z \quad (7.198)$$

and hence

$$\mathbb{P}(\exists t \geq 1 : M^1(t) \geq \varepsilon) = \mathbb{E}[\mathbb{P}(\exists t \geq 1 : M^1(t) \geq \varepsilon \mid W(1))] \quad (7.199)$$

$$= \mathbb{E}[\mathbb{P}(\exists t \geq 1 : M^{1-}(t-1) \geq \varepsilon \mid Z)] \quad (7.200)$$

$$= \mathbb{E}[\mathbb{P}(\exists s \geq 0 : M^{1-}(s) \geq \varepsilon \mid Z)] \quad (7.201)$$

$$= \mathbb{E}[(\varepsilon^{-1} \exp \{Z^2/2\}) \wedge 1] \quad (7.202)$$

which completes the proof of Step 2.

### Step 3: Integrating out the distribution of $W(1)$ to obtain $\mathbb{P}(\exists t \geq 1 : M^1(t) \geq \varepsilon)$

Writing out the desired probability  $\mathbb{P}(\exists t \geq 1 : M^1(t) \geq \varepsilon)$  as the final expectation from Step 2, we have

$$\mathbb{P}(\exists t \geq 1 : M^1(t) > \varepsilon) = \mathbb{E}[(\varepsilon^{-1} \exp \{W(1)^2/2\}) \wedge 1] \quad (7.203)$$

$$= \int_{z \in \mathbb{R}} (\varepsilon^{-1} \exp \{w^2/2\}) \wedge 1 \cdot d\mathbb{P}(W(1) \leq w) \quad (7.204)$$

$$= \int_{|w| > \sqrt{2 \log \varepsilon}} \frac{1}{\sqrt{2\pi}} \exp \{-w^2/2\} dw \quad (7.205)$$

$$+ \int_{|w| \leq \sqrt{2 \log \varepsilon}} (\varepsilon^{-1} \exp \{w^2/2\}) \frac{1}{\sqrt{2\pi}} \exp \{-w^2/2\} dw \quad (7.206)$$

$$= 2 \left( 1 - \Phi(\sqrt{2 \log \varepsilon}) \right) + \frac{2}{\varepsilon \sqrt{2\pi}} \sqrt{2 \log \varepsilon} \quad (7.207)$$

$$= 2(1 - \Phi(a) + a\phi(a)), \quad (7.208)$$

where  $a := \sqrt{2 \log \varepsilon}$ . This completes the proof of Lemma 7.A.12.  $\square$

The proof of Lemma 7.A.12 is similar in spirit to the derivation of Robbins [217, Eq. (20)] but for continuous-time Wiener processes. Moreover, we only relied on a simple application of Ville's inequality and iterated expectation. Given the above result, we are ready to prove Proposition 7.2.3.

*Proof of Proposition 7.2.3.* The proof is a straightforward application of the ideas in the proof of Theorem 7.2.5 (some of the relevant notation can be found therein) combined with Lemma 7.A.12. Writing out the crossing probability of  $|S_t|$  for any  $t \geq m$  for a fixed  $m$ , we have that

$$\mathbb{P} \left( \exists t \geq m : |S_t| \geq \sqrt{t\hat{\sigma}_t^2 \left[ a^2 + \log \left( \frac{t\hat{\sigma}_t^2}{m\hat{\sigma}_m^2} \right) \right]} \right) \quad (7.209)$$

$$= \mathbb{P} \left( \exists t \geq m : |S'(t)| \geq \sqrt{\hat{V}'_t \left[ a^2 + \log \left( \hat{V}'_t / \hat{V}'_m \right) \right]} \right) \quad (7.210)$$

$$= \mathbb{P} \left( \exists v \geq v_m : |\hat{S}(v)| \geq \sqrt{v [a^2 + \log(v/v_m)]} \right) \quad (7.211)$$

$$= \mathbb{P} \left( \exists sv_m \geq v_m : |\hat{S}(sv_m)| \geq \sqrt{sv_m [a^2 + \log(sv_m/v_m)]} \right) \quad (7.212)$$

$$= \mathbb{P} \left( \exists s \geq 1 : \left| W(sv_m) + o((sv_m)^{\kappa/2}) \right| \geq \sqrt{sv_m [a^2 + \log s]} \right) \quad (7.213)$$

$$= \mathbb{P} \left( \exists s \geq 1 : \left| \sqrt{v_m} W(s) + o((sv_m)^{\kappa/2}) \right| \geq \sqrt{sv_m [a^2 + \log s]} \right) \quad (7.214)$$

$$= \mathbb{P} \left( \exists s \geq 1 : \left| W(s) + o(s^{\kappa/2} v_m^{\kappa-1/2}) \right| \geq \sqrt{s [a^2 + \log s]} \right). \quad (7.215)$$

Now, since  $\sup_{s \geq 1} \left\{ |W(s) + o(s^{\kappa/2} v_m^{(\kappa-1)/2})| - \sqrt{s [a^2 + \log s]} \right\} \xrightarrow{d} \sup_{s \geq 1} \left\{ |W(s)| - \sqrt{s [a^2 + \log s]} \right\}$  as  $m \rightarrow \infty$ , we have that

$$\limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists t \geq m : |S_t| \geq \sqrt{t\hat{\sigma}_t^2 \left[ a^2 + \log \left( \frac{t\hat{\sigma}_t^2}{m\hat{\sigma}_m^2} \right) \right]} \right) \quad (7.216)$$

$$= \mathbb{P} (\exists s \geq 1 : |W(s)| \geq \sqrt{s [a^2 + \log s]}). \quad (7.217)$$

Writing out the event  $\{M^1(s) > \exp\{a^2/2\}\}$  with  $M^1(s)$  defined as in Lemma 7.A.12, we note that it is equivalent to the event in the above probability statement:

$$\left\{ \exists s \geq 1 : M^1(s) > \exp\{a^2/2\} \right\} \quad (7.218)$$

$$= \left\{ \exists s \geq 1 : \frac{\exp\{\frac{1}{2s} W(s)^2\}}{\sqrt{s}} > \exp\{a^2/2\} \right\} \quad (7.219)$$

$$= \left\{ \exists s \geq 1 : \frac{1}{2s} W(s)^2 - \log \sqrt{s} > a^2/2 \right\} \quad (7.220)$$

$$= \left\{ \exists s \geq 1 : |W(s)| > \sqrt{s [a^2 + \log s]} \right\}, \quad (7.221)$$

and hence by Lemma 7.A.12, we have that

$$\limsup_{m \rightarrow \infty} \mathbb{P} \left( \exists t \geq m : |S_t| \geq \sqrt{t\hat{\sigma}_t^2 \left[ a^2 + \log \left( \frac{t\hat{\sigma}_t^2}{m\hat{\sigma}_m^2} \right) \right]} \right) = 2(1 - \Phi(a) + a\phi(a)), \quad (7.222)$$

as desired. This completes the proof.  $\square$

## 7.B Additional discussions

### 7.B.1 One-sided asymptotic confidence sequences

In Sections 7.2.2 and 7.2.4, we derived universal two-sided AsympCSs for the means of independent random variables in the i.i.d. and time-varying settings, respectively. Here, we give analogous one-sided bounds for the aforementioned settings. First, let us derive a one-sided AsympCS for the mean of i.i.d. random variables analogous to Theorem 7.2.2.

**Proposition 7.B.1.** *Given the same setup as in Theorem 7.2.2, we have that*

$$\hat{\mu}_t - \hat{\sigma}_t \sqrt{\frac{2(t\rho^2 + 1)}{t^2\rho^2} \log \left( 1 + \frac{\sqrt{t\rho^2 + 1}}{2\alpha} \right)} \quad (7.223)$$

*forms a lower  $(1 - \alpha)$ -AsympCS for  $\mu$  with the same rates as given in Theorem 7.2.2.*

Notice that the  $(1 - \alpha)$ -AsympCS of Proposition 7.B.1 resembles the  $(1 - 2\alpha)$ -AsympCS of Theorem 7.2.2 but with an additional additive 1 inside the log. A similar phenomenon appears in the one- and two-sided sub-Gaussian CSs of Howard et al. [125]. Recall, however, that their bounds are nonasymptotic and require much stronger assumptions (and in particular are not applicable to the observational causal inference setup of this chapter).

Similar to the relationship between i.i.d. (Theorem 7.2.2) and martingale (Proposition 7.2.2) two-sided AsympCSs, an analogue of Proposition 7.B.1 can be derived under martingale dependence with time-varying means and variances.

**Proposition 7.B.2.** *Given the same setup and assumptions as Proposition 7.2.2, we have that*

$$\hat{\mu}_t - \sqrt{\frac{2(t\hat{\sigma}_t^2\rho^2 + 1)}{t^2\rho^2} \log \left( 1 + \frac{\sqrt{t\hat{\sigma}_t^2\rho^2 + 1}}{2\alpha} \right)} \quad (7.224)$$

*forms a lower  $(1 - \alpha)$ -AsympCS for the time-varying average  $\tilde{\mu}_t := \frac{1}{t} \sum_{i=1}^t \mu_i$ .*

We will first prove a lemma concerning one-sided boundaries for sums of independent *Gaussian* random variables, which in turn will be used to prove Propositions 7.B.1 and 7.B.2 shortly.

**Lemma 7.B.1.** *Suppose  $(G_t)_{t=1}^\infty \sim N(0, 1)$  is an i.i.d. sequence of standard Gaussian random*

variables. Then,

$$\mathbb{P} \left( \forall t \in \mathbb{N}, \tilde{\mu}_t \geq \underbrace{\frac{1}{t} \sum_{i=1}^t (\sigma_i G_i + \mu_i) - \sqrt{\frac{2(t\rho^2\tilde{\sigma}_t^2 + 1)}{(t\rho)^2} \log \left( 1 + \frac{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}}{2\alpha} \right)}}_{L_t^*} \right) \geq 1 - \alpha. \quad (7.225)$$

In other words,  $L_t^*$  forms a nonasymptotic lower  $(1 - \alpha)$ -CS for  $\tilde{\mu}_t$ .

*Proof.* The proof begins similarly to that of Theorem 7.2.2 but with a modified mixing distribution, and proceeds in four steps. First, we derive a sub-Gaussian nonnegative supermartingale (NSM) indexed by a parameter  $\lambda \in \mathbb{R}$  identical to that of Theorem 7.2.2. Second, we mix this NSM over  $\lambda$  using a *folded* Gaussian density (rather than the classical Gaussian density used in the proof of Theorem 7.2.2), and justify why the resulting process is also an NSM. Third, we derive an implicit lower CS for  $(\tilde{\mu}_t^*)_{t=1}^\infty$ . Fourth and finally, we compute a closed-form lower bound for the implicit CS.

**Step 1: Constructing the  $\lambda$ -indexed NSM** Similar to the proof of Theorem 7.2.2, let  $(G_t)_{t=1}^\infty$  be an infinite sequence of i.i.d. standard Gaussian random variables, and let  $S_t := \sum_{i=1}^t \sigma_i G_i$ . Then, we have that for any  $\lambda \in \mathbb{R}$ ,

$$M_t(\lambda) := \exp \left\{ \lambda S_t - t\tilde{\sigma}_t^2 \lambda^2 / 2 \right\}, \quad (7.226)$$

forms an NSM with respect to the filtration given by  $\mathcal{F}_t := \sigma(G_1^t)$ .

**Step 2: Mixing over  $\lambda \in (0, \infty)$  to obtain a mixture NSM** Let us now construct a one-sided sub-Gaussian mixture NSM. First, note that the mixture of an NSM with respect to a probability density is itself an NSM [217, 124] and is a simple consequence of Fubini's theorem. For our purposes, we will consider the density of a *folded* Gaussian distribution with location zero and scale  $\rho^2$ . In particular, if  $\Lambda \sim N(0, \rho^2)$ , let  $\Lambda_+ := |\Lambda|$  be the folded Gaussian. Then  $\Lambda_+$  has a probability density function  $f_{\rho^2}^+(\lambda)$  given by

$$f_{\rho^2}^+(\lambda) := \mathbb{1}(\lambda > 0) \frac{2}{\sqrt{2\pi\rho^2}} \exp \left\{ \frac{-\lambda^2}{2\rho^2} \right\}. \quad (7.227)$$

Note that  $f_{\rho^2}^+$  is simply the density of a mean-zero Gaussian with variance  $\rho^2$ , but truncated from below by zero, and multiplied by two to ensure that  $f_{\rho^2}^+(\lambda)$  integrates to one.

Then, since mixtures of NSMs are themselves NSMs, the process  $(M)_{t=1}^\infty$  given by

$$M_t := \int_\lambda M_t(\lambda) f_{\rho^2}^+(\lambda) d\lambda \quad (7.228)$$

is an NSM. We will now find a closed-form expression for  $M_t$ . Some of the algebraic steps are the same as those in the proof of Theorem 7.2.2, but we repeat them here for completeness. Writing out the definition of  $M_t$ , we have

$$M_t := \int_{\lambda \in \mathbb{R}} \exp \left\{ \lambda S_t - t\tilde{\sigma}_t^2 \lambda^2 / 2 \right\} f_{\rho^2}^+(\lambda) d\lambda \quad (7.229)$$

$$= \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \lambda S_t - t\tilde{\sigma}_t^2 \lambda^2 / 2 \right\} \frac{2}{\sqrt{2\pi\rho^2}} \exp \left\{ \frac{-\lambda^2}{2\rho^2} \right\} d\lambda \quad (7.230)$$

$$= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \lambda S_t - t\tilde{\sigma}_t^2 \lambda^2 / 2 \right\} \exp \left\{ \frac{-\lambda^2}{2\rho^2} \right\} d\lambda \quad (7.231)$$

$$= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \lambda S_t - \frac{\lambda^2(t\rho^2\tilde{\sigma}_t^2 + 1)}{2\rho^2} \right\} d\lambda \quad (7.232)$$

$$= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \frac{-\lambda^2(t\rho^2\tilde{\sigma}_t^2 + 1) + 2\lambda\rho^2 S_t}{2\rho^2} \right\} d\lambda \quad (7.233)$$

$$= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \underbrace{\exp \left\{ \frac{-a(\lambda^2 - \frac{b}{a}2\lambda)}{2\rho^2} \right\}}_{(*)} d\lambda, \quad (7.234)$$

where we have set  $a := t\rho^2\tilde{\sigma}_t^2 + 1$  and  $b := \rho^2 S_t$ . Completing the square in  $(*)$ , we have that

$$\exp \left\{ \frac{-a(\lambda^2 - \frac{b}{a}2\lambda)}{2\rho^2} \right\} = \exp \left\{ \frac{-\lambda^2 + 2\lambda\frac{b}{a} + (\frac{b}{a})^2 - (\frac{b}{a})^2}{2\rho^2/a} \right\} \quad (7.235)$$

$$= \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} + \frac{a(b/a)^2}{2\rho^2} \right\} \quad (7.236)$$

$$= \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} \right\} \exp \left\{ \frac{b^2}{2a\rho^2} \right\}. \quad (7.237)$$

Plugging this back into our derivation of  $M_t$  and multiplying the entire quantity by  $a^{-1/2}/a^{-1/2}$ , we have

$$M_t = \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \underbrace{\exp \left\{ \frac{-a(\lambda^2 + \frac{b}{a}2\lambda)}{2\rho^2} \right\}}_{(*)} d\lambda \quad (7.238)$$

$$= \frac{2}{\sqrt{2\pi\rho^2}} \int_{\lambda} \mathbb{1}(\lambda > 0) \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} \right\} \exp \left\{ \frac{b^2}{2a\rho^2} \right\} d\lambda \quad (7.239)$$

$$= \frac{2}{\sqrt{a}} \exp \left\{ \frac{b^2}{2a\rho^2} \right\} \underbrace{\int_{\lambda} \mathbb{1}(\lambda > 0) \frac{1}{\sqrt{2\pi\rho^2/a}} \exp \left\{ \frac{-(\lambda - b/a)^2}{2\rho^2/a} \right\} d\lambda}_{(\star\star)} . \quad (7.240)$$

Now, notice that  $(\star\star) = \mathbb{P}(N(b/a, \rho^2/a) \geq 0)$ , which can be rewritten as  $\Phi(b/\rho\sqrt{a})$ , where  $\Phi$  is the CDF of a standard Gaussian. Putting this all together and plugging in  $a = t\rho^2\tilde{\sigma}_t^2 + 1$  and  $b = \rho^2S_t$ , we have the following expression for  $M_t$ ,

$$\begin{aligned} M_t &= \frac{2}{\sqrt{a}} \exp \left\{ \frac{b^2}{2a\rho^2} \right\} \Phi \left( \frac{b}{\rho\sqrt{a}} \right) \\ &= \frac{2}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \exp \left\{ \frac{\rho^4 S_t^2}{2(t\rho^2\tilde{\sigma}_t^2 + 1)\rho^2} \right\} \Phi \left( \frac{\rho^2 S_t}{\rho\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \right) \\ &= \frac{2}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \exp \left\{ \frac{\rho^2 S_t^2}{2(t\rho^2\tilde{\sigma}_t^2 + 1)} \right\} \Phi \left( \frac{\rho^2 S_t}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \right). \end{aligned} \quad (7.241)$$

**Step 3: Deriving a  $(1 - \alpha)$ -lower CS  $(L'_t)_{t=1}^\infty$  for  $(\tilde{\mu}_t)_{t=1}^\infty$**  Now that we have computed the mixture NSM  $(M)_{t=1}^\infty 0$ , we apply Ville's inequality to it and "invert" a family of processes — one of which is  $(M_t)_{t=1}^\infty$  — to obtain an *implicit* lower CS (we will further derive an *explicit* lower CS in Step 4).

First, let  $(m)_{t=1}^\infty 1$  be an arbitrary real-valued process — i.e. not necessarily equal to  $(\mu)_{t=1}^\infty 1$  — and define their running average  $\tilde{m}_t := \frac{1}{t} \sum_{i=1}^t m_i$ . Define the partial sum process in terms of  $(\tilde{m})_{t=1}^\infty 1$ ,

$$S_t(\tilde{m}_t) := S_t + t\tilde{\mu}_t - t\tilde{m}_t$$

and the resulting nonnegative process,

$$M_t(\tilde{m}_t) := \frac{2}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \exp \left\{ \frac{\rho^2 S_t(\tilde{m}_t)^2}{2(t\rho^2\tilde{\sigma}_t^2 + 1)} \right\} \Phi \left( \frac{\rho S_t(\tilde{m}_t)}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \right). \quad (7.242)$$

Notice that if  $\tilde{m}_t = \tilde{\mu}_t$ , then  $S_t(\tilde{\mu}_t) = S_t = \sum_{i=1}^t \sigma_i G_i$  and  $M_t(\tilde{\mu}_t) = M_t$  from Step 2. Importantly,  $(M_t(\tilde{\mu}_t))_{t=0}^\infty$  is an NSM. Indeed, by Ville's inequality, we have

$$\mathbb{P}(\exists t : M_t(\tilde{\mu}_t) \geq 1/\alpha) \leq \alpha. \quad (7.243)$$

We will now "invert" this family of processes to obtain an implicit lower boundary given by

$$L'_t := \inf\{\tilde{\mu}_t : M_t(\tilde{\mu}_t) < 1/\alpha\}, \quad (7.244)$$

and justify that  $(L'_t)_{t=1}^\infty$  is indeed a lower  $(1 - \alpha)$ -CS for  $\tilde{\mu}_t$ . Writing out the probability of miscoverage at any time  $t$ , we have

$$\mathbb{P}(\exists t : \tilde{\mu}_t < L'_t) \equiv \mathbb{P} \left( \exists t : \tilde{\mu}_t < \inf_{\tilde{m}_t} \{M_t(\tilde{m}_t) < 1/\alpha\} \right) \quad (7.245)$$

$$= \mathbb{P}(\exists t : M_t(\tilde{\mu}_t) \geq 1/\alpha) \quad (7.246)$$

$$\leq \alpha, \quad (7.247)$$

where the last line follows from Ville's inequality applied to  $(M_t(\tilde{\mu}_t))_{t=0}^{\infty}$ . In particular,  $L'_t$  forms a  $(1 - \alpha)$ -lower CS, meaning

$$\mathbb{P}(\forall t, \tilde{\mu}_t \geq L'_t) \geq 1 - \alpha.$$

**Step 4: Obtaining a closed-form lower bound  $(\tilde{L}'_t)_{t=1}^{\infty}$  for  $(L'_t)_{t=1}^{\infty}$**  The lower CS of Step 3 is simple to evaluate via line- or grid-searching, but a closed-form expression may be desirable in practice, and for this we can compute a sharp lower bound on  $L'_t$ .

First, take notice of two key facts:

- (a) When  $\tilde{m}_t = S_t/t + \tilde{\mu}_t$ , we have that  $S_t(\tilde{m}_t) = 0$  and hence  $M_t(\tilde{m}_t) < 1$ , and
- (b)  $S_t(\tilde{m}_t)$  is a strictly decreasing function of  $\tilde{m}_t \leq S_t/t + \tilde{\mu}_t$ , and hence so is  $M_t(\tilde{m}_t)$ .

Property (a) follows from the fact that  $\Phi(0) = 1/2$ , and that  $\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1} > 1$  for any  $\rho > 0$ . Property (b) follows from property (a) combined with the definitions of  $S_t(\cdot)$ ,

$$S_t(\tilde{m}_t) := S_t + t\tilde{\mu}_t - t\tilde{m}_t$$

and of  $M_t(\cdot)$ ,

$$M_t(\tilde{m}_t) := \frac{2}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \exp \left\{ \frac{\rho^2 S_t(\tilde{m}_t)^2}{2(t\rho^2\tilde{\sigma}_t^2 + 1)} \right\} \Phi \left( \frac{\rho S_t(\tilde{m}_t)}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \right),$$

In particular, by facts (a) and (b), the infimum in (7.244) must be attained when  $S_t(\cdot) \geq 0$ . That is,

$$S_t(L'_t) \geq 0. \quad (7.248)$$

Using (7.248) combined with the inequality  $1 - \Phi(x) \leq \exp\{-x^2/2\}$  (a straightforward consequence of the Cramér-Chernoff technique), we have the following lower bound on  $M_t(L'_t)$ :

$$M_t(L'_t) = \frac{2}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \exp \left\{ \frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2\tilde{\sigma}_t^2 + 1)} \right\} \Phi \left( \frac{\rho S_t(L'_t)}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \right) \quad (7.249)$$

$$\geq \frac{2}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \exp \left\{ \frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2\tilde{\sigma}_t^2 + 1)} \right\} \left( 1 - \exp \left\{ -\frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2\tilde{\sigma}_t^2 + 1)} \right\} \right) \quad (7.250)$$

$$= \frac{2}{\sqrt{t\rho^2\tilde{\sigma}_t^2 + 1}} \left( \exp \left\{ \frac{\rho^2 S_t(L'_t)^2}{2(t\rho^2\tilde{\sigma}_t^2 + 1)} \right\} - 1 \right) \quad (7.251)$$

$$=: \tilde{M}_t(L'_t). \quad (7.252)$$

Finally, the above lower bound on  $M_t(L'_t)$  implies that  $1/\alpha \geq M_t(L'_t) \geq \tilde{M}_t(L'_t)$  which yields the following lower bound on  $L'_t$ :

$$\tilde{M}_t(L'_t) \leq 1/\alpha \iff \frac{2}{\sqrt{tp^2\tilde{\sigma}_t^2 + 1}} \left( \exp \left\{ \frac{\rho^2 S_t(L'_t)^2}{2(tp^2\tilde{\sigma}_t^2 + 1)} \right\} - 1 \right) \leq 1/\alpha \quad (7.253)$$

$$\iff \exp \left\{ \frac{\rho^2 S_t(L'_t)^2}{2(tp^2\tilde{\sigma}_t^2 + 1)} \right\} \leq 1 + \frac{\sqrt{tp^2\tilde{\sigma}_t^2 + 1}}{2\alpha} \quad (7.254)$$

$$\iff \frac{\rho^2 S_t(L'_t)^2}{2(tp^2\tilde{\sigma}_t^2 + 1)} \leq \log \left( 1 + \frac{\sqrt{tp^2\tilde{\sigma}_t^2 + 1}}{2\alpha} \right) \quad (7.255)$$

$$\iff S_t(L'_t) \leq \sqrt{\frac{2(tp^2\tilde{\sigma}_t^2 + 1)}{\rho^2} \log \left( 1 + \frac{\sqrt{tp^2\tilde{\sigma}_t^2 + 1}}{2\alpha} \right)} \quad (7.256)$$

$$\iff \sum_{i=1}^t \sigma_i G_i + t\tilde{\mu}_t - tL'_t \leq \sqrt{\frac{2(tp^2\tilde{\sigma}_t^2 + 1)}{\rho^2} \log \left( 1 + \frac{\sqrt{tp^2\tilde{\sigma}_t^2 + 1}}{2\alpha} \right)} \quad (7.257)$$

$$\iff tL'_t \geq \sum_{i=1}^t (\sigma_i G_i + \mu_i) - \sqrt{\frac{2(tp^2\tilde{\sigma}_t^2 + 1)}{\rho^2} \log \left( 1 + \frac{\sqrt{tp^2\tilde{\sigma}_t^2 + 1}}{2\alpha} \right)} \quad (7.258)$$

$$\iff L'_t \geq \underbrace{\frac{1}{t} \sum_{i=1}^t (\sigma_i G_i + \mu_i) - \sqrt{\frac{2(tp^2\tilde{\sigma}_t^2 + 1)}{(tp)^2} \log \left( 1 + \frac{\sqrt{tp^2\tilde{\sigma}_t^2 + 1}}{2\alpha} \right)}}_{L_t^*}, \quad (7.259)$$

and hence  $\mathbb{P}(\forall t \in \mathbb{N}, \tilde{\mu}_t \geq L_t^*) \geq 1 - \alpha$ .  $\square$

*Proof of Proposition 7.B.1.* In this case, the data  $(Y_t)_{t=1}^\infty$  are i.i.d., and hence we have that  $\sigma_1 = \sigma_2 = \dots = \sigma$  and  $\mu_1 = \mu_2 = \dots = \mu$ . First, notice that if we define  $\beta = \rho\sigma$ , then we can write  $L_t^*$  as

$$L_t^* := \frac{\sigma}{t} \sum_{i=1}^t (G_i + \mu) - \sigma \sqrt{\frac{2(t\beta^2 + 1)}{(t\beta)^2} \log \left( 1 + \frac{\sqrt{t\beta^2 + 1}}{2\alpha} \right)}. \quad (7.260)$$

Now, by the strong approximation of Strassen [248], we have that

$$L_t^* = \frac{1}{t} \sum_{i=1}^t Y_i - \sigma \sqrt{\frac{2(t\beta^2 + 1)}{(t\beta)^2} \log \left( 1 + \frac{\sqrt{t\beta^2 + 1}}{2\alpha} \right)} + \varepsilon'_t \quad (7.261)$$

where  $\varepsilon'_t = o(\sqrt{\log \log t/t})$ . Writing the above in terms of an empirical standard deviation  $\hat{\sigma}_t$ , we have by the proof of Lemma 7.A.1 that

$$L_t^* = \frac{1}{t} \sum_{i=1}^t Y_i - \hat{\sigma}_t \sqrt{\frac{2(t\beta^2 + 1)}{(t\beta)^2} \log \left(1 + \frac{\sqrt{t\beta^2 + 1}}{2\alpha}\right)} + \varepsilon_t \quad (7.262)$$

where  $\varepsilon_t = o(\sqrt{\log \log t/t})$  as well, and hence

$$\frac{1}{t} \sum_{i=1}^t Y_i - \hat{\sigma}_t \sqrt{\frac{2(t\beta^2 + 1)}{(t\beta)^2} \log \left(1 + \frac{\sqrt{t\beta^2 + 1}}{2\alpha}\right)} \quad (7.263)$$

forms a lower  $(1 - \alpha)$ -AsympCS for  $\mu$  with approximation rate  $\varepsilon_t$ . This completes the proof.<sup>10</sup>

□

*Proof of Proposition 7.B.2.* Similar to the proof of Proposition 7.2.2, Lemma 7.A.2 yields the following strong invariance principle

$$\sum_{i=1}^t Y_i = \sum_{i=1}^t \sigma_i(G_i + \mu_i) + o(V_t^{3/8} \log V_t). \quad (7.264)$$

Therefore, with probability at least  $(1 - \alpha)$ ,

$$\forall t \geq 1, \tilde{\mu}_t \geq \frac{1}{t} \sum_{i=1}^t Y_i - \sqrt{\frac{2(t\rho^2 \tilde{\sigma}_t^2 + 1)}{(t\rho)^2} \log \left(1 + \frac{\sqrt{t\rho^2 \tilde{\sigma}_t^2 + 1}}{2\alpha}\right)} + o(V_t^{3/8} \log V_t/t).$$

In particular, we have that

$$\left( \hat{\mu}_t \pm \sqrt{\frac{2(t\tilde{\sigma}_t^2 \rho^2 + 1)}{t^2 \rho^2} \log \left(1 + \frac{\sqrt{t\tilde{\sigma}_t^2 \rho^2 + 1}}{2\alpha}\right)} \right) \quad (7.265)$$

forms a  $(1 - \alpha)$ -AsympCS for  $\tilde{\mu}_t$ . The derivation of an analogous lower AsympCS in terms of the empirical variance  $\hat{\sigma}_t^2$  proceeds similarly to Step 3 of the proof of Proposition 7.2.2. In particular, we get that

$$\hat{\mu}_t - \tilde{\mathfrak{B}}_t^* := \hat{\mu}_t - \sqrt{\frac{2(t\hat{\sigma}_t^2 \rho^2 + 1)}{t^2 \rho^2} \log \left(1 + \frac{\sqrt{t\hat{\sigma}_t^2 \rho^2 + 1}}{2\alpha}\right)} + o\left(\frac{\sqrt{V_t \log V_t}}{t}\right)$$

---

<sup>10</sup>While we wrote this final bound in terms of  $\beta$ , we leave the statement of the original result in terms of  $\rho$  to maintain consistency with other boundaries throughout the chapter. The change from  $\beta$  to  $\rho$  and back is entirely cosmetic, and does not affect the interpretation of the final result.

forms a nonasymptotic  $(1 - \alpha)$ -CS for  $\tilde{\mu}_t$ , meaning  $\mathbb{P} \left( \forall t \in \mathbb{N}, \tilde{\mu}_t \geq \hat{\mu}_t - \tilde{\mathfrak{B}}_t^* \right) \leq \alpha$ . Combined with Condition L-3, we have that

$$\hat{\mu}_t - \tilde{\mathfrak{B}}_t := \hat{\mu}_t - \sqrt{\frac{2(t\hat{\sigma}_t^2\rho^2 + 1)}{t^2\rho^2} \log \left( 1 + \frac{\sqrt{t\hat{\sigma}_t^2\rho^2 + 1}}{2\alpha} \right)}$$

forms a  $(1 - \alpha)$ -AsympCS for  $\tilde{\mu}_t$  since  $\tilde{\mathfrak{B}}_t \asymp \sqrt{V_t \log V_t}/t$ . This completes the proof.  $\square$

### 7.B.2 Optimizing Robbins' normal mixture for $(t, \alpha)$

In this section, we outline how one can choose  $\rho$  to optimize the boundary  $\bar{\mathfrak{B}}_t$  in Theorem 7.2.2 for a specific time  $t^*$  and type-I error level  $\alpha \in (0, 1)$ .<sup>11</sup> We will outline both the (computationally inexpensive) exact solution, and the closed-form approximate solution. Note that the derivations that follow are essentially the same as those in Howard et al. [125, Section 3.5] but we repeat them here to keep our results self-contained.

**The exact solution** Let  $W_{-1}$  be the lower branch of the Lambert  $W$  function [65]. Then,

$$\operatorname{argmin}_{\rho > 0} \bar{\mathfrak{B}}_{t^*}(\alpha) = \sqrt{\frac{-W_{-1}(-\alpha^2 \exp \{-1\}) - 1}{t^*}}. \quad (7.266)$$

*Proof.* Consider the boundary in Theorem 7.2.2 at time  $t$ ,

$$\bar{\mathfrak{B}}_t(\alpha) := \sqrt{\frac{t\rho^2 + 1}{t^2\rho^2} \log \left( \frac{t\rho^2 + 1}{\alpha^2} \right)}.$$

Defining  $x := \rho^2$  and after some simple algebra, notice that

$$\operatorname{argmin}_{\rho > 0} \bar{\mathfrak{B}}_t(\alpha) = \sqrt{\operatorname{argmin}_{x > 0} f(x)},$$

where  $f(x) := \frac{tx + 1}{t^2x} \log \left( \frac{tx + 1}{\alpha^2} \right)$ .

Notice that  $\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow \infty} f(x) = \infty$  and thus if we find that  $df/dx = 0$  has exactly one positive solution, we know that it must be the minimizer of  $f$ .

To that end, it is straightforward to show that

$$\frac{df}{dx} = -\frac{1}{t^2x^2} \log \left( \frac{tx + 1}{\alpha^2} \right) + \frac{1}{tx}.$$

---

<sup>11</sup>We will discuss choosing  $\rho$  for the two-sided AsympCS in Theorem 7.2.2 but for the one-sided AsympCSs of Section 7.B.1, we suggest repeating the same argument but with  $\alpha$  replaced by  $2\alpha$ .

Setting the above to 0, we obtain

$$\alpha^2 \exp\{tx\} = tx + 1,$$

which, after some algebra, can be rewritten as

$$-\alpha^2 \exp\{-1\} = -(tx + 1) \exp\{-(tx + 1)\} \quad (7.267)$$

Notice that if we rewrite  $y := -(tx + 1)$ , we have that  $y = W_{-1}(-\alpha^2 \exp\{-1\})$  where  $W_{-1}$  is the lower branch of the Lambert  $W$  function. Furthermore,  $y = W_{-1}(z)$  only has a solution if  $z \geq -e^{-1}$ , requiring that  $\alpha^2 \leq 1$ , which we have trivially by the definition of  $\alpha \in (0, 1)$ . In summary, we have that

$$\underset{\rho > 0}{\operatorname{argmin}} \bar{\mathfrak{B}}_{t^*}(\alpha) = \sqrt{\frac{-W_{-1}(-\alpha^2 \exp\{-1\}) - 1}{t^*}}. \quad (7.268)$$

This completes the proof.  $\square$

**An approximate solution** We can derive a closed-form approximation to (7.266) by considering the Taylor series expansion to the Lambert  $W$  function [65],

$$W_{-1}(z) = \log(-z) - \log(-\log(-z)) + o(1). \quad (7.269)$$

Replacing  $W_{-1}(z)$  by  $\log(-z) - \log(-\log(-z))$  in (7.266), we obtain the following approximate solution,

$$\rho'(t^*) := \sqrt{\frac{-2 \log \alpha + \log(-2 \log \alpha + 1)}{t^*}}. \quad (7.270)$$

In practice, we find that using (7.270) over (7.266) has negligible downstream effects on the resulting CSs, but both are inexpensive to compute. Moreover, notice that  $\rho'(t^*)$  is quite similar to  $\sqrt{2 \log(1/\alpha)/t^*}$ , which is precisely what one would choose when sharpening a sub-Gaussian confidence interval based on the Cramér-Chernoff technique for a fixed sample size  $t^*$ .

### 7.B.3 Time-uniform convergence in probability is equivalent to almost sure convergence

In Theorems 7.2.2, 7.3.1, and 7.3.2, we justified the asymptotic validity of our AsympCSs by showing that the approximation error

$$\varepsilon_t \xrightarrow{a.s.} 0 \quad (7.271)$$

at a particular rate. At first glance, this may seem like a slightly stronger statement than required since we only need the approximation error  $\varepsilon_t$  to vanish *time-uniformly in probability*:

$$\sup_{k \geq t} |\varepsilon_k| \xrightarrow{p} 0. \quad (7.272)$$

It turns out that (7.271) and (7.272) are equivalent. This is not new, but we present a proof for completeness.

**Proposition 7.B.3.** *Let  $(X_n)_{n=1}^{\infty}$  be a sequence of random variables. Then,*

$$X_n \xrightarrow{a.s.} 0 \iff \sup_{k \geq n} |X_k| \xrightarrow{p} 0.$$

*Proof.* First, we prove ( $\implies$ ). By the continuous mapping theorem,  $|X_n| \xrightarrow{a.s.} 0$ . Thus,

$$1 = \mathbb{P} \left( \lim_n |X_n| = 0 \right) \leq \mathbb{P} \left( \limsup_n |X_n| = 0 \right) \leq 1. \quad (7.273)$$

In other words,  $\sup_{k \geq n} |X_k| \xrightarrow{a.s.} 0$ , which implies  $\sup_{k \geq n} |X_k| \xrightarrow{p} 0$ .

Now, consider ( $\impliedby$ ). Suppose for the sake of contradiction that  $\mathbb{P}(\lim_n |X_n| = 0) < 1$ . Then with some probability  $\delta > 0$ , we have that  $\lim_n |X_n| \neq 0$ , meaning there exists some  $\epsilon > 0$  such that  $|X_k| > \epsilon$  for some  $k \geq n$  no matter how large  $n$  is. In other words,

$$\delta < \mathbb{P} \left( \lim_{n \rightarrow \infty} \sup_{k \geq n} |X_k| > \epsilon \right) \quad (7.274)$$

$$\leq \mathbb{P} \left( \sup_{k \geq n} |X_k| > \epsilon \right) \text{ for any } n \geq 1. \quad (7.275)$$

In particular,  $\mathbb{P}(\sup_{k \geq n} |X_k| > \epsilon) \rightarrow 0$ , which would imply that  $\sup_{k \geq n} |X_k| \not\xrightarrow{p} 0$ , a contradiction. This completes the proof.  $\square$

#### 7.B.4 Comparing AsympCSs to group-sequential repeated confidence intervals

Another approach to sequential inference is via so-called *group-sequential trials* and *repeated confidence intervals* (see the now-classical text of Jennison and Turnbull [134]). In brief, group-sequential trials allow the analyst to peek at CIs at certain prespecified times  $(t_1, t_2, \dots, t_K)$ . They differ from the “anytime-valid” approach of CSs and (and AsympCSs) in that they do not permit continuous monitoring (i.e. updating of inferences for each new data point collected) and require a (fixed, data-independent) maximum sample size  $t_K$ . In this way, they can be thought of as fixed-time CLT-based CIs that permit a fixed number of peeks prior to  $t_K$  for early stopping. AsympCSs by contrast can be continuously monitored indefinitely, allowing the study to continue for as long as needed by the analyst.

Here, we provide a simulation comparing the widths and cumulative miscoverage rates of AsympCSs to two popular repeated CIs from the group-sequential literature — namely those of Pocock [204] and O’Brien and Fleming [194] and display the classical CLT-based CI alongside them for reference (Figure 7.10). Perhaps unsurprisingly, group-sequential methods (and especially those using the boundary of Pocock [204]) tend to lie somewhere in between

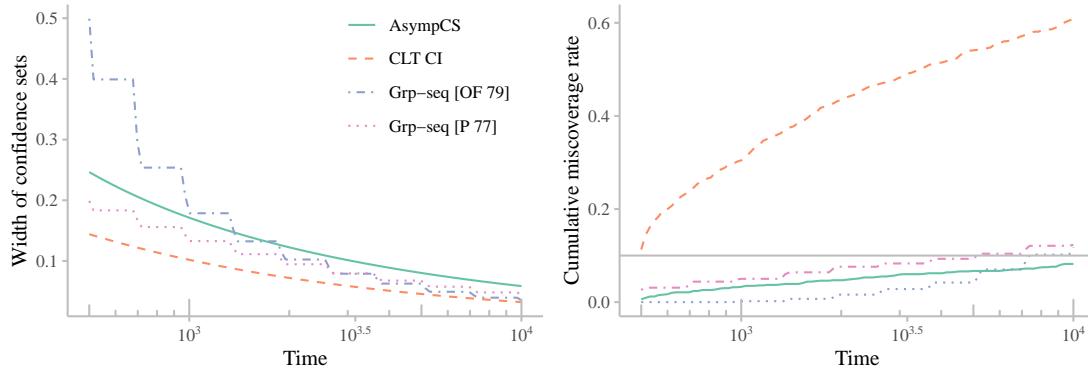


Figure 7.10: A comparison of 90% confidence sets for the average treatment effect in a simulated sequential experiment using (a) the AsympCS of Theorem 7.3.1, (b) the CLT-based CI, and the group-sequential repeated CIs of (c) O'Brien and Fleming [194], and (d) Pocock [204]. The group-sequential methods used a start time of  $t_1 = 500$  and a maximal sample size of  $t_{10} = 10,000$  with 10 logarithmically-spaced peeking times. The jagged drops in their widths correspond to CIs being updated at each peek. Notice in the left-hand side plot the group-sequential repeated CIs lie somewhere in between the AsympCS and the CLT CI (especially those using the Pocock boundary). Furthermore, notice in the right-hand side plot that the cumulative miscoverage rate of the AsympCS lies below  $\alpha = 0.1$  while the group-sequential CIs slightly exceed it. Nevertheless, those of the CLT-based CI's miscoverage rates greatly exceed the other three, diverging quickly beyond  $\alpha = 0.1$  and exceeding 0.6 by  $t_{10} = 10,000$ .

CLT CIs and AsympCSs. This added tightness over AsympCSs of course comes at the cost of less flexibility, especially in regards to continuous monitoring and unbounded time horizons discussed above.

### 7.B.5 The Lyapunov-type condition implies the Lindeberg-type condition

Here, we give a brief proof of the fact that the Lyapunov-type condition discussed in the paragraph following Proposition 7.2.2 indeed implies Condition L-2.

*Proof.* Suppose that the following Lyapunov-type condition holds with some  $\delta > 0$ :

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}(|X_t|^{2+\delta} | Y_1^{t-1})}{\sqrt{V_t}^{2+\delta}} < \infty, \quad (7.276)$$

and we want to show that Condition L-2 holds for some  $\kappa \in (0, 1)$ . Indeed, set  $\kappa := \frac{1+\delta/2}{1+\delta}$ . Note that  $1 \leq V_t^{-\kappa\delta/2} |X_t|^\delta$  whenever  $X_t^2 > V_t^\kappa$ , and hence

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}(X_t^2 \mathbb{1}\{X_t^2 > V_t^\kappa\} | Y_1^{t-1})}{V_t^\kappa} \leq \sum_{t=1}^{\infty} \frac{\mathbb{E}(V_t^{-\kappa\delta/2} |X_t|^{2+\delta} \mathbb{1}\{X_t^2 > V_t^\kappa\} | Y_1^{t-1})}{V_t^\kappa} \quad (7.277)$$

$$= \sum_{t=1}^{\infty} \frac{\mathbb{E}(|X_t|^{2+\delta} \mathbb{1}\{X_t^2 > V_t^\kappa\} | Y_1^{t-1})}{V_t^{\kappa(1+\delta/2)}} \quad (7.278)$$

$$= \sum_{t=1}^{\infty} \frac{\mathbb{E}(|X_t|^{2+\delta} | Y_1^{t-1})}{\sqrt{V_t}^{2+\delta}} < \infty, \quad (7.279)$$

which completes the proof.  $\square$

### 7.B.6 On martingale AsympCSs for running average treatment effects

In Remark 20, we mentioned that it is possible to derive AsympCSs for the running average treatment effect  $\tilde{\psi}_t$  in randomized experiments without sample splitting. The following proposition is a corollary of Proposition 7.2.2 applied to a particular sequence of estimated influence functions. In what follows, we use  $R_t$  for the outcome of subject  $t$  instead of  $Y_t$  so that it is not confused with the appearance of  $Y_t$  in Conditions L-1, L-2, and L-3 which we will refer to.

**Proposition 7.B.4.** *Let  $(Z_t)_{t=1}^{\infty} \equiv (X_t, A_t, R_t)_{t=1}^{\infty}$  be independent triplets as in Theorem 7.3.3 and let  $\hat{\mu}_t^a$  be an estimator for  $\mu^a$  derived from  $Z_1^{t-1}; a \in \{0, 1\}$ . If Conditions L-1, L-2, and L-3 hold but with  $Y_t$  replaced by  $f_t(Z_t)$  everywhere given by*

$$f_t(Z_t) := \{\hat{\mu}_t^1(X_t) - \hat{\mu}_t^0(X_t)\} + \left( \frac{A_t}{\pi(X_t)} - \frac{1-A_t}{1-\pi(X_t)} \right) \{R_t - \bar{\mu}^a(X_t)\}. \quad (7.280)$$

Then for any  $\rho > 0$ ,

$$\tilde{C}_t := \frac{1}{t} \sum_{i=1}^t f_i(Z_i) \pm \sqrt{\frac{t\hat{\sigma}_t^2\rho^2 + 1}{t^2\rho^2} \log\left(\frac{t\hat{\sigma}_t^2\rho^2 + 1}{\alpha^2}\right)} \quad (7.281)$$

forms a  $(1 - \alpha)$ -AsympCS for the running average treatment effect  $\tilde{\psi}_t$ .

For certain practical purposes, however, we still recommend sequential sample-splitting or cross fitting as described in Section 7.3 since the conditions of Proposition 7.B.4 are somewhat less transparent.

### 7.B.7 Explicit connections between “delayed start” boundaries and other works

In Section 7.2.5, we elaborated on certain connections between Bibaut et al. [34] and the results in Proposition 7.2.3 and its corollary in (7.31). In Claims 7.B.1 and 7.B.2 we explicitly show that  $\tilde{C}_t^{\text{DS}*}$  is almost surely tighter than the bound found in version 1 and identical to that found in version 2 of Bibaut et al. [34]. In what follows, we let  $C_t^{\text{Bv1}}$  and  $C_t^{\text{Bv2}}$  denote the bounds found in versions 1 and 2 of Bibaut et al. [34], respectively.

**Claim 7.B.1** ( $\tilde{C}_t^{\text{DS}*}$  is almost surely tighter than  $C_t^{\text{Bv1}}(\lambda)$ ). *Let  $(X_t)_{t=1}^\infty$  be i.i.d. random variables with unknown mean  $\mu$  and known variance  $\sigma$ . Let  $\hat{\mu}_t := \frac{1}{t} \sum_{i=1}^t X_i$  be the sample mean and  $\alpha \in (0, 1)$  the desired type-I error level. For any fixed  $\lambda > 0$ , consider  $C_t^{\text{Bv1}}(\lambda)$  from version 1 of Bibaut et al. [34, Example 1] given by*

$$C_t^{\text{Bv1}}(\lambda) := \left[ \hat{\mu}_t \pm \sigma t^{-1} \cdot (t + \lambda)^{1/2} (\log(t + \lambda) - \log m - 2 \log \tilde{\alpha}(\alpha))^{1/2} \right], \quad (7.282)$$

where  $\tilde{\alpha}(\alpha)$  is the unique solution to  $2\tilde{\alpha}(\alpha)\sqrt{-\log \tilde{\alpha}(\alpha)/\pi} + 2 \left[ 1 - \Phi\left(\sqrt{-2 \log \tilde{\alpha}(\alpha)}\right) \right] = \alpha$ .<sup>12</sup> Then,  $C_t^{\text{Bv1}}(\lambda)$  in (7.282) can be equivalently re-written as

$$C_t^{\text{Bv1}}(\lambda) \equiv \left[ \hat{\mu}_t \pm \sigma \sqrt{\frac{t + \lambda}{t^2} \left( a^2 + \log\left(\frac{t + \lambda}{m}\right) \right)} \right], \quad (7.283)$$

where  $a \geq 0$  is the unique solution to  $2(1 - \Phi(a) + a\phi(a)) = \alpha$ , and thus  $\tilde{C}_t^{\text{DS}*}$  is a strict subset of  $C_t^{\text{Bv1}}(\lambda)$  for any  $\lambda > 0$ .

*Proof.* Begin by defining  $b \geq 0$  as

$$b := \sqrt{-2 \log \tilde{\alpha}(\alpha)} \quad \text{and thus} \quad \tilde{\alpha}(\alpha) = \exp\{-b^2/2\}, \quad (7.284)$$

---

<sup>12</sup>Note that in version 1 of Bibaut et al. [34], Algorithm 1 includes slightly different constants (specifically some multiples of 2 are moved around), but what we have written here matches their code at [github.com/nathankallus/UniversalSPRT](https://github.com/nathankallus/UniversalSPRT) which we believe contains the correct bound. Note that in version 2 of Bibaut et al. [34], their constants agree with what is written here.

noting that the transformation  $x \mapsto \sqrt{-2 \log x}$  is a bijection for  $x \in [0, 1]$ . We will now show that  $b$  solves  $2(1 - \Phi(b) + b\phi(b)) = \alpha$  if and only if  $\tilde{\alpha}$  solves the equation given in the definition of  $C_t^{\text{Bv1}}$ . Indeed, writing out the equation that  $\tilde{\alpha}$  solves, we have

$$2\tilde{\alpha}\sqrt{-\log \tilde{\alpha}/\pi} + 2\left[1 - \Phi\left(\sqrt{-2 \log \tilde{\alpha}}\right)\right] = \alpha \quad (7.285)$$

$$\iff 2\exp\{-b^2/2\}\sqrt{b^2/(2\pi)} + 2[1 - \Phi(b)] = \alpha \quad (7.286)$$

$$\iff 2b\underbrace{\frac{1}{\sqrt{2\pi}}\exp\{-b^2/2\}}_{\equiv \phi(b)} + 2[1 - \Phi(b)] = \alpha \quad (7.287)$$

$$\iff 2[1 - \Phi(b) + b\phi(b)] = \alpha. \quad (7.288)$$

Now, it remains to show that  $C_t^{\text{Bv1}}(\lambda)$  can be written in the form given in (7.283). Indeed, writing out the boundary in (7.282), we have

$$\sigma t^{-1} \cdot (t + \lambda)^{1/2} (\log(t + \lambda) - \log m - 2 \log \tilde{\alpha})^{1/2} \quad (7.289)$$

$$= \sigma t^{-1} \cdot \sqrt{(t + \lambda)(\log(t + \lambda) - \log m + b^2)} \quad (7.290)$$

$$= \sigma \sqrt{\frac{t + \lambda}{t^2} \left( b^2 + \log\left(\frac{t + \lambda}{m}\right) \right)}, \quad (7.291)$$

and since  $b$  solves  $2[1 - \Phi(b) + b\phi(b)] = \alpha$  as demonstrated above, this completes the proof of the claim.  $\square$

**Claim 7.B.2** ( $\tilde{C}_t^{\text{DS}*}$  is equivalent to  $C_t^{\text{Bv2}}$ ). Consider  $\tilde{C}_t^{\text{DS}*}$  as before and consider  $C_t^{\text{Bv2}}$  as in version 2 of Bibaut et al. [34]:

$$C_t^{\text{Bv2}} := \left[ \hat{\mu}_t \pm \frac{\sigma}{t} \cdot \sqrt{t \cdot \left( -2 \log \tilde{\alpha} + \log\left(\frac{t}{m}\right) \right)} \right], \quad (7.292)$$

where  $\tilde{\alpha}$  solves  $h_1(-\log \tilde{\alpha}) = \alpha$  and  $h_1(x)$  is given by  $h_1(x) = 2\exp\{-x\}\sqrt{x/\pi} + 2(1 - \Phi(\sqrt{2x}))$ ;  $x \geq 0$ . Then  $C_t^{\text{Bv2}} = \tilde{C}_t^{\text{DS}*}$ .

*Proof.* Similar to the proof of Claim 7.B.1, we define  $b := \sqrt{-2 \log \tilde{\alpha}}$  and show that  $b$  solves  $2[1 - \Phi(b) + b\phi(b)] = \alpha$  if and only if  $\tilde{\alpha}$  solves  $h_1(-\log \tilde{\alpha}) = \alpha$ . Indeed,

$$h_1(-\log \tilde{\alpha}) = \alpha \quad (7.293)$$

$$\iff 2\exp\{\log \tilde{\alpha}\}\sqrt{-\log \tilde{\alpha}/\pi} + 2(1 - \Phi(\sqrt{-2 \log \tilde{\alpha}})) = \alpha \quad (7.294)$$

$$\iff 2\exp\{-b^2/2\}\sqrt{b/(2\pi)} + 2(1 - \Phi(\sqrt{b^2})) \quad (7.295)$$

$$\iff 2b\underbrace{\frac{1}{\sqrt{2\pi}}\exp\{-b^2/2\}}_{\equiv \phi(b)} + 2(1 - \Phi(b)) \quad (7.296)$$

$$\iff 2[1 - \Phi(b) + b\phi(b)]. \quad (7.297)$$

Finally, noting that  $b^2 = -2 \log \tilde{\alpha}$ , we have that

$$C_t^{\text{Bv2}} \equiv \left[ \hat{\mu}_t \pm \sigma \sqrt{\frac{1}{t} (b^2 + \log(t/m))} \right], \quad (7.298)$$

and since  $b$  solves  $2[1 - \Phi(b) + b\phi(b)] = \alpha$ , we have that  $C_t^{\text{Bv2}} = \tilde{C}_t^{\text{DS}\star}$  for each  $t$ , which completes the proof of the claim.  $\square$

### 7.8 A brief review of efficient estimators

For a detailed account of efficient estimation in semiparametric models, we refer readers to Bickel et al. [35], van der Vaart [261], van der Laan and Robins [257], Tsiatis [254] and Kennedy [153], but we provide a brief overview of their fundamental relevance to estimation of the ATE here.

A central goal of semiparametric efficiency theory is to characterize the set of *influence functions* of  $\psi$  – the summands found in sample averages forming consistent and asymptotically normal regular estimators of  $\psi$ . Of particular interest is finding the *efficient influence function* (EIF) as it is the one with the smallest variance (which itself acts as a semiparametric analogue of the Cramer-Rao lower bound), hence providing a benchmark for constructing optimal estimators, at least in an asymptotic local minimax sense. In the case of  $\psi$ , the (uncentered) EIF is given by

$$f(z) \equiv f(x, a, y) := \{\mu^1(x) - \mu^0(x)\} + \left( \frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)} \right) \{y - \mu^a(x)\}, \quad (7.299)$$

where  $\mu^a(x) := \mathbb{E}(Y | X = x, A = a)$  is the regression function among those treated at level  $a \in \{0, 1\}$  and  $\pi(x) := \mathbb{P}(A = 1 | X = x)$  is the propensity score (i.e. probability of treatment) for an individual with covariates  $x$ . In particular, this means that no estimator of  $\psi$  based on  $t$  observations can have asymptotic mean squared error smaller than  $\text{Var}(f(Z))/t$  without imposing additional assumptions.

Of course, the exact values of  $\eta := (\mu^1, \mu^0, \pi)$  are not known in general –  $\pi$  is only known in randomized experiments, but not observational studies, while  $(\mu^1, \mu^0)$  are typically not known in either. As such, we will need to replace these “nuisance functions”  $\eta$  with data-dependent estimates  $\hat{\eta}$ , but using the same data to construct estimators  $\hat{\eta}$  and  $\hat{f}$  for both  $\eta$  and  $f$  (sometimes referred to as “double-dipping”) complicates the analysis of the downstream estimator of  $\psi$ . A clever and simple remedy used throughout the semiparametric literature is to split the sample, whereby a random subset of the data are used to estimate  $\eta$ , while the remaining data are used to construct  $\hat{f}$ , greatly simplifying downstream analysis [226, 296, 56].

In a randomized experiment, the joint distribution of  $(X, Y)$  is unknown but the conditional distribution of  $A | X = x$  is known to be Bernoulli( $\pi(x)$ ) by design. In this case, our statistical model for  $Z$  is a proper semiparametric model, and hence there are infinitely many influence

functions, all of which take the form,

$$\bar{f}(z) \equiv \bar{f}(x, a, y) := \{\bar{\mu}^1(x) - \bar{\mu}^0(x)\} + \left( \frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)} \right) \{y - \bar{\mu}^a(x)\}, \quad (7.300)$$

where  $\bar{\mu}^a : \mathbb{R}^d \mapsto \mathbb{R}$  is any function. However, when the joint distribution of  $(X, A, Y)$  is left completely unspecified (such as in an observational study with unknown propensity scores), our statistical model for  $\mathbb{P}$  is nonparametric, and hence there is only one influence function, the EIF given in (7.299).

Not only does the EIF  $f(z)$  provide us with a benchmark against which to compare estimators, but it hints at the first step in deriving the most efficient estimator. Namely,  $\frac{1}{t} \sum_{i=1}^t f(Z_i)$  is a consistent estimator for  $\psi$  with asymptotic variance equal to the efficiency bound,  $\text{Var}(f)$  by construction. However,  $f(Z)$  depends on possibly unknown nuisance functions  $\eta := (\mu^1, \mu^0, \pi)$ . A natural next step would be to simply estimate  $\eta$  from the data  $(Z)_{t=1}^\infty$ . Crucially, it is possible to ensure that only a negligible amount of additional estimation error is incurred by replacing  $\eta$  by a data-dependent estimate  $\hat{\eta}_t$  — the essential technique here being sample splitting and cross fitting [226, 296, 56].

### 7.B.9 On the sharpness of AsympCSs using efficient influence functions

Consider the AsympCS of Theorem 7.3.2,

$$\hat{\psi}_t^\times \pm \underbrace{\sqrt{\widehat{\text{Var}}_t(\hat{f})}}_{(iii)} \cdot \underbrace{\sqrt{\frac{t\rho^2+1}{t^2\rho^2} \log\left(\frac{t\rho^2+1}{\alpha^2}\right)}}_{(i)} \quad \text{with rate } o\left(\underbrace{\sqrt{\frac{\log \log t}{t}}}_{(ii)}\right). \quad (7.301)$$

It is natural to wonder whether (7.301) can be tightened. In a certain sense, (7.301) inherits optimality from its three main components: (i) Robbins' normal mixture boundary, (ii) the approximation error rate, and (iii) the estimated standard deviation  $\sqrt{\widehat{\text{Var}}_t(\hat{f})}$  of the efficient influence function  $f$ .

**Term (i)** Starting with the width, we have that in the case of i.i.d. Gaussian data  $G_1, G_2, \dots \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ , Robbins' normal mixture confidence sequence [217] is obtained by first showing that

$$M_t(\mu) := \exp \left\{ \frac{\rho^2(\sum_{i=1}^t (G_i - \mu))^2}{2(t\rho^2 + 1)} \right\} (t\rho^2 + 1)^{-1/2}$$

is a nonnegative martingale starting at one, and hence by Ville's inequality [264],

$$\mathbb{P}(\exists t \geq 1 : M_t(\mu) \geq 1/\alpha) \leq \alpha.$$

The resulting confidence sequence  $\bar{C}_t^{\mathcal{N}}$  at each time  $t$  is defined as the set of  $m$  such that  $M_t(m) < 1/\alpha$ , i.e.  $\bar{C}_t^{\mathcal{N}} := \{m \in \mathbb{R} : M_t(m) < 1/\alpha\}$  and consequently,

$$\mathbb{P}(\exists t \geq 1 : \mu \notin \bar{C}_t^{\mathcal{N}}) = \mathbb{P}(\exists t \geq 1 : M_t(\mu) \geq 1/\alpha) \leq \alpha.$$

This inequality is extremely tight, since Ville's inequality almost holds with equality for nonnegative martingales. Technically, the paths of the martingale need to be continuous for equality to hold, which can only happen in continuous time (such as for a Wiener process). However, any deviation from equality only holds because of this "overshoot" and in practice, the error probability is almost exactly  $\alpha$ . This means that the normal mixture confidence sequence  $\bar{C}_t^{\mathcal{N}}$  cannot be uniformly tightened: any improvement for some times will necessarily result in looser bounds for others. For a precise characterization of this optimality for the (sub)-Gaussian case, see Howard et al. [125, Section 3.6], or Ramdas et al. [209] for a more general discussion of admissible confidence sequences.

**Term (ii)** The error incurred from almost-surely approximating a sample average  $\frac{1}{t} \sum_{i=1}^t f(Z_i)$  of influence functions by Gaussian random variables is a direct consequence of Strassen [248] and improvements under  $q > 2$  finite absolute moments by Komlós, Major, and Tusnády [160, 161] and Major [181], which are unimprovable without additional assumptions. Further approximation errors result from using  $\widehat{\text{Var}}_t(\hat{f})$  to estimate  $\text{Var}(f)$ , where almost-sure law of the iterated logarithm rates appear, and are themselves unimprovable.

**Term (iii)** Using the approximations mentioned in (ii) permits the use of Robbins' normal mixture confidence sequence in (i). However, a factor of  $\sqrt{\widehat{\text{Var}}_t(\hat{f})}$  necessarily appears in front of the width as an estimate of the standard deviation  $\sqrt{\text{Var}(f)}$  of the efficient influence function  $f$  discussed in Section 7.3. Importantly,  $\sqrt{\text{Var}(f)}$  corresponds to the semiparametric efficiency bound, so that no estimator of  $\psi$  can have asymptotic mean squared error smaller than  $\text{Var}(f(Z))/t$  without imposing additional assumptions [261].

### 7.B.10 Multivariate asymptotic confidence sequences

We have thus far focused on univariate AsympCSs since even this simple setting encompasses several areas of application including some of our main causal inference-related motivations found in Section 7.3. One may nevertheless wonder if the notion of an AsympCS can be generalized to  $\mathbb{R}^d$  for  $d \geq 2$ , and if so, whether constructions thereof are possible. In this section, we provide a definition for multivariate AsympCSs and construct explicit examples of them, relying on certain multivariate strong invariance principles due to Einmahl [99] and nonasymptotic confidence sets for means of Gaussian vectors due to Manole and Ramdas [183] and Chugg et al. [61]. In what follows, let  $\nu$  be the Lebesgue measure on  $\mathbb{R}^d$ .

*Definition 7.B.1* (Multivariate asymptotic confidence sequences). Let  $\mathcal{T}$  be a totally ordered infinite set including a minimum value  $t_0 \in \mathcal{T}$ . We say that the sequence of  $\mathbb{R}^d$ -valued random sets  $(C_t)_{t \in \mathcal{T}}$  is a  $(1 - \alpha)$ -asymptotic confidence sequence (AsympCS) for a sequence of parameters  $(\theta_t)_{t \in \mathcal{T}}$  taking values in  $\mathbb{R}^d$  if there exists a nonasymptotic  $(1 - \alpha)$ -CS  $(C_t^*)_{t \in \mathcal{T}}$  for  $(\theta_t)_{t \in \mathcal{T}}$

meaning that

$$\mathbb{P}(\forall t \in \mathcal{T}, \theta_t \in C_t^*) \geq 1 - \alpha, \quad (7.302)$$

so that the normalized measure of the symmetric difference between  $(C_t^*)_{t \in \mathcal{T}}$  and  $(C_t)_{t \in \mathcal{T}}$  a.s. vanishes:

$$\frac{\nu(C_t \Delta C_t^*)}{\nu(C_t^*)} \rightarrow 0 \text{ almost surely.} \quad (7.303)$$

In the univariate case, any AsympCS satisfies Definition 7.B.1 (in fact, Definition 7.B.1 is still more general since it encompasses sets other than intervals — such as disjoint unions thereof — but we ignore this technicality since most confidence sets we are interested in are in fact compact and connected). Now, let us use a strong invariance principle due to Einmahl [99] combined with the nonasymptotic confidence sequences for means of sub-Gaussian random vectors due to Manole and Ramdas [183] to derive multivariate AsympCSs for means of i.i.d. random vectors with certain finite moments.

**Proposition 7.B.5** (Multivariate AsympCSs for means of i.i.d. random vectors). *Let  $(X_t)_{t=1}^\infty$  be i.i.d. random vectors with mean  $\mu \in \mathbb{R}^d$  and covariance matrix  $\Sigma \in \mathbb{R}^{d \times d}$  so that*

$$\mathbb{E}\|X\|_2^{2+\delta} < \infty \quad (7.304)$$

for some  $\delta > 0$ . Let  $\hat{\mu}_t := \frac{1}{t} \sum_{i=1}^t X_i$  be the sample mean and  $\hat{\gamma}_t \equiv \gamma(\hat{\Sigma}_t)$  the maximum eigenvalue of the empirical covariance matrix  $\hat{\Sigma}_t := \frac{1}{t} \sum_{i=1}^t (X_i - \hat{\mu}_t)(X_i - \hat{\mu}_t)^\top$ . Then,

$$C_t := \left\{ \mu' \in \mathbb{R}^d : \|\hat{\mu}_t - \mu'\|_2 < \sqrt{\hat{\gamma}_t 16 [2 \log(\log_2(t) + 1) + \log(1/\alpha) + d \log 5] / t} \right\} \quad (7.305)$$

forms a multivariate  $(1 - \alpha)$ -AsympCS for  $\mu$ . The set  $C_t$  can be alternatively written as

$$B_d(\hat{\mu}_t, r_t), \quad \text{where } r_t := \sqrt{\hat{\gamma}_t 16 [2 \log(\log_2(t) + 1) + \log(1/\alpha) + d \log 5] / t} \quad (7.306)$$

and  $B_d(a, r)$  is the ball in  $\mathbb{R}^d$  centered at  $a \in \mathbb{R}^d$  with radius  $r > 0$ .

Similar to the relationship between Proposition 7.2.1 and Theorem 7.2.2, one could have replaced  $\sqrt{16[2 \log(\log_2(t) + 1) + \log(1/\alpha) + d \log 5]}$  with any other time-uniform  $(1 - \alpha)$  boundary for means of 1-sub-Gaussian random vectors. For example, replacing the bound of Manole and Ramdas [183] with that of Chugg et al. [61] in Step 2 of the proof of Proposition 7.B.5, we have that for any  $\rho > 0$ ,

$$C_t := \left\{ \mu' \in \mathbb{R}^d : \|\hat{\mu}_t - \mu'\|_2 < \sqrt{\frac{\hat{\gamma}_t 9d}{2} \cdot \frac{1 + t\rho^2}{t^2 \rho^2} \cdot \left[ 2 + \log \left( \frac{\sqrt{1 + t\rho^2}}{\alpha} \right) \right]} \right\} \quad (7.307)$$

forms a  $(1 - \alpha)$ -AsympCS for  $\mu$  for the same reason that Proposition 7.2.2 reduces to Theorem 7.2.2 for an appropriate tuning parameter  $\rho > 0$ . The bounds in (7.305) and (7.307) can be thought of as multivariate analogues of those found in Proposition 7.2.1 and Theorem 7.2.2, respectively.

*Proof of Proposition 7.B.5.* Like many of the proofs found throughout this chapter, multivariate AsympCSs are constructed from three key ingredients: strongly approximating partial sums of Gaussian vectors, a nonasymptotic time-uniform concentration result, and a justification for nuisance parameter estimation. We deal with these three ingredients in Steps 1, 2, and 3, respectively, ultimately combining them in Step 4.

**Step 1: Multivariate strong invariance via Einmahl [99]** Using the strong invariance principle of Einmahl [99] combined with the condition in (7.304), we have that on a sufficiently rich probability space, there exist i.i.d. multivariate Gaussian random vectors  $(Y_t)_{t=1}^\infty$  with mean zero and covariance matrix  $\Sigma := \text{cov}(X)$  so that for some  $q > 2$ ,

$$(\hat{\mu}_t - \mu) - \bar{Y}_t = o(t^{1/q-1}) \quad (7.308)$$

coordinate-wise almost-surely where  $\bar{Y}_t := \frac{1}{t} \sum_{i=1}^t Y_i$ .

**Step 2: Nonasymptotic time-uniform concentration for Gaussian random vectors**  
We will rely on a result due to Manole and Ramdas [183, Corollary 23] which states that if for some  $\gamma > 0$ ,

$$\sup_{u \in \mathbb{R}^d : \|u\|_2=1} \log \mathbb{E} \exp \{\lambda \langle u, Y \rangle\} \leq \gamma \lambda^2 / 2; \quad \lambda \in \mathbb{R}, \quad (7.309)$$

then with probability at least  $(1 - \alpha)$ ,

$$\forall t \geq 1, \|\bar{Y}_t\| < \sqrt{\gamma 16 [2 \log(\log_2 t + 1) + \log(1/\alpha) + d \log 5] / t}. \quad (7.310)$$

Now, notice that the i.i.d. random variables  $(Y_t)_{t=1}^\infty$  from Step 1 are multivariate Gaussian and hence their moment generating function is given by

$$\mathbb{E} \exp \{\langle u, Y \rangle\} = \exp \left\{ \frac{1}{2} u^T \Sigma u \right\}; \quad u \in \mathbb{R}^d. \quad (7.311)$$

Consequently, we have that for any  $\lambda \in \mathbb{R}$ ,

$$\sup_{u \in \mathbb{R}^d : \|u\|_2=1} \log \mathbb{E} \exp \{\lambda \langle u, Y \rangle\} = \gamma(\Sigma) \cdot \lambda^2 / 2 \quad (7.312)$$

where  $\gamma \equiv \gamma(\Sigma)$  is the largest eigenvalue value of  $\Sigma$ . Applying Manole and Ramdas [183, Corollary 23] as discussed above, we have

$$\mathbb{P} \left( \exists t \geq 1 : \|\bar{Y}_t\|_2 \geq \sqrt{\gamma(\Sigma) 16 [2 \log(\log_2(t) + 1) + \log(1/\alpha) + d \log 5] / t} \right) \leq \alpha. \quad (7.313)$$

**Step 3: Strongly consistent nuisance parameter estimation** Applying the strong law of large numbers coordinate-wise, we have that the sample covariance matrix  $\hat{\Sigma}_t$  is strongly consistent for the covariance matrix  $\Sigma$  in the Frobenius norm  $\|\cdot\|_F$ , meaning that  $\|\hat{\Sigma}_t - \Sigma\|_F =$

$o(1)$  almost surely. By Courant-Fischer,

$$|\gamma(\hat{\Sigma}_t) - \gamma(\Sigma)| \leq \|\hat{\Sigma}_t - \Sigma\|_F = o(1), \quad (7.314)$$

and hence we have that  $\gamma(\hat{\Sigma}_t) \rightarrow \gamma(\Sigma)$  almost surely.

**Step 4: Proving that  $B_d(\hat{\mu}_t, r_t)$  is a  $(1 - \alpha)$ -AsympCS for  $\mu$**  Combining Steps 2 and 3, there exists a radius  $r_t^* \equiv r_t + o(\sqrt{\log \log t/t})$  so that  $\mathbb{P}(\exists t \geq 1 : \|\hat{\mu}_t - \mu\|_2 \geq r_t^*) \leq \alpha$ , or in other words, we have with probability at least  $(1 - \alpha)$ ,

$$\forall t \geq 1, \quad \mu \in B_d(\hat{\mu}_t, r_t^*). \quad (7.315)$$

Therefore, it remains to show that  $\nu(B_d(\hat{\mu}_t, r_t^*) \Delta B_d(\hat{\mu}_t, r_t)) / \nu(B_d(\hat{\mu}_t, r_t^*)) \rightarrow 0$  where

$$r_t := \sqrt{\gamma(\hat{\Sigma}_t) 16 [2 \log(\log_2(t) + 1) + \log(1/\alpha) + d \log 5] / t}. \quad (7.316)$$

Indeed, writing out the aforementioned normalized symmetric difference, we notice that

$$\frac{\nu(B_d(\hat{\mu}_t, r_t^*) \Delta B_d(\hat{\mu}_t, r_t))}{\nu(B_d(\hat{\mu}_t, r_t^*))} \quad (7.317)$$

$$= \frac{\left| \frac{\pi^{d/2}}{\Gamma(d/2+1)} \cdot (r_t^*)^d - \frac{\pi^{d/2}}{\Gamma(d/2+1)} \cdot r_t^d \right|}{\frac{\pi^{d/2}}{\Gamma(d/2+1)} \cdot (r_t^*)^d} \quad (7.318)$$

$$= \left| 1 - \left( \frac{r_t}{r_t + o(\sqrt{\log \log t/t})} \right)^d \right| \quad (7.319)$$

$$= \left| 1 - \left( \frac{C}{C + o(1)} \right)^d \right| = o(1) \quad (7.320)$$

almost surely for some constant  $C > 0$  by the continuous mapping theorem and since  $r_t \asymp \sqrt{\log \log t/t}$ . This completes the proof of Proposition 7.B.5.  $\square$

# Chapter 8

## Distribution-uniform anytime-valid sequential inference

### 8.1 Introduction

Some of the simplest and most efficient statistical inference tools are asymptotic ones that rely on large-sample theory such as the central limit theorem (CLT). However, there is a sharp distinction between asymptotics that are only valid for a single distribution  $P$  and those that are *uniformly valid* over a large collection of distributions  $\mathcal{P}$ . To elaborate, consider the classical CLT which states that for independent and identically distributed random variables  $X_1, \dots, X_n \sim P$  with mean  $\mu_P$  and finite variance  $\sigma_P^2 < \infty$ , their scaled partial sums  $\hat{Z}_n := \sum_{i=1}^n (X_i - \mu_P)/(\sigma_P \sqrt{n})$  are asymptotically standard Gaussian, meaning for any real  $x$ , we have  $\mathbb{P}_P(\hat{Z}_n \leq x) \rightarrow \Phi(x)$  where  $\Phi$  is the cumulative distribution function (CDF) of a standard Gaussian. However, this is a *distribution-pointwise* statement in the sense that the limit holds for a single  $P \in \mathcal{P}$ . An unsettling consequence of  $P$ -pointwise statements is that no matter how large  $n$  is,  $|\mathbb{P}_{P'}(\hat{Z}_n \leq x) - \Phi(x)|$  can be far from zero for some  $P' \in \mathcal{P}$  – or more informally, asymptotics may be “kicking in” arbitrarily late.

By contrast, *distribution-uniformity* (or more specifically  $\mathcal{P}$ -uniformity) rules out the aforementioned unsettling scenario so that convergence occurs simultaneously for all  $P \in \mathcal{P}$ . Concretely, consider the difference between  $P$ -pointwise versus  $\mathcal{P}$ -uniform convergence in distribution when written out side-by-side:

$$\underbrace{\sup_{P \in \mathcal{P}} \lim_{n \rightarrow \infty} \left| \mathbb{P}_P(\hat{Z}_n \leq x) - \Phi(x) \right| = 0}_{P\text{-pointwise convergence in distribution}} \quad \text{versus} \quad \underbrace{\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \left| \mathbb{P}_P(\hat{Z}_n \leq x) - \Phi(x) \right| = 0}_{\mathcal{P}\text{-uniform convergence in distribution}}, \quad (8.1)$$

where the essential difference lies in the order of limits and suprema. The initial study of  $\mathcal{P}$ -uniformity is often attributed to Li [171] and many papers have emphasized its importance in recent years; see Kasy [147], Rinaldo et al. [213], Tibshirani et al. [253], Shah and Peters

[238], Kuchibhotla et al. [165], and Lundborg et al. [179]. Note that this literature sometimes refers to distribution-uniformity as “honesty” [171, 165] or simply “uniformity” [147, 213, 253, 238, 179]. We opt for the phrase “distribution-uniform” — or “ $\mathcal{P}$ -uniform” when we want to specify that uniformity is with respect to  $\mathcal{P}$  — since there are many other notions of uniformity throughout probability and statistics, including time-uniformity and quantile-uniformity, both of which will become relevant throughout this chapter. We do not use the term “honesty” as it has also been used to refer to other properties of estimators in statistical inference [270, 11] and is sometimes used in the sense of parameter-uniformity [224].

Simultaneously, there is a parallel literature on *time-uniform* (typically called “anytime-valid”) inference where the goal is to derive confidence sequences (CSs) — sequences of confidence intervals (CIs) that are uniformly valid for all sample sizes — as well as anytime  $p$ -values and sequential hypothesis tests (to be defined more formally later) that can be continuously monitored and adaptively stopped. This literature has historically taken a mostly nonasymptotic approach to inference so that the type-I errors and coverage probabilities hold in finite samples; see the early work of Wald, Robbins, and colleagues [271, 76, 217, 168], as well as the review paper of Ramdas, Grünwald, Vovk, and Shafer [211] which gives a broad overview of this literature. However, nonasymptotic approaches generally require strong assumptions on the random variables such as lying in a parametric family, *a priori* known bounds on their support, or on their moments. On the other hand, this chapter takes an *asymptotic* view of anytime-valid inference where type-I errors and coverage probabilities hold in the limit; see Chapter 7, Robbins and Siegmund [220], and Bibaut et al. [34]. An advantage of this regime is that the resulting methods take simple, universal forms and allow for substantially weaker conditions (for example, requiring only that absolute moments exist and are finite but for which *a priori* bounds are not known).

To illustrate time-uniformity in the asymptotic regime, suppose that random variables  $X_1, \dots, X_n$  have finite mean  $\mu$  and variance  $\sigma^2$  and we would like to derive a CI for  $\mu$ . A classical asymptotic CI  $\dot{C}_n$  has the guarantee that  $\limsup_{n \rightarrow \infty} \mathbb{P}_P(\mu \notin \dot{C}_n) \leq \alpha$ , but its asymptotic validity hinges on the sample size  $n$  being fixed and pre-specified in advance. By contrast, an asymptotically valid CS  $(\bar{C}_k^{(m)})_{k=m}^{\infty}$  can elicit a much stronger property written in juxtaposition with the classical asymptotic CI as follows:

$$\underbrace{\limsup_{n \rightarrow \infty} \mathbb{P}_P(\mu \notin \dot{C}_n) \leq \alpha}_{(\text{Asymptotic}) \text{ fixed-}n \text{ CI}} \quad \text{versus} \quad \underbrace{\limsup_{m \rightarrow \infty} \mathbb{P}_P(\exists k \geq m : \mu \notin \bar{C}_k^{(m)}) \leq \alpha}_{(\text{Asymptotic}) \text{ anytime-valid CS}} \quad (8.2)$$

where the main difference lies in the fact that the right-hand side probability holds uniformly in  $k \geq m$  for sufficiently large  $m$ . From a practical perspective, the right-hand side permits a researcher to continuously monitor the outcome of an experiment, for example, updating their CIs as each new data point is collected as long as the starting sample size  $m$  is sufficiently large. Importantly, these anytime-valid procedures allow for the experiment to stop *as soon as* the researcher has sufficient evidence to reject some null hypothesis (e.g. as soon as  $0 \notin \bar{C}_k^{(m)}$  for a null effect of 0). Note that while CSs and anytime  $p$ -values are typically studied from a

*nonasymptotic* viewpoint, we will henceforth omit the “asymptotic” phrasing when referring to asymptotic procedures such as those in (8.2) since we are solely interested in asymptotics in this chapter (and distribution-uniformity is always trivially satisfied for nonasymptotics).

In this chapter, our main goal is to define and derive distribution-uniform anytime-valid tests,  $p$ -values, and confidence sequences. However, the time-uniform guarantee in the right-hand side of (8.2) is a *distribution-pointwise* statement, and to the best of our knowledge, there currently exist no distribution-uniform guarantees for time-uniform asymptotics. The reason for this is subtle and has led us to identify a gap in the probability literature. To elaborate, while fixed- $n$  asymptotics are based on the CLT, Chapter 7 analyzed asymptotic analogues of nonasymptotic CSs using strong Gaussian approximations (sometimes called “strong invariance principles” or “strong embeddings”) such as the seminal results of Strassen [248] and Komlós, Major, and Tusnády [160, 161] — see also Chatterjee [54] and the references therein. Not only have strong approximations not yet been studied from a distribution-uniform perspective, it is not even clear what the right definition of “distribution-uniform strong Gaussian approximation” ought to be. We give both a definition and a corresponding result satisfying it in Section 8.5, and this serves as a probabilistic foundation for the rest of our statistical results.

### 8.1.1 Outline of the chapter

Below we outline how the chapter will proceed, highlighting our key contributions.

- We begin in Section 8.2 by defining  $\mathcal{P}$ -uniform anytime-valid inference in the form of anytime hypothesis tests, anytime  $p$ -values, and confidence sequences (Definition 8.2.1). This definition serves as context for Section 8.2.1 where we state our main result in Proposition 8.2.1 (initially without proof). The remaining sections are focused on providing the necessary machinery to prove a stronger version of Proposition 8.2.1 which is ultimately given in Theorem 8.3.3.
- Section 8.2.2 lays some foundations for distribution-, time-, and boundary-uniform central limit theory for centered partial sums, culminating in Proposition 8.2.2. The results therein are new to the literature even in the distribution-pointwise regime. However, Proposition 8.2.2 is stated in terms of the true (rather than empirical) variance in standardizing the partial sums, motivating the following section on distribution-uniform almost-sure consistency.
- Section 8.3 discusses what it means for a sequence of random variables to converge almost-surely *and* uniformly in a class of distributions. The section culminates in a result showing that the empirical variance is a distribution-uniform almost-surely consistent estimator for the true variance and its convergence rate is polynomial in the sample size (Proposition 8.3.1), which, when combined with Proposition 8.2.2 from Section 8.2 yields our main result in Theorem 8.3.3.
- Section 8.4 applies the content of the previous sections to the problem of anytime-valid conditional independence testing. We first show that distribution-uniform anytime-valid tests of conditional independence are impossible to derive without imposing structural assumptions, a fact that can be viewed as a time-uniform analogue of the hardness result

due to Shah and Peters [238, §2]. We then develop a sequential version of the Generalized Covariance Measure test due to Shah and Peters [238, §3] and show that it distribution- and time-uniformly controls the type-I error (and has nontrivial power) as long as certain regression functions are estimated at sufficiently fast rates. To the best of our knowledge, this is the first anytime-valid test of conditional independence that does not rely on Model-X assumptions.

- Section 8.5 highlights that all of the preceding results fundamentally rely on a distribution-uniform strong Gaussian approximation (Theorem 8.5.2) that serves as a (purely probabilistic) foundational piece of our main results and it is the first result of its kind in the literature (to the best of our knowledge). This strong approximation is itself a consequence of a *nonasymptotic* high-probability coupling inequality (Lemma 8.5.1). Finally, we illustrate how these couplings and approximations give rise to a distribution-uniform law of the iterated logarithm. All three of these results may be of independent interest.

### 8.1.2 Notation

Throughout, we will let  $\Omega$  be a sample space,  $\mathcal{F}$  the Borel sigma-algebra, and  $\mathcal{P}$  a collection of probability measures so that  $(\Omega, \mathcal{F}, P)$  is a probability space for each  $P \in \mathcal{P}$ . We will often write  $(\Omega, \mathcal{F}, \mathcal{P})$  to refer to the collection of probability spaces  $(\Omega, \mathcal{F}, P)_{P \in \mathcal{P}}$ . Note that  $P \in \mathcal{P}$  are defined with respect to the same sample space  $\Omega$  and sigma-algebra  $\mathcal{F}$  but do not need to have a common dominating measure (e.g.  $\mathcal{P}$  can consist of infinitely many discrete and continuous distributions as well as their mixtures).

Throughout, we will work with random variables that are defined on the collection of probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  (unless otherwise specified, as will be the case in Section 8.5). For any event  $A \in \mathcal{F}$ , we use  $\mathbb{P}_P(A)$  to denote the probability of that event and  $\mathbb{E}_P(\cdot)$  to denote the expectation of a random variable with respect to  $P \in \mathcal{P}$ , meaning for  $X$  defined on  $(\Omega, \mathcal{F}, P)$ ,

$$\mu_P \equiv \mathbb{E}_P(X) = \int x \, dP(x). \quad (8.3)$$

Similarly,  $\sigma_P^2 \equiv \text{Var}_P(X)$  will be shorthand for  $\mathbb{E}_P(X - \mathbb{E}_P(X))^2$ , and so on.

## 8.2 What is distribution-uniform anytime-valid inference?

Recalling the  $\mathcal{P}$ -uniform convergence in distribution guarantee provided in (8.1), a fixed- $n$   $p$ -value  $\dot{p}_n$  defined on  $(\Omega, \mathcal{F}, \mathcal{P})$  is said to be  $\mathcal{P}_0$ -uniform for the null hypothesis  $\mathcal{P}_0 \subseteq \mathcal{P}$  if

$$\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \mathbb{P}_P(\dot{p}_n \leq \alpha) \leq \alpha, \quad (8.4)$$

and it is easy to see how such a  $p$ -value can be constructed given a statistic satisfying the right-hand side of (8.1). Similarly, Chapter 7 contains a definition of ( $P$ -pointwise) time-uniform coverage of asymptotic CSs, which is also implicit in Robbins and Siegmund [220] and Bibaut et al. [34]. Adapting that definition to anytime  $p$ -values, we say that  $(\bar{p}_k^{(m)})_{k=m}^\infty$  has asymptotic

time-uniform type-I error control under the null  $\mathcal{P}_0$  if

$$\forall P \in \mathcal{P}_0, \limsup_{m \rightarrow \infty} \mathbb{P}_P(\exists k \geq m : \bar{p}_k^{(m)} \leq \alpha) \leq \alpha. \quad (8.5)$$

Juxtaposing (8.4) and (8.5), we can intuit the right definition of distribution- *and* time-uniform type-I error control, where we simply place a supremum over  $\mathcal{P}_0$  inside the limit in (8.5). We lay this definition out formally alongside corresponding definitions for anytime hypothesis tests, confidence sequences, and sharpness thereof below.

*Definition 8.2.1* ( $\mathcal{P}$ -uniform anytime-valid statistical inference). Let  $\mathcal{P}$  be a collection of distributions and let  $\mathcal{P}_0 \subseteq \mathcal{P}$  be the null hypothesis. We say that  $(\bar{\Gamma}_k^{(m)})_{k=m}^\infty$  is a  $\mathcal{P}_0$ -uniform anytime hypothesis test if

$$\limsup_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \mathbb{P}_P \left( \exists k \geq m : \bar{\Gamma}_k^{(m)} = 1 \right) \leq \alpha \quad (8.6)$$

and that  $(\bar{p}_k^{(m)})_{k=m}^\infty$  is a  $\mathcal{P}_0$ -uniform anytime *p*-value if

$$\limsup_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \mathbb{P}_P \left( \exists k \geq m : \bar{p}_k^{(m)} \leq \alpha \right) \leq \alpha. \quad (8.7)$$

Moreover, we say that  $(\bar{C}_k^{(m)})_{k=m}^\infty$  is a  $\mathcal{P}$ -uniform  $(1 - \alpha)$ -confidence sequence for  $\theta(P)$  if

$$\limsup_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : \theta(P) \notin \bar{C}_k^{(m)} \right) \leq \alpha. \quad (8.8)$$

Finally, we say that all of these procedures are sharp if the limit suprema are limits and the inequalities ( $\leq \alpha$ ) are equalities ( $= \alpha$ ).

As one may expect, any  $\mathcal{P}$ -uniform anytime-valid test, *p*-value, or CS satisfying Definition 8.2.1 is also  $\mathcal{P}$ -uniform for a fixed sample size  $n$  in the sense of (8.4) as well as  $P$ -pointwise anytime-valid for any  $P \in \mathcal{P}$  in the sense of (8.5). With Definition 8.2.1 in mind, we will now derive distribution-uniform anytime hypothesis tests, *p*-values, and confidence sequences for the mean of independent and identically distributed random variables.

### 8.2.1 Our primary goal: Inference for the mean

While Definition 8.2.1 is a natural extension of distribution-uniform inference to the anytime-valid setting, it is deceptively challenging to derive procedures satisfying Definition 8.2.1 even for the simplest of statistical problems such as tests for the mean of independent and identically distributed random variables and the main results of this section themselves rely on certain technical underpinnings such as distribution-uniform almost-sure consistency and strong Gaussian approximations. Rather than laboriously discuss these technical details here, let us instead articulate our main goal and result — distribution-uniform anytime inference for the mean — and defer more in-depth discussions to Sections 8.2.2, 8.3, and 8.5.

In many of the results that follow, we will rely on a monotonically increasing function  $\Psi : \mathbb{R}^{\geq 0} \rightarrow [0, 1]$  given by

$$\Psi(r) := 1 - 2[1 - \Phi(\sqrt{r}) + \sqrt{r}\phi(\sqrt{r})]; \quad r \geq 0. \quad (8.9)$$

This function happens to be the cumulative distribution function of a particular probability distribution that we have opted to call the *Robbins-Siegmund distribution* since it was implicitly computed by Robbins and Siegmund [220] in the context of boundary-crossing probabilities for Wiener processes. For now, we will only rely on the fact that  $\Psi$  is invertible and leave more detailed discussions of its properties to Section 8.B.1. As we will see shortly,  $\Psi$  plays a role in asymptotic anytime-valid inference similar to that of the Gaussian cumulative distribution function in asymptotic fixed- $n$  inference. Indeed, define the process  $(\bar{p}_k^{(m)})_{k=m}^\infty$  given by

$$\bar{p}_k^{(m)} := 1 - \Psi(k\hat{\mu}_k^2/\hat{\sigma}_k^2 - \log(k/m)) \quad (8.10)$$

and the intervals  $(\bar{C}_k^{(m)}(\alpha))_{k=m}^\infty$  given by

$$\bar{C}_k^{(m)}(\alpha) := \hat{\mu}_k \pm \hat{\sigma}_k \sqrt{[\Psi^{-1}(1 - \alpha) + \log(k/m)]/k}, \quad (8.11)$$

where  $\hat{\sigma}_k^2 := \frac{1}{k} \sum_{i=1}^k (X_i - \hat{\mu}_k)^2$  is the sample variance. The following result gives conditions under which  $(\bar{p}_k^{(m)})_{k=m}^\infty$  is a  $\mathcal{P}_0$ -uniform anytime  $p$ -value for the null of  $\mu_P = 0$  and  $(\bar{C}_k^{(m)}(\alpha))_{k=m}^\infty$  is a  $\mathcal{P}$ -uniform  $(1 - \alpha)$ -CS for  $\mu_P$  in the sense of Definition 8.2.1.

**Proposition 8.2.1** (Distribution-uniform anytime-valid inference for the mean). *Let  $X_1, X_2, \dots$  be random variables defined on  $(\Omega, \mathcal{F}, \mathcal{P})$ , and suppose that for some  $\delta > 0$ , the  $(2 + \delta)^{th}$  moment is  $\mathcal{P}$ -uniformly upper-bounded and the variance is  $\mathcal{P}$ -uniformly positive, i.e.*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P |X - \mathbb{E}_P(X)|^{2+\delta} < \infty \text{ and } \inf_{P \in \mathcal{P}} \text{Var}_P(X) > 0. \quad (8.12)$$

If  $\mathcal{P}_0 \subseteq \mathcal{P}$  is a subcollection of distributions so that  $\mathbb{E}_P(X) = 0$  for each  $P \in \mathcal{P}_0$ , then  $(\bar{p}_k^{(m)})_{k=m}^\infty$  is a sharp  $\mathcal{P}_0$ -uniform anytime  $p$ -value:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \mathbb{P}_P \left( \exists k \geq m : \bar{p}_k^{(m)} \leq \alpha \right) = \alpha, \quad (8.13)$$

and  $(\bar{C}_k^{(m)}(\alpha))_{k=m}^\infty$  is a sharp  $\mathcal{P}$ -uniform CS for the mean:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : \mu_P \notin \bar{C}_k^{(m)}(\alpha) \right) = \alpha. \quad (8.14)$$

Rather than prove Proposition 8.2.1 directly, we will spend the next few sections laying the groundwork to prove a more general result, culminating in Theorem 8.3.3. Clearly, one can obtain a sharp  $\mathcal{P}_0$ -uniform level- $\alpha$  anytime hypothesis test  $(\bar{\Gamma}_k^{(m)})_{k=m}^\infty$  in the sense of

Definition 8.2.1 from Proposition 8.2.1 by setting  $\bar{\Gamma}_k^{(m)} := \mathbb{1}\{\bar{p}_k^{(m)} \leq \alpha\}$  or  $\bar{\Gamma}_k^{(m)} := \mathbb{1}\{0 \notin \bar{C}_k^{(m)}\}$ . Notice that uniformly bounded  $(2 + \delta)^{\text{th}}$  moment conditions are precisely what appear in several distribution-uniform central limit theorems [238, 179].

### 8.2.2 Time- and $\mathcal{P}$ -uniform central limit theory for partial sums

Recall that in the batch (fixed- $n$ , non-sequential) setting, the CLT is typically stated for a single quantile, meaning that the survival function  $\mathbb{P}_P(S_n/\sqrt{n} \geq x)$  – equivalently, the CDF<sup>1</sup> – of  $\sqrt{n}$ -scaled normalized partial sums  $S_n := \sigma^{-1} \sum_{i=1}^n [X_i - \mathbb{E}_P(X)]$  converge to that of a standard Gaussian:

$$\forall x \in \mathbb{R}, \lim_{n \rightarrow \infty} |\mathbb{P}_P(S_n/\sqrt{n} \geq x) - [1 - \Phi(x)]| = 0. \quad (8.15)$$

Under no additional assumptions, however, the above holds *quantile-uniformly* [260, Lemma 2.11], meaning

$$\lim_{n \rightarrow \infty} \sup_{x \in \mathbb{R}} |\mathbb{P}_P(S_n/\sqrt{n} \geq x) - [1 - \Phi(x)]| = 0. \quad (8.16)$$

Clearly, (8.16) is strictly stronger than (8.15). Particularly relevant to this chapter, distribution-uniform *fixed-n* tests and CIs are also stated with quantile-uniformity and their proofs typically rely on this property. Even in the  $P$ -pointwise case, however, there is no result showing that an analogous property exists for time-uniform *boundaries* (and it is not clear in what sense such a statement should be formulated). The following theorem provides such a result in both the  $P$ -pointwise and  $\mathcal{P}$ -uniform settings.

**Proposition 8.2.2** (( $\mathcal{P}, n, x$ )-uniform boundaries for centered partial sums). *Let  $X_1, X_2, \dots$  be random variables defined on probability spaces  $(\Omega, \mathcal{F}, \mathcal{P}^\star)$  with finite  $(2 + \delta)^{\text{th}}$  moments, i.e.  $\mathbb{E}_P |X - \mathbb{E}_P(X)|^{2+\delta} < \infty$  for every  $P \in \mathcal{P}^\star$ . Letting  $S_n := \sum_{i=1}^n (X_i - \mathbb{E}_P(X_i))/\sigma_P$  be their centered partial sums, we have*

$$\forall P \in \mathcal{P}^\star, \lim_{m \rightarrow \infty} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \exists k \geq m : |S_k|/\sqrt{k} \geq \sqrt{x + \log(k/m)} \right) - [1 - \Psi(x)] \right| = 0. \quad (8.17)$$

Furthermore, if  $\mathcal{P} \subseteq \mathcal{P}^\star$  is a sub-collection of distributions for which the  $(2 + \delta)^{\text{th}}$  moment is  $\mathcal{P}$ -uniformly upper-bounded and the variance is  $\mathcal{P}$ -uniformly positive, then the above limit holds  $\mathcal{P}$ -uniformly:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \exists k \geq m : |S_k|/\sqrt{k} \geq \sqrt{x + \log(k/m)} \right) - [1 - \Psi(x)] \right| = 0. \quad (8.18)$$

The proof of Proposition 8.2.2 in Section 8.A.1 relies on our novel distribution-uniform strong Gaussian approximation discussed in Section 8.5. After some algebraic manipulations, one can see that (8.18) is equivalent to saying that  $\sup_{k \geq m} \{S_k^2/(\sigma_P^2 k) - \log(k/m)\}$  converges

---

<sup>1</sup>This discussion is in terms of the survival function  $\mathbb{P}_P(S_n/\sqrt{n} \geq x)$  instead of the CDF  $\mathbb{P}_P(S_n/\sqrt{n} \leq x)$  to aid transparent comparisons with boundary-crossing inequalities in Proposition 8.2.2.

$\mathcal{P}$ -uniformly in distribution to the Robbins-Siegmund distribution, i.e.

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{S_k^2}{\sigma_P^2 k} - \log(k/m) \right\} \leq x \right) - \Psi(x) \right| = 0, \quad (8.19)$$

and similarly for the  $P$ -pointwise case in (8.17) but with the above limit over  $m$  and supremum over  $P \in \mathcal{P}$  swapped.

Note that Proposition 8.2.2 does not quite yield Proposition 8.2.1 as a direct consequence since the variance  $\sigma_P^2$  used in the latter is the true (rather than empirical) variance. Moving to a fully empirical version of Proposition 8.2.2 will require that the variance  $\sigma_P^2$  is not only consistently estimated but *almost surely* at a *polynomial rate* and for the  $\mathcal{P}$ -uniform result in (8.18), we will require that this consistency also holds uniformly in  $\mathcal{P}$ . But what does it mean for a sequence of random variables (such as a sequence of estimators) to converge to a limit almost surely *and uniformly in  $\mathcal{P}$* ? The next section provides an answer to this question alongside sufficient conditions for the sample variance to be  $\mathcal{P}$ -uniformly almost surely consistent.

## 8.3 Almost-sure consistency and time-uniform asymptotics

In Proposition 8.2.2, we stated a  $\mathcal{P}$ -, time-, and boundary-uniform convergence result for centered partial sums, but this depended on those partial sums  $S_n(P) := \sum_{i=1}^n (X_i - \mu_P)/\sigma_P$  being weighted by the true standard deviation  $\sigma_P$ . The natural next step is to replace the true variance  $\sigma_P^2$  by an *empirical* variance  $\hat{\sigma}_n^2$  so that the results of Proposition 8.2.2 still hold with  $\hat{\sigma}_n^2$  in place of  $\sigma_P^2$ , thereby providing tools that can be used to derive  $\mathcal{P}$ -uniform anytime-valid tests,  $p$ -values, and confidence sequences. However, the conditions that must be placed on  $\hat{\sigma}_n^2$  are different from what one may encounter in a classical asymptotic inference analysis – indeed we will require  $\hat{\sigma}_n^2$  to be  $\mathcal{P}$ -uniformly almost-surely consistent for  $\sigma_P^2 := \mathbb{E}_P(X - \mathbb{E}_P X)^2$  at a faster-than-logarithmic rate. However, the notion of  $\mathcal{P}$ -uniform almost-sure convergence is not commonly encountered in the statistical literature, so this section is dedicated to reviewing this.

### 8.3.1 What is $\mathcal{P}$ -uniform almost-sure consistency?

Recall the classical notion of convergence in  $P$ -probability for a single  $P \in \mathcal{P}$  and its natural extension to  $\mathcal{P}$ -uniform convergence in probability. That is, a sequence of random variables  $Y_1, Y_2, \dots$  defined on a probability space  $(\Omega, \mathcal{F}, P)$  is said to converge *in probability* to 0 (or  $Y_n = o_P(1)$  for short) if for any  $\varepsilon > 0$ ,

$$\sup_{P \in \mathcal{P}} \lim_{n \rightarrow \infty} \mathbb{P}_P(|Y_n| \geq \varepsilon) = 0, \quad (8.20)$$

and that this convergence holds uniformly in  $\mathcal{P}$  (or  $Y_n = o_{\mathcal{P}}(1)$  for short) if

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P(|Y_n| \geq \varepsilon) = 0. \quad (8.21)$$

The extension of (8.20) to (8.21) is very natural, but at first glance, an analogous extension for almost-sure convergence is less obvious. Indeed, recall that a sequence of random variables  $Y_1, Y_2, \dots$  is said to converge  $P$ -almost surely to 0 for every  $P \in \mathcal{P}$  if

$$\forall P \in \mathcal{P}, \mathbb{P}_P \left( \lim_{n \rightarrow \infty} |Y_n| = 0 \right) = 1. \quad (8.22)$$

It is not immediately obvious what the “right” notion of  $\mathcal{P}$ -uniform almost-sure consistency ought to be since taking an infimum over  $P \in \mathcal{P}$  of the above probabilities does not change the statement of (8.22) whatsoever. Intuitively, it is not possible to simply swap limits and suprema in (8.22) as was done when (8.20) was extended to (8.21). However, it is possible to make such a leap when using an equivalent definition of almost-sure convergence using an idea attributable to Chung [64] and Beck and Giesy [22]. To elaborate, it is a well-known fact that for any  $P \in \mathcal{P}$ ,

$$\mathbb{P}_P \left( \lim_{n \rightarrow \infty} |Y_n| = 0 \right) = 1 \quad \text{if and only if} \quad \forall \varepsilon > 0, \lim_{m \rightarrow \infty} \mathbb{P}_P \left( \sup_{k \geq m} |Y_k| \geq \varepsilon \right) = 0, \quad (8.23)$$

and for this reason, instead of writing  $Y_n = o_{\text{a.s.}}(1)$  as a shorthand for  $P$ -almost-sure convergence, we write  $Y_n = \bar{o}_P(1)$  with the overhead bar  $\bar{o}$  to emphasize time-uniformity and the subscript  $o_P$  to emphasize the distribution  $P$  that this convergence is with respect to. As such, a natural notion of  $\mathcal{P}$ -uniform almost-sure convergence is one that places a supremum over  $P \in \mathcal{P}$  in the right-hand limit of (8.23) which we make precise in the following definition.

*Definition 8.3.1 ( $\mathcal{P}$ -uniform almost-sure convergence [64, 288]).* We say that a sequence of random variables  $Y_1, Y_2, \dots$  defined on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  converges  $\mathcal{P}$ -uniformly and almost surely to 0 if

$$\forall \varepsilon > 0, \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |Y_k| \geq \varepsilon \right) = 0, \quad (8.24)$$

and we write  $Y_n = \bar{o}_P(1)$  for short, where the overhead bar  $\bar{o}$  emphasizes time-uniformity and the subscript  $o_P$  emphasizes  $\mathcal{P}$ -uniformity. Finally, we write  $Y_n = \bar{o}_P(r_n)$  for a monotonically nonincreasing sequence  $(r_n)_{n=1}^{\infty}$  if  $r_n \cdot Y_n = \bar{o}_P(1)$ .

The expression in (8.24) initially appeared in a paper by Chung [64] in a proof of a  $\mathcal{P}$ -uniform strong law of large numbers, and later in a more explicit form by Beck and Giesy [22]. Table 8.1 summarizes the four notions of convergence  $\dot{o}_P(\cdot)$ ,  $\bar{o}_P(\cdot)$ ,  $\dot{\bar{o}}_P(\cdot)$ , and  $\bar{\dot{o}}_P(\cdot)$  and the implications between them.

Adapting (8.24) to the discussion of consistency in parameter estimation, however, requires some additional care since the parameter of interest may itself depend on the distribution  $P \in \mathcal{P}$ . That is, let  $(\hat{\theta}_n)_{n=1}^{\infty}$  be a sequence of estimators and for each  $P \in \mathcal{P}$ , let  $\theta(P) \in \mathbb{R}$  be a real-valued parameter. We will consider  $\hat{\theta}_n$  to be a  $\mathcal{P}$ -uniformly consistent estimator for

Table 8.1: Four notions of convergence with implications between them. Recall that  $\bar{o}_P(\cdot)$  is equivalent to  $P$ -a.s. convergence. Clearly, if a sequence of random variables converges with respect to one of the four cells below, it also does so with respect to the cell above and/or to the left of it. This section is concerned with the strongest of the four, found in the bottom right cell with the **bolded** frame:  $\mathcal{P}$ -uniform almost-sure convergence.

	$P$ -pointwise	$\mathcal{P}$ -uniform
<b>In probability</b>	$\dot{o}_P(\cdot)$	$\dot{o}_{\mathcal{P}}(\cdot)$
<b>Almost surely</b>	$\bar{o}_P(\cdot)$	$\bar{o}_{\mathcal{P}}(\cdot)$

$\theta \equiv \{\theta(P)\}_{P \in \mathcal{P}}$  if

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |\hat{\theta}_k - \theta(P)| \geq \varepsilon \right) = 0, \quad (8.25)$$

and as a shorthand, we will write  $\hat{\theta}_n - \theta = \bar{o}_{\mathcal{P}}(1)$ . Similarly to Definition 8.3.1, we write  $\hat{\theta}_n - \theta = \bar{o}_{\mathcal{P}}(r_n)$  if  $r_n \cdot (\hat{\theta}_n - \theta) = \bar{o}_{\mathcal{P}}(1)$ .

Following the relationship between  $\dot{o}_P$  and  $\dot{O}_P$  notation in the fixed- $n$  in-probability setting, we now provide an analogous definition of time- and  $\mathcal{P}$ -uniform stochastic boundedness. To the best of our knowledge, this definition is new to the literature.

*Definition 8.3.2.* We say that a sequence of random variables  $Y_1, Y_2, \dots$  defined on  $(\Omega, \mathcal{F}, \mathcal{P})$  is time- and  $\mathcal{P}$ -uniformly stochastically bounded if for any  $\delta > 0$ , there exists some  $C \equiv C(\delta) > 0$  and  $M \equiv M(C, \delta) > 1$  so that for all  $m \geq M$ ,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P (\exists k \geq m : |X_k| > C) < \delta, \quad (8.26)$$

and we write  $Y_n = \bar{O}_{\mathcal{P}}(1)$  as a shorthand for the above. Similar to Definition 8.3.1, we write  $Y_n = \bar{O}_{\mathcal{P}}(r_n)$  if  $r_n \cdot Y_n = \bar{O}_{\mathcal{P}}(1)$ .

Note that we do *not* refer to Definition 8.3.2 as  $\mathcal{P}$ -uniform “almost sure” boundedness since even in the  $P$ -pointwise case, almost-sure boundedness and time-uniform stochastic boundedness are not equivalent despite the relationship in (8.23) for almost-sure and time-uniform *convergence*. A related condition has also appeared in the context of conditional local independence testing as in Christgau et al. [60]. As one may expect, there is a calculus of  $\bar{o}_P(\cdot)$  and  $\bar{O}_{\mathcal{P}}(\cdot)$  analogous to that for  $\dot{o}_P(\cdot)$  and  $\dot{O}_{\mathcal{P}}(\cdot)$ . We lay this out formally in the following lemma, but the proofs are routine and can be found in Section 8.A.2.

**Lemma 8.3.1** (Calculus of  $\bar{O}_{\mathcal{P}}(\cdot)$  and  $\bar{o}_{\mathcal{P}}(\cdot)$ ). *Let  $Y_1, Y_2, \dots$  be random variables defined on  $(\Omega, \mathcal{F}, \mathcal{P})$ . Let  $(a_n)_{n=1}^{\infty}$  and  $(b_n)_{n=1}^{\infty}$  be positive and monotonically nonincreasing sequences.*

Then we have the following basic implications:

$$Y_n = \bar{o}_{\mathcal{P}}(a_n) \implies Y_n = \bar{O}_{\mathcal{P}}(a_n) \quad (8.27)$$

$$Y_n = \bar{o}_{\mathcal{P}}(a_n)\bar{O}_{\mathcal{P}}(b_n) \implies Y_n = \bar{o}_{\mathcal{P}}(a_n b_n) \quad (8.28)$$

$$Y_n = \bar{O}_{\mathcal{P}}(a_n)\bar{O}_{\mathcal{P}}(b_n) \implies Y_n = \bar{O}_{\mathcal{P}}(a_n b_n) \quad (8.29)$$

$$Y_n = \bar{o}_{\mathcal{P}}(a_n) + \bar{O}_{\mathcal{P}}(a_n) \implies Y_n = \bar{O}_{\mathcal{P}}(a_n) \quad (8.30)$$

$$Y_n = \bar{o}_{\mathcal{P}}(a_n) + \bar{o}_{\mathcal{P}}(b_n) \implies Y_n = \bar{o}_{\mathcal{P}}(\max\{a_n, b_n\}). \quad (8.31)$$

Furthermore, (8.31) holds with  $\bar{o}_{\mathcal{P}}(\cdot)$  replaced by  $\bar{O}_{\mathcal{P}}(\cdot)$  on both sides. Finally, if  $Y_n = \bar{O}_{\mathcal{P}}(a_n)$  and  $a_n/b_n \rightarrow 0$ , then  $Y_n = \bar{o}_{\mathcal{P}}(b_n)$ .

The calculus provided in Lemma 8.3.1 will appear frequently throughout the proofs of our main results. In the next section, we discuss  $\mathcal{P}$ -uniform, almost-sure, polynomial-rate variance estimation and its implications for deriving an empirical version of Proposition 8.2.2.

### 8.3.2 $\mathcal{P}$ -uniform almost-sure variance estimation

In Section 8.2.2, we alluded to the fact that arriving at a fully empirical version of Proposition 8.2.2 would require  $\mathcal{P}$ -uniformly almost-surely consistent estimation of the variance  $\sigma^2 := \mathbb{E}(X - \mathbb{E}X)^2$  at a faster-than-logarithmic rate. With Definition 8.3.1 and the expression (8.24) in mind, we now provide sufficient conditions for this consistency.

**Proposition 8.3.1** ( $\mathcal{P}$ -uniform almost-surely consistent variance estimation). *Consider the same setup as in Proposition 8.2.2 where  $(X_n)_{n=1}^\infty$  have  $\mathcal{P}$ -uniformly upper-bounded  $(2 + \delta)^{\text{th}}$  moments and  $\mathcal{P}$ -uniformly positive variances. Then the sample variance  $\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_n)^2$  is a  $\mathcal{P}$ -uniformly almost-surely consistent estimator of the variance  $\sigma^2$  at a polynomial rate, meaning there exists  $\beta > 0$  so that*

$$\hat{\sigma}_n^2 = \sigma^2 + \bar{o}_{\mathcal{P}}(n^{-\beta}), \quad (8.32)$$

or more formally, for all  $\varepsilon > 0$ , we have

$$\lim_{m \rightarrow \infty} \mathbb{P}_P \left( \sup_{k \geq m} k^\beta |\hat{\sigma}_k^2 - \sigma_P^2| \geq \varepsilon \right) = 0. \quad (8.33)$$

Proposition 8.3.1 is an immediate consequence of Chapter 9 combined with the de la Vallée Poussin criterion for uniform integrability (see Chong [58] and Hu and Rosalsky [126]).

### 8.3.3 The main result: $(\mathcal{P}, n, \alpha)$ -uniform statistical inference

Pairing together Proposition 8.2.2 and Proposition 8.3.1, we obtain the following ( $\mathcal{P}$ -uniform) anytime  $p$ -values and CSs whose type-I errors converge to the nominal level  $\alpha \in (0, 1)$  uniformly in  $\alpha$ . We present this in the following result on distribution-, time-, and  $\alpha$ -uniform – or  $(\mathcal{P}, n, \alpha)$ -uniform for short – statistical inference. This is our main result and it implies both Proposition 8.2.1 and Proposition 8.2.2 as special cases.

**Theorem 8.3.3** ( $(\mathcal{P}, n, \alpha)$ -uniform statistical inference). *Let  $X_1, X_2, \dots$  be defined on  $(\Omega, \mathcal{F}, \mathcal{P})$  and suppose that for some  $\delta > 0$ , the  $(2 + \delta)^{\text{th}}$  moment is  $\mathcal{P}$ -uniformly upper-bounded and the variance is  $\mathcal{P}$ -uniformly positive. Recall the definitions of  $(\bar{p}_k^{(m)})_{k=m}^{\infty}$  and  $(\bar{C}_k^{(m)}(\alpha))_{k=m}^{\infty}$  from Proposition 8.2.1:*

$$\bar{p}_k^{(m)} := 1 - \Psi(k\hat{\mu}_k^2/\hat{\sigma}_k^2 - \log(k/m)) \quad (8.34)$$

$$\text{and } \bar{C}_k^{(m)}(\alpha) := \hat{\mu}_k \pm \hat{\sigma}_k \sqrt{[\Psi^{-1}(1 - \alpha) + \log(k/m)]/k}. \quad (8.35)$$

Let  $\mathcal{P}_0 \subseteq \mathcal{P}$  be a subcollection of distributions so that  $\mathbb{E}_P(X) = 0$  for each  $P \in \mathcal{P}_0$ . Then the time-uniform type-I error of  $(\bar{p}_k^{(m)})_{k=m}^{\infty}$  and the time-uniform miscoverage of  $(\bar{C}_k^{(m)})_{k=m}^{\infty}$  converge to  $\alpha \in (0, 1)$  uniformly in  $\alpha$ , meaning

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sup_{\alpha \in (0, 1)} \left| \mathbb{P}_P \left( \exists k \geq m : \bar{p}_k^{(m)} \leq \alpha \right) - \alpha \right| = 0, \quad \text{and} \quad (8.36)$$

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{\alpha \in (0, 1)} \left| \mathbb{P}_P \left( \exists k \geq m : \mathbb{E}_P(X) \notin \bar{C}_k^{(m)}(\alpha) \right) - \alpha \right| = 0. \quad (8.37)$$

The full proof of Theorem 8.3.3 can be found in Section 8.A.3. As alluded to at the beginning of Section 8.2, its proof relies on a  $\mathcal{P}$ -uniform strong Gaussian approximation theorem discussed in Section 8.5. Before that, we will discuss how the results derived thus far can be used to conduct distribution-uniform anytime-valid tests of conditional independence.

## 8.4 Illustration: Sequential conditional independence testing

In this section, we aim to derive anytime-valid tests for the null hypothesis,  $X \perp\!\!\!\perp Y | Z$  given  $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ -valued triplets  $(X_n, Y_n, Z_n)_{n=1}^{\infty}$  on probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$ . Several works on conditional independence testing operate under the so-called ‘‘Model-X’’ assumption where the conditional distribution of  $X | Z$  is known exactly [47]. We do not work under the Model-X assumption in this illustration. It is well-known that testing for conditional independence is much simpler under Model-X, and indeed the recent works of Duan et al. [88], Shaer et al. [233], and Grünwald et al. [112] derive powerful anytime-valid tests in that paradigm. Borrowing a quote from the recent work of Grünwald et al. [112], the authors write ‘‘it is an open question to us how to construct general sequential tests of conditional independence without the [Model-X] assumption’’. This section gives an answer to this question, deriving tests that draw inspiration from the batch tests found in Shah and Peters [238] — a pair of authors we will henceforth refer to as S&P. Before giving a brief refresher on batch conditional independence testing and the main results of S&P, let us review some basic concepts in weak regression consistency since nuisance function estimation will form key conditions for our results.

### 8.4.1 Prelude: weak regression consistency

An important part of conditional independence testing (in both batch and sequential settings as we will see) is the ability to consistently estimate certain regression functions. Recall that the (potentially random) squared  $L_2(P)$  risk of a regression estimator  $\hat{f}_n : \mathbb{R}^d \rightarrow \mathbb{R}$  for a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  is given by

$$\|\hat{f}_n - f\|_{L_2(P)}^2 := \int_{z \in \mathbb{R}^d} (\hat{f}_n(z) - f(z))^2 dP(z). \quad (8.38)$$

Importantly, if sample splitting is used to construct  $\hat{f}_n$ , the norm  $\|\cdot\|_{L_2(P)}$  is to be interpreted as conditional on that “training” data. Recall from Györfi et al. [114, Definition 1.1] that a regression estimator  $\hat{f}_n : \mathbb{R}^d \rightarrow \mathbb{R}$  is  $P$ -weakly consistent for a function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  in  $L_2(P)$  at a rate of  $r_n$  if its expected  $L_2(P)$  risk vanishes at that rate, meaning

$$\mathbb{E}_P \|\hat{f}_n - f\|_{L_2(P)} = o(r_n), \quad (8.39)$$

and hence we will say that  $\hat{f}_n$  is  $\mathcal{P}$ -weakly consistent at the rate  $r_n$  if the above convergence occurs uniformly in the class of distributions  $\mathcal{P}$ :

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \|\hat{f}_n - f\|_{L_2(P)} = o(r_n). \quad (8.40)$$

At times, we may omit  $L_2(P)$  from the norm  $\|\cdot\|_{L_2(P)}$  in (8.39) and write  $\|\cdot\|$  when the norm is clear from context.

### 8.4.2 A brief refresher on batch conditional independence testing

Given  $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ -valued triplets  $(X_i, Y_i, Z_i)_{i=1}^n$  from some distribution in a class  $\mathcal{P}$ , the problem of conditional independence testing is concerned with the null

$$H_0 : X \perp\!\!\!\perp Y | Z \quad \text{versus the alternative} \quad H_1 : X \not\perp\!\!\!\perp Y | Z. \quad (8.41)$$

As alluded to before, without the Model-X assumption, powerful tests for the conditional independence null  $H_0$  in (8.41) are *impossible* to derive (even in the batch and asymptotic settings) unless additional distributional or structural assumptions are imposed [S&P, §2]. Indeed, S&P show that even in the bounded setting where  $(X, Y, Z) \sim P \in \mathcal{P}^*$  take values in  $[0, 1] \times [0, 1] \times [0, 1]$ , any test with distribution-uniform type-I error control under  $H_0$  is powerless against *any* alternative in  $H_1$ . Formally, if  $\mathcal{P}_0^* \subset \mathcal{P}^*$  is the subset of distributions satisfying  $H_0$  (and hence  $\mathcal{P}_1^* := \mathcal{P}^* \setminus \mathcal{P}_0^*$  satisfies  $H_1$ ), then

$$\underbrace{\sup_{P \in \mathcal{P}_1^*} \limsup_{n \rightarrow \infty} \mathbb{P}_P (\dot{\Gamma}_n = 1)}_{\text{Best-case } \mathcal{P}_1^*\text{-pointwise power}} \leq \underbrace{\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0^*} \mathbb{P}_P (\dot{\Gamma}_n = 1)}_{\text{Worst-case } \mathcal{P}_0^*\text{-uniform type-I error}}. \quad (8.42)$$

As a consequence of (8.42), one cannot derive a more powerful test than the trivial one that ignores all of the data  $(X_i, Y_i, Z_i)_{i=1}^n$  and randomly outputs 1 with probability  $\alpha$ .

Despite the rather pessimistic result in (8.42), S&P derive the Generalized Covariance Measure (GCM) test which manages to achieve nontrivial power while still uniformly controlling the type-I error. The caveat here is that they are controlling the type-I error in a restricted (but nevertheless rich and nonparametric) class of nulls  $\mathcal{P}_0 \subseteq \mathcal{P}_0^*$ , and the restriction they impose is that certain nuisance functions are sufficiently estimable, a requirement commonly appearing in other literatures including semiparametric functional estimation [155, 15]. Let us now review the key aspects of their test. S&P introduce the estimated residuals  $R_{i,n}$  for each  $i \in [n]$ :

$$R_{i,n} := \{X_i - \hat{\mu}_n^x(Z_i)\} \{Y_i - \hat{\mu}_n^y(Z_i)\} \quad (8.43)$$

where  $\hat{\mu}_n^x(z)$  and  $\hat{\mu}_n^y(z)$  are estimates of the regression functions  $\mu^x(z) := \mathbb{E}(X | Z = z)$  and  $\mu^y(z) := \mathbb{E}(Y | Z = z)$ . For the remainder of the discussion on batch conditional independence testing, we will assume that  $\hat{\mu}_n^x(Z_i)$  and  $\hat{\mu}_n^y(Z_i)$  are constructed from an independent sample (e.g. through sample-splitting or cross-fitting, in which case we may assume access to  $2n$  triplets of  $(X, Y, Z)$ ) for mathematical simplicity, but S&P do not always suggest doing so. However, we will not dwell on arguments for or against sample splitting here. From the residuals in (8.43), they construct the test statistic  $\dot{\text{GCM}}_n$  taking the form

$$\dot{\text{GCM}}_n := \frac{1}{n\hat{\sigma}_n} \sum_{i=1}^n R_{i,n} \quad (8.44)$$

where  $\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n R_{i,n}^2 - \left(\frac{1}{n} \sum_{i=1}^n R_{i,n}\right)^2$  and they show that if the regression functions  $(\mu^y, \mu^x)$  are estimated sufficiently fast (and under some other mild regularity conditions) then  $\sqrt{n}\dot{\text{GCM}}_n$  has a standard Gaussian limit, enabling asymptotic (fixed- $n$ ) inference. We formally recall a minor simplification of their main result here. Consider the following three assumptions for a class of distributions  $\mathcal{P}_0$ .

**Assumption GCM-1** (Product regression error decay). *The weak convergence rate of the average of product residuals is faster than  $n^{-1/2}$ , i.e.*

$$\sup_{P \in \mathcal{P}_0} \|\mu^x - \hat{\mu}_n^x\|_{L_2(P)} \cdot \|\mu^y - \hat{\mu}_n^y\|_{L_2(P)} = o(n^{-1/2}). \quad (8.45)$$

**Assumption GCM-2** ( $\mathcal{P}_0$ -uniform regularity of regression errors). *Letting  $\xi^x := \{X - \mu^x(Z)\}$  and  $\xi^y := \{Y - \mu^y(Z)\}$  denote the true residuals, the variances of  $\{\hat{\mu}_n^x(Z) - \mu^x(Z)\}\xi^y$  and  $\{\hat{\mu}_n^y(Z) - \mu^y(Z)\}\xi^x$  are  $\mathcal{P}_0$ -uniformly vanishing, i.e.*

$$\sup_{P \in \mathcal{P}_0} \text{Var}_P (\{\hat{\mu}_n^x(Z) - \mu^x(Z)\} \cdot \xi^y) = o(1) \quad (8.46)$$

$$\text{and} \quad \sup_{P \in \mathcal{P}_0} \text{Var}_P (\{\hat{\mu}_n^y(Z) - \mu^y(Z)\} \cdot \xi^x) = o(1). \quad (8.47)$$

**Assumption GCM-3 ( $\mathcal{P}_0$ -uniformly bounded moments).** *The true product residuals defined above have  $\mathcal{P}_0$ -uniformly upper-bounded  $(2 + \delta)^{\text{th}}$  moments for some  $\delta > 0$  and uniformly lower-bounded second moments:*

$$\sup_{P \in \mathcal{P}_0} \mathbb{E}_P |\xi^x \xi^y|^{2+\delta} < \infty \quad (8.48)$$

$$\text{and } \inf_{P \in \mathcal{P}_0} \text{Var}_P(\xi^x \xi^y) > 0. \quad (8.49)$$

With these three assumptions in mind, we are ready to recall a simplified version of Shah and Peters [238, Theorem 6].

**Theorem (S&P:  $\mathcal{P}_0$ -uniform validity of the GCM test).** *Suppose  $(X_i, Y_i, Z_i)_{i=1}^n$  are  $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ -valued random variables on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  and let  $\mathcal{P}_0 \subset \mathcal{P}$  be the collection of distributions in  $\mathcal{P}$  satisfying the conditional independence null  $H_0$  and Assumptions GCM-1, GCM-2, and GCM-3. Then,*

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_P(\sqrt{n} \dot{\text{GCM}}_n \leq x) - \Phi(x) \right| = 0. \quad (8.50)$$

and hence the function given by  $\Gamma_k^{(m)} := \mathbf{1} \left\{ |\sqrt{n} \dot{\text{GCM}}_n| \geq \Phi^{-1}(1 - \alpha/2) \right\}$  is a  $\mathcal{P}_0$ -uniform level- $\alpha$  test.

We will now shift our focus to *sequential* conditional independence testing with anytime-valid type-I error guarantees. Before deriving an explicit test, we first demonstrate in Proposition 8.4.1 that the hardness of conditional independence testing highlighted in (8.42) has a similar analogue in the anytime-valid regime.

### 8.4.3 On the hardness of anytime-valid conditional independence testing

As mentioned in Section 8.4.2, S&P illustrated the fundamental hardness of conditional independence testing by showing that unless additional restrictions are placed on the null hypothesis  $\mathcal{P}_0^*$ , any  $\mathcal{P}_0^*$ -uniformly valid (fixed- $n$ ) test is powerless against any alternative, i.e.

$$\underbrace{\sup_{P \in \mathcal{P}_1^*} \limsup_{n \rightarrow \infty} \mathbb{P}_P(\dot{\Gamma}_n = 1)}_{\text{Best-case } \mathcal{P}_1^*\text{-pointwise power}} \leq \underbrace{\limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}_0^*} \mathbb{P}_P(\dot{\Gamma}_n = 1)}_{\text{Worst-case } \mathcal{P}_0^*\text{-uniform type-I error}}. \quad (8.42 \text{ revisited})$$

Does an analogous result hold if  $\dot{\Gamma}_n$  is replaced by an anytime-valid hypothesis test  $\bar{\Gamma}_k^{(m)}$  as in Definition 8.2.1? The following proposition gives an answer to this question, confirming that anytime-valid conditional independence testing is fundamentally hard in a sense similar to (8.42).

**Proposition 8.4.1 (Hardness of anytime-valid conditional independence testing).** *Suppose  $(X_n, Y_n, Z_n)_{n=1}^\infty$  are  $[0, 1]^3$ -valued triplets on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P}^*)$  where  $\mathcal{P}^*$  consists of all distributions supported on  $[0, 1]^3$ . Let  $\mathcal{P}_0^* \subseteq \mathcal{P}^*$  be the subset of distributions satisfying the*

conditional independence null  $H_0$  and denote  $\mathcal{P}_1^* := \mathcal{P}^* \setminus \mathcal{P}_0^*$ . Then for any potentially randomized test  $(\bar{\Gamma}_k^{(m)})_{k=m}^\infty$ ,

$$\sup_{P \in \mathcal{P}_1^*} \limsup_{m \rightarrow \infty} \mathbb{P}_P \left( \exists k \geq m : \bar{\Gamma}_k^{(m)} = 1 \right) \leq \limsup_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0^*} \mathbb{P}_P \left( \exists k \geq m : \bar{\Gamma}_k^{(m)} = 1 \right). \quad (8.51)$$

In other words, no  $\mathcal{P}_0^*$ -uniform anytime-valid test can have power against any alternative in  $\mathcal{P}_1^*$  at any  $\{m, m+1, \dots\}$ -valued stopping time no matter how large  $m$  is.

The proof can be found in Section 8.A.4. It should be noted that Proposition 8.4.1 is not an immediate consequence of S&P's fixed- $n$  hardness result in (8.42) since while it is true that the time-uniform *type-I error* in the right-hand side of (8.51) is always larger than its fixed- $n$  counterpart, the time-uniform *power* in the left-hand side of (8.51) is typically much larger than the fixed- $n$  power. Indeed, while an important facet of hypothesis testing is to find tests with power as close to 1 as possible, the time-uniform power of anytime-valid tests is typically *equal to 1*, and such tests are sometimes referred to explicitly as "tests of power 1" for this reason [222]. This should not be surprising since the ability to reject at any stopping time (data-dependent sample size) larger than  $m$  introduces a great deal of flexibility. The fact that this flexibility is insufficient to overcome  $\mathcal{P}_0^*$ -uniform control of the time-uniform type-I error is what makes Proposition 8.4.1 nontrivial.

Using the techniques of Section 8.2, we will now derive an anytime-valid analogue of S&P's GCM test with similar distribution-uniform guarantees, allowing the tests and  $p$ -values to be continuously monitored and adaptively stopped.

#### 8.4.4 SeqGCM: The sequential generalized covariance measure test

We will now lay out the assumptions required for our SeqGCM test to have distribution-uniform anytime-validity. Similar to our discussion of the batch GCM test in the previous section, we will assume that for each  $n$ ,  $\hat{\mu}_n^x$  and  $\hat{\mu}_n^y$  are trained from an independent sample. This can be achieved easily by supposing that at each time  $n$ , we observe pairs  $(X_1^{(n)}, Y_1^{(n)}, Z_1^{(n)}), (X_2^{(n)}, Y_2^{(n)}, Z_2^{(n)})$  where the first is used for training  $(\hat{\mu}_i^x, \hat{\mu}_i^y)_{i=n}^\infty$  and the second is used for evaluating  $\{X_n - \hat{\mu}_n^x(Z_n)\} \cdot \{Y_n - \hat{\mu}_n^y(Z_n)\}$ .

Recall that in S&P's GCM test, the test statistic  $\dot{GCM}_n := \frac{1}{n} \sum_{i=1}^n R_{i,n} / \hat{\sigma}_n^2$  was built from the product residuals  $R_{i,n}$  that were defined in (8.43) as

$$R_{i,n} := \{X_i - \hat{\mu}_n^x(Z_i)\} \{Y_i - \hat{\mu}_n^y(Z_i)\}. \quad (8.52)$$

In particular, note that the regression estimators  $\hat{\mu}_n^x$  and  $\hat{\mu}_n^y$  are trained *once* on a held-out sample of size  $n$  and then evaluated on  $Z_1, \dots, Z_n$ , which is perfectly natural in the batch setting. By contrast, we will evaluate the product residual

$$R_n := \{X_n - \hat{\mu}_n^x(Z_n)\} \{Y_n - \hat{\mu}_n^y(Z_n)\} \quad (8.53)$$

to arrive at the test statistic

$$\overline{\text{GCM}}_n := \frac{1}{n\widehat{\sigma}_n} \sum_{i=1}^n R_i, \quad (8.54)$$

where we will abuse notation slightly and redefine  $\widehat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n R_i^2 - (\frac{1}{n} \sum_{i=1}^n R_i)^2$ . The main difference between (8.52) and (8.53) is that in the latter case, the index for regression estimators  $(\widehat{\mu}_n^x, \widehat{\mu}_n^y)$  is the same as those on which these functions are evaluated. Notice that while  $\overline{\text{GCM}}_n$  is more amenable to online updates than  $\text{GCM}_n$ , it does less to exploit the most up-to-date regression estimates. Nevertheless, as we will see shortly, it is still possible to control the distribution- and time-uniform asymptotic behavior of  $\overline{\text{GCM}}_n$  under *weak* regression consistency conditions on  $(\widehat{\mu}_n^x, \widehat{\mu}_n^y)$ . This is in contrast to some earlier work found in Chapter 7 that also considered asymptotic time-uniform inference with nuisance estimation (focusing on the problem of average treatment effect estimation), but relied on *strong* regression consistency conditions. It should be noted that the weak consistency rates we impose here are polylogarithmically faster than those considered in Chapter 7. The key technique that will allow us to derive *strong* convergence behavior of certain sample averages of nuisances from *weak* consistency of regression functions is a distribution-uniform strong law of large numbers found in Chapter 9. This will be discussed further after the statement of Theorem 8.4.1.

Since the assumptions required for our SeqGCM test are similar in spirit to those of S&P's batch GCM test (Assumptions **GCM-1**, **GCM-2**, and **GCM-3**) we correspondingly name them "Assumptions **SeqGCM-1** and **SeqGCM-2**" and underline certain keywords to highlight their differences (we do not need to make additional moment assumptions beyond those found in Assumption **GCM-3**, and thus there is no "SeqGCM-3" to introduce).

**Assumption SeqGCM-1** (Product regression error decay). *The weak convergence rate of the product of average squared residuals is no slower than  $(n \log^{2+\delta} n)^{-1/2}$  for some  $\delta > 0$ , i.e.*

$$\sup_{P \in \mathcal{P}_0} \|\widehat{\mu}_n^x - \mu^x\|_{L_2(P)} \cdot \|\widehat{\mu}_n^y - \mu^y\|_{L_2(P)} = O\left(\frac{1}{\sqrt{n \log^{2+\delta}(n)}}\right), \quad (8.55)$$

**Assumption SeqGCM-2** ( $\mathcal{P}_0$ -uniform regularity of regression errors). *Both  $\text{Var}(\{\widehat{\mu}_n^x(Z) - \mu^x(Z)\} \cdot \xi_n^y)$  and  $\text{Var}(\{\widehat{\mu}_n^y(Z) - \mu^y(Z)\} \cdot \xi_n^x)$  are  $\mathcal{P}_0$ -uniformly vanishing to 0 no slower than  $1/(\log n)^{2+\delta}$  for some  $\delta > 0$ , i.e.*

$$\sup_{P \in \mathcal{P}_0} \text{Var}_P(\{\widehat{\mu}_n^x(Z) - \mu^x(Z)\} \cdot \xi_n^y) = O\left(\frac{1}{(\log n)^{2+\delta}}\right) \quad (8.56)$$

$$\text{and} \quad \sup_{P \in \mathcal{P}_0} \text{Var}_P(\{\widehat{\mu}_n^y(Z) - \mu^y(Z)\} \cdot \xi_n^x) = O\left(\frac{1}{(\log n)^{2+\delta}}\right). \quad (8.57)$$

With Assumptions **SeqGCM-1**, **SeqGCM-2**, and **GCM-3** in mind, we are ready to state the  $\mathcal{P}_0$ -uniform type-I error guarantees of the SeqGCM test.

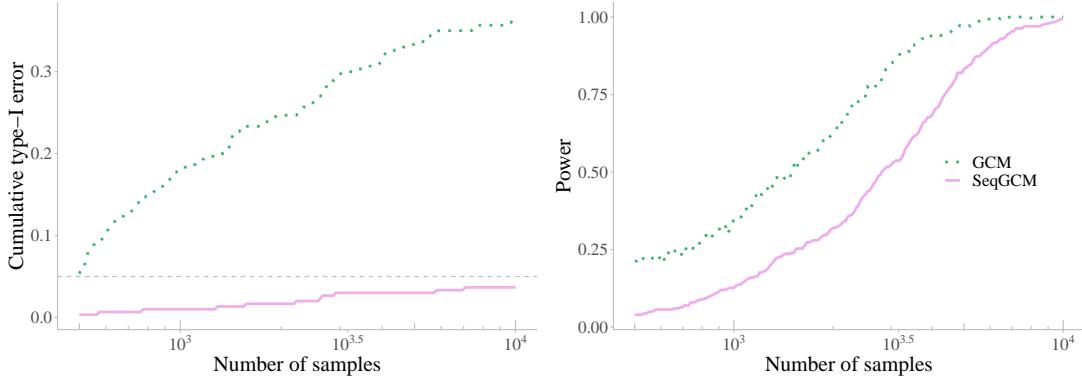


Figure 8.1: Empirical cumulative type-I error rates and power for the fixed- $n$  GCM test of S&P versus the sequential GCM test (SeqGCM) in Theorem 8.4.1 with a target type-I error of  $\alpha = 0.05$  in a simulated conditional independence testing problem. Notice that in the left-hand side plot, the type-I error rate for the GCM starts at around  $\alpha = 0.05$  but steadily grows as more samples are collected. By contrast, the SeqGCM test remains below  $\alpha = 0.05$  for all  $k \geq m = 300$ . In the right-hand side plot, we see that the power of the GCM test is higher than that of SeqGCM. This is unsurprising given that SeqGCM has a stronger (time-uniform) type-I error guarantee, but both have power near 1 after 10,000 samples.

**Theorem 8.4.1** ( $\mathcal{P}_0$ -uniform type-I error control of the SeqGCM). *Suppose  $(X_i, Y_i, Z_i)_{i=1}^\infty$  are  $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ -valued triplets defined on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  and let  $\mathcal{P}_0 \subseteq \mathcal{P}$  be a collection of distributions in  $\mathcal{P}$  satisfying the conditional independence null  $H_0$  and Assumption SeqGCM-1, SeqGCM-2, and GCM-3. Define*

$$\bar{p}_{k,m}^{\text{GCM}} := 1 - \Psi(k(\overline{\text{GCM}}_k)^2 - \log(k/m)). \quad (8.58)$$

*Then  $(\bar{p}_{k,m}^{\text{GCM}})_{k=m}^\infty$  forms a  $\mathcal{P}_0$ -uniform anytime p-value for the conditional independence null:*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sup_{\alpha \in (0,1)} |\mathbb{P}_P(\exists k \geq m : \bar{p}_{k,m}^{\text{GCM}} \leq \alpha) - \alpha| = 0. \quad (8.59)$$

The proof can be found in Section 8.A.5 and uses the results from the previous sections combined with a distribution-uniform strong laws of large numbers (SLLNs) for independent but non-identically distributed random variables (Theorem 9.2.2 in Chapter 9). The latter is crucial to analyzing the (uniform) almost sure convergence properties of sample averages with online regression estimators under weak consistency assumptions (SeqGCM-1 and SeqGCM-2).

To give some intuition as to when Assumption SeqGCM-1 may be satisfied, suppose that  $\mu^x$  and  $\mu^y$  are  $d$ -dimensional and Hölder  $s$ -smooth [114, §3.2]. Note that the minimax rate for estimating such functions in the resulting class of distributions  $\mathcal{P}(s)$  is given by

$$\inf_{\hat{\mu}_n^x} \sup_{P \in \mathcal{P}(s)} \mathbb{E}_P \|\hat{\mu}_n^x - \mu^x\|_{L_2(P)}^2 \asymp n^{-2s/(2s+d)}, \quad (8.60)$$

and similarly for  $\mu^y$ . In particular, if  $d < 2s$  so that the dimension is not too large relative to the smoothness, then minimax-optimal local polynomial estimators  $\hat{\mu}_n^x$  and  $\hat{\mu}_n^y$  for  $\mu^x$  and  $\mu^y$  can be constructed and will be  $\mathcal{P}(s)$ -weakly consistent at rates of  $o((n \log^{2+\delta} n)^{-1/4})$ . In this case, Assumption SeqGCM-1 (and Assumption GCM-1) will be satisfied as long as  $\mathcal{P}_0 \subseteq \mathcal{P}(s)$ . More broadly, any regression algorithms can be used to construct  $\hat{\mu}_n^x$  and  $\hat{\mu}_n^y$  (e.g. using random forests, neural networks, nearest neighbors, etc.) and they can further be selected via cross-validation or aggregated [42, 255].

The left-hand side plot of Figure 8.1 demonstrates how the SeqGCM test controls the type-I error rate under the null uniformly over time while the standard GCM test fails to. The right-hand side plot compares their empirical power under one alternative.

## 8.5 Distribution-uniform strong Gaussian approximation

In this section, we both articulate what it means for a strong (almost-sure) coupling to be “ $\mathcal{P}$ -uniform” and then provide such a coupling in the form of a strong Gaussian approximation in Theorem 8.5.2. Before that, however, let us give a brief historical overview of weak and strong Gaussian approximations in the  $P$ -pointwise setting to contextualize and motivate the result to come. Given iid random variables  $(X_1, \dots, X_n)$  with mean  $\mu$  and finite variance  $\sigma^2$  on a probability space  $(\Omega, \mathcal{F}, P)$ , the CLT states that standardized partial sums  $S_n := \sum_{i=1}^n (X_i - \mu)/\sigma$  converge in distribution to a standard Gaussian with CDF  $\Phi(x)$  after  $\sqrt{n}$ -rescaling:

$$\forall x \in \mathbb{R}, \lim_{n \rightarrow \infty} \mathbb{P}_P(S_n/\sqrt{n} \leq x) \rightarrow \Phi(x). \quad (8.61)$$

Note that (8.61) is only a statement about the *distribution* of  $S_n$ , but a stronger statement can be made in terms of a *coupling* between  $S_n$  and a partial sum of iid Gaussians [101, Eq. (1.2)]. Concretely, one can define a new probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$  containing random vectors  $((\tilde{X}_1, Y_1), (\tilde{X}_2, Y_2), \dots, (\tilde{X}_n, Y_n))$  where  $(Y_1, \dots, Y_n)$  are marginally standard Gaussian and  $(\tilde{X}_1, \dots, \tilde{X}_n)$  have the same marginal distribution as  $(X_1, \dots, X_n)$  so that

$$\tilde{S}_n - G_n = o_{\tilde{P}}(\sqrt{n}), \quad (8.62)$$

where  $\tilde{S}_n := \sum_{i=1}^n \tilde{X}_i$  and  $G_n := \sum_{i=1}^n Y_i$  and without loss of generality, we may simply write  $S_n - G_n = o_P(\sqrt{n})$ . Indeed, we could have simply started with a probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$  rich enough to describe  $(X, Y)$  jointly and for this reason, some authors write “without loss of generality” to refer to this probability space construction [249]. Crucially,  $Y_1, \dots, Y_n$  are independent of each other, but the random variables  $S_n$  and  $G_n$  are highly dependent, and clearly (8.62)  $\implies$  (8.61).

For the purposes of obtaining a *time-uniform* guarantee, however, neither (8.61) nor (8.62) are sufficient since they only hold for a single sample size  $n$ , and naive union bounds over  $n \in \mathbb{N}$  are not sharp enough to remedy the issue. Fortunately, there do exist analogues of (8.62) that hold *almost-surely* and hence uniformly for all  $n$  simultaneously. The study of such results – *strong Gaussian approximations* – began with the seminal results of Strassen [248] who used

the Skorokhod embedding [243] (see also [36, p. 513]) to obtain an almost-sure analogue of (8.62) but with an iterated logarithm rate:

$$S_n - G_n = o_{\text{a.s.}}(\sqrt{n \log \log n}), \quad (8.63)$$

where  $o_{\text{a.s.}}(\cdot)$  denotes  $P$ -a.s. convergence. As noted in (8.23), the above is equivalent to saying that for any  $\varepsilon > 0$ , we have  $\lim_{m \rightarrow \infty} \mathbb{P}_P(\exists k \geq m : |S_k - G_k| > \varepsilon) = 0$ , and we write this as

$$S_n - G_n = \bar{o}_P(\sqrt{n \log \log n}), \quad (8.64)$$

following the notation laid out in Section 8.3. Improvements to the iterated logarithm rate in (8.63) and (8.64) were made by Strassen [249] under higher moment assumptions, with the optimal rates uncovered in the famous papers by Komlós, Major, and Tusnády [160, 161] and Major [181].

Here, we do not focus on attaining optimal coupling rates since error rates incurred from estimation of nuisances (such as the variance) typically dominate them and optimal rates would not change our main statistical results in any meaningful way (much like they do not “improve” CLT-based confidence intervals). However, we do highlight the fact that the results of Strassen [248, 249], Komlós, Major, and Tusnády [160, 161], Major [181], and every other work on strong approximation to our knowledge only hold  $P$ -a.s. for a fixed  $P$ , and hence are not  $\mathcal{P}$ -uniform in any sense. We will now define “distribution-uniform strongly coupled processes” in Definition 8.5.1 and subsequently provide one such coupling in Theorem 8.5.2.

*Definition 8.5.1* (( $\mathcal{P}, n$ )-uniformly coupled stochastic processes). For each probability measure  $P$  in a collection  $\mathcal{P}$ , let  $(S_n(P))_{n=1}^\infty$  be a stochastic process defined on the probability space  $(\Omega, \mathcal{F}, P)$ . Let  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}(P))_{P \in \mathcal{P}}$  be a new collection of probability spaces containing stochastic processes  $(\tilde{S}_n(P))_{n=1}^\infty$  and  $(G_n)_{n=1}^\infty$  so that  $(\tilde{S}_n(P))_{n=1}^\infty$  has the same distribution as  $(S_n(P))_{n=1}^\infty$  for each  $P \in \mathcal{P}$ . We say that  $(S_n(P))_{n=1}^\infty$  and  $(G_n)_{n=1}^\infty$  are ( $\mathcal{P}, n$ )-uniformly coupled at a rate of  $r_n$  if for every  $\varepsilon > 0$ ,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_{\tilde{P}(P)} \left( \exists k \geq m : \frac{|\tilde{S}_k(P) - G_k|}{r_k} \geq \varepsilon \right) = 0, \quad (8.65)$$

and we write  $S_n - G_n = \bar{o}_{\mathcal{P}}(r_n)$  as a shorthand for (8.65).

Since time-uniform convergence with high probability and almost-sure convergence — denoted by  $o_{\text{a.s.}}(\cdot)$  and  $\bar{o}_P(\cdot)$  respectively — are equivalent, observe that Definition 8.5.1 reduces to the standard notion of  $P$ -a.s. strong approximation when  $\mathcal{P} = \{P\}$  is a singleton. To avoid repeating the technicalities of constructing a new probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}(P))$  with equidistributed random variables and so on, some authors in the strong approximation literature refer to this procedure as “the construction” [99, 100] and they will say that “there exists a construction such that  $S_n - G_n = o_{\text{a.s.}}(r_n)$ ” as a shorthand. We henceforth adopt and extend this convention to the  $\mathcal{P}$ -uniform setting by writing “there exists a construction such that  $S_n - G_n = \bar{o}_{\mathcal{P}}(r_n)$ ”. Let us now give a strong Gaussian approximation for partial sums of

random variables with finite  $(2 + \delta)^{\text{th}}$  finite absolute moments.

**Theorem 8.5.2** (Distribution-uniform strong Gaussian approximation). *Let  $(X_n)_{n=1}^{\infty}$  be independent and identically distributed random variables defined on the collection of probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  with means  $\mu_P := \mathbb{E}_P(X)$  and variances  $\sigma_P^2 := \mathbb{E}_P(X - \mu_P)^2$ . If  $X$  has  $q > 2$  uniformly upper-bounded moments, and a uniformly positive variance, i.e.*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P |X - \mu_P|^q < \infty \quad \text{and} \quad \inf_{P \in \mathcal{P}} \sigma_P^2 > 0, \quad (8.66)$$

*then there exists a construction with independent standard Gaussians  $(Y_n)_{n=1}^{\infty} \sim N(0, 1)$  so that*

$$\left| \sum_{i=1}^n \frac{X_i - \mu_P}{\sigma_P} - \sum_{i=1}^n Y_i \right| = \bar{o}_{\mathcal{P}}(n^{1/q} \log^{2/q}(n)). \quad (8.67)$$

We remark that Theorem 8.5.2 is a purely probabilistic result that may be of interest outside of statistical inference altogether. To the best of our knowledge, Theorem 8.5.2 serves as the first *distribution-uniform strong Gaussian approximation* in the literature. Note that the rate (8.67) is optimal up to a factor of  $\log^{2/q}(n)$  compared to the best rate possible in the  $P$ -pointwise setting and in fact  $\log^{2/q}(n)$  can be replaced by any  $f(n)^{1/q}$  as long as  $\sum_{n=1}^{\infty} (nf(n))^{-1} < \infty$ . As we alluded to before, improvements to this rate would *not* advance the statistical inference goals of this chapter. The reason behind this is that strong approximation rates are often dominated by the rates of errors incurred from estimating nuisance functions such as the variance (which is often of order  $\sqrt{\log \log n/n}$  or slower). Nevertheless, in future work we will explore rate-optimal analogues of Theorem 8.5.2 in a thorough study of distribution-uniform strong approximations but we keep the current version here because it is sufficient for the current chapter's objectives.

In fact, the strong approximation of Theorem 8.5.2 is a corollary of the following more general *nonasymptotic* high-probability strong Gaussian coupling inequality for independent (but not necessarily identically distributed) random variables that depends on features of the distribution of  $X$  in transparent ways.

**Lemma 8.5.1** (Strong Gaussian coupling inequality). *Let  $(X_n)_{n=1}^{\infty}$  be independent random variables on the probability space  $(\Omega, \mathcal{F}, \mathcal{P})$ . Suppose that for some  $q \geq 2$ , we have  $\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q < \infty$  for each  $k \in \mathbb{N}$ . Let  $f(\cdot)$  be a positive and increasing function so that  $\sum_{n=1}^{\infty} (nf(n))^{-1} < \infty$  and*

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q / \sigma_k^q}{kf(k)} < \infty, \quad (8.68)$$

*where  $\sigma_k^2 := \text{Var}_P(X_k)$ . Then one can construct a probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathcal{P}}(P))$  rich enough to define  $(\tilde{X}_n, Y_n)_{n=1}^{\infty}$  where  $\tilde{X}_n$  and  $X_n$  are equidistributed for each  $n$  and  $(Y_n)_{n=1}^{\infty}$  are marginally*

independent standard Gaussians so that for any  $\varepsilon > 0$ ,

$$\mathbb{P}_{\tilde{P}(P)} \left( \exists k \geq m : \left| \frac{\sum_{i=1}^k (\tilde{X}_i - Y_i)}{k^{1/q} f(k)^{1/q}} \right| > \varepsilon \right) \leq \frac{C_{q,f}}{\varepsilon^q} \left\{ \sum_{k=2^{m-1}}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q / \sigma_k^q}{k f(k)} + \frac{1}{2^m} \sum_{k=1}^{2^m-1} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q / \sigma_k^q}{k f(k)} \right\}, \quad (8.69)$$

where  $C_{q,f}$  is a constant that depends only on  $q$  and  $f$ .

Instantiating Lemma 8.5.1 in the identically distributed case with  $q = 2 + \delta$  for some  $\delta > 0$  and taking suprema over  $P \in \mathcal{P}$  on both sides of (8.69) yields Theorem 8.5.2. The proofs of Lemma 8.5.1 and Theorem 8.5.2 can be found in Section 8.A.7.

A straightforward consequence of Lemma 8.5.1 and Theorem 8.5.2 is that the law of the iterated logarithm holds uniformly in a class of distributions with uniformly bounded  $(2 + \delta)^{\text{th}}$  moments.

**Corollary 8.5.1** (A  $\mathcal{P}$ -uniform law of the iterated logarithm). *Suppose  $(X_n)_{n=1}^{\infty}$  are defined on probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  where  $\mathcal{P}$  is a collection of distributions such that*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P |X - \mathbb{E}_P X|^{2+\delta} < \infty \text{ and } \inf_{P \in \mathcal{P}} \text{Var}_P(X) > 0 \quad (8.70)$$

for some  $\delta > 0$ . Then,

$$\sup_{k \geq n} \frac{|\sum_{i=1}^k (X_i - \mathbb{E}_P(X))|}{\sqrt{2\text{Var}_P(X)k \log \log k}} = 1 + \bar{o}_{\mathcal{P}}(1). \quad (8.71)$$

A proof of Corollary 8.5.1 is provided in Section 8.A.6 and follows from Theorem 8.5.2 combined with Kolmogorov's  $P$ -pointwise law of the iterated logarithm.

## 8.6 Summary & discussion

We gave a definition of “distribution-uniform anytime-valid inference” as a time-uniform analogue of distribution-uniform fixed- $n$  inference and then derived explicit hypothesis tests,  $p$ -values, and confidence sequences satisfying that definition. Our methods relied on a novel boundary for centered partial sums that is uniformly valid in a class of distributions, in time, and in a family of boundaries. Along the way, we discussed what it meant for a sequence of random variables to converge distribution-uniformly almost-surely, and provided definitions for distribution- and time-uniform stochastic boundedness alongside a calculus for manipulating sequences with these types of asymptotics. At their core, all of our results relied on a novel strong Gaussian approximation that allows a partial sum process to be tightly coupled with an implicit Gaussian process uniformly in time and in a class of distributions. We believe this is

the first result of its kind in the literature. Zooming out, we believe that this strong Gaussian approximation forms the tip of the iceberg for distribution-uniform strong laws. In future work, we plan to study these problems in depth.

## 8.A Proofs of the main results

In the proofs to come, we will make extensive use of the notions of convergence in Table 8.1, especially  $\bar{o}_{\mathcal{P}}(\cdot)$  and  $\bar{O}_{\mathcal{P}}(\cdot)$ . However, some of our terms will be converging or asymptotically bounded with respect to different indices – e.g. there may be two sequences  $(X_n)_{n=1}^{\infty}$  and  $(Y_k)_{k=1}^{\infty}$  with indices  $n$  and  $k$  that are diverging to  $\infty$  not necessarily together (e.g., imagine  $k = n^2$ ). Writing  $X_n = \bar{o}_{\mathcal{P}}(r_n)$  and  $Y_k = \bar{o}_{\mathcal{P}}(r_k)$  is unambiguous, for example, but when no rate is specified, we will remove ambiguity with respect to indices  $n$  or  $k$  by saying  $X_n = \bar{o}_{\mathcal{P}}^{(n)}(1)$  and  $Y_k = \bar{o}_{\mathcal{P}}^{(k)}(1)$ .

### 8.A.1 Proof of Proposition 8.2.2

**Proposition 8.2.2** (( $\mathcal{P}, n, x$ )-uniform boundaries for centered partial sums). *Let  $X_1, X_2, \dots$  be random variables defined on probability spaces  $(\Omega, \mathcal{F}, \mathcal{P}^*)$  with finite  $(2 + \delta)^{th}$  moments, i.e.  $\mathbb{E}_P|X - \mathbb{E}_P(X)|^{2+\delta} < \infty$  for every  $P \in \mathcal{P}^*$ . Letting  $S_n := \sum_{i=1}^n (X_i - \mathbb{E}_P(X_i))/\sigma_P$  be their centered partial sums, we have*

$$\forall P \in \mathcal{P}^*, \lim_{m \rightarrow \infty} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \exists k \geq m : |S_k|/\sqrt{k} \geq \sqrt{x + \log(k/m)} \right) - [1 - \Psi(x)] \right| = 0. \quad (8.17)$$

Furthermore, if  $\mathcal{P} \subseteq \mathcal{P}^*$  is a sub-collection of distributions for which the  $(2 + \delta)^{th}$  moment is  $\mathcal{P}$ -uniformly upper-bounded and the variance is  $\mathcal{P}$ -uniformly positive, then the above limit holds  $\mathcal{P}$ -uniformly:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \exists k \geq m : |S_k|/\sqrt{k} \geq \sqrt{x + \log(k/m)} \right) - [1 - \Psi(x)] \right| = 0. \quad (8.18)$$

*Proof.* Let  $\sigma^2 > 0$  be a uniform lower bound on  $\inf_{P \in \mathcal{P}} \text{Var}_P(X)$ . Writing out  $\sup_{k \geq m} \{S_k^2/\sigma_P^2 k - \log(k/m)\}$  and invoking the strong Gaussian coupling of Theorem 8.5.2, we have on a potentially enriched probability space a partial sum  $G_n := \sum_{i=1}^n Y_i$  of standard Gaussians  $Y_1, \dots, Y_n \sim N(0, 1)$  so that for some  $q = 2 + \delta/2$  (say),

$$\sup_{k \geq m} \{S_k^2/\sigma_P^2 k - \log(k/m)\} = \sup_{k \geq m} \left\{ \left( \sigma_P G_k + \bar{o}_{\mathcal{P}}(k^{1/q}) \right)^2 / (\sigma_P^2 k) - \log(k/m) \right\} \quad (8.72)$$

$$= \sup_{k \geq m} \left\{ \frac{\sigma_P^2 G_k^2 + \bar{O}_{\mathcal{P}}(k^{1/q} \sqrt{k \log \log k}) + \bar{o}_{\mathcal{P}}(k^{2/q})}{\sigma_P^2 k} - \log(k/m) \right\} \quad (8.73)$$

$$= \sup_{k \geq m} \left\{ \frac{G_k^2}{k} + \frac{1}{\sigma^2} \bar{O}_{\mathcal{P}} \left( \sqrt{\frac{\log \log k}{k^{1-2/q}}} \right) + \frac{1}{\sigma^2} \bar{o}_{\mathcal{P}} \left( \frac{k^{2/q}}{k} \right) - \log(k/m) \right\} \quad (8.74)$$

$$= \sup_{k \geq m} \left\{ \frac{G_k^2}{k} - \log(k/m) + \bar{o}_{\mathcal{P}}^{(k)}(1) \right\}, \quad (8.75)$$

where (8.73) expands the square and applies the ( $\mathcal{P}$ -uniform) law of the iterated logarithm (Corollary 8.5.1) to  $(G_n)_{n=1}^\infty$ , (8.74) uses the  $\mathcal{P}$ -uniform lower-boundedness of the variance, and (8.75) consolidates the  $\bar{o}_{\mathcal{P}}(\cdot)$  terms. Now, notice that  $\sup_{k \geq m} \{G_k^2/k - \log(k/m)\}$  converges uniformly to the Robbins-Siegmund distribution (Lemma 8.B.2) since the distribution of the supremum does not depend on any measure  $P$ . That is,

$$\forall x \geq 0, \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{G_k^2}{k} - \log(k/m) \right\} \leq x \right) - \Psi(x) \right| = 0. \quad (8.76)$$

Applying van der Vaart [260, Lemma 2.11] and using the fact that  $\Psi$  is continuous, we have that the above also holds uniformly in  $x \geq 0$ :

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{G_k^2}{k} - \log(k/m) \right\} \leq x \right) - \Psi(x) \right| = 0. \quad (8.77)$$

Some algebraic manipulations will reveal that the above is equivalent to the desired result:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \exists k \geq m : |S_k|/\sqrt{k} \geq \sqrt{x + \log(k/m)} \right) - [1 - \Psi(x)] \right| = 0, \quad (8.78)$$

which completes the proof.  $\square$

### 8.A.2 Proof of Lemma 8.3.1

**Lemma 8.3.1** (Calculus of  $\bar{O}_{\mathcal{P}}(\cdot)$  and  $\bar{o}_{\mathcal{P}}(\cdot)$ ). *Let  $Y_1, Y_2, \dots$  be random variables defined on  $(\Omega, \mathcal{F}, \mathcal{P})$ . Let  $(a_n)_{n=1}^\infty$  and  $(b_n)_{n=1}^\infty$  be positive and monotonically nonincreasing sequences. Then we have the following basic implications:*

$$Y_n = \bar{o}_{\mathcal{P}}(a_n) \implies Y_n = \bar{O}_{\mathcal{P}}(a_n) \quad (8.27)$$

$$Y_n = \bar{o}_{\mathcal{P}}(a_n) \bar{O}_{\mathcal{P}}(b_n) \implies Y_n = \bar{o}_{\mathcal{P}}(a_n b_n) \quad (8.28)$$

$$Y_n = \bar{O}_{\mathcal{P}}(a_n) \bar{O}_{\mathcal{P}}(b_n) \implies Y_n = \bar{O}_{\mathcal{P}}(a_n b_n) \quad (8.29)$$

$$Y_n = \bar{o}_{\mathcal{P}}(a_n) + \bar{O}_{\mathcal{P}}(a_n) \implies Y_n = \bar{O}_{\mathcal{P}}(a_n) \quad (8.30)$$

$$Y_n = \bar{o}_{\mathcal{P}}(a_n) + \bar{o}_{\mathcal{P}}(b_n) \implies Y_n = \bar{o}_{\mathcal{P}}(\max\{a_n, b_n\}). \quad (8.31)$$

Furthermore, (8.31) holds with  $\bar{o}_{\mathcal{P}}(\cdot)$  replaced by  $\bar{O}_{\mathcal{P}}(\cdot)$  on both sides. Finally, if  $Y_n = \bar{O}_{\mathcal{P}}(a_n)$  and  $a_n/b_n \rightarrow 0$ , then  $Y_n = \bar{o}_{\mathcal{P}}(b_n)$ .

**Proof of (8.27)** Suppose that  $Y_n = \bar{o}_{\mathcal{P}}(a_n)$ . We want to show that for any  $\delta$ , there exists  $C \equiv C(\delta)$  and  $M \equiv M(\delta)$  so that for all  $m \geq M$ ,

$$\text{Goal: } \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |a_k^{-1} Y_k| \geq C \right) < \delta. \quad (8.79)$$

*Proof.* This is immediate from the definition of  $\bar{o}_{\mathcal{P}}(a_n)$ . Indeed, fix any  $\varepsilon > 0$  and choose  $M \equiv M(\varepsilon)$  so that for any  $m \geq M$ ,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |a_k^{-1} Y_k| \geq \varepsilon \right) < \delta. \quad (8.80)$$

Identifying  $C$  with  $\varepsilon$  completes the proof.  $\square$

**Proof of (8.28).** Suppose that  $Y_n = A_n B_n$  with  $A_n = \bar{o}_{\mathcal{P}}(a_n)$  and  $B_n = \bar{o}_{\mathcal{P}}(b_n)$ . We want to show that  $a_n^{-1} b_n^{-1} Y_n = \bar{o}_{\mathcal{P}}(1)$ . More formally, our goal is to show that for arbitrary  $\varepsilon, \delta > 0$ , there exists  $M \equiv M(\varepsilon, \delta) \geq 1$  so that for all  $m \geq M$ ,

$$\text{Goal: } \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : |a_k^{-1} b_k^{-1} Y_k| \geq \varepsilon \right) < \delta. \quad (8.81)$$

*Proof.* Choose  $M$  sufficiently large so that for all  $m \geq M$ ,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |a_k^{-1} A_k| \geq \sqrt{\varepsilon/2} \right) < \delta \quad \text{and} \quad \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |b_k^{-1} B_k| \geq \sqrt{\varepsilon/2} \right) < \delta. \quad (8.82)$$

Then, writing out the equation in (8.81), we have that

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : |a_k^{-1} b_k^{-1} Y_k| \geq \varepsilon \right) \quad (8.83)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : |a_k^{-1} A_k| |b_k^{-1} B_k| \geq \varepsilon \right) \quad (8.84)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : |a_k^{-1} A_k| |b_k^{-1} B_k| \geq \varepsilon \mid \sup_{k \geq m} |a_k^{-1} A_k| < \sqrt{\varepsilon/2} \text{ and } \sup_{k \geq m} |b_k^{-1} B_k| < \sqrt{\varepsilon/2} \right) + \quad (8.85)$$

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |a_k^{-1} A_k| < \sqrt{\varepsilon/2} \text{ and } \sup_{k \geq m} |b_k^{-1} B_k| < \sqrt{\varepsilon/2} \right) \quad (8.86)$$

$$\leq \underbrace{\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : \varepsilon/2 \geq \varepsilon \right)}_{=0} + \quad (8.87)$$

$$\underbrace{\max \left\{ \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |a_k^{-1} A_k| < \sqrt{\varepsilon/2} \right), \mathbb{P} \left( \sup_{k \geq m} |b_k^{-1} B_k| < \sqrt{\varepsilon/2} \right) \right\}}_{< \delta} \quad (8.88)$$

$$< \delta, \quad (8.89)$$

which completes the proof.  $\square$

**Proof of (8.29).** Suppose that  $Y_n = A_n B_n$  with  $A_n = \bar{O}_{\mathcal{P}}(a_n)$  and  $B_n = \bar{O}_{\mathcal{P}}(b_n)$ . Our goal is to show that for any  $\delta > 0$ , there exists some  $C \equiv C(\delta)$  and  $M \equiv M(C, \delta)$  so that

$$\text{Goal: } \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |a_n^{-1} b_n^{-1} Y_n| > C \right) < \delta. \quad (8.90)$$

*Proof.* Fix  $\delta > 0$ . Let  $C_a, M_a, C_b, M_b$  be sufficiently large so that for all  $m \geq \max\{M_a, M_b\}$ ,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |a_k^{-1} A_k| \geq M_a \right) < \delta \quad \text{and} \quad \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |b_k^{-1} B_k| \geq M_b \right) < \delta. \quad (8.91)$$

Now, set  $C = C_a C_b + 1$ . Then,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |a_k^{-1} b_k^{-1} Y_k| \geq C \right) \quad (8.92)$$

$$\leq \sup_{P \in \mathcal{P}} \left( \sup_{k \geq m} |a_k^{-1} A_k| |b_k^{-1} B_k| \geq C \right) \quad (8.93)$$

$$\leq \sup_{P \in \mathcal{P}} \left( \sup_{k \geq m} C_a C_b \geq C \right) + \sup_{P \in \mathcal{P}} \left( \sup_{k \geq m} |a_k^{-1} A_k| > C_a \text{ and } |b_k^{-1} B_k| > C_b \right) \quad (8.94)$$

$$\leq \underbrace{\sup_{P \in \mathcal{P}} \left( \sup_{k \geq m} C_a C_b \geq C_a C_b + 1 \right)}_{=0} + \quad (8.95)$$

$$\underbrace{\max \left\{ \sup_{P \in \mathcal{P}} \left( \sup_{k \geq m} |a_k^{-1} A_k| > C_a \right), \sup_{P \in \mathcal{P}} \left( \sup_{k \geq m} |b_k^{-1} B_k| > C_b \right) \right\}}_{< \delta}, \quad (8.96)$$

which completes the proof.  $\square$

**Proof of (8.30).** Suppose  $Y_n = A_n + A'_n$  with both  $A_n = \bar{o}_{\mathcal{P}}(a_n)$  and  $A'_n = \bar{O}_{\mathcal{P}}(a_n)$ . The goal is to show that for every  $\delta > 0$ , there exists  $C > 0$  and  $M \geq 1$  so that for all  $m \geq M$ ,

$$\text{Goal: } \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} a_k^{-1} |Y_k| > C \right) < \delta. \quad (8.97)$$

*Proof.* Fix  $\delta > 0$ . Let  $C'$  and  $M'$  be so that  $\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} a_k^{-1} |A'_k| > C' \right) < \delta/2$ . Fix any  $\varepsilon \in (0, C')$  and let  $M^*$  be so that  $\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} a_k^{-1} |A_k| \geq \varepsilon \right) < \delta/2$  for all  $m \geq M^*$ . Choose  $M > \max\{M', M^*\}$ . Then, for all  $m \geq M$ ,

$$\sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} a_k^{-1} |A_k + A'_k| \geq C' \right) \quad (8.98)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} a_k^{-1} |A_k| + a_k |A'_k| \geq C' \right) \quad (8.99)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} a_k^{-1} |A_k| \geq C' \right) + \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} a_k^{-1} |A'_k| \geq C' \right) \quad (8.100)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} a_k^{-1} |A_k| \geq \varepsilon \right) + \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} a_k^{-1} |A'_k| \geq C' \right) \quad (8.101)$$

$$< \delta, \quad (8.102)$$

which completes the proof.  $\square$

**Proof of (8.31).** Suppose  $Y_n = A_n + B_n$  with  $A_n = \bar{o}_{\mathcal{P}}(a_n)$  and  $B_n = \bar{o}_{\mathcal{P}}(b_n)$ . The goal is to show that for every  $\varepsilon, \delta > 0$ , there exists  $M \geq 1$  so that for all  $m \geq M$ ,

$$\text{Goal: } \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} c_k^{-1} |Y_k| \geq \varepsilon \right) < \delta, \quad (8.103)$$

where  $c_k = \max\{a_k, b_k\}$ .

*Proof.* Fix  $\varepsilon, \delta > 0$ . Let  $M$  be so that  $\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} a_k |A_k| > \varepsilon \right) < \delta/2$  and  $\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} b_k |B_k| > \varepsilon \right) < \delta/2$  for all  $m \geq M$ . Then, for all  $m \geq M$ ,

$$\sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} c_k^{-1} |A_k + B_k| \geq \varepsilon \right) \quad (8.104)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} c_k^{-1} |A_k| + c_k^{-1} |B_k| \geq \varepsilon \right) \quad (8.105)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} a_k^{-1} |A_k| + b_k^{-1} |B_k| \geq \varepsilon \right) \quad (8.106)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} a_k^{-1} |A_k| \geq \varepsilon \right) + \sup_{P \in \mathcal{P}} \mathbb{P} \left( \sup_{k \geq m} b_k^{-1} |B_k| \geq \varepsilon \right) \quad (8.107)$$

$$< \delta, \quad (8.108)$$

which completes the proof.  $\square$

**Proof that if  $Y_n = \bar{o}_{\mathcal{P}}(a_n)$  and  $a_n/b_n \rightarrow 0$ , then  $Y_n = \bar{o}_{\mathcal{P}}(b_n)$ .** Let  $\varepsilon, \delta > 0$ . The goal is to show that there exists  $M \equiv M(\varepsilon, \delta) \geq 1$  so that for all  $m \geq M$ ,

$$\text{Goal: } \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} b_k^{-1} |Y_k| \geq \varepsilon \right) < \delta. \quad (8.109)$$

*Proof.* Let  $C > 0$  and  $M_1 \geq 1$  be constants so that  $\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} a_k^{-1} |Y_k| \geq C \right) < \delta$  for all  $m \geq M_1$ . Moreover, choose  $M_2 \geq 1$  so that  $a_k/b_k < \varepsilon/C$  for all  $k \geq M_2$ . Set

$M := \max\{M_1, M_2\}$ . Then, for all  $m \geq M$ ,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} b_k^{-1} |Y_k| \geq \varepsilon \right) \quad (8.110)$$

$$= \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : b_k^{-1} |Y_k| \geq \varepsilon \right) \quad (8.111)$$

$$= \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : a_k^{-1} |Y_k| \geq (b_k/a_k) \varepsilon \right) \quad (8.112)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : a_k^{-1} |Y_k| \geq (C/\varepsilon) \cdot \varepsilon \right) \quad (8.113)$$

$$= \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq m : a_k^{-1} |Y_k| \geq C \right) \quad (8.114)$$

$$< \delta, \quad (8.115)$$

which completes the proof.  $\square$

### 8.A.3 Proof of Theorem 8.3.3

**Theorem 8.3.3** ( $(\mathcal{P}, n, \alpha)$ -uniform statistical inference). *Let  $X_1, X_2, \dots$  be defined on  $(\Omega, \mathcal{F}, \mathcal{P})$  and suppose that for some  $\delta > 0$ , the  $(2 + \delta)^{\text{th}}$  moment is  $\mathcal{P}$ -uniformly upper-bounded and the variance is  $\mathcal{P}$ -uniformly positive. Recall the definitions of  $(\bar{p}_k^{(m)})_{k=m}^{\infty}$  and  $(\bar{C}_k^{(m)}(\alpha))_{k=m}^{\infty}$  from Proposition 8.2.1:*

$$\bar{p}_k^{(m)} := 1 - \Psi \left( k \hat{\mu}_k^2 / \hat{\sigma}_k^2 - \log(k/m) \right) \quad (8.34)$$

$$\text{and } \bar{C}_k^{(m)}(\alpha) := \hat{\mu}_k \pm \hat{\sigma}_k \sqrt{[\Psi^{-1}(1 - \alpha) + \log(k/m)]/k}. \quad (8.35)$$

Let  $\mathcal{P}_0 \subseteq \mathcal{P}$  be a subcollection of distributions so that  $\mathbb{E}_P(X) = 0$  for each  $P \in \mathcal{P}_0$ . Then the time-uniform type-I error of  $(\bar{p}_k^{(m)})_{k=m}^{\infty}$  and the time-uniform miscoverage of  $(\bar{C}_k^{(m)})_{k=m}^{\infty}$  converge to  $\alpha \in (0, 1)$  uniformly in  $\alpha$ , meaning

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sup_{\alpha \in (0, 1)} \left| \mathbb{P}_P \left( \exists k \geq m : \bar{p}_k^{(m)} \leq \alpha \right) - \alpha \right| = 0, \quad \text{and} \quad (8.36)$$

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{\alpha \in (0, 1)} \left| \mathbb{P}_P \left( \exists k \geq m : \mathbb{E}_P(X) \notin \bar{C}_k^{(m)}(\alpha) \right) - \alpha \right| = 0. \quad (8.37)$$

*Proof.* Throughout, denote  $S_n := \sum_{k=1}^n (X_k - \mathbb{E}_P(X))$ . The proof is broken up into two steps. The first (and main) step of the proof shows that  $\sup_{k \geq m} \{S_k^2 / (\hat{\sigma}_k^2 k) - \log(k/m)\}$  converges  $(\mathcal{P}, x)$ -uniformly to the Robbins-Siegmund distribution  $\Psi$ . The second step of the proof uses the first to show how such convergence is equivalent to  $\bar{p}_k^{(m)}$  and  $\bar{C}_k^{(m)}(\alpha)$  forming distribution-uniform anytime-valid  $p$ -values and confidence sequences, respectively, in the senses of Definition 8.2.1.

**Step 1: Establishing the asymptotic distribution of  $\sup_{k \geq m} \{S_k^2/(\hat{\sigma}_k^2 k) - \log(k/m)\}$ .** First, notice that by Proposition 8.3.1,

$$|\hat{\sigma}_n^2 - \sigma^2| = \bar{o}_{\mathcal{P}_0}(1/\log n). \quad (8.116)$$

Letting  $\underline{\sigma}^2 > 0$  be the  $\mathcal{P}$ -uniform lower-bound on the variance so that  $\inf_{P \in \mathcal{P}} \sigma_P^2 \geq \underline{\sigma}^2$ , we therefore have

$$\frac{1}{\hat{\sigma}_n^2} = \frac{1}{\sigma_P^2 + \bar{o}_{\mathcal{P}}(1/\log n)} \quad (8.117)$$

$$= \frac{1}{\sigma_P^2(1 + \underline{\sigma}^{-2} \cdot \bar{o}_{\mathcal{P}}(1/\log n))} \quad (8.118)$$

$$= \frac{1}{\sigma_P^2(1 + \bar{o}_{\mathcal{P}}(1/\log n))}. \quad (8.119)$$

Now, let  $\gamma_n$  be the  $1 + \bar{o}_{\mathcal{P}}(1/\log n)$  term in the above denominator so that  $\hat{\sigma}_n^{-2} = \sigma_P^{-2} \gamma_n^{-1}$ . Writing out  $\sup_{k \geq m} \{S_k^2/(\hat{\sigma}_k^2 k) - \log(k/m)\}$  and using the above, we then have

$$\sup_{k \geq m} \{S_k^2/\hat{\sigma}_n^2 k - \log(k/m)\} = \sup_{k \geq m} \left\{ \frac{S_k^2}{\sigma_P^2 \gamma_k k} - \log(k/m) \right\} \quad (8.120)$$

$$= \sup_{k \geq m} \left\{ \left( \frac{S_k^2}{\sigma_P^2 k} - \gamma_k \log(k/m) \right) \frac{1}{\gamma_k} \right\} \quad (8.121)$$

$$= \sup_{k \geq m} \left\{ \left( \frac{S_k^2}{\sigma_P^2 k} - \log(k/m) + \log(k/m) \cdot \bar{o}_{\mathcal{P}}(1/\log k) \right) \frac{1}{\gamma_k} \right\} \quad (8.122)$$

$$= \sup_{k \geq m} \left\{ \left( \frac{S_k^2}{\sigma_P^2 k} - \log(k/m) + \bar{o}_{\mathcal{P}}^{(k)}(1) \right) \frac{1}{1 + \bar{o}_{\mathcal{P}}^{(k)}(1)} \right\}, \quad (8.123)$$

where (8.123) uses the fact that  $k/m \leq k$  for any  $m \geq 1$ . We will now justify why the above converges  $\mathcal{P}$ - and quantile-uniformly to the Robbins-Siegmund distribution  $\Psi(\cdot)$ . First, by Lemma 8.B.1, we have that  $\sup_{k \geq m} \{S_k^2/\sigma^2 k - \log(k/m)\}$  converges  $\mathcal{P}$ - and quantile-uniformly in distribution to  $\Psi$  as  $m \rightarrow \infty$ . That is,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{S_k^2}{\sigma_P^2 k} - \log(k/m) \right\} \leq x \right) - \Psi(x) \right| = 0. \quad (8.124)$$

By the fact that  $(\mathcal{P}, n, x)$ -uniform convergence to Lipschitz CDFs is preserved under additive  $\bar{o}_{\mathcal{P}}(1)$ -perturbations (Lemma 8.B.4) and the fact that  $\Psi(\cdot)$  is Lipschitz (Lemma 8.B.3), we have that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{S_k^2}{\sigma_P^2 k} - \log(k/m) + \bar{o}_{\mathcal{P}}^{(k)}(1) \right\} \leq x \right) - \Psi(x) \right| = 0. \quad (8.125)$$

Finally, using the fact that  $(\mathcal{P}, n, x)$ -uniform convergence in distribution is preserved under multiplicative  $(1 + \bar{o}_{\mathcal{P}}(1))^{-1}$ -perturbations (Lemma 8.B.5), we have that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \left( \frac{S_k^2}{\sigma_P^2 k} - \log(k/m) + \bar{o}_{\mathcal{P}}^{(k)}(1) \right) \frac{1}{1 + \bar{o}_{\mathcal{P}}^{(k)}(1)} \right\} \leq x \right) - \Psi(x) \right| = 0. \quad (8.126)$$

**Step 2: Establishing validity of  $\bar{p}_k^{(m)}$  and  $\bar{C}_k^{(m)}(\alpha)$ .** Writing out the definition of  $\bar{p}_k^{(m)}$ , we have for any  $P \in \mathcal{P}_0$  and  $\alpha \in (0, 1)$ ,

$$\mathbb{P}_P \left( \exists k \geq m : \bar{p}_k^{(m)} \leq \alpha \right) \quad (8.127)$$

$$= \mathbb{P}_P \left\{ \exists k \geq m : 1 - \Psi \left( k \hat{\mu}_k^2 / \hat{\sigma}_k^2 - \log(k/m) \right) \leq \alpha \right\} \quad (8.128)$$

$$= \mathbb{P}_P \left\{ \exists k \geq m : \Psi \left( k \hat{\mu}_k^2 / \hat{\sigma}_k^2 - \log(k/m) \right) \geq 1 - \alpha \right\} \quad (8.129)$$

$$= \mathbb{P}_P \left( \exists k \geq m : k \hat{\mu}_k^2 / \hat{\sigma}_k^2 - \log(k/m) \geq \Psi^{-1}(1 - \alpha) \right) \quad (8.130)$$

$$= \mathbb{P}_P \left( \sup_{k \geq m} \left\{ k \hat{\mu}_k^2 / \hat{\sigma}_k^2 - \log(k/m) \right\} \geq \Psi^{-1}(1 - \alpha) \right). \quad (8.131)$$

Recalling that  $x \mapsto \Psi(x)$  is a bijection between  $\mathbb{R}^{\geq 0}$  and  $[0, 1]$  and invoking Step 1, we have the desired result:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sup_{\alpha \in (0, 1)} \left| \mathbb{P}_P \left( \exists k \geq m : \bar{p}_k^{(m)} \leq \alpha \right) - \alpha \right| = 0, \quad (8.132)$$

completing the justification for  $\bar{p}_k^{(m)}$ . Moving on to  $\bar{C}_k^{(m)}(\alpha)$ , we have for any  $P \in \mathcal{P}$  and any  $\alpha \in (0, 1)$  that

$$\mathbb{P}_P \left( \exists k \geq m : \mathbb{E}_P(X) \notin \bar{C}_k^{(m)}(\alpha) \right) \quad (8.133)$$

$$= \mathbb{P}_P \left( \exists k \geq m : \mathbb{E}_P(X) \notin \left( \hat{\mu}_k \pm \hat{\sigma}_k \sqrt{[\Psi^{-1}(1 - \alpha) + \log(k/m)]/k} \right) \right) \quad (8.134)$$

$$= \mathbb{P}_P \left( \exists k \geq m : \left| \sum_{i=1}^k (X_i - \mathbb{E}_P(X)) \right| \geq \hat{\sigma}_k \sqrt{k[\Psi^{-1}(1 - \alpha) + \log(k/m)]} \right) \quad (8.135)$$

$$= \mathbb{P}_P \left( \exists k \geq m : S_k^2 / \hat{\sigma}_k^2 k - \log(k/m) \geq \Psi^{-1}(1 - \alpha) \right) \quad (8.136)$$

$$= \mathbb{P}_P \left( \sup_{k \geq m} \left\{ S_k^2 / \hat{\sigma}_k^2 k - \log(k/m) \right\} \geq \Psi^{-1}(1 - \alpha) \right), \quad (8.137)$$

and thus we have that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{\alpha \in (0, 1)} \left| \mathbb{P}_P \left( \exists k \geq m : \mathbb{E}_P(X) \notin \bar{C}_k^{(m)}(\alpha) \right) - \alpha \right| = 0 \quad (8.138)$$

via the same reasoning as was used for the anytime  $p$ -value. This completes the proof.

□

### 8.A.4 Proof of Proposition 8.4.1

**Proposition 8.4.1** (Hardness of anytime-valid conditional independence testing). *Suppose  $(X_n, Y_n, Z_n)_{n=1}^\infty$  are  $[0, 1]^3$ -valued triplets on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P}^*)$  where  $\mathcal{P}^*$  consists of all distributions supported on  $[0, 1]^3$ . Let  $\mathcal{P}_0^* \subseteq \mathcal{P}^*$  be the subset of distributions satisfying the conditional independence null  $H_0$  and denote  $\mathcal{P}_1^* := \mathcal{P}^* \setminus \mathcal{P}_0^*$ . Then for any potentially randomized test  $(\bar{\Gamma}_k^{(m)})_{k=m}^\infty$ ,*

$$\sup_{P \in \mathcal{P}_1^*} \limsup_{m \rightarrow \infty} \mathbb{P}_P \left( \exists k \geq m : \bar{\Gamma}_k^{(m)} = 1 \right) \leq \limsup_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0^*} \mathbb{P}_P \left( \exists k \geq m : \bar{\Gamma}_k^{(m)} = 1 \right). \quad (8.51)$$

In other words, no  $\mathcal{P}_0^*$ -uniform anytime-valid test can have power against any alternative in  $\mathcal{P}_1^*$  at any  $\{m, m+1, \dots\}$ -valued stopping time no matter how large  $m$  is.

*Proof.* Suppose for the sake of contradiction that there exists a potentially randomized test  $(\bar{\Gamma}_k^{(m)})_{k=m}^\infty$  so that for some  $\alpha \in (0, 1)$ , we have both

$$\limsup_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0^*} \mathbb{P}_P \left( \exists k \geq m : \bar{\Gamma}_k^{(m)} = 1 \right) \leq \alpha \quad (8.139)$$

and

$$\sup_{P \in \mathcal{P}_1^*} \limsup_{m \rightarrow \infty} \mathbb{P}_P \left( \exists k \geq m : \bar{\Gamma}_k^{(m)} = 1 \right) > \alpha. \quad (8.140)$$

Then there must exist  $\varepsilon > 0$  so that we can always find  $m_1$  arbitrarily large and nevertheless satisfy

$$\sup_{P \in \mathcal{P}_1^*} \mathbb{P}_P \left( \exists k \geq m_1 : \bar{\Gamma}_k^{(m_1)} = 1 \right) > \alpha + \varepsilon. \quad (8.141)$$

Furthermore, by (8.139), there exists  $m_0 \geq 1$  large enough so that for all  $m \geq m_0$ ,

$$\sup_{P \in \mathcal{P}_0^*} \mathbb{P}_P \left( \exists k \geq m : \bar{\Gamma}_k^{(m)} = 1 \right) < \alpha + \varepsilon. \quad (8.142)$$

In particular, choose some  $m_1 \geq m_0$  so that (8.141) holds. Notice that the events

$$A_M := \{\bar{\Gamma}_k^{(m_1)} = 1 \text{ for some } m_1 \leq k \leq M\} \quad (8.143)$$

are nested for  $M = m_1, m_1 + 1, \dots$  and that  $A_M \rightarrow A := \{\exists k \geq m_1 : \bar{\Gamma}_k^{(m_1)} = 1\}$  as  $M \rightarrow \infty$ . Consequently, there must exist some  $M^*$  such that

$$\sup_{P \in \mathcal{P}_1^*} \mathbb{P}_P \left( \max_{m_1 \leq k \leq M^*} \bar{\Gamma}_k^{(m_1)} = 1 \right) > \alpha + \varepsilon. \quad (8.144)$$

On the other hand, notice that by virtue of being a  $\mathcal{P}_0^*$ -uniform anytime valid test and the fact

that  $m_1 \geq m_0$ , we have that  $\max_{m_1 \leq k \leq M^*} \bar{\Gamma}_k^{(m_1)}$  uniformly controls the type-I error under  $\mathcal{P}_0^*$ , i.e.

$$\sup_{P \in \mathcal{P}_0^*} \mathbb{P}_P \left( \max_{m_1 \leq k \leq M^*} \bar{\Gamma}_k^{(m_1)} = 1 \right) \leq \sup_{P \in \mathcal{P}_0^*} \mathbb{P}_P \left( \exists k \geq m_1 : \bar{\Gamma}_k^{(m_1)} = 1 \right) < \alpha + \varepsilon. \quad (8.145)$$

Combining the above with the hardness result of Shah and Peters [238, Theorem 2] applied to the test  $\max_{m_1 \leq k \leq M^*} \bar{\Gamma}_k^{(m_1)}$ , we have that

$$\sup_{P \in \mathcal{P}_1^*} \mathbb{P}_P \left( \max_{m_1 \leq k \leq M^*} \bar{\Gamma}_k^{(m_1)} = 1 \right) < \alpha + \varepsilon, \quad (8.146)$$

contradicting (8.144), and thus completing the proof of Proposition 8.4.1.  $\square$

### 8.A.5 Proof of Theorem 8.4.1

**Theorem 8.4.1** ( $\mathcal{P}_0$ -uniform type-I error control of the SeqGCM). *Suppose  $(X_i, Y_i, Z_i)_{i=1}^\infty$  are  $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$ -valued triplets defined on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  and let  $\mathcal{P}_0 \subseteq \mathcal{P}$  be a collection of distributions in  $\mathcal{P}$  satisfying the conditional independence null  $H_0$  and Assumption SeqGCM-1, SeqGCM-2, and GCM-3. Define*

$$\bar{p}_{k,m}^{\text{GCM}} := 1 - \Psi(k(\bar{\text{GCM}}_k)^2 - \log(k/m)). \quad (8.58)$$

*Then  $(\bar{p}_{k,m}^{\text{GCM}})_{k=m}^\infty$  forms a  $\mathcal{P}_0$ -uniform anytime p-value for the conditional independence null:*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sup_{\alpha \in (0,1)} |\mathbb{P}_P (\exists k \geq m : \bar{p}_{k,m}^{\text{GCM}} \leq \alpha) - \alpha| = 0. \quad (8.59)$$

Before proceeding with the proof, notice that the estimated residual  $R_i$  can be written as

$$R_i = \xi_i + b_i + \nu_i \quad (8.147)$$

where  $\xi_i := \xi_i^x \cdot \xi_i^y$  is a true product residual with

$$\xi_i^x := \{X_i - \mu^x(Z_i)\} \quad \text{and} \quad \xi_i^y := \{Y_i - \mu^y(Z_i)\}, \quad (8.148)$$

$b_i$  is a product regression error term given by

$$b_i := \{\hat{\mu}_i^x(Z_i) - \mu^x(Z_i)\} \{\hat{\mu}_i^y(Z_i) - \mu^y(Z_i)\}, \quad (8.149)$$

and  $\nu_i := \nu_i^{x,y} + \nu_i^{y,x}$  is a cross-term where

$$\nu_i^{x,y} := \{\hat{\mu}_i^x(Z_i) - \mu^x(Z_i)\} \xi_i^y, \quad \text{and} \quad (8.150)$$

$$\nu_i^{y,x} := \{\hat{\mu}_i^y(Z_i) - \mu^y(Z_i)\} \xi_i^x. \quad (8.151)$$

Furthermore, define their averages as  $\bar{b}_n := \frac{1}{n} \sum_{i=1}^n b_i$  and similarly for  $\bar{\nu}_n^{x,y}$ ,  $\bar{\nu}_n^{y,x}$ , and  $\bar{\xi}_n$ . We

may at times omit the argument  $(Z_i)$  from  $\hat{\mu}_i^x(Z_i) \equiv \hat{\mu}_i^x$  or  $\mu^x(Z_i) \equiv \mu^x$  etc. when it is clear from context. With these shorthands in mind, we are ready to prove Theorem 8.4.1.

*Proof of Theorem 8.4.1.* Note that by some simple algebraic manipulations, it suffices to show that  $\sup_{k \geq m} \left\{ k \overline{\text{GCM}}_k^2 - \log(k/m) \right\}$  converges  $\mathcal{P}_0$ -uniformly to the Robbins-Siegmund distribution as  $m \rightarrow \infty$ . Begin by writing  $\overline{\text{GCM}}_n$  as

$$\overline{\text{GCM}}_n := \frac{1}{n\hat{\sigma}_n^2} \sum_{i=1}^n R_i \quad (8.152)$$

$$\equiv \hat{\sigma}_n^{-1} (\bar{\xi}_n + \bar{\nu}_n + \bar{b}_n) \quad (8.153)$$

and through a direct calculation, notice that our squared GCM statistic can be written as

$$\overline{\text{GCM}}_n^2 = \frac{\bar{\xi}_n^2 + 2\bar{\xi}_n(\bar{\nu}_n + \bar{b}_n) + (\bar{\nu}_n + \bar{b}_n)^2}{\hat{\sigma}_n^2} \quad (8.154)$$

$$= \underbrace{\frac{\bar{\xi}_n^2}{\hat{\sigma}_n^2}}_{(i)} + \underbrace{\frac{2\bar{\xi}_n(\bar{\nu}_n + \bar{b}_n)}{\hat{\sigma}_n^2}}_{(ii)} + \underbrace{\frac{(\bar{\nu}_n + \bar{b}_n)^2}{\hat{\sigma}_n^2}}_{(iii)}. \quad (8.155)$$

In the discussion to follow, we analyze these three terms separately (in Steps 1, 2, and 3, respectively) and combine them to yield the desired result in Step 4.

**Step 1: Analyzing (i).** In Lemma 8.A.1, we show that under the assumptions of Theorem 8.4.1, the estimator  $\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n R_i^2$  is  $\mathcal{P}_0$ -uniformly consistent for  $\text{Var}(\xi) \equiv \mathbb{E}(\xi^2)$  at a rate faster than  $1/\log n$ , meaning

$$\hat{\sigma}_n^2 - \text{Var}(\xi) = \bar{o}_{\mathcal{P}_0}(1/\log n). \quad (8.156)$$

Invoking Assumption **GCM-3**, let  $\underline{\sigma}^2$  be a uniform lower bound on the variance. Then, for any  $P \in \mathcal{P}_0$ ,

$$(i) \equiv \frac{\bar{\xi}_n^2}{\hat{\sigma}_n^2} \quad (8.157)$$

$$= \frac{\bar{\xi}_n^2}{\sigma_P^2 + \bar{o}_{\mathcal{P}_0}(1/\log n)} \quad (8.158)$$

$$= \frac{\bar{\xi}_n^2}{\sigma_P^2 \cdot (1 + \sigma_P^{-2} \cdot \bar{o}_{\mathcal{P}_0}(1/\log n))} \quad (8.159)$$

$$= \frac{\bar{\xi}_n^2}{\sigma_P^2 \cdot (1 + \underline{\sigma}^{-2} \cdot \bar{o}_{\mathcal{P}_0}(1/\log n))} \quad (8.160)$$

$$= \frac{\bar{\xi}_n^2}{\sigma_P^2 \cdot (1 + \bar{o}_{\mathcal{P}_0}(1/\log n))}. \quad (8.161)$$

The final form of (i) above will be used later in Step 4 of the proof.

**Step 2: Analyzing (ii).** In Lemmas 8.A.2 and 8.A.3, we show that under the assumptions of Theorem 8.4.1,  $\bar{b}_n = \bar{o}_{\mathcal{P}_0}(1/\sqrt{n \log \log n})$  and  $\bar{\nu}_n = \bar{o}_{\mathcal{P}_0}(1/\sqrt{n \log \log n})$ , respectively. Recall that  $\hat{\sigma}_n^2 - \mathbb{E}(\xi_n^2) = \bar{o}_{\mathcal{P}_0}(1/\log n)$  by Lemma 8.A.1. Furthermore, we have by the uniform law of the iterated logarithm in Corollary 8.5.1 that  $\xi_n = \bar{O}_{\mathcal{P}_0}(\sqrt{\log \log n/n})$ . Combining these four convergence results together with the calculus outlined in Lemma 8.3.1, we have

$$(ii) = \frac{2\bar{\xi}_n \cdot (\bar{\nu}_n + \bar{b}_n)}{\hat{\sigma}_n^2} \quad (8.162)$$

$$= \frac{\bar{O}_{\mathcal{P}_0}(\sqrt{\log \log n/n}) \cdot \bar{o}_{\mathcal{P}_0}(1/\sqrt{n \log \log n})}{\sigma^2 \cdot (1 + \bar{o}_{\mathcal{P}_0}(1))} \quad (8.163)$$

$$= \bar{o}_{\mathcal{P}_0}(1/n). \quad (8.164)$$

**Step 3: Analyzing (iii).** Again by Lemmas 8.A.2 and 8.A.3 and the calculus of Lemma 8.3.1, we have that

$$(iii) \leq \left| \frac{(\bar{\nu}_n + \bar{b}_n)^2}{\hat{\sigma}_n^2} \right| \quad (8.165)$$

$$= \left| \frac{\bar{o}_{\mathcal{P}_0}(1/n)}{\sigma^2 \cdot (1 + \bar{o}_{\mathcal{P}_0}(1))} \right| \quad (8.166)$$

$$= \bar{o}_{\mathcal{P}_0}(1/n). \quad (8.167)$$

**Step 4: Putting (i)–(iii) together.** Writing out  $\overline{\text{GCM}}_n^2$  and noting the forms of (i), (ii), and (iii) displayed above, we have that for any  $P \in \mathcal{P}_0$ ,

$$\overline{\text{GCM}}_n^2 = \underbrace{\frac{\bar{\xi}_n^2}{\hat{\sigma}_n^2}}_{(i)} + \underbrace{\frac{2\bar{\xi}_n(\bar{\nu}_n + \bar{b}_n)}{\hat{\sigma}_n^2}}_{(ii)} + \underbrace{\frac{(\bar{\nu}_n + \bar{b}_n)^2}{\hat{\sigma}_n^2}}_{(iii)} \quad (8.168)$$

$$= \frac{\bar{\xi}_n^2 / \sigma_P^2}{1 + \bar{o}_{\mathcal{P}_0}(1/\log n)} + \bar{o}_{\mathcal{P}_0}(1/n). \quad (8.169)$$

Similar to the proof of Theorem 8.3.3, let  $\gamma_n$  be the  $1 + \bar{o}_{\mathcal{P}_0}(1/\log n)$  denominator of the first term above. Then for any  $P \in \mathcal{P}_0$  and any  $x \geq 0$ ,

$$\mathbb{P}_P \left( \sup_{k \geq m} \left\{ k \overline{\text{GCM}}_k^2 - \log(k/m) \right\} \leq x \right) \quad (8.170)$$

$$= \mathbb{P}_P \left( \sup_{k \geq m} \left\{ k \left( \frac{\bar{\xi}_k^2 / \sigma_P^2}{\gamma_k} + \bar{o}_{\mathcal{P}_0}^{(k)}(1) \right) - \log(k/m) \right\} \leq x \right) \quad (8.171)$$

$$= \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{k}{\gamma_k} \left( \bar{\xi}_k^2 / \sigma_P^2 + \gamma_k \cdot \bar{o}_{\mathcal{P}_0}^{(k)}(1) - \gamma_k \log(k/m) \right) \right\} \leq x \right) \quad (8.172)$$

$$= \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{k}{1 + \bar{o}_{\mathcal{P}_0}^{(k)}(1)} \left( \bar{\xi}_k^2 / \sigma_P^2 + \bar{o}_{\mathcal{P}_0}^{(k)}(1) - \log(k/m) \right) \right\} \leq x \right), \quad (8.173)$$

and hence similar to the proof of Theorem 8.3.3, we apply Proposition 8.2.1, Lemma 8.B.4, and Lemma 8.B.5 in succession to arrive at the desired result:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sup_{x \geq 0} \left| \mathbb{P}_P \left( \exists k \geq m : k \overline{\text{GCM}}_k^2 - \log(k/m) \geq x \right) - [1 - \Psi(x)] \right| = 0, \quad (8.174)$$

which completes the proof of Theorem 8.4.1.  $\square$

**Lemma 8.A.1** ( $\mathcal{P}_0$ -uniformly strongly consistent variance estimation). *Let  $\hat{\sigma}_n^2$  be the sample variance of  $R_i$ :*

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{i=1}^n R_i^2 - \left( \frac{1}{n} \sum_{i=1}^n R_i \right)^2. \quad (8.175)$$

Then,

$$\hat{\sigma}_n^2 - \mathbb{E}(\xi^2) = \bar{o}_{\mathcal{P}_0} \left( \frac{1}{\log n} \right). \quad (8.176)$$

*Proof of Lemma 8.A.1.* First, consider the following decomposition:

$$R_i^2 = [\xi_i^x \xi_i^y + \xi_i^x \{\mu^y - \hat{\mu}_i^y\} + \xi_i^y \{\mu^x - \hat{\mu}_i^x\} + (\hat{\mu}_i^x - \mu^x)(\hat{\mu}_i^y - \mu^y)]^2 \quad (8.177)$$

$$= \xi_i^2 + \quad (8.178)$$

$$\underbrace{2(\xi_i^x)^2 \xi_i^y \{\mu^y - \hat{\mu}_i^y\} + 2(\xi_i^y)^2 \xi_i^x \{\hat{\mu}_i^x - \mu^x\}}_{I_i} + \quad (8.179)$$

$$\underbrace{4\xi_i \{\mu^x - \hat{\mu}_i^x\} \{\mu^y - \hat{\mu}_i^y\}}_{II_i} + \quad (8.180)$$

$$\underbrace{2\xi_i^x \{\mu^y - \hat{\mu}_i^y\}^2 \{\mu^x - \hat{\mu}_i^x\} + 2\xi_i^y \{\mu^x - \hat{\mu}_i^x\}^2 \{\mu^y - \hat{\mu}_i^y\}}_{III_i} + \quad (8.181)$$

$$\underbrace{\{\mu^x - \hat{\mu}_i^x\}^2 \{\mu^y - \hat{\mu}_i^y\}^2}_{IV_i}. \quad (8.182)$$

Letting  $\bar{I}_n := \frac{1}{n} \sum_{i=1}^n I_i$  and similarly for  $\bar{II}_n$ ,  $\bar{III}_n$ , and  $\bar{IV}_n$ , we have that

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^n \xi_i^2 + \bar{I}_n + \bar{II}_n + \bar{III}_n + \bar{IV}_n - (\bar{R}_n)^2 \quad (8.183)$$

and we will separately show that  $\bar{I}_n$ ,  $\bar{II}_n$ ,  $\bar{III}_n$ ,  $\bar{IV}_n$ , and  $(\bar{R}_n)^2$  are all  $\bar{o}_{\mathcal{P}_0}(1/\log n)$ .

**Step 1: Convergence of  $\bar{I}_n$ .** By the Cauchy-Schwarz inequality, we have that

$$\frac{1}{n} \sum_{i=1}^n (\xi_i^x)^2 \xi_i^y \{\mu^y - \hat{\mu}_i^y\} \leq \underbrace{\sqrt{\frac{1}{n} \sum_{i=1}^n (\xi_i^x \xi_i^y)^2}}_{(\star)} \cdot \underbrace{\sqrt{\frac{1}{n} \sum_{i=1}^n (\xi_i^x)^2 \{\mu^y - \hat{\mu}_i^y\}^2}}_{(\dagger)}. \quad (8.184)$$

Now, writing  $\xi_i := \xi_i^x \xi_i^y$ , notice that

$$(\star) \equiv \frac{1}{n} \sum_{i=1}^n \xi_i^2 \quad (8.185)$$

$$\leq \left( \frac{1}{n} \sum_{i=1}^n \xi_i^2 - \mathbb{E}(\xi_i^2) \right) + \mathbb{E}(\xi_i^2) \quad (8.186)$$

$$= \bar{o}_{\mathcal{P}_0}(1) + \mathbb{E} \left[ \left( |\xi_i|^{2+\delta} \right)^{\frac{2}{2+\delta}} \right] \quad (8.187)$$

$$\leq \bar{o}_{\mathcal{P}_0}(1) + \left( \mathbb{E} |\xi_i|^{2+\delta} \right)^{\frac{2}{2+\delta}} \quad (8.188)$$

$$\leq \bar{O}_{\mathcal{P}_0}(1), \quad (8.189)$$

where the last line follows from Assumption GCM-3. Moreover, by Lemma 8.A.4, we have that  $(\dagger) = \bar{o}_{\mathcal{P}_0}(1/\log n)$ , and hence by Lemma 8.3.1,  $\bar{I}_n \leq (\star) \cdot (\dagger) = \bar{o}_{\mathcal{P}_0}(1/\log n)$ .

**Step 2: Convergence of  $\bar{\Pi}_n$ .** Again by Cauchy-Schwarz, we have

$$\frac{1}{n} \sum_{i=1}^n \xi_i^x \xi_i^y \{\mu^x - \hat{\mu}_i^x\} \{\mu^y - \hat{\mu}_i^y\} \leq \sqrt{\frac{1}{n} \sum_{i=1}^n (\xi_i^x)^2 \{\mu^y - \hat{\mu}_i^y\}^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (\xi_i^y)^2 \{\mu^x - \hat{\mu}_i^x\}^2}, \quad (8.190)$$

and hence again by Lemma 8.A.4, we have  $\bar{\Pi}_n = \bar{o}_{\mathcal{P}_0}(1/\log n)$ .

**Step 3: Convergence of  $\bar{\Pi}\Pi_n$ .** Following Shah and Peters [238, Section D.1] and using the inequality  $2|ab| \leq a^2 + b^2$  for any  $a, b \in \mathbb{R}$ , we have

$$\frac{2}{n} \sum_{i=1}^n \xi_i^x \{\mu^y - \hat{\mu}_i^y\}^2 \{\mu^x - \hat{\mu}_i^x\} \quad (8.191)$$

$$\leq \frac{1}{n} \sum_{i=1}^n (\xi_i^x)^2 \{\mu^y - \hat{\mu}_i^y\}^2 + \frac{1}{n} \sum_{i=1}^n \{\mu^y - \hat{\mu}_i^y\}^2 \{\mu^x - \hat{\mu}_i^x\}^2, \quad (8.192)$$

and hence by Lemmas 8.A.4 and 8.A.2, we have  $\bar{\Pi}\Pi_n = \bar{o}_{\mathcal{P}_0}(1/\log n)$ .

**Step 4: Convergence of  $\bar{IV}_n$ .** First, notice that

$$\bar{IV}_n := \frac{1}{n} \sum_{i=1}^n \{\mu^x - \hat{\mu}_i^x\}^2 \cdot \{\mu^y - \hat{\mu}_i^y\}^2 \quad (8.193)$$

$$\leq n \cdot \frac{1}{n} \sum_{i=1}^n \{\mu^x - \hat{\mu}_i^x\}^2 \cdot \frac{1}{n} \sum_{i=1}^n \{\mu^y - \hat{\mu}_i^y\}^2. \quad (8.194)$$

Applying Lemmas 8.A.2 and 8.3.1, we have that  $\bar{IV}_n = \bar{o}_{\mathcal{P}_0}(1/\log n)$ .

**Step 5: Convergence of  $(\bar{R}_n)^2$  to 0.** We will show that  $(\bar{R}_n)^2 = \bar{o}_{\mathcal{P}_0}(1/\log n)$ . Using the decomposition in (8.147) at the outset of the proof of Theorem 8.4.1, we have that

$$\bar{R}_n := \bar{\xi}_n + \bar{b}_n + \bar{\nu}_n. \quad (8.195)$$

Therefore, we can write its square as

$$(\bar{R}_n)^2 = (\bar{\xi}_n)^2 + 2\bar{\xi}_n \cdot (\bar{b}_n + \bar{\nu}_n) + (\bar{b}_n + \bar{\nu}_n)^2. \quad (8.196)$$

By Assumption GCM-3, we have that there exists a  $\delta > 0$  so that  $\sup_{P \in \mathcal{P}_0} \mathbb{E}_P |\xi|^{2+\delta} < \infty$ . By the de la Vallée-Poussin criterion for uniform integrability [58, 126, 52], we have that the  $(1+\delta)^{\text{th}}$  moment of  $\xi$  is uniformly integrable:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \mathbb{E}_P \left( |\xi|^{1+\delta} \mathbf{1}\{|\xi|^{1+\delta} \geq m\} \right) = 0. \quad (8.197)$$

By Theorem 9.2.1 in Chapter 9, we have that  $\bar{\xi}_n = \bar{o}_{\mathcal{P}_0}(n^{1/(1+\delta)-1})$ , and in particular,

$$\bar{\xi}_n = \bar{o}_{\mathcal{P}_0} \left( 1/\sqrt{\log n} \right). \quad (8.198)$$

Using Lemma 8.3.1, we observe that

$$(\bar{\xi}_n)^2 = \bar{o}_{\mathcal{P}_0}(1/\log n), \quad (8.199)$$

and hence it now suffices to show that  $\bar{b}_n + \bar{\nu}_n = \bar{o}_{\mathcal{P}_0}(1/\log n)$ . Indeed, by Lemmas 8.A.2 and 8.A.3, we have that  $\bar{b}_n = \bar{o}_{\mathcal{P}_0}(1/\sqrt{n \log \log n})$  and  $\bar{\nu}_n = \bar{o}_{\mathcal{P}_0}(1/\sqrt{n \log \log n})$ , respectively. Putting these together, we have

$$(\bar{R}_n)^2 = (\bar{\xi}_n)^2 + 2\bar{\xi}_n \cdot (\bar{b}_n + \bar{\nu}_n) + (\bar{b}_n + \bar{\nu}_n)^2 = \bar{o}_{\mathcal{P}_0}(1/\log n), \quad (8.200)$$

completing the argument for Step 5.

**Step 6: Convergence of  $\hat{\sigma}_n^2$  to  $\mathbb{E}(\xi^2)$ .** Putting Steps 1–5 together, notice that

$$\hat{\sigma}_n^2 - \mathbb{E}(\xi^2) = \frac{1}{n} \sum_{i=1}^n \xi_i^2 - \mathbb{E}(\xi^2) + \bar{I}_n + \bar{II}_n + \bar{III}_n + \bar{IV}_n - (\bar{R}_n)^2 \quad (8.201)$$

$$= \frac{1}{n} \sum_{i=1}^n \xi_i^2 - \mathbb{E}(\xi^2) + \bar{o}_{\mathcal{P}_0}(1/\log n). \quad (8.202)$$

Now, since  $\sup_{P \in \mathcal{P}} \mathbb{E}_P |\xi^2|^{1+\delta/2} < \infty$ ,  $\xi^2$  we have by the de la Vallée criterion for uniform integrability that  $\xi^2$  has a  $\mathcal{P}_0$ -uniformly integrable  $(1 + \delta/4)^{\text{th}}$  moment meaning that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \mathbb{E}_P \left[ (\xi^2)^{1+\delta/4} \mathbf{1}_{\{(\xi^2)^{1+\delta/4} > m\}} \right] = 0, \quad (8.203)$$

and hence by Theorem 9.2.1 in Chapter 9, we have that

$$\frac{1}{n} \sum_{i=1}^n \xi_i^2 - \mathbb{E}(\xi^2) = \bar{o}_{\mathcal{P}} \left( n^{1/(1+\delta/4)-1} \right), \quad (8.204)$$

and in particular,  $\frac{1}{n} \sum_{i=1}^n \xi_i^2 - \mathbb{E}(\xi^2) = \bar{o}_{\mathcal{P}}(1/\log n)$ , so that  $\hat{\sigma}_n^2 - \mathbb{E}(\xi^2) = \bar{o}_{\mathcal{P}}(1/\log n)$ , completing the proof.  $\square$

**Lemma 8.A.2** (Convergence of the average bias term). *Under Assumption SeqGCM-1, we have that*

$$\bar{b}_n \equiv \frac{1}{n} \sum_{i=1}^n b_i = \bar{o}_{\mathcal{P}} \left( 1/\sqrt{n \log \log n} \right). \quad (8.205)$$

*Proof.* Under Assumption SeqGCM-1, we have that

$$\sup_{P \in \mathcal{P}_0} \|\hat{\mu}_n^x - \mu^x\|_{L_2(P)} \cdot \|\hat{\mu}_n^y - \mu^y\|_{L_2(P)} = O \left( \frac{1}{\sqrt{n \log^{2+\delta}(n)}} \right), \quad (8.206)$$

and hence let  $C_{\mathcal{P}_0} > 0$  be a constant depending only on  $\mathcal{P}_0$  so that

$$\sup_{P \in \mathcal{P}_0} \|\hat{\mu}_n^x - \mu^x\|_{L_2(P)} \cdot \|\hat{\mu}_n^y - \mu^y\|_{L_2(P)} \leqslant \frac{C_{\mathcal{P}_0}}{\sqrt{(n+1) \log^{2+\delta/2}(n+1) \log \log(n+1)}} \quad (8.207)$$

for all  $n$  sufficiently large. Consider the following series for all  $k \geq m$  for any  $m \geq 3$

$$\sup_{P \in \mathcal{P}_0} \sum_{k=m}^{\infty} \frac{\mathbb{E}_P |\{\hat{\mu}_{k-1}^x(Z_k) - \mu^x(Z_k)\} \cdot \{\hat{\mu}_{k-1}^y(Z_k) - \mu^y(Z_k)\}|}{\sqrt{k/\log \log k}} \quad (8.208)$$

$$\leqslant \sup_{P \in \mathcal{P}_0} \sum_{k=m}^{\infty} \frac{\|\hat{\mu}_{k-1}^x - \mu^x\|_{L_2(P)} \cdot \|\hat{\mu}_{k-1}^y - \mu^y\|_{L_2(P)}}{\sqrt{k/\log \log k}} \quad (8.209)$$

$$= \sum_{k=m}^{\infty} \frac{C_{\mathcal{P}_0}}{\sqrt{k \log^{2+\delta/2}(k) \log \log k} \cdot \sqrt{k \log \log k}} \quad (8.210)$$

$$= \sum_{k=m}^{\infty} \frac{C_{\mathcal{P}_0}}{k \log^{1+\delta/4}(k)}, \quad (8.211)$$

and since  $(k \log^{1+\delta/4}(k))^{-1}$  is summable for any  $\delta > 0$ , we have that the above vanishes as  $m \rightarrow \infty$ , hence

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sum_{k=m}^{\infty} \frac{\mathbb{E}_P |\{\hat{\mu}_{k-1}^x(Z_k) - \mu^x(Z_k)\} \cdot \{\hat{\mu}_{k-1}^y(Z_k) - \mu^y(Z_k)\}|}{\sqrt{k/\log \log k}} = 0. \quad (8.212)$$

Applying Theorem 9.2.2 in Chapter 9, we have that

$$\bar{b}_n \equiv \frac{1}{n} \sum_{k=1}^n \{\hat{\mu}_{k-1}^x(Z_k) - \mu^x(Z_k)\} \cdot \{\hat{\mu}_{k-1}^y(Z_k) - \mu^y(Z_k)\} = \bar{o}_{\mathcal{P}_0} \left( \frac{1}{\sqrt{n \log \log n}} \right), \quad (8.213)$$

which completes the proof.  $\square$

**Lemma 8.A.3** (Convergence of average cross-terms). *Suppose that for some  $\delta > 0$ , and some independent  $Z$  with the same distribution as  $Z_n$ ,*

$$\sup_{P \in \mathcal{P}_0} \mathbb{E}_P \left[ (\{\hat{\mu}_n^x(Z_n) - \mu^x(Z_n)\} \xi_n^y)^2 \right] = O \left( \frac{1}{(\log n)^{2+\delta}} \right). \quad (8.214)$$

Then,

$$\frac{1}{n} \sum_{i=1}^n \nu_i^{x,y} = \bar{o}_{\mathcal{P}_0} (1/\sqrt{n \log \log n}), \quad (8.215)$$

with an analogous statement holding when  $x$  and  $y$  are swapped in the above condition and conclusion.

*Proof.* We will only prove the result for  $\nu_i^{x,y}$  but the same argument goes through for  $\nu_i^{y,x}$ . Appealing to (8.214), let  $C_{\mathcal{P}_0}$  be a constant so that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ (\{\hat{\mu}_n^x(Z_n) - \mu^x(Z_n)\} \xi_n^y)^2 \right] \leq \frac{C_{\mathcal{P}_0}}{(\log n)^{2+\delta}}. \quad (8.216)$$

Then notice that for all  $m$  sufficiently large

$$\sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{\mathbb{E}_P \left[ (\{\hat{\mu}_k^x(Z_k) - \mu^x(Z_k)\} \xi_k^y)^2 \right]}{k/\log \log k} \quad (8.217)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{C_{\mathcal{P}_0}}{k(\log k)^{2+\delta}/\log \log k} \quad (8.218)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{C_{\mathcal{P}_0}}{k(\log k)^{1+\delta}} \quad (8.219)$$

$$= 0, \quad (8.220)$$

and hence

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{\mathbb{E}_P \left[ (\{\hat{\mu}_{k-1}^x(Z_k) - \mu^x(Z_k)\} \xi_k^y)^2 \right]}{k / \log \log k} = 0. \quad (8.221)$$

By Theorem 9.2.2 in Chapter 9, we have that

$$\frac{1}{n} \sum_{i=1}^n \nu_i^{x,y} = \bar{o}_{\mathcal{P}}(1/\sqrt{n \log \log n}), \quad (8.222)$$

completing the proof.  $\square$

**Lemma 8.A.4** (Convergence of average squared cross-terms). *Under Assumption SeqGCM-2, we have that*

$$\frac{1}{n} \sum_{i=1}^n (\nu_i^{x,y})^2 \equiv \frac{1}{n} \sum_{i=1}^n (\xi_i^x)^2 \{\mu^y - \hat{\mu}_i^y\}^2 = \bar{o}_{\mathcal{P}_0}(1/\log n). \quad (8.223)$$

An analogous statement holds with  $\xi_n^x$  replaced by  $\xi_n^y$  and  $\{\mu^y(Z_n) - \hat{\mu}_n^y(Z_n)\}$  replaced by  $\{\mu^x(Z_n) - \hat{\mu}_n^x(Z_n)\}$ .

*Proof.* Using Assumption SeqGCM-2, let  $C_{\mathcal{P}_0} > 0$  be a constant so that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P [(\xi_n^x)^2 \{\mu^y(Z_n) - \hat{\mu}_{n-1}^y(Z_n)\}^2] \leq \frac{C_{\mathcal{P}_0}}{(\log n)^{2+\delta}}. \quad (8.224)$$

Therefore, we have that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}_0} \sum_{k=m}^{\infty} \frac{\mathbb{E}_P [(\xi_k^x)^2 \{\mu^y(Z_k) - \hat{\mu}_{k-1}^y(Z_k)\}^2]}{k(\log k)^{-1}} \quad (8.225)$$

$$\leq \lim_{m \rightarrow \infty} \sum_{k=m}^{\infty} \frac{C_{\mathcal{P}_0}}{k(\log k)^{2+\delta-1}} \quad (8.226)$$

$$= \lim_{m \rightarrow \infty} \sum_{k=m}^{\infty} \frac{C_{\mathcal{P}_0}}{k(\log k)^{1+\delta}} \quad (8.227)$$

$$= 0. \quad (8.228)$$

Combining the above with Theorem 9.2.2 in Chapter 9, we have that

$$\frac{1}{n} \sum_{i=1}^n (\xi_i^x)^2 \{\mu^y - \hat{\mu}_i^y(Z_i)\}^2 = \bar{o}_{\mathcal{P}_0}(1/\log n), \quad (8.229)$$

completing the proof.  $\square$

### 8.A.6 Proof of Corollary 8.5.1

**Corollary 8.5.1** (A  $\mathcal{P}$ -uniform law of the iterated logarithm). *Suppose  $(X_n)_{n=1}^\infty$  are defined on probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  where  $\mathcal{P}$  is a collection of distributions such that*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P |X - \mathbb{E}_P X|^{2+\delta} < \infty \text{ and } \inf_{P \in \mathcal{P}} \text{Var}_P(X) > 0 \quad (8.70)$$

for some  $\delta > 0$ . Then,

$$\sup_{k \geq n} \frac{|\sum_{i=1}^k (X_i - \mathbb{E}_P(X))|}{\sqrt{2\text{Var}_P(X)k \log \log k}} = 1 + \bar{o}_P(1). \quad (8.71)$$

*Proof.* This is a consequence of our distribution-uniform strong Gaussian coupling given in Theorem 8.5.2. Letting  $\mu_P := \mathbb{E}_P(X)$  and  $\sigma_P := \sqrt{\text{Var}_P(X)}$  to reduce notational clutter, note that by Theorem 8.5.2, we have that there exists a construction with a sequence of standard Gaussians  $Y_1, Y_2, \dots$  such that

$$\sum_{i=1}^n (X_i - \mu_P)/\sigma_P = \sum_{i=1}^n Y_i + \bar{o}_P(n^{1/q}(\log n)^{2/q}), \quad (8.230)$$

where  $q := 2 + \delta$ , or more formally that for any  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq n : \frac{|\sum_{i=1}^k (X_i - \mu_P)/\sigma_P - \sum_{i=1}^k Y_i|}{k^{1/q}(\log k)^{2/q}} > \varepsilon \right) = 0. \quad (8.231)$$

Now, by the law of the iterated logarithm, we also have that

$$\sup_{\ell \geq n} \frac{|\sum_{i=1}^\ell Y_i|}{\sqrt{2\ell \log \log \ell}} = 1 + \bar{o}_P(1), \quad (8.232)$$

for each  $P \in \mathcal{P}$ , and since  $Y$  has the same distribution on every element of  $P \in \mathcal{P}$ , the above also holds with  $\bar{o}_P(1)$  replaced by  $\bar{o}_{\mathcal{P}}(1)$ . Now, to prove the final result, we have that

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq n : \left| \sup_{\ell \geq k} \frac{|\sum_{i=1}^\ell (X_i - \mu_P)|}{\sqrt{2\sigma_P^2 \ell \log \log \ell}} - 1 \right| > \varepsilon \right) \quad (8.233)$$

$$\leq \lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq n : \left| \sup_{\ell \geq k} \left\{ \frac{|\sum_{i=1}^\ell (X_i - \mu_P)/\sigma_P - \sum_{i=1}^\ell Y_i|}{\sqrt{2\ell \log \log \ell}} + \frac{|\sum_{i=1}^\ell Y_i|}{\sqrt{2\ell \log \log \ell}} \right\} - 1 \right| > \varepsilon \right) \quad (8.234)$$

$$\leq \lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq n : \left| \sup_{\ell \geq k} \left\{ \varepsilon/2 + \frac{|\sum_{i=1}^\ell Y_i|}{\sqrt{2\ell \log \log \ell}} \right\} - 1 \right| > \varepsilon/2 \right) + \quad (8.235)$$

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq n : \frac{|\sum_{i=1}^k (X_i - \mu_P)/\sigma_P - \sum_{i=1}^k Y_i|}{\sqrt{2k \log \log k}} > \varepsilon/2 \right) \quad (8.236)$$

$$\leq \lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq n : \left| \sup_{\ell \geq k} \left\{ \frac{|\sum_{i=1}^{\ell} Y_i|}{\sqrt{2\ell \log \log \ell}} \right\} - 1 \right| > \varepsilon/2 \right) + \quad (8.237)$$

$$\underbrace{\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq n : \frac{|\sum_{i=1}^k (X_i - \mu_P)/\sigma_P - \sum_{i=1}^k Y_i|}{k^{1/q}(\log k)^{2/q}} > \varepsilon/2 \right)}_{=0} \quad (8.238)$$

$$= \lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \exists k \geq n : \left| \sup_{\ell \geq k} \left\{ \frac{|\sum_{i=1}^{\ell} Y_i|}{\sqrt{2\ell \log \log \ell}} \right\} - 1 \right| > \varepsilon/2 \right) \quad (8.239)$$

$$= \sup_{P \in \mathcal{P}} \lim_{n \rightarrow \infty} \mathbb{P}_P \left( \exists k \geq n : \left| \sup_{\ell \geq k} \left\{ \frac{|\sum_{i=1}^{\ell} Y_i|}{\sqrt{2\ell \log \log \ell}} \right\} - 1 \right| > \varepsilon/2 \right) \quad (8.240)$$

$$= 0, \quad (8.241)$$

where the second inequality follows from Theorem 8.5.2 and the third follows from the triangle inequality and the fact that  $k^{1/q}(\log k)^{2/q} \leq 2k \log \log k$  for all  $k$  sufficiently large. The second-last equality follows from the fact that the probability does not depend on features of the distribution  $P$  and the last equality follows from the  $P$ -pointwise law of the iterated logarithm.  $\square$

### 8.A.7 Proof of Lemma 8.5.1 and Theorem 8.5.2

**Lemma 8.5.1** (Strong Gaussian coupling inequality). *Let  $(X_n)_{n=1}^{\infty}$  be independent random variables on the probability space  $(\Omega, \mathcal{F}, P)$ . Suppose that for some  $q \geq 2$ , we have  $\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q < \infty$  for each  $k \in \mathbb{N}$ . Let  $f(\cdot)$  be a positive and increasing function so that  $\sum_{n=1}^{\infty} (nf(n))^{-1} < \infty$  and*

$$\sum_{k=1}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q / \sigma_k^q}{kf(k)} < \infty, \quad (8.68)$$

where  $\sigma_k^2 := \text{Var}_P(X_k)$ . Then one can construct a probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}(P))$  rich enough to define  $(\tilde{X}_n, Y_n)_{n=1}^{\infty}$  where  $\tilde{X}_n$  and  $X_n$  are equidistributed for each  $n$  and  $(Y_n)_{n=1}^{\infty}$  are marginally independent standard Gaussians so that for any  $\varepsilon > 0$ ,

$$\begin{aligned} \mathbb{P}_{\tilde{P}(P)} \left( \exists k \geq m : \left| \frac{\sum_{i=1}^k (\tilde{X}_i - Y_i)}{k^{1/q} f(k)^{1/q}} \right| > \varepsilon \right) &\leq \frac{C_{q,f}}{\varepsilon^q} \left\{ \sum_{k=2^{m-1}}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q / \sigma_k^q}{kf(k)} + \right. \\ &\quad \left. \frac{1}{2^m} \sum_{k=1}^{2^{m-1}-1} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q / \sigma_k^q}{kf(k)} \right\}, \end{aligned} \quad (8.69)$$

where  $C_{q,f}$  is a constant that depends only on  $q$  and  $f$ .

First, we need the following result due to Lifshits [175, Theorem 3.3] which is itself a refinement of an inequality due to Sakhnenko [230].

**Lemma 8.A.5** (Sakhanenko-Lifshits inequality). *Let  $X_1, X_2, \dots : \Omega \rightarrow \mathbb{R}$  be independent mean-zero random variables on a probability space  $(\Omega, \mathcal{F}, P)$  and let  $q \geq 2$ . Then one can construct a new probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$  rich enough to contain  $(\tilde{X}_n, Y_n)_{n=1}^\infty$  so that  $(X_1, X_2, \dots)$  and  $(\tilde{X}_1, \tilde{X}_2, \dots)$  are equidistributed and  $(Y_1, Y_2, \dots)$  are standard Gaussian random variables so that*

$$\mathbb{E}_P \left( \max_{1 \leq k \leq n} \left| \sum_{i=1}^k X_i / \sigma_P(X_i) - \sum_{i=1}^k Y_i \right| \right)^q \leq C_q \sum_{i=1}^n \frac{\mathbb{E}_P |X_i|^q}{\sigma_P(X_i)^q} \quad (8.242)$$

where  $C_q$  is a constant depending only on  $q$ .

Notice that  $\sigma_P(X_i) = \sigma_{\tilde{P}}(\tilde{X}_i)$  and  $\mathbb{E}_P |X_i|^q = \mathbb{E}_{\tilde{P}} |\tilde{X}_i|^q$  in the above lemma so we may use them interchangeably.

### Proof of the main result

*Proof of Lemma 8.5.1.* Throughout the proof, we will use  $\sigma_i$  in place of  $\sigma_P(X_i)$  whenever the distribution  $P$  is clear from context. We will also let  $S_k(P)$  and  $G_k$  be the partial sums given by

$$S_k(P) := \sum_{i=1}^k (X_i - \mathbb{E}_P(X_i)) / \sigma_P(X_i) \text{ and } G_k := \sum_{i=1}^k Y_i. \quad (8.243)$$

For any  $P \in \mathcal{P}$ , we appeal to Lemma 8.A.5 and let  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$  be a construction so that for any  $n$ ,

$$\mathbb{E}_{\tilde{P}} \left( \max_{1 \leq k \leq n} |S_k - G_k| \right)^q \leq C_q \sum_{k=1}^n \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q}{\sigma_P(X_k)^q}, \quad (8.244)$$

By Markov's inequality, we have that for any  $z > 0$ ,

$$\mathbb{P}_{\tilde{P}} \left( \max_{1 \leq k \leq n} |S_k(P) - G_k| > z \right) \leq C_q \frac{\sum_{k=1}^n \mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q / \sigma_k^q}{z^q}, \quad (8.245)$$

noting that the right-hand side does not depend on the new probability space, but only on the original  $P$ . Defining  $\Delta_k \equiv \Delta_k(P) := S_k(P) - G_k$ , we have that for any  $k$  and  $n$ ,

$$\max_{\mathcal{D}(n-1) \leq k < \mathcal{D}(n)} \{\Delta_k\} = \underbrace{\max_{\mathcal{D}(n-1) \leq k < \mathcal{D}(n)} \{\Delta_k - \Delta_{\mathcal{D}(n-1)-1}\}}_{(a)} + \underbrace{\Delta_{\mathcal{D}(n-1)-1}}_{(b)}, \quad (8.246)$$

where  $\mathcal{D}(n) := 2^n$  are exponentially spaced demarcation points that will become important in the arguments to follow. We will proceed by separately bounding (a) and (b) time-uniformly with high-probability.

**Step 1: Bounding (a) time-uniformly with high probability.** Let  $a_k := k^{1/q} f(k)^{1/q}$ . By (8.245) applied to  $\Delta_k - \Delta_{\mathcal{D}(n-1)-1} \equiv \sum_{i=\mathcal{D}(n-1)}^k (X_i - \mathbb{E}_P(X_i)) / \sigma_i$  with  $z := \varepsilon a_{\mathcal{D}(n-1)}$ ,

we have that

$$\mathbb{P}_{\tilde{P}} \left( \max_{\mathcal{D}(n-1) \leq k < \mathcal{D}(n)} \{\Delta_k - \Delta_{\mathcal{D}(n-1)-1}\} > \varepsilon a_{\mathcal{D}(n-1)} \right) \quad (8.247)$$

$$\leq C_q \sum_{k=\mathcal{D}(n-1)}^{\mathcal{D}(n)-1} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q / \sigma_k^q}{a_{\mathcal{D}(n-1)}^q \varepsilon^q} \quad (8.248)$$

$$\leq \varepsilon^{-q} C_q \sum_{k=\mathcal{D}(n-1)}^{\mathcal{D}(n)-1} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q / \sigma_k^q}{a_k^q}. \quad (8.249)$$

Union bounding over  $n = m, m+1, \dots$  we have that

$$\mathbb{P}_{\tilde{P}} (\exists n \geq m \text{ and } k \in \{\mathcal{D}(n-1), \dots, \mathcal{D}(n)\} : \{\Delta_k - \Delta_{\mathcal{D}(n-1)-1}\} > \varepsilon a_k) \quad (8.250)$$

$$\leq \mathbb{P}_{\tilde{P}} \left( \exists n \geq m : \max_{\mathcal{D}(n-1) \leq k < \mathcal{D}(n)} \{\Delta_k - \Delta_{\mathcal{D}(n-1)-1}\} > \varepsilon a_{\mathcal{D}(n-1)} \right) \quad (8.251)$$

$$\leq \sum_{n=m}^{\infty} \varepsilon^{-q} C_q \sum_{k=\mathcal{D}(n-1)}^{\mathcal{D}(n)-1} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q / \sigma_k^q}{a_k^q}. \quad (8.252)$$

$$\leq \varepsilon^{-q} C_q \sum_{n=\mathcal{D}(m-1)}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q}{a_k^q}. \quad (8.253)$$

$$(8.254)$$

**Step 2: Bounding (b) time-uniformly with high probability.** Applying (8.245) to  $\Delta_{\mathcal{D}(n-1)-1}$  with  $z = \varepsilon a_{\mathcal{D}(n-1)}$ , we have that

$$\mathbb{P}_{\tilde{P}} (|\Delta_{\mathcal{D}(n-1)-1}| > \varepsilon a_{\mathcal{D}(n-1)}) \quad (8.255)$$

$$\leq C_q \sum_{k=1}^{\mathcal{D}(n-1)-1} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q / \sigma_k^q}{a_{\mathcal{D}(n-1)}^q \varepsilon^q} \quad (8.256)$$

$$\leq \frac{C_q}{\varepsilon^q a_{\mathcal{D}(n-1)}^q} \sum_{k=1}^{\mathcal{D}(n-1)-1} \mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q / \sigma_k^q. \quad (8.257)$$

Union bounding again over  $n = m, m+1, \dots$ , we have

$$\mathbb{P}_{\tilde{P}} (\exists n \geq m \text{ and } k \in \{\mathcal{D}(n-1), \dots, \mathcal{D}(n)\} : |\Delta_{\mathcal{D}(n-1)-1}| > \varepsilon a_k) \quad (8.258)$$

$$\leq \mathbb{P}_{\tilde{P}} (\exists n \geq m \text{ and } k \in \{\mathcal{D}(n-1), \dots, \mathcal{D}(n)\} : |\Delta_{\mathcal{D}(n-1)-1}| > \varepsilon a_{\mathcal{D}(n-1)}) \quad (8.259)$$

$$= \mathbb{P}_{\tilde{P}} (\exists n \geq m : |\Delta_{\mathcal{D}(n-1)-1}| > \varepsilon a_{\mathcal{D}(n-1)}) \quad (8.260)$$

$$\leq \sum_{n=m}^{\infty} \frac{C_q}{\varepsilon^q a_{\mathcal{D}(n-1)}^q} \sum_{k=1}^{\mathcal{D}(n-1)-1} \mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q / \sigma_k^q. \quad (8.261)$$

**Step 3: Union bounding over the results from Steps 1 and 2.** Putting Steps 1 and 2 together, we have the following time-uniform crossing inequality for  $|\Delta_k|$ :

$$\mathbb{P}_{\tilde{P}}(\exists k \geq m : |\Delta_k| > 2\varepsilon a_k) \quad (8.262)$$

$$\leq \mathbb{P}_{\tilde{P}}(\exists k \geq m : |\Delta_k - \Delta_{\mathcal{D}(n-1)-1}| + |\Delta_{\mathcal{D}(n-1)-1}| > a_k + a_k) \quad (8.263)$$

$$\leq \varepsilon^{-q} C_q \left[ \sum_{n=\mathcal{D}(m-1)}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q}{a_k^q} + \sum_{n=m}^{\infty} \frac{1}{a_{\mathcal{D}(n-1)}^q} \sum_{k=1}^{\mathcal{D}(n-1)-1} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q}{\sigma_k^q} \right] \quad (8.264)$$

$$\leq \varepsilon^{-q} C_q \left[ \sum_{n=\mathcal{D}(m-1)}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q}{a_k^q} + \sum_{n=m}^{\infty} \frac{1}{a_{\mathcal{D}(n-1)}^q} \sum_{k=1}^{\mathcal{D}(n-1)-1} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q}{\sigma_k^q} \right]. \quad (8.265)$$

Letting  $\rho_k^q := \mathbb{E}_P |X_k - \mathbb{E}_P(X_k)|^q$  and further simplifying the above expression so that it does not depend on the demarcation points  $\mathcal{D}(n)$ , we have

$$\mathbb{P}_{\tilde{P}}(\exists k \geq m : |\Delta_k| > 2\varepsilon a_k) \quad (8.266)$$

$$\leq \varepsilon^{-q} C_q \left[ \sum_{k=\mathcal{D}(m-1)}^{\infty} \frac{\rho_k^q / \sigma_k^q}{a_k^q} + \sum_{n=m}^{\infty} \frac{1}{a_{\mathcal{D}(n-1)}^q} \sum_{k=1}^{\mathcal{D}(n-1)-1} \rho_k^q / \sigma_k^q \right] \quad (8.267)$$

$$\leq \varepsilon^{-q} C_q \left[ \sum_{k=\mathcal{D}(m-1)}^{\infty} \frac{\rho_k^q / \sigma_k^q}{k f(k)} + \sum_{n=m}^{\infty} \frac{1}{\mathcal{D}(n-1) f(\mathcal{D}(n-1))} \sum_{k=1}^{\mathcal{D}(n-1)-1} \rho_k^q / \sigma_k^q \right] \quad (8.268)$$

$$\leq \varepsilon^{-q} C_q \left[ \sum_{k=\mathcal{D}(m-1)}^{\infty} \frac{\rho_k^q / \sigma_k^q}{k f(k)} + \sum_{n=m}^{\infty} \frac{1}{\mathcal{D}(n-1)} \sum_{k=1}^{\mathcal{D}(n-1)-1} \frac{\rho_k^q / \sigma_k^q}{f(\mathcal{D}(k))} \right] \quad (8.269)$$

$$= \varepsilon^{-q} C_q \left[ \sum_{k=2^{m-1}}^{\infty} \frac{\rho_k^q / \sigma_k^q}{k f(k)} + \sum_{n=m}^{\infty} \frac{1}{2^{n-1}} \sum_{k=1}^{\mathcal{D}(n-1)-1} \frac{\rho_k^q / \sigma_k^q}{f(2^k)} \right] \quad (8.270)$$

$$\leq \varepsilon^{-q} C_q \left[ \sum_{k=2^{m-1}}^{\infty} \frac{\rho_k^q / \sigma_k^q}{k f(k)} + \sum_{n=m}^{\infty} \frac{C_f^{-1}}{2^{n-1}} \sum_{k=1}^{2^n} \frac{\rho_k^q / \sigma_k^q}{k f(k)} \right], \quad (8.271)$$

where (8.268) follows from the definition of  $a_k := k^{1/q} f(k)^{1/q}$ , (8.269) follows from the fact that  $f$  is increasing and that  $\mathcal{D}(k) := 2^k$ , (8.270) follows from the definition of  $\mathcal{D}(\cdot)$ , and (8.271) from the fact that  $f(k) \geq C_f \log(k)$  for all  $k \geq 1$  and some constant  $C_f$  depending only on  $f$  (if this were not true, then  $\sum_{k=1}^{\infty} [k f(k)]^{-1}$  would not be summable).

The final result follows from observing that  $\sum_{k=1}^{2^n} \rho_k^q / (\sigma_k^q k f(k)) \leq \sum_{k=1}^{\infty} \rho_k^q / (\sigma_k^q k f(k))$  and absorbing constants only depending on  $q$  and  $f$  into  $C_{q,f}$ :

$$\mathbb{P}_{\tilde{P}}(\exists k \geq m : |\Delta_k| > \varepsilon a_k) \quad (8.272)$$

$$\leq 2^q \varepsilon^{-q} C_q \left[ \sum_{k=2^{m-1}}^{\infty} \frac{\rho_k^q / \sigma_k^q}{kf(k)} + \sum_{n=m}^{\infty} \frac{C_f^{-1}}{2^{n-1}} \sum_{k=1}^{2^n} \frac{\rho_k^q / \sigma_k^q}{kf(k)} \right] \quad (8.273)$$

$$\leq 2^q \varepsilon^{-q} C_q C_f^{-1} \left[ C_f \sum_{k=2^{m-1}}^{\infty} \frac{\rho_k^q / \sigma_k^q}{kf(k)} + 2^{-m} \left( \sum_{k=1}^{2^{m-1}-1} \frac{\rho_k^q / \sigma_k^q}{kf(k)} + \sum_{k=2^{m-1}}^{\infty} \frac{\rho_k^q / \sigma_k^q}{kf(k)} \right) \right] \quad (8.274)$$

$$= 2^q \varepsilon^{-q} C_q C_f^{-1} \left[ (C_f + 2^{-m}) \sum_{k=2^{m-1}}^{\infty} \frac{\rho_k^q / \sigma_k^q}{kf(k)} + \frac{1}{2^m} \sum_{k=1}^{2^{m-1}-1} \frac{\rho_k^q / \sigma_k^q}{kf(k)} \right] \quad (8.275)$$

$$\leq \varepsilon^{-q} C_{q,f} \left[ \sum_{k=2^{m-1}}^{\infty} \frac{\rho_k^q / \sigma_k^q}{kf(k)} + \frac{1}{2^m} \sum_{k=1}^{2^{m-1}-1} \frac{\rho_k^q / \sigma_k^q}{kf(k)} \right], \quad (8.276)$$

which completes the proof  $\square$

Let us now show how Theorem 8.5.2 is a consequence of the above.

**Theorem 8.5.2** (Distribution-uniform strong Gaussian approximation). *Let  $(X_n)_{n=1}^{\infty}$  be independent and identically distributed random variables defined on the collection of probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  with means  $\mu_P := \mathbb{E}_P(X)$  and variances  $\sigma_P^2 := \mathbb{E}_P(X - \mu_P)^2$ . If  $X$  has  $q > 2$  uniformly upper-bounded moments, and a uniformly positive variance, i.e.*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P |X - \mu_P|^q < \infty \quad \text{and} \quad \inf_{P \in \mathcal{P}} \sigma_P^2 > 0, \quad (8.66)$$

*then there exists a construction with independent standard Gaussians  $(Y_n)_{n=1}^{\infty} \sim N(0, 1)$  so that*

$$\left| \sum_{i=1}^n \frac{X_i - \mu_P}{\sigma_P} - \sum_{i=1}^n Y_i \right| = \bar{o}_{\mathcal{P}}(n^{1/q} \log^{2/q}(n)). \quad (8.67)$$

*Proof.* The proof of Theorem 8.5.2 amounts to analyzing the  $\mathcal{P}$ -uniform tail behavior of the probability bound in Lemma 8.5.1. Indeed, for each  $P \in \mathcal{P}$ , let  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P}(P))$  be the construction that yields

$$\mathbb{P}_{\tilde{P}} \left( \exists k \geq m : \left| \frac{\sum_{i=1}^k (X_i - \mu_P) / \sigma_P - \sum_{i=1}^k Y_i}{kf(k)} \right| > \varepsilon \right) \quad (8.277)$$

$$\leq \varepsilon^{-q} C_{q,f} \left[ \sum_{k=2^{m-1}}^{\infty} \frac{\rho_P^q / \sigma_P^q}{kf(k)} + \frac{1}{2^m} \sum_{k=1}^{2^{m-1}-1} \frac{\rho_P^q / \sigma_P^q}{kf(k)} \right], \quad (8.278)$$

where  $\rho_P^q := \mathbb{E}_P |X - \mathbb{E}_P X|^q$  and  $\sigma_P^2 := \mathbb{E}_P(X - \mathbb{E}_P)^2$ . Let  $\bar{\rho} < \infty$  be the uniform upper bound so that  $\sup_{P \in \mathcal{P}} \rho_P^q \leq \bar{\rho}^q$  and  $\underline{\sigma}^2 > 0$  be the uniform lower bound on the variance so that  $\inf_{P \in \mathcal{P}} \sigma_P^2 \geq \underline{\sigma}^2$ . Replacing the above finite sum by its infinite extension and taking suprema

over  $P$  on both sides, we have that

$$\sup_{P \in \mathcal{P}} \mathbb{P}_{\tilde{P}(P)} \left( \exists k \geq m : \left| \frac{\sum_{i=1}^k (X_i - \mu_P)/\sigma_P - \sum_{i=1}^k Y_i}{kf(k)} \right| > \varepsilon \right) \quad (8.279)$$

$$\leq \frac{\bar{\rho}^q C_{q,f}}{\sigma^q \varepsilon^q} \left[ \underbrace{\sum_{k=2^{m-1}}^{\infty} \frac{1}{kf(k)}}_{(*)} + \underbrace{\frac{1}{2^m} \sum_{k=1}^{\infty} \frac{1}{kf(k)}}_{(\dagger)} \right]. \quad (8.280)$$

Now,  $(*) \rightarrow 0$  and  $(\dagger) \rightarrow 0$  as  $m \rightarrow \infty$ , both of which follow from the fact that  $1/[kf(k)]$  is summable. Instantiating the above for  $f(k) := \log^2(k)$  completes the proof.  $\square$

## 8.B Additional theoretical discussions and results

### 8.B.1 The Robbins-Siegmund distribution

Fundamental to this chapter is a probability distribution that describes the supremum of a transformed Wiener process with a delayed start (see Lemma 8.B.1). As far as we can tell, the distribution was first (implicitly) discovered by Robbins and Siegmund [220] and as such we refer to it as the *Robbins-Siegmund distribution*. In this section, we provide its cumulative distribution function (CDF)  $\Psi$  and show how the suprema of scaled Wiener processes have this distribution  $\Psi$ . The Robbins-Siegmund distribution has also been implicitly used in Chapter 7 and Bibaut et al. [34] for the sake of  $P$ -pointwise anytime-valid inference.

*Definition 8.B.1* (The Robbins-Siegmund distribution). We say that a nonnegative random variable  $\mathfrak{R}$  follows the Robbins-Siegmund (R-S) distribution if its CDF is given by

$$\Psi(r) := 1 - 2 \left[ 1 - \Phi(\sqrt{r}) + \sqrt{r}\phi(\sqrt{r}) \right]; \quad r \geq 0, \quad (8.281)$$

where  $\Phi$  and  $\phi$  are the CDF and density of a standard Gaussian, respectively.

The following lemma demonstrates how the supremum of a transformed Wiener process follows the Robbins-Siegmund distribution.

**Lemma 8.B.1.** *Let  $(W(t))_{t \geq 0}$  be a standard Wiener process and define*

$$\mathfrak{R} := \sup_{t \geq 1} \left\{ \frac{W(t)^2}{t} - \log t \right\}. \quad (8.282)$$

*Then,  $\mathfrak{R}$  follows the Robbins-Siegmund distribution given in Definition 8.B.1.*

*Proof.* Rather than derive its CDF for a given  $r$ , we will derive the survival function  $\mathbb{P}(\mathfrak{R} \geq a^2)$  for any  $a \geq 0$ , showing that  $\mathbb{P}(\mathfrak{R} \geq a^2) = 1 - \Psi(a^2)$  as given in Definition 8.B.1, which will yield the desired result.

$$\mathbb{P}(\mathfrak{R} \geq a^2) = \mathbb{P}(\exists t \geq 1 : W(t)^2/t - \log t \geq a^2) \quad (8.283)$$

$$= \mathbb{P}(\exists t \geq 1 : |W(t)| \geq \sqrt{t[a^2 + \log t]}) \quad (8.284)$$

$$= 2[1 - \Phi(a) + a\phi(a)] = 1 - \Psi(a^2). \quad (8.285)$$

where the last line follows from Robbins and Siegmund [220] but with their value of  $\tau$  set to 1. Alternatively, a different proof found in Lemma 7.A.12 in Chapter 7 yields the desired result.  $\square$

The following lemma demonstrates that appropriately scaled discrete Gaussian partial sums converge to the Robbins-Siegmund distribution.

**Lemma 8.B.2** (Transformed Gaussian partial sums converge to the Robbins-Siegmund distribution). *Let  $G_k$  be a sum of iid Gaussian random variables with mean zero and variance  $\sigma^2$ . Then,*

$$\sup_{k \geq m} \left\{ \frac{G_k^2}{k\sigma^2} - \log(k/m) \right\} \xrightarrow{d} \Psi \quad \text{as } m \rightarrow \infty. \quad (8.286)$$

*Proof.* Since  $G_k$  is a sum of iid Gaussian random variables with mean zero and variance  $\sigma^2$ , we have by Komlós, Major, and Tusnády [160, 161] that  $G_k = \sigma W(k) + \bar{O}_P(\log k)$  where  $(W(t))_{t \geq 0}$  is a standard Wiener process. We will now show that  $\sup_{k \geq n} \left\{ G_k^2/k\sigma^2 - \log(k/n) \right\}$  converges to the Robbins-Siegmund distribution.

$$\sup_{k \geq n} \left\{ \frac{G_k^2}{k\sigma^2} - \log(k/n) \right\} \quad (8.287)$$

$$= \sup_{k \in [n, \infty)} \left\{ \frac{(W(k) + O(\log k))^2}{k\sigma^2} - \log(k/n) + \bar{O}_P \left( \frac{\log(k)\sqrt{k \log \log k}}{k+1} \right) + O \left( \log \left[ \frac{n+1}{n} \right] \right) \right\} \quad (8.288)$$

$$= \sup_{k \in [n, \infty)} \left\{ \frac{W(k)^2}{k\sigma^2} + \bar{O}_P \left( \frac{\log k}{k} \cdot \sqrt{k \log \log k} \right) - \log(k/n) \right\} + \bar{o}_P(1) \quad (8.289)$$

$$= \sup_{k \in [n, \infty)} \left\{ \frac{W(k)^2}{k\sigma^2} - \log(k/n) \right\} + \bar{o}_P(1) \quad (8.290)$$

$$= \sup_{tn \in [n, \infty)} \left\{ W(tn)^2/tn\sigma^2 - \log(tn/n) \right\} + \bar{o}_P(1) \quad (8.291)$$

$$= \sup_{t \in [1, \infty)} \left\{ nW(t)^2/tn\sigma^2 - \log(t) \right\} + \bar{o}_P(1) \quad (8.292)$$

$$= \sup_{t \in [1, \infty)} \left\{ W(t)^2/t\sigma^2 - \log(t) \right\} + \bar{o}_P(1), \quad (8.293)$$

where (8.288) results from the discrete-to-continuous overshoot in  $1/k$  and  $\log(k/n)$  when taking a supremum over  $k \in [n, \infty)$  instead of over  $k \in \{n, n+1, \dots\}$  and (8.292) follows

from elementary properties of the Wiener process. It follows that

$$\sup_{k \geq n} \left\{ \frac{G_k^2}{k\sigma^2} - \log(k/n) \right\} \xrightarrow{d} \Psi \quad \text{as } n \rightarrow \infty. \quad (8.294)$$

□

The following lemma establishes that the Robbins-Siegmund distribution has a Lipschitz CDF.

**Lemma 8.B.3.** *The cumulative distribution function  $\Psi(r)$  of a Robbins-Siegmund-distributed random variable is  $L$ -Lipschitz with  $L \leq 1/4$ . In other words,*

$$\sup_{r \geq 0} \left| \frac{d}{dr} \Psi(r) \right| \leq 1/4. \quad (8.295)$$

*Proof.* Clearly, it suffices to show that  $1 - \Psi(r)$  is  $L$ -Lipschitz. Defining  $f(r) := 1 - \Psi(r)$ , we have that

$$f(r) := 2(1 - \Phi(\sqrt{r}) + \sqrt{r}\phi(\sqrt{r})) \quad (8.296)$$

$$= 2 - 2\Phi(\sqrt{r}) + 2\sqrt{r}\phi(\sqrt{r}). \quad (8.297)$$

A direct calculation reveals that

$$f'(r) = -\phi(\sqrt{r}) \frac{1}{\sqrt{r}} + \frac{1}{\sqrt{r}}\phi(\sqrt{r}) + \sqrt{r}\phi'(\sqrt{r}) \frac{1}{\sqrt{r}} \quad (8.298)$$

$$= \phi'(\sqrt{r}) \quad (8.299)$$

$$= \frac{-\sqrt{r}}{\sqrt{2\pi}} \exp\{-r/2\}, \quad (8.300)$$

from which it is easy to check that  $\sup_{r \geq 0} |f'(r)| \leq 1/4$ , completing the proof. □

## 8.B.2 Uniform convergence of perturbed random variables

Throughout many of our proofs, we rely on facts about convergence of random variables under  $\mathcal{P}$ -uniformly small perturbations. Similar results are common in the proofs of  $\mathcal{P}$ -uniform fixed- $n$  central limit theorems but are only discussed in the context of Gaussian limiting distributions and for time-pointwise convergence. We show here that similar results hold for *Robbins-Siegmund* limiting distributions (in fact, for any continuous and Lipschitz distribution) under time- and  $\mathcal{P}$ -uniformly small perturbations to random variables inside suprema over time.

**Lemma 8.B.4** (Time-uniform closure under additive  $\bar{o}_{\mathcal{P}}(1)$ -perturbations). *Let  $((A_{k,m})_{k=m}^{\infty})_{m=1}^{\infty}$  be a doubly indexed sequence of random variables on  $(\Omega, \mathcal{F}, \mathcal{P})$ . Let  $Z \sim F(z)$  with where the*

CDF  $F$  is  $L$ -Lipschitz and does not depend on  $P \in \mathcal{P}$ . Suppose that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{z \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \geq z \right) - \mathbb{P}_P(Z \geq z) \right| = 0. \quad (8.301)$$

If  $R_n = \bar{o}_{\mathcal{P}}(1)$ , then

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{z \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \right) - \mathbb{P}_P(Z \geq z) \right| = 0. \quad (8.302)$$

*Proof.* Let  $\varepsilon > 0$  be any positive constant. Using (8.301) and the fact that  $R_n = \bar{o}_{\mathcal{P}}(1)$ , let  $M$  be large enough so that for all  $m \geq M$ , we have

$$\sup_{P \in \mathcal{P}} \sup_{z \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \geq z \right) - \mathbb{P}_P(Z \geq z) \right| < \varepsilon \quad (8.303)$$

and so that

$$\sup_{P \in \mathcal{P}} \sup_{z \in \mathbb{R}} \mathbb{P}_P \left( \sup_{k \geq m} |R_k| \geq \varepsilon \right) < \varepsilon. \quad (8.304)$$

Then, writing out  $\mathbb{P}_P(\sup_{k \geq m} \{A_{k,m} + R_k\} \geq z)$  for any  $P \in \mathcal{P}$ ,  $z \in \mathbb{R}$ , and  $m \geq M$ , we find the following upper bound,

$$\mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \right) \quad (8.305)$$

$$= \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \mid \sup_{k \geq m} |R_k| < \varepsilon \right) \underbrace{\mathbb{P}_P(|R_n| < \varepsilon)}_{\leq 1} + \quad (8.306)$$

$$\mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \mid \sup_{k \geq m} |R_k| \geq \varepsilon \right) \underbrace{\mathbb{P}_P \left( \sup_{k \geq m} |R_k| \geq \varepsilon \right)}_{\leq \varepsilon} \quad (8.307)$$

$$\leq \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \geq z - \varepsilon \right) + \varepsilon, \quad (8.308)$$

and via a similar argument, the corresponding lower bound,

$$\mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \right) \quad (8.309)$$

$$\geq \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \geq z + \varepsilon \right) - \varepsilon. \quad (8.310)$$

Using first the upper bound, we thus have that

$$\mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \right) - \mathbb{P}_P(Z \geq z) \quad (8.311)$$

$$\leq \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \geq z - \varepsilon \right) - \mathbb{P}_P(Z \geq z) + \varepsilon \quad (8.312)$$

$$\leq \mathbb{P}_P(Z \geq z - \varepsilon) - \mathbb{P}_P(Z \geq z) + 2\varepsilon \quad (8.313)$$

$$= 1 - F(z - \varepsilon) - (1 - F(z)) + 2\varepsilon \quad (8.314)$$

$$= F(z) - F(z - \varepsilon) + 2\varepsilon \quad (8.315)$$

$$\leq (L + 2)\varepsilon, \quad (8.316)$$

where the last line used the fact that  $F$  is  $L$ -Lipschitz for some  $L > 0$ . Similarly,

$$\mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \right) - \mathbb{P}_P(Z \geq z) \quad (8.317)$$

$$\geq -(L + 2)\varepsilon, \quad (8.318)$$

Putting the two together, we have that

$$\left| \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \right) - \mathbb{P}_P(Z \geq z) \right| \leq (L + 2)\varepsilon, \quad (8.319)$$

and since  $L$  neither depends on  $z$  nor on  $P$ , we have that

$$\left| \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \right) - \mathbb{P}_P(Z \geq z) \right| \leq (L + 2)\varepsilon. \quad (8.320)$$

Since  $\varepsilon$  was arbitrary, it follows that

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{z \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m} + R_k\} \geq z \right) - \mathbb{P}_P(Z \geq z) \right| = 0, \quad (8.321)$$

which completes the proof.  $\square$

**Lemma 8.B.5** (Time-uniform closure under multiplicative  $\bar{o}_{\mathcal{P}}(1)$ -perturbations). *Let  $((A_{k,m})_{k=m}^{\infty})_{m=1}^{\infty}$  be a doubly indexed sequence of random variables on  $(\Omega, \mathcal{F}, \mathcal{P})$ . Let  $Z \sim F(z)$  with CDF  $F$  not depending on  $P \in \mathcal{P}$ . Suppose that*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \leq x \right) - F(x) \right| = 0 \quad (8.322)$$

and suppose that  $R_n = \bar{o}_{\mathcal{P}}(1)$ . Then,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \right) - F(x) \right| = 0. \quad (8.323)$$

*Proof.* The proof proceeds in four steps. In First, we ensure that the CDF of  $\sup_{k \geq m} A_{k,m}$  is

$(\mathcal{P}, x)$ -uniformly close to  $F$ . Second, we use a result of van der Vaart [260] and Slutsky's theorem to justify why deterministically perturbed continuous random variables converge quantile-uniformly in distribution. Third, we use the fact that  $R_n = \bar{o}_{\mathcal{P}}(1)$  to ensure that  $R_n$  is  $\mathcal{P}$ - and time-uniformly smaller than a certain radius. The fourth and final steps puts these results together to arrive at the desired result.

Let  $\varepsilon > 0$  be arbitrary. Our goal is to show that there exists  $M$  sufficiently large so that for all  $m \geq M$ ,

$$\text{Goal: } \sup_{P \in \mathcal{P}} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \right) - F(x) \right| < 2\varepsilon. \quad (8.324)$$

(Here, the multiplication by 2 is only for algebraic convenience later on.)

**Step 1: Ensuring that the CDF of  $\sup_{k \geq m} A_{k,m}$  is  $(\mathcal{P}, x)$ -uniformly close to  $F(x)$ .** By the assumption in (8.322), choose  $M_1$  large enough so that whenever  $m \geq M_1$ , we have

$$\sup_{P \in \mathcal{P}} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \leq x \right) - F(x) \right| < \varepsilon \quad (8.325)$$

**Step 2: CDFs of deterministically perturbed random variables are close to  $F$ .** Letting  $X \sim F$  be a continuous random variable with CDF  $F$ , note that

$$\frac{X}{1 + h} \xrightarrow{d} X \quad (8.326)$$

as  $h \rightarrow 0$  by Slutsky's theorem. Consequently, by van der Vaart [260, Lemma 2.11] combined with the fact that  $F(x)$  is continuous in  $x \in \mathbb{R}$ , we have that

$$\lim_{h \rightarrow 0} \sup_x |F(x(1 + h)) - F(x)| = \lim_{h \rightarrow 0} \sup_x \left| \mathbb{P} \left( \frac{X}{1 + h} \leq x \right) - \mathbb{P}(X \leq x) \right| \quad (8.327)$$

$$= 0. \quad (8.328)$$

As such, let  $h_2 > 0$  be a positive number so that whenever  $|h| \leq h_2$ ,

$$\sup_{x \in \mathbb{R}} |F(x(1 + h)) - F(x)| < \varepsilon. \quad (8.329)$$

**Step 3: Ensuring that  $R_n$  is  $\mathcal{P}$ - and time-uniformly close to 0.** Given the assumption that  $R_n = \bar{o}_{\mathcal{P}}(1)$ , choose  $M_3$  large enough so that for all  $m \geq M_3$ , we have

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left( \sup_{k \geq m} |R_k| \geq h_2 \right) < \varepsilon, \quad (8.330)$$

where  $h_2$  is as in Step 2.

**Step 4: Putting Steps 1–3 together to obtain the final bound.** Set  $M = \max\{M_1, M_3\}$ . First, consider the following upper bound on  $\mathbb{P}_P(\sup_{k \geq m} \{A_{k,m}/(1 + R_k)\} \leq x)$  for any  $m \geq M$ :

$$\mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \right) \quad (8.331)$$

$$= \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \mid \sup_{k \geq m} |R_k| < h_2 \right) \underbrace{\mathbb{P}_P \left( \sup_{k \geq m} |R_k| < h_2 \right)}_{\leq 1} + \quad (8.332)$$

$$\mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \mid \sup_{k \geq m} |R_k| \geq h_2 \right) \underbrace{\mathbb{P}_P \left( \sup_{k \geq m} |R_k| \geq h_2 \right)}_{\leq \varepsilon} \quad (8.333)$$

$$\leq \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + h_2} \right\} \leq x \right) + \varepsilon. \quad (8.334)$$

By a similar argument, we have that for all  $m \geq M$ ,

$$\mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \right) \quad (8.335)$$

$$\geq \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 - h_2} \right\} \leq x \right) - \varepsilon. \quad (8.336)$$

Keeping these upper and lower bounds in mind, we have that

$$\mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \right) - F(x) \quad (8.337)$$

$$\leq \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \leq x(1 + h_2) \right) - F(x) + \varepsilon \quad (8.338)$$

$$\leq F(x(1 + h_2)) - F(x) + \varepsilon \quad (8.339)$$

$$\leq 2\varepsilon. \quad (8.340)$$

and

$$\mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \right) - F(x) \quad (8.341)$$

$$\geq \mathbb{P}_P \left( \sup_{k \geq m} \{A_{k,m}\} \leq x(1 - h_2) \right) - F(x) - \varepsilon \quad (8.342)$$

$$\geq F(x(1 - h_2)) - F(x) - \varepsilon \quad (8.343)$$

$$\geq -2\varepsilon. \quad (8.344)$$

Putting these upper and lower bounds on the difference of probabilities together and noting that their bounds do not depend on  $P \in \mathcal{P}$  nor on  $x \in \mathbb{R}$ , we have

$$\sup_{P \in \mathcal{P}} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_P \left( \sup_{k \geq m} \left\{ \frac{A_{k,m}}{1 + R_k} \right\} \leq x \right) - F(x) \right| \leq 2\varepsilon, \quad (8.345)$$

which completes the proof.

□

# Chapter 9

## Distribution-uniform strong laws of large numbers

### 9.1 Introduction

In his 1951 Berkeley Symposium paper titled “The strong law of large numbers” [62], Kai Lai Chung writes “*For use in certain statistical applications Professor Wald raised the question of the uniformity of the strong [law of large numbers] with respect to a family of [distributions]*”. Chung’s paper proceeds to provide a concrete answer to that question, yielding a generalization of Kolmogorov’s strong law of large numbers (SLLN) that holds uniformly in a rich family of distributions having a uniformly integrable first absolute moment. Let us recall (a minor refinement of) Chung’s distribution-uniform SLLN here.

**Theorem** (Chung’s  $\mathcal{P}$ -uniform strong law of large numbers [62, 229]). *Let  $\mathcal{P}$  be a collection of probability distributions and  $(X_n)_{n=1}^\infty$  be independent and identically distributed random variables defined on the probability spaces  $(\Omega, \mathcal{F}, P) := (\Omega, \mathcal{F}, P)_{P \in \mathcal{P}}$  satisfying the  $\mathcal{P}$ -uniform integrability ( $\mathcal{P}$ -UI) condition*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|X - \mathbb{E}_P(X)| \cdot \mathbb{1}\{|X - \mathbb{E}_P(X)| > m\}) = 0. \quad (9.1)$$

*Then for every  $\varepsilon > 0$ ,*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{k} \sum_{i=1}^k X_i - \mathbb{E}_P(X) \right| \geq \varepsilon \right) = 0. \quad (9.2)$$

Notice that Chung’s SLLN recovers Kolmogorov’s as a special case when the class of distributions  $\dot{\mathcal{P}} = \{P\}$  is taken to be a singleton such that  $\mathbb{E}_P|X| < \infty$  since for any sequence

of random variables  $(Y_n)_{n=1}^\infty$ ,

$$P\left(\lim_{n \rightarrow \infty} Y_n = 0\right) = 1 \quad \text{if and only if} \quad \forall \varepsilon > 0, \lim_{m \rightarrow \infty} P\left(\sup_{k \geq m} |Y_k| \geq \varepsilon\right) = 0. \quad (9.3)$$

The equivalence in (9.3) highlights why Chung's original characterization of the SLLN holding " $\mathcal{P}$ -uniformly" in (9.2) is a natural one. Despite Chung's advance, there are four open questions that we aim to address in this chapter:

- (i) Can the convergence rate in (9.2) be improved in the presence of higher moments in the sense of Marcinkiewicz and Zygmund [184]? That is, can it be shown that for all  $\varepsilon > 0$ ,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\sup_{k \geq m} \left| \frac{1}{k^{1/q}} \sum_{i=1}^k (X_i - \mathbb{E}_P(X)) \right| \geq \varepsilon\right) = 0 \quad (9.4)$$

under certain  $\mathcal{P}$ -UI conditions on the  $q^{\text{th}}$  moment for  $1 < q < 2$ ?

- (ii) Is it possible to restrict the *divergence* rate when  $X$  has fewer than 1 but more than 0 finite absolute moments, again in the sense of Marcinkiewicz and Zygmund [184]? That is, can it be shown that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\sup_{k \geq m} \left| \frac{1}{k^{1/q}} \sum_{i=1}^k X_i \right| \geq \varepsilon\right) = 0, \quad (9.5)$$

under similar  $\mathcal{P}$ -UI conditions but for  $0 < q < 1$  even when  $\mathbb{E}_P|X| = \infty$  for some  $P \in \mathcal{P}$ ?

- (iii) Are  $\mathcal{P}$ -UI conditions *necessary* for  $\mathcal{P}$ -uniform SLLNs to hold (in addition to being sufficient)? That is, if the condition in (9.1) does not hold, can it be shown that for some positive constant  $C > 0$ ,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\sup_{k \geq m} \left| \frac{1}{k} \sum_{i=1}^k (X_i - \mathbb{E}_P(X)) \right| \geq C\right) > 0, \quad (9.6)$$

with analogous questions in the case of higher or lower  $\mathcal{P}$ -UI moments as in (i) and (ii)?

- (iv) Does an analogue of Chung's SLLN exist for independent but *non-identically distributed* random variables, such as in the sense of Petrov [202, §IX, Theorem 12]? That is, can it be shown that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\sup_{k \geq m} \left| \frac{1}{a_k} \sum_{i=1}^k (X_i - \mathbb{E}_P(X_i)) \right| \geq \varepsilon\right) = 0 \quad (9.7)$$

for some appropriately chosen sequence  $a_n \nearrow \infty$ , and if so, under what conditions on  $(X_n)_{n=1}^\infty$ ?

We provide positive answers to (i), (ii), (iii), and (iv) in Theorems 9.2.1(i), 9.2.1(ii), 9.2.1(iii), and

[9.2.2](#), respectively.

*Remark 22* (On centered versus uncentered uniform integrability). As outlined by Ruf et al. [229, Remark 4.5], the assumption displayed in (9.1) is a minor refinement of Chung [62] whose original result made the (stronger) uncentered  $\mathcal{P}$ -UI assumption,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|X| \mathbb{1}\{|X| > m\}) = 0 \quad (9.8)$$

in place of (9.1) but yielding the same conclusion in (9.2). While the difference between (9.1) and (9.8) may seem minor – indeed, for a single  $P \in \mathcal{P}$ ,  $\mathbb{E}_P|X| < \infty$  and  $\mathbb{E}_P|X - \mathbb{E}_P(X)| < \infty$  are equivalent – we highlight in Theorem 9.2.1(iii) how (9.1) is both sufficient *and necessary* for the SLLN to hold, while the same cannot be said for (9.8), drawing an important distinction between the two.

*Remark 23* (On the phrase “uniform integrability”). Note that the phrase “uniform integrability” is commonly used to refer to an analogue of (9.1) holding for a family of *random variables*  $(X_n)_{n=1}^\infty$  on the *same* probability space  $(\Omega, \mathcal{F}, P)$  (as in Chung [63, §4.5], Chong [58], Chandra and Goswami [53], Hu and Rosalsky [126], and Hu and Zhou [127] among others) in the sense that

$$\lim_{m \rightarrow \infty} \sup_{k \in \mathbb{N}} \mathbb{E}_P(|X_k - \mathbb{E}_P(X_k)| \cdot \mathbb{1}\{|X_k - \mathbb{E}_P(X_k)| > m\}) = 0, \quad (9.9)$$

while the presentation in (9.1) is a statement about a *single* random variable on a *collection* of probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  (as also seen in Chow and Teicher [59, pp. 93–94] and Ruf et al. [229, Section 4.2]). Clearly, these two presentations communicate a similar underlying property, but they are used in conceptually different contexts and for this reason, we deliberately write “ $\mathcal{P}$ -UI” to emphasize adherence to (9.1) and avoid ambiguity.

The discussion surrounding (9.3) motivates the following definition which summarizes, extends, and makes succinct Chung’s notion of sequences that vanish both  $\mathcal{P}$ -uniformly and almost surely.

*Definition 9.1.1* (Distribution-uniformly and almost surely vanishing sequences). Let  $\mathcal{P}$  be a collection of distributions and  $(Y_n(P))_{n=1}^\infty$  be random variables defined on  $(\Omega, \mathcal{F}, P)$  for each  $P \in \mathcal{P}$ . We say that  $(Y_n)_{n=1}^\infty \equiv (Y_n(P))_{n=1}^\infty$   $\mathcal{P}$ -uniformly vanishes almost surely if for any  $\varepsilon > 0$ ,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} |Y_k(P)| \geq \varepsilon \right) = 0, \quad (9.10)$$

and as a shorthand for (9.10), we write

$$Y_n = \bar{o}_{\mathcal{P}}(1). \quad (9.11)$$

Moreover, for a monotonic and real sequence  $(r_n)_{n=1}^\infty$ , we say that  $Y_n = \bar{o}_{\mathcal{P}}(r_n)$  if  $Y_n/r_n = \bar{o}_{\mathcal{P}}(1)$ .

Clearly, if a sequence is  $\mathcal{P}$ -uniformly vanishing almost surely, then it is both  $\mathcal{P}$ -uniformly vanishing in probability for the same class  $\mathcal{P}$  as well as vanishing  $P$ -almost surely for every

$P \in \mathcal{P}$ . Using the notation of Definition 9.1.1, the desideratum in (9.4) can be rewritten as  $\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}(X)) = \bar{o}_{\mathcal{P}}(n^{1/q-1})$  with similar presentations for (9.5)–(9.7).

While Definition 9.1.1 is sufficient to *state* our main results (presented in Theorems 9.2.1 and 9.2.2), intermediate steps of their proofs centrally rely on generalizing the notion of almost surely *convergent* (but not necessarily vanishing) sequences to a class of distributions  $\mathcal{P}$  as well as a notion of distribution-uniform stochastic nonincreasingness. Indeed, classical proofs of SLLNs in the  $P$ -pointwise setting rely on showing that certain weighted sums are  $P$ -almost surely convergent (to potentially random quantities depending on  $P$ ), to which the deterministic Kronecker lemma is applied on the same set of  $P$ -probability one to argue that an appropriate sequence *vanishes*. To facilitate a similar argument uniformly in a class  $\mathcal{P}$ , we introduce so-called “ $\mathcal{P}$ -uniform Cauchy sequences” as well as  $\mathcal{P}$ -uniformly stochastically nonincreasing sequences.

**Definition 9.1.2** (Distribution-uniform Cauchy sequence). Let  $\mathcal{P}$  be a collection of distributions and  $(Y_n)_{n=1}^\infty$  a sequence of random variables defined on  $(\Omega, \mathcal{F}, \mathcal{P})$ . We say that  $(Y_n)_{n=1}^\infty$  is a  $\mathcal{P}$ -uniform Cauchy sequence if for any  $\varepsilon > 0$ ,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \sup_{k, n \geq m} |Y_k - Y_n| \geq \varepsilon \right) = 0. \quad (9.12)$$

It is easy to check that a  $\mathcal{P}$ -uniform Cauchy sequence is  $P$ -almost surely a Cauchy sequence (and hence is  $P$ -almost surely convergent) for every  $P \in \mathcal{P}$ . Moreover, if  $Y_n - C = \bar{o}_{\mathcal{P}}(1)$  for any fixed  $C \in \mathbb{R}$ , then  $(Y_n)_{n=1}^\infty$  is  $\mathcal{P}$ -uniformly Cauchy, but the limit point of a  $\mathcal{P}$ -uniform Cauchy sequence can more generally be random and depend on the individual distributions  $P \in \mathcal{P}$ . To control the distribution- and time-uniform stochastic nonincreasingness of sequences (which will routinely appear in the context of limit points of  $\mathcal{P}$ -uniform Cauchy sequences), we introduce the following definition.

**Definition 9.1.3** ( $\mathcal{P}$ -uniformly stochastically nonincreasing). Let  $(Y_n)_{n=1}^\infty$  be a sequence of random variables on  $(\Omega, \mathcal{F}, \mathcal{P})$ . We say that  $(Y_n)_{n=1}^\infty$  is  $\mathcal{P}$ -uniformly stochastically nonincreasing if for every  $\delta > 0$ , there exists  $B_\delta > 0$  so that for every  $n \geq 1$ ,

$$\sup_{P \in \mathcal{P}} P(|Y_n| \geq B_\delta) < \delta, \quad (9.13)$$

and a shorthand for (9.13), we write

$$Y_n = \bar{o}_{\mathcal{P}}(1). \quad (9.14)$$

For a monotonic and real sequence  $(r_n)_{n=1}^\infty$ , we say that  $Y_n = \bar{o}_{\mathcal{P}}(r_n)$  if  $Y_n/r_n = \bar{o}_{\mathcal{P}}(1)$ .

Definition 9.1.3 should be contrasted with the weaker and more familiar notion of  $\mathcal{P}$ -uniform *asymptotic* stochastic boundedness which states that for every  $\delta > 0$ , there exist some  $B_\delta > 0$  and  $N_\delta > 0$  so that for every  $n \geq N_\delta$ , (9.13) holds. In this chapter, we will only be concerned with Definition 9.1.3 which will play an important role when applying Lemma 9.3.1

as the final step in the proofs of Theorems 9.2.1 and 9.2.2.

### 9.1.1 Notation and conventions

Let us now make explicit some notation and conventions that will be used throughout the chapter.

- Individual distributions are denoted by the capital letter  $P$  and collections of distributions are denoted by calligraphic capital letters (typically  $\mathcal{P}$ ).
- We write “ $\mathcal{P}$ -UI” (or simply “UI”) for “ $\mathcal{P}$ -uniformly integrable” when it is clear from context that the phrase is used as an adjective and “ $\mathcal{P}$ -uniform integrability” when used as a noun.
- Collections of probability spaces are written as  $(\Omega, \mathcal{F}, \mathcal{P})$  to denote  $(\Omega, \mathcal{F}, P)_{P \in \mathcal{P}}$ .
- If the  $q^{\text{th}}$  absolute central moment of  $X$  is  $\mathcal{P}$ -UI, i.e.

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|X - \mathbb{E}_P(X)|^q \mathbf{1}\{|X - \mathbb{E}_P(X)|^q > m\}) = 0, \quad (9.15)$$

we condense this to “the  $q^{\text{th}}$  moment of  $X$  is  $\mathcal{P}$ -UI” and omit the qualifiers “absolute” and “central”.

- The phrase “independent and identically distributed” is abbreviated to “i.i.d.”.
- For real numbers  $a, b \in \mathbb{R}$ , we use  $a \wedge b$  to denote  $\min\{a, b\}$  and  $a \vee b$  to denote  $\max\{a, b\}$ .
- We write  $b_n \nearrow \infty$  for a real sequence  $(b_n)_{n=1}^\infty$  if it is nondecreasing and diverging to  $\infty$ .
- We omit the subscript  $P$  from  $\mathbb{E}_P(X)$  when using the shorthand notation in (9.11). For example, we write  $\frac{1}{n} \sum_{i=1}^n X_i - \mathbb{E}(X) = \bar{o}_{\mathcal{P}}(1)$  if in fact

$$\forall \varepsilon > 0, \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} |X_k - \mathbb{E}_P(X)| \geq \varepsilon \right) = 0. \quad (9.16)$$

- If  $\mathbb{E}_P|X_n|$  is finite for every  $P \in \mathcal{P}$  and every  $n \in \mathbb{N}$ , we say that “the SLLN holds at a rate of  $\bar{o}_{\mathcal{P}}(a_n/n)$ ” to mean that  $a_n^{-1} \sum_{i=1}^n (X_i - \mathbb{E}(X_i)) = \bar{o}_{\mathcal{P}}(1)$  for the sequence  $a_n \nearrow \infty$ . Similarly, if  $\mathbb{E}_P|X_n| = \infty$  for some  $P \in \mathcal{P}$  and some  $n \in \mathbb{N}$ , then we use the same phrase “the SLLN holds at a rate of  $\bar{o}_{\mathcal{P}}(a_n/n)$ ” if  $a_n^{-1} \sum_{i=1}^n X_i = \bar{o}_{\mathcal{P}}(1)$ . For example, Chung [62] gives conditions under which the SLLN holds at a rate of  $\bar{o}_{\mathcal{P}}(1)$ .

### 9.1.2 Outline and summary of contributions

Below we outline how the chapter will proceed, summarizing our main contributions.

- Section 9.2 contains our main results — Theorems 9.2.1(i), 9.2.1(ii), 9.2.1(iii), and 9.2.2 — which provide answers to the questions posed in (9.4), (9.5), (9.6), and (9.7), respectively. In short, these theorems show that the SLLN holds at a rate of  $\bar{o}_{\mathcal{P}}(n^{1/q-1})$  in the i.i.d. case *if and only if* they have a  $\mathcal{P}$ -UI  $q^{\text{th}}$  moment, and that it holds at a rate of  $\bar{o}_{\mathcal{P}}(a_n/n)$  in the non-i.i.d. case if  $\sum_{k=m}^\infty \mathbb{E}|X_k - \mathbb{E}X_k|^q/a_k^q$  vanishes  $\mathcal{P}$ -uniformly as  $m \rightarrow \infty$ .

- Section 9.3 contains distribution-uniform analogues of several almost sure convergence results that are commonly used in the proofs of SLLNs. These include analogues of the Khintchine-Kolmogorov convergence theorem (Section 9.3.1), the Kolmogorov three-series theorem (Section 9.3.2), Kronecker's lemma (Section 9.3.4), and the Borel-Cantelli lemmas (Section 9.3.5). These results rely on the notion of a distribution-uniform Cauchy sequence, whose definition is provided in Definition 9.1.2 and which serves as a  $\mathcal{P}$ -uniform generalization of a sequence that is  $P$ -almost surely convergent.
- Section 9.4 contains complete proofs to Theorems 9.2.1 and 9.2.2. After considering the “right” generalizations of distribution-uniform convergence (in Definitions 9.1.1 and 9.1.2), the high-level structure of the proofs to Theorems 9.2.1(i), 9.2.1(ii), and 9.2.2 largely mirror those of their  $P$ -pointwise counterparts due to Kolmogorov, Marcinkiewicz, and Zygmund in the sense that they use analogous technical theorems and lemmas from Section 9.3 in similar succession. One exception to this is the combination of Kolmogorov's three-series theorem and Kronecker's lemma — certain subtleties surrounding uniform stochastic nonincreasingness of  $\mathcal{P}$ -uniform Cauchy sequences requires the introduction of another three-series theorem provided in Theorem 9.3.3. Furthermore, our proofs noticeably deviate from their  $P$ -pointwise counterparts in satisfying the conditions of our  $\mathcal{P}$ -uniform three series theorems (Theorems 9.3.2 and 9.3.3). These require additional care in both cases, relying for example on a few delicate applications of the de la Vallée Poussin criterion of uniform integrability; details can be found in Lemmas 9.4.1–9.4.7.
- Section 9.5 illustrates an application of Theorem 9.2.1(i) to the derivation of rates of uniform consistency in the statistical problem of variance estimation. While the literature has seen statistical applications relying on quantitative rates of strong consistency of the sample variance, they have thus far been distribution-*pointwise* results; to the best of our knowledge, Corollary 9.5.1 is the first to quantify such rates uniformly in a class of distributions.

## 9.2 Distribution-uniform strong laws of large numbers

We begin by presenting our first main result which gives both necessary and sufficient conditions for the SLLN to hold at a rate of  $\bar{o}_{\mathcal{P}}(n^{1/q-1})$  in the i.i.d. setting, providing answers to the questions posed in (9.4), (9.5), and (9.6).

**Theorem 9.2.1** ( $\mathcal{P}$ -uniform Marcinkiewicz-Zygmund strong law of large numbers). *Let  $(X_n)_{n=1}^{\infty}$  be independent and identically distributed random variables. Consider the following  $\mathcal{P}$ -UI condition for some  $0 < q < 2$ :*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|X - \mu(P; q)|^q \mathbf{1}\{|X - \mu(P; q)|^q > m\}) = 0, \quad (9.17)$$

where  $\mu(P; q) = \mathbb{E}_P(X)$  if  $1 \leq q < 2$  and  $\mu(P; q) = 0$  if  $0 < q < 1$ .

(i) If (9.17) holds with  $q \in [1, 2)$ , then

$$\frac{1}{n^{1/q}} \sum_{i=1}^n (X_i - \mathbb{E}(X)) = \bar{o}_{\mathcal{P}}(1). \quad (9.18)$$

(ii) If (9.17) holds with  $q \in (0, 1)$ , then

$$\frac{1}{n^{1/q}} \sum_{i=1}^n X_i = \bar{o}_{\mathcal{P}}(1). \quad (9.19)$$

(iii) If (9.17) does not hold, then

$$\frac{1}{n^{1/q}} \sum_{i=1}^n (X_i - \mu(P; q)) \neq \bar{o}_{\mathcal{P}}(1). \quad (9.20)$$

In other words, the  $\mathcal{P}$ -uniform SLLN holds for the average  $\frac{1}{n} \sum_{i=1}^n X_i$  with a rate of  $\bar{o}_{\mathcal{P}}(n^{1/q-1})$  if and only if the  $q^{\text{th}}$  moment of  $X$  is  $\mathcal{P}$ -UI.

In the same way that Chung's  $\mathcal{P}$ -uniform SLLN for UI first moments generalizes Kolmogorov's  $P$ -pointwise SLLN for finite first moments, Theorems 9.2.1(i) and 9.2.1(ii) generalize the Marcinkiewicz-Zygmund [184]  $P$ -pointwise SLLN for finite  $q^{\text{th}}$  moments when  $0 < q < 2$ ;  $q \neq 1$ , painting a fuller picture of sufficiency for  $\mathcal{P}$ -uniform SLLNs in the i.i.d. case.

Turning to Theorem 9.2.1(iii), the *necessity* of  $\mathcal{P}$ -UI appears to be new to the literature even in the case of  $q = 1$ . In fact, Chung's original paper [62] studied necessary conditions for the  $\mathcal{P}$ -uniform SLLN but only considered *uncentered* UI as in (9.8) which turns out not to be necessary in general. Concretely, he showed that if the SLLN in (9.18) holds for  $q = 1$  and the median of  $X$  is uniformly bounded, then the uncentered  $\mathcal{P}$ -UI condition in (9.8) holds; in other words, if  $\sup_{P \in \mathcal{P}} |\text{med}_P(X)| < \infty$  where  $\text{med}_P(X) := \sup\{x : P(X \leq x) \leq 1/2\}$ , then

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}(X)) = \bar{o}_{\mathcal{P}}(1) \implies \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|X| \mathbf{1}\{|X| > m\}) = 0. \quad (9.21)$$

Chung [62, Remark 2] uses a simple counterexample to point out that without uniform boundedness of the medians, uncentered  $\mathcal{P}$ -UI is *not* necessary. Indeed, letting  $\mathcal{P}_{\mathbb{N}} := \{P_n : n \in \mathbb{N}\}$  where  $P_n$  is the distribution of  $X$  with a point mass at  $x = n$ , we obviously have that the uniform SLLN holds (since the centered sample average is always 0 with  $P$ -probability one for all  $P \in \mathcal{P}_{\mathbb{N}}$ ) and yet  $X$  does not satisfy uncentered  $\mathcal{P}_{\mathbb{N}}$ -uniform integrability. Clearly, this counterexample does not apply to the *centered* uniform integrability condition we are considering in (9.17).

Theorem 9.2.1(iii) also highlights that uniform *boundedness* of the  $q^{\text{th}}$  moment is not sufficient for the SLLN to hold  $\mathcal{P}$ -uniformly at a rate of  $o(n^{1/q-1})$ . Going further, by the de la Vallée Poussin criterion for uniform integrability [58], the SLLN holding uniformly at this rate

is equivalent to the uniform boundedness of  $\mathbb{E}_P \varphi(|X|^q)$  for some positive and nondecreasing function  $\varphi$  growing faster than  $x \mapsto x$ , i.e.  $\lim_{x \rightarrow \infty} \varphi(x)/x = \infty$ . Let us now give rough outlines of the proofs of Theorems 9.2.1(i), 9.2.1(ii), and 9.2.1(iii) (with a diagrammatic overview of the former displayed in Figure 9.1), leaving most technical details for Section 9.4.1.

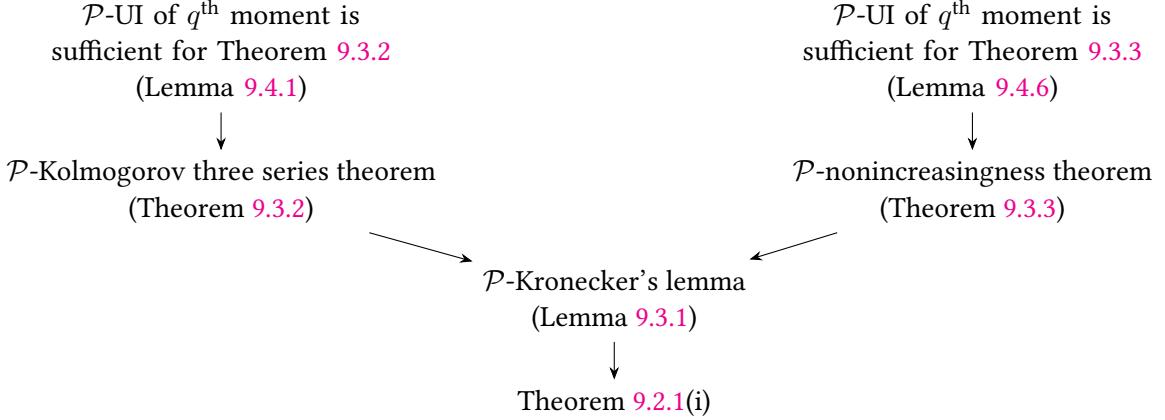


Figure 9.1: A diagrammatic summary of the theorems and lemmas required to prove Theorem 9.2.1(i). Note that when  $\mathcal{P} = \{P\}$  is a singleton, the proof due to Marcinkiewicz and Zygmund [184] only involves the left branch (Lemma 9.4.1 → Theorem 9.3.2 → Lemma 9.3.1) since the right one is trivially satisfied in that case (more discussion can be found in Section 9.3.3). The above structure similarly applies to the proof of Theorem 9.2.1(ii) but we replace Lemmas 9.4.1 and 9.4.6 with Lemmas 9.4.2 and 9.4.7 for the sake of satisfying the conditions of Theorems 9.3.2 and 9.3.3, respectively.

*Proof outline of Theorem 9.2.1(i).* Since  $q = 1$  corresponds to the SLLN of Chung [62], we focus on  $1 < q < 2$ . Similar to classical SLLN proofs, we focus our attention on the weighted random variables  $(Z_n)_{n=1}^\infty$  given by  $Z_n := (X_n - \mathbb{E}_P(X_n))/n^{1/q}$ . First, in Theorem 9.3.2 we develop a  $\mathcal{P}$ -uniform analogue of the Kolmogorov three-series theorem which states that if for some  $c > 0$ ,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} |\mathbb{E}_P Z_n^{\leq c}| = 0, \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \text{Var}_P Z_n^{\leq c} = 0, \quad \text{and} \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} P(|Z_n| > c) = 0, \quad (9.22)$$

where  $Z_n^{\leq c} := Z_n \mathbf{1}\{Z_n \leq c\}$ , then  $S_n := \sum_{i=1}^n Z_i$  is a  $\mathcal{P}$ -uniform Cauchy sequence. Indeed, Lemma 9.4.1 focuses on exploiting  $\mathcal{P}$ -UI of the  $q^{\text{th}}$  moment to show that (9.22) holds with  $c = 1$ .

We then introduce the  $\mathcal{P}$ -uniform stochastic Kronecker lemma (Lemma 9.3.1) which states that if  $S_n$  is  $\mathcal{P}$ -uniformly Cauchy and  $\mathcal{P}$ -uniformly stochastically nonincreasing, then for any  $b_n \nearrow \infty$ , we have

$$\frac{1}{b_n} \sum_{i=1}^n b_i Z_i = \bar{o}_{\mathcal{P}}(1). \quad (9.23)$$

To apply Lemma 9.3.1 to our setting, we show that  $S_n$  is  $\mathcal{P}$ -uniformly Cauchy as a consequence of the three-series theorem discussed above combined with Lemma 9.4.1, and to show that  $S_n$  is  $\mathcal{P}$ -uniformly stochastically nonincreasing, we introduce another three-series-type theorem in Theorem 9.3.3 which states that if

$$\lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} |\mathbb{E}_P [(Z_n/B) \mathbf{1}\{|Z_n/B| \leq 1\}]| = 0, \quad (9.24)$$

$$\lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \text{Var}_P [(Z_n/B) \mathbf{1}\{|Z_n/B| \leq 1\}] = 0, \text{ and} \quad (9.25)$$

$$\lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} P(|Z_n/B| > 1) = 0, \quad (9.26)$$

then  $S_n := \sum_{i=1}^n Z_i$  is  $\mathcal{P}$ -uniformly stochastically nonincreasing in the sense of Definition 9.1.3. Lemma 9.4.6 indeed shows that the above three series conditions are satisfied as long as  $X$  has a  $\mathcal{P}$ -UI  $q^{\text{th}}$  moment. Taking the sequence  $(b_n)_{n=1}^{\infty}$  to be given by  $b_n = n^{1/q}$  and invoking the  $\mathcal{P}$ -uniform Kronecker lemma yields the desired result:

$$\frac{1}{n^{1/q}} \sum_{i=1}^n (X_i - \mathbb{E}(X)) = \bar{o}_{\mathcal{P}}(1), \quad (9.27)$$

which completes the proof outline of Theorem 9.2.1(i).  $\square$

*Proof outline of Theorem 9.2.1(ii).* The proof in the case of  $0 < q < 1$  proceeds in the same manner as that of  $1 < q < 2$  but instead of Lemma 9.4.1 showing that the three-series conditions in (9.22) are satisfied for  $(X_n - \mathbb{E}_P(X_n))/n^{1/q}$ , it is Lemma 9.4.2 that shows that these conditions are satisfied for  $X_n/n^{1/q}$ , thereby demonstrating that  $\sum_{k=1}^n X_k/k^{1/q}$  is  $\mathcal{P}$ -uniformly Cauchy. Similarly, rather than using Lemma 9.4.6 to satisfy the  $\mathcal{P}$ -uniform stochastic nonincreasingness three series above, we use Lemma 9.4.7. Again, invoking the  $\mathcal{P}$ -uniform stochastic Kronecker lemma yields the desired result.  $\square$

*Proof outline of Theorem 9.2.1(iii).* We will describe the proof outline for the case where  $1 \leq q < 2$  but a similar argument goes through for  $0 < q < 1$  (with all details provided in Section 9.4.1). The proof relies on a  $\mathcal{P}$ -uniform generalization of the second Borel-Cantelli lemma (Lemma 9.3.2) which states that for independent events  $(E_n)_{n=1}^{\infty}$  in  $\mathcal{F}$ , if the tails of the sums of  $(P(E_n))_{n=1}^{\infty}$  do not uniformly vanish, then the probability of the tails of the *unions* of  $(E_n)_{n=1}^{\infty}$  do not uniformly vanish; more succinctly:

$$0 < \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P(E_k) \leq \infty \implies \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\bigcup_{k=m}^{\infty} E_k\right) > 0. \quad (9.28)$$

To make use of this lemma, we highlight that for any  $P \in \mathcal{P}$ ,

$$P\left(\sup_{k \geq m-1} \frac{1}{k^{1/q}} |S_k| \geq 1/2\right) \geq P\left(\sup_{k \geq m} \frac{1}{k^{1/q}} |X - \mathbb{E}_P(X)| \geq 1\right) \quad (9.29)$$

where  $S_k := \sum_{i=1}^k (X_i - \mathbb{E}_P(X_i))$  are the centered partial sums, and hence once paired with (9.28), it suffices to show that

$$0 < \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P(|X_k - \mathbb{E}_P(X_k)|^q > k) \leq \infty. \quad (9.30)$$

Indeed by Hu and Zhou [127, Theorem 2.1], (9.30) is equivalent to the  $\mathcal{P}$ -UI condition in (9.17) being violated, i.e. (9.30) holds if and only if

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|X - \mathbb{E}_P(X)|^q \mathbf{1}\{|X - \mathbb{E}_P(X)|^q > m\}) > 0, \quad (9.31)$$

which completes the proof outline of Theorem 9.2.1(iii).  $\square$

Let us now consider the setting of independent but *non-identically distributed* random variables. The following theorem serves as a distribution-uniform generalization of the well-known strong law of large numbers for independent random variables (see Petrov [202, §IX, Theorem 12]).

**Theorem 9.2.2** ( $\mathcal{P}$ -uniform strong law for non-identically distributed random variables). *Let  $(X_n)_{n=1}^{\infty}$  be independent random variables and suppose that for some  $q \in [1, 2]$ , they each have a finite absolute  $q^{\text{th}}$  central moment. Suppose that for some  $a_n \nearrow \infty$ , we have*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q}{a_k^q} = 0. \quad (9.32)$$

*Then the SLLN holds  $\mathcal{P}$ -uniformly at a rate of  $\bar{o}_{\mathcal{P}}(a_n/n)$ , meaning for any  $\varepsilon > 0$ ,*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\sup_{k \geq m} \left|\frac{1}{a_k} \sum_{i=1}^k (X_i - \mathbb{E}_P(X_i))\right| \geq \varepsilon\right) = 0. \quad (9.33)$$

*Proof outline of Theorem 9.2.2.* The proof of Theorem 9.2.2 is identical to that of Theorem 9.2.1(i) but instead of using Lemma 9.4.1 and Lemma 9.4.6 to satisfy the conditions of the  $\mathcal{P}$ -uniform Kolmogorov and stochastic nonincreasingness three-series theorems (Theorems 9.3.2 and 9.3.3), we use different arguments found in Lemmas 9.4.9 and 9.4.10, respectively. (In fact, the latter two lemmas are simpler and require much softer arguments.) This completes the proof outline of Theorem 9.2.2.  $\square$

After some inspection, the reader will notice that when instantiated in the identically distributed setting, Theorem 9.2.2 does not recover Theorem 9.2.1, meaning that it cannot

attain an SLLN rate as fast as  $o(n^{1/q-1})$  in the presence of only  $q \in [1, 2)$   $\mathcal{P}$ -UI moments. This is not surprising and it directly mirrors the relationship between the  $P$ -pointwise non-i.i.d. SLLNs [202, §IX, Theorem 12] and the strong laws of Kolmogorov, Marcinkiewicz, and Zygmund in the i.i.d. case. The latter proofs in the  $P$ -pointwise case (and now ours provided in Section 9.4.1 for the  $\mathcal{P}$ -uniform case) all crucially exploit the fact that  $(X_n)_{n=1}^\infty$  are identically distributed.

As alluded to in the proof outlines of Theorems 9.2.1 and 9.2.2, our results rely on  $\mathcal{P}$ -uniform analogues of several familiar almost sure convergence results. We present all of these in the next section.

### 9.3 Other distribution-uniform convergence results

In this section, we provide  $\mathcal{P}$ -uniform analogues of various almost sure convergence results including the Khintchine-Kolmogorov convergence theorem, the Kolmogorov three-series theorem, and a stochastic generalization of Kronecker's lemma. These are instrumental to the proofs of Theorems 9.2.1 and 9.2.2. Note that the classical ( $P$ -pointwise) forms of these results are stated (either in their assumptions or in their conclusions) in terms of a sequence of random variables *converging  $P$ -almost surely*, and hence we will rely heavily on the notion of  $\mathcal{P}$ -uniform Cauchy sequences provided in Definition 9.1.2. We begin in the following section with a  $\mathcal{P}$ -uniform generalization of the Khintchine-Kolmogorov convergence theorem.

#### 9.3.1 A distribution-uniform Khintchine-Kolmogorov convergence theorem

In the classical  $P$ -pointwise case, the Khintchine-Kolmogorov convergence theorem states that for a sequence of independent random variables  $(X_n)_{n=1}^\infty$ , if the sum of their variances is finite, i.e.

$$\sum_{k=1}^{\infty} \text{Var}_P(X_k) < \infty, \quad (9.34)$$

then  $\sum_{k=1}^{\infty} X_k$  is  $P$ -almost surely finite. With Definition 9.1.2 in mind, we are ready to state and prove a  $\mathcal{P}$ -uniform generalization of the Khintchine-Kolmogorov convergence theorem, establishing that the sum  $\sum_{k=1}^{\infty} X_k$  is  $\mathcal{P}$ -uniformly Cauchy whenever the series in (9.34) has  $\mathcal{P}$ -uniformly vanishing tails.

**Theorem 9.3.1** ( $\mathcal{P}$ -uniform Khintchine-Kolmogorov convergence theorem). *Let  $(X_n)_{n=1}^\infty$  be independent random variables on  $(\Omega, \mathcal{F}, \mathcal{P})$ . If*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \text{Var}_P X_k = 0 \quad (9.35)$$

*then  $S_n := \sum_{i=1}^n (X_i - \mathbb{E}_P(X_i))$  is  $\mathcal{P}$ -uniformly Cauchy (Definition 9.1.2).*

*Proof.* First note that for any  $m \geq 1$ , we have that

$$\left\{ \sup_{k,n \geq m} |S_k - S_n| \geq \varepsilon \right\} \subseteq \left\{ \sup_{k \geq m} |S_k - S_m| \geq \varepsilon/2 \right\} \cup \left\{ \sup_{n \geq m} |S_n - S_m| \geq \varepsilon/2 \right\} \quad (9.36)$$

and hence for any  $P \in \mathcal{P}$ , Kolmogorov's inequality yields

$$P \left( \sup_{k,n \geq m} |S_k - S_n| \geq \varepsilon \right) \leq P \left( \sup_{k \geq m} |S_k - S_m| \geq \varepsilon/2 \right) + P \left( \sup_{n \geq m} |S_n - S_m| \geq \varepsilon/2 \right) \quad (9.37)$$

$$\leq \frac{8}{\varepsilon^2} \cdot \sum_{k=m+1}^{\infty} \text{Var}_P(X_k). \quad (9.38)$$

Taking suprema over  $P \in \mathcal{P}$  and limits as  $m \rightarrow \infty$  and noting the condition in (9.35) yields the desired result, completing the proof.  $\square$

### 9.3.2 A distribution-uniform Kolmogorov three-series theorem

Now that we have a  $\mathcal{P}$ -uniform Khintchine-Kolmogorov convergence theorem, we will use it to prove a  $\mathcal{P}$ -uniform analogue of Kolmogorov's three-series theorem. To begin, define the truncated version  $X^{\leq c}$  of a random variable  $X$  at a constant  $c$  as

$$X^{\leq c} := X \cdot \mathbb{1}\{|X| \leq c\}. \quad (9.39)$$

In the  $P$ -pointwise case, recall that Kolmogorov's three-series theorem states that if the following three series are convergent for some  $c > 0$ :

$$\sum_{n=1}^{\infty} \mathbb{E}_P X_n^{\leq c}, \quad \sum_{n=1}^{\infty} \text{Var}_P X_n^{\leq c}, \quad \text{and} \quad \sum_{n=1}^{\infty} P(|X_n| > c), \quad (9.40)$$

then  $\sum_{k=1}^{\infty} X_k$  is  $P$ -almost surely convergent. Similarly to Theorem 9.3.1 in the previous section, our  $\mathcal{P}$ -uniform analogue of Kolmogorov's three series theorem will conclude that  $\sum_{k=1}^{\infty} X_k$  is  $\mathcal{P}$ -uniformly Cauchy as long as the tails of a certain three series are  $\mathcal{P}$ -uniformly vanishing.

**Theorem 9.3.2** ( $\mathcal{P}$ -uniform Kolmogorov three-series theorem). *Let  $(X_n)_{n=1}^{\infty}$  be a sequence of independent random variables. Suppose that the following three summation tails decay  $\mathcal{P}$ -uniformly for some  $c > 0$ :*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} |\mathbb{E}_P X_n^{\leq c}| = 0, \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \text{Var}_P X_n^{\leq c} = 0, \quad \text{and} \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} P(|X_n| > c) = 0.$$

*Then  $S_n := \sum_{i=1}^n X_i$  is  $\mathcal{P}$ -uniformly Cauchy.*

Notice that the first series in (9.40) does not have an exact analogue in Theorem 9.3.2 since the former is not a sum of *absolute values* of  $(\mathbb{E}_P X_n^{\leq c})_{n=1}^\infty$  while that of the latter is. In particular, Theorem 9.3.2 is not a *strict* generalization of Kolmogorov's three-series theorem in general, but this distinction is inconsequential for the sake of proving ( $\mathcal{P}$ -uniform or  $P$ -pointwise) SLLNs, at least in the i.i.d. and independent but non-i.i.d. settings considered by Kolmogorov, Marcinkiewicz, and Zygmund, as well as Petrov [202, Theorem 12]. Indeed, all of their (and our) proofs ultimately upper bound  $\mathbb{E}_P X_n^{\leq c}$  for a mean-zero  $X_n$  by  $\mathbb{E}_P(|X_n| \cdot \mathbf{1}\{|X_n| \leq c\})$  or by  $\mathbb{E}_P(|X_n| \cdot \mathbf{1}\{|X_n| > c\})$ , and hence one can simply analyze  $|\mathbb{E}_P X_n^{\leq c}|$  from the outset. Detailed discussions and proofs can be found in Section 9.4.1.1. Let us now return to and prove Theorem 9.3.2.

*Proof.* Abusing notation slightly, let  $S_n^{\leq c} := \sum_{i=1}^n X_i^{\leq c}$ . Note that for any  $m \geq 1$  and any  $\varepsilon > 0$ , we have

$$\sup_{P \in \mathcal{P}} P \left( \sup_{n, k \geq m} |S_n - S_k| \geq \varepsilon \right) \quad (9.41)$$

$$\leq \sup_{P \in \mathcal{P}} P \left( \sup_{n \geq k \geq m} |S_n^{\leq c} - S_k^{\leq c}| \geq \varepsilon \right) + \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P(|X_k| > c). \quad (9.42)$$

The second term above vanishes asymptotically by the third series, so it suffices to show that the first term goes to 0 as  $m \rightarrow \infty$ . Indeed,

$$\sup_{P \in \mathcal{P}} P \left( \sup_{n \geq k \geq m} |S_n^{\leq c} - S_k^{\leq c}| \geq \varepsilon \right) \quad (9.43)$$

$$\leq \underbrace{\sup_{P \in \mathcal{P}} \left( \sup_{n \geq k \geq m} |S_n^{\leq c} - \mathbb{E}_P S_n^{\leq c} - (S_k^{\leq c} - \mathbb{E}_P(S_k^{\leq c}))| \geq \varepsilon/2 \right)}_{(\star)} + \quad (9.44)$$

$$\underbrace{\sup_{P \in \mathcal{P}} \mathbf{1} \left\{ \sup_{k \geq n \geq m} |\mathbb{E}_P S_n^{\leq c} - \mathbb{E}_P S_k^{\leq c}| \geq \varepsilon/2 \right\}}_{(\dagger)}. \quad (9.45)$$

Now,  $(\star) \rightarrow 0$  by the second series combined with the  $\mathcal{P}$ -uniform Khintchine-Kolmogorov convergence theorem (Theorem 9.3.1). Turning to  $(\dagger)$ , we have

$$\sup_{P \in \mathcal{P}} \mathbf{1} \left\{ \sup_{k \geq n \geq m} |\mathbb{E}_P S_n^{\leq c} - \mathbb{E}_P S_k^{\leq c}| \geq \varepsilon/2 \right\} \quad (9.46)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbf{1} \left\{ \sum_{i=m}^{\infty} |\mathbb{E}_P X_i^{\leq c}| \geq \varepsilon/4 \right\} + \sup_{P \in \mathcal{P}} \mathbf{1} \left\{ \sum_{i=m}^{\infty} |\mathbb{E}_P X_i^{\leq c}| \geq \varepsilon/4 \right\} \quad (9.47)$$

which is zero for sufficiently large  $m$  by the first of the three series, completing the proof.  $\square$

### 9.3.3 A three-series theorem for stochastic nonincreasingness

In addition to certain partial sums being  $\mathcal{P}$ -uniformly Cauchy, an important condition that will appear throughout our proofs — namely in the application of the uniform Kronecker lemma (Lemma 9.3.1) — is that of  $\mathcal{P}$ -uniform stochastic nonincreasingness (see Definition 9.1.3). Here, we provide a three-series theorem providing sufficient conditions for partial sums to be  $\mathcal{P}$ -uniformly stochastically nonincreasing and whose conditions are similar in spirit to those of Theorem 9.3.2.

**Theorem 9.3.3** (A three-series theorem for  $\mathcal{P}$ -uniform stochastic nonincreasingness). *Let  $(Z_n)_{n=1}^\infty$  be independent random variables on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$ . Define the truncated random variables  $(Z_{n,B}^{\leq 1})_{n=1}^\infty$  given by*

$$Z_{n,B}^{\leq 1} := (Z_n/B) \cdot \mathbb{1}\{|Z_n/B| \leq 1\}. \quad (9.48)$$

Suppose that the following three series uniformly vanish as  $B \rightarrow \infty$ :

$$\lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^\infty \left| \mathbb{E}_P Z_{n,B}^{\leq 1} \right| = 0, \quad \lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^\infty \text{Var}_P Z_{n,B}^{\leq 1} = 0, \quad \text{and} \quad \lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^\infty P(|Z_n| > B) = 0.$$

Then,  $S_n := \sum_{k=1}^n Z_k$  is  $\mathcal{P}$ -uniformly stochastically nonincreasing, meaning that for any  $\delta > 0$ , there exists  $B_\delta > 0$  so that for all  $n \geq 1$ ,

$$\sup_{P \in \mathcal{P}} P(|S_n| \geq B_\delta) < \delta. \quad (9.49)$$

*Proof of Theorem 9.3.3.* Let  $\delta > 0$  be arbitrary. Our goal is to show that there exists  $B_\delta$  large enough so that for all  $n \geq 1$ , (9.49) holds. Notice that for any  $B > 0$ , we have that

$$\sup_{P \in \mathcal{P}} P(|S_n| > B) \quad (9.50)$$

$$= \sup_{P \in \mathcal{P}} P \left( \left| \sum_{k=1}^n Z_k \right| > B \right) \quad (9.51)$$

$$\leq \underbrace{\sup_{P \in \mathcal{P}} P \left( \left| \sum_{k=1}^n (Z_k/B) \mathbb{1}\{|Z_k/B| \leq 1\} \right| > 1 \right)}_{(*)} + \underbrace{\sup_{P \in \mathcal{P}} \sum_{k=1}^n P(|Z_k/B| > 1)}_{(\dagger)}. \quad (9.52)$$

Letting  $S_{n,B}^{\leq 1} := \sum_{k=1}^n Z_{k,B}^{\leq 1}$  be the partial sums of the truncated random variables  $Z_{n,B}^{\leq 1} := (Z_n/B) \cdot \mathbb{1}\{|Z_n/B| \leq 1\}$ , notice that we can write  $(*)$  as

$$\sup_{P \in \mathcal{P}} P \left( \left| \sum_{k=1}^n (Z_k/B) \cdot \mathbb{1}\{|Z_k/B| \leq 1\} \right| > 1 \right) \quad (9.53)$$

$$= \sup_{P \in \mathcal{P}} P \left( \left| S_{n,B}^{\leq 1} - \mathbb{E}_P S_{n,B}^{\leq 1} + \mathbb{E}_P S_{n,B}^{\leq 1} \right| > 1 \right) \quad (9.54)$$

$$\leq \sup_{P \in \mathcal{P}} P \left( |S_{n,B}^{\leq} - \mathbb{E}_P S_{n,B}^{\leq}| > 1/2 \right) + \sup_{P \in \mathcal{P}} P \left( |\mathbb{E}_P S_{n,B}^{\leq}| > 1/2 \right) \quad (9.55)$$

$$= \underbrace{\sup_{P \in \mathcal{P}} P \left( |S_{n,B}^{\leq} - \mathbb{E}_P S_{n,B}^{\leq}| > 1/2 \right)}_{(\star i)} + \underbrace{\sup_{P \in \mathcal{P}} \mathbf{1} \left\{ |\mathbb{E}_P S_{n,B}^{\leq}| > 1/2 \right\}}_{(\star ii)}. \quad (9.56)$$

By Kolmogorov's inequality, we have that

$$(\star i) \leq 4 \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \text{Var}_P ((Z_k/B) \cdot \{|Z_k/B| \leq 1\}). \quad (9.57)$$

Furthermore, by the triangle inequality and upper bounding the finite sum by an infinite one, we have

$$(\star ii) \leq \sup_{P \in \mathcal{P}} \mathbf{1} \left\{ \sum_{n=1}^{\infty} |\mathbb{E}_P [(Z_n/B) \mathbf{1}\{|Z_n/B| \leq 1\}]| > 1/2 \right\}. \quad (9.58)$$

Once again upper bounding a finite sum by an infinite one, we have

$$(\dagger) \leq \sum_{k=1}^{\infty} P(|Z_k/B| > 1). \quad (9.59)$$

Therefore, using the first, second, and third series conditions, we can find  $B_\delta > 0$  so that for all  $n \geq 1$ , we have  $(\star i) \leq \delta/2$ ,  $(\star ii) = 0$ , and  $(\dagger) \leq \delta/2$ , respectively, and thus

$$\sup_{P \in \mathcal{P}} P(|S_n| > B_\delta) \leq \delta, \quad (9.60)$$

completing the proof.  $\square$

### 9.3.4 A distribution-uniform stochastic generalization of Kronecker's lemma

In the classical  $P$ -pointwise setting, proofs of strong laws of large numbers rely on a (non-stochastic) convergence result known as *Kronecker's lemma* which states that if  $(x_n)_{n=1}^{\infty}$  is a sequence of real numbers so that  $\sum_{i=1}^{\infty} x_i = \ell \in \mathbb{R}$ , then for any positive sequence  $b_n \nearrow \infty$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{b_n} \sum_{i=1}^n b_i x_i = 0. \quad (9.61)$$

This lemma is typically used as follows (consider the Marcinkiewicz-Zygmund SLLN with  $1 < q < 2$  for the sake of example). One first shows via the  $P$ -pointwise Kolmogorov three-series theorem that the sum

$$\sum_{k=1}^n \frac{X_k - \mathbb{E}_P(X)}{k^{1/q}} \quad (9.62)$$

is  $P$ -almost surely convergent as  $n \rightarrow \infty$ , at which point one applies Kronecker's lemma (on the same set of  $P$ -probability 1) to justify that

$$P \left( \lim_{n \rightarrow \infty} \frac{1}{n^{1/q}} \sum_{i=1}^n (X_i - \mathbb{E}_P(X)) = 0 \right) = 1. \quad (9.63)$$

However, it is not clear how Kronecker's lemma can be used to derive a  $\mathcal{P}$ -uniform analogue of (9.63) if (9.62) is only shown to be  $\mathcal{P}$ -uniformly Cauchy and especially if the limiting value  $\ell \equiv \ell(P)$  of (9.62) is a potentially random quantity whose behavior depends on the distribution  $P \in \mathcal{P}$  itself. Indeed, for the  $\mathcal{P}$ -uniform case, we introduce an additional uniform stochastic nonincreasingness condition given in Definition 9.1.3. Satisfying Definition 9.1.3 in pursuit of proving Theorems 9.2.1(i), 9.2.1(ii), and 9.2.2 requires additional care (the details of which can be found in Section 9.3.3), while this subtlety is easily sidestepped in the  $P$ -pointwise setting. Nevertheless, the following lemma serves as a stochastic and  $\mathcal{P}$ -uniform generalization of Kronecker's lemma that lends itself naturally to our goals and reduces to the usual  $P$ -almost sure application of Kronecker's lemma when  $\mathcal{P} = \{P\}$  is a singleton.

**Lemma 9.3.1** (A  $\mathcal{P}$ -uniform stochastic generalization of Kronecker's lemma). *Let  $(Z_n)_{n=1}^\infty$  be a sequence of random variables so that their partial sums  $S_n := \sum_{i=1}^n Z_i$  form a  $\mathcal{P}$ -uniform Cauchy sequence in the sense of Definition 9.1.2 and which is  $\mathcal{P}$ -uniformly stochastically nonincreasing in the sense of Definition 9.1.3. Let  $b_n \nearrow \infty$  be a positive, nondecreasing, and diverging sequence. Then,  $b_n^{-1} \sum_{i=1}^n b_i Z_i$  vanishes  $\mathcal{P}$ -uniformly in the sense of Definition 9.1.1.*

*Proof.* Fix any  $\varepsilon > 0$  and any  $\delta > 0$ . Our goal is to show that for all  $m$  sufficiently large,

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{b_k} \sum_{i=1}^k b_i Z_i \right| \geq \varepsilon \right) < 4\delta, \quad (9.64)$$

where the factor of 4 is only included for mathematical convenience later on. Using the assumption that  $S_n$  is  $\mathcal{P}$ -uniformly Cauchy and stochastically nonincreasing, let  $B > 0$  and choose  $N$  sufficiently large so that for any  $m \geq N$ ,

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k, n \geq m} |S_n - S_k| \geq \varepsilon/6 \right) < \delta, \quad (9.65)$$

and so that

$$\sup_{P \in \mathcal{P}} P(|S_m| \geq B) < \delta. \quad (9.66)$$

Again using stochastic nonincreasingness and the assumption that  $b_n \nearrow \infty$ , let  $N^* \equiv N^*(\varepsilon, B, N) \geq N$  be sufficiently large so that

$$\frac{\varepsilon b_{N^*}}{6b_N} \geq B, \quad (9.67)$$

and so that

$$\sup_{P \in \mathcal{P}} P \left( \sum_{i=1}^{N-1} |S_i| \geq \frac{\varepsilon b_{N^*}}{6b_N} \right) < \delta, \quad (9.68)$$

where we can impose the latter condition since  $\sum_{i=1}^{N-1} |S_i|$  is  $\mathcal{P}$ -uniformly bounded in probability for any fixed  $N$  and we can take  $\varepsilon b_{N^*}/6b_N$  to be arbitrarily large for any fixed  $N$  and  $\varepsilon$ . Then for all  $m \geq N^*$ ,

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{b_k} \sum_{i=1}^k b_i Z_i \right| \geq \varepsilon \right) \quad (9.69)$$

$$= \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| S_k - \frac{1}{b_k} \sum_{i=1}^{k-1} (b_{i+1} - b_i) S_i \right| \geq \varepsilon \right) \quad (9.70)$$

$$\leq \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| S_k - \frac{b_k - b_N}{b_k} S_m \right| \geq \varepsilon/3 \right) + \quad (9.71)$$

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{b_k} \sum_{i=1}^{N-1} (b_{i+1} - b_i) S_i \right| \geq \varepsilon/3 \right) + \quad (9.72)$$

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{b_k} \sum_{i=N}^{k-1} (b_{i+1} - b_i) (S_i - S_m) \right| \geq \varepsilon/3 \right), \quad (9.73)$$

where (9.70) follows from summation by parts and (9.71) follows from the triangle inequality. We will now bound the terms in (9.71), (9.72), and (9.73) separately.

**Bounding (9.71) by  $2\delta$ .** For any  $m \geq N^*$ , we have

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| S_k - \frac{b_k - b_N}{b_k} S_m \right| \geq \varepsilon/3 \right) \quad (9.74)$$

$$\leq \underbrace{\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} |S_k - S_m| \geq \varepsilon/6 \right)}_{<\delta} + \underbrace{\sup_{P \in \mathcal{P}} P \left( |S_m| \geq \frac{\varepsilon b_m}{6b_N} \right)}_{<\delta} < 2\delta. \quad (9.75)$$

where the last inequality follows from the conditions imposed on  $N^* \geq N$  in (9.65) and (9.66) combined with the fact that  $\varepsilon b_m/6b_N \geq B$  for all  $m \geq N^*$  as in (9.67).

**Bounding (9.72) by  $\delta$ .** For any  $m \geq N^*$ , we have

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{b_k} \sum_{i=1}^{N-1} (b_{i+1} - b_i) S_i \right| \geq \varepsilon/3 \right) \quad (9.76)$$

$$\leq \sup_{P \in \mathcal{P}} P \left( \sum_{i=1}^{N-1} |b_{i+1} - b_i| \cdot |S_i| \geq \varepsilon b_m/3 \right) \quad (9.77)$$

$$\leq \sup_{P \in \mathcal{P}} P \left( \sum_{i=1}^{N-1} |S_i| \geq \frac{\varepsilon b_m}{3b_N} \right) < \delta, \quad (9.78)$$

which follows from the condition imposed on  $N^*$  in (9.68).

**Bounding (9.73) by  $\delta$ .** For any  $m \geq N$ , we have

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{b_k} \sum_{i=N}^{k-1} (b_{i+1} - b_i)(S_i - S_m) \right| \geq \varepsilon/3 \right) \quad (9.79)$$

$$\leq \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \frac{1}{b_k} \sum_{i=N}^{k-1} (b_{i+1} - b_i)\varepsilon/6 \geq \varepsilon/3 \right) + \underbrace{\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq N} |S_k - S_m| \geq \varepsilon/6 \right)}_{< \delta} \quad (9.80)$$

$$< \underbrace{\sup_{P \in \mathcal{P}} \mathbf{1} \left\{ \sup_{k \geq m} \frac{b_k - b_N}{b_k} \geq 2 \right\}}_{=0} + \delta \quad (9.81)$$

which follows from the conditions imposed on  $N$  in (9.65) and the fact that  $\sup_{k \geq m} (b_k - b_N)/b_k \leq 1$ . Putting the bounds in (9.71), (9.72), and (9.73) together, we have that for any  $m \geq N^*$ ,

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{b_k} \sum_{i=1}^k b_i Z_i \right| \geq \varepsilon \right) < 4\delta, \quad (9.82)$$

which yields the desired result, completing the proof.  $\square$

### 9.3.5 Distribution-uniform Borel-Cantelli lemmas

In order to show that  $\mathcal{P}$ -UI of certain finite absolute moments is in fact *necessary* for the  $\mathcal{P}$ -uniform SLLN to hold — i.e. the result of Theorem 9.2.1(iii) — we rely on a  $\mathcal{P}$ -uniform generalization of the *second* Borel-Cantelli lemma. Before discussing the second Borel-Cantelli lemma, let us briefly discuss the first. A natural desideratum for a  $\mathcal{P}$ -uniform first Borel-Cantelli lemma would be to say that for events  $(E_n)_{n=1}^\infty$  in  $\mathcal{F}$ , if  $\lim_m \sup_{P \in \mathcal{P}} \sum_{k=m}^\infty P(E_n) = 0$ , then  $\lim_m \sup_{P \in \mathcal{P}} P(\bigcup_{k=m}^\infty E_k) = 0$ . Indeed, this is trivially satisfied since

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \bigcup_{k=m}^\infty E_k \right) \leq \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^\infty P(E_k). \quad (9.83)$$

For this reason, we do not dwell on the first Borel-Cantelli lemma, but instead shift our attention to the second since its  $\mathcal{P}$ -uniform generalization (and the proof thereof) is nontrivial in comparison and is central to the proof of Theorem 9.2.1(iii).

**Lemma 9.3.2** (The second  $\mathcal{P}$ -uniform Borel-Cantelli lemma). *Let  $(E_n)_{n=1}^\infty$  be independent*

events such that

$$0 < \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P(E_k) \leq \infty. \quad (9.84)$$

Then the probability of infinitely many of them occurring does not  $\mathcal{P}$ -uniformly vanish, i.e.

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\bigcup_{k=m}^{\infty} E_k\right) > 0. \quad (9.85)$$

*Proof.* The proof proceeds by a direct calculation. Writing out the limit in (9.85), we have

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P\left(\bigcup_{k=m}^{\infty} E_k\right) = \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \left\{ 1 - P\left(\bigcap_{k=m}^{\infty} E_k^c\right) \right\} \quad (9.86)$$

$$= 1 - \lim_{m \rightarrow \infty} \inf_{P \in \mathcal{P}} \lim_{t \rightarrow \infty} P\left(\bigcap_{k=m}^t E_k^c\right) \quad (9.87)$$

$$= 1 - \lim_{m \rightarrow \infty} \inf_{P \in \mathcal{P}} \lim_{t \rightarrow \infty} \prod_{k=m}^t (1 - P(E_k)) \quad (9.88)$$

$$= 1 - \exp \left\{ - \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P(E_k) \right\} > 0, \quad (9.89)$$

where (9.87) follows from the fact that the intersections  $(\bigcap_k^t E_k)_{t=1}^{\infty}$  are nested, (9.88) exploits independence of  $(E_n)_{n=1}^{\infty}$ , and (9.89) follows from the assumption in (9.84). This completes the proof.  $\square$

## 9.4 Proof details for Theorems 9.2.1 and 9.2.2

With the introduction of the distribution-uniform analogues of Kolmogorov's three-series theorem (Theorem 9.3.2), Kronecker's lemma (Lemma 9.3.1), and the second Borel-Cantelli lemma (Lemma 9.3.2), we are ready to complete the proof details for our main results in Theorems 9.2.1 and 9.2.2.

### 9.4.1 Proof details for Theorem 9.2.1

*Proof details for Theorem 9.2.1(i).* Given the proof outline following Theorem 9.2.1, it only remains to show that the  $\mathcal{P}$ -UI condition

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P(|X - \mathbb{E}_P(X)|^q \cdot \mathbf{1}\{|X - \mathbb{E}_P(X)|^q > m\}) = 0 \quad (9.90)$$

is sufficient to satisfy the conditions of both the  $\mathcal{P}$ -uniform Kolmogorov three series theorem (Theorem 9.3.2) and those of the  $\mathcal{P}$ -uniform nonincreasingness theorem (Theorem 9.3.3) for appropriately truncated and scaled versions of  $(X_n)_{n=1}^{\infty}$ , ultimately showing that  $S_n := \sum_{i=1}^n X_i$

is both  $\mathcal{P}$ -uniformly Cauchy and stochastically nonincreasing. These conditions are shown in Lemmas 9.4.1 and 9.4.10, respectively.  $\square$

*Proof details for Theorem 9.2.1(ii).* Similar to the proof of Theorem 9.2.1(i), it suffices to show that the *uncentered*  $\mathcal{P}$ -UI condition

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|X|^q \mathbf{1}\{|X|^q > m\}) = 0 \quad (9.91)$$

is sufficient to satisfy the conditions of Theorems 9.3.2 and 9.3.3 for appropriately truncated and scaled versions of  $(X_n)_{n=1}^\infty$ , the details of which are provided in Lemmas 9.4.2 and 9.4.7.  $\square$

*Proof details for Theorem 9.2.1(iii).* Suppose that  $\mathcal{P}$  is a class of distributions for which the  $\mathcal{P}$ -UI condition does not hold, i.e.

$$0 < \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|X - \mu(P; q)|^q \mathbf{1}\{|X - \mu(P; q)|^q > m\}) \leq \infty, \quad (9.92)$$

recalling that  $\mu(P; q) = \mathbb{E}_P(X)$  when  $1 \leq q < 2$  and  $\mu(P; q) = 0$  when  $0 < q < 1$ . Then we will show that

$$\text{Goal: } \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left| \frac{1}{k^{1/q}} \sum_{i=1}^k (X_i - \mu(P; q)) \right| \geq \frac{1}{2} \right) > 0. \quad (9.93)$$

Indeed, pre-multiplying the above probability by 2 for any  $P \in \mathcal{P}$  and any  $m \geq 1$ , consider the partial sums  $S_n := \sum_{i=1}^n (X_i - \mu(P; q))$  and note that

$$2P \left( \sup_{k \geq m-1} \frac{1}{k^{1/q}} |S_k| \geq 1/2 \right) \geq P \left( \sup_{k \geq m} \frac{1}{k^{1/q}} |S_k| \geq 1/2 \right) + P \left( \sup_{k \geq m-1} \frac{1}{k^{1/q}} |S_k| \geq 1/2 \right) \quad (9.94)$$

$$\geq P \left( \sup_{k \geq m} \frac{1}{k^{1/q}} (|S_k| + |S_{k-1}|) \geq 1 \right) \quad (9.95)$$

$$\geq P \left( \sup_{k \geq m} \frac{1}{k^{1/q}} |X_k - \mu(P; q)| \geq 1 \right) \quad (9.96)$$

and hence by the  $\mathcal{P}$ -uniform second Borel-Cantelli lemma (Lemma 9.3.2), it suffices to show that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P(|X - \mu(P; q)| > k^{1/q}) > 0, \quad (9.97)$$

from which we will obtain (9.93). Indeed, by Hu and Zhou [127, Theorem 2.1] — or as shown directly in Lemma 9.4.8 — we have that for any random variable  $Y$ ,

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P(|Y| > k) = 0 \quad \text{if and only if} \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y| \mathbf{1}\{|Y| > m\}) = 0, \quad (9.98)$$

and hence if the  $q^{\text{th}}$  moment is not  $\mathcal{P}$ -UI, we must have that (9.97) holds. This completes the proof.  $\square$

#### 9.4.1.1 Sufficient conditions for the $\mathcal{P}$ -uniform Kolmogorov three-series theorem

In what follows, we verify that the tails of the three series of Theorem 9.3.2 vanish  $\mathcal{P}$ -uniformly when  $X$  has a  $\mathcal{P}$ -UI  $q^{\text{th}}$  moment, thereby enabling the application of the  $\mathcal{P}$ -uniform Kolmogorov three-series theorem. Lemmas 9.4.1 and 9.4.2 consider the cases of  $1 < q < 2$  and  $0 < q < 1$ , respectively.

**Lemma 9.4.1** (Sufficient conditions in the identically distributed case when  $1 < q < 2$ ). *Let  $(X_n)_{n=1}^\infty$  be i.i.d. random variables on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  and let  $Y_n := X_n - \mathbb{E}_P X$  be their centered versions for each  $P \in \mathcal{P}$ . Suppose that the  $q^{\text{th}}$  moment is UI for some  $1 < q < 2$ . Then, the three conditions of the  $\mathcal{P}$ -uniform Kolmogorov three-series theorem are satisfied for  $Z_n := Y_n/n^{1/q}$  with  $c = 1$ .*

*Proof.* Throughout the proofs for the three series, consider the truncated random variable  $Z_n^{\leq 1}$  given by

$$Z_n^{\leq 1} := Z_n \mathbf{1}\{Z_n \leq 1\}. \quad (9.99)$$

Let us now separately show that the tails of the three series vanish  $\mathcal{P}$ -uniformly.

**The first series.** Writing out the first series  $\sup_{P \in \mathcal{P}} \sum_{n=m}^\infty |\mathbb{E}_P Z_n^{\leq 1}|$  for any  $m \geq 1$ , we have

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^\infty |\mathbb{E}_P Z_n^{\leq 1}| = \sup_{P \in \mathcal{P}} \sum_{n=m}^\infty \left| \mathbb{E}_P \left( \frac{Y \mathbf{1}(|Y| \leq n^{1/q})}{n^{1/q}} \right) \right| \quad (9.100)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{n=m}^\infty \mathbb{E}_P \left( \frac{|Y| \mathbf{1}(|Y| > n^{1/q})}{n^{1/q}} \right) \quad (9.101)$$

$$= \sup_{P \in \mathcal{P}} \sum_{n=m}^\infty \sum_{k=n}^\infty \mathbb{E}_P \left( \frac{|Y| \mathbf{1}(k^{1/q} < |Y| \leq (k+1)^{1/q})}{n^{1/q}} \right) \quad (9.102)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{k=m}^\infty \mathbb{E}_P \left( |Y| \mathbf{1}(k^{1/q} < |Y| \leq (k+1)^{1/q}) \right) \cdot \sum_{n=1}^k \frac{1}{n^{1/q}}. \quad (9.103)$$

Now, there exists some constant  $C_q > 0$  depending only on  $q$  so that  $\sum_{n=1}^k 1/n^{1/q} \leq C_q k/(k+1)^{1/q}$ , thus

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^\infty |\mathbb{E}_P Z_n^{\leq 1}| \leq \sup_{P \in \mathcal{P}} \sum_{k=m}^\infty \mathbb{E}_P \left( |Y| \mathbf{1}(k^{1/q} < |Y| \leq (k+1)^{1/q}) \right) \cdot C_q \frac{k}{(k+1)^{1/q}} \quad (9.104)$$

$$\leq C_q \sup_{P \in \mathcal{P}} \sum_{k=m}^\infty k P \left( k^{1/q} < |Y| \leq (k+1)^{1/q} \right) \quad (9.105)$$

$$= C_q \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} P(|Y|^q > (m \vee n)) \quad (9.106)$$

$$\leq C_q \sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q \mathbf{1}\{|Y|^q > m\}), \quad (9.107)$$

where the last inequality follows from the fact that  $|Y|^q > (m \vee n)$  if and only if  $\{|Y|^q > m \vee n\} > (m \vee n)$  and using the expectation-tail-probability identity. Taking limits as  $m \rightarrow \infty$  completes the argument for the first series.

**The second series.** Writing out the second series  $\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \text{Var}_P(Z_n^{\leq 1})$  for any  $m \geq 1$  and performing a direct calculation, we have

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \text{Var}_P Z_n^{\leq 1} \leq \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \mathbb{E}_P[(Z_n^{\leq 1})^2] \quad (9.108)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \mathbf{1}\{n \geq m\} \cdot \mathbb{E}_P \left[ \frac{Y^2}{n^{2/q}} \mathbf{1}\{|Y| \leq n^{1/q}\} \right] \quad (9.109)$$

$$= \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \mathbb{E}_P \left[ Y^2 \mathbf{1}\{(k-1)^{1/q} < |Y| \leq k^{1/q}\} \right] \sum_{n=k \vee m}^{\infty} \frac{1}{n^{2/q}}. \quad (9.110)$$

Now, there exists a constant  $C_q > 0$  depending only on  $q$  so that  $\sum_{n=k \vee m}^{\infty} 1/n^{2/q} \leq (k \vee m)/(k \vee m)^{2/q}$ , and hence we have

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \text{Var}_P Z_n^{\leq 1} \leq C_q \underbrace{\left( \sup_{P \in \mathcal{P}} \sum_{k=1}^m \frac{m}{m^{2/q}} \cdot \mathbb{E}_P \left[ Y^2 \mathbf{1}\{(k-1)^{1/q} < |Y| \leq k^{1/q}\} \right] \right)}_{(\star_{\leq})} + \quad (9.111)$$

$$\underbrace{\sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{k}{k^{2/q}} \mathbb{E}_P \left[ Y^2 \mathbf{1}\{(k-1)^{1/q} < |Y| \leq k^{1/q}\} \right]}_{(\star_{\geq})}, \quad (9.112)$$

and we will now separately show that  $(\star_{\leq}) \rightarrow 0$  and  $(\star_{\geq}) \rightarrow 0$  as  $m \rightarrow \infty$  using different arguments. Focusing first on  $(\star_{\leq})$ , we invoke Lemma 9.4.3 with  $p = 2$  and let  $\tilde{\varphi}(x) \equiv x\tilde{h}(x)$ ;  $x \geq 0$  be a function where  $\tilde{h}$  has the following three key properties: (1) it is diverging to  $\infty$ , (2) it satisfies  $a^{2/q}/b^{2/q} \leq a\tilde{h}(a)/(b\tilde{h}(b)) = \tilde{\varphi}(a)/\tilde{\varphi}(b)$  whenever  $0 \leq a \leq b$  and  $b > 0$ , and (3) it satisfies  $\sup_{P \in \mathcal{P}} \mathbb{E}_P [|Y|^q h(|Y|^q)] < \infty$ . Writing out  $(\star_{\leq})$  and exploiting these three properties, we have

$$(\star_{\leq}) = m \cdot \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \frac{Y^2}{m^{2/q}} \mathbf{1}\{|Y| \leq m^{1/q}\} \right] \quad (9.113)$$

$$\leq m \cdot \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \frac{\tilde{\varphi}(|Y|^q)}{\tilde{\varphi}(m)} \mathbf{1}\{|Y| \leq m^{1/q}\} \right] \quad (9.114)$$

$$\leq \underbrace{\frac{1}{\tilde{h}(m)}}_{\rightarrow 0} \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_P \tilde{\varphi}(|Y|^q)}_{<\infty}, \quad (9.115)$$

where (9.114) follows from property (2) of  $\tilde{h}$  described above, and the last uses properties (1) and (3). Therefore,  $\lim_{m \rightarrow \infty} (\star_{\leq}) = 0$ . Turning our focus to  $(\star_{\geq})$ ,

$$(\star_{\geq}) \leq \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} k P \left( (k-1)^{1/q} < |Y| \leq k^{1/q} \right), \quad (9.116)$$

and the above vanishes as  $m \rightarrow \infty$  by the proof for the first series beginning from (9.105). Putting the analyses for  $(\star_{\leq})$  and  $(\star_{\geq})$  together and taking limits as  $m \rightarrow \infty$  completes the argument for the second series.

**The third series.** Once again using the fact that  $|Y|^q > k$  if and only if  $|Y|^q \mathbf{1}\{|Y|^q > k\} > k$  and writing out the third series for any  $m \geq 1$ , we have

$$\sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P \left( \left| \frac{Y}{k^{1/q}} \right| > 1 \right) = \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \mathbf{1}(k \geq m) P(|Y|^q \mathbf{1}\{|Y|^q > k\} > k) \quad (9.117)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \mathbf{1}(k \geq m) P(|Y|^q \mathbf{1}\{|Y|^q > m\} > k) \quad (9.118)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q \mathbf{1}\{|Y|^q > m\}), \quad (9.119)$$

which used the fact that  $P(|Y|^q \mathbf{1}\{|Y|^q > k\} > k) \leq P(|Y|^q \mathbf{1}\{|Y|^q > m\} > k)$  whenever  $k \geq m$  combined with the expectation-tail-probability identity. Taking limits as  $m \rightarrow \infty$  completes the argument for the third series, concluding the proof of Lemma 9.4.1.  $\square$

**Lemma 9.4.2** (Sufficient conditions for the three series with  $0 < q < 1$ ). *Given the same setup as Lemma 9.4.1, suppose that  $X$  has a  $\mathcal{P}$ -UI (uncentered)  $q^{\text{th}}$  moment for some  $0 < q < 1$ :*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|X|^q \mathbf{1}\{|X|^q > m\}) = 0. \quad (9.120)$$

*Then, the three conditions of the  $\mathcal{P}$ -uniform Kolmogorov three-series theorem are satisfied for  $Z_n := X_n/n^{1/q}$  with  $c = 1$ .*

*Proof.* Once again, consider the truncated random variable  $Z_n^{\leq 1}$  given by

$$Z_n^{\leq 1} := Z_n \mathbf{1}\{Z_n \leq 1\}. \quad (9.121)$$

Let us now separately show that the tails of the three series vanish  $\mathcal{P}$ -uniformly.

**The first series.** Writing out the first series for any  $m \geq 1$  and using a similar argument to that of the second series in Lemma 9.4.1, we have that

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} |\mathbb{E}_P Z_n^{\leq 1}| \leq \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \mathbb{E}_P \left( \frac{|X| \mathbf{1}(|X| \leq n^{1/q})}{n^{1/q}} \right) \quad (9.122)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \mathbb{E}_P \left( |X| \mathbf{1}\{(k-1)^{1/q} < |X| \leq k^{1/q}\} \right) \cdot \sum_{n=k \vee m}^{\infty} \frac{1}{n^{1/q}}, \quad (9.123)$$

and thus there exists  $C_q > 0$  depending only on  $q \in (0, 1)$  so that

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \mathbb{E}_P Z_n^{\leq 1} \leq C_q \underbrace{\sup_{P \in \mathcal{P}} \sum_{k=1}^m \frac{m}{m^{1/q}} \cdot \mathbb{E}_P \left( |X| \mathbf{1}\{(k-1)^{1/q} < |X| \leq k^{1/q}\} \right)}_{(\star \leq)} + \quad (9.124)$$

$$C_q \underbrace{\sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{k}{k^{1/q}} \cdot \mathbb{E}_P \left( |X| \mathbf{1}\{(k-1)^{1/q} < |X| \leq k^{1/q}\} \right)}_{(\star \geq)}, \quad (9.125)$$

and thus similarly to the proof for the second series in Lemma 9.4.1, it suffices to show that  $(\star \leq) \rightarrow 0$  and  $(\star \geq) \rightarrow 0$  as  $m \rightarrow \infty$ . Focusing on  $(\star \leq)$  first, we have

$$(\star \leq) \leq \sup_{P \in \mathcal{P}} m \cdot \mathbb{E}_P \left( \frac{|X|}{m^{1/q}} \mathbf{1}\{|X| \leq m^{1/q}\} \right). \quad (9.126)$$

By Lemma 9.4.3 applied with  $p = 1$ , there exists a function  $\tilde{\varphi}(x) = x\tilde{h}(x)$ ;  $x \geq 0$  where  $\tilde{h} > 0$  is a function that (1) is diverging to  $\infty$ , (2) satisfies  $a^{1/q}/b^{1/q} \leq a\tilde{h}(a)/(b\tilde{h}(b)) \equiv \tilde{\varphi}(a)/\tilde{\varphi}(b)$  whenever  $0 \leq a \leq b$  and  $b > 0$ , and (3) satisfies  $\sup_{P \in \mathcal{P}} \mathbb{E}_P \tilde{\varphi}(|X|^q) < \infty$ . Writing out  $(\star \leq)$  and exploiting properties (1)–(3), we have

$$(\star \leq) \leq \sup_{P \in \mathcal{P}} m \cdot \mathbb{E}_P \left( \frac{|X|}{m^{1/q}} \mathbf{1}\{|X| \leq m^{1/q}\} \right) \quad (9.127)$$

$$\leq \sup_{P \in \mathcal{P}} m \cdot \mathbb{E}_P \left( \frac{\tilde{\varphi}(|X|^q)}{\tilde{\varphi}(m)} \right) \quad (9.128)$$

$$= \frac{\mathcal{M}}{p\tilde{h}(m)} \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_P \tilde{\varphi}(|X|^q)}_{<\infty}, \quad (9.129)$$

and since  $\lim_{m \rightarrow \infty} \tilde{h}(m) = \infty$ , we have that  $\lim_{m \rightarrow \infty} (\star \leq) = 0$ . Moving to  $(\star \geq)$ , we have that

$$(\star \geq) = \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{k}{k^{1/q}} \cdot \mathbb{E}_P \left( |X| \mathbf{1}\{(k-1)^{1/q} < |X| \leq k^{1/q}\} \right) \quad (9.130)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} k \cdot P \left( (k-1)^{1/q} < |X| \leq k^{1/q} \right) \quad (9.131)$$

$$\leq \sup_{P \in \mathcal{P}} P(|X|^q \mathbb{1}\{|X|^q > m\} > m) + \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} P(|X|^q \mathbb{1}\{|X|^q > n \vee m\} > n \vee m) \quad (9.132)$$

$$\leq \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_P |X|^q / m}_{<\infty} + \sup_{P \in \mathcal{P}} (\mathbb{E}_P (|X|^q \mathbb{1}\{|X|^q > m\})), \quad (9.133)$$

and thus  $\lim_{m \rightarrow \infty} (\star_{\geq}) = 0$ . Putting  $(\star_{\leq})$  and  $(\star_{\geq})$  together and taking limits as  $m \rightarrow \infty$  completes the argument for the first series.

**The second series.** The second series proceeds similarly to that of Lemma 9.4.1. Indeed, using the same arguments therein, notice that there exists a constant  $C_q > 0$  depending only on  $q \in (0, 1)$  so that

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \text{Var}_P Z_n^{\leq 1} \leq C_q \left\{ \underbrace{\sup_{P \in \mathcal{P}} \sum_{k=1}^m \frac{m}{m^{2/q}} \cdot \mathbb{E}_P [Y^2 \mathbb{1}\{(k-1)^{1/q} < |Y| \leq k^{1/q}\}]}_{(\dagger_{\leq})} + \right. \quad (9.134)$$

$$\left. \underbrace{\sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{k}{k^{2/q}} \mathbb{E}_P [Y^2 \mathbb{1}\{(k-1)^{1/q} < |Y| \leq k^{1/q}\}]}_{(\dagger_{\geq})} \right\}, \quad (9.135)$$

Focusing first on  $(\dagger_{\leq})$ , since  $a/b \leq (a/b)^2$  whenever  $0 \leq a \leq b$  and  $b > 0$ , we have that

$$(\dagger_{\leq}) = m \cdot \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \frac{Y^2}{m^{2/q}} \mathbb{1}\{|Y| \leq m^{1/q}\} \right] \quad (9.136)$$

$$\leq m \cdot \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \frac{|Y|}{m^{1/q}} \mathbb{1}\{|Y| \leq m^{1/q}\} \right], \quad (9.137)$$

and now by the same argument used to show that  $(\star_{\leq}) \rightarrow 0$  in the first series above, we have that  $(\dagger_{\leq}) \rightarrow 0$  as  $m \rightarrow 0$ . Turning to  $(\dagger_{\geq})$ , we have by the same argument as in the second series of Lemma 9.4.1 (and in particular, starting from (9.105)) that

$$(\dagger_{\geq}) \leq \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} k P \left( (k-1)^{1/q} < |Y| \leq k^{1/q} \right) \quad (9.138)$$

$$\leq \sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q \mathbb{1}\{|Y|^q > m\}), \quad (9.139)$$

and hence by the  $\mathcal{P}$ -UI of the  $q^{\text{th}}$  moment, we have that  $(\dagger_{\geq}) \rightarrow 0$  as  $m \rightarrow \infty$ . Putting the limits for  $(\dagger_{\leq})$  and  $(\dagger_{\geq})$  together completes the argument for the second series.

**The third series.** The proof for the series proceeds identically to that of Lemma 9.4.1 when  $1 < q < 2$ . This completes the proof of Lemma 9.4.2.  $\square$

**Lemma 9.4.3.** Let  $0 < q < p \leq 2$  and suppose that the (potentially uncentered)  $q^{\text{th}}$  moment of  $Y$  is UI, meaning that

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q \mathbf{1}(|Y|^q > m)) = 0. \quad (9.140)$$

Then there exists a function  $\tilde{\varphi}$  that can be written as  $\tilde{\varphi}(x) = x\tilde{h}(x)$  for any  $x \geq 0$  where  $\tilde{h}$  is diverging to  $\infty$  and satisfies

$$\frac{a^{p/q}}{b^{p/q}} \leq \frac{a\tilde{h}(a)}{b\tilde{h}(b)} \quad \text{for all } 0 \leq a \leq b \text{ and } b > 0, \quad (9.141)$$

and so that  $\tilde{\varphi}(|Y|^q)$  is  $\mathcal{P}$ -uniformly bounded in expectation:

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \tilde{\varphi}(|Y|^q) < \infty. \quad (9.142)$$

*Proof.* By the criterion of uniform integrability due to Charles de la Vallée Poussin [58, 126, 52], we have that there exists a function  $h : [0, \infty) \rightarrow [0, \infty)$  diverging to  $\infty$  so that

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P [|Y|^q h(|Y|^q)] < \infty. \quad (9.143)$$

Moreover,  $h$  can be assumed to be concave, strictly increasing, and starting at  $h(0) = 0$  (see Proposition 9.A.1). Let  $h^*(x) := h(x) + 1$  for each  $x$ , and use  $h^*$  in Lemma 9.4.4 obtain a function  $\tilde{h}$  that concave, strictly increasing, and starting at  $\tilde{h}(0) \geq 1$ , and in addition,  $\tilde{h}(x) \leq h(x) + 1$  for all  $x \geq 0$  so that

$$\frac{a^{p/q}}{b^{p/q}} \leq \frac{a\tilde{h}(a)}{b\tilde{h}(b)} \quad (9.144)$$

whenever  $0 \leq a \leq b$  and  $b > 0$ . Noticing that since  $\tilde{h} \leq h + 1$ ,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P [|Y|^q \tilde{h}(|Y|^q)] \leq \sup_{P \in \mathcal{P}} \mathbb{E}_P [|Y|^q] + \sup_{P \in \mathcal{P}} \mathbb{E}_P [|Y|^q h(|Y|^q)] < \infty \quad (9.145)$$

completes the proof.  $\square$

**Lemma 9.4.4.** Let  $h^* : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}^{\geq 0}$  be a function that is concave, strictly increasing, diverging to  $\infty$ , and beginning at  $h^*(0) \geq 1$ . Let  $1 < q < p \leq 2$ . Then there exists a function  $\tilde{h}$  (depending on  $p$  and  $q$ ) with all the aforementioned properties and in addition,  $\tilde{h}(x) \leq h^*(x)$  for all  $x \geq 0$  so that for any real  $a \geq 0$  and  $b > 0$  such that  $a \leq b$ ,

$$\frac{a^{p/q}}{b^{p/q}} \leq \frac{a\tilde{h}(a)}{b\tilde{h}(b)}. \quad (9.146)$$

*Proof.* Throughout, denote  $\delta := p/q - 1 \in (0, 1)$ . Choose  $\tilde{h}(x) := (h^*(x))^\delta$  and notice that since  $\delta \in (0, 1)$  and  $h^*(x) \geq 1$ , we have that  $\tilde{h}$  is also concave, strictly increasing, diverging to  $\infty$ , and beginning at  $\tilde{h}(0) \geq 1$ . Clearly the result of the lemma holds for this choice of  $\tilde{h}$  when  $a = 0$  since  $\tilde{h}(0) \geq 1$  and  $\tilde{h}$  is strictly increasing so let us focus on the case where  $0 < a \leq b$ . Showing the desired result is equivalent to showing that

$$a^{-\delta} \tilde{h}(a) \geq b^{-\delta} \tilde{h}(b) \quad \text{for all } a \leq b \quad (9.147)$$

or in other words, that  $x^{-\delta} \tilde{h}(x)$  is nonincreasing on  $x > 0$ . By Lemma 9.4.5, we have that  $x^{-1} h(x)$  is nonincreasing on  $x > 0$ , and hence so is  $x^{-\delta} \tilde{h}(x) \equiv (x^{-1} h(x))^\delta$ . This completes the proof.  $\square$

**Lemma 9.4.5.** *Let  $h : \mathbb{R}^{>0} \rightarrow \mathbb{R}^{>0}$  be a function that is concave and beginning at  $h(0) \geq 0$ . Then, the function  $h(x)/x$  is nonincreasing on  $x > 0$ .*

*Proof.* Let  $0 < a < b$ . Our goal is to show that  $h(a)/a \geq h(b)/b$ . Indeed, notice that by concavity of  $h$  and the fact that  $h(0) \geq 0$ , we have that

$$\frac{h(b) - h(a)}{b - a} \leq \frac{h(b) - h(0)}{b - 0} \leq \frac{h(b)}{b}. \quad (9.148)$$

Rearranging the terms above, we easily obtain the desired result.  $\square$

**Lemma 9.4.6** (Satisfying the three series for stochastic nonincreasingness when  $1 < q < 2$ ). *Suppose that  $(X_n)_{n=1}^\infty$  are i.i.d. and have a  $\mathcal{P}$ -UI  $q^{th}$  moment for  $1 < q < 2$ . Then the three series for uniform stochastic nonincreasingness in Theorem 9.3.3 are satisfied for the random variable  $Z_n := (X_n - \mathbb{E}_P(X))/n^{1/q}$ .*

*Proof.* We will handle the three series separately below. Throughout, let  $Y_n := X_n - \mathbb{E}_P(X)$  and  $Z_{n,B}^{\leq 1} := (Y/B)\mathbf{1}\{|(Y/B)| \leq n^{1/q}\}/n^{1/q}$ .

**The first series.** Writing out the first series  $\sup_{P \in \mathcal{P}} \sum_{n=1}^\infty |\mathbb{E}_P(Z_{n,B}^{\leq 1})|$  and performing calculations analogous to those for the first series in Lemma 9.4.1, we have

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^\infty |\mathbb{E}_P Z_{n,B}^{\leq 1}| = \sup_{P \in \mathcal{P}} \sum_{n=1}^\infty \left| \mathbb{E}_P \left( \frac{(Y/B)\mathbf{1}\{|(Y/B)| \leq n^{1/q}\}}{n^{1/q}} \right) \right| \quad (9.149)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{k=1}^\infty \mathbb{E}_P \left( |Y/B| \mathbf{1}(k^{1/q} < |Y/B| \leq (k+1)^{1/q}) \right) \cdot \sum_{n=1}^k \frac{1}{n^{1/q}}. \quad (9.150)$$

Now, there exists some constant  $C_q > 0$  depending only on  $q$  so that  $\sum_{n=1}^k 1/n^{1/q} \leq C_q k/(k+1)^{1/q}$  and thus

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} |\mathbb{E}_P Z_{n,B}^{\leq 1}| \leq \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \mathbb{E}_P \left( |Y/B| \mathbf{1}(k^{1/q} < |Y/B| \leq (k+1)^{1/q}) \right) \cdot C_q \frac{k}{(k+1)^{1/q}} \quad (9.151)$$

$$\leq C_q \cdot \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} k P \left( k^{1/q} < |Y/B| \leq (k+1)^{1/q} \right) \quad (9.152)$$

$$= C_q \cdot \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} P(|Y/B|^q > n) \quad (9.153)$$

$$\leq C_q \cdot B^{-q} \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q)}_{<\infty}. \quad (9.154)$$

Taking limits as  $B \rightarrow \infty$ , we have the desired result:

$$\lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} |\mathbb{E}_P Z_{n,B}^{\leq 1}| = 0, \quad (9.155)$$

completing the proof for the first series for  $\mathcal{P}$ -uniform stochastic nonincreasingness.

**The second series.** Writing out the second series  $\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \text{Var}_P(Z_{n,B}^{\leq 1})$  for any  $B > 0$ , we have

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \text{Var}_P Z_{n,B}^{\leq 1} = \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \left\{ \mathbb{E}_P[(Z_{n,B}^{\leq 1})^2] - [\mathbb{E}_P Z_{n,B}^{\leq 1}]^2 \right\} \quad (9.156)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \mathbb{E}_P \left[ \frac{(Y/B)^2}{n^{2/q}} \mathbf{1}\{|Y/B| \leq n^{1/q}\} \right] \quad (9.157)$$

$$= \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \mathbb{E}_P \left[ (Y/B)^2 \mathbf{1}\{(k-1)^{1/q} < |Y/B| \leq k^{1/q}\} \right] \sum_{n=k}^{\infty} \frac{1}{n^{2/q}}, \quad (9.158)$$

and since there exists a constant  $C_q > 0$  depending only on  $q$  so that  $\sum_{n=k}^{\infty} 1/n^{2/q} \leq C_q k/k^{2/q}$ , we have

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \text{Var}_P Z_{n,B}^{\leq 1} \leq C_q \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \frac{k}{k^{2/q}} \cdot \mathbb{E}_P \left[ (Y/B)^2 \mathbf{1}\{(k-1)^{1/q} < |Y/B| \leq k^{1/q}\} \right]. \quad (9.159)$$

Separating the first term from the rest of the sum, we can write the above as

$$C_q \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \frac{k}{k^{2/q}} \cdot \mathbb{E}_P \left[ (Y/B)^2 \mathbb{1}\{(k-1)^{1/q} < |Y/B| \leq k^{1/q}\} \right] \quad (9.160)$$

$$\leq \underbrace{C_q \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ (Y/B)^2 \mathbb{1}\{|Y/B| \leq 1\} \right]}_{(\star)} + \quad (9.161)$$

$$\underbrace{C_q \sup_{P \in \mathcal{P}} \sum_{k=2}^{\infty} \frac{k}{k^{2/q}} \cdot \mathbb{E}_P \left[ (Y/B)^2 \mathbb{1}\{(k-1)^{1/q} \leq |Y/B| \leq k^{1/q}\} \right]}_{(\dagger)}. \quad (9.162)$$

Notice that  $(Y/B)^2 \mathbb{1}\{|Y/B| \leq 1\} \leq |Y/B|^q \mathbb{1}\{|Y/B| \leq 1\}$  with  $P$ -probability one for every  $P \in \mathcal{P}$  since  $0 < q < 2$ . Therefore, the first term  $(\star)$  of the aforementioned sum can be upper-bounded as

$$(\star) = C_q \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ (Y/B)^2 \mathbb{1}\{|Y/B| \leq 1\} \right] \leq \frac{C_q}{B^q} \sup_{P \in \mathcal{P}} \mathbb{E}_P |Y|^q. \quad (9.163)$$

Turning now to the second term  $(\dagger)$ , we have

$$(\dagger) = C_q \sup_{P \in \mathcal{P}} \sum_{k=2}^{\infty} \frac{k}{k^{2/q}} \cdot \mathbb{E}_P \left[ (Y/B)^2 \mathbb{1}\{(k-1)^{1/q} < |Y/B| \leq k^{1/q}\} \right] \quad (9.164)$$

$$\leq C_q \sup_{P \in \mathcal{P}} \sum_{k=2}^{\infty} \frac{k}{k^{2/q}} \mathbb{E}_P \left[ k^{2/q} \mathbb{1}\{(k-1)^{1/q} < |Y/B| \leq k^{1/q}\} \right] \quad (9.165)$$

$$= C_q \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} \mathbb{1}\{k \geq 2\} P \left( (k-1)^{1/q} < |Y/B| \leq k^{1/q} \right) \quad (9.166)$$

$$= C_q \sup_{P \in \mathcal{P}} \left[ \sum_{k=2}^{\infty} P \left( (k-1)^{1/q} < |Y/B| \leq k^{1/q} \right) + \sum_{n=2}^{\infty} \sum_{k=n}^{\infty} P \left( (k-1)^{1/q} < |Y/B| \leq k^{1/q} \right) \right] \quad (9.167)$$

$$\leq C_q \sup_{P \in \mathcal{P}} \left[ 2 \sum_{n=2}^{\infty} \sum_{k=n}^{\infty} P \left( (k-1)^{1/q} < |Y/B| \leq k^{1/q} \right) \right] \quad (9.168)$$

$$\leq 2C_q \sup_{P \in \mathcal{P}} \sum_{n=2}^{\infty} P \left( |Y/B| > (n-1)^{1/q} \right) \quad (9.169)$$

$$\leq \frac{2C_q}{B^q} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q). \quad (9.170)$$

Putting the bounds on  $(\star)$  and  $(\dagger)$  together, we have

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \text{Var}_P Z_{n,B}^{\leq 1} \leq (\star) + (\dagger) \leq \frac{3C_q}{B^q} \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_P |Y|^q}_{<\infty}, \quad (9.171)$$

so that when we take limits as  $B \rightarrow \infty$ , we obtain the desired result,

$$\lim_{B \rightarrow \infty} \sum_{n=1}^{\infty} \text{Var}_P Z_{n,B}^{\leq 1} = 0, \quad (9.172)$$

which completes the proof for the second series.

**The third series.** Writing out the series  $\sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} P(|Y/k^{1/q}| > B)$  for any  $B > 0$  and using the expectation-tail sum identity, we have

$$\sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} P(|Y/k^{1/q}| > B) = \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} P(|Y/B|^q > k) \leq \frac{1}{B^q} \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q)}_{<\infty}, \quad (9.173)$$

and we note that  $\sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q) < \infty$  above since uniform integrability implies uniform boundedness. Consequently, we have that

$$\lim_{B \rightarrow \infty} \sum_{k=1}^{\infty} P(|Y/k^{1/q}| > B) = 0, \quad (9.174)$$

completing the proof for the third series and hence the entire lemma.  $\square$

**Lemma 9.4.7** (Satisfying the three series for stochastic nonincreasingness when  $0 < q < 1$ ). *Suppose that  $X$  has a  $\mathcal{P}$ -UI (uncentered)  $q^{\text{th}}$  moment for  $0 < q < 1$ . Then the three series for uniform stochastic nonincreasingness in Theorem 9.3.3 are satisfied for the random variable  $Z_n := X_n/n^{1/q}$ .*

*Proof.* Similar to satisfying the conditions of the  $\mathcal{P}$ -uniform Kolmogorov three-series theorem as in Lemma 9.4.2, satisfying the conditions of the  $\mathcal{P}$ -uniform stochastic nonincreasingness three-series theorem proceeds identically for the second and third series, and thus we focus solely on the first series here. Throughout, let  $Z_{n,B}^{\leq 1} := (X/B)\mathbf{1}\{|(X/B)| \leq n^{1/q}\}/n^{1/q}$ . Writing out  $\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} |\mathbb{E}_P(Z_{n,B}^{\leq 1})|$  for any  $B > 0$ ,

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} |\mathbb{E}_P Z_{n,B}^{\leq 1}| \quad (9.175)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \mathbb{E}_P \left( \frac{|X/B|\mathbf{1}\{|X/B| \leq n^{1/q}\}}{n^{1/q}} \right) \quad (9.176)$$

$$= \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \mathbb{E}_P \left( |X/B| \mathbb{1}\{(k-1)^{1/q} < |X/B| \leq k^{1/q}\} \right) \cdot \sum_{n=k}^{\infty} \frac{1}{n^{1/q}}. \quad (9.177)$$

Now, following techniques from earlier proofs, notice that since  $0 < q < 1$ , there exists a constant  $C_q$  depending only on  $q$  so that  $\sum_{n=k}^{\infty} 1/n^{1/q} \leq C_q k/k^{1/q}$ . Breaking up the sum similarly to how we did for the second series of uniform stochastic nonincreasingness in Lemma 9.4.6, we have

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \left| \mathbb{E}_P Z_{n,B}^{\leq 1} \right| \leq \sup_{P \in \mathcal{P}} \sum_{k=1}^{\infty} \mathbb{E}_P \left( |X/B| \mathbb{1}\{(k-1)^{1/q} < |X/B| \leq k^{1/q}\} \right) \cdot C_q \frac{k}{k^{1/q}} \quad (9.178)$$

$$\leq C_q \underbrace{\sup_{P \in \mathcal{P}} \mathbb{E}_P (|X/B| \mathbb{1}\{|X/B| \leq 1\})}_{(\star)} + \quad (9.179)$$

$$\underbrace{C_q \sup_{P \in \mathcal{P}} \sum_{k=2}^{\infty} \mathbb{E}_P \left( |X/B| \mathbb{1}\{(k-1)^{1/q} < |X/B| \leq k^{1/q}\} \right) \frac{k}{k^{1/q}}}_{(\dagger)} \quad (9.180)$$

First looking at the first term  $(\star)$ , we notice that  $|X/B| \mathbb{1}\{|X/B| \leq 1\} \leq |X/B|^q \mathbb{1}\{|X/B| \leq 1\}$  with  $P$ -probability one for every  $P \in \mathcal{P}$  since  $0 < q < 1$ , and thus

$$(\star) \leq \frac{C_q}{B^q} \sup_{P \in \mathcal{P}} \mathbb{E}_P |X|^q. \quad (9.181)$$

Turning to the second term  $(\dagger)$ , we have that

$$(\dagger) \leq C_q \sup_{P \in \mathcal{P}} \sum_{k=2}^{\infty} \mathbb{E}_P \left( k^{1/q} \mathbb{1}\{(k-1)^{1/q} < |X/B| \leq k^{1/q}\} \right) \frac{k}{k^{1/q}} \quad (9.182)$$

$$\leq \frac{2C_q}{B^q} \sup_{P \in \mathcal{P}} \mathbb{E}_P |X|^q. \quad (9.183)$$

Combining the upper bounds on  $(\star)$  and  $(\dagger)$  and taking limits as  $B \rightarrow \infty$ , we have

$$\lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \left| \mathbb{E}_P Z_{n,B}^{\leq 1} \right| = 0, \quad (9.184)$$

completing the proof of Lemma 9.4.7. □

#### 9.4.1.2 An equivalent criterion of uniform integrability

**Lemma 9.4.8** (Uniform integrability is equivalent to uniformly vanishing sums of tail probabilities). *Let  $Y$  be a random variable on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$ . Then  $Y$  has a  $\mathcal{P}$ -UI  $q^{\text{th}}$*

moment if and only if the tail sum of its tail probability is  $\mathcal{P}$ -uniformly vanishing, meaning

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q \mathbf{1}(|Y|^q > m)) = 0 \quad \text{if and only if} \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P (|Y|^q > k) = 0. \quad (9.185)$$

The above lemma gives an equivalent criterion of  $\mathcal{P}$ -UI and was shown in Hu and Zhou [127, Theorem 2.1] where they refer to the property in the right-hand side of (9.17) as “ $W^*$  uniform integrability” [127, Definition 1.6]. Since Hu and Zhou [127, Theorem 2.1] is written in the context of uniform integrability for a family of random variables  $(X_n)_{n=1}^{\infty}$  defined on a single probability space (as contrasted with  $\mathcal{P}$ -UI in Section 9.1), we provide a self-contained proof for completeness.

*Proof.* The forward implication is not hard to show since for any  $P \in \mathcal{P}$ ,

$$\sum_{k=m}^{\infty} P (|Y|^q > k) = \sum_{k=m}^{\infty} P (|Y|^q \mathbf{1}(|Y|^q > m) > k) \quad (9.186)$$

$$\leq \int_0^{\infty} P (|Y|^q \mathbf{1}(|Y|^q > m) > k) dk \quad (9.187)$$

$$= \mathbb{E}_P (|Y|^q \mathbf{1}(|Y|^q > m)). \quad (9.188)$$

The reverse implication is more involved. Suppose that there exists a collection of distributions  $\mathcal{P}$  so that right-hand side of (9.185) holds but the left-hand side does not. First, note that we can write the supremum over expectations in (9.185) as

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P (|Y|^q \mathbf{1}(|Y|^q > m)) = \sup_{P \in \mathcal{P}} \int_0^{\infty} P (|Y|^q \mathbf{1}(|Y|^q > m) > k) dk \quad (9.189)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{k=0}^{\infty} P (|Y|^q \mathbf{1}(|Y|^q > m) > k) \quad (9.190)$$

$$\leq \underbrace{\sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} P (|Y|^q \mathbf{1}(|Y|^q > m) > k)}_{(\star)} + \quad (9.191)$$

$$\underbrace{\sup_{P \in \mathcal{P}} \sum_{k=0}^{m-1} P (|Y|^q \mathbf{1}(|Y|^q > m) > k)}_{(\dagger)}, \quad (9.192)$$

and since  $(\star) \rightarrow 0$  as  $m \rightarrow \infty$ , we must have that  $(\dagger) \rightarrow 0$ , and hence

$$\limsup_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=0}^{m-1} P (|Y|^q \mathbf{1}(|Y|^q > m) > k) > \varepsilon \quad (9.193)$$

for some  $\varepsilon > 0$ , or in other words, no matter how large we take  $M$  to be, we can always find some  $M^* \geq M$  and  $P^* \in \mathcal{P}$  so that  $\sum_{k=0}^{M^*-1} P^*(|Y|^q \mathbf{1}\{|Y|^q > M^*\} > k) > \varepsilon$ . Writing out the above sum, we have for any  $m, P$ ,

$$\sum_{k=0}^{m-1} P(|Y|^q \mathbf{1}\{|Y|^q > m\} > k) = \sum_{k=0}^{m-1} P(|Y|^q > m) \quad (9.194)$$

since  $|Y|^q \mathbf{1}\{|Y|^q > m\} > k$  if and only if  $|Y|^q > m$  whenever  $k \leq m$ . Carrying on with the above calculation, we have as a consequence of (9.193) that

$$\limsup_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} m P(|Y|^q > m) > \varepsilon. \quad (9.195)$$

Simultaneously, by the assumed uniform tail vanishing property in the right-hand side of (9.185), we can choose an  $M$  sufficiently large so that

$$\sup_{P \in \mathcal{P}} \sum_{k=M}^{\infty} P(|Y|^q \mathbf{1}\{|Y|^q > M\} > k) < \varepsilon/2. \quad (9.196)$$

By (9.195), find some  $M^* > 3M$  and  $P^* \in \mathcal{P}$  so that

$$M^* \cdot P^*(|Y|^q > M^*) > \varepsilon. \quad (9.197)$$

We will now show that (9.196) and (9.197) are incompatible, leading to a contradiction. Indeed,

$$\sup_{P \in \mathcal{P}} \sum_{k=M}^{\infty} P(|Y|^q \mathbf{1}\{|Y|^q > M\} > k) \quad (9.198)$$

$$\geq \sum_{k=M}^{\infty} P^*(|Y|^q \mathbf{1}\{|Y|^q > M\} > k) \quad (9.199)$$

$$\geq \sum_{k=M}^{M^*} P^*(|Y|^q \mathbf{1}\{|Y|^q > M\} > k) \quad (9.200)$$

$$= \sum_{k=M}^{M^*} P^*(|Y|^q > k), \quad (9.201)$$

where the first inequality follows by definition of a supremum, the second since we are taking only a smaller sum over finitely many elements, and the last equality follows from the fact that whenever  $k \geq M$ ,  $|Y|^q \mathbf{1}\{|Y|^q > M\} > k$  if and only if  $|Y|^q > k$ . Carrying on with the above calculation, we finally have

$$\sup_{P \in \mathcal{P}} \sum_{k=M}^{\infty} P(|Y|^q \mathbf{1}\{|Y|^q > M\} > k) \quad (9.202)$$

$$\geq \sum_{k=M}^{M^*} P^*(|Y|^q > k) \quad (9.203)$$

$$\geq \sum_{k=M}^{M^*} P^*(|Y|^q > M^*) \quad (9.204)$$

$$= (M^* - M)P^*(|Y|^q > M^*) \quad (9.205)$$

$$\geq (M^* - M^*/3)P^*(|Y|^q > M^*) \quad (9.206)$$

$$> 2\varepsilon/3, \quad (9.207)$$

which contradicts the fact that the same sum was assumed to be less than  $\varepsilon/2$  in (9.196), completing the proof.  $\square$

#### 9.4.2 Proof details for Theorem 9.2.2

*Proof of Theorem 9.2.2.* By Lemma 9.4.9, we have that the following  $\mathcal{P}$ -uniform moment regularity condition

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{k=m}^{\infty} \frac{\mathbb{E}_P |X_k - \mathbb{E}_P X_k|^q}{a_k^q} = 0 \quad (9.208)$$

implies that the three series in Theorem 9.3.2 vanish  $\mathcal{P}$ -uniformly with  $c = 1$  for the random variables  $Z_k^{\leq 1}$  which are scaled and truncated versions of  $X_k$  given by

$$Z_k^{\leq 1} := \frac{X_k - \mathbb{E}_P(X_k)}{a_k} \cdot \mathbf{1}\{|X_k - \mathbb{E}_P(X)| \leq a_k\}. \quad (9.209)$$

Therefore, by the  $\mathcal{P}$ -uniform three series theorem (Theorem 9.3.2) we have that  $S_n \equiv S_n(P) := \sum_{k=1}^n (X_k - \mathbb{E}_P(X))/a_k$  forms a  $\mathcal{P}$ -uniform Cauchy sequence. Moreover, by Theorem 9.3.3 and Lemma 9.4.10, we have that the  $\mathcal{P}$ -uniform moment regularity condition (9.208) implies that  $S_n$  is  $\mathcal{P}$ -uniformly stochastically nonincreasing. Combining the facts that  $(S_n)_{n=1}^{\infty}$  is  $\mathcal{P}$ -uniformly Cauchy and stochastically nonincreasing, we invoke the  $\mathcal{P}$ -uniform Kronecker lemma (Lemma 9.3.1) similar to the proof of Theorem 9.2.1(i) but now with the sequence  $(a_n)_{n=1}^{\infty}$  to yield the desired result, completing the proof of Theorem 9.2.2.  $\square$

**Lemma 9.4.9** (Satisfying the three series for independent random variables). *Let  $X_1, \dots, X_n$  be independent random variables on  $(\Omega, \mathcal{F}, \mathcal{P})$  and let  $Y_n := X_n - \mathbb{E}X_n$  be their centered versions. Suppose that for some increasing sequence  $(a_n)_{n=1}^{\infty}$  that diverges to  $\infty$ ,*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \frac{\mathbb{E}_P |Y_n|^q}{a_n^q} = 0. \quad (9.210)$$

*Then the three series conditions of Theorem 9.3.2 are satisfied for  $Z_k := Y_k/a_k$  with  $c = 1$ .*

*Proof.* First, define the truncated random variables  $Z_n^{\leq 1} := Z_n \mathbf{1}\{|Z_n| \leq 1\}$  for any  $n$ . We will satisfy the three series separately.

**The first series.** Writing out  $\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} |\mathbb{E}_P Z_n^{\leq 1}|$ , we have

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} |\mathbb{E}_P Z_n^{\leq 1}| = \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \left| \mathbb{E}_P \left( \frac{Y_n \mathbf{1}(|Y_n| \leq a_n)}{a_n} \right) \right| \quad (9.211)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \mathbb{E}_P \left( \frac{|Y_n| \mathbf{1}(|Y_n| > a_n)}{a_n} \right) \quad (9.212)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \frac{\mathbb{E}_P |Y_n|^q}{a_n^q} \rightarrow 0 \quad (9.213)$$

where the last inequality follows from the fact that  $(|Y_n|/a_n) \mathbf{1}\{|Y_n| > a_n\} \leq (|Y_n|^q/a_n^q) \mathbf{1}\{|Y_n| > a_n\}$  since  $q \geq 1$ .

**The second series.** Writing out  $\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \text{Var}_P Z_n^{\leq 1}$ , we have

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \text{Var}_P Z_n^{\leq 1} \leq \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \mathbb{E}_P [(Z_n^{\leq 1})^2] \quad (9.214)$$

$$= \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \mathbb{E}_P \left( \frac{Y_n^2 \mathbf{1}(|Y_n| \leq a_n)}{a_n^2} \right) \quad (9.215)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \frac{\mathbb{E}_P |Y_n|^q}{a_n^q} \rightarrow 0, \quad (9.216)$$

where the last inequality follows from the fact that  $(Y_n^2/a_n^2) \mathbf{1}\{|Y_n| \leq a_n\} \leq (|Y_n|^q/a_n^q) \mathbf{1}\{|Y_n| \leq a_n\}$  with  $P$ -probability one for each  $P \in \mathcal{P}$  since  $q \leq 2$ .

**The third series.** Finally, writing out the third series  $\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} P(|Y_n/a_n| > 1)$ , we have by Markov's inequality,

$$\sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} P \left( \left| \frac{Y_n}{a_n} \right| > 1 \right) \leq \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \frac{\mathbb{E}_P |Y_n|^q}{a_n^q} \rightarrow 0, \quad (9.217)$$

which completes the proof.  $\square$

**Lemma 9.4.10** (Satisfying the three series of  $\mathcal{P}$ -uniform stochastic nonincreasingness under independence). *Let  $X_1, \dots, X_n$  be independent random variables and let  $Y_n := X_n - \mathbb{E}X_n$  be their centered versions. Suppose that for some increasing sequence  $(a_n)_{n=1}^{\infty}$  that diverges to  $\infty$ ,*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \frac{\mathbb{E}_P |Y_n|^q}{a_n^q} = 0. \quad (9.218)$$

*Then the three series conditions of Theorem 9.3.3 are satisfied for  $Z_k := Y_k/a_k$ .*

*Proof.* First, define the truncated random variables  $Z_{n,B}^{\leq 1}$  as

$$Z_{n,B}^{\leq 1} := \frac{(Y_n/B)\mathbb{1}\{|Y_n/B| \leq a_n\}}{a_n}. \quad (9.219)$$

We will satisfy the three series separately.

**The first series.** Writing out  $\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} |\mathbb{E}_P Z_{n,B}^{\leq 1}|$  for any  $B > 0$ , we have

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} |\mathbb{E}_P Z_{n,B}^{\leq 1}| = \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \left| \mathbb{E}_P \left( \frac{(Y_n/B)\mathbb{1}\{|Y_n/B| \leq a_n\}}{a_n} \right) \right| \quad (9.220)$$

$$\leq \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \mathbb{E}_P \left( \frac{|Y_n/B|\mathbb{1}\{|Y_n/B| > a_n\}}{a_n} \right) \quad (9.221)$$

$$\leq \frac{1}{B^q} \underbrace{\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \frac{\mathbb{E}_P |Y_n|^q}{a_n^q}}_{<\infty}, \quad (9.222)$$

and we note that  $\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \mathbb{E}_P |Y_n|^q / a_n^q < \infty$  since  $\lim_m \sup_{P \in \mathcal{P}} \sum_{n=m}^{\infty} \mathbb{E}_P |Y_n|^q / a_n^q = 0$ . Therefore,

$$\lim_{B \rightarrow \infty} \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} |\mathbb{E}_P Z_{n,B}^{\leq 1}| = 0, \quad (9.223)$$

completing the proof of the first series.

**The second series.** Writing out  $\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \text{Var}_P Z_{n,B}^{\leq 1}$ , we have

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \text{Var}_P Z_{n,B}^{\leq 1} \leq \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \mathbb{E}_P [(Z_{n,B}^{\leq 1})^2] \quad (9.224)$$

$$= \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \mathbb{E}_P \left( \frac{(Y_n/B)^2 \mathbb{1}\{|Y_n/B| \leq a_n\}}{a_n^2} \right) \quad (9.225)$$

$$\leq \frac{1}{B^q} \underbrace{\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \frac{\mathbb{E}_P |Y_n|^q}{a_n^q}}_{<\infty}. \quad (9.226)$$

Therefore,  $\lim_B \sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \text{Var}_P Z_{n,B}^{\leq 1} = 0$ , completing the proof for the second series.

**The third series.** Finally, writing out the third series  $\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} P(|Y_n/B| > a_n)$ , we have by Markov's inequality,

$$\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} P(|Y_n/B| > a_n) \leq \frac{1}{B^q} \underbrace{\sup_{P \in \mathcal{P}} \sum_{n=1}^{\infty} \frac{\mathbb{E}|Y_n|^q}{a_n^q}}_{<\infty} \rightarrow 0 \quad (9.227)$$

as  $B \rightarrow \infty$  which completes the proof.  $\square$

## 9.5 Application to uniformly consistent variance estimation

Estimating a parameter from a random sample is a fundamental task in the field of statistics. Indeed, using the language of statistical estimation, Kolmogorov's strong law of large numbers is synonymous with saying that the sample average  $\frac{1}{n} \sum_{i=1}^n X_i$  is a strongly  $P$ -consistent estimator of the population mean  $\mathbb{E}_P(X)$  (meaning that the estimator converges to the estimand  $P$ -a.s.) while the theorems of Marcinkiewicz and Zygmund [184] show that the *rate* of strong  $P$ -consistency can be improved to  $o_{\text{a.s.}}(n^{1/q-1})$  when  $\mathbb{E}_P|X|^q < \infty$  for  $1 < q < 2$ . Using this same language, the SLLNs of Chung [62] and Theorem 9.2.1(i) are simply alternative ways of stating that the sample mean is strongly  $\mathcal{P}$ -uniformly consistent for the mean in a class of distributions  $\mathcal{P}$  under the appropriate  $\mathcal{P}$ -UI conditions.

For the above reasons, applications of SLLNs to statistical estimation of the *mean* are immediate and transparent. In the context of *variance* estimation, however, such applications require more care and have important implications for sequential hypothesis testing and statistical inference. For example, consider Chapter 7 where we derived coverage guarantees of so-called “asymptotic confidence sequences”. Let us summarize a simplified version of those goals here. Given a potentially infinite stream of i.i.d. observations  $X_1, X_2, \dots$  on a single probability space  $(\Omega, \mathcal{F}, P)$  and any desired coverage probability  $(1 - \alpha)$  for  $\alpha \in (0, 1)$ , we wish to construct a sequence of intervals  $[L_k^{(\alpha)}, U_k^{(\alpha)}]_{k=m}^{\infty}$  — where  $L_k^{(\alpha)}$  and  $U_k^{(\alpha)}$  are  $\sigma(X_1, \dots, X_k)$ -measurable — so that these intervals cover the mean  $\mathbb{E}_P(X)$  uniformly for all  $k \geq m$  with probability tending to  $(1 - \alpha)$ , meaning that

$$\lim_{m \rightarrow \infty} P\left(\forall k \geq m, \mathbb{E}_P(X) \in [L_k^{(\alpha)}, U_k^{(\alpha)}]\right) = 1 - \alpha. \quad (9.228)$$

The guarantee in (9.228) can be viewed as a time-uniform<sup>1</sup> analogue of the coverage guarantee enjoyed by large-sample confidence intervals for a fixed sample size. Chapter 7 derives intervals satisfying (9.228) which rely on the variance  $\text{Var}_P(X)$  being strongly  $P$ -consistently estimated at a polynomial rate. That is, those intervals involved an estimator  $\hat{\sigma}_n^2$  so that  $\hat{\sigma}_n^2 - \text{Var}_P(X) = o(n^{-\beta})$   $P$ -almost surely for some  $\beta > 0$ , and to achieve this, in Chapter 7, we let  $\hat{\sigma}_n^2$  be the sample variance and showed via the Marcinkiewicz-Zygmund SLLN that this strong polynomial-rate consistency holds. Notice, however, that (9.228) is a distribution-pointwise

---

<sup>1</sup>Here, we are using “time” to refer to the sample size as is common in the sequential inference literature [124, 125].

convergence result, and hence a distribution-pointwise SLLN is sufficient for the goals of that chapter. It is plausible that any hope of deriving a distribution-*uniform* analogue of (9.228) will rely on showing that  $\hat{\sigma}_n^2$  is strongly  $\mathcal{P}$ -uniformly consistent for the variance in the sense that  $\hat{\sigma}_n^2 - \text{Var}(X) = \bar{o}_{\mathcal{P}}(n^{-\beta})$  for some  $\beta > 0$  (and indeed, Chapter 8 relies heavily on the results of this section). The following corollary shows that such a guarantee follows from Theorem 9.2.1(i), laying some of the groundwork for potential extensions of (9.228) to the uniform setting.

**Corollary 9.5.1** ( $\mathcal{P}$ -uniformly strongly consistent variance estimation). *Let  $X_1, \dots, X_n$  be i.i.d. random variables defined on the collection of probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$ . Suppose that the  $(2 + \delta)^{\text{th}}$  moment of  $X$  is  $\mathcal{P}$ -UI for some  $\delta \in (0, 2)$  so that*

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left( |X - \mathbb{E}_P(X)|^{2+\delta} \mathbf{1}\{|X - \mathbb{E}_P(X)|^{2+\delta} > m\} \right) = 0. \quad (9.229)$$

*Then, the sample variance  $\hat{\sigma}_n^2$  is a  $\mathcal{P}$ -uniformly strongly consistent estimator for the variance at a rate of  $\bar{o}_{\mathcal{P}}(n^{2/(2+\delta)-1})$ , meaning  $\hat{\sigma}_n^2 - \text{Var}(X) = \bar{o}_{\mathcal{P}}(n^{2/(2+\delta)-1})$ , or more formally,*

$$\forall \varepsilon > 0, \quad \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} \left\{ \frac{|\hat{\sigma}_k^2 - \text{Var}_P(X)|}{k^{2/(2+\delta)-1}} \right\} \geq \varepsilon \right) = 0. \quad (9.230)$$

*Proof.* First, letting  $\bar{X}_n := \frac{1}{n} \sum_{i=1}^n X_i$ , notice that for any fixed  $P \in \mathcal{P}$ , we can write  $\hat{\sigma}_n^2$  as

$$\hat{\sigma}_n^2 - \text{Var}_P(X) \equiv \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 - \text{Var}_P(X) \quad (9.231)$$

$$= \underbrace{\frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}_P(X))^2}_{(\star)} - \text{Var}_P(X) - \underbrace{(\bar{X}_n - \mathbb{E}_P(X))^2}_{(\dagger)}. \quad (9.232)$$

Letting  $q := 2 + \delta$ , note that by Theorem 9.2.1(i) we have that  $\bar{X}_n - \mathbb{E}(X) = \bar{o}_{\mathcal{P}}(n^{1/\gamma-1})$  for  $\gamma := 2/(2/q+1)$  since  $1 < \gamma < 2$  and in particular,  $(\dagger) = \bar{o}_{\mathcal{P}}(n^{2/q-1})$ . To analyze  $(\star)$ , we write  $Y := X - \mathbb{E}_P(X)$  and note that  $Y^2 \equiv (X - \mathbb{E}_P(X))^2$  has a  $\mathcal{P}$ -UI  $(q/2)^{\text{th}}$  moment:

$$\lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left( |Y^2 - \mathbb{E}_P(Y^2)|^{q/2} \cdot \mathbf{1}\{|Y^2 - \mathbb{E}_P(Y^2)|^{q/2} > m\} \right) \quad (9.233)$$

$$\leq \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P \left( |Y^2|^{q/2} \cdot \mathbf{1}\{|Y^2|^{q/2} > m\} \right) \quad (9.234)$$

$$= \lim_{m \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P (|X - \mathbb{E}_P(X)|^q \cdot \mathbf{1}\{|X - \mathbb{E}_P(X)|^q > m\}) \quad (9.235)$$

$$= 0, \quad (9.236)$$

and hence we have that  $(\star) = \bar{o}_{\mathcal{P}}(n^{2/q-1})$ . Putting this together with the analysis for  $(\dagger)$ , we

have by the triangle inequality that

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} k^{-2/q+1} |\hat{\sigma}_n^2 - \text{Var}_P(X)| \geq \varepsilon \right) \quad (9.237)$$

$$\leq \sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} k^{-2/q+1} \left| \frac{1}{k} \sum_{i=1}^k (X_i - \mathbb{E}_P(X))^2 - \text{Var}_P(X) \right| \geq \varepsilon \right) + \quad (9.238)$$

$$\sup_{P \in \mathcal{P}} P \left( \sup_{k \geq m} k^{-1/\gamma+1} |\bar{X}_k - \mathbb{E}_P(X)| \geq \sqrt{\varepsilon} \right), \quad (9.239)$$

and the first supremum vanishes as  $m \rightarrow \infty$  by (9.236) while the second vanishes as  $m \rightarrow \infty$  by the fact that  $1 < \gamma < 2$ , and hence  $\hat{\sigma}_n^2 - \text{Var}(X) = \bar{o}_{\mathcal{P}}(n^{2/q-1})$ , completing the proof.  $\square$

## 9.6 Summary

In this chapter, we introduced a set of tools and techniques to derive distribution-uniform strong laws of large numbers, culminating in extensions of Chung's i.i.d. strong law to uniformly integrable  $q^{\text{th}}$  moments for  $0 < q < 2$ ;  $q \neq 1$  in the sense of Marcinkiewicz and Zygmund [184] as well as to independent but non-identically distributed random variables. Furthermore, we showed that  $\mathcal{P}$ -uniform integrability of the  $q^{\text{th}}$  moment is both sufficient *and necessary* for the strong law to hold at the Kolmogorov-Marcinkiewicz-Zygmund rates of  $\bar{o}_{\mathcal{P}}(n^{1/q-1})$ , shedding new light on uniform strong laws even in Chung's case when  $q = 1$ .

As alluded to in the introduction, Ruf et al. [229] were able to prove Chung's strong law using an argument not resembling those typically found in almost sure convergence theorems. In short, they derive a novel high-probability line-crossing inequality for sums of i.i.d. random variables whose first moments are finite and show how this inequality can be uniformly controlled in a family with a uniformly integrable first moment. It is not obvious to us whether their proof techniques can be adapted to higher (or lower) finite moments or to non-identically distributed settings, but this would be interesting to see.

We anticipate that the proof techniques found in Sections 9.3 and 9.4 may open vistas for understanding other distribution-uniform almost sure behavior. In particular, we plan to explore their use in the development of distribution-uniform analogues of strong invariance principles such as the Komlós-Major-Tusnády embeddings [160, 161] in the presence of  $q^{\text{th}}$  uniformly integrable moments when  $q > 2$ .

## 9.A A note on de la Vallée-Poussin's criterion

In Lemma 9.4.3, we used the fact that in de la Vallée-Poussin's criterion of uniform integrability, the function  $h$  diverging to  $\infty$  can be taken to be concave, strictly increasing, and starting at  $h(0) = 0$ . Here, we provide a self-contained proof of this fact, using the same argument found in some notes on the website of José A. Cañizo [45].

**Proposition 9.A.1.** *Let  $Y$  be a random variable on the probability spaces  $(\Omega, \mathcal{F}, \mathcal{P})$  so that  $\mathbb{E}_P|Y| < \infty$  for every  $P \in \mathcal{P}$ . Then  $Y$  is (uncentered) uniformly integrable if and only if there exists a measurable function  $h : [0, \infty) \rightarrow [0, \infty)$  that is concave, strictly increasing, and starting at  $h(0) = 0$  so that*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P [|Y|h(|Y|)] < \infty. \quad (9.240)$$

*Proof.* The “if” direction follows from the usual statement of de la Vallée-Poussin’s theorem [58] so we will focus on the “only if” direction. That is, we are tasked with finding a function  $h$  satisfying the conditions above. To begin, define the uniform integrability tail  $\mathbb{U}_P(y)$  with lower truncation level  $y \geq 0$  as

$$\mathbb{U}_P(y) := \mathbb{E}_P (|Y|\mathbf{1}\{|Y| \geq y\}). \quad (9.241)$$

and let  $\mathbb{U}(y) := \sup_{P \in \mathcal{P}} \mathbb{U}_P(y)$ , noting that  $\mathbb{U}$  is nonincreasing and  $\mathbb{U}(y) \rightarrow 0$  as  $y \rightarrow \infty$  by  $Y$  being uniformly integrable. Now, for every  $n = 1, 2, \dots$ , define

$$a_n := \inf\{y > 0 : \mathbb{U}(x) < 1/n^2\} \quad (9.242)$$

and consider the increasing sequence  $(x_n)_{n=0,1,\dots}$  starting at  $x_0 = 0$  and defined recursively as

$$x_{n+1} := \max\{x_n + 1, a_{n+1} + 1\}, \quad (9.243)$$

noting that  $x_n$  is strictly increasing in  $n$  and diverging to  $\infty$ . Further, observe that  $\mathbb{U}(x_n) \leq 1/n^2$  since  $x_n \geq a_n$  for each  $n$ . Now define the function  $\phi$  as the step function given by

$$\phi(x) := n + 1 \quad \text{when } x \in [x_n, x_{n+1}) \text{ for } n = 0, 1, 2, \dots, \quad (9.244)$$

and notice that  $\phi(x)$  diverges in  $x$  because  $(x_n)_n$  does in  $n$ . We now observe that  $\sup_{P \in \mathcal{P}} \mathbb{E}_P [|Y|\phi(|Y|)]$  is bounded:

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P [|Y|\phi(|Y|)] = \sup_{P \in \mathcal{P}} \mathbb{E}_P \left[ \sum_{n=0}^{\infty} |Y|\mathbf{1}\{|Y| \geq x_n\} \right] \quad (9.245)$$

$$= \sup_{P \in \mathcal{P}} \sum_{n=0}^{\infty} \mathbb{E}_P [|Y|\mathbf{1}\{|Y| \geq x_n\}] \quad (9.246)$$

$$= \sup_{P \in \mathcal{P}} \sum_{n=0}^{\infty} \mathbb{U}_P(x_n) \quad (9.247)$$

$$\leq \sup_{P \in \mathcal{P}} \underbrace{\mathbb{U}_P(0)}_{\mathbb{E}_P|Y|} + \sup_{P \in \mathcal{P}} \underbrace{\sum_{n=1}^{\infty} \frac{1}{n^2}}_{\pi^2/6} < \infty. \quad (9.248)$$

Now, we will construct  $h$  as a piecewise-linear, nonnegative, and continuous concave minorant of  $\phi$ . Following Cañizo [45], we will let  $(d_n)_{n=0}^{\infty}$  denote the piecewise derivatives of  $h$  and define both  $d_n$  and  $h(x)$  recursively as  $d_0 = 1$ ,  $h(0) = 0$  and

$$d_{n+1} := \min \left\{ d_n, \frac{n+1 - h(x_n)}{x_{n+1} - x_n} \right\}; \quad n = 0, 1, 2, \dots, \quad (9.249)$$

$$h(x) := h(x_n) + d_{n+1}(x - x_n); \quad x \in [x_n, x_{n+1}]. \quad (9.250)$$

Clearly  $h$  is continuous. Moreover, notice that it is differentiable on each interval  $(x_n, x_{n+1})$  with decreasing derivative  $d_{n+1}$  so  $h$  is concave. This completes the proof.  $\square$

# Bibliography

- [1] Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296, 2020. [244](#)
- [2] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL <https://proceedings.neurips.cc/paper/2011/file/e1d5be1c7f2f456670de3d53c7b54f4a-Paper.pdf>. [51](#)
- [3] Jayadev Acharya, Clément L Canonne, Yanjun Han, Ziteng Sun, and Himanshu Tyagi. Domain compression and its application to randomness-optimal distributed goodness-of-fit. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 3–40. PMLR, 2020. [197](#)
- [4] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: Lower bounds from Chi-square contraction. *IEEE Transactions on Information Theory*, 66(12):7835–7855, 2020. [197](#)
- [5] Jayadev Acharya, Clément L. Canonne, Cody Freitag, Ziteng Sun, and Himanshu Tyagi. Inference under information constraints III: Local privacy constraints. *IEEE Journal on Selected Areas in Information Theory*, 2(1):253–267, 2021. [197](#)
- [6] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private Assouad, Fano, and Le Cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021. [145](#), [197](#)
- [7] Jayadev Acharya, Clément L. Canonne, Yuhan Liu, Ziteng Sun, and Himanshu Tyagi. Interactive inference under information constraints. *IEEE Transactions on Information Theory*, 68(1):502–516, 2022. [197](#)
- [8] Kareem Amin, Matthew Joseph, and Jieming Mao. Pan-private uniformity testing. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 183–218. PMLR, 2020. [197](#)
- [9] Theodore Anderson. Confidence limits for the expected value of an arbitrary bounded random variable with a continuous distribution function. Technical report, Stanford University Department of Statistics, 1969. [98](#), [99](#)

- [10] Apple Inc. Differential privacy overview. [https://www.apple.com/privacy/docs/Differential\\_Privacy\\_Overview.pdf](https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf), 2022. Accessed: 2022-02-01. 144, 145, 152
- [11] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. 326
- [12] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In *Algorithmic Learning Theory*, 2007. 45, 56, 171
- [13] Jordan Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. *Advances in Neural Information Processing Systems*, 31:4208–4218, 2018. 145, 197
- [14] Raghu R Bahadur and Leonard J Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956. 7, 248, 276
- [15] Sivaraman Balakrishnan, Edward H Kennedy, and Larry Wasserman. The fundamental limits of structure-agnostic functional estimation. *arXiv preprint arXiv:2305.04116*, 2023. 338
- [16] Borja Balle, James Bell, Adrià Gascón, and Kobbi Nissim. The privacy blanket of the shuffle model. In *Advances in Cryptology–CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part II* 39, pages 638–667. Springer, 2019. 147, 148
- [17] Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv:1405.2639*, 2014. 3, 51
- [18] Akshay Balsubramani and Aaditya Ramdas. Sequential Nonparametric Testing with the Law of the Iterated Logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, 2016. 3, 22
- [19] Rémi Bardenet and Odalric-Ambrym Maillard. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361–1385, 2015. 9, 15, 20, 21, 38, 40, 45, 66
- [20] Leighton Pate Barnes, Wei-Ning Chen, and Ayfer Özgür. Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3):645–659, 2020. 197
- [21] RCM Barnes, EH Cooke-Yarborough, and DGA Thomas. An electronic digital computer using cold cathode counting tubes for storage. *Electronic Engineering*, 23:286–91, 1951. 147
- [22] Anatole Beck and Daniel P Giesy. P-uniform convergence and a vector-valued strong law of large numbers. *Transactions of the American Mathematical Society*, 147(2):541–559, 1970. 333

- [23] Robert Bell and Thomas M Cover. Game-theoretic optimal portfolios. *Management Science*, 34(6):724–733, 1988. [71](#), [122](#)
- [24] Robert M Bell and Thomas M Cover. Competitive optimality of logarithmic investment. *Mathematics of Operations Research*, pages 161–166, 1980. [71](#), [122](#)
- [25] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. [228](#)
- [26] Yoav Benjamini and Daniel Yekutieli. False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005. [228](#)
- [27] George Bennett. Probability Inequalities for the Sum of Independent Random Variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962. [44](#), [100](#)
- [28] Vidmantas Bentkus. On Hoeffding’s inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004. [45](#), [98](#), [171](#)
- [29] Vidmantas Bentkus, N Kalosha, and M Van Zuijlen. On domination of tail probabilities of (super) martingales: explicit bounds. *Lithuanian Mathematical Journal*, 46(1):1–43, 2006. [98](#), [100](#)
- [30] Sergei Bernstein. *Theory of probability*. Gostehizdat Publishing House, 1927. [45](#)
- [31] Thomas Berrett and Cristina Butucea. Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms. In *Advances in Neural Information Processing Systems*, volume 33, pages 3164–3173, 2020. [197](#)
- [32] Tom Berrett and Yi Yu. Locally private online change point detection. *Advances in Neural Information Processing Systems*, 34, 2021. [198](#)
- [33] Aurélien Bibaut, Maria Dimakopoulou, Nathan Kallus, Antoine Chambaz, and Mark van der Laan. Post-contextual-bandit inference. *Advances in Neural Information Processing Systems*, 34:28548–28559, 2021. [200](#), [201](#), [205](#), [218](#)
- [34] Aurélien Bibaut, Nathan Kallus, and Michael Lindon. Near-optimal non-parametric sequential tests and confidence sequences with possibly dependent observations. *arXiv preprint arXiv:2212.14411*, 2022. [8](#), [260](#), [262](#), [263](#), [317](#), [318](#), [326](#), [328](#), [371](#)
- [35] Peter J Bickel, Chris AJ Klaassen, Ya’acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993. [319](#)
- [36] Patrick Billingsley. *Probability and measure (3rd ed.)*. John Wiley & Sons, 1995. [253](#), [256](#), [344](#)

- [37] Michelle Blom, Jurlind Budurushi, Ronald L Rivest, Philip B Stark, Peter J Stuckey, Vanessa Teague, and Damjan Vukcevic. Assertion-based approaches to auditing complex elections, with application to party-list proportional elections. In *International Joint Conference on Electronic Voting*. Springer, 2021. [125](#)
- [38] Bokeh Development Team. *Bokeh: Python library for interactive visualization*, 2018. URL <https://bokeh.pydata.org/en/latest/>. [138](#)
- [39] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, Oxford, 1st edition, 2013. [47](#), [49](#)
- [40] Legislative Services Branch. Canada elections act, part 12: Counting votes, 2019. URL <https://laws-lois.justice.gc.ca/eng/acts/e-2.01/page-37.html#h-206023>. [138](#), [139](#)
- [41] Leo Breiman. Optimal gambling systems for favorable games. *Berkeley Symposium on Mathematical Statistics and Probability*, 4.1(65-78), 1961. [71](#)
- [42] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996. [268](#), [343](#)
- [43] Cristina Butucea and Yann Issartel. Locally differentially private estimation of functionals of discrete distributions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24753–24764. Curran Associates, Inc., 2021. [197](#)
- [44] Cristina Butucea, Amandine Dubois, Martin Kroll, and Adrien Saumard. Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli*, 26(3):1727–1764, 2020. [145](#), [197](#)
- [45] José A. Cañizo. The lemma of de la Vallée-Poussin. Accessed: 5-14-2024. [418](#), [419](#)
- [46] Elections Canada. 43rd general election: Official voting results (raw data), 2019. URL <https://www.elections.ca/content.aspx?section=res&dir=rep/off/43gedata&document=index&lang=e>. [138](#), [139](#)
- [47] Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold:‘model-x’ knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):551–577, 2018. [336](#)
- [48] Clément L Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 310–321, 2019. [145](#), [197](#)
- [49] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013. [84](#)

- [50] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006. [71](#), [122](#), [123](#)
- [51] Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:27475–27490, 2021. [201](#), [203](#), [205](#), [226](#), [227](#)
- [52] Tapas Kumar Chandra. de la Vallée Poussin’s theorem, uniform integrability, tightness and moments. *Statistics & Probability Letters*, 107:136–141, 2015. [361](#), [404](#)
- [53] Tapas Kumar Chandra and A Goswami. Cesaro uniform integrability and the strong law of large numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 215–231, 1992. [381](#)
- [54] Sourav Chatterjee. A new approach to strong embeddings. *Probability Theory and Related Fields*, 152(1-2):231–264, 2012. [253](#), [327](#)
- [55] Haoyu Chen, Wenbin Lu, and Rui Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255, 2021. [205](#)
- [56] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. [208](#), [247](#), [264](#), [265](#), [319](#), [320](#)
- [57] Yo Joong Choe and Aaditya Ramdas. Comparing sequential forecasters. *Operations Research*, 2023. [216](#), [222](#)
- [58] Kong-Ming Chong. On a theorem concerning uniform integrability. *Publ Inst Math (Beograd)(NS)*, 25(39):8–10, 1979. [335](#), [361](#), [381](#), [385](#), [404](#), [418](#)
- [59] Yuan Shih Chow and Henry Teicher. Integration in a Probability Space. In *Probability Theory: Independence, Interchangeability, Martingales*, pages 84–112. Springer New York, New York, NY, 1997. ISBN 978-1-4612-1950-7. doi: 10.1007/978-1-4612-1950-7\_4. URL [https://doi.org/10.1007/978-1-4612-1950-7\\_4](https://doi.org/10.1007/978-1-4612-1950-7_4). [381](#)
- [60] Alexander Mangulad Christgau, Lasse Petersen, and Niels Richard Hansen. Nonparametric conditional local independence testing. *The Annals of Statistics*, 51(5):2116–2144, 2023. [334](#)
- [61] Ben Chugg, Hongjian Wang, and Aaditya Ramdas. Time-uniform confidence spheres for means of random vectors. *arXiv preprint arXiv:2311.08168*, 2023. [321](#), [322](#)
- [62] Kai Lai Chung. The strong law of large numbers. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 341–353. University of California Press, January 1951. [379](#), [381](#), [383](#), [385](#), [386](#), [415](#)

- [63] Kai Lai Chung. *A course in probability theory*. Academic press, 2001. 381
- [64] KL Chung. The strong law of large numbers. *Selected Works of Kai Lai Chung*, pages 145–156, 2008. 333
- [65] Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the Lambert W function. *Advances in Computational mathematics*, 5(1):329–359, 1996. 312, 313
- [66] Simon Couch, Zeki Kazan, Kaiyan Shi, Andrew Bray, and Adam Groce. Differentially private nonparametric hypothesis testing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 737–751, 2019. 145, 197
- [67] Thomas M Cover. Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin. *Technical Report, no. 12*, 1974. 71, 122
- [68] Thomas M Cover. An algorithm for maximizing expected log investment return. *IEEE Transactions on Information Theory*, 30(2):369–373, 1984. 71, 122
- [69] Thomas M Cover. Log optimal portfolios. In *Chapter in “Gambling Research: Gambling and Risk Taking,” Seventh International Conference*, volume 4, 1987. 71, 122
- [70] Thomas M Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991. 71, 122
- [71] Christian Covington, Xi He, James Honaker, and Gautam Kamath. Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*, 2021. 145, 197
- [72] Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Proceedings of the 31st Conference On Learning Theory*, volume 75, 2018. 71, 72, 86, 87, 88, 122
- [73] D. A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967. 48, 51, 71, 121
- [74] D. A. Darling and Herbert Robbins. Inequalities for the Sequence of Sample Means. *Proceedings of the National Academy of Sciences*, 57(6):1577–1580, 1967. 71, 121
- [75] D. A. Darling and Herbert Robbins. Iterated Logarithm Inequalities. *Proceedings of the National Academy of Sciences*, 57(5):1188–1192, 1967. 71, 121
- [76] D. A. Darling and Herbert E. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58 1:66–8, 1967. 3, 202, 217, 247, 250, 252, 326
- [77] Open Data. Federal electoral districts - Canada, 2019. URL <https://open.canada.ca/data/en/dataset/5931f6f0-0008-4b0c-94d7-a1ff596182c5>. 138

- [78] A Philip Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984. [53](#), [122](#), [123](#)
- [79] A Philip Dawid. Prequential analysis. *Encyclopedia of Statistical Sciences*, 1:464–470, 1997. [122](#), [123](#)
- [80] Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *The Annals of Probability*, 32(3):1902–1933, 2004. ISSN 0091-1798, 2168-894X. [51](#)
- [81] Victor H. de la Peña, Michael J. Klass, and Tze Leung Lai. Pseudo-maximization and self-normalized processes. *Probability Surveys*, 4:172–192, 2007. [50](#), [51](#)
- [82] Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: limit theory and statistical applications*. Springer, Berlin, 2009. [51](#)
- [83] Iván Díaz and Mark van der Laan. Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549, 2012. [277](#)
- [84] Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34:1939–1951, 2021. [205](#)
- [85] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3574–3583, 2017. [145](#), [148](#), [167](#)
- [86] Joseph Leo Doob. *Stochastic Processes*, volume 10. New York Wiley, 1953. [120](#)
- [87] Joerg Drechsler, Ira Globus-Harris, Audra McMillan, Jayshree Sarathy, and Adam Smith. Non-parametric differentially private confidence intervals for the median. *arXiv preprint arXiv:2106.10333*, 2021. [145](#), [197](#)
- [88] Boyan Duan, Aaditya Ramdas, and Larry Wasserman. Interactive rank testing by betting. In *Conference on Causal Learning and Reasoning*, pages 201–235. PMLR, 2022. [336](#)
- [89] John C Duchi and Feng Ruan. The right complexity measure in locally private estimation: It is not the Fisher information. *arXiv preprint arXiv:1806.05756*, 2018. [197](#)
- [90] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013. [144](#), [145](#), [146](#), [152](#), [197](#)
- [91] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and minimax bounds: sharp rates for probability estimation. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1*, pages 1529–1537, 2013. [145](#), [197](#)

- [92] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018. [11](#), [144](#), [145](#), [197](#)
- [93] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1097–1104, 2011. [199](#), [207](#), [208](#)
- [94] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014. [199](#), [207](#), [208](#)
- [95] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, (3):642–669, 1956. [227](#)
- [96] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014. [144](#), [146](#), [150](#)
- [97] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006. [143](#), [144](#), [146](#), [147](#)
- [98] Bradley Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971. [203](#), [220](#), [265](#)
- [99] Uwe Einmahl. Strong invariance principles for partial sums of independent random vectors. *Annals of Probability*, 15(4):1419–1440, 1987. [321](#), [322](#), [323](#), [344](#)
- [100] Uwe Einmahl. Extensions of results of Komlós, Major, and Tusnády to the multivariate case. *Journal of multivariate analysis*, 28(1):20–68, 1989. [253](#), [344](#)
- [101] Uwe Einmahl. A new strong invariance principle for sums of independent random vectors. *Journal of Mathematical Sciences*, 163(4):311–327, 2009. [251](#), [343](#)
- [102] Xiequan Fan, Ion Grama, and Quansheng Liu. Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20(1):1–22, 2015. [35](#), [54](#), [75](#), [79](#), [83](#), [86](#), [187](#), [195](#), [212](#), [216](#), [229](#), [230](#)
- [103] Cecilia Ferrando, Shufan Wang, and Daniel Sheldon. Parametric bootstrap for differentially private confidence intervals. In *International Conference on Artificial Intelligence and Statistics*, pages 1598–1618. PMLR, 2022. [145](#), [197](#)
- [104] Daniel Fink. A compendium of conjugate priors, 1997. [18](#)
- [105] Ronald A Fisher. Design of experiments. *British Medical Journal*, 1(3923):554, 1936. [244](#)
- [106] Ronald A Fisher. Mathematics of a lady tasting tea. *The world of mathematics*, 3(part 8):1514–1521, 1956. [28](#)

- [107] George E Forsythe. Reprint of a note on rounding-off errors. *SIAM review*, 1(1):66, 1959. [147](#)
- [108] Marco Gaboardi, Hyun Lim, Ryan Rogers, and Salil Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International conference on machine learning*, pages 2111–2120. PMLR, 2016. [145](#), [197](#)
- [109] Marco Gaboardi, Ryan Rogers, and Or Sheffet. Locally private mean estimation:  $z$ -test and tight confidence intervals. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2545–2554. PMLR, 2019. [197](#)
- [110] Evan Greene and Jon A Wellner. Exponential bounds for the hypergeometric distribution. *Bernoulli*, 23(3):1911, 2017. [15](#), [20](#)
- [111] Peter Grünwald. *The minimum description length principle*. MIT press, 2007. [122](#)
- [112] Peter Grünwald, Alexander Henzi, and Tyron Lardy. Anytime-valid tests of conditional independence under model-x. *Journal of the American Statistical Association*, pages 1–12, 2023. [336](#)
- [113] Peter Grünwald, Rianne de Heide, and Wouter M Koolen. Safe testing. *Journal of the Royal Statistical Society Series B - Methodology (with discussion)*, 2024. [26](#), [50](#), [51](#), [70](#), [72](#), [85](#), [122](#), [176](#), [177](#), [222](#)
- [114] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002. [337](#), [342](#)
- [115] Vitor Hadad, David A Hirshberg, Ruohan Zhan, Stefan Wager, and Susan Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the national academy of sciences*, 118(15):e2014602118, 2021. [205](#)
- [116] Peter Hall and Barbara La Scala. Methodology and algorithms of empirical likelihood. *International Statistical Review*, pages 109–127, 1990. [117](#)
- [117] Dae Woong Ham, Iavor Bojinov, Michael Lindon, and Martin Tingley. Design-based confidence sequences for anytime-valid causal inference. *arXiv preprint arXiv:2210.08639*, 2022. [244](#)
- [118] Sebastian Haneuse and Andrea Rotnitzky. Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine*, 32(30):5260–5277, 2013. [244](#)
- [119] Harrie Hendriks. Test martingales for bounded random variables. *arXiv preprint arXiv:1801.09418*, 2018. [45](#), [50](#)
- [120] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. [6](#), [15](#), [20](#), [22](#), [44](#), [47](#), [52](#), [66](#), [74](#), [98](#), [99](#), [151](#), [152](#), [162](#), [164](#), [167](#), [180](#), [211](#), [254](#)

- [121] Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 2010. [84](#)
- [122] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952. [156](#)
- [123] Steven R Howard and Aaditya Ramdas. Sequential estimation of quantiles with applications to A/B testing and best-arm identification. *Bernoulli*, 28(3):1704–1728, 2022. [3](#), [73](#), [205](#), [225](#), [226](#), [238](#), [247](#), [248](#)
- [124] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. [7](#), [15](#), [21](#), [30](#), [31](#), [40](#), [45](#), [47](#), [48](#), [50](#), [51](#), [52](#), [53](#), [56](#), [70](#), [71](#), [83](#), [86](#), [98](#), [120](#), [121](#), [123](#), [131](#), [153](#), [155](#), [159](#), [164](#), [172](#), [190](#), [216](#), [226](#), [229](#), [234](#), [236](#), [254](#), [280](#), [306](#), [415](#)
- [125] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. [3](#), [4](#), [5](#), [6](#), [15](#), [21](#), [22](#), [23](#), [26](#), [38](#), [40](#), [45](#), [48](#), [50](#), [52](#), [53](#), [54](#), [55](#), [64](#), [70](#), [71](#), [83](#), [86](#), [97](#), [98](#), [107](#), [120](#), [121](#), [123](#), [155](#), [174](#), [175](#), [177](#), [202](#), [205](#), [211](#), [215](#), [216](#), [218](#), [221](#), [226](#), [229](#), [230](#), [232](#), [233](#), [234](#), [236](#), [240](#), [244](#), [247](#), [248](#), [252](#), [253](#), [254](#), [273](#), [275](#), [280](#), [305](#), [312](#), [321](#), [415](#)
- [126] Tien-Chung Hu and Andrew Rosalsky. A note on the de la Vallée Poussin criterion for uniform integrability. *Statistics & probability letters*, 81(1):169–174, 2011. [335](#), [361](#), [381](#), [404](#)
- [127] Ze-Chun Hu and Qian-Qian Zhou. A note on uniform integrability of random variables in a probability space and sublinear expectation space. *Chinese Journal of Applied Probability and Statistics*, 34(6):577–586, 2017. URL <https://arxiv.org/abs/1705.08333>. [381](#), [388](#), [398](#), [410](#)
- [128] Audrey Huang, Liu Leqi, Zachary Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. *Advances in Neural Information Processing Systems*, 34:23714–23726, 2021. [201](#), [203](#), [205](#), [226](#), [227](#)
- [129] Zhuoqun Huang, Ronald L Rivest, Philip B Stark, Vanessa J Teague, and Damjan Vukcevic. A unified evaluation of two-candidate ballot-polling election auditing methods. In *International Joint Conference on Electronic Voting*, pages 112–128. Springer, 2020. [129](#)
- [130] Thomas E Hull and J Richard Swenson. Tests of probabilistic models for propagation of roundoff errors. *Communications of the ACM*, 9(2):108–113, 1966. [147](#)
- [131] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015. [244](#)
- [132] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil’ UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014. [3](#), [217](#)

- [133] Wojciech Jamroga, Peter B Røenne, Peter YA Ryan, and Philip B Stark. Risk-limiting tallies. In *International Joint Conference on Electronic Voting*, pages 183–199. Springer, 2019. [139](#), [141](#)
- [134] Christopher Jennison and Bruce W Turnbull. *Group sequential methods with applications to clinical trials*. CRC Press, 1999. [314](#)
- [135] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017. [3](#), [26](#), [175](#), [250](#)
- [136] Alistair Johnson and Tom Pollard. sepsis3-mimic, May 2018. URL <https://doi.org/10.5281/zenodo.1256723>. [277](#)
- [137] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016. [277](#)
- [138] Matthew Joseph, Janardhan Kulkarni, Jieming Mao, and Steven Z Wu. Locally private Gaussian estimation. *Advances in Neural Information Processing Systems*, 32:2984–2993, 2019. [145](#), [197](#)
- [139] Kwang-Sung Jun and Francesco Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Conference on Learning Theory*, pages 1802–1823. PMLR, 2019. [45](#), [50](#), [71](#), [72](#), [87](#), [122](#), [145](#), [198](#)
- [140] Kwang-Sung Jun, Francesco Orabona, Stephen Wright, and Rebecca Willett. Improved strongly adaptive online learning using coin betting. In *Artificial Intelligence and Statistics*, pages 943–951. PMLR, 2017. [71](#), [87](#), [122](#)
- [141] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Advances in neural information processing systems*, 27, 2014. [147](#)
- [142] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016. [147](#), [148](#)
- [143] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, pages 2204–2235. PMLR, 2020. [145](#)
- [144] Nikos Karampatziakis, Paul Mineiro, and Aaditya Ramdas. Off-policy confidence sequences. *International Conference on Machine Learning*, 2021. [199](#), [200](#), [201](#), [202](#), [203](#), [204](#), [205](#), [206](#), [209](#), [211](#), [218](#), [219](#), [230](#)
- [145] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018. [145](#), [197](#)

- [146] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011. [144](#)
- [147] Maximilian Kasy. Uniformity and the delta method. *Journal of Econometric Methods*, 8(1):20180001, 2018. [325](#), [326](#)
- [148] Masahiro Kato, Takuya Ishihara, Junya Honda, and Yusuke Narita. Adaptive experimental design for efficient treatment effect estimation. *arXiv preprint arXiv:2002.05308*, 2020. [203](#), [220](#)
- [149] Emilie Kaufmann and Wouter Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *arXiv:1811.11419*, 2018. [45](#), [51](#)
- [150] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016. [217](#)
- [151] Michael Kearns and Lawrence Saul. Large deviation methods for approximate probabilistic inference. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI’98, pages 311–319, 1998. [98](#)
- [152] John L Kelly Jr. A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926, 1956. [71](#), [84](#), [121](#), [133](#)
- [153] Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016. [247](#), [264](#), [319](#)
- [154] Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019. [243](#), [244](#)
- [155] Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022. [207](#), [209](#), [338](#)
- [156] Edward H Kennedy, Sivaraman Balakrishnan, and Max G’Sell. Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030, 2020. [287](#), [294](#)
- [157] Koulik Khamaru, Yash Deshpande, Lester Mackey, and Martin J Wainwright. Near-optimal inference in adaptive linear regression. *arXiv preprint arXiv:2107.02266*, 2021. [205](#)
- [158] Aleksandr Khintchine. über einen satz der wahrscheinlichkeitsrechnung. *Fundamenta Mathematicae*, 6(1):9–20, 1924. [5](#)
- [159] Andrei Kolmogorov. Über das gesetz des iterierten logarithmus. *Mathematische Annalen*, 101(1):126–135, 1929. [5](#)

- [160] János Komlós, Péter Major, and Gábor Tusnády. An approximation of partial sums of independent rv's, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32(1-2):111–131, 1975. 8, 12, 251, 253, 321, 327, 344, 372, 417
- [161] János Komlós, Péter Major, and Gábor Tusnády. An approximation of partial sums of independent rv's, and the sample df. ii. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 34(1):33–58, 1976. 8, 12, 251, 253, 321, 327, 344, 372, 417
- [162] Wojciech Kotłowski, Peter Grünwald, and Steven De Rooij. Following the flattened leader. In *Conference on Learning Theory*, pages 106–118. Citeseer, 2010. 54, 72
- [163] Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981. 71, 122
- [164] Arun Kumar Kuchibhotla and Qinqing Zheng. Near-optimal confidence sequences for bounded random variables. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2021. 98, 100, 173
- [165] Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, and Larry Wasserman. Median regularity and honest inference. *Biometrika*, 110(3):831–838, 2023. 326
- [166] Masayuki Kumon, Akimichi Takemura, and Kei Takeuchi. Sequential optimizing strategy in multi-dimensional bounded forecasting games. *Stochastic Processes and their Applications*, 121(1):155–183, 2011. 72, 84
- [167] Tze Leung Lai. Boundary crossing probabilities for sample sums and confidence sequences. *The Annals of Probability*, pages 299–312, 1976. 250, 252
- [168] Tze Leung Lai. On confidence sequences. *The Annals of Statistics*, 4(2):265–280, 1976. 3, 71, 121, 202, 250, 252, 326
- [169] John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007. 199
- [170] Erik Learned-Miller and Philip S Thomas. A new confidence interval for the mean of a bounded random variable. *arXiv preprint arXiv:1905.06208*, 2019. 99
- [171] Ker-Chau Li. Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008, 1989. 325, 326
- [172] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010. 199
- [173] Qiang Jonathan Li. *Estimation of mixture models*. PhD thesis, Yale University, 1999. 122
- [174] Zitao Li, Tianhao Wang, Milan Lopuhaä-Zwakenberg, Ninghui Li, and Boris Škoric. Estimating numerical distributions under local differential privacy. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 621–635, 2020. 147, 148

- [175] M Lifshits. Lecture notes on strong approximation. *Pub. IRMA Lille*, 53(13), 2000. [366](#)
- [176] Jarl Waldemar Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225, 1922. [256](#)
- [177] Mark Lindeman, Philip B Stark, and Vincent S Yates. BRAVO: Ballot-polling risk-limiting audits to verify outcomes. In *EVT/WOTE*, 2012. [125](#), [129](#)
- [178] Gary Lorden and Moshe Pollak. Nonanticipating estimation applied to sequential analysis and changepoint detection. *The Annals of Statistics*, 33(3):1422–1454, June 2005. ISSN 0090-5364, 2168-8966. [53](#)
- [179] Anton Rask Lundborg, Rajen D Shah, and Jonas Peters. Conditional independence testing in Hilbert spaces with applications to functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(5):1821–1850, 2022. [326](#), [331](#)
- [180] Akash Maharaj, Ritwik Sinha, David Arbour, Ian Waudby-Smith, Simon Z Liu, Moumita Sinha, Raghavendra Addanki, Aaditya Ramdas, Manas Garg, and Viswanathan Swaminathan. Anytime-valid confidence sequences in an enterprise a/b testing platform. In *Companion Proceedings of the ACM Web Conference 2023*, pages 396–400, 2023. [12](#)
- [181] Péter Major. The approximation of partial sums of independent rv's. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 35(3):213–220, 1976. [253](#), [321](#), [344](#)
- [182] Ivana Malenica, Yongyi Guo, Kyra Gan, and Stefan Konigorski. Anytime-valid inference in n-of-1 trials. In *Machine Learning for Health (ML4H)*, pages 307–322. PMLR, 2023. [2](#)
- [183] Tudor Manole and Aaditya Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 2023. [290](#), [321](#), [322](#), [323](#)
- [184] J Marcinkiewicz and A Zygmund. Sur les fonctions indépendantes. *Fundamenta Mathematicae*, 29:60–90, 1937. [260](#), [380](#), [385](#), [386](#), [415](#), [417](#)
- [185] Per Martin-Löf. The definition of random sequences. *Information and control*, 9(6):602–619, 1966. [120](#)
- [186] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The annals of Probability*, 18(3):1269–1283, 1990. [227](#)
- [187] Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the Conference on Learning Theory*, 2009. [15](#), [22](#), [45](#), [55](#), [56](#), [98](#), [99](#), [171](#)
- [188] Michael McKerns, Leif Strand, Tim Sullivan, Alta Fang, and Michael Aivazis. Building a framework for predictive science. *Proceedings of the 10th Python in Science Conference*, 2011. [97](#)

- [189] Ilya Mironov. On significance of the least significant bits for differential privacy. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 650–661, 2012. [149](#)
- [190] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, pages 672–679. ACM, 2008. [98](#)
- [191] Gregory Morrow and Walter Philipp. An almost sure invariance principle for hilbert space valued martingales. *Transactions of the American Mathematical Society*, pages 231–251, 1982. [253](#)
- [192] Per Aslak Mykland. Dual likelihood. *The Annals of Statistics*, pages 396–421, 1995. [60](#), [119](#)
- [193] Jerzy Neyman. On the application of probability theory to agricultural experiments, essay on principles, section 9. *Statistical Science*, 5(4):465–472, 1923/1990. [242](#), [244](#)
- [194] Peter C O’Brien and Thomas R Fleming. A multiple testing procedure for clinical trials. *Biometrics*, pages 549–556, 1979. [314](#), [315](#)
- [195] Francesco Orabona and Kwang-Sung Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *arXiv preprint arXiv:2110.14099*, 2021. [171](#), [173](#)
- [196] Francesco Orabona and David Pal. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29:577–585, 2016. [71](#), [122](#)
- [197] Francesco Orabona and Tatiana Tommasi. Training deep networks without learning rates through coin betting. In *Advances in Neural Information Processing Systems*, pages 2160–2170, 2017. [71](#), [72](#), [87](#), [122](#)
- [198] Kellie Ottoboni, Philip B Stark, Mark Lindeman, and Neal McBurnett. Risk-limiting audits by stratified union-intersection tests of elections (SUITE). In *International Joint Conference on Electronic Voting*, pages 174–188. Springer, 2018. [129](#)
- [199] Kellie Ottoboni, Matthew Bernhard, J Alex Halderman, Ronald L Rivest, and Philip B Stark. Bernoulli ballot polling: a manifest improvement for risk-limiting audits. In *International Conference on Financial Cryptography and Data Security*, pages 226–241. Springer, 2019. [125](#), [129](#)
- [200] Art B Owen. *Empirical likelihood*. CRC press, 2001. [60](#), [117](#), [118](#), [119](#)
- [201] Luigi Pace and Alessandra Salvan. Likelihood, replicability and robbins’ confidence sequences. *International Statistical Review*, 88(3):599–615, 2020. [250](#)
- [202] Valentin V. Petrov. *Laws of Large Numbers*, pages 256–291. Springer Berlin Heidelberg, Berlin, Heidelberg, 1975. ISBN 978-3-642-65809-9. doi: 10.1007/978-3-642-65809-9\_9. URL [https://doi.org/10.1007/978-3-642-65809-9\\_9](https://doi.org/10.1007/978-3-642-65809-9_9). [285](#), [286](#), [380](#), [388](#), [389](#), [391](#)

- [203] My Phan, Philip S Thomas, and Erik Learned-Miller. Towards practical mean bounds for small samples. *International Conference on Machine Learning*, 2021. [100](#), [101](#)
- [204] Stuart J Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977. [314](#), [315](#)
- [205] Tom J Pollard and Alistair EW Johnson. The MIMIC-III Clinical Database. <http://dx.doi.org/10.13026/C2XW26>, 2016. [277](#)
- [206] Eric C Polley and Mark J van der Laan. Super learner in prediction. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (222), 2010. [268](#)
- [207] Eric C Polley, Sherri Rose, and Mark J van der Laan. Super learning. In *Targeted learning*, pages 43–66. Springer, 2011. [268](#)
- [208] Alexander Rakhlin and Karthik Sridharan. On equivalence of martingale tail bounds and deterministic regret inequalities. In *Conference on Learning Theory*, pages 1704–1722. PMLR, 2017. [71](#), [72](#), [122](#)
- [209] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020. [48](#), [49](#), [51](#), [60](#), [70](#), [210](#), [321](#)
- [210] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter M Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 2021. [50](#), [61](#), [71](#), [176](#)
- [211] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. [6](#), [207](#), [326](#)
- [212] Pratik Ramprasad, Yuantong Li, Zhuoran Yang, Zhaoran Wang, Will Wei Sun, and Guang Cheng. Online bootstrap inference for policy evaluation in reinforcement learning. *Journal of the American Statistical Association*, 118(544):2901–2914, 2023. [205](#)
- [213] Alessandro Rinaldo, Larry Wasserman, and Max G’Sell. Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438–3469, 2019. [325](#), [326](#)
- [214] Jorma Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory*, 30(4):629–636, 1984. [53](#), [122](#), [123](#)
- [215] Jorma Rissanen. *Stochastic Complexity in Statistical Inquiry*, volume 15. World Scientific, 1998. [122](#), [123](#)
- [216] Ronald L Rivest. ClipAudit: A simple risk-limiting post-election audit. *arXiv preprint arXiv:1701.08312*, 2017. [125](#)

- [217] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970. 3, 5, 6, 48, 51, 71, 121, 155, 159, 164, 190, 215, 247, 252, 254, 256, 259, 260, 262, 263, 273, 275, 280, 303, 306, 320, 326
- [218] Herbert Robbins and David Siegmund. Iterated logarithm inequalities and related statistical procedures. In *Mathematics of the Decision Sciences, Part II*, pages 267–279. American Mathematical Society, Providence, 1968. 3, 48, 51, 71, 121
- [219] Herbert Robbins and David Siegmund. Probability distributions related to the law of the iterated logarithm. *Proc. of the National Academy of Sciences*, 62(1):11–13, January 1969. ISSN 0027-8424. 71, 121
- [220] Herbert Robbins and David Siegmund. Boundary crossing probabilities for the Wiener process and sample sums. *The Annals of Mathematical Statistics*, 41(5):1410–1429, 1970. 8, 51, 71, 121, 252, 260, 262, 263, 326, 328, 330, 371, 372
- [221] Herbert Robbins and David Siegmund. A class of stopping rules for testing parametric hypotheses. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 37–41, 1972. 51, 71, 121
- [222] Herbert Robbins and David Siegmund. The expected sample size of some tests of power one. *The Annals of Statistics*, 2(3):415–436, 1974. 48, 51, 53, 71, 121, 340
- [223] James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986. 243, 244
- [224] James Robins and Aad van der Vaart. Adaptive nonparametric confidence sets. *Annals of statistics*, 34(1):229–253, 2006. 326
- [225] James Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994. 156, 207, 208, 247
- [226] James Robins, Lingling Li, Eric Tchetgen Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008. 264, 265, 319, 320
- [227] Andrea Rotnitzky, James Robins, and Daniel O Scharfstein. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339, 1998. 265
- [228] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974. 242
- [229] Johannes Ruf, Martin Larsson, Wouter M Koolen, and Aaditya Ramdas. A composite generalization of Ville’s martingale theorem. *Electronic Journal of Probability*, 2023. 379, 381, 417

- [230] Aleksandr Ivanovich Sakhnenko. Estimates in an invariance principle. *Matematicheskie Trudy*, 5:27–44, 1985. 366
- [231] Daniel O Scharfstein, Andrea Rotnitzky, and James Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999. 265
- [232] Robert J Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48, 1974. 9, 15, 20, 45, 66
- [233] Shalev Shaer, Gal Maman, and Yaniv Romano. Model-X sequential testing for conditional independence via testing by betting. In *International Conference on Artificial Intelligence and Statistics*, pages 2054–2086. PMLR, 2023. 336
- [234] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021. 45, 51, 72, 177, 222
- [235] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* John Wiley & Sons, February 2001. ISBN 978-0-471-46171-5. 45, 58, 70, 86, 120, 122
- [236] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, 2019. 26, 45, 70, 120, 122
- [237] Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011. 50, 159, 222
- [238] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. 13, 326, 328, 331, 336, 337, 338, 339, 340, 341, 342, 356, 360
- [239] Zach Shahn, Nathan I Shapiro, Patrick D Tyler, Daniel Talmor, and H Lehman Li-wei. Fluid-limiting treatment strategies among sepsis patients in the icu: a retrospective causal analysis. *Critical Care*, 24(1):1–9, 2020. 277
- [240] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. 121
- [241] Ye Shen, Hengrui Cai, and Rui Song. Doubly robust interval estimation for optimal policy evaluation in online learning. *arXiv preprint arXiv:2110.15501*, 2021. 205
- [242] Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016. 277
- [243] Anatoliy Skorokhod. Research on the theory of random processes. *Kiev Univ*, 1961. 253, 344

- [244] Philip B Stark. CAST: Canvass audits by sampling and testing. *IEEE Transactions on Information Forensics and Security*, 4(4):708–717, 2009. [129](#)
- [245] Philip B Stark. Risk-limiting postelection audits: Conservative  $p$ -values from common probability inequalities. *IEEE Transactions on Information Forensics and Security*, 4(4):1005–1014, 2009. [129](#)
- [246] Philip B Stark. Sets of half-average nulls generate risk-limiting audits: SHANGRLA. In *International Conference on Financial Cryptography and Data Security*, pages 319–336. Springer, 2020. [10](#), [45](#), [66](#), [125](#), [127](#), [129](#), [130](#), [135](#), [141](#)
- [247] Philip B Stark et al. Conservative statistical post-election audits. *The Annals of Applied Statistics*, 2(2):550–581, 2008. [129](#)
- [248] Volker Strassen. An invariance principle for the law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 3(3):211–226, 1964. [8](#), [12](#), [252](#), [253](#), [256](#), [258](#), [278](#), [279](#), [310](#), [321](#), [327](#), [343](#), [344](#)
- [249] Volker Strassen. Almost sure behavior of sums of independent random variables and martingales. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 3, page 315. University of California Press, 1967. [8](#), [12](#), [253](#), [258](#), [262](#), [282](#), [283](#), [297](#), [298](#), [343](#), [344](#)
- [250] Terence Tao. E pluribus unum: from complexity, universality. *Daedalus*, 141(3):23–34, 2012. [249](#)
- [251] Judith ter Schure and Peter Grünwald. All-in meta-analysis: breathing life into living systematic reviews. *arXiv preprint arXiv:2109.12141*, 2021. [177](#), [222](#)
- [252] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. [199](#), [201](#), [203](#), [211](#), [213](#), [214](#), [218](#)
- [253] Ryan J Tibshirani, Alessandro Rinaldo, Rob Tibshirani, and Larry Wasserman. Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, 46(3):1255–1287, 2018. [325](#), [326](#)
- [254] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007. [272](#), [319](#)
- [255] Alexandre B Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003. [268](#), [343](#)
- [256] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022. [207](#), [209](#)
- [257] Mark J van der Laan and James Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003. [319](#)

- [258] Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011. 208, 247
- [259] Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007. 268
- [260] Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 247, 331, 348, 376
- [261] Aad W van der Vaart. Semiparametric statistics. *Lecture Notes in Math.*, 2002. 319, 321
- [262] Aad W van der Vaart, Sandrine Dudoit, and Mark J van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics and Decisions*, 24(3):351–371, 2006. 268
- [263] Carl van Walraven, Peter C Austin, Alison Jennings, Hude Quan, and Alan J Forster. A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Medical care*, pages 626–633, 2009. 277
- [264] Jean Ville. *Étude Critique de la Notion de Collectif*. PhD thesis, Paris, 1939. 6, 9, 30, 31, 48, 50, 52, 59, 70, 120, 131, 159, 162, 165, 172, 182, 187, 197, 207, 209, 212, 231, 237, 280, 281, 320
- [265] Vladimir Vovk. Testing randomness online. *Statistical Science*, 36(4):595–611, 2021. 51
- [266] Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. 176, 177, 222
- [267] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005. 50
- [268] Vladimir G Vovk. A logic of probability, with application to the foundations of statistics. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):317–341, 1993. 70
- [269] Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009. 145, 197
- [270] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. 326
- [271] Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of mathematical statistics*, 16(2):117–186, 1945. 25, 48, 121, 198, 326
- [272] Abraham Wald. *Sequential Analysis*. John Wiley & Sons, New York, 1947. 53, 70
- [273] Hongjian Wang and Aaditya Ramdas. Catoni-style confidence sequences for heavy-tailed mean estimation. *Stochastic Processes and their Applications*, 2023. 247, 248

- [274] Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649. IEEE, 2019. [145](#)
- [275] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *arXiv preprint arXiv:2009.02824*, 2020. [177](#), [222](#), [228](#)
- [276] Yu Wang, Hussein Sibai, Mark Yen, Sayan Mitra, and Geir E Dullerud. Differentially private algorithms for statistical verification of cyber-physical systems. *IEEE Open Journal of Control Systems*, 1:294–305, 2022. [145](#), [198](#)
- [277] Yue Wang, Jaewoo Lee, and Daniel Kifer. Differentially private hypothesis testing, revisited. *arXiv preprint arXiv:1511.03376*, 2015. [145](#), [197](#)
- [278] Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. [11](#), [144](#), [146](#), [148](#), [228](#)
- [279] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. [145](#), [197](#)
- [280] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. [51](#), [71](#), [248](#)
- [281] Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. *Advances in Neural Information Processing Systems*, 33, 2020. [9](#), [66](#), [83](#), [177](#), [222](#)
- [282] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1): 1–27, 2024. [10](#), [155](#), [229](#), [247](#), [248](#)
- [283] Ian Waudby-Smith, Philip B Stark, and Aaditya Ramdas. RiLACS: risk limiting audits via confidence sequences. In *Electronic Voting: 6th International Joint Conference, E-Vote-ID 2021, Virtual Event, October 5–8, 2021, Proceedings 6*, pages 124–139. Springer, 2021. [10](#)
- [284] Ian Waudby-Smith, Lili Wu, Aaditya Ramdas, Nikos Karampatziakis, and Paul Mineiro. Anytime-valid off-policy inference for contextual bandits. *ACM/JMS Journal of Data Science*, 2022. [11](#)
- [285] Ian Waudby-Smith, Edward H. Kennedy, and Aaditya Ramdas. Distribution-uniform anytime-valid inference. *arXiv preprint arXiv:2311.03343*, 2023. [13](#)
- [286] Ian Waudby-Smith, Zhiwei Steven Wu, and Aaditya Ramdas. Nonparametric extensions of randomized response for private confidence sets. *International Conference on Machine Learning*, 202:36748–36789, 2023. [11](#)

- [287] Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H Kennedy, and Aaditya Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *The Annals of Statistics (to appear)*, 2024. 12
- [288] Ian Waudby-Smith, Martin Larsson, and Aaditya Ramdas. Distribution-uniform strong laws of large numbers. *arXiv preprint arXiv:2402.00713*, 2024. 13, 333
- [289] Xiao Wu, Kate R Weinberger, Gregory A Wellenius, Francesca Dominici, and Danielle Braun. Assessing the causal effects of a stochastic intervention in time series data: are heat alerts effective in preventing deaths and hospitalizations? *Biostatistics*, 25(1):57–79, 2024. 2
- [290] Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. Post-selection inference for e-value based confidence intervals. *arXiv preprint arXiv:2203.12572*, 2022. 228
- [291] Miao Yu, Wenbin Lu, and Rui Song. A new framework for online testing of heterogeneous treatment effect. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10310–10317, 2020. 250
- [292] Ruohan Zhan, Vitor Hadad, David A Hirshberg, and Susan Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2125–2135, 2021. 201, 205, 218
- [293] Kelly Zhang, Lucas Janson, and Susan Murphy. Inference for batched bandits. *Advances in neural information processing systems*, 33:9818–9829, 2020. 205
- [294] Kelly Zhang, Lucas Janson, and Susan Murphy. Statistical inference with m-estimators on adaptively collected data. *Advances in neural information processing systems*, 34: 7460–7471, 2021. 205
- [295] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. Adaptive concentration inequalities for sequential decision problems. *Advances in Neural Information Processing Systems*, 29, 2016. 98, 177
- [296] Wenjing Zheng and Mark J van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, (273), 2010. 264, 265, 319, 320
- [297] Julian Zimmert and Tor Lattimore. Connections between mirror descent, Thompson sampling and the information ratio. *Advances in Neural Information Processing Systems*, 32, 2019. 211
- [298] Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. 211