PH.D. THESIS

Comparing Forecasters and Abstaining Classifiers

Yo Joong Choe

June 15, 2023

Department of Statistics and Data Science
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA

Thesis Committee:
Aaditya Ramdas (Chair)
Aarti Singh
Edward H. Kennedy
Johanna F. Ziegel (University of Bern)
Alexander D'Amour (Google DeepMind)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics and Machine Learning.

Keywords: evaluation, black-box machine learning, nonparametric statistics, anytime-valid inference, game-theoretic statistics, confidence sequences, e-processes, sequential inference, forecasting, proper scoring rules, abstaining classifiers, reliable machine learning, missing data, causal inference, double robustness

For my parents and my partner

Abstract

This thesis concerns nonparametric statistical methods for comparing black-box predictors, namely sequential forecasters and abstaining classifiers.

In the first part of the thesis, we develop anytime-valid estimation and testing approaches for comparing probability forecasters on sequentially occurring events. Our main contribution is the development of confidence sequences (CS) that estimate the time-varying average score difference between the forecasters. Unlike classical confidence intervals, CSs can be continuously monitored over time while retaining their coverage guarantees. The CSs also do not require any distributional assumptions on the dynamics of the outcomes or on the forecasting models. We additionally develop e-processes and p-processes, which are testing counterparts to CSs that are anytime-valid, i.e., valid at any data-dependent stopping times.

In the second part of the thesis, we consider the problem of evaluating and comparing black-box abstaining classifiers. Abstaining classifiers have the option to withhold predictions on inputs that they are uncertain about, making them increasingly popular in safety-critical applications. We introduce a novel approach and perspective to the evaluation problem by treating the abstentions of a classifier as missing data. Our approach is centered around defining the counterfactual score, which measures the expected performance of the classifier had it not been allowed to abstain. The missing data perspective clarifies the precise identifying conditions for the counterfactual score, requiring independent evaluation data and stochastic abstentions, and paves the way for a nonparametrically efficient and doubly robust estimator for the score. The approach also straightforwardly extends to estimating the difference in two counterfactual scores under distinct abstention mechanisms.

Acknowledgements

My journey through the Ph.D. program involved a highly nonlinear path, and I would not have been able to go through it without the help of many people in the process.

First and foremost, I express my heartfelt gratitude to Aaditya Ramdas, who advised this thesis. I was fortunate to have met such a passionate, caring, and relentlessly positive mentor, especially when I needed structure and guidance after coming back from a long leave. Aaditya introduced me to the exciting field of game-theoretic statistics, and he was patient throughout my switch from applied to theoretical research. His mentorship has directly shaped my identity as a researcher, in everything from formulating new ideas, rigorously deriving proofs, and meticulously editing drafts to maintaining a positive and healthy mindset in research and life.

Next, I express my sincere appreciation to Aarti Singh, who served as my advisor during the early parts of my Ph.D. program and as my ML mentor during my thesis work. Aarti is a great role model, not only as a leader but as an encouraging, gracious, and wise mentor. I benefited immensely from her unwavering support throughout the program, including the ADA project before my leave, the materials science project upon return, and the gradual switch to my thesis project with Aaditya.

I am truly grateful to my thesis committee members, Edward Kennedy, Johanna Ziegel, and Alexander D'Amour, along with Aaditya and Aarti. I could not have gotten a better group of leading researchers in the relevant fields of my research. Every committee member gave constructive feedback and valuable suggestions, many of which are reflected in this thesis.

I also extend my gratitude to my ADA co-advisors, Sivaraman Balakrishnan and Timothy Verstynen, for their helpful support during the early part of my Ph.D. program. Along with Aarti, Siva and Tim shaped my research experience as mentors for my first publication. I learned a lot in my experiences handling challenging neuroscience datasets and carving out meaningful scientific questions, and the skillset has proven to be useful in my later endeavors during and after the leave.

After the first two years in the program, I had to take an unusually long leave from the program, but there were many people in the StatDS and ML departments who made my leave and return seamless. I thank Chad Schafer for understanding my situation and helping with my leave of absence;

Alessandro Rinaldo for accepting my (late) request to return; and Ann Lee and Valerie Ventura for helping with the return amidst an ongoing pandemic. I also thank Christopher Genovese and Rebecca Nugent for overseeing these requests as the StatDS department heads during their respective times, and Danielle Hamilton and Diane Stidle for handling the administrative matters.

At CMU, I also met amazing mentors and colleagues that helped with my learning and research experiences. I thank the students and postdocs of Aaditya's group (2021-2023): Ian Waudby-Smith, Sasha Podkopaev, Chirag Gupta, Neil Xu, Aditya Gangrade, Shubhanshu Shekhar, Ojash Neopane, Justin Whitehouse, James Leiner, Ben Chugg, and Hongjian Wang. The group had a great synergy in spearheading new advances in anytime-valid inference, and I have been lucky to be a part of the process. A special thanks to Aditya Gangrade for his mentorship and collaboration on Chapter 4.

I would like to further thank members of the StatML reading group, including faculty members Larry Wasserman, Ryan Tibshirani, Arun Kuchibhotla, Aaditya, Siva, and Ale, as well as all the student contributors, for educating me and others about current research topics in this exciting field. I also learned a lot from the courses that many of the faculty members taught during my first two years.

One of the few good things about taking a long leave was that I got to be part of two cohorts, both of whom were instrumental in my life as a graduate student. First, I thank the entire 2015 StatDS cohort — Kayla Frisoli, Mike Stanley, Neil Spencer, Ilmun Kim, Jaehyeok Shin, Xiao Hui Tai, Alden Green, Daren Wang, and Richard Wang — with whom I had such great memories, including the "wild wave," the wine tasting nights, the Pirates and Penguins games, the countless dinners in FMS, and many more. I am also thankful for all the joyful moments with other friends and colleagues I met during 2015-2017, including the StatDS friends, Sangwon (Justin), Jisu, Yotam, Collin, Kevin, Bryan, and many others, as well as the MLD/SCS friends, Maruan, Jay-Yoon, Lisa, Xun, Simon, and more.

After my return, I met new (and old) friends who greatly helped me relieve the stress of handling both thesis research and job search. I am thankful for the rotating hangout crew of Mike, Ian, Sasha, Alec, Nick, Luca, Neil, Maya, Lucas, Kayla, Alden, Maria, Ben, Shannon, Todd, Kate, Spyros, and Alexa. A special thanks to Sasha, Ian, Neil, Luca, and Chirag, as well as my office mates, Mike, Alec, and Nick, for listening to my job market stories (again and again). Another set of thanks goes to the new Korean StatDS group, Beomjo, Heejong, and Woonyoung, who also had to listen to those stories (and, more importantly, are wonderful friends). I also appreciate my new MLD colleagues,

Youngseog, Tom, Jeremy, Ojash, Euxhen, Chris, and others, for the friendly banters during the few times I showed up to MLD events. Last but not least, I would like to thank some of my old friends, Jun Ho, Chanwook, and Jun Won from high school and Junhyung, Youngsoo, Sangwoo, and Min Jae from college, who helped me throughout the uncertain times in my final year.

I want to further extend my appreciation to all my colleagues, friends, and mentors at Kakao and Kakao Brain for helping me continue to grow as a researcher during my leave.

Finally, I dedicate this thesis to my parents and my partner, who are my biggest supporters. I am always grateful to my parents for their unconditional love and support, even as I continue on with my journey across the ocean from home. And none of this work would have been possible without Minji, my partner, wife, and best friend, who stood by me every step of the way, trusted in my decisions even when they did not work out, and made me laugh every day, even as she has been working toward her own doctorate.

Contents

Chapter 1

Introduction

1.1 Problem Statement

As more black-box machine learning (ML) predictors become readily available across domains, the practitioner faces the task of choosing between these predictors by estimating their performance differences on a desired use case. Despite the focus on (small) accuracy increases among ML researchers, it is often unclear whether these improvements will translate to better performance in the practitioner's actual use case. In particular, the black-box nature of many ML predictors, whether it is because they are expensive to train and uninterpretable, or because their training data is proprietary, only adds to the practitioner's challenge of figuring out which predictor is the most useful and reliable for them. This motivates us to develop principled answers to the following general question:

Given (a pair of) black-box predictors, test data, and a scoring rule,
how do we compare their expected scores on the test distribution,
while accurately accounting for the sampling uncertainty of the test data?

This problem is well-studied in a standard setup where the predictors each give their predictions on an independent and identically distributed (i.i.d.) test data, of some fixed sample size n, and the evaluation is done once. However, modern applications of ML introduce new challenging scenarios that differ from this standard setup in one or more ways. One example is the case of comparing *sequential* predictors, where two forecasters make predictions on sequentially occurring events, such as weather outcomes, and the practitioner seeks to continuously monitor the forecasters' scores. An-

other example is the case of comparing *abstaining* predictors, where each predictor may occasionally withhold its predictions according to an unknown abstention mechanism. In both cases, standard methods of comparison are no longer readily applicable. In this thesis, we seek to develop statistically rigorous methods for comparing their expected prediction scores, without requiring minimal assumptions about the underlying data distributions or the predictors.

As we will illustrate throughout the thesis, the departure from the standard i.i.d. setup introduces new conceptual and technical challenges that can be addressed by ideas from seemingly disparate topics, such as game-theoretic statistics, anytime-valid inference, missing data, and causal inference. A complementary goal of this thesis is to make these connections clear within each problem context.

We now proceed with a brief overview of the main contributions in this thesis to the problem of comparing black-box predictors.

1.2 Overview of Contributions

Figure 1.1 places the main contributions of this thesis in the broader context of *nonparametric methods for comparing black-box predictors*. Our focus here is on confidence intervals (CI) that can estimate the expected (average) score difference between the two predictors, according to a wide range of scoring rules; hypothesis tests for the score differences are also mentioned when relevant. We categorize various comparison settings into groups using two characteristics:

- *Evaluation data:* i.i.d. data, sequentially observed data, and potentially missing evaluation data due to abstentions;
- *Anytime-validity:* whether or not the methods are valid at arbitrary data-dependent stopping times. For confidence sequences, this is equivalent to the *time-uniform* guarantee, i.e., that the coverage guarantee holds at all (fixed or random) times (Howard et al., 2021). This is explained in greater detail in Chapter 2.

The **i.i.d. evaluation** setting is completely standard, at least for a fixed sample size. Once we choose a scoring rule S (say, the accuracy or the Brier score for classifiers and the squared error for regressors), the problem reduces to that of i.i.d. mean comparison using paired samples (of i.i.d. scores). This means that the central limit theorem (CLT) can be applied to the score differences to yield asymp-

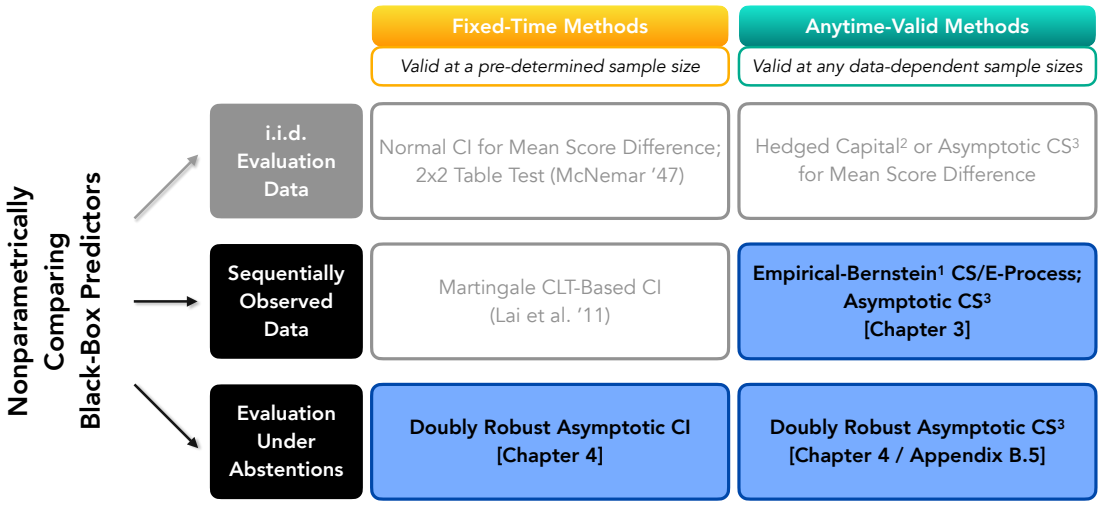


Figure 1.1: Overview of the main contributions in this thesis, in the context of nonparametric methods for comparing black-box predictors. Novel contributions in this thesis are highlighted in bold and blue cells. See text for an explanation of each cell. [1]Howard et al. (2021); [2]Waudby-Smith and Ramdas (2023); [3]Waudby-Smith et al. (2021).

totic CIs that can estimate the score difference. Analogously, any statistical tests of paired difference, such as the paired t-test/z-test, are applicable. There also exist specialized methods for certain setups, such as McNemar (1947)'s test for 2x2 contingency tables (when comparing binary classifiers) and paired t-tests using K-fold cross-validation (when comparing *learning algorithms* and not just their predictions). See Dietterich (1998) for a review of these methods.

To achieve anytime-validity in the i.i.d. evaluation setting, we can apply any time-uniform confidence sequence (CS) for i.i.d. means. For example, if the scores are bounded in the setup, then any CS for a bounded i.i.d. mean, such as the empirical-Bernstein (EB) CS (Howard et al., 2021) and the hedged capital CS (Waudby-Smith and Ramdas, 2023), can tightly estimate the mean difference at data-dependent sample sizes. If the scores are assumed to have at least $m > 2$ moments (bounded or not), then the asymptotic CS (Waudby-Smith et al., 2021) can tightly estimate the mean score difference (while trading off finite-sample validity).

The rest of the thesis primarily concerns challenging settings beyond the simple i.i.d. evaluation

setup. In Chapter 3, we focus on comparing sequential forecasters, i.e., black-box probabilistic predictors on **sequentially occurring events**, such as outcomes in meteorology, sports, and economics. Previously, Lai et al. (2011) derived fixed-time CIs for the time-varying mean score difference, where the expectation is taken term-by-term given all available information at the time of prediction. In our work, we introduce tight CSs for this time-varying average score difference that are valid under *continuous monitoring*, which is a common scenario when comparing time series forecasters, as well as at data-dependent stopping times. The validity of the CSs, like Lai et al. (2011)'s fixed-time CI, does not rely on stationarity (Diebold and Mariano, 1995) or restrictive modeling (Giacomini and White, 2006) assumptions. Our results include both the (nonasymptotic) EB CS and the asymptotic CS.

In Chapter 4, we focus on comparing black-box predictors **under abstentions**, i.e., where each predictor can selectively withhold their predictions given each evaluation input. We first propose a novel evaluation metric called the *counterfactual score*, defined as the expected score of an abstaining predictor *had it not been allowed to abstain*. We then show how the problem of comparing abstaining predictors w.r.t. the counterfactual score reduces to the problem of evaluating each predictor under missing-at-random predictions under Rubin (1974)'s missing data framework. Note that, for this score, standard methods for estimating i.i.d. means are no longer applicable, even when the evaluation set itself is i.i.d., because we do not observe predictions on any abstentions. Given that the evaluation metric itself is novel, we first develop an asymptotically valid CI for a fixed sample size, and we later extend our results to an asymptotic CS (Appendix B.5).

Before delving into these contributions, in Chapter 2, we also include an exposition of anytime-valid inference methods, with a focus on their game-theoretic formulations and their applications to sequential inference involving time-varying means. The chapter serves as a prelude to Chapter 3.

1.3 Bibliographical Notes

The main chapters of this thesis are based on the author's recent preprints:

- Chapter 3 is based on Choe and Ramdas (2021), a joint work with Aaditya Ramdas. This work is currently under revision in a journal.
- Chapter 4 is based on Choe et al. (2023), a joint work with Aditya Gangrade and Aaditya Ram-

das. This work is currently in submission for review at a conference.

During the early part of his Ph.D., the author also published an applied work on high-dimensional correlation analysis using neuroscience data, demonstrating the structure-function relationship between the local white matter and the functional connectivity of the human brain (Choe et al., 2018). This is joint work with Sivaraman Balakrishnan, Aarti Singh, Jean Vettel, and Timothy Verstynen, and it was published in the Proceedings of the 2018 IEEE Conference on Systems, Man, and Cybernetics (SMC). The work is excluded from this thesis in favor of topical coherence.

Chapter 2

A Prelude on Game-Theoretic Statistics and Anytime-Valid Sequential Inference

Before jumping into the main chapters of this thesis, we give a selective exposition of the core tools in anytime-valid inference, including e-processes and confidence sequences (CS). We specifically focus on their uses in sequential settings involving time-varying means as well as their game-theoretic interpretations, both directly relevant to Chapter 3. For a comprehensive survey, see Ramdas et al. (2022a); other key references include Ville (1939); Wald (1945); Darling and Robbins (1967, 1968); Robbins and Siegmund (1970); Robbins (1970); Dawid and Vovk (1999); Shafer and Vovk (2005); Shafer et al. (2011); Balsubramani and Ramdas (2016); Johari et al. (2022); Shafer and Vovk (2019); Shafer (2021); Grünwald et al. (2019); Vovk and Wang (2021); Wasserman et al. (2020); Howard et al. (2020, 2021); Waudby-Smith and Ramdas (2023); Ramdas et al. (2020, 2022b); Ruf et al. (2022).

We note that game-theoretic (betting) ideas are prevalent in fields beyond probability theory and statistical inference, namely in information theory (Kelly, 1956; Krichevsky and Trofimov, 1981; Cover, 1974, 1991) and online learning (Cesa-Bianchi and Lugosi, 2006; Orabona and Pál, 2016; Rakhlin and Sridharan, 2017; Jun and Orabona, 2019). We refer the reader to Waudby-Smith and Ramdas (2023, Section 6 and Appendix F) for a summary of how betting ideas have been utilized in these fields. In

	Fixed-time	**Anytime-valid**
Testing (Quantifying evidence)	p-values	e-processes & test supermartingales; p-processes
Estimation (Quantifying uncertainty)	confidence intervals	confidence sequences

Table 2.1: A brief comparison of tools for fixed-time and anytime-valid statistical inference.

our exposition, we focus on concepts directly relevant to statistical inference, which is the primary goal of this thesis.

In Table 2.1 (right column), we summarize the key methods and objects of anytime-valid inference, categorized by their uses in statistical inference. E-processes, test supermartingales, and p-processes are anytime-valid methods applicable to hypothesis testing as measures of evidence, whereas confidence sequences can be used to estimate a (possibly time-varying) parameter while accounting for the sampling uncertainty. These contrast with their "fixed-time" counterparts (left column), i.e., p-values for testing and confidence intervals (CI) for estimation, whose validity is restricted to a fixed, pre-specified sample size.

It is worth pointing out that the game-theoretic view is not merely an interpretation. Game-theoretic probability, in the senses of Shafer and Vovk (2005, 2019), should be viewed as an alternative to (and a generalization of) measure-theoretic probability itself. Game-theoretic statistics (Ramdas et al., 2022a) then builds upon the ideas and intuitions from game-theoretic probability and applies them to statistical inference problems, i.e., hypothesis testing and parameter estimation. Although our general approach here is to define concepts using measure-theoretic probability and intuit them using their game-theoretic equivalents, we note that a fully game-theoretic exposition is also possible. Such connections have been made precise in, e.g., Ville (1939); Dawid and Vovk (1999); Shafer and Vovk (2005, 2019); Ruf et al. (2022).

2.1 Test Supermartingales and Testing by Betting

Test Supermartingales. The theory of martingales and its interpretation as a gambler's wealth in a betting game are often the starting point for deriving anytime-valid methods (although we will shortly

see that it is the e-process that is the central object). To set the stage, let $(\mathcal{X}, \mathcal{G})$ be a measurable space, and let $\mathfrak{G} = (\mathcal{G}_t)_{t\geq 0}$ be a filtration where $\mathcal{G}_0 = \{\emptyset, \mathcal{X}\}$ and $\mathcal{G}_t$ represents the accumulated information up to time $t \geq 0$. Hereafter, we restrict ourselves to a discrete time scale ($t = 0, 1, 2, \dots$). A stochastic process $(X_t)_{t\geq 0}$ is *adapted* to (or *non-anticipating* w.r.t.) $\mathfrak{G}$ if X_t is $\mathcal{G}_t$-measurable for each $t \geq 0$; it is further *predictable* if X_t is $\mathcal{G}_{t-1}$-measurable for each $t \geq 1$.

Let P denote a probability distribution on $(\mathcal{X}, \mathcal{G})$. In the context of hypothesis testing, P represents a point null hypothesis. (In the next section, we will generalize this to composite null testing for a family of distributions $\mathcal{P}$.) An adapted and integrable process $(X_t)_{t\geq 0}$ is a *P-supermartingale* if

$$\mathbb{E}_P\left[X_t \mid \mathcal{G}_{t-1}\right] \leq X_{t-1}, \quad \forall t \geq 1. \tag{2.1}$$

$(X_t)_{t\geq 0}$ is a *P-martingale* if it the inequality is replaced with an equality. If a P-(super)martingale is nonnegative P-almost surely, then it is a *nonnegative P-(super)martingale*. Finally, a nonnegative P-(super)martingale $(L_t)_{t\geq 0}$ with initial value one, i.e., $L_0 = 1$, is called a **test (super)martingale for** P. Any nonnegative P-(super)martingale can be rescaled by $1/L_0$ into a test (super)martingale for P.

Game-Theoretic Interpretation of Test Supermartingales. Test (super)martingales have an important game-theoretic interpretation[1]: a test (super)martingale for a probability P corresponds to a gambler's wealth against a casino who proposes P as a bet. If P correctly describes the bet's outcome distribution, then a test supermartingale for P describes a *betting strategy* with which the wealth does not increase over time in expectation under P. (A test martingale for P describes one with which the wealth stays constant in expectation.) The connection to hypothesis testing is then immediate. If the null hypothesis P correctly describes the data, then a test supermartingale for P is not expected to grow large; conversely, if a test supermartingale for P grows large, then we can discredit P.

To give a concrete example, suppose that a casino table offers a game that costs $\mu \in [0, 1]$ dollars to enter. At each round $t = 1, 2, \dots$ of the game, after a bet is made, a biased coin is flipped, and the bettor receives a payoff Y_t of \$1 if the coin turns heads and \$0 otherwise.[2] A gambler, also known as

[1]The relevance of martingales to betting games is sometimes discussed in probability courses, but it is not often formalized in the game-theoretic statistical framework.

[2]Shafer and Vovk (2019, Section 1.3) notes that the price μ can be referred to as a *probability* for $Y_t = 1$, "because it invites Skeptic [the Gambler] to bet on this outcome at odds $\mu : 1 - \mu$." The price of this game is in fact the *definition* of probability in subjective probability theory (de Finetti, 1970).

the *skeptic*, sits at the table with \$1 and places a fraction of their money at each round. We formalize this game in Game 2.1:

Game 2.1 (Testing a probability by betting)**.**

Players: Casino and Gambler

Protocol:

1. Casino announces the price of the game, $\mu \in [0, 1]$.
2. Gambler enters the game with an initial wealth of $L_0 = 1$.
3. For rounds $t = 1, 2, \ldots$:
   (a) Gambler chooses the amount of bet $\lambda_t \in [-\frac{1}{1-\mu}, \frac{1}{\mu}]$.
   (b) Casino reveals the payoff $Y_t \in \{0, 1\}$.
   (c) Gambler's wealth is updated according to the payoff as follows:

$$L_t = L_{t-1} \cdot \{1 + \lambda_t(Y_t - \mu)\}. \tag{2.2}$$

Result: Gambler wins if L_t grows large; Casino wins otherwise.

The payoff function (2.2) says that the gambler is placing bets on the difference $(Y_t - \mu)$ between the realized payoff and the price. Thus, if μ correctly describes the probability of the payoffs, then the gambler is not expected to make or lose money in the long run. On the other hand, if μ underestimates the probability of the payoffs being \$1, then the gambler can increase their wealth by placing positive bets ($\lambda_t > 0$). The fact that the gambler has a strategy to grow their wealth suggests that the price μ does not adequately describe the probability of actual payoffs being one, and the gambler's wealth can thus be a measure of evidence against the proposed probability (in the form of the price μ).

We can now translate this game into the measure-theoretic framework by defining the appropriate filtration. The filtration $\mathfrak{G} = (\mathcal{G}_t)_{t\geq 0}$ corresponding to this game is defined as:

$$\mathcal{G}_{t-1} = \text{all available information up to round } t-1 \textit{ and} \text{ the gambler's bet } \lambda_t.$$

Under this filtration, the gambler's wealth (2.2) is a test martingale for μ, in the sense that $\mathbb{E}_P[L_t \mid \mathcal{G}_{t-1}] = L_{t-1}$ for each $t \geq 1$, where P satisfies $P(Y_t = 1 \mid \mathcal{G}_{t-1}) = \mu$ for each $t \geq 1$. The bounds on λ_t

are given to ensure that the gambler cannot bet more than the wealth they have at any given round, such that $L_t \geq 0$ for each round t. Note that, if we restrict the gambler to make only nonnegative bets ($\lambda_t \geq 0$), then $(L_t)_{t\geq 0}$ is also a test *super*martingale for any P that satisfies $P(Y_t = 1 \mid \mathcal{G}_{t-1}) \leq \mu$, $\forall t$. We also remark that everything generalizes straightforwardly to the case where the payoff is now a continuous variable in $[0, 1]$ and μ models the conditional mean $\mathbb{E}_P[Y_t \mid \mathcal{G}_{t-1}]$. Finally, see Waudby-Smith and Ramdas (2023) for ways to effectively choose the betting strategy $(\lambda_t)_{t\geq 0}$ for Game 2.1.

The betting interpretation is "as old as probability itself," and yet it provides a scientifically meaningful notion of evidence in hypothesis testing: if P describes a null hypothesis of a proposed test, then the gambler's wealth expressed as test supermartingale for P quantifies the amount of evidence against the null. This interpretation is the basis for the testing-by-betting framework proposed by Shafer (2021), who refers to the gambler's wealth as the *betting score*.

***A Clarifying Note on the Term *Forecaster*.** In Shafer and Vovk (2019)'s parlance, Casino in Game 2.1 plays both the role of a Forecaster, who announces the price/probability of the game (μ), and Reality, who decides and reveals the outcome (Y_t). Then, Forecaster is a separate entity from Reality that merely gives a hypothesis for the (unknown) probability of the biased coin in the game. This use of the term *Forecaster* is consistent with the term's usage in the literature of subjective probability theory and forecast evaluation; nevertheless, in Chapters 1 and 3, we reserve the term *forecaster* to only refer to an entity that produces probabilistic predictions for future outcomes in a sequence of events (synonymous to a "prophet"). This choice is mainly due to the fact that, when *comparing* multiple forecasters, each forecaster alone does not fully describe the hypothesis being tested by the gambler (e.g., that one forecaster outperforms the other). An alternative terminology choice would have been to call these forecasters as *Predictors*, and Forecaster would be the one that proposes a hypothesis describing the behaviors of any involved Predictor as well as Reality. See Shafer and Vovk (2019, Chapter 12.9) for further clarification.

2.2 E-Processes, Anytime-Validity, and Composite Nulls

E-Processes and Anytime-Validity. The central object of anytime-valid inference is the e-process (Ramdas et al., 2022b), which is a sequence of nonnegative random variables that are, under the

(possibly composite) null hypothesis, bounded by one in expectation at arbitrary stopping times. Designing useful e-processes is the key to deriving modern anytime-valid procedures.

To define an e-process, first recall that a *stopping time* τ (w.r.t. $\mathfrak{G}$) is a random variable that takes values in $\mathbb{N} \cup \{\infty\}$ and satisfies $\{\tau = t\} \in \mathcal{G}_t$ for each $t \geq 0$. Then, given a family of distributions $\mathcal{P}$, a nonnegative adapted process $(E_t)_{t\geq 0}$ with $E_0 = 1$ is defined as an **e-process** for $\mathcal{P}$ if

$$\text{for any } P \in \mathcal{P} \text{ and any arbitrary stopping time } \tau, \quad \mathbb{E}_P[E_\tau] \leq 1, \tag{2.3}$$

where we take $E_\infty = \limsup_{t\to\infty} E_t$ for infinite stopping times. The term 'process' is used to emphasize the fact that the condition $\mathbb{E}_P[E_\tau] \leq 1$ is true under any $P \in \mathcal{P}$ at arbitrary *stopping* times. We refer to this validity at stopping times as *anytime-validity* (Johari et al., 2022; Howard et al., 2021).

For a fixed value t, E_t is also referred to as an *e-variable*, or *e-value* for its instantiation (Vovk and Wang, 2021; Grünwald et al., 2019), and if $\mathcal{P} = \{P\}$ then it is a betting score after t rounds in a betting game against P (as in Game 2.1). E-values are alternatives to the familiar (but problematic) p-values and can be defined outside of the sequential framework. Recent work has shown that e-values have important uses in combining results that can be arbitrarily dependent, particularly in multiple testing (Wang and Ramdas, 2022; Xu et al., 2021) and meta-analysis (ter Schure and Grünwald, 2022).

Importantly, e-processes are strict generalizations (Ramdas et al., 2022b) of test supermartingales to a family of distributions $\mathcal{P}$, which corresponds to a *composite* null in hypothesis testing.[3] To see this, we first extend the definition of a test supermartingale to a family of distributions $\mathcal{P}$ as follows: $(L_t)_{t\geq 0}$ is a *test (super)martingale for* $\mathcal{P}$ if it is a test (super)martingale for each $P \in \mathcal{P}$. Then, by the supermartingale stopping theorem (Durrett, 2019, Theorem 4.8.4), for any $P \in \mathcal{P}$ and any (possibly infinite) stopping time τ,

$$\mathbb{E}_P[L_\tau] \leq 1,$$

which coincides with (2.3). In fact, this connection further leads to an *equivalent* characterization of an e-process involving test supermartingales (Ramdas et al., 2020, Section 8.2.3). That is, $(E_t)_{t\geq 0}$ is an e-process for $\mathcal{P}$ *if and only if* for each $P \in \mathcal{P}$, there exists a test supermartingale $(L_t^P)_{t\geq 0}$ for P that

[3]Composite nulls are also related to imprecise probabilities, as the null set for testing imprecise probabilities can be expressed as a family of distributions (e.g., when testing $H_0 : \mathbb{E}[Y_t] \in [0.3, 0.7]$).

upper-bounds the e-process uniformly over time, under P:

$$E_t \leq L_t^P, \quad P\text{-almost surely}, \quad \forall t \geq 0,\ \forall P \in \mathcal{P}. \tag{2.4}$$

Ramdas et al. (2022b) further formalized the betting-based definition of an e-process: an e-process is the *minimum* wealth of a gambler who bets on each game corresponding to $P \in \mathcal{P}$.

An Example of a Game-Theoretic Formulation Involving an E-Process. To illustrate how an e-process can be utilized for composite null hypothesis testing, we consider a generalization of Game 2.1 where the casino can choose a different price for each round of the game. This example combines ideas from Howard et al. (2021); Ramdas et al. (2022b) and Chapter 3 of this thesis.

Game 2.2 (Testing time-varying probabilities by betting)**.**

Players: Casino and Gambler

Protocol:

1. Gambler enters the game with an initial wealth of $L_0 = 1$.
2. Gambler chooses the amount of bet $\lambda \in [0, 1]$ *(fixed for each round)*.
3. For rounds $t = 1, 2, \ldots$:
   (a) Casino announces the price of the game *for this round*, $\mu_t \in [0, 1]$.
   (b) Casino reveals the payoff $Y_t \in \{0, 1\}$.
   (c) Gambler's wealth is updated according to the payoff as follows:

$$L_t := L_{t-1} \cdot \exp\left\{\lambda(Y_t - \mu_t) - \lambda^2/8\right\}. \tag{2.5}$$

Result: Gambler wins if L_t grows large; Casino wins otherwise.

There are two modifications made from Game 2.1. First, the casino now announces a different price for each round, allowing for the possibility that the bias of each coin changes over time. If we were to define the analogous filtration, then we would additionally include this price in it:

$\mathcal{G}_{t-1}$ = all available information up to round $t-1$, including the gambler's bet λ, *and* the price μ_t.

Then, μ_t is interpreted as the casino's proposed probability for $P(Y_t = 1 \mid \mathcal{G}_{t-1})$.

Second, the gambler now chooses a single bet amount $\lambda \in [0, 1]$ for all rounds, and their wealth (2.5) is updated differently. At each round t, the gambler still bets proportionally to the difference $(Y_t - \mu_t)$, but the bet is now offset by a "hedge" on a conservative estimate of the variance, i.e., $1/4$, of each outcome. The wealth is also an example of an *exponential supermartingale*, as it can be expressed as an exponential of sum and variance terms:

$$L_t = \prod_{i=1}^{t} \exp\{\lambda(Y_i - \mu_i) - \lambda^2/8\} = \exp\left\{\lambda \sum_{i=1}^{t}(Y_i - \mu_i) - t\lambda^2/8\right\}. \tag{2.6}$$

To clarify in what sense this is a supermartingale, we define a family of distributions $\mathcal{P}^\mu$, each defined over all variables in Game 2.2 $(\lambda, \mu_1, Y_1, \mu_2, Y_2, \dots)$, as follows:

$$\mathcal{P}^\mu = \{P : \mathbb{E}_P[Y_t \mid \mathcal{G}_{t-1}] = P(Y_t = 1 \mid \mathcal{G}_{t-1}) = \mu_t, \forall t \geq 1\}. \tag{2.7}$$

In words, $\mathcal{P}^\mu$ consists of any distribution on the entire game sequence such that the conditional probability for each round's payoff being \$1 matches the casino's proposed price for that round.[4] Then, it can be shown that $(L_t)_{t\geq 0}$ is a test supermartingale for any $P \in \mathcal{P}^\mu$, by Hoeffding (1963)'s lemma (using the boundedness/sub-Gaussianity of Y_t):

$$\mathbb{E}_P\left[L_t \mid \mathcal{G}_{t-1}\right] = L_{t-1} \cdot \mathbb{E}_P[\exp\left(\lambda(Y_t - \mu_t) - \lambda^2/8\right) \mid \mathcal{G}_{t-1}] \leq L_{t-1}, \tag{2.8}$$

given that, under any $P \in \mathcal{P}^\mu$, μ_t is precisely the conditional mean of Y_t w.r.t. $\mathcal{G}_{t-1}$. Thus, the test supermartingale $(L_t)_{t\geq 0}$ for $\mathcal{P}^\mu$ can test whether the casino's announced price matches the actual conditional probability of the outcomes. Similarly, we can further show that $(L_t)_{t\geq 0}$ is also a test supermartingale for the one-sided family defined as $\mathcal{P}^{\leq\mu} = \{P : P(Y_t = 1 \mid \mathcal{G}_{t-1}) \leq \mu_t, \forall t \geq 1\}$. This wealth process would grow large if the casino consistently overprices its bet at each round.[5]

So far, we did not introduce an e-process that is not a test supermartingale. To do this, suppose

[4]Given that Y_t is binary and the only random outcome in the game, the "family" $\mathcal{P}^\mu$ is actually a singleton in this case. But we can also straightforwardly generalize the game to one where Y_t is a bounded random variable within $[0, 1]$; in this case, $\mathcal{P}^\mu$ can be a composite family of distributions whose conditional means are specified by μ.

[5]If we are only interested in testing $\mathcal{P}^\mu$ or $\mathcal{P}^{\leq\mu}$, a variant of the game in which the gambler can place different bets per round ($\lambda_t \in [0, 1]$) would also work.

now there are reasons to believe that the casino's announced bets accurately model the outcomes, i.e., $\mu_t = P(Y_t = 1 \mid \mathcal{G}_{t-1})$, but the gambler does *not* know what the values μ_t would be. Recall that, in Game 2.2, the gambler has to announce their bet *before* any of the rounds begin. In this case, let $\bar{\mu}_t = t^{-1} \sum_{i=1}^{t} \mu_i$ be the running average of the outcome probabilities, and consider the case where the gambler wants to test whether the *chance of winning on average* is better than $1/2$ at some point. This leads to the following family of distributions for testing:

$$\mathcal{P} = \left\{ P : \frac{1}{t} \sum_{i=1}^{t} P\left(Y_i = 1 \mid \mathcal{G}_{i-1}\right) \leq \frac{1}{2} \quad \text{and} \quad P(Y_t = 1 \mid \mathcal{G}_{t-1}) = \mu_t, \, \forall t \geq 1 \right\}. \tag{2.9}$$

Note that $\mathcal{P} \subseteq \mathcal{P}^{\mu}$, and the gambler does not know the exact values of $(\mu_t)_{t \geq 0}$ when placing the bet.

In order to test (2.9), the gambler can agree to play a variant[6] of Game 2.2 where the wealth is updated as follows: $E_0 = 1$ and

$$E_t = E_{t-1} \cdot \exp\left\{ \lambda \left(Y_t - \frac{1}{2} \right) - \frac{\lambda^2}{8} \right\} = \exp\left\{ \lambda \sum_{i=1}^{t} \left(Y_t - \frac{1}{2} \right) - \frac{t\lambda^2}{8} \right\}. \tag{2.10}$$

Then, under any $P \in \mathcal{P}$, notice that E_t is upper-bounded by L_t for every t:

$$\begin{aligned}
E_t &= \exp\left\{ \lambda \sum_{i=1}^{t} \left(Y_t - \frac{1}{2} \right) - \frac{t\lambda^2}{8} \right\} && (2.11) \\
&= \exp\left\{ \lambda \sum_{i=1}^{t} (Y_t - \mu_t) - \frac{t\lambda^2}{8} \right\} \cdot \exp\left\{ \lambda t \left(\bar{\mu}_t - \frac{1}{2} \right) \right\} && (2.12) \\
&= L_t \cdot \exp\left\{ \lambda t \left(\bar{\mu}_t - \frac{1}{2} \right) \right\} && (2.13) \\
&\leq L_t \quad (P\text{-a.s.}). && (2.14)
\end{aligned}$$

The equality in (2.13) follows under any $P \in \mathcal{P}^{\mu}$; the inequality in (2.14) follows from the condition in (2.9), which simplifies to $\bar{\mu}_t \leq 1/2$, $\forall t$. Given that $(L_t)_{t \geq 0}$ is a test supermartingale for $\mathcal{P}$ from (2.8), we can use the equivalent definition (2.4) to prove that $(E_t)_{t \geq 0}$ is an e-process for $\mathcal{P}$.

Therefore, assuming that the price of each bet (μ_t) correctly describes the payoff distribution, the

[6] As implied earlier, the precise game corresponding to an e-process would be one where the gambler plays *many* games, each corresponding to a member of $\mathcal{P}$, and then taking the *minimum* of the wealth across all games. We omit the exact formulation for the sake of clarity of the current exposition. See Ramdas et al. (2022b, Section 5.4) for a general formulation, including when there is no single upper-bounding test supermartingale.

e-process is expected to be small (at most one) at any time τ when the gambler decides to stop (2.3), as long as the running average of the outcome probabilities is at most half. In particular, the e-process can be used to then test whether the average outcome probability is favorable to the gambler, *just by seeing the outcomes* (Y_t), even when the exact prices at each round are not known. In Chapter 3, we show how an improved version of this e-process can be utilized to tightly *estimate* the running average $\bar{\mu}_t$, uniformly across time.

***A Comparison with the Frequentist Hypothesis Testing Framework.** In the standard frequentist framework, we would first start with a (modeling) assumption that the payoffs are generated as $Y_t \mid \mathcal{G}_{t-1} \sim \mathsf{Ber}(\theta_t)$, for some unknown "true" parameter sequence $(\theta_t)_{t\geq 0}$. Then, we can define the null hypothesis as $H_0 : \bar{\theta}_t = t^{-1}\sum_{i=1}^{t}\theta_i \leq 1/2$, $\forall t$ and try to come up with a test, say involving a likelihood according to the assumed model, such that the test's type I error is controlled.

In contrast, the game-theoretic approach can often bypass the need to assume a parametrized model but still retain frequentist type I error control or coverage guarantee via the e-process (see Section 2.3). For example, in Game 2.2, we do not require fully parametrizing the proposed distributions from the beginning, as $(\theta_t)_{t\geq 0}$ can be chosen *after* the earlier outcomes are realized. In sequential settings, this makes it easier to deal with composite nulls involving imprecise probability statements or allowing for "adversarial" choices by Reality. The approach is also more amenable to non-i.i.d. settings, as games are constructed sequentially. There are many composite nulls for which the model involves unspecified parameters besides the mean or a parametrized model is not assumed at all (Ramdas et al., 2022b; Shekhar and Ramdas, 2021; Shaer et al., 2023; Podkopaev et al., 2023).

Another key difference is that the game-theoretic approach is more *evidential* than the frequentist approach, as it places a notion of quantified evidence (via e-processes) at the center. This is particularly in contrast with the Neyman-Pearson framework, which forces a binary decision. The equivalent of the alternative hypothesis is the gambler's betting strategy, which can be chosen flexibly. More generally, the game-theoretic framework can be viewed as a middle ground between frequentist and Bayesian inference. See Ramdas et al. (2022a, Appendix A) for further details.

2.3 Ville's Inequality, P-Processes, and Confidence Sequences

Ville's Inequality for Test Supermartingales and E-Processes. Ville (1939)'s inequality for test supermartingales is the primary tool for constructing confidence sequences. If $(L_t)_{t=0}^{\infty}$ is a test supermartingale for P, then Ville's inequality states that, for any $\alpha \in (0,1)$,

$$P\left(\exists t \geq 1 : L_t > 1/\alpha\right) \leq \alpha. \tag{2.15}$$

Given the definition (2.4), it is immediate that Ville's inequality also holds for e-processes $(E_t)_{t\geq 0}$ for $\mathcal{P}$, w.r.t. each $P \in \mathcal{P}$: for any $\alpha \in (0,1)$,

$$P\left(\exists t \geq 1 : E_t > 1/\alpha\right) \leq \alpha, \quad \forall P \in \mathcal{P}. \tag{2.16}$$

Furthermore, by Howard et al. (2021, Lemma 3), each statement has an equivalent stopping time version. For any $\alpha \in (0,1)$,

$$P\left(L_\tau > 1/\alpha\right) \leq \alpha, \quad \forall \text{ stopping time } \tau, \tag{2.17}$$

and

$$P\left(E_\tau > 1/\alpha\right) \leq \alpha, \quad \forall \text{ stopping time } \tau, \quad \forall P \in \mathcal{P}. \tag{2.18}$$

These versions reveal that Ville's inequality is an anytime-valid generalization of Markov's inequality.

Ville's inequality is a powerful statement. It states that the probability of a test supermartingale or an e-process exceeding $1/\alpha$ at *any* (fixed or random) time is at most α, under P or $\mathcal{P}$ respectively. By Ville's inequality, any e-process for $\mathcal{P}$ immediately yields a level-α *sequential test* for $\mathcal{P}$, which we can formally define as a binary decision function ϕ_t such that $P(\phi_\tau = 1) \leq \alpha$ for any stopping time τ, $P \in \mathcal{P}$, and $\alpha \in (0,1)$. Given an e-process $(E_t)_{t\geq 0}$, we have from (2.18) that $\phi_t = \mathbb{1}\left(E_t > 1/\alpha\right)$, as well as the more powerful variant $\phi_t = \mathbb{1}\left(\sup_{i=1,\ldots,t} E_i > 1/\alpha\right)$, are both valid level-$\alpha$ sequential tests. Ville's inequality thus formalizes the betting interpretation into a concrete sequential test using an e-process: if P correctly describes the data generating distribution, then the wealth of a gambler betting against $\mathcal{P}$ cannot ever exceed a large threshold (e.g., 20) with high probability (e.g., 95%).

P-Processes. Given a family of distributions $\mathcal{P}$, representing a possibly composite null hypothesis, we say that a nonnegative adapted process $(\mathsf{p}_t)_{t\geq 0}$ is a **p-process** (or an *anytime-valid p-value*) for $\mathcal{P}$ if

$$P(\mathsf{p}_\tau \leq \alpha) \leq \alpha, \quad \forall \text{ stopping time } \tau, \quad \forall \alpha \in [0,1], \quad \forall P \in \mathcal{P}. \tag{2.19}$$

A p-process is the anytime-valid counterpart to a p-value, for which the definition does not require stopping time validity. As evident from equations (2.16) and (2.18), we can convert an e-process into a p-process via either

$$\mathsf{p}_t = \frac{1}{E_t} \quad \text{or} \quad \mathsf{p}_t = \frac{1}{\sup_{i=1,\ldots,t} E_i}. \tag{2.20}$$

There are also ways to "calibrate" p-processes into e-processes, e.g., via $E_t = \left(2\sqrt{\mathsf{p}_t}\right)^{-1}$ (Vovk and Wang, 2021). In this thesis, we generally favor e-processes over p-processes for their usefulness (in constructing CSs or combining results under arbitrary dependence) and their betting interpretation.

Confidence Sequences. When a test supermartingale or an e-process involves a (time-varying) parameter, such as in Game 2.2, we can utilize Ville's inequality to construct confidence sequences that estimate the parameter. Formally, given $\alpha \in (0,1)$, a $(1-\alpha)$-level **confidence sequence (CS)** (Darling and Robbins, 1967; Howard et al., 2021) for a time-varying target parameter $(\theta_t)_{t=1}^\infty$ is a sequence of confidence intervals (CIs) $(C_t)_{t=1}^\infty$ such that

$$P\left(\exists t \geq 1 : \theta_t \notin C_t\right) \leq \alpha, \text{ or equivalently, } P\left(\forall t \geq 1 : \theta_t \in C_t\right) \geq 1-\alpha. \tag{2.21}$$

A CS is *time-uniform*, in the sense that it covers the target parameter uniformly across all (fixed or random) times, even when it changes over time. Thus, a CS can be *continuously monitored* as more data is collected, and its coverage guarantee remains valid without requiring any corrections. This crucially differentiates a CS from a fixed-time CI, C_n, which only has the following (much) weaker guarantee:

$$\forall n \geq 1,\ P\left(\theta_n \notin C_n\right) \leq \alpha, \text{ or equivalently, } \forall n \geq 1,\ P\left(\theta_n \in C_n\right) \geq 1-\alpha. \tag{2.22}$$

Hereafter, n represents a fixed (pre-specified) sample size.

	Fixed-time CI & p-value	**Anytime-valid** CS, e-process, sequential test, & p-process
Validity at a fixed sample size n	Yes	Yes
Validity at an arbitrary stopping time τ ("anytime-valid")	No	Yes
Validity under continuous monitoring ("time-uniform")	No	Yes (CS, sequential test, & p-process)
Inference w/ composite nulls (imprecise probabilities)	Nontrivial	Yes
Game-theoretic interpretation	No	Yes

Table 2.2: A summary of the characterizing properties of anytime-valid methods. This is shown in comparison with existing "fixed-time" methods, whose validity is limited to a fixed sample size.

Furthermore, as with Ville's inequality, the time-uniform guarantee (2.21) is *equivalent* to the following anytime-valid property of the CS (Howard et al., 2021, Lemma 3):

$$P(\theta_\tau \in C_\tau) \geq 1 - \alpha, \quad \forall \text{ stopping time } \tau. \tag{2.23}$$

This means that a CS also remains valid under *optional stopping or continuation* (Grünwald et al., 2019), i.e., deciding to stop or continue collecting data after seeing the data, without having to pre-specify a sample size *a priori*. This property is sometimes referred to as the *safety*[7] of an inference procedure, leading to the umbrella term *safe, anytime-valid inference (SAVI)*. Note that a fixed-time CI is not anytime-valid or safe in this sense.

When it comes to sequential inference, particularly involving a time-varying parameter and adaptive data collection (e.g., A/B testing), CSs can be much more useful than CIs.

2.4 Summary

Table 2.2 summarizes the characterizing properties of the aforementioned anytime-valid inference methods, in comparison with their fixed-time counterparts. In Chapter 3, we illustrate these characteristics on the problem of comparing sequential forecasters using real-world examples.

[7]This is unrelated to the notion of safety for an ML predictor (e.g., robustness and alignment).

Chapter 3

Comparing Sequential Forecasters

This chapter is based on Choe and Ramdas (2021).

3.1 Introduction

Forecasts of future outcomes are widely used across domains, including meteorology, economics, epidemiology, elections, and sports. Often, we encounter multiple forecasters making probability forecasts on a regularly occurring event, such as whether it will rain the next day and whether a sports team will win its next game. Yet, despite the ubiquity of forecasts, it is not obvious how we can formally compare different forecasters on their predictive ability, particularly in a sequential setting where they each make a prediction on a sequence of outcomes (once for each outcome).

As an illustrative example, consider the probability forecasts made on each game of the 2019 World Series by real-world (and fictitious) forecasters in Table 3.1. It is not clear how we can effectively model the sequence of baseball game outcomes over time, and we also do not have full information on how each forecaster comes up with their predictions. As we observe these forecasts and outcomes game-by-game, we may see one forecaster appearing to be better than the other, according to some scoring rule. But how much of that difference can be attributed to chance or luck? How much evidence do we have that one forecaster has been "genuinely" better than another, even after accounting for chance, and can we quantify this evidence without having to make assumptions about reality or

[1]Source: https://projects.fivethirtyeight.com/2019-mlb-predictions/games/.
[2]Source: https://sports-statistics.com/sports-data/mlb-historical-odds-scores-datasets/.

Forecasts on Nationals Win	1	2	3	4	5	6	7
FiveThirtyEight[1]	37.9%	41.0%	52.7%	58.7%	37.3%	40.5%	48.5%
Vegas-Odds.com[2]	34.9%	37.7%	41.0%	50.7%	33.7%	37.4%	43.1%
Adjusted Win Percentage	47.1%	47.4%	47.6%	47.4%	47.2%	47.0%	47.2%
K29 Defensive Forecast	50.0%	50.0%	50.9%	51.6%	50.7%	49.9%	49.1%
Constant Baseline	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%	50.0%
Average Joe	40.0%	50.0%	60.0%	50.0%	30.0%	40.0%	50.0%
Nationals Fan	70.0%	70.0%	80.0%	70.0%	60.0%	60.0%	70.0%
Did the Nationals Win?	Yes	Yes	No	No	No	Yes	Yes

Table 3.1: Probability forecasts (%) on whether the Washington Nationals will win each game of the 2019 World Series. The first two forecasts are taken from publicly available websites online. The next three forecasts are baselines computed using the 10-year win/loss records (win probability is rescaled with the opponent's win probability to sum to 1). The last two forecasts are imaginary (but not unrealistic) casual sports fans making their own forecasts using different heuristics. All forecasts are made prior to the beginning of each game. See Section 3.5.2 for more details.

how the forecasts are made?

In this work, we derive statistically rigorous procedures for *sequentially* comparing forecasters via the powerful tool of *confidence sequences (CS)* (Darling and Robbins, 1967; Lai, 1976b; Howard et al., 2021). CSs are sequences of confidence intervals (CIs) that provide time-uniform coverage guarantees, which allow valid sequential inference under continuous monitoring and at data-dependent stopping times. The parameter of interest in this work is the time-varying mean difference in forecast scores up to time t. Most CSs we develop in our work are also nonasymptotically valid, meaning that their coverage guarantee holds at every time point $t \geq 1$.

In addition, we derive *e-processes* and *p-processes* (Ramdas et al., 2022b) for testing whether one forecaster outperforms the other on average, which is a composite null that we formally define in Section 3.4.4. An e-process E_t is a nonnegative process such that under the null, its expectation at any stopping time is at most one. It quantifies the amount of accumulated evidence against the null up to time t: a larger E_t is more evidence against the null. Further, $p_t = 1/\sup_{i \leq t} E_i$ is a p-process — its realization at any stopping time is a valid p-value, a property referred to as *anytime-valid* or *always-valid* (Johari et al., 2022; Howard et al., 2021). These are also formally defined in Section 3.4.4. Throughout the chapter, we define *safe, anytime-valid inference (SAVI)* methods as ones that satisfy either the time-uniform coverage guarantee (CS) or the anytime-valid guarantee (e- or p-processes).

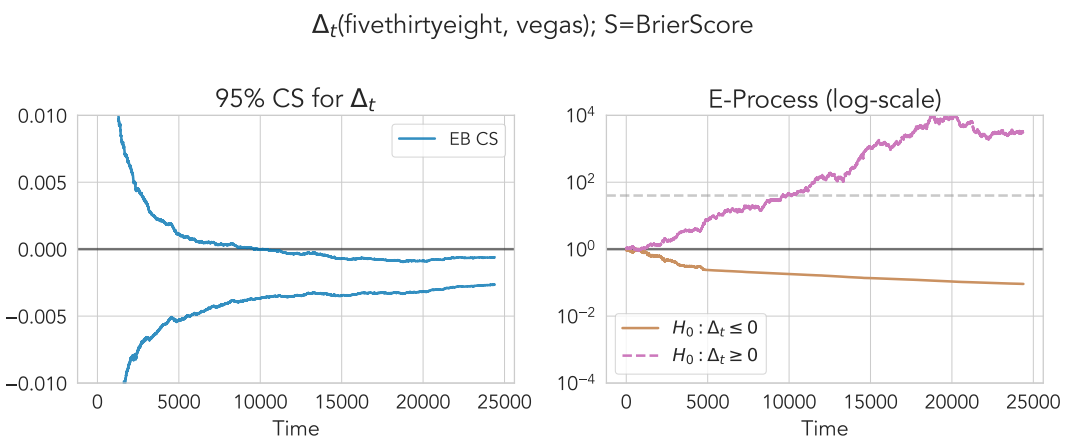


Figure 3.1: *Left:* A 95% CS (Theorem 3.2) for the average Brier score differentials $(\Delta_t)_{t=1}^T$ between *FiveThirtyEight* and *Vegas*, two real-world forecasters that made game-by-game probability forecasts on Major League Baseball (MLB) games from 2010 to 2019 ($T = 25,165$). Positive values of Δ_t indicate that the first forecaster is better than the second on average. Unlike a classical CI, a CS covers the time-varying parameter Δ_t uniformly over all t with high probability. In this case, we find that, with 95% probability, the sequence Δ_t trends negative for $t \geq 10,000$, indicating that *Vegas* outperformed *FiveThirtyEight* on average across most of the time horizon. *Right:* E-processes (Theorem 3.3) for the null hypotheses, $\mathcal{H}_0 : \Delta_t \leq 0, \ \forall t$ (brown) and $\mathcal{H}_0 : \Delta_t \geq 0, \ \forall t$ (purple), respectively. An e-process quantifies the accumulated evidence against the null, and it has a direct correspondence to the CS. In this example, larger values in the e-process for $\mathcal{H}_0 : \Delta_t \geq 0, \ \forall t$ indicate evidence of *Vegas* outperforming *FiveThirtyEight* on average. The gray horizontal line plots the value $2/\alpha = 40$, and the time at which an e-process upcrosses this line is also when the $(1-\alpha)$-CS moves entirely below or above zero. See Sections 3.4 and 3.5 for details.

The setup in which we develop our methods is game-theoretic (Shafer and Vovk, 2019): we posit that two players participate in a forecasting game on a sequence of outcomes with an unknown distribution. This game-theoretic setup naturally leads to "distribution-free" inference procedures — other than requiring bounded scoring rules, we make no distributional assumptions on the time-varying dynamics of the outcomes and forecasts, such as stationarity. We further discuss how to relax even the assumption of bounded scores using asymptotic CSs (App. A.3) and normalized scores (App. A.4).

In Figure 3.1, we show an example of a CS and its corresponding e-processes applied to a forecasting game between two real-world forecasters, *FiveThirtyEight* and *Vegas*, on the outcomes of Major League Baseball (MLB) games. The CS in the left plot continuously tracks the expected average score differential over time and effectively visualizes the time-varying trend along with the uncertainty on its estimation. The two e-processes in the right plot each measure the accumulated evidence favoring each forecaster over time. In this example, both the CS and the e-processes show that *Vegas* has

outperformed *FiveThirtyEight* on average. We return to this example in Section 3.5.2.

The rest of the chapter is organized as follows. After discussing related work (Section 3.2) and preliminaries (Section 3.3), we derive CSs for the time-varying average forecast score differentials between two probabilistic forecasters in Sections 3.4.1-3.4.3, with the case of binary outcomes as a working example. In Section 3.4.4, we also derive e-processes and p-processes as duals to our CSs, providing alternative sequential inference procedures for forecast comparison. In Section 3.5.1, we empirically validate our CSs and compare them against fixed-time and asymptotic confidence intervals (CIs) on simulated data. Finally, in Sections 3.5.2 and 3.5.3, we apply our methods to real-world forecast comparison tasks, namely comparing game-by-game predictions in Major League Baseball (MLB) and comparing statistical postprocessing methods of ensemble weather forecasts. For further details, Section A.1 contains omitted proofs; Section A.2 contains technical details about the time-uniform boundary choices; Section A.3 contains an alternative forecast comparison approach using an asymptotic CS; Sections A.4-A.6 contain extensions to normalized scores, lag-h forecasts, and predictable conditions/bounds, respectively; Section A.7 contains extensions from binary outcomes to categorical and continuous outcomes; Section A.8 contains detailed comparisons with existing forecast comparison methods; and Section A.9 contains additional experimental results and details.

3.2 Related Work

Evaluation and Comparison of Forecasts. Forecast evaluation is a well-studied subject in the literature of statistics, economics, finance, and climatology, dating back to the works of Brier (1950); Good (1952); DeGroot and Fienberg (1983); Dawid (1984); Schervish (1989). The primary tool for evaluating forecasts is proper scoring rules, of which the literature is extensive. Many characterization theorems for proper scoring rules exist across different forecasting scenarios, notably including the case of probability forecasts for binary and categorical outcomes, point forecasts (e.g., mean, quantiles, and prediction intervals) for continuous outcomes, and fully probabilistic forecasts (e.g., densities and CDFs) for continuous outcomes. See, e.g., McCarthy (1956); Savage (1971); Schervish (1989); Winkler et al. (1996); Grünwald and Dawid (2004); Gneiting and Raftery (2007); Gneiting (2011); Abernethy and Frongillo (2012); Dawid and Musio (2014); Ehm et al. (2016); Ovcharov (2018);

Frongillo and Kash (2021); Waggoner (2021), for both classical and recent developments.

The problem of comparing forecasts while accounting for sampling uncertainty was first popularized in the case of probability forecasts by Diebold and Mariano (1995) (DM), who proposed tests of equal (historical) forecast accuracy using the differences in forecast errors. The DM test is based on the asymptotic normality of the average forecast score differentials, and it makes stationarity assumptions about the outcomes. Giacomini and White (2006) (GW) developed tests of *conditional* predictive accuracy given past information, allowing for the comparison of "which forecaster is more accurate given the information available at the time of forecasting." The GW test thus allows for nonstationarity, although it restricts the forecasters to a fixed window size m and its validity depends on mixing assumptions. Lai et al. (2011) presented a comprehensive overview of the aforementioned methods of forecast comparison and developed a martingale-based theory of scoring rules whose differentials are linear in the outcome, such as proper scoring rules. They proved the asymptotic normality of both forecast scores and score differentials, leading to an asymptotic and fixed-time CI that we use as a point of comparison in our work. More recent work by Ehm and Krüger (2018); Ziegel et al. (2020); Yen and Yen (2021) derive fixed-time tests of forecast dominance under all consistent scoring functions (Gneiting, 2011). In comparison with all of these previous methods that presuppose a fixed sample size, the key difference in our work is that we develop inference methods that are valid at arbitrary data-dependent stopping times, while making virtually no assumption on the time-varying dynamics of the data generating process. The resulting graphical representations of CSs and e-processes also convey information about the entire time-varying trend of score differences, as in Figure 3.1, unlike classical tests and CIs that concern a single comparison at a fixed time point.

Recently, Henzi and Ziegel (2022) constructed sequential tests of conditional forecast dominance based on e-processes (Howard et al., 2020; Grünwald et al., 2019; Shafer, 2021; Ramdas et al., 2022b; Vovk and Wang, 2021). These methods are also anytime-valid and nonasymptotic; yet, they test a "strong[3] null," which states that one forecaster is better than the other at *every* point in time, something we rarely believe a priori. Thus, rejecting the strong null only suggests that there exists *some* time point where the latter forecaster is better than the former, which may not come as much of a

[3]This distinction of strong and weak nulls come from the discussion of randomized experiments in causal inference; see, e.g., Lehmann (1975); Rosenbaum (1995). Within the context of forecast comparison, Ehm and Krüger (2018) distinguish between tests of average and step-by-step conditional predictive ability, which mirrors that of weak and strong nulls.

Method & Key Result	**Null Hypothesis $\mathcal{H}_0$**	**Weak**	**CI**	**SAVI**	**NA**	**DF**
Diebold and Mariano (1995) $\sqrt{n}(\hat{\Delta}_n - \delta) \rightsquigarrow N(0, 2\pi f_d(0))$	$\delta = 0$	✗	✓	✗	✗	✗
Giacomini and White (2006) $T_m(\hat{\Delta}_n) \rightsquigarrow \chi^2$ (m: max. forecasting window)	$\mathbb{E}_{n-1}[\hat{\delta}_{m,n}] = 0,\ \forall n$	✗	✗	✗	✗	✗
Lai et al. (2011) $\sqrt{n}(\hat{\Delta}_n - \Delta_n)/s_n \rightsquigarrow N(0,1)$	$\frac{1}{n}\sum_{i=1}^{t} \mathbb{E}_{i-1}[\hat{\delta}_i] = 0,\ \forall n$	✓	✓	✗	✓	✗
Henzi and Ziegel (2022) $E_t = \prod_{i=1}^{t}\left(1 + \lambda \frac{\delta_i(y_i)}{\delta_i(\mathbb{1}(p_i > q_i))}\right)$ is an e-process, $\lambda > 0$	$\mathbb{E}_{t-1}[\hat{\delta}_t] \leq 0,\ \forall t$	✗	✗	✓	✓	✓
Ours $t(\hat{\Delta}_t - \Delta_t)$ is sub-exponential, yielding a CS & an e-process	$\frac{1}{t}\sum_{i=1}^{t} \mathbb{E}_{i-1}[\hat{\delta}_i] \leq 0,\ \forall t$	✓	✓	✓	✓	✓

Table 3.2: Inference methods for comparing probability forecasts for binary outcomes. This table is meant to be a quick summary only; see each referenced paper for the precise definitions, conditions, and guarantees for the method. *Notations:* for each $t \in \mathbb{N}$, p_t and q_t are two probability forecasts on the outcome y_t; $\delta_t(y) = S(p_t, y) - S(q_t, y)$; $\hat{\delta}_t = \delta_t(y_t)$; $\hat{\Delta}_t = t^{-1}\sum_{i=1}^{t} \hat{\delta}_i$; $\Delta_t = t^{-1}\sum_{i=1}^{t} \mathbb{E}_{i-1}[\hat{\delta}_i]$. We also use t to refer to a time index varying over time, and n to denote a fixed sample size that must be determined before the experiment. **Weak**: whether the method tests a weak null (involving a time-varying average). **CI**: whether the method provides a confidence interval for the score difference (as opposed to only deriving a test). **SAVI**: whether inference is valid at arbitrary data-dependent stopping times (as opposed to only fixed times). **NA**: whether the method has a nonasymptotic guarantee. **DF**: whether the method has a distribution-free guarantee (as opposed to requiring distributional assumptions like stationarity/mixing/i.i.d.). The last two methods are the only ones that are anytime-valid, nonasymptotic, and distribution-free — both of which develop e-processes. Among the two, only our method tests the weak null and provides a CS for *estimating* Δ_t.

surprise. (One case where the strong null is appropriate is if we test two sets of forecasts produced by the same data scientist, with one forecaster using more features or more sophisticated models; but for two unrelated forecasters, we rarely expect the strong null to be true.) In contrast, our e-processes test whether one forecaster dominates the other *on average* over time (thus requiring consistent outperformance), and the CSs can even test such averaged nulls in a two-sided fashion (equivalently, it tests both one-sided nulls). We examine this distinction further in Sections 3.4.4 and 3.5.3; other methodological differences are summarized in Section A.8.1.

Table 3.2 summarizes the aforementioned methods of forecast comparison in terms of whether they have a stopping time (or equivalently, time-uniform; see Section 3.4.4 for further details) guarantee, a non-asymptotic guarantee, and a distribution-free guarantee.

Time-Uniform Confidence Sequences. Confidence sequences were developed by Robbins and coauthors (Darling and Robbins, 1967; Robbins, 1970; Robbins and Siegmund, 1970; Lai, 1976a). Recent renewed interests on CSs are partly due to best-arm identification in multi-armed bandits (Jamieson et al., 2014; Jamieson and Jain, 2018), where CSs are sometimes referred to as always-valid or anytime confidence intervals. CSs are also duals to sequential hypothesis tests, analogously to CIs being dual to fixed-time hypothesis tests, and one can further derive a sequence of e-processes and p-processes given the CSs (more precisely, its underlying exponential process) (Ramdas et al., 2022b). In Section 3.4.4, we make this connection explicit and discuss how our approach also leads to p-processes, or anytime-valid p-values (Johari et al., 2022), for weak nulls.

The recent work by Howard et al. (2021) is of particular importance in our work, as it develops tight confidence sequences that are uniformly valid over time under nonparametric assumptions and has widths that shrink to zero. This work and its underlying technique of developing exponential test (super)martingales (Howard et al., 2020; Darling and Robbins, 1967; Ville, 1939) have led to several interesting results, including state-of-the-art concentration inequalities for IID mean estimation (Waudby-Smith and Ramdas, 2023) and sequential quantile estimation (Howard and Ramdas, 2022). Our work makes the connection between the empirical Bernstein (EB) CSs derived in Howard et al. (2021) and the martingale property of forecast score differentials (Lai et al., 2011), leading to a novel sequential inference procedure for forecaster comparison.

3.3 Preliminaries

3.3.1 Test Supermartingales, Ville's Inequality, and Confidence Sequences

See Chapter 2 for a detailed introduction to these central concepts.

3.3.2 Forecast Evaluation via Scoring Rules

Let $\mathcal{Y}$ be the space of all possible outcomes equipped with a σ-field $\mathcal{G}$. Let $\Delta(\mathcal{Y})$ be the set of all probability distributions on $(\mathcal{Y}, \mathcal{G})$ and $\mathcal{P} \subseteq \Delta(\mathcal{Y})$. To facilitate our discussion, the primary working example in this chapter will be the space of binary outcomes $\mathcal{Y} = \{0, 1\}$ and probability forecasts parametrized by their means in $\mathcal{P} = [0, 1]$. But our setup can be generalized to any finite sample space

$\mathcal{Y} = \{1, \dots, K\}$ with K-dimensional probability forecasts $\mathcal{P} = \Delta^{K-1}$, for $K \geq 2$, and d-dimensional sample space $\mathcal{Y} \subseteq \mathbb{R}^d$, for $d \geq 1$, with point (e.g., mean and quantile) or probabilistic (e.g., CDF) forecasts. (We defer our discussion of these general cases to Section A.7.)

A *scoring rule* is any extended real-valued function[4] $S : \mathcal{P} \times \mathcal{Y} \to \overline{\mathbb{R}}$ and can be used to evaluate the performance of a (probabilistic) forecast $p \in \mathcal{P}$ given an observation $y \in \mathcal{Y}$. Following Gneiting and Raftery (2007), we take scoring rules to be *positively oriented*, meaning that higher scores reflect better forecasts. A prominent example is the Brier score (Brier, 1950), which in the binary case can be expressed as $S(p, y) = 1 - (p - y)^2$ for $p \in [0, 1]$ and $y \in \{0, 1\}$.

Given a forecast $p \in \mathcal{P}$ and a probability distribution $q \in \Delta(\mathcal{Y})$, we can naturally extend the definition of a scoring rule S to its *expected score* w.r.t. $y \sim q$ (conditional on p):

$$S(p; q) = \mathbb{E}_{y \sim q}\left[S(p, y)\right]. \tag{3.1}$$

Here, we make the distinction between the scoring rule S on $\mathcal{P} \times \mathcal{Y}$ and its expected score S defined on $\mathcal{P} \times \Delta(\mathcal{Y})$ by the notations $S(p, y)$ and $S(p; q)$, respectively. We can recover the scoring rule from the expected score definition via $S(p, y) = S(p; \delta_y)$, where δ_y is a point measure on y.

A scoring rule S is *proper* if any probability $q \in \Delta(\mathcal{Y})$ maximizes the expected score $S(\cdot; q)$:

$$q \in \operatorname*{argmax}_{p \in \mathcal{P}} S(p; q). \tag{3.2}$$

S is *strictly proper* if the argmax in (3.2) is unique. Intuitively, a proper scoring rule encourages forecasters to be honest, because if a forecaster believes that the outcome follows the distribution $q \in \mathcal{P}$, then they are incentivized to honestly forecast q, instead of any other distribution $p \neq q$, as q maximizes the expected score (uniquely, if S is strictly proper) according to their belief. Proper scoring rules are often considered as the primary means of evaluating probabilistic forecasts, as they assess both calibration and sharpness (Winkler et al., 1996; Gneiting et al., 2007).

Classical examples of proper scoring rules for probability forecasts $p \in \mathcal{P} = [0, 1]$ on binary outcomes $y \in \mathcal{Y} = \{0, 1\}$ include the following:

[4]More formally, the scoring rule S is required to be *$\mathcal{P}$-quasi-integrable* in its second argument, meaning that for every $p \in \mathcal{P}$, $S(p, \cdot)$ is measurable and, for all $q \in \mathcal{P}$, the integral $\int_y S(p, y) dq(y)$ exists as a possibly infinite but not indeterminate value (Bauer, 2001; Abernethy and Frongillo, 2012).

- The Brier score or the quadratic score (Brier, 1950): $S(p, y) = 1 - (p - y)^2$.
- The spherical score (Good, 1971): $S(p, y) = \frac{py+(1-p)(1-y)}{\sqrt{p^2+(1-p)^2}}$.
- The logarithmic score (Good, 1952): $S(p, y) = y \log(p) + (1 - y) \log(1 - p)$.
- The zero-one score or the success rate: $S(p, y) = y\mathbb{1}(p \geq 0.5) + (1 - y)\mathbb{1}(p < 0.5)$.

The Brier, spherical, and logarithmic scores are examples of strictly proper scoring rules, while the zero-one score is an example of a proper but not strictly proper scoring rule. An example of an improper scoring rule for probability forecasts is the absolute score, $S(p, y) = 1 - |p - y|$. Also note that all of the examples except the logarithmic score are bounded for $p \in [0, 1]$ and $y \in \{0, 1\}$.

3.4 Anytime-Valid Inference for Average Forecast Score Differentials

In this section, we derive CSs and e-processes, as well as their corresponding sequential tests and p-processes, for the time-varying average difference in the quality of forecasts, as measured by a scoring rule. Our intuition comes from the extensive literature on evaluating and comparing probability forecasts via scoring rules (Winkler et al., 1996; Gneiting and Raftery, 2007; DeGroot and Fienberg, 1983; Schervish, 1989; Gneiting, 2011; Lai et al., 2011), combined with the powerful tool of time-uniform CSs (Darling and Robbins, 1967; Howard et al., 2021). For now, our working example in this section will be the case of comparing probability forecasts on binary outcomes; we further discuss extensions to categorical and certain continuous outcomes in Section A.7.

3.4.1 A Game-Theoretic Formulation

The intuition behind our SAVI methods for forecast score differentials comes from the game-theoretic statistical framework (Shafer, 2021; Ramdas et al., 2022a). Consider a forecasting game where two players make probabilistic forecasts on an event that happens over time (e.g., whether it will rain on each day, whether a sports team will win its game each week, and more) and an unknown player named reality chooses a sequence of distributions that generates the outcomes that the forecasters are trying to predict. Let $t = 1, 2, \ldots$ denote each round of the game. Though not required, we can also optionally allow having any historical data $y_{-(H-1)}, \ldots, y_{-1}, y_0$ for some $H \geq 0$. The forecasting game can be formulated in general as follows — the case of probability forecasts on binary outcomes

is obtained by setting $\mathcal{P} = \Delta(\mathcal{Y}) = [0,1]$ ($y_t \sim r_t$ would refer to $y_t \sim \text{Bernoulli}(r_t)$).

Game 3.1 (Comparing Sequential Forecasters)**.** For rounds $t = 1, 2, \dots$:

1. Forecasters 1 and 2 make their forecasts, $p_t, q_t \in \mathcal{P}$, respectively. *The order in which the forecasters make their forecasts is not specified.*
2. Reality chooses $r_t \in \Delta(\mathcal{Y})$. *$r_t$ is not revealed to the forecasters.*
3. $y_t \sim r_t$ is sampled and revealed to the forecasters.

We now elaborate on the role of each player in Game 3.1.

Forecasters 1 & 2. At each round t, the two forecasters can make their forecasts using any information available to them. This includes historical and previous outcomes $y_{-(H-1)}, \dots, y_0, y_1, \dots, y_{t-1}$, any of the previous forecasts made, $p_1, \dots, p_{t-1}, q_1, \dots, q_{t-1}$, as well as any other side information available to either forecaster. They cannot, however, make their predictions using any of $r_1, \dots, r_t$'s (or information from the future). For example, when predicting the outcome of the next baseball game, the forecasters' filtration may include not only all of previous games' results but also any side information that either forecaster may have, such as which players are starting the game and whether there are injuries. The setup also allows for the case where two forecasters have different side information, as our results are completely agnostic to such details.

This game-theoretic framework for forecast comparison is *prequential* (Dawid, 1984), in the sense that we put no restrictions on how these forecasts are generated, and we only evaluate forecasters based on the forecasts they did make and the outcomes that did occur, as opposed to forecasts they would have made had the outcomes been different.

Reality. In our game, Reality is the player that determines the unknown distribution r_t of the eventual outcome y_t conditioned on its past, which notably includes the forecasters' choices p_t and q_t. In the binary case, for example, Reality chooses the conditional mean sequence of the outcomes y_t given everything it has seen. Reality can essentially choose r_t "however they want," and they can even choose r_t after seeing p_t or q_t, although in practice Reality is usually not influenced by the forecasters. Put differently, the framework is agnostic to what information Reality sees: Reality may only see its past choices $r_1, \dots, r_{t-1}$ and (optionally) the past outcomes $y_1, \dots, y_{t-1}$, or it may act adversarially after

seeing p_t and q_t. In particular, r_t could also be a point distribution at y_t.

We note that the distribution-free property of our methods corresponds to the fact that the game places no distributional assumptions on the time-varying dynamics of $(r_t)_{t=1}^{\infty}$, such as stationarity, Markovian or other conditional independence assumptions.

The Statistician. The statistician, who stands outside of the game, has the goal of comparing the predictive performance of the two forecasters according to a chosen scoring rule and based only on the observed data $(p_t, q_t, y_t)_{t=1}^{\infty}$, without making any assumptions about the behavior of any player involved.[5] The statistician may choose to update their inferential conclusions as the game progresses. How the statistician achieves such a goal will be the focus of the subsequent sections.

3.4.2 The Measure-Theoretic Setup

We now formalize Game 3.1 in the context of comparing the two probabilistic forecasters over time. Let $(p_t)_{t=1}^{\infty}$ and $(q_t)_{t=1}^{\infty}$ be two sequences of forecasts in $\mathcal{P}$, for a sequence of outcomes $(y_t)_{t=1}^{\infty}$ in $\mathcal{Y}$. In the binary case, the forecasts will take values in $\mathcal{P} = [0, 1]$ and the outcomes in $\mathcal{Y} = \{0, 1\}$. We can define Game 3.1 in a measure-theoretic sense by specifying the associated filtrations, i.e., a sequence of "information sets" with which we perform inference. Our formulation is closely related to the setup of Lai et al. (2011), although we make the game-theoretic intuitions explicit.

The "Observable" Forecaster Filtration $\mathfrak{F}$. We first define the filtration with which the two forecasters generate their forecasts, denoted as $\mathfrak{F} := (\mathcal{F}_t)_{t=0}^{\infty}$. For each $t \geq 1$, let $\mathcal{F}_{t-1}$ represent *any* information available to the forecasters before making their predictions at time t, as described in the previous subsection. Mathematically, this means that $(p_t)_{t=1}^{\infty}$, $(q_t)_{t=1}^{\infty}$, and $(y_t)_{t=1}^{\infty}$ are predictable w.r.t. $\mathfrak{F}$. Note that $\mathfrak{F}$ also represents the information available to the statistician, making this the "observable" filtration that contrasts with the "oracle" filtration (defined below).

The "Oracle" Game Filtration $\mathfrak{G}$. The game filtration, denoted as $\mathfrak{G} := (\mathcal{G}_t)_{t=0}^{\infty}$, represents *all* sets of information associated with Game 3.1. The parameter of interest (unknown to the statistician)

[5]Specifically, we do not explicitly consider strategic issues arising from (say) the choice of the scoring rule or the method of comparison. In other words, we consider the comparison problem separately from the elicitation problem (how to elicit honest forecasts). A separate line of work considers these important, but orthogonal, issues.

is defined in terms of this filtration, making it the "oracle" filtration. More precisely, for each $t \geq 1$, $\mathcal{G}_{t-1}$ includes not only everything in $\mathcal{F}_{t-1}$ but also any information available to Reality before the outcome y_t is realized, including Reality's choice r_t. Mathematically, this implies that $(p_t)_{t=1}^{\infty}$, $(q_t)_{t=1}^{\infty}$, and $(r_t)_{t=1}^{\infty}$ are predictable w.r.t. $\mathfrak{G}$. The setup allows for the flexible choices of Reality described in the previous subsection, as it does not preclude Reality's actions in any way.

In the remainder of the chapter, we use the notation $\mathbb{E}_{t-1}[\cdot] = \mathbb{E}[\cdot \mid \mathcal{G}_{t-1}]$ to denote the conditional expectation with respect to the game filtration for each t. In the case of binary (and categorical) outcomes, because the outcome distribution is completely specified by their mean, we simply let r_t denote the (unknown) conditional mean of the outcome y_t given $\mathcal{G}_{t-1}$ for each t, with a slight abuse of notation. In such cases, we have that

$$r_t = \mathbb{E}_{t-1}[y_t] \quad \forall t = 1, 2, \ldots, \tag{3.3}$$

where $\mathbb{E}_{t-1}$ refers to the conditional expectation over $y_t \sim r_t \mid \mathcal{G}_{t-1}$.

Comparing Sequential Forecasters via Average Forecast Score Differentials. With the aforementioned setup, we can now use scoring rules to assess and compare the quality of the two forecasters over time. We define the ***average (forecast) score differential*** Δ_t between the sequences of forecasts $(p_i)_{i=1}^{\infty}$ and $(q_i)_{i=1}^{\infty}$, up to time t, as the average difference in *expected scores*:

$$\Delta_t := \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}_{i-1}\left[S(p_i, y_i) - S(q_i, y_i)\right], \quad t \geq 1, \tag{3.4}$$

where $\mathbb{E}_{i-1}$ denotes the expectation over $y_i \sim r_i$ *conditioned on* the game filtration $\mathcal{G}_{i-1}$, which includes both forecasts p_i and q_i as well as r_i. The time-varying parameter Δ_t provides an intuitive way of quantifying the difference in the quality of forecasts made up to time t. We highlight that Δ_t helps us infer whether one forecaster is better than the other *on average* (over time), as opposed to one strictly dominating the other (Giacomini and White, 2006; Henzi and Ziegel, 2022). This estimand is also used in Lai et al. (2011)'s asymptotic CI.

The parameter Δ_t is not observable to the statistician or the forecasters, because reality's moves $r_1, \ldots, r_t$ are unknown and never observed. We thus define the ***empirical average (forecast) score***

differential $\hat{\Delta}_t$ as the unbiased estimate of each summand in (3.4), also averaged over time:

$$\hat{\Delta}_t := \frac{1}{t}\sum_{i=1}^{t}\left[S\left(p_i, y_i\right) - S\left(q_i, y_i\right)\right], \quad t \geq 1. \tag{3.5}$$

$\hat{\Delta}_t$ is completely observable to the statistician after time t.

The statistician's goal then becomes quantifying how far $\hat{\Delta}_t$ is from Δ_t, while accounting for the uncertainty associated with sampling y_t at each time t. To this end, we define the *pointwise (forecast) score differential* $\delta_i := \mathbb{E}_{i-1}[S(p_i, y_i) - S(q_i; y_i)]$ and its empirical counterpart $\hat{\delta}_i := S(p_i, y_i) - S(q_i, y_i)$. Then, it is immediate that the cumulative sums of deviations, defined by $S_0 = 1$ and

$$S_t := t\left(\hat{\Delta}_t - \Delta_t\right) = \sum_{i=1}^{t}\left(\hat{\delta}_i - \delta_i\right), \quad t \geq 1, \tag{3.6}$$

forms a martingale, i.e., $\mathbb{E}_{t-1}[S_t] = S_{t-1},\ \forall t \geq 1$. Previous work including Seillier-Moiseiwitsch and Dawid (1993); Lai et al. (2011) use this property to derive the asymptotic normality of empirical average score differentials. In the following sections, we illustrate how $(S_t)_{t=0}^{\infty}$ can further be uniformly and non-asymptotically bounded by constructing *exponential* test supermartingales. As a result, we will be able to estimate and cover Δ_t using CSs and also test its sign using e-processes.

3.4.3 Time-Uniform Confidence Sequences for Average Score Differentials

Time-Uniform Boundaries and Exponential Test Supermartingales We now show that we can uniformly bound the difference between $\hat{\Delta}_t$ and Δ_t over time using uniform boundaries and test supermartingales. To do this, we start with a *cumulative sum* process $S_t := \sum_{i=1}^{t}(\hat{\delta}_i - \delta_i)$ as well as its *intrinsic time* $\hat{V}_t$, which is the variance process for S_t (to be defined later). Our goal is then to uniformly bound the sum S_t over the intrinsic time $\hat{V}_t$, which corresponds to bounding the difference between $\hat{\Delta}_t$ and Δ_t over time due to (3.6).

Following Howard et al. (2020), for any sum process $(S_t)_{t=0}^{\infty}$ and its intrinsic times $(\hat{V}_t)_{t=0}^{\infty}$, we define a *(one-sided) uniform boundary* $u = u_\alpha$ *with crossing probability* $\alpha \in (0, 1)$ as any function of the intrinsic time that gives a time-uniform bound on the sums:

$$P\left(\forall t \geq 1 : S_t \leq u_\alpha(\hat{V}_t)\right) \geq 1 - \alpha, \tag{3.7}$$

that is, with probability at least $1-\alpha$, the sums S_t are upper-bounded by $u(\hat{V}_t)$ *at all times* t. By similarly computing uniform boundary to $(-S_t, \hat{V}_t)_{t=0}^{\infty}$, we can also obtain a time-uniform lower bound on S_t. (Alternatively, we can directly define a *two-sided* sub-ψ uniform boundary, which satisfies $P(\forall t \geq 1 : -u_\alpha(\hat{V}_t) \leq S_t \leq u_\alpha(\hat{V}_t)) \geq 1-\alpha$. An example is Robbins (1970)'s two-sided normal mixture that we describe later.) The upper and lower bounds then jointly form a time-uniform CS on $(\Delta_t)_{t=1}^{\infty}$ by rearranging the terms.

How do we show that there exists such a uniform boundary for our definitions of $(S_t, \hat{V}_t)_{t=0}^{\infty}$? Howard et al. (2020, 2021) show that there exists such a uniform boundary if, for each $\lambda \in [0, \lambda_{\max})$, the *exponential process* defined by $L_0(\lambda) = 1$ and

$$L_t(\lambda) = \exp\left\{\lambda S_t - \psi(\lambda)\hat{V}_t\right\}, \quad t \geq 1, \tag{3.8}$$

is a test supermartingale w.r.t. $\mathfrak{G}$. Here, $\psi : [0, \lambda_{\max}) \to \mathbb{R}$ is a "CGF-like" function (Howard et al., 2020), with a scale parameter $c > 0$, that controls how fast S_t can grow relative to the intrinsic time $\hat{V}_t$. It is called a "CGF-like" function because it closely resembles (or equals) a cumulant generating function (CGF) of a mean-zero random variable. In this work, we use two ψ functions:

- $\psi_{N,c}(\lambda) = c^2\lambda^2/2, \ \forall \lambda \in [0, \infty)$, which is the CGF of a centered Gaussian with variance c^2;
- $\psi_{E,c}(\lambda) = c^{-2}(-\log(1-c\lambda) - c\lambda), \ \forall \lambda \in [0, 1/c)$, which is a rescaled CGF of a centered Exponential with scale c.

If $L_t(\lambda)$ is a test supermartingale for each $\lambda \in [0, \lambda_{\max})$ for some ψ, then we say that $(S_t)_{t=0}^{\infty}$ is *sub-*ψ *with variance process* $(\hat{V}_t)_{t=0}^{\infty}$. In particular, we say that $(S_t)_{t=0}^{\infty}$ is sub-Gaussian or sub-exponential, with variance process $(\hat{V}_t)_{t=0}^{\infty}$ and scale c, if it is sub-$\psi_{N,c}$ or sub-$\psi_{E,c}$ respectively; these generalize the definitions of sub-Gaussian and sub-exponential random variables to cumulative sums w.r.t. intrinsic time. The uniform boundary u defined using ψ is then called a *sub-*ψ *uniform boundary*.

Our goal is now to identify the conditions with which $(L_t(\lambda))_{t=0}^{\infty}$ is indeed a test supermartingale and use different ψ functions to obtain different uniform boundaries and hence CSs.

Warmup: Hoeffding-Style Confidence Sequences We first derive an illustrative example of a CS for Δ_t solely based on the sub-Gaussianity of the empirical pointwise score differentials $(\hat{\delta}_i)_{i=1}^{\infty}$.

While the resulting CS is not the tightest one in our case, its derivation is simple enough to showcase the general pipeline for deriving CSs.

Recall the problem setup in Section 3.4.2, and for each $i \geq 1$, consider two probability forecasts $p_i, q_i \in [0,1]$ on a binary outcome $y_i \in \{0,1\}$ with unknown mean $r_i \in [0,1]$. Since p_i, q_i, and y_i are all bounded, we know that the pointwise score differentials $\hat{\delta}_i$ for $i \geq 1$ are also bounded for many of the scoring rules we've discussed (e.g., $|\hat{\delta}_i| \leq 1$ for the Brier, spherical, and zero-one scores). If $|\hat{\delta}_i| \leq c$ for some $c > 0$, we know that $\hat{\delta}_i$ is c-sub-Gaussian (Hoeffding, 1963) conditioned on the game filtration $\mathcal{G}_{i-1}$, meaning that $\mathbb{E}_{i-1}[e^{\lambda(\hat{\delta}_i - \delta_i)}] \leq e^{\lambda^2 c^2/2} = \exp\{\psi_{N,c}(\lambda)\}$ for all $\lambda \in \mathbb{R}$.

Now, for each t, define the cumulative sum $S_t = \sum_{i=1}^{t}(\hat{\delta}_i - \delta_i)$ and the intrinsic time $\hat{V}_t = \sum_{i=1}^{t} 1 = t$. It then follows that, for each $\lambda \in [0, \infty)$, the exponential process $(L_t(\lambda))_{t=0}^{\infty}$ given by $L_t(\lambda) = \exp\{\lambda S_t - \psi_{N,c}(\lambda)\hat{V}_t\}$ is a test supermartingale:

$$\mathbb{E}_{t-1}[L_t(\lambda)] = L_{t-1}(\lambda) \cdot \mathbb{E}_{t-1}\left[\exp\left\{\lambda\left(\hat{\delta}_t - \delta_t\right) - \psi_{N,c}(\lambda)\right\}\right] \leq L_{t-1}(\lambda). \tag{3.9}$$

Hence, there exists a sub-Gaussian uniform boundary for $(S_t, \hat{V}_t)$ such that the time-uniform guarantee in (3.7) holds. By rearranging terms and also using the analogous argument for $(-S_t, \hat{V}_t)$, we arrive at our first CS. Hereafter, the notation $(a \pm b)$ denotes the interval $(a - b, a + b)$.

Theorem 3.1 (Hoeffding-style confidence sequences for Δ_t). *Suppose that $\hat{\delta}_i$ is c-sub-Gaussian conditioned on $\mathcal{G}_{i-1}$ for $i \geq 1$, for some $c \in (0, \infty)$. Then, for any $\alpha \in (0,1)$,*

$$C_t^{\mathsf{H}} := \left(\hat{\Delta}_t \pm \frac{u(t)}{t}\right) \quad \textit{forms a } (1-\alpha)\textit{-CS for } \Delta_t, \tag{3.10}$$

where $u = u_{\alpha/2,c}$ is any (one-sided) sub-Gaussian uniform boundary with crossing probability $\frac{\alpha}{2}$ and scale c (or alternatively, a two-sided version with crossing probability α and scale c).

The statement (3.10) is equivalent to saying that, with probability at least $1 - \alpha$, Δ_t is contained in C_t^{H} *for all time t*, or that $P(\forall t \geq 1 : \Delta_t \in C_t^{\mathsf{H}}) \geq 1 - \alpha$. This CS is called a Hoeffding-style CS, as it extends Hoeffding (1963)'s inequality for the sums of independent sub-Gaussian random variables to the sequential case. In the sub-Gaussian case, it is also possible to construct a two-sided boundary without separately constructing one-sided boundary. This is due to a classical result by Robbins (1970) that we restate later in (3.13), so the upper and lower confidence bounds need not be constructed

separately; in practice, the one-sided and two-sided variants are nearly identical (Howard et al., 2021). We further discuss the possible choices of the uniform boundary later in this subsection.

The condition for Theorem 3.1 (and for Theorem 3.2 that will follow shortly) is satisfied by many scoring rules for probability forecasts on binary or categorical outcomes, including the Brier, spherical, and zero-one scores. For the unbounded logarithmic score, one can use its truncated variant $S(p, y) = y \log(p \vee \epsilon) + (1 - y) \log((1 - p) \vee \epsilon)$ for some small $\epsilon > 0$; although the score is no longer proper, our methods remain valid. The condition is also satisfied for scoring rules on bounded continuous outcomes, such as Brier and quantile scores on $[0, 1]$-valued outcomes (See Section A.7).

Main Result: Empirical Bernstein Confidence Sequences Now we are ready to present our main result, which is the derivation of a tight CS for Δ_t. The key difference from the Hoeffding-style CS is that we now use an empirical estimate of the variance process for the cumulative sums, leading to a variance-adaptive CS that is often much tighter in practice.[6] Recall the problem setup in Section 3.4.2 once again.

Theorem 3.2 (Empirical Bernstein confidence sequences for Δ_t)**.** *Suppose that* $|\hat{\delta}_i| \leq \frac{c}{2}$ *for each* $i \geq 1$*, for some* $c \in (0, \infty)$*. Also, let* $\hat{V}_t = \sum_{i=1}^{t} (\hat{\delta}_i - \gamma_i)^2$*, where* $(\gamma_i)_{i=1}^{\infty}$ *is any* $\left[-\frac{c}{2}, \frac{c}{2}\right]$*-valued predictable sequence w.r.t.* $\mathfrak{G}$*. Then, for any* $\alpha \in (0, 1)$*,*

$$C_t^{\mathsf{EB}} := \left(\hat{\Delta}_t \pm \frac{u(\hat{V}_t)}{t} \right) \quad \text{forms a } (1 - \alpha)\text{-CS for } \Delta_t, \tag{3.11}$$

where $u = u_{\alpha/2,c}$ *is any sub-exponential uniform boundary with crossing probability* $\frac{\alpha}{2}$ *and scale* c*.*

As before, the statement (3.11) is equivalent to saying that, with probability at least $1 - \alpha$, Δ_t is contained in C_t^{EB} *for all time* t, or that $P\left(\forall t \geq 1 : \Delta_t \in C_t^{\mathsf{EB}}\right) \geq 1 - \alpha$. The proof is provided in Section A.1.2. Theorem 3.2 (and its proof) can be viewed as an extension of Theorem 4 in Howard et al. (2021) to our setup of sequential forecast comparison.

Like the Hoeffding-style CS in Theorem 3.1, the EB CS estimates the conditional predictive ability in an anytime-valid and distribution-free manner. The EB CS is further variance-adaptive because its width is a function of the empirical variance process $(\hat{V}_t)_{t=0}^{\infty}$, and we illustrate this empirically in Sec-

[6]The improvement from a Hoeffding-style CS to an empirical Bernstein CS mirrors the improvement from Hoeffding's inequality to empirical Bernstein's inequality for bounded random variables in the fixed-sample case.

Type	CS C_t	Intrinsic Time $\hat{V}_t$	Uniform Boundary u
Hoeffding-Style (Theorem 3.1)	$\left(\hat{\Delta}_t \pm \frac{u(\hat{V}_t)}{t}\right)$	t	Normal Mixture Polynomial Stitching
Emp. Bernstein (Theorem 3.2)	$\left(\hat{\Delta}_t \pm \frac{u(\hat{V}_t)}{t}\right)$	$\sum_{i=1}^{t}(\hat{\delta}_i - \gamma_i)^2$, $(\gamma_i)_{i=1}^{\infty}$ predictable	Gamma-Exponential Mixture Polynomial Stitching

Table 3.3: Summary of confidence sequences and their uniform boundary choices.

tion 3.5. As before, we can use any bounded scoring rules, which in the binary and categorical cases include the Brier, spherical, and zero-one scores (proper), as well as the truncated logarithmic score (improper); scoring rules for bounded continuous outcomes can similarly be used. In addition, for unbounded *proper* scores for binary forecasts, such as the logarithmic score, we show in Section A.4 that a normalized version of the average score differential, due to Winkler (1994), can be used.

The choice of the uniform boundary u is discussed in the following subsection. A reasonable choice for the predictable sequence $(\gamma_i)_{i=1}^{\infty}$ is the average of previous score differentials, i.e., $\gamma_i = \hat{\Delta}_{i-1}$, although a smarter choice may lead to tighter CS. For the rest of this chapter, our default choice of CS for Δ_t will be that of Theorem 3.2, using $\hat{V}_t = \sum_{i=1}^{t}(\hat{\delta}_i - \hat{\Delta}_{i-1})^2$, unless specified otherwise.

Choosing the Uniform Boundary via the Method of Mixtures The specific choice of the uniform boundary u controls the tightness of the CS across time, and an extensive list of choices for u is covered in detail in Howard et al. (2021). While the simplest uniform boundaries are given as linear functions of the intrinsic time (Howard et al., 2020), curved uniform boundaries can produce CSs that are tighter across time. Here, we focus on a type of curved boundary called the conjugate-mixture boundary; another option, called the polynomial stitching boundary, is also discussed in App. A.2.2. Either boundary type is applicable to Theorems 3.1 and 3.2.

The conjugate-mixture (CM) boundary (Howard et al., 2021), denoted as u_α^{CM}, represents a class of uniform boundaries arising from the method of mixtures, the first instance of which was derived by Darling and Robbins (1967). The key idea is summarized as follows. Since $L_t(\lambda) = \exp\{\lambda S_t - \psi(\lambda)\hat{V}_t\}$ is a test supermartingale for every $\lambda \in [0, \lambda_{\max})$, it follows that for any distribution F on $[0, \lambda_{\max})$, the mixture $L_t^{\mathrm{mix}} := \int L_t(\lambda) dF(\lambda)$ is also a test supermartingale. Choosing F to be conjugate (in the Bayesian sense) to ψ then gives a closed-form expression for L_t^{mix}. For example, if $(S_t)_{t=0}^{\infty}$ is sub-

Gaussian with $(\hat{V}_t)_{t=0}^{\infty}$ (Theorem 3.1), then choosing F to be a Gaussian results in the *normal mixture* boundary (Robbins, 1970); if $(S_t)_{t=0}^{\infty}$ is sub-exponential with $(\hat{V}_t)_{t=0}^{\infty}$ (Theorem 3.2), then choosing F as a Gamma results in a *gamma-exponential mixture* boundary.

To elaborate, by Lemma 2 of Howard et al. (2021), if $L_t(\lambda) = \exp\{\lambda S_t - \psi(\lambda)\hat{V}_t\}$ is a test supermartingale for each $\lambda \in [0, \lambda_{\max})$ and F is any probability distribution on $[0, \lambda_{\max})$, then the following function is a sub-ψ uniform boundary with crossing probability $\alpha \in (0,1)$:

$$u_\alpha^{\mathsf{CM}}(v) := \sup\left\{ s \in \mathbb{R} : \underbrace{\int \exp\{\lambda s - \psi(\lambda) v\}\, dF(\lambda)}_{=: m(s,v)} < \frac{1}{\alpha} \right\}, \quad v \geq 0. \tag{3.12}$$

Because $m(S_t, \hat{V}_t) = L_t^{\mathsf{mix}}$ is a test supermartingale, Ville's inequality says that $P(\forall t \geq 1 : m(S_t, \hat{V}_t) < 1/\alpha) \geq 1 - \alpha$, which in turn implies that $P(\forall t \geq 1 : S_t \leq u_\alpha^{\mathsf{CM}}(\hat{V}_t)) \geq 1 - \alpha$. Similarly, if $(-S_t, \hat{V}_t)_{t=0}^{\infty}$ is also sub-ψ, then the above procedure also gives the lower bound on S_t.

Importantly, the uniform boundary (3.12) can be used for both Theorems 3.1 and 3.2, with the choice of F differing in each case. For the Hoeffding-style CS in Theorem 3.1, a two-sided normal mixture boundary can be computed directly in closed-form by choosing F to be $\mathcal{N}(0, \rho^{-1})$ (Robbins, 1970):

$$u_\alpha^{\mathsf{CM}}(v; \psi_N) = \sqrt{(v + \rho) \log\left(\frac{v + \rho}{\alpha^2 \rho}\right)} \tag{3.13}$$

where $\rho > 0$ is a free parameter. In practice, ρ can be chosen to optimize the width of the resulting CS at a pre-specified intrinsic time. A one-sided normal mixture boundary can also be derived in closed-form (Howard et al., 2021).

For the EB CS in Theorem 3.2, a one-sided gamma-exponential mixture boundary $u_\alpha^{\mathsf{CM}}(v; \psi_E)$, with F as a Gamma, can be computed efficiently using a numerical root finder ($m(s,v)$ has a closed form; the boundary u_α^{CM} is obtained numerically. See App. A.2.1 for details). The one-sided boundary can be used for computing both the upper and lower confidence bounds of the EB CS. If a closed-form boundary is needed, then the polynomial stitching boundary (App. A.2.2) can be used. Also, while the CM boundary has an asymptotic rate of $O(\sqrt{v \log v})$ as illustrated in (3.13), it is usually tighter than the polynomial stitched boundary in practice. In fact, the CM boundary is unimprovable in the

case of sub-Gaussian random variables without extra assumptions (Howard et al., 2021, Ppn. 4).

Table 3.3 summarizes the choice of uniform boundaries and the CSs we derived for estimating Δ_t. In our experiments, we use the conjugate-mixture uniform boundary by default, although we also perform an empirical comparison between the different choices as well as their hyperparameters in Section A.9.4. We use the publicly available implementation of the polynomial stitching and CM uniform boundaries by Howard et al. (2021).[7]

3.4.4 Sequential Tests, E-Processes and P-Processes

While our derivation so far has focused on confidence sequences, we can also derive e-processes and p-processes (Shafer and Vovk, 2019; Vovk and Wang, 2021; Grünwald et al., 2019; Ramdas et al., 2020). In particular, an e-process can be derived as a lower bound on the exponential test supermartingale (3.8) that we used to construct the CS in the previous section. This correspondence is general to any exponential process upper-bounded by a test supermartingale, as noted in, e.g., Ramdas et al. (2020); Howard et al. (2021); our work utilizes this fact to introduce alternative sequential inference procedures with the same anytime-valid and distribution-free guarantees.

Weak and Strong Null Hypotheses. Before deriving e- and p-processes, we first make clear the null hypotheses that correspond to the CS derived in Theorem 3.2. We define the *weak one-sided null* $\mathcal{H}_0^{\mathrm{w}}(p,q)$ as

$$\mathcal{H}_0^{\mathrm{w}}(p,q) : \Delta_t = \frac{1}{t}\sum_{i=1}^{t}\delta_i \leq 0, \quad \forall t = 1,2,.... \tag{3.14}$$

$\mathcal{H}_0^{\mathrm{w}}(p,q)$ implies that, across all times t, the first forecaster (p) is no better than the second forecaster (q) *on average*. Note that $\mathcal{H}_0^{\mathrm{w}}(p,q)$ is a composite null, in the sense that it consists of all joint distributions P on $\mathfrak{G}$ such that $\Delta_t \leq 0$ for all $t \geq 1$ under P. $\mathcal{H}_0^{\mathrm{w}}(q,p)$ is analogously defined as $\mathcal{H}_0^{\mathrm{w}}(q,p) : \Delta_t = \frac{1}{t}\sum_{i=1}^{t}\delta_i \geq 0$.

We now illustrate how the CSs derived in Theorem 3.1 and Theorem 3.2 would correspond to sequential tests of the weak one-sided nulls $\mathcal{H}_0^{\mathrm{w}}(p,q)$ and $\mathcal{H}_0^{\mathrm{w}}(q,p)$, drawing from the duality between CSs and sequential tests (Johari et al., 2022; Howard et al., 2021; Ramdas et al., 2020). Specifically, because the upper and lower confidence bounds are often constructed separately, the $(1-\alpha)$-level

[7] https://github.com/gostevehoward/confseq

CS for Δ_t denoted as $C_t = (L_t, U_t)$ satisfies $\Delta_t \leq U_t$ with probability at least $1 - \frac{\alpha}{2}$ *and* that $\Delta_t \geq L_t$ with probability at least $1 - \frac{\alpha}{2}$. Thus, if for any time t we find that $L_t > 0$ or $U_t < 0$, then we can reject either $\mathcal{H}_0^{\mathsf{w}}(p, q)$ or $\mathcal{H}_0^{\mathsf{w}}(q, p)$ with high probability. More generally, the CSs readily provide a valid stopping rule for rejecting $\mathcal{H}_0^{\mathsf{w}}$, a fact that we summarize in the following corollary. Below, we follow Robbins' power-one testing framework which uses one-sided stopping rules that only stop on rejecting the null (and do not stop otherwise).

Corollary 3.1 (A sequential test for $\mathcal{H}_0^{\mathsf{w}}$ using a CS)**.** *Given a* $(1-\alpha)$*-CS* $C_t = (L_t, U_t)$ *obtained using either Theorem 3.1 or 3.2, the following stopping rule provides a valid level-*α *sequential test for* $\mathcal{H}_0^{\mathsf{w}}(p, q)$ *and* $\mathcal{H}_0^{\mathsf{w}}(q, p)$ *(jointly):*

$$\textit{Reject } \mathcal{H}_0^{\mathsf{w}}(p, q) \textit{ if } L_t > 0\textit{; reject } \mathcal{H}_0^{\mathsf{w}}(q, p) \textit{ if } U_t < 0. \tag{3.15}$$

This means that:

$$\sup_{P \in \mathcal{H}_0^{\mathsf{w}}(p,q)} P(\exists t \geq 1 : \textit{Reject } \mathcal{H}_0^{\mathsf{w}}(p, q)) + \sup_{P \in \mathcal{H}_0^{\mathsf{w}}(q,p)} P(\exists t \geq 1 : \textit{Reject } \mathcal{H}_0^{\mathsf{w}}(q, p)) \leq \alpha. \tag{3.16}$$

The stopping rule (3.15) is equivalent to *deciding that p has been better (worse) than q if* C_t *is entirely above (below) zero.* The anytime-validity of this rule implies that the statistician can, e.g., periodically perform the test as t increases and update their decision accordingly. On one extreme, the statistician can choose to perform the test after every round t, or on the other extreme, they can test just once at a designated time t^* (while leaving open the possibility of revisiting the experiment some time later). Compared to a standard hypothesis test for a stationary mean, the underlying Δ_t can change its course over time, so in general it may not be sufficient to test once at t^* in order to have power against the weak null. See Section 3.5 for an illustration and Section 3.6 for a further discussion.

We note that separately testing for both $\mathcal{H}_0^{\mathsf{w}}(p, q)$ and $\mathcal{H}_0^{\mathsf{w}}(q, p)$ is not equivalent to simply testing for $\Delta_t = 0, \forall t$, which is equivalent to $\delta_t = 0, \forall t$. Rather, the sequential test (3.15) is the combination of two separate sequential tests in (3.15) for $\mathcal{H}_0^{\mathsf{w}}(p, q)$ and $\mathcal{H}_0^{\mathsf{w}}(q, p)$, each at the significance level $\alpha/2$. The interpretation of the CS as two simultaneous sequential tests allows the user to continuously monitor the score differential on both sides via the CS-based stopping rule (3.15).

For the sake of comparison, we also define the *strong one-sided null* $\mathcal{H}_0^{\mathsf{s}} = \mathcal{H}_0^{\mathsf{s}}(p, q)$ as

$$\mathcal{H}_0^{\mathsf{s}}(p, q) : \delta_t \leq 0, \quad \forall t = 1, 2, \dots . \tag{3.17}$$

$\mathcal{H}_0^{\mathsf{s}}(q, p)$ is defined analogously as $\mathcal{H}_0^{\mathsf{s}}(q, p) : \delta_t \geq 0, \ \forall t = 1, 2, \dots$. The recent work by Henzi and Ziegel (2022) develops e-processes (defined in the next paragraph) and sequential tests for this null. In contrast to $\mathcal{H}_0^{\mathsf{w}}$, $\mathcal{H}_0^{\mathsf{s}}$ corresponds to saying that the first forecaster (p) is no better than the second forecaster (q) at *every* time step $t = 1, 2, \dots$. Thus, the strong null $\mathcal{H}_0^{\mathsf{s}}$ implies the weak null $\mathcal{H}_0^{\mathsf{w}}$, but not vice versa. The critical distinction here is that rejecting $\mathcal{H}_0^{\mathsf{s}}$ only tells us that p outperformed q at *some* time step t, but it does not tell us if either was better on average over time. To give a concrete example, fix $k > 2$ (say, $k = 7$ indicating Sundays), and define

$$\delta_t = +0.1 \text{ if } t = k, 2k, 3k, \dots ; \quad \delta_t = -1 \text{ otherwise.} \tag{3.18}$$

In other words, p is generally worse than q but marginally better than q every kth time step (e.g., every Sunday). Because the strong null is false, any (powerful) sequential test for the strong null will reject it, and yet this may be a confusing conclusion as q is generally a better forecaster.

Sub-Exponential E-Processes for the Weak Null. We now show that the exponential test supermartingale underlying the CS in Theorem 3.2 can also be transformed to directly measure evidence against the weak one-sided null (rather than make a decision at a level α). Formally, an *e-process* (Ramdas et al., 2022b) for a (possibly composite) null hypothesis $\mathcal{H}_0$ is defined as a nonnegative process $(E_t)_{t=0}^{\infty}$, starting at one ($E_0 = 1$), such that:

$$\text{for any } P \in \mathcal{H}_0 \text{ and any arbitrary stopping time } \tau, \quad \mathbb{E}_P[E_\tau] \leq 1, \tag{3.19}$$

where we define $E_\infty := \limsup_{t \to \infty} E_t$. The larger the value of E_t, the more the evidence against the null. In particular, if the null is true, then it is unlikely to observe large values of the process at any stopping times (by Markov's inequality, $P(E_\tau \geq 1/\alpha) \leq \alpha$). An e-process is anytime-valid by definition (3.19) (validity at arbitrary stopping times), analogous to the anytime-validity of a CS,

and the term 'process' is also used to emphasize this property. An e-process can also be interpreted in a fully game-theoretic statistical sense: an e-process for a composite null measures the *minimum* wealth among bets against each member of the null (Ramdas et al., 2022b), such that it only grows large when there is evidence against all members. At a fixed t, E_t is also called an e-variable, and its realization is called an e-value (Vovk and Wang, 2021; Grünwald et al., 2019).

We can now define and show an e-process that corresponds to Theorem 3.2. (We can also define an analogous e-process corresponding to Theorem 3.1, but this is omitted due to space constraints.) The following e-process is for the weak one-sided null $\mathcal{H}_0^{\mathsf{w}}(p,q)$ and is related to the lower confidence bound of the CS from Theorem 3.2; the e-process for $\mathcal{H}_0^{\mathsf{w}}(q,p)$ is analogous and related to the upper confidence bound of the CS. Recall once again the problem setup in Section 3.4.2.

Theorem 3.3 (Sub-exponential e-processes for $\mathcal{H}_0^{\mathsf{w}}$)**.** *Assume the same conditions as Theorem 3.2. Then, for each $\lambda \in [0, 1/c)$,*

$$E_t(\lambda) := \exp\left\{\lambda \sum_{i=1}^{t} \hat{\delta}_i - \psi_{E,c}(\lambda)\hat{V}_t\right\} \qquad \text{is an e-process for } \mathcal{H}_0^{\mathsf{w}}(p,q). \tag{3.20}$$

Furthermore, given a probability distribution F on $[0, 1/c)$, the mixture process $E_t^{mix} := \int E_t(\lambda)dF(\lambda)$ is an e-process for $\mathcal{H}_0^{\mathsf{w}}(p,q)$.

The proof, provided in Section A.1.3, shows that under each $P \in \mathcal{H}_0^{\mathsf{w}}$, $E_t(\lambda)$ is upper-bounded by a exponential test supermartingale for P, namely $L_t(\lambda)$ in (3.8). Because a process is upper-bounded by a test supermartingale for $P \in \mathcal{H}_0$ if and only if it is an e-process for $\mathcal{H}_0$ (Ramdas et al., 2020), this establishes that $E_t(\lambda)$ is an e-process in the sense of (3.19). It then follows that $E_t^{\mathsf{mix}} \leq \int L_t(\lambda)dF(\lambda) = L_t^{\mathsf{mix}}\ \forall t$, so E_t^{mix} is also an e-process.

The e-process of Theorem 3.3 is an anytime-valid inference procedure that provides a measure of accumulated evidence against the weak one-sided null $\mathcal{H}_0^{\mathsf{w}}(p,q)$ at any stopping time. By definition, it is expected to be small under the weak null, and we only expect to see it grow large when the weak null does not hold. In comparison with Henzi and Ziegel (2022)'s e-process for the *strong* null, we see that our e-process provides a more useful notion of evidence for saying that one forecaster outperforms another. In the example of (3.18), an e-process for the strong null can grow large, even though q is generally a better forecaster; in contrast, our e-process (3.20) for the weak null is expected

to remain small. In Section 3.5.3, we provide an empirical comparison of the two e-processes.

Choosing λ (or F) for E-Processes. Theorem 3.3 tells us that the expected value of $E_t(\lambda)$ and E_t^{mix} are bounded by 1 at all stopping times under the null, for any choice of λ or any mixture distribution F. In practice, we default to using a mixture e-process with the conjugate distribution F, as in Section 3.4.3. For the sub-exponential e-process, the gamma-exponential mixture as before provides a closed form for the function $m(s,v)$ in (3.12), so that $E_t^{\text{mix}} = m(\sum_{i=1}^t \hat{\delta}_i, \hat{V}_t)$ can be computed efficiently. The expression for $m(s,v)$ is included in Section A.2.1.

P-Processes. Finally, we remark that any e-process for $\mathcal{H}_0$ can also be converted into an *p-process* for $\mathcal{H}_0$, i.e., the sequence $(\mathsf{p}_t)_{t=0}^\infty$ that satisfies: for any $\alpha \in (0,1)$,

$$\text{for any } P \in \mathcal{H}_0 \text{ and for any arbitrary stopping time } \tau, \quad P(\mathsf{p}_\tau \leq \alpha) \leq \alpha. \tag{3.21}$$

A p-process evaluated at any stopping time τ, i.e. p_τ, is a p-value, but unlike a classical p-value, a p-process is anytime-valid. Any e-process $(E_t)_{t=0}^\infty$ can be converted into a p-process via $\mathsf{p}_t := 1/\sup_{i\leq t} E_i$, following derivations from, e.g., Ramdas et al. (2020, 2022b).

We also remark that p_t can alternatively be defined from a CS as the smallest α for which the $(1-\alpha)$-level CS does not include zero (Howard et al., 2021), so all three notions (CS, e-process, and p-process) are closely related.

3.5 Experiments

In this section, we run both simulated and real-data experiments for sequential forecast comparison using our CSs as well as e-processes. All code and data sources for the experiments are made publicly available online at `https://github.com/yjchoe/ComparingForecasters`.

3.5.1 Numerical Simulations

As our first experiment, we compare our Hoeffding-style and EB CSs (Theorems 3.1 and 3.2, respectively) on simulated data with the asymptotic fixed-time CIs due to Theorem 2 of Lai et al. (2011). The

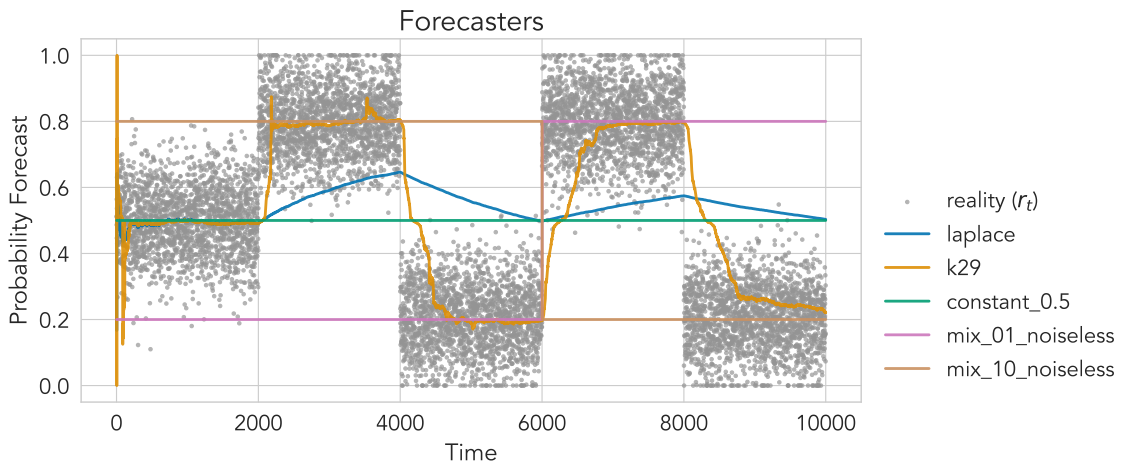


Figure 3.2: Various forecasters on a simulated non-IID data ($T = 10^4$) with sharp changepoints across time. Note that, instead of plotting the binary outcomes $y_t \in \{0, 1\}$, we plot the Reality's choices $(r_t)_{t=1}^T$ that generates the outcome sequence. See text for details about the forecasters.

main goal is to confirm that the CSs cover time-varying average score differentials uniformly, unlike the fixed-time CI, and are also nearly as tight as the CI.

In our simulated experiments, we also include an asymptotic CS for time-varying means, recently developed by Waudby-Smith et al. (2021), as an additional tool for anytime-valid inference. Asymptotic CSs can be viewed as alternatives to their non-asymptotic counterparts, including the ones we introduced in Section 3.4, and they trade off non-asymptotic validity to achieve versatility and also comparatively smaller widths at smaller sample sizes. A formal review of asymptotic CSs in the context of sequential forecast comparison is included in Section A.3.

As for our simulated data, we generate a sequence of non-IID binary outcomes and compare different forecasters using our CSs. The overall simulation pipeline closely follows Game 3.1, with $\mathcal{P} = \Delta(\mathcal{Y}) = [0, 1]$, $\mathcal{Y} = \{0, 1\}$, and $T = 10^4$. At each round $t = 1, \dots, T$, each forecaster makes a probability forecast $p_t, q_t \in \mathcal{P}$, then reality chooses r_t, and finally $y_t \sim \text{Bernoulli}(r_t)$ is sampled. The forecasts p_t and q_t are made only using the previous outcomes, i.e., $y_1, \dots, y_{t-1}$. The Reality's choices $(r_t)_{t=1}^T$ is specifically chosen to be non-IID and contain sharp changepoints, as shown in Figure 3.2. This serves as a challenging test case for the EB CS, as the sharp changepoints make it difficult to quickly adapt to the underlying variance. See Section A.9.1 for further details.

At the end of each round $t = 1, \dots, T$, we compute the 95% Hoeffding-style and EB CS for Δ_t, using

Theorems 3.1 and 3.2 respectively. We use the Brier score $S(p,q) = 1-(p-q)^2$ as our default scoring rule, but we also explore other scoring rules later in the section. As for the hyperparameter choices for sub-ψ uniform boundaries, we are guided by preliminary experiments in Section A.9.4.

We consider several forecasters, which are drawn with lines in Figure 3.2. These include the constant baseline, i.e., $p_t = 0.5$ (`constant_0.5`), as well as the Laplace forecasting algorithm (`laplace`) $p_t = \frac{k+0.5}{t+1}$, where $k = \#\{i \in [t] : y_i = 1\}$. We further add predictions using the K29 defensive forecasting algorithm (`k29`) (Vovk et al., 2005), which is a game-theoretic forecasting method that yields calibrated forecasts. The method depends on the choice of a kernel function, and here we use the Gaussian RBF $K(p,q) = \exp\left(-\frac{(p-q)^2}{2\sigma^2}\right)$ with bandwidth $\sigma = 0.01$. The `mix_01_noiseless` forecaster is defined as $p_t = 0.8$ for $t \le 6000$ and $p_t = 0.2$ for $t > 6000$; the `mix_01` forecaster is a noisy version that adds an independent noise to p_t by $\tilde{p}_t = p_t + 0.5 \cdot \epsilon_t$ (clipped at 0 and 1), where ϵ_t is drawn IID from Student's t distribution with 1 degree of freedom. The `mix_10_noiseless` forecaster is defined as $q_t = 1 - p_t$ and the `mix_10` forecaster $\tilde{q}_t$ is analogously defined.

The choices of forecasters and Reality are made in such a way that the unknown parameter Δ_t can not only change its sign but also have different variances over time. For example, the `mix_10` forecaster outperforms ($\Delta_t > 0$) the `mix_01` forecaster on average during $t \in (2000, 6000)$, while the sign then reverses ($\Delta_t < 0$) for $t \in (6000, 10000)$. Among the algorithmic forecasters, the K29 variants consistently perform better than the Laplace algorithm, especially when using sharper kernels, because they are better at modeling the sharp changepoints over time.

In Figure 3.3, we plot the 95% Hoeffding-style CS (Theorem 3.1), EB CS (Theorem 3.2), and a fixed-time CI for Δ_t (top left), as well as their widths (top right), the corresponding e-process (bottom left), and the cumulative miscoverage rates (bottom right). First, both CSs successfully cover Δ_t at any given time point, and their widths decrease as more outcomes are observed. As expected, the width of the EB CS decays more quickly than the width of the Hoeffding CS due to its use of the empirical variance term ($\hat{V}_t$) but more slowly than the fixed-time CI, matching the patterns observed in Howard et al. (2021); Waudby-Smith et al. (2021). As noted before, the fixed-time CI is only valid at a fixed time t and not uniformly over time, despite its tighter width, and this is illustrated by its large cumulative miscoverage rate, i.e., $\alpha_t = P(\exists i \le t : \Delta_i \notin C_i)$ (estimated over the repeated sampling of

$y_1, \ldots, y_t$ under P). In contrast, the EB CS[8] keeps its cumulative miscoverage rate well below α (it is in fact zero, as it is constructed using supermartingales and not martingales). In Section A.8.2, we also include an analogous plot comparing our methods with other classical tests (Diebold and Mariano, 1995; Giacomini and White, 2006).

The sub-exponential e-processes for $\mathcal{H}_0(p, q)$ (solid green) and $\mathcal{H}_0(q, p)$ (dotted purple) show how they accurately track the accumulated evidence for/against each forecaster over time. For example, the e-process for $\mathcal{H}_0(p, q)$ stays below 1 during $t < 2000$, when neither forecaster outperforms the other, and grows large during $t \in (2000, 6000)$ when data shows more evidence against the null hypothesis that $\Delta_t \leq 0, \ \forall t$ because the true Δ_t in fact becomes positive. It then decreases back to values below 1 during $t \in (6000, 10000)$, when the true Δ_t becomes negative. We note that the gray dotted line indicates the value $2/\alpha = 40$; testing whether an e-process exceeds $2/\alpha$ corresponds to a level-$(\alpha/2)$ sequential test equivalent to the one stated in Corollary 3.1. In fact, the plots show that the points at which the $(1 - \alpha)$-level EB CS excludes zero (on either side) are precisely when either e-process exceeds $2/\alpha$, illustrating the duality between the CS and the e-process.

In Figure 3.4, we now plot the 95% CSs (left), their widths (middle), and also the corresponding e-processes (right) for comparing the `k29_poly3` forecaster against the `laplace` baseline, using the spherical score (strictly proper), zero-one score (proper), the ϵ-truncated logarithmic score ($\epsilon = 10^{-8}$) (improper). We observe that all variants of CSs always cover the true Δ_t over time, at $\alpha = 0.05$, and its width decreases similarly to the case of Brier scores and eventually approaches that of the asymptotic CS. In terms of the width comparison between EB and Hoeffding CSs, we see that the EB CS is generally much tighter than the Hoeffding CS, and it decreases more slowly around time steps when there are sharp changepoints in Δ_t. This can be explained by the variance-adaptive nature of the EB CS, which would use larger values of intrinsic time $\hat{V}_t$ at sharp changepoints, whereas the Hoeffding CS simply uses $\hat{V}_t = t$ irrespective of the variance process. The sub-exponential e-processes for $\mathcal{H}_0^{\mathrm{w}}(p, q)$ and $\mathcal{H}_0^{\mathrm{w}}(q, p)$ illustrate the accumulated evidence for the first forecaster in all three cases around the same time the CS moves entirely above zero, illustrating the duality between the two methods.

We include a plot of all pairwise comparisons between four of the forecasters in Section A.9.1.

[8]The EB CS is computed with the polynomial stitching bound for computational efficiency.

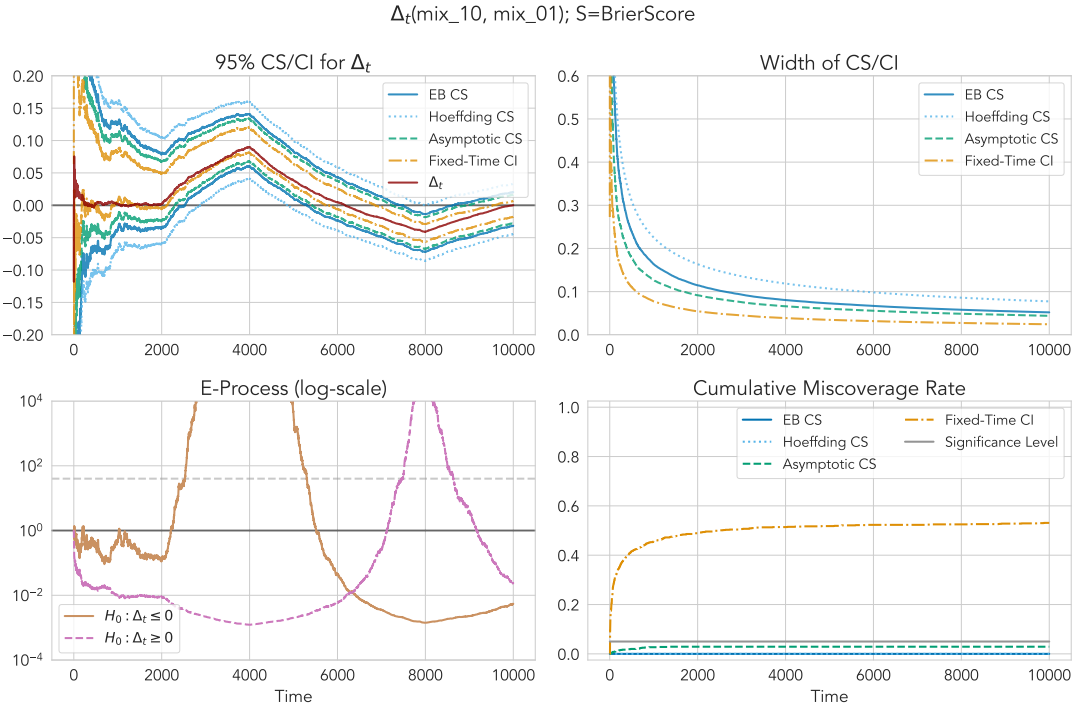


Figure 3.3: *Top Left:* 95% EB CS (blue, solid), Hoeffding-style CS (skyblue, dotted), asymptotic CS (green, dashed; Section A.3), and a fixed-time asymptotic CI (orange, dash-dotted) for simulated time-varying average score differentials $(\Delta_t)_{t=1}^T$ between the `mix_10` and `mix_01` forecasters ($T = 10^4$). The Brier score is used. *All CSs, but not the CI, uniformly cover the true score differential sequence, which changes signs sharply multiple times across the horizon. Top Right:* Widths of the CSs and the CI across time steps. The variance-adaptive EB CS is tighter than the Hoeffding CS and slightly looser than the asymptotic CS; the fixed-time CI is the tightest, but it does not have the time-uniform guarantee. *Bottom Left:* Sub-exponential e-processes (Theorem 3.3) that measures the accumulated evidence against either forecaster (solid green: first forecaster; dashed purple: second). Testing whether the e-process exceeds the dashed gray line at $2/0.05 = 40$ corresponds to a sequential test at $\alpha = 0.05$ (Corollary 3.1). *Bottom Right:* The cumulative miscoverage rate, which estimates $\alpha_t = P(\exists i \leq t : \Delta_i \notin C_i)$ over repeated sampling of $y_1, \ldots, y_t$ under P. For a $(1-\alpha)$-CS, this rate is controlled at α by definition; it is in fact always zero for the non-asymptotic CSs in our experiments. For a fixed-time CI, this rate exceeds well above α and continues to increase (in log-scale of time).

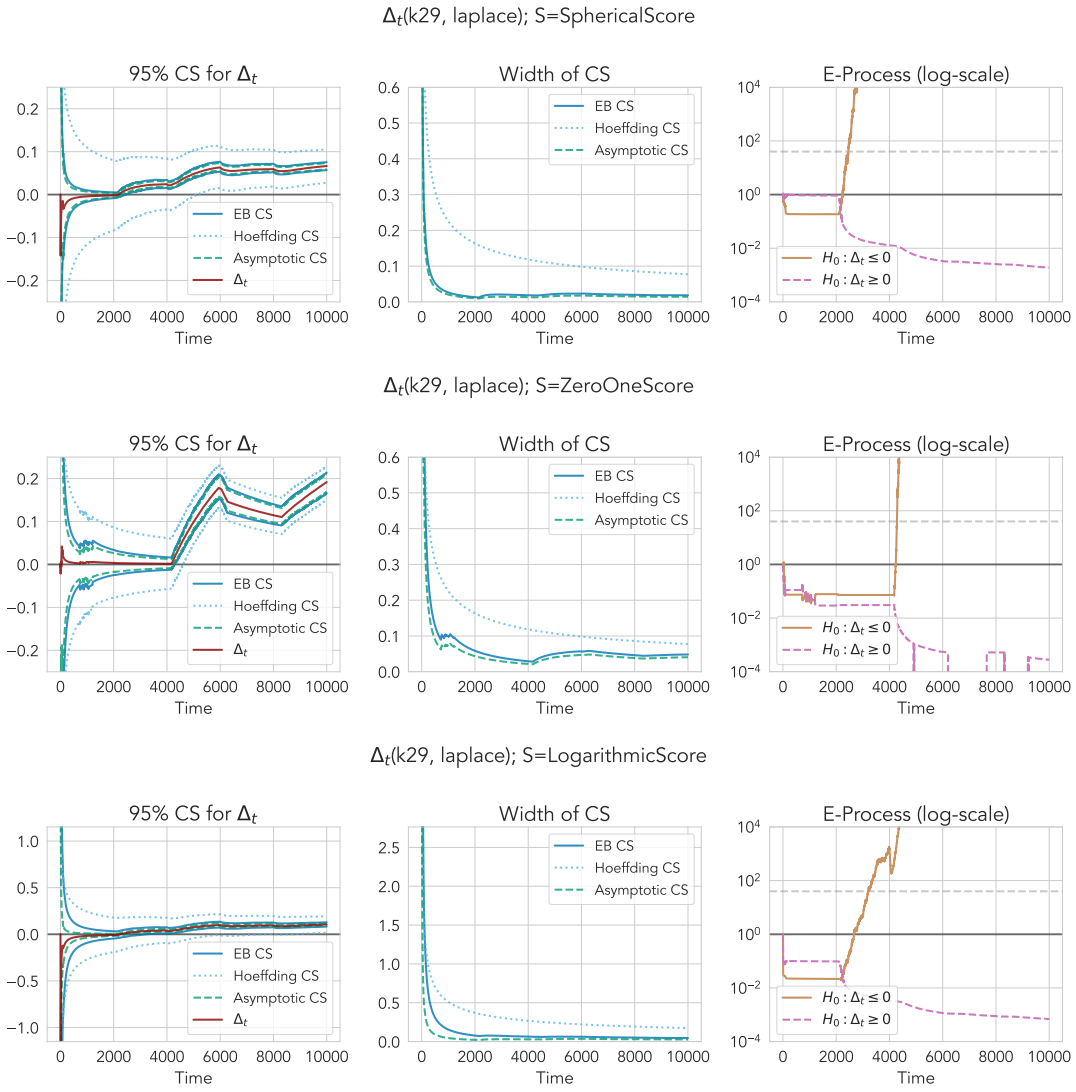


Figure 3.4: 95% EB, Hoeffding-style, and asymptotic CSs (left), their widths (middle), and the sub-exponential e-processes (right) between the K29 forecaster and the Laplace forecaster. Three different scoring rules are used here: the spherical (top), the zero-one (middle), and the ϵ-truncated logarithmic ($\epsilon = 0.01$) (bottom) scores. All scoring rules are positively oriented, such that positive values of Δ_t indicate that the first forecaster is better than the second. Even when the scoring rule is not strictly proper (zero-one) or not proper at all (truncated logarithmic), all CSs still cover Δ_t uniformly, and in general the width of the EB CS shrinks close to the asymptotic CS than the Hoeffding-style CS, which is wider. The e-processes for $\mathcal{H}_0^{\mathsf{w}} : \Delta_t \leq 0$ (green) cross the $2/\alpha$ line (gray) as the lower confidence bound of the EB CS crosses zero.

3.5.2 Comparing Forecasters on Major League Baseball Games

As our first real-world application of the CSs, we consider the problem of predicting wins and losses for baseball games played in the Major League Baseball (MLB). Sports game prediction is particularly suitable for our setting, because there are multiple publicly available probability forecasts on the outcome of each game (e.g., FiveThirtyEight, betting odds, and pundits/experts), that are frequently updated across time. There is also no obvious assumption to be reasonably made about the outcome of the games, such as stationarity or assumptions of parametric models. Recall Table 3.1 for an illustration of various probability forecasts made on MLB games.

We specifically focus on predicting the outcome of MLB games over ten years (2010-2019), culminating in the 2019 World Series between the Houston Astros and the Washington Nationals. We use every regular season and postseason MLB game from 2010 to 2019 as our dataset. We convert each game as a single time point in chronological order, leading to a total of $T = 25,165$ games. As for the forecasters, we consider the following:

- `538`: Game-by-game probability forecasts by FiveThirtyEight on every MLB game since 1871, available at `https://data.fivethirtyeight.com/#mlb-elo`.
- `vegas`: Pre-game closing odds made on each game by online sports bettors, converted and scaled to probabilities, as reported by `https://Vegas-Odds.com`.[9]
- `constant`: a constant baseline corresponding to $p_t = 0.5$ for each t.
- `laplace`: A seasonally adjusted Laplace algorithm, representing the season win percentage for each team. The final adjust win percentage from the previous season, reverted to the mean by one-third, is used as the baseline probability for the next season. The final probability forecast for a game between two teams is rescaled to sum to 1.
- `k29`: The K29 algorithm applied to each team, using the Gaussian kernel with $\sigma = 0.1$, computed using data from the current season only. The final probability forecast for a game between two teams is rescaled to sum to 1.

In Section A.9.2, we give further details about the five forecasters and also plot their forecasts on the last 200 games of 2019.

[9] `https://sports-statistics.com/sports-data/mlb-historical-odds-scores-datasets/`

We perform all pairwise comparisons of the five aforementioned forecasters on the 10-year win/loss predictions. See Sections A.9.4 for details on tuning the free hyperparameter on the uniform boundary. First, as we showed in Figure 3.1, we compare the two publicly available forecasters in `538` (p) and `vegas` (q), finding that the `vegas` forecaster has marginally outperformed the `538` forecaster: after $T = 25,165$ games, 95% EB CS for Δ_T is $(-0.00265, -0.00062)$, and the e-value for $\mathcal{H}_0^{\mathsf{w}}(q,p) : \Delta_t \geq 0,\ \forall t$ is 2979.0. The fact that the `vegas` forecaster (marginally) outperformed the `538` forecaster is interesting, especially given that the primary goal of sports bettors is not to maximize predictive accuracy but their overall profit.[10] Yet, given the relatively small score difference and also the inherent uncertainty in sports game outcomes,[11] more fine-grained comparisons between real-world sports forecasters (e.g., regular season vs. playoffs, team-specific comparisons, and comparisons with or without specific side information) remain interesting future work.

In Table 3.4, we further compare every other forecaster against the `vegas` forecaster by estimating the average Brier score differential Δ_T using the 95% EB CS. We also show the corresponding sub-exponential e-processes (Theorem 3.3) for the null of $\mathcal{H}_0^{\mathsf{w}}(q,p) : \Delta_t \geq 0,\ \forall t$, which translates to saying that `vegas` is not assumed to be better under the null, evaluated at time T. Furthermore, we include comparisons involving the logarithmic score, namely via the average Winkler score $W_T(p,q)$ (Proposition A.4, Section A.4) that quantifies the relative "skill" of forecasters (Winkler, 1994; Lai et al., 2011) as measured by a scoring rule (the logarithmic score, in this case). The Winkler score approach allows us to utilize unbounded proper scoring rules, such as the logarithmic score, when dealing with binary outcomes. Because the score is normalized and thus always maximized at 1, we can construct a one-sided CS with an upper confidence bound (UCB), and also construct an e-process against the null $\mathcal{H}_0^{\mathsf{ww}} : W_t \geq 0,\ \forall t$. A negative UCB or a high value in the e-process indicates that p is significantly worse than q in relative skill.

Our results show that none of the other forecasters, including the `538` forecaster, have outperformed `vegas`, both in terms of the Brier score and the Winkler-logarithmic score.

We include a plot of all pairwise comparisons between the five forecasters in Section A.9.2.

[10] https://fivethirtyeight.com/features/the-imperfect-pursuit-of-a-perfect-baseball-forecast/
[11] https://projects.fivethirtyeight.com/checking-our-work/mlb-games/

Forecaster	C_T^{EB}	E_T
538	(-0.00265, -0.00061)	2979.0
laplace	(-0.00980, -0.00596)	$> 10^4$
k29	(-0.01392, -0.00905)	$> 10^4$
constant	(-0.01115, -0.00713)	$> 10^4$

(a) Δ_T (Brier) against vegas

Forecaster	C_T^{EB}	E_T
538	$(-\infty$, -0.01012)	$> 10^4$
laplace	$(-\infty$, -0.04723)	$> 10^4$
k29	$(-\infty$, -0.14684)	$> 10^4$
constant	$(-\infty$, -0.05165)	$> 10^4$

(b) W_T (Winkler-logarithmic) against vegas

Table 3.4: Comparing forecasters against the vegas forecaster. In (a), we present 95% EB CS for the average Brier score differential $(\Delta_t)_{t=0}^{\infty}$, evaluated at time $T = 25,165$ (i.e., C_T^{EB}), as well as the e-process for the null of $\mathcal{H}_0^{\text{w}}(q,p) : \Delta_t \geq 0, \ \forall t$, also evaluated at time T (i.e., E_T). In (b), we present the analogous table for the average Winkler score W_T (Section A.4), with the logarithmic score as the base score. Note that C_T^{EB} is one-sided due to the one-sided boundedness of W_T. Positive (negative) values of Δ_T and W_T indicate that the forecaster is better (worse) than the baseline. We find that none of the other forecasters, including 538, have outperformed vegas from 2010 to 2019.

3.5.3 Comparing Statistical Postprocessing Methods for Weather Forecasts

As our second real-data experiment, we compare a set of statistical postprocessing methods for weather forecasts (Vannitsem et al., 2021), following the recent work by Henzi and Ziegel (2022). Statistical postprocessing here refers to the process of correcting for biases and dispersion errors in ensemble weather forecasts, which are produced by perturbing the initial conditions of numerical weather prediction (NWP) methods. As ensemble forecasts are commonly used in state-of-the-art weather forecasting systems as a means of producing probabilistic forecasts, statistical postprocessing is considered a key component of modern weather forecasting.

Given 24-hour precipitation data from 2007 to 2017 at four locations (Brussels, Frankfurt, London Heathrow, and Zurich), our goal is to compare three postprocessing methods over time: isotonic distributional regression (IDR; Henzi et al. (2021)), heteroscedastic censored logistic regression (HCLR; Messner et al. (2014)), and a variant of HCLR without its scale parameter (HCLR_). We use the Brier score throughout this section. See Section A.9.3 for details regarding data as well as a plot of the three forecasting methods.

Our main goal here is to sequentially compare the three statistical postprocessing methods using the EB CS and the sub-exponential e-process. As noted in Sections 3.2 and 3.4.4, the inferential conclusions drawn from the sub-exponential e-process (Theorem 3.3) are different from Henzi and Ziegel (2022)'s e-process, which provides a test of conditional forecast dominance at all times (i.e.,

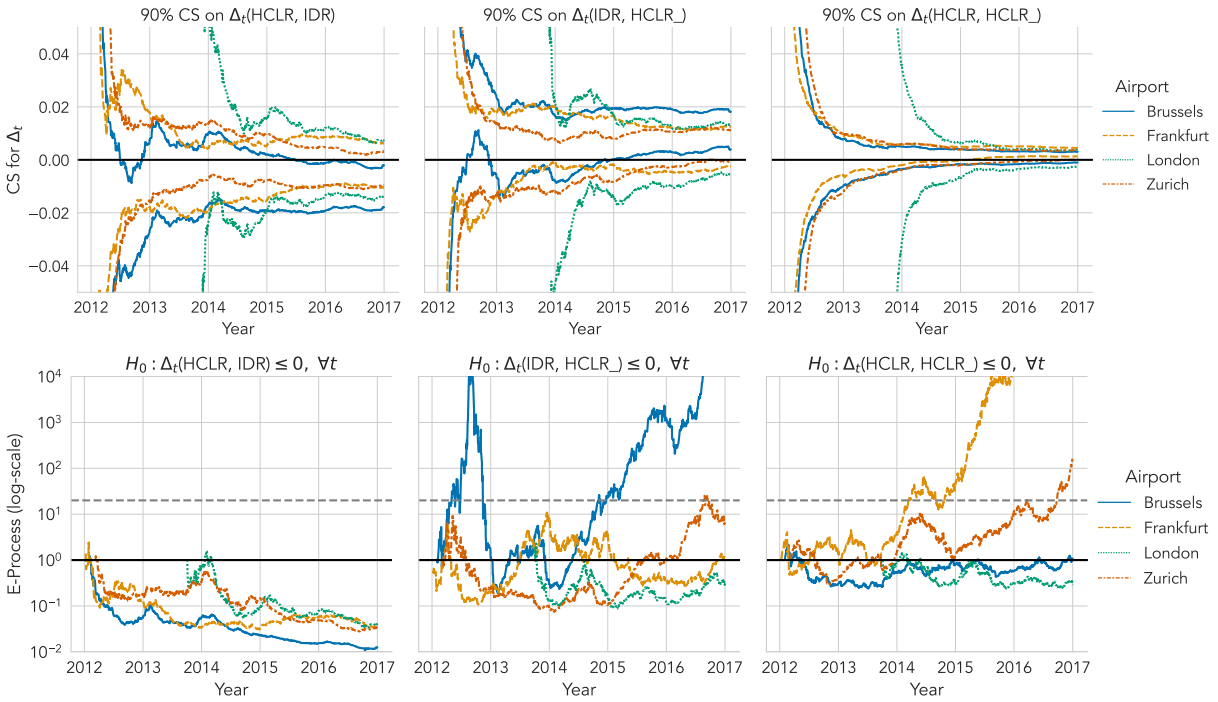


Figure 3.5: *Top:* 90% EB CSs for Δ_t between pairs of statistical postprocessing methods (HCLR and IDR; IDR and HCLR_; HCLR and HCLR_) for 1-day ensemble forecasts using Theorem 3.2, computed and plotted separately for each airport: Brussels ($T = 1,703$), Frankfurt ($T = 1,809$), London ($T = 1,128$), and Zurich ($T = 1,621$). Positive/negative scores of $\Delta_t(p, q)$ indicate that forecaster p is better/worse than forecaster q. Overall, the CSs capture the time-varying score gap on average between the two forecasters across the years. *Bottom:* E-processes for the null that $\mathcal{H}_0^{\mathsf{w}} : \Delta_t \leq 0, \ \forall t$, corresponding to (the lower bound of) the 90% CSs above. These e-processes are the *weak* (average) counterpart to Henzi and Ziegel (2022)'s e-processes for the *strong* (step-by-step) null that $\mathcal{H}_0^{\mathsf{s}} : \delta_t \leq 0 \ \forall t$. Note that the e-processes exceed 10 approximately when the lower bound of the 90% CS exceeds 0. Both procedures use the Brier score as the scoring rule.

the strong null), instead of average (i.e., the weak null). Given that the weak null is larger than the strong null, we would generally expect the sub-exponential e-process for the weak null to be smaller than Henzi and Ziegel (2022)'s e-process for the strong null. On the other hand, the two methods are similar in that they are both valid at arbitrary (data-dependent) stopping times.

In Figure 3.5, we plot both the 90% EB CS on Δ_t (top) as well as the sub-exponential e-processes for the weak one-sided null $\mathcal{H}_0^{\mathsf{w}}$ (bottom), between HCLR and IDR, IDR and HCLR_, and HCLR and HCLR_ on 1-day PoP forecasts at the four airport locations. Note that we compare the same three pairs as Henzi and Ziegel (2022), who compare e-processes for the strong one-sided null $\mathcal{H}_0^{\mathsf{s}}$. The EB CS is computed using Theorem 3.2 and the gamma-exponential mixture boundary (3.12); the analogous

mixture e-processes are then computed using Theorem 3.3. We use the significance level of $\alpha = 0.1$ for the EB CS, corresponding the threshold of $2/\alpha = 20$ for each one-sided e-process.

We first note from Figure 3.5 that the lower bound of our 90% EB CS on $\Delta_t(p, q)$ and the e-process for $\mathcal{H}_0^{\mathsf{w}} : \Delta_t(p, q) \leq 0$ share a similar trend over time, where the e-process grows large when the lower bound grows significantly larger than zero, implying that the forecaster p is better than the forecaster q, using the stopping rule (3.15). Whereas the CS provides a (two-sided) estimate of $\Delta_t(p, q)$ with uncertainty, the e-process explicitly gives the amount of evidence for whether one is better than the other. This illustrates how the two procedures complement each other for anytime-valid inference on Δ_t. We also remark that, although we only plot the e-processes for one-sided null $\mathcal{H}_0^{\mathsf{w}}(p, q)$, we can further compute the e-processes for $\mathcal{H}_0^{\mathsf{w}}(q, p) : \Delta_t(q, p) \leq 0$, and they would correspond to the upper confidence bounds of the EB CSs.

Based on these results, we find from the 90% EB CSs that IDR forecasts are found to outperform both HCLR and HCLR_ 1-day forecasts for Brussels and that HCLR forecasts outperform HCLR_ forecasts for Frankfurt and Zurich, but we do not find significant differences at other locations between other pairs. The e-processes (thresholded at 20) lead to the same conclusions, and they clearly visualize at which point in time is one forecaster first found to outperform the other and how that pattern changes. For example, when comparing IDR to HCLR_ for Brussels, IDR is found to be better as early as 2012, and it also shows the period between late 2012 and late 2015 where it is no longer found to be better, before eventually regaining evidence favoring IDR starting 2016.

When we compare the sub-exponential e-processes for the weak null $\mathcal{H}_0^{\mathsf{w}}$ with the e-processes for the strong null $\mathcal{H}_0^{\mathsf{s}}$, which are drawn in Figure 3 of Henzi and Ziegel (2022), we find that e-processes for the strong null are large whenever e-processes for the weak null are also large, but not vice versa. For example, the comparison of IDR against HCLR_ in Frankfurt is only found to have strong evidence against the strong null, but not the weak null. This is consistent with our previous discussion in Section 3.4.4 that the strong null implies the weak null and thus is easier to "reject" (or gather evidence against). For example, in Frankfurt, we can infer we only have strong evidence that IDR has outperformed HCLR_ *at some point in time* between 2012 and 2017, but we do not have sufficient evidence that IDR has outperformed HCLR_ *on average* in the same time period.

In Section A.5, we include e-processes for comparing lag-h forecasts in the same setting.

3.6 Extensions and Discussion

In the following, we discuss some related points that were not highlighted in previous sections.

On the Use of Unbounded Scoring Rules. Our main results in Theorems 3.2 and 3.3 require the use of bounded scoring rules, which may be restrictive in certain use cases. If the score differentials are unbounded, a general solution would be to use the asymptotic CS (Section A.3), which assumes that only $2 + \delta$ moments are bounded. When it comes to unbounded proper scores for binary outcomes, such as the logarithmic score, the Winkler score (Section A.4), which we used in Section 3.5.2, offers a nonasymptotic and anytime-valid solution.

Comparing Forecasts of Lag $h > 1$. In general forecasting scenarios, we may encounter forecasts that are made $h > 1$ rounds ahead of when the outcome is revealed at time t. In these cases, the expected score differential we seek to estimate should be conditioned on the filtration available at the time of forecasting, rather than the filtration at round $t-1$. We formally derive methods for comparing lag-h forecasts in Section A.5. These include lagged sequential e-values (Arnold et al., 2021), which are not e-processes themselves but can nevertheless quantify the evidence against the weak null (and a "less weak" variant), as well as p-processes and e-processes that are more conservative. The technical details follow the recent discussions by Arnold et al. (2021); Henzi and Ziegel (2022). Constructing a more powerful e-process and also a CS for the lagged weak null remains a challenging problem.

On "Looking Ahead" in Distribution-Free Sequential Inference on Time-Varying Means. Our methods are valid without any assumptions about the time-varying dynamics of the forecast score differentials $(\hat{\delta}_i)_{i=1}^{\infty}$, and in particular we avoid conditions involving stationarity or mixing. A large e-value against $\mathcal{H}_0 : \Delta_t(p, q) \leq 0, \ \forall t$ at some stopping time τ tells us that p has achieved a better conditional predictive performance than q up to τ on average. The utility of comparing forecasters in such a descriptive sense is often significant in the real world: determining a winner in real-world forecasting competitions can often land significant cash prizes (e.g., financial forecasting[12]) and/or media attention (e.g., election and sports forecasting).

[12] https://m6competition.com

This also means that the inferential conclusions drawn from our methods need not extrapolate to *future* time steps, because hypothetically the forecasters or Reality (from Game 3.1) can completely change their behaviors going forward. Indeed, there is a distinction between saying that one *has done* better than the other and that one *is going to be* better than the other in the future — the former is descriptive, while the latter is predictive. All our methods provide evidence and uncertainty related to the former statement. Because we do not make any assumption that says "the future will resemble the past," no method can make conclusive statements about the latter without clairvoyance. Our setup highlights that past performance can be compared in a distribution-free manner, while predictions of future performance will require nontrivial distributional assumptions.

Ultimately, the decision to take the inferential conclusion and extrapolate it toward the future is (and should be) left to the practitioner's own beliefs. If a practitioner opts to make additional assumptions about Reality, then in principle, the conclusions drawn from our methods can extend to settings that the assumptions allow. If one is willing to assume, say, that the score differentials are constant, then the inferential conclusions will straightforwardly extrapolate to future time steps (in the assumed setting). Furthermore, the variance-adaptive EB CS will remain tight, because the underlying variance remains constant. It should be noted that, even under such assumptions, which are often made by classical methods like the Diebold and Mariano (1995) test, anytime-valid approaches avoid the "p-hacking" problem that the classical methods are susceptible to.

Acknowledgements for this Chapter

We thank Alexander Henzi, Johanna F. Ziegel, and Rafael M. Frongillo for their valuable feedback on this chapter. We acknowledge funding from NSF DMS 1916320. Research reported in this chapter was sponsored in part by the DEVCOM Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196 (ARL IoBT CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Chapter 4

Counterfactaully Comparing Abstaining Classifiers

This chapter is based on Choe et al. (2023).

4.1 Introduction

Abstaining classifiers (Chow, 1957; El-Yaniv and Wiener, 2010), also known as selective classifiers or classifiers with a reject option, are classifiers that have the option to abstain from making predictions on certain inputs. As their use continues to grow in safety-critical applications, such as medical imaging and autonomous driving, it is natural to ask how a practitioner should *evaluate and compare* the predictive performance of abstaining classifiers under black-box access to their decisions.

In this chapter, we introduce the *counterfactual score* as a new evaluation metric for black-box abstaining classifiers. The counterfactual score is defined as the expected score of an abstaining classifier's predictions, *had it not been allowed to abstain*. This score is of intrinsic importance when the potential predictions on abstaining inputs are relevant. We proceed with an illustrative example:

Example 4.1 (Free-trial ML APIs)**.** Suppose that we compare different image classification APIs. Each API has two versions: a free version that abstains, and a paid one that does not. Before paying for the full service, the user can query the free version for up to n sample predictions on a user-provided dataset, although it may choose to reject any input that it deems as requiring the paid service. Given

two such APIs, how can the practitioner determine which of the two paid (non-abstaining) versions would be better on the population data source, given their abstaining predictions on a sample?

Example 4.1 exhibits why a user of a black-box abstaining classifier would be interested in its counterfactual score. Although this is a hypothetical example, we can imagine variants of popular application settings in which the counterfactual score is relevant. Specifically, these are settings in which the hidden prediction of an abstaining classifier is meaningful for future uses, or it may be utilized as a backup option in a failure mode. We include three additional examples in Appendix B.1.1.

To formally define, identify, and estimate the counterfactual score, we cast the evaluation problem in Rubin (1976)'s missing data framework and treat abstentions as *missing predictions*. This novel viewpoint directly yields nonparametric methods for estimating the counterfactual score of an abstaining classifier, drawing upon methods for causal inference in observational studies (Rubin, 1974; Robins et al., 1994; van der Vaart, 2000), and represents an interesting yet previously unutilized theoretical connection between selective classification, model evaluation, and causal inference.

The identifiability of the counterfactual score is guaranteed under two standard assumptions: the missing at random (MAR) condition, which is satisfied as long as the evaluation data is independent of the classifier (or its training data), and the positivity condition, both of which are provably unavoidable. We later discuss each condition in detail, including when the positivity condition is met and how a policy-level approach may be necessary for safety-critical applications.

The counterfactual score can be viewed as an alternative to the selective score (mean score on nonabstentions) and the coverage (1 minus the abstention rate) (El-Yaniv and Wiener, 2010), which are the main existing metrics for evaluating black-box abstaining classifiers. As a two-dimensional metric, comparison on the basis of these is non-trivial. A common approach is to assume a fixed cost for each abstention (Chow, 1970), but this is not always satisfactory since determining how to weigh abstentions and errors against one another is a nontrivial question. Thus, in settings such as Example 4.1, the notion of counterfactual score becomes necessary. Importantly, selective scores are not representative of the counterfactual performance, except in the (unrealistic) case wherein predictions are missing completely at random (MCAR).[1] Figure 4.1 gives an overview of scenarios

[1]MCAR means the missing observations are simply a uniformly random subset of all observations, independently of the input/output. In contrast, MAR means there can be systematic differences between the missing and observed values, but these can be explained by the input. Our method only requires MAR.

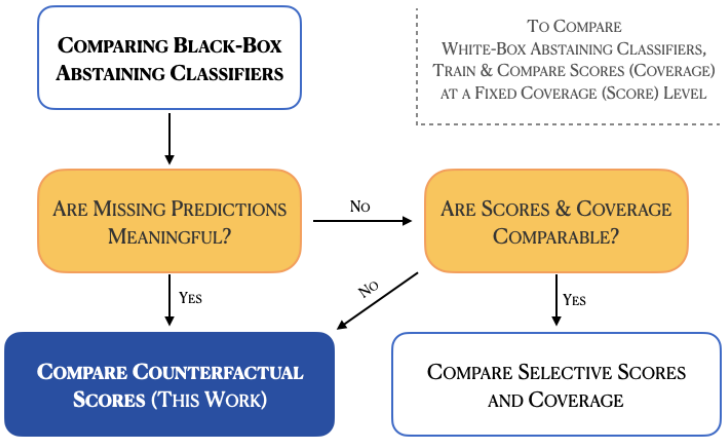


Figure 4.1: A schematic flowchart of comparing abstaining classifiers. In a black-box setting where the evaluator does not have access to the training algorithms or the resources to train them, the task can be viewed as a nontrivial missing data problem. This work proposes the counterfactual score as an evaluation metric.

where different metrics may be appropriate to compare abstaining classifiers.

The counterfactual score also offers practical benefits when comparing abstaining classifiers. Counterfactual scores are comparable even if abstaining classifiers are tuned to different abstention rates or selective scores. Moreover, compared to evaluation methods using the selective score-coverage curve (equivalent to re-training the classifier several times at different score/coverage levels), estimating the counterfactual score does not require re-training the classifier. Instead, we only require estimating a pair of nuisance functions that can be learned using the observed predictions (nonabstentions) in the evaluation set. Let us further note that the setup is applicable generally to any form of prediction that can be scored, including regression and structured prediction. In this chapter, we restrict our attention to classification for concreteness, given that it is the most well-studied abstention framework.

Summary of Contributions. We first formalize the problem of comparing abstaining classifiers as a missing data problem and introduce the counterfactual score as a metric for abstaining classifier comparison. Next, we discuss how the counterfactual score can be identified under the MAR and positivity conditions. Then, we develop efficient nonparametric estimators for the counterfactual scores and their differences, namely doubly robust confidence intervals. We analyze our approach in simulated and real-data experiments. Table 4.1 summarizes the methods developed in our work.

	Evaluation	Comparison
Classifier(s)	(f, π)	$(f^{\mathsf{A}}, \pi^{\mathsf{A}})$ & $(f^{\mathsf{B}}, \pi^{\mathsf{B}})$
Target	$\psi = \mathbb{E}[S]$	$\Delta^{\mathsf{AB}} = \mathbb{E}\left[S^{\mathsf{A}} - S^{\mathsf{B}}\right]$
Identification	MAR & Positivity	
Estimation	Doubly Robust CI	
Optimality	Nonparametrically Efficient	

Table 4.1: A summary of problem formulations and proposed approaches for evaluation and comparison of abstaining classifiers. Our approaches avoid parametric assumptions and allow for black-box classifiers.

Related Work on Abstaining Classifiers. The *training* of abstaining classifiers has seen significant interest in the literature. We refer to Hendrickx et al. (2021); Zhang et al. (2023) for recent surveys.

For *evaluation*, aside from using some combination of selective score and coverage, the most pertinent reference is the work of Condessa et al. (2017), who propose the metric of 'classifier quality' that is a somewhat inverse version of our counterfactual accuracy. This metric is the sum of the prediction accuracy when the classifier predicts, and prediction *inaccuracy* when it abstains, the idea being that if a classifier is not abstaining needlessly, then it must hold that the underlying predictions on points it abstains on are very poor. While this view is relevant to the training of abstention rules, it is at odds with black-box settings where the underlying predictions may still be executed even when the method abstains, motivating the counterfactual score.

Related Work on Missing Data, Causal Inference, and Doubly Robust Estimation. Our main approach to the estimability of counterfactual scores is driven by a reduction to an inference problem under missing data (or censoring) (Rubin, 1976; Little and Rubin, 2019). A missing data problem can be viewed equivalently as a causal inference problem in an observational study (Rubin, 1976; Pearl, 2000; Shpitser et al., 2015; Ding and Li, 2018), and there exist well-established theories and methods for both identifying the counterfactual quantity of interest and efficiently estimating the identified target functional. For *identification*, unlike in some observational settings for estimating treatment effects, our setup straightforwardly satisfies the standard assumption of consistency (non-interference). The MAR assumption is satisfied as long as independent evaluation data is used, and the positivity

assumption translates to abstentions being stochastic. We discuss various implications of these conditions in Sections 4.2.2 and 4.5. For *estimation*, efficient methods for estimating targets such as the average treatment effect (ATE) have long been studied under semiparametric and nonparametric settings. Doubly robust (DR) estimators, in particular, are known to achieve the asymptotic minimax lower bound on the mean squared error. For details, we refer the reader to Bickel et al. (1993); Robins et al. (1994); van der Vaart (2000, 2002); van der Laan and Robins (2003); Bang and Robins (2005); Tsiatis (2006); Chernozhukov et al. (2018); Kennedy (2022). Unlike the standard ATE estimation setup in observational studies, our setup contrasts two separate causal estimands, the counterfactual scores of the two competing classifiers, that operate under their distinct missingness mechanisms.

4.2 Definition and Identification of the Counterfactual Score

We formulate the problem of evaluating and comparing abstaining classifiers under the missing data framework (Rubin, 1976). We follow the standard approach of defining the target parameter (§4.2.1), identifying it with observable quantities (§4.2.2), and estimating the identified parameter using data (§4.3). In each step, we first consider evaluating one abstaining classifier and then extend to comparing two abstaining classifiers. In the following, $\mathcal{X}$ denotes the input space and $\mathcal{Y} = \{1, \dots, C\}$ is the set of possible classes, while Δ^{C-1} denotes the C-dimensional probability simplex on $\mathcal{Y}$.

Abstaining Classifiers. We define an abstaining classifier as a pair of functions (f, π), representing its *base classifier* $f : \mathcal{X} \to \Delta^{C-1}$ and *abstention mechanism* $\pi : \mathcal{X} \to [0, 1]$, respectively. Given a query X, the classifier first forms a preliminary (probabilistic) prediction $f(X)$. Then, potentially using the output $f(X)$, the classifier determines $\pi(X)$, i.e., the abstention probability. Using $\pi(X)$, the classifier then makes the binary abstention decision $R \mid \pi(X) \sim \mathsf{Ber}(\pi(X))$, so that if $R = 1$ ("rejection"), the classifier abstains on the query, and if $R = 0$, it reveals its prediction $f(X)$. In some cases, we will explicitly define the source of randomness ξ (independent of the data) in deciding R, such that $R = \mathsf{r}(\pi(X), \xi)$ for a deterministic function r.[2] Neither f nor π is assumed to be known to the evaluator, modeling the black-box access typically available to practitioners.

[2]Specifically, let $\xi \sim \mathsf{Unif}[0, 1]$ and $R = \mathbb{1}(\xi \leq \pi(X))$. Then, R is a function of only $\pi(X)$ and ξ.

Scoring Rules (Higher Scores Are Better). We measure the quality of a prediction $f(x)$ for a label y via a positively oriented *scoring rule* $\mathsf{s} : \Delta^{C-1} \times \mathcal{Y} \to \mathbb{R}$. One simple scoring rule is classification accuracy, i.e., $\mathsf{s}(f(x), y) = \mathbb{1}\left(\operatorname{argmax}_{c \in \mathcal{Y}} f(x)_c = y\right)$, but a plethora of scores exist in the literature, such as the Brier (1950) score, defined as $\mathsf{s}(f(x), y) = 1 - \sum_{c \in \mathcal{Y}} (f(x)_c - \mathbb{1}(y = c))^2$.

The Evaluation Setup. For each labeled data point (X, Y) in an evaluation set, we observe the abstention decision $R = \mathsf{r}(\pi(X), \xi)$ for some independent source of randomness ξ used by the abstaining classifier. Then, its prediction $f(X)$ is observed by the evaluator if and only if $R = 0$. Let $S := \mathsf{s}(f(X), Y)$ denote the score of the prediction f on the query X, irrespective of R. Because S is not observable when $R = 1$, we refer to S as the *potential score* that *would have been seen* had the classifier not abstained. (See Appendix B.1.2 for equivalent formulations that explicitly invoke Rubin (1974)'s potential outcomes model.) Since our evaluation is based only on the score S, we can suppress the role of Y and assume that S is observed directly when $R = 0$. Similarly, we can suppress the role of ξ, which is independent of the data. We let $\mathbb{P}$ denote the law of $Z := (X, R, S)$.

4.2.1 Definition of the Counterfactual Score

We propose to assess an abstaining classifier (f, π) through its *(expected) counterfactual score*:

$$\psi := \mathbb{E}[S], \tag{4.1}$$

where the expectation is taken w.r.t. $\mathbb{P}$. In words, ψ refers to the expected score of the abstaining classifier had it not been given the option to abstain. The counterfactual score captures the performance of an abstaining classifier via the score of its base classifier, making it suitable for cases where the evaluator is interested in the predictions without using an abstention mechanism.

Note that ψ does *not* in general equal the *selective score*, i.e., $\mathbb{E}[S \mid R = 0]$. For example, when a classifier abstains from making predictions on its "weak points," i.e., inputs on which the classifier performs poorly, the counterfactual score will be lower than the selective score. Also see Appendix B.1.3 for a direct comparison with Condessa et al. (2017)'s score.

Comparison. Counterfactual scores may also be used to compare two abstaining classifiers, $(f^{\mathsf{A}}, \pi^{\mathsf{A}})$ and $(f^{\mathsf{B}}, \pi^{\mathsf{B}})$, in the form of their *counterfactual score difference*: $\Delta^{\mathsf{AB}} := \psi^{\mathsf{A}} - \psi^{\mathsf{B}} = \mathbb{E}[S^{\mathsf{A}} - S^{\mathsf{B}}]$. Here, the expectation is now taken over the joint law of $Z^{\mathsf{AB}} := (X, R^{\mathsf{A}}, S^{\mathsf{A}}, R^{\mathsf{B}}, S^{\mathsf{B}})$.

4.2.2 Identification of the Counterfactual Score

Having defined the target parameters ψ and Δ^{AB}, we now discuss the assumptions under which these quantities become identifiable using only the observed random variables. In other words, these assumptions establish when the counterfactual quantity equals a statistical quantity. As in standard settings of counterfactual inference under missing data, the identifiability of counterfactual scores in this setting depends on two standard conditions: (i) the missing at random condition and (ii) positivity.

The *missing at random (MAR)* condition, also known as the *ignorability* or *no unmeasured confounding* condition, requires that the score S is conditionally independent of the abstention decision R given X, meaning that there are no unobserved confounders U that affect both the abstention decision R as well as the score S. Note that S is the *potential* score of what the classifier would get had it not abstained — it is only observed when $R = 0$. We formally state the MAR condition as follows:

Assumption 4.1 (Scores are missing at random)**.** $S \perp\!\!\!\perp R \mid X$.

In standard ML evaluation scenarios, where the evaluation set is independent of the training set for the classifier, Assumption 4.1 is always met. We formalize this sufficient condition for MAR in the following proposition. Let $\mathcal{D}_{\text{train}}$ denote the collection of any training data used to learn the abstaining classifier (f, π) and, as before, (X, Y) denote an (i.i.d.) data point in the evaluation set.

Proposition 4.1 (Independent evaluation data ensures MAR)**.** *If* $(X, Y) \perp\!\!\!\perp \mathcal{D}_{\text{train}}$*, then* $S \perp\!\!\!\perp R \mid X$.

This result is intuitive: given an independent test input X, the score $S = \mathsf{s}(f(X), Y)$ is a deterministic function of the test label Y, and the abstention decision R of a classifier cannot depend on Y simply because the classifier has no access to it. A short proof is given in Appendix B.2.1 for completeness. In Appendix B.3, we also include causal graphs that visually illustrate how the MAR condition is met.

If the evaluation data is not independent of $\mathcal{D}_{\text{train}}$, then the classifier already has information about the data on which it is tested, so generally speaking, no evaluation score will not accurately

reflect its generalization performance. Although the independence between the training and evaluation data is expected in standard ML applications, it may not be guaranteed when, e.g., using a publicly available dataset that is used during the training of the classifier. These issues can be prevented by ensuring that the evaluation set is held out (e.g., a hospital can use its own patient data to evaluate APIs).

The second condition, the *positivity* condition, is more substantial in our setting:

Assumption 4.2 (Positivity)**.** There exists $\epsilon > 0$ such that $\pi(X) = \mathbb{P}(R = 1 \mid X) \leq 1 - \epsilon$.

Assumption 4.2 says that, for each input X, there has to be at least a small probability that the classifier will *not* abstain ($R = 0$). Indeed, if the classifier deterministically abstains from making predictions on a specific input that has nonzero marginal density, then we have no hope of estimating an expectation over all possible values that X can take. When it comes to evaluating abstaining classifiers on safety-critical applications, we argue that this condition may need to be enforced at a policy level — we elaborate on this point in Section 4.5 and Appendix B.4. In practice, the exact value of ϵ is problem-dependent, and in Appendix B.6.4, we include additional experiments illustrating how our methods retain validity as long as the abstention rate is capped at $1 - \epsilon$ for some $\epsilon > 0$.

Another justification for the positivity condition is that stochastically abstaining classifiers can achieve better performances than their deterministic counterparts. Kalai and Kanade (2021) illustrate how stochastic abstentions can improve the out-of-distribution (OOD) performance w.r.t. the Chow (1970) score (i.e., selective score + $\alpha \cdot$ coverage). Schreuder and Chzhen (2021) also introduce randomness in their abstaining classifiers, which leverage abstentions as a means to improve their accuracy while satisfying a fairness constraint. The role of random abstentions in these examples mirrors the role of randomization in the fairness literature (Barocas et al., 2019), where the optimal randomized fair predictors are known to outperform their deterministic counterparts (Agarwal and Deshpande, 2022; Grgić-Hlača et al., 2017). Given the effectiveness of randomized classifiers for fairness, it would not be surprising if a fair abstaining classifier was randomized (in its decisions and abstentions).

With MAR and positivity in hand, we can show that the counterfactual score is indeed identifiable. Define $\mu_0(x) := \mathbb{E}[S \mid R = 0, X = x]$ as the regression function for the score under $R = 0$.

Proposition 4.2 (Identification)**.** *Under Assumptions 4.1 and 4.2, ψ is identified as* $\mathbb{E}[\mu_0(X)]$.

The proof, included in Appendix B.2.2, follows a standard argument in causal inference. The identification of the target parameter ψ using μ_0 implies that we can estimate ψ, the expectation of a potential outcome, using only the observed inputs and scores. Specifically, the task of estimating ψ consistently reduces to the problem of estimating the regression function μ_0, which only involves predictions that the classifier did not abstain from making. We note that, as in standard causal inference, the task of identification, which concerns *what* to estimate, is largely orthogonal to the task of estimation, which concerns *how* to estimate the quantity. We discuss the latter problem in Section 4.3.

Comparison. For the comparison task, given $\Delta^{\mathsf{AB}} = \psi^{\mathsf{A}} - \psi^{\mathsf{B}}$, it immediately follows that if the MAR and positivity assumptions hold for each of $(X, R^{\mathsf{A}}, S^{\mathsf{A}})$ and $(X, R^{\mathsf{B}}, S^{\mathsf{B}})$, then Δ^{AB} is also identified, and it can be consistently estimated via $\Delta^{\mathsf{AB}} = \mathbb{E}[\mu_0^{\mathsf{A}}(X) - \mu_0^{\mathsf{B}}(X)]$, where $\mu_0^{\bullet}(x) := \mathbb{E}[S^{\bullet} | R^{\bullet} = 0, X = x]$ for $\bullet \in \{\mathsf{A}, \mathsf{B}\}$. A sufficient condition for the identification of Δ^{AB} is that (i) the evaluation data is independent of the training data for either classifier (MAR) and (ii) each classifier has at least a small chance of not abstaining on each input (positivity).

4.3 Nonparametric and Doubly Robust Estimation of the Counterfactual Score

Having identified the counterfactual scores, we now focus on the problem of consistently estimating them. We estimate these quantities without resorting to parametric assumptions about the underlying black-box abstention mechanisms. Instead, we reduce the problem to that of functional estimation and leverage techniques from nonparametric statistics. See Kennedy (2022) for a recent review.

4.3.1 Estimating the Counterfactual Score

Task. Let $\{(X_i, R_i, S_i)\}_{i=1}^n \sim \mathbb{P}$ denote an i.i.d. evaluation set of size n. As before, we assume that we are given access to the censored version of this sample, i.e., that we observe S_i if and only if $R_i = 0$. Using the observables, we seek to form an estimate $\hat{\psi}$ of the counterfactual score $\psi = \mathbb{E}[S]$.

Doubly Robust Estimation. Under identification (Ppn. 4.2), we can estimate ψ by estimating the regression function $\mu_0(X)$ on the data $\{(X_i, S_i) : R_i = 0\}$. However, the naïve "plug-in" estimate suffers from an inflated bias due to structure present in the abstention patterns. (See §B.1.4 for details.) We instead develop a doubly robust (DR) estimator (Robins et al., 1994; Bang and Robins, 2005), which is known to consistently estimate ψ at the optimal *nonparametric efficiency rates*, meaning that no other estimator based on the n observations can asymptotically achieve a smaller mean squared error (van der Vaart, 2002). The derivation below is relatively standard, explaining our brevity.

Formally, the DR estimator is defined using the (uncentered) *efficient influence function (EIF)* for the identified target functional $\psi(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\mu_0(X)]$: $\mathsf{IF}(x, r, s) := \mu_0(x) + \frac{1-r}{1-\pi(x)}(s - \mu_0(x))\,(0/0 := 0)$. Here, π and μ_0 are the "nuisance" functions, representing the abstention mechanism and the score regression function under $R = 0$, respectively. The EIF can be computed as long as s is available when $r = 0$. An intuition for the EIF is that it is the first-order "distributional Taylor approximation" (Fisher and Kennedy, 2021) of the target functional, such that its bias is second-order.

Given that π and μ_0 are unknown, we define an estimate of the EIF, denoted as $\hat{\mathsf{IF}}$, by plugging in estimates $\hat{\pi}$ for π and $\hat{\mu}_0$ for μ_0. Then, the DR estimator is simply the empirical mean of the EIF:

$$\hat{\psi}_{\mathsf{dr}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mathsf{IF}}(X_i, R_i, S_i) = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_0(X_i) + \frac{1 - R_i}{1 - \hat{\pi}(X_i)}(S_i - \hat{\mu}_0(X_i))\right]. \tag{4.2}$$

This estimator is well-defined because S_i is available precisely when $R_i = 0$. Note that the first term is the (biased) plug-in estimator, and the second term represents the first-order correction term, which involves inverse probability weighting (IPW) (Horvitz and Thompson, 1952; Rosenbaum, 1995). In our experiments, we show how the DR estimator improves upon both the plug-in estimator, in terms of the bias, and the IPW-based estimator, which we formally define in §B.1.4, in terms of the variance.

The "double robustness" of $\hat{\psi}_{\mathsf{dr}}$ translates to the following useful property: $\hat{\psi}_{\mathsf{dr}}$ retains the parametric rate of convergence, $O_{\mathbb{P}}(1/\sqrt{n})$, even when the estimators $\hat{\mu}_0$ and $\hat{\pi}$ themselves converge at slower rates. This allows us to use nonparametric function estimators to estimate μ_0 and π, such as stacking ensembles (Breiman, 1996) like the super learner (van der Laan et al., 2007) and regularized estimators like the Lasso (Tibshirani, 1996; Belloni et al., 2014). Even for nonparametric models whose rates of convergence are not fully understood, such as random forests (Breiman, 2001) and deep

neural networks (LeCun et al., 2015), we can empirically demonstrate valid coverage and efficiency (§4.4).

In practice, the nuisance functions can be estimated via *cross-fitting* (Robins et al., 2008; Zheng and van der Laan, 2011; Chernozhukov et al., 2018), which is a K-fold generalization of sample splitting. First, randomly split the data into K folds; then, fit $\hat{\pi}$ and $\hat{\mu}_0$ on $K-1$ folds and use them to estimate the EIF on the remaining "evaluation" fold; repeat the process K times with each fold being the evaluation fold; finally, average the EIFs across all data points. The key benefit of using cross-fitting is to avoid any complexity restrictions on individual nuisance functions without sacrificing sample efficiency. In the following, we let $\hat{\psi}_{\mathsf{dr}}$ be the estimator (4.2) obtained via cross-fitting.

Now we are ready to present our first result, which states the asymptotic validity and efficiency of the DR estimator for ψ under identification and the DR condition.

Theorem 4.1 (DR estimation of the counterfactual score for an abstaining classifier)**.** *Suppose that Assumptions 4.1 and 4.2 hold. Also, suppose that*

$$\|\hat{\pi} - \pi\|_{L_2(\mathbb{P})} \|\hat{\mu}_0 - \mu_0\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1/\sqrt{n}) \tag{4.3}$$

and that $\|\widehat{\mathsf{IF}} - \mathsf{IF}\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1)$*. Then,*

$$\sqrt{n}\left(\hat{\psi}_{\mathsf{dr}} - \psi\right) \rightsquigarrow \mathcal{N}\left(0, \mathsf{Var}_{\mathbb{P}}\left(\mathsf{IF}\right)\right),$$

where $\mathsf{Var}_{\mathbb{P}}\left(\mathsf{IF}\right)$ *matches the nonparametric efficiency bound.*

The proof adapts standard arguments in mathematical statistics as found in, e.g., van der Vaart (2002); Kennedy (2022), to the abstaining classifier evaluation setup. We include a proof sketch in Appendix B.2.3. Theorem 4.1 tells us that, under the identification and the DR condition (4.3), we can construct a closed-form asymptotic confidence interval (CI) at level $\alpha \in (0, 1)$ as follows:

$$C_{n,\alpha} = \left(\hat{\psi}_{\mathsf{dr}} \pm z_{\alpha/2}\sqrt{n^{-1}\mathsf{Var}_{\mathbb{P}_n}\left(\widehat{\mathsf{IF}}\right)}\right), \tag{4.4}$$

where $z_{\alpha/2} = \Phi(1 - \frac{\alpha}{2})$ is the $(1 - \frac{\alpha}{2})$-quantile of a standard normal (e.g., 1.96 for $\alpha = 0.05$). In §B.5, we describe a version that is also valid under continuous monitoring (e.g., as more data is collected).

4.3.2 Estimating Counterfactual Score Differences

Task. Next, we return to the problem of comparing two abstaining classifiers, $(f^{\mathsf{A}}, \pi^{\mathsf{A}})$ and $(f^{\mathsf{B}}, \pi^{\mathsf{B}})$, that each makes a decision to make a prediction on each input X_i or abstain from doing so. That is, $R_i^\bullet \mid \pi^\bullet(X_i) \sim \mathsf{Ber}(\pi^\bullet(X_i))$, and we observe $S_i^\bullet = \mathsf{s}(f^\bullet(X_i), Y_i)$ if and only if $R_i^\bullet = 0$, for $\bullet \in \{\mathsf{A}, \mathsf{B}\}$. Recall that the target here is the score difference $\Delta^{\mathsf{AB}} = \psi^{\mathsf{A}} - \psi^{\mathsf{B}} = \mathbb{E}\left[S^{\mathsf{A}} - S^{\mathsf{B}}\right]$.

Doubly Robust Difference Estimation. If the parameters ψ^{A} and ψ^{B} are each identified according to Ppn. 4.2, then we can estimate Δ^{AB} as $\hat{\Delta}^{\mathsf{AB}} = \hat{\psi}^{\mathsf{A}} - \hat{\psi}^{\mathsf{B}}$, for individual estimates $\hat{\psi}^{\mathsf{A}}$ and $\hat{\psi}^{\mathsf{B}}$. The resulting EIF is simply the difference in the EIF for A and B: $\mathsf{IF}^{\mathsf{AB}}(x, r^{\mathsf{A}}, r^{\mathsf{B}}, s^{\mathsf{A}}, s^{\mathsf{B}}) = \mathsf{IF}^{\mathsf{A}}(x, r^{\mathsf{A}}, s^{\mathsf{A}}) - \mathsf{IF}^{\mathsf{B}}(x, r^{\mathsf{B}}, s^{\mathsf{B}})$, where IF^{A} and IF^{B} denote the EIF of the respective classifier. Thus, we arrive at an analogous theorem that involves estimating the nuisance functions of each abstaining classifier and utilizing $\mathsf{IF}^{\mathsf{AB}}$ to obtain the limiting distribution of $\hat{\Delta}^{\mathsf{AB}}_{\mathsf{dr}} = \hat{\psi}^{\mathsf{A}}_{\mathsf{dr}} - \hat{\psi}^{\mathsf{B}}_{\mathsf{dr}}$.

Theorem 4.2 (DR estimation of the counterfactual score difference)**.** *Suppose that Assumptions 4.1 and 4.2 hold for both* $(X_i, R_i^{\mathsf{A}}, S_i^{\mathsf{A}})$ *and* $(X_i, R_i^{\mathsf{B}}, S_i^{\mathsf{B}})$*. Also, suppose that*

$$\|\hat{\pi}^{\mathsf{A}} - \pi^{\mathsf{A}}\|_{L_2(\mathbb{P})} \|\hat{\mu}_0^{\mathsf{A}} - \mu_0^{\mathsf{A}}\|_{L_2(\mathbb{P})} + \|\hat{\pi}^{\mathsf{B}} - \pi^{\mathsf{B}}\|_{L_2(\mathbb{P})} \|\hat{\mu}_0^{\mathsf{B}} - \mu_0^{\mathsf{B}}\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1/\sqrt{n}) \tag{4.5}$$

and that $\|\hat{\mathsf{IF}}^{\mathsf{AB}} - \mathsf{IF}^{\mathsf{AB}}\|_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1)$*. Then,*

$$\sqrt{n}\left(\hat{\Delta}^{\mathsf{AB}}_{\mathsf{dr}} - \Delta^{\mathsf{AB}}\right) \rightsquigarrow \mathcal{N}\left(0, \mathsf{Var}_{\mathbb{P}}\left(\mathsf{IF}^{\mathsf{AB}}\right)\right),$$

where $\mathsf{Var}_{\mathbb{P}}[\mathsf{IF}^{\mathsf{AB}}]$ *matches the nonparametric efficiency bound.*

A proof is given in Appendix B.2.4. As with evaluation, Theorem 4.2 yields a closed-form asymptotic CI of the form (4.4) using the analogous estimate of EIF under MAR, positivity, and DR (4.5). Inverting this CI in the standard manner further yields a hypothesis test for $H_0 : \psi^{\mathsf{A}} = \psi^{\mathsf{B}}$ vs. $H_1 : \psi^{\mathsf{A}} \neq \psi^{\mathsf{B}}$.

4.4 Experiments

In our experiments, we first present results on simulated data to examine the validity of our proposed inference methods (CIs and hypothesis tests). We then study three scenarios on the CIFAR-100 dataset that illustrate the practical use of our approach to real data settings.

4.4.1 Simulated Experiments: Abstentions Near the Decision Boundary

Setup (MAR but Not MCAR). We first consider comparing two abstaining classifiers according to their accuracy scores, on a simulated binary classification dataset with 2-dimensional inputs. Given $n = 2,000$ i.i.d. inputs $\{X_i\}_{i=1}^n \sim \mathsf{Unif}([0,1]^2)$, each label Y_i is decided using a linear boundary, $f_*(x_1, x_2) = \mathbb{1}(x_1 + x_2 \geq 1)$, along with a 15% i.i.d. label noise. Importantly, each classifier abstains near its decision boundary, such that its predictions and scores are *MAR but not MCAR* (because abstentions depend on the inputs). As a result, while the counterfactual score of A ($\psi^{\mathsf{A}} = 0.86$) is much higher than B ($\psi^{\mathsf{B}} = 0.74$), their selective scores are more similar ($\mathsf{Sel}^{\mathsf{A}} = 0.86$, $\mathsf{Sel}^{\mathsf{B}} = 0.81$) and coverage is lower for A ($\mathsf{Cov}^{\mathsf{A}} = 0.55$) than for B ($\mathsf{Cov}^{\mathsf{B}} = 0.62$). Another point to note here is that, even though both the outcome model and classifier A are linear, both the abstention mechanism[3] π^{A} and the selective score function[4] μ_0^{A} are *non*linear functions of the inputs (similarly for π^{B} and μ_0^{B}). More generally, if a classifier abstains near its decision boundary, then both π and μ_0 could be at least as complex as the base classifier f itself. Further details of the setup, including a plot of the data, predictions, and abstentions, are provided in Appendix B.6.1.

Miscoverage Rates and Widths. As our first experiment, we compare the miscoverage rates and widths of the 95% DR CIs (Theorem 4.2) against two baseline estimators: the plug-in and the IPW (Rosenbaum, 1995) estimators (§B.1.4). For each method, the miscoverage rate of the CI C_n is approximated via $\mathbb{P}(\Delta^{\mathsf{AB}} \notin C_n) \approx m^{-1} \sum_{j=1}^m \mathbf{1}(\Delta^{\mathsf{AB}} \notin C_n^{(j)})$, where m is the number of simulations over repeatedly sampled data. If the CI is valid, then this rate should approximately be 0.05. The miscoverage rate and the width of a CI, respectively, capture its bias and variance components. For

[3]The abstention mechanism $\pi(x) = \mathbb{P}(R = 1 \mid X = x)$ here separates the region below *and* above the decision boundary from the region near the boundary. Thus π is nonlinear even when the boundary is linear.

[4]Given any input X for which the classifier did not abstain ($R = 0$) and its output Y, the score $S = \mathsf{s}(f(X), Y)$ is nonlinear if either $\mathsf{s}(\cdot, y)$ or f is nonlinear. Thus, even for linear f, nonlinear scores like the Brier score automatically make the selective score function $\mu_0(x) = \mathbb{E}[S \mid R = 0, X = x]$ nonlinear.

Nuisance Function Estimators	**Plug-in**	**IPW**	**DR**
Linear/Logistic Regression	1.00 ± 0.00 (0.00)	0.76 ± 0.01 (0.09)	1.00 ± 0.00 (0.04)
Random Forest	0.64 ± 0.02 (0.02)	0.14 ± 0.01 (0.13)	$\mathbf{0.05 \pm 0.01}$ **(0.07)**
Super Learner	0.91 ± 0.01 (0.01)	$\mathbf{0.03 \pm 0.01}$ **(0.12)**	$\mathbf{0.05 \pm 0.01}$ **(0.06)**

Table 4.2: Miscoverage rates (and widths) of 95% CIs using three estimation approaches and three nuisance function (π and μ_0) estimators in a simulated experiment. Mean and standard error computed over $m = 1,000$ runs are shown; those within 2 standard errors of the intended level (0.05) are boldfaced. The sample size is $n = 2,000$ in each run. The mean widths of CIs are shown in parentheses. DR estimation with either a random forest or a super learner achieves control over the miscoverage rate, and the DR-based CI is twice as tight as the IPW-based CI in terms of their width.

the nuisance functions, we try linear predictors (L2-regularized linear/logistic regression for $\hat{\mu}_0/\hat{\pi}$), random forests, and super learners with k-NN, kernel SVM, and random forests.

We present our results in Table 4.2. First, using either the random forest or the super learner, the DR CIs consistently achieve the intended coverage level of 0.95, over $m = 1,000$ repeated simulations (standard error 0.01). This validates the asymptotic normality result of (4.2). Note that the version with linear estimators does not achieve the intended coverage level: this is expected as neither $\hat{\pi}$ nor $\hat{\mu}_0$ can consistently estimate π or μ_0, which are nonlinear functions, and thus violates (4.5).

Second, when considering both the miscoverage rate and CI width, the DR estimator outperforms both the plug-in and IPW estimators. The plug-in estimator, despite having small CI width, has a very high miscoverage rate (0.91 with the super learner), meaning that it is biased even when flexible nuisance learners are used. On the other hand, the IPW estimator has double the width of the DR estimator (0.12 to 0.06, with the super learner), meaning that it is not as efficient. Also, while the IPW estimator achieves the desired coverage level with the super learner (0.03), it fails with the random forest (0.14), which tends to make overconfident predictions of the abstention pattern and biases the resulting estimate. In contrast, the DR estimator retains its intended coverage level of 0.05 with the random forest, suggesting that it is amenable to overconfident nuisance learners.

Power Analysis. We further conduct a power analysis of the statistical test for $H_0 : \Delta^{AB} = 0$ vs. $H_1 : \Delta^{AB} \neq 0$ by inverting the DR CI. The results confirm that the power reaches 1 as either the sample size (n) or the absolute difference ($|\Delta^{AB}|$) increases. This experiment is included in App. B.6.2.

4.4.2 Comparing Abstaining Classifiers on CIFAR-100

To illustrate a real data use case, we compare abstaining classifiers on the CIFAR-100 image classification dataset (Krizhevsky, 2009). Observe that abstaining classifiers can behave differently not just when their base classifiers are different but also when their abstention mechanisms are different. In fact, two abstaining classifiers can have a similarly effective base classifier but substantially different abstention mechanisms (e.g., one more confident than the other). In such a case, the counterfactual score difference between the two classifiers is zero, but their selective scores and coverages are different. We examine such scenarios by comparing image classifiers that use the same pre-trained representation model but have different output layers and abstention mechanisms.

We start with the 512-dimensional final-layer representations of a VGG-16 convolutional neural network (Simonyan and Zisserman, 2015), pre-trained[5] on the CIFAR-100 training set, and compare different output layers and abstention mechanisms on the validation set. Generally, in a real data setup, we cannot verify whether a statistical test or a CI. Yet, in this experiment, we still have access to the base model of each abstaining classifier. This means that (a) if we compare abstaining classifiers that share the base classifier but differ in their abstention patterns, then we actually know that their counterfactual scores are exactly the same ($\Delta^{AB} = 0$); (b) if we compare abstaining classifiers with different base classifiers, then we can compute their counterfactual scores accurately up to an i.i.d. sampling error. This estimate is denoted by $\bar{\Delta}^{AB} := n^{-1} \sum_{i=1}^{n} [\mathrm{s}(f^A(X_i), Y_i) - \mathrm{s}(f^B(X_i), Y_i)]$.

For all comparisons, we use the DR estimator, where the nuisance functions $\hat{\pi}$ and $\hat{\mu}_0$ for both classifiers are each an L2-regularized linear layer learned on top of the pre-trained VGG-16 features. The use of pre-trained representations for learning the nuisance functions is motivated by Shi et al. (2019), who demonstrated the effectiveness of the approach in causal inference contexts. We also use the Brier score in this experiment. We defer other experiment details to Appendix B.6.3.

In scenario `I`, we compare two abstaining classifiers that use the same softmax output layer but

[5]Reproduced version, accessed from `https://github.com/chenyaofo/pytorch-cifar-models`.

Scenarios	Base Classifier	Abstention Rule	$\bar{\Delta}^{\mathsf{AB}}$	95% DR CI	Reject H_0?
I	Same	Different	0.000	(-0.005, 0.018)	No
II	Same	Different	0.000	(-0.014, 0.008)	No
III	Different	Same	−0.029	(-0.051, -0.028)	Yes

Table 4.3: The 95% DR CIs and their corresponding hypothesis tests for $H_0 : \Delta^{\mathsf{AB}} = 0$ at significance level $\alpha = 0.05$, for three different comparison scenarios on (half of) the CIFAR-100 test set ($n = 5,000$). The three scenarios compare different abstention mechanisms or predictors, as detailed in text; all comparisons use the Brier score. $\bar{\Delta}^{\mathsf{AB}}$ is the empirical counterfactual score difference without any abstentions. The result of each statistical test agrees with whether $\bar{\Delta}^{\mathsf{AB}}$ is 0.

use a different threshold for abstentions. Specifically, both classifiers use the softmax response (SR) thresholding (Geifman and El-Yaniv, 2017), i.e., abstain if $\max_{c \in y} f(X)_c < \tau$ for a threshold $\tau > 0$, but A uses a more conservative threshold ($\tau = 0.8$) than B ($\tau = 0.5$). As a result, while their counterfactual scores are identical ($\Delta^{\mathsf{AB}} = 0$), A has a higher selective score (+0.06) and a lower coverage (−0.20) than B. This is also a deterministic abstention mechanism, potentially challenging the premises of our setup. As shown in Table 4.3, we see that the 95% DR CI is $(-0.005, 0.018)$ ($n = 5,000$), confirming that there is no difference in counterfactual scores.[6]

Scenario II is similar to scenario I, except that the abstention mechanisms are now stochastic: A uses the SR as the probability of making a prediction, i.e., $\pi^{\mathsf{A}}(x) = 1 - \max_{c \in [C]} f(X)_c$, while B uses the Gini impurity as the probability of abstention, i.e., $\pi^{\mathsf{B}}(x) = 1 - \sum_{c=1}^{C} p_c^2$, both clipped to $(0.2, 0.8)$. This results in a higher coverage for A and a higher selective score for B because the Gini impurity is typically smaller than the SR. The 95% DR CI is $(-0.014, 0.008)$, confirming that there is once again no difference in counterfactual scores. These two scenarios correspond to case (a).

In scenario III, we now examine a case where there *is* a difference in counterfactual scores between the two abstaining classifiers (case (b)). Specifically, we compare the pre-trained VGG-16 model's output layers (512-512-100) with the single softmax output layer that we considered in earlier scenarios. It turns out that the original model's multi-layer output model achieves a worse Brier score (0.758) than one with a single output layer (0.787), likely because the probability predictions of the multi-layer model are too confident (miscalibrated). When using the same abstention mech-

[6]The reason why the DR CI correctly estimates the true $\Delta^{\mathsf{AB}} = 0$, despite the fact that positivity is violated in this case, is because the two classifiers happen to abstain on similar examples (B abstains whenever A does) *and* their scores on their abstentions happen to be relatively similar (0.604 for A; 0.576 for B).

anism (stochastic abstentions using SR, as in II), the overconfident original model correspondingly achieves a higher coverage (and worse selective Brier score) than the single-output-layer model. The Monte Carlo estimate of the true counterfactual score difference is given by $\bar{\Delta}^{\mathsf{AB}} = -0.029$, and the 95% DR CI falls entirely negative with $(-0.051, -0.028)$, rejecting the null of $\Delta^{\mathsf{AB}} = 0$ at $\alpha = 0.05$.

4.5 Limitations and Discussion

This chapter lays the groundwork for addressing the challenging problem of comparing black-box abstaining classifiers in a counterfactual sense. Our solution casts the problem in Rubin (1976)'s missing data framework, where we treat abstentions as missing-at-random predictions of the classifier(s). This allows us to leverage nonparametrically efficient, doubly-robust tools from causal inference.

There are important future directions and limitations stemming from our framework. On a conceptual level, the largest challenge arises from the positivity condition, which requires the classifiers to deploy a non-deterministic abstention mechanism. As mentioned in §4.2.2, the counterfactual score is unidentifiable without this assumption. We argue that, especially in auditing scenarios, this issue calls for a policy-level treatment, in which vendors must supply evaluators with a classifier that is can abstain but must have at least an $\epsilon > 0$ chance of nonabstention, a level that can be mutually agreed upon by both parties. This achieves a middle ground where the vendors are not required to fully reveal their proprietary classifiers. In §B.4, we discuss this policy-level treatment in further detail.

An alternative is to explore existing techniques that aim to address positivity violations directly (Petersen et al., 2012; Ju et al., 2019; Léger et al., 2022). Due to the unidentifiability result, these techniques have their own limitations. For example, we may consider applying sample trimming (Crump et al., 2009) and making inferences on a subpopulation, although the conclusions would not hold for the entire input domain (even when relevant). Exploring these options is left as future work.

Acknowledgements for this Chapter

We thank Edward H. Kennedy and anonymous reviewers for their helpful comments and suggestions. The work in this chapter used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges-2 system, which is supported by NSF award number ACI-1928147, at the Pittsburgh Supercomputing Center (PSC).

Appendix A

Supplementary Materials for "Comparing Sequential Forecasters"

A.1 Main Proofs

A.1.1 Sub-exponential Test Supermartingales for Time-Varying Means

The proofs of Theorems 3.2 and 3.3 are both based on a variance-adaptive test supermartingale that uniformly bounds sums of random variables that are bounded from below. We first derive this test supermartingale (which, by definition, is also an e-process itself) and use the result for the proofs of the main theorems in the following subsections.

We start by revisiting a useful lemma for the sub-exponential processes. Recall from Section 3.4.3 that $\psi_{E,c}(\lambda) = c^{-2}(-\log(1-c\lambda) - c\lambda),\ \forall\lambda \in [0, 1/c)$ is the exponential CGF-like function. By the proof of Lemma 4.1 in Fan et al. (2015), for any $\lambda \in [0, 1/c)$ and any $\xi \geq -c$,

$$\exp\left\{\lambda\xi - \psi_{E,c}(\lambda)\xi^2\right\} \leq 1 + \lambda\xi. \tag{A.1}$$

Note that the original proof uses $c = 1$, but it straightforwardly generalizes to any value of $c > 0$. To see this, for any $c > 0$, set $\tilde{\lambda} = c\lambda \in [0, 1)$ and $\tilde{\xi} = c^{-1}\xi \geq -1$. Then, applying the lemma with $c = 1$ using $(\tilde{\lambda}, \tilde{\xi})$ gives the desired result.

Now, we show a time-uniform sub-exponential boundary that is generally applicable to sums of

random variables that are bounded from below. This is an extension of Lemma 3(e) from Howard et al. (2020), which also utilizes (A.1). We note that a similar extension is utilized in the recent work of Waudby-Smith et al. (2022) but without the predictable bounds $(c_i)_{i=1}^{\infty}$.

In the following, let $(X_i)_{i=1}^{\infty}$ be any process whose conditional means $\mu_i := \mathbb{E}_{i-1}[X_i]$ exist. Let $(S_t)_{t=0}^{\infty}$ be its cumulative deviations from the conditional means, i.e., $S_0 = 0$ and $S_t = \sum_{i=1}^{t}(X_i - \mu_i)$. Note that S_t is a martingale, i.e., $\mathbb{E}_{t-1}[S_t] = S_{t-1}$. Also, let $(\hat{V}_t)_{t=0}^{\infty}$ be a nondecreasing variance process of the form $\hat{V}_0 = 0$ and $\hat{V}_t = \sum_{i=1}^{t}(X_i - \gamma_i)^2$, where $(\gamma_i)_{i=1}^{\infty}$ is a predictable process. Also, we take $1/\infty = 0$ and, with a slight abuse of notation, $[0, 0) = \{0\}$.

Proposition A.1 (Sub-exponential test supermartingales for time-varying means)**.** *Suppose that there exists a predictable positive sequence $(c_i)_{i=1}^{\infty}$ such that $X_i - \gamma_i \geq -c_i$ a.s. for all $i \geq 1$. Then,*

$$L_t(\lambda) = \prod_{i=1}^{t} \exp\left\{\lambda(X_i - \mu_i) - \psi_{E,c_i}(\lambda)\,(X_i - \gamma_i)^2\right\} \tag{A.2}$$

is a test supermartingale for each $\lambda \in [0, 1/c_0)$, where $c_0 = \sup_{i\geq 1} c_i$.

Proof. For each $i \geq 1$, it suffices to show that

$$\mathbb{E}_{i-1}\left[\exp\left\{\lambda(X_i - \mu_i) - \psi_{E,c_i}(\lambda)(X_i - \gamma_i)^2\right\}\right] \leq 1. \tag{A.3}$$

Let $\tilde{X}_i = X_i - \mu_i$ and $\tilde{\gamma}_i = \gamma_i - \mu_i$. Then, $\tilde{X}_i - \tilde{\gamma}_i = X_i - \gamma_i \geq -c_i$ a.s. by assumption. By (A.1),

$$\exp\left\{\lambda(\tilde{X}_i - \tilde{\gamma}_i) - \psi_{E,c_i}(\lambda)(\tilde{X}_i - \tilde{\gamma}_i)^2\right\} \leq 1 + \lambda(\tilde{X}_i - \tilde{\gamma}_i). \tag{A.4}$$

Multiplying each side by $\exp\{\lambda\tilde{\gamma}_i\}$ and rearranging terms, we get

$$\exp\left\{\lambda\tilde{X}_i - \psi_{E,c_i}(\lambda)(\tilde{X}_i - \tilde{\gamma}_i)^2\right\} \leq e^{\lambda\tilde{\gamma}_i}(1 - \lambda\tilde{\gamma}_i) + e^{\lambda\tilde{\gamma}_i}\lambda\tilde{X}_i \leq 1 + e^{\lambda\tilde{\gamma}_i}\lambda\tilde{X}_i, \tag{A.5}$$

where in the second inequality we used the fact that $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$.

Finally, we take the conditional expectation $\mathbb{E}_{i-1}$ on each side. Because $\mathbb{E}_{i-1}[\tilde{X}_i] = \mathbb{E}_{i-1}[X_i - \mu_i] =$

0, and also because $(\gamma_i)_{i=1}^{\infty}$ and $(c_i)_{i=1}^{\infty}$ are predictable, we get

$$\mathbb{E}_{i-1}\left[\exp\left\{\lambda\tilde{X}_i - \psi_{E,c_i}(\lambda)(\tilde{X}_i - \tilde{\gamma}_i)^2\right\}\right] \leq 1 + e^{\lambda\tilde{\gamma}_i}\lambda\mathbb{E}_{i-1}\left[\tilde{X}_i\right] = 1. \tag{A.6}$$

Substituting back in $\tilde{X}_i = X_i - \mu_i$ and $\tilde{X}_i - \tilde{\gamma}_i = X_i - \gamma_i$, we get the desired result. □

Proposition A.1 is stated for a general setting in which bounds on the pointwise score differentials can vary across time, as long as they form a predictable sequence. If there is a constant $c \in (0, \infty)$ such that $|\hat{\delta}_i| \leq \frac{c}{2}$, such as in Theorems 3.2 and 3.3, then we can simply choose $c_i = c$ for all i and further simplify the expression (A.2) to

$$L_t(\lambda) = \exp\left\{\lambda S_t - \psi_{E,c}(\lambda)\hat{V}_t\right\}, \quad \forall\lambda \in [0, 1/c). \tag{A.7}$$

We return to the case of using non-constant predictable bounds in Section A.6.2.

A.1.2 Proof of Theorem 3.2

The proof is a direct consequence of Proposition A.1, applied once each to the lower and upper confidence bounds.

The stated conditions imply that $\hat{\delta}_i - \gamma_i \geq -c$ a.s. for all $i \geq 1$. Define $S_t = \sum_{i=1}^{t}(\hat{\delta}_i - \delta_i)$. Then, by Proposition A.1, the process

$$L_t^{\mathsf{lcb}}(\lambda) = \exp\left\{\lambda S_t - \psi_{E,c}(\lambda)\hat{V}_t\right\} \tag{A.8}$$

is a test supermartingale for $\lambda \in [0, 1/c)$. By definition, this implies that $(S_t)_{t=0}^{\infty}$ is sub-$\psi_{E,c}$ ("sub-exponential with scale c") with variance process $(\hat{V}_t)_{t=0}^{\infty}$, and thus we have

$$\mathbb{P}\left(\exists t \geq 1 : S_t \geq u_{\alpha/2}(\hat{V}_t)\right) \leq \alpha/2, \tag{A.9}$$

for any sub-exponential uniform boundary (3.7) with crossing probability $\alpha/2$ and scale c, denoted here as $u_{\alpha/2}$. Using the fact that $\frac{1}{t}S_t = \frac{1}{t}\sum_{i=1}^{t}\hat{\delta}_i - \frac{1}{t}\sum_{i=1}^{t}\delta_i = \hat{\Delta}_t - \Delta_t$, we can divide each side of the inequality by t to obtain the lower confidence bound (LCB).

Similarly, the conditions also imply that $-\hat{\delta}_i + \gamma_i \geq -c$, so Proposition A.1 also implies that the process

$$L_t^{\mathrm{ucb}}(\lambda) = \exp\left\{\lambda(-S_t) - \psi_{E,c}(\lambda)\hat{V}_t\right\} \tag{A.10}$$

is also a test supermartingale for $\lambda \in [0, 1/c)$, or equivalently, $(-S_t)_{t=0}^{\infty}$ is sub-$\psi_{E,c}$ with the same variance process $(\hat{V}_t)_{t=0}^{\infty}$. Applying the same argument to $L_t^{\mathrm{ucb}}(\lambda)$ gives the analogous upper confidence bound (UCB) using the *same* uniform boundary $u_{\alpha/2}$.

Finally, combining the lower and upper confidence bounds with a union bound, we obtain the CS:

$$\mathbb{P}\left(\forall t \geq 1 : \left|\hat{\Delta}_t - \Delta_t\right| < \frac{u(\hat{V}_t)}{t}\right) \geq 1 - \alpha. \tag{A.11}$$

A.1.3 Proof of Theorem 3.3

We state and prove a slightly more general version of Theorem 3.3 that only assumes the empirical score differentials $\hat{\delta}_i$ are bounded from *below* and the predictable estimates γ_i are bounded (or truncated) from *above*. Theorem 3.3 assumes that the score differentials are bounded from below *and* above, so applying the following proposition twice to $(\hat{\delta}_i, \gamma_i)_{i=1}^{\infty}$ and $(-\hat{\delta}_i, -\gamma_i)_{i=1}^{\infty}$ will give us the result.

Proposition A.2. *Suppose that $\hat{\delta}_i \geq -\frac{c}{2}$ for each $i \geq 1$, for some $c \in (0, \infty)$. Also, let $(\gamma_i)_{i=1}^{\infty}$ be any predictable sequence and $\hat{V}_t = \sum_{i=1}^{t}(\hat{\delta}_i - \overline{\gamma}_i)^2$, where $\overline{\gamma}_i = \gamma_i \wedge \frac{c}{2}$. Then, for each $\lambda \in [0, 1/c)$, the process $(E_t(\lambda))_{t=0}^{\infty}$ defined as $E_0(\lambda) = 1$ and*

$$E_t(\lambda) := \exp\left\{\lambda \sum_{i=1}^{t} \hat{\delta}_i - \psi_{E,c}(\lambda)\hat{V}_t\right\} \quad \text{is an e-process for } \mathcal{H}_0^{\mathrm{w}}(p, q). \tag{A.12}$$

Proposition A.2 tells us that, if the pointwise empirical score differentials are bounded from below (or above), then we can derive a sub-exponential e-process for $\mathcal{H}_0(p, q)$ (or $\mathcal{H}_0(q, p)$). An important use case for the more general scenario is when using the Winkler score (Winkler, 1994), which is bounded from above by 1 but unbounded from below, as we describe in Section A.4.

Proof of Proposition A.2. First, note that $(E_t(\lambda))_{t=0}^{\infty}$ is an adapted process w.r.t. $\mathfrak{G}$ (and also consists of empirical quantities only). Let $S_t = \sum_{i=1}^{t}(\hat{\delta}_i - \delta_i) = t(\hat{\Delta}_t - \Delta_t)$. Since $\hat{\delta}_i - \overline{\gamma}_i \geq -c$ for all $i \geq 1$,

Proposition A.1 implies that

$$L_t(\lambda) := \exp\{\lambda S_t - \psi_E(\lambda)\hat{V}_t\} \tag{A.13}$$

is a test supermartingale for each $\lambda \in [0, 1/c)$.

Now, under any $P \in \mathcal{H}_0^{\mathrm{w}}(p, q)$, we have that $\exp\left\{-\lambda \sum_{i=1}^t \delta_i\right\} \geq 1$, so for any $t \geq 1$,

$$\begin{aligned} L_t(\lambda) &= \exp\left\{\lambda \sum_{i=1}^t \hat{\delta}_i - \psi_E(\lambda)\hat{V}_t\right\} \exp\left\{-\lambda \sum_{i=1}^t \delta_i\right\} \\ &\geq \exp\left\{\lambda \sum_{i=1}^t \hat{\delta}_i - \psi_E(\lambda)\hat{V}_t\right\} = E_t(\lambda). \end{aligned} \tag{A.14}$$

In other words, for each $P \in \mathcal{H}_0^{\mathrm{w}}(p, q)$, the process $(E_t(\lambda))_{t=0}^{\infty}$ is upper-bounded by the test supermartingale $(L_t(\lambda))_{t=0}^{\infty}$ at all times t. This implies that $(E_t(\lambda))_{t=0}^{\infty}$ is an e-process for $\mathcal{H}_0^{\mathrm{w}}(p, q)$, by Corollary 22 of Ramdas et al. (2020). □

A.2 Details on Time-Uniform Boundary Choices

A.2.1 Computing the Gamma-Exponential Mixture

Here, we derive a closed-form expression (up to efficiently computable gamma functions) for the gamma-exponential mixture, which is used in both the mixture boundary for the CS (Equation (3.12)) and in the mixture e-process for the weak null (Theorem 3.3). The mixture takes the following form:

$$m(s, v) := \int \exp\{\lambda s - \psi_{E,c}(\lambda)v\} f_\rho(\lambda) d\lambda, \tag{A.15}$$

where f_ρ, for any $\rho > 0$, is a reparametrized Gamma density $f_\rho(\lambda) = C(\rho)(1-\lambda)^{\rho-1}e^{-\rho(1-\lambda)}$, $\lambda \in [0, 1/c)$, where $C(\rho) = \frac{\rho^\rho}{\underline{\gamma}(\rho,\rho)\Gamma(\rho)}$ is the normalization constant, $\Gamma(a, z) := \int_z^\infty u^{a-1}e^{-u}du$ is the upper incomplete gamma function, $\Gamma(a) := \Gamma(a, 0)$ is the gamma function, and $\underline{\gamma}$ is the regularized lower incomplete gamma function:

$$\underline{\gamma}(a, z) := \frac{1}{\Gamma(a)} \int_0^z u^{a-1}e^{-u}du, \quad \forall a, z > 0. \tag{A.16}$$

Both Γ and $\underline{\gamma}$ can be computed efficiently in standard scientific computing software. (E.g., $\underline{\gamma}$ can be computed using `boost::math::gamma_p` in C++ and `scipy.special.gammainc` in Python.)

We note here that all time-uniform boundaries have a "tradeoff of tightness" across different (intrinsic) times (Howard et al., 2021), so that it is natural to have a hyperparameter that controls at what intrinsic time we want the resulting CS width to be optimized. In the above, the single hyperparameter, $\rho > 0$, can be related to the user-specified optimal intrinsic time v_{opt} (and the significance level α) via the mapping $\rho = -v_{\text{opt}}(W_{-1}(-\alpha^2/e) + 1)$, where W_{-1} is the lower branch of the Lambert W function. As described in Proposition 3 of Howard et al. (2021), this choice of ρ uniquely minimizes the width function $v \longmapsto u(v)/\sqrt{v}$, when u is the two-sided normal mixture boundary, and it is also known to also provide a good approximation for the (one-sided) gamma-exponential mixture boundary in practice.

The first part of the following proposition is essentially a restatement of Proposition 9 in Howard et al. (2021); the second part additionally provides an upper bound for the mixture when $s \ll 0$ (e.g., the mixture e-process when data supports the null).

Proposition A.3 (Gamma-exponential mixture for e-processes)**.** *Fix $c > 0$ and $\rho > 0$. Consider any values of $s \in \mathbb{R}$ and $v \geq 0$. If $\frac{cs+v+\rho}{c^2} > 0$, then*

$$m(s,v) = C\left(\frac{\rho}{c^2}\right) \frac{\Gamma\left(\frac{v+\rho}{c^2}\right)\underline{\gamma}\left(\frac{v+\rho}{c^2}, \frac{cs+v+\rho}{c^2}\right)}{\left(\frac{cs+v+\rho}{c^2}\right)^{\frac{v+\rho}{c^2}}} \exp\left\{\frac{cs+v}{c^2}\right\}; \tag{A.17}$$

otherwise, if $\frac{cs+v+\rho}{c^2} < 0$, then

$$m(s,v) \leq C\left(\frac{\rho}{c^2}\right) \frac{\exp\left\{-\frac{\rho}{c^2}\right\}}{\frac{v+\rho}{c^2}} \leq 1. \tag{A.18}$$

This is precisely the formula for the sub-exponential mixture e-process in Theorem 3.3: $E_t^{\text{mix}} = m(\sum_{i=1}^t \hat{\delta}_i, \hat{V}_t)$ with f_ρ being the mixture density. It makes sense that $m(s,v)$ is upper-bounded by 1 when $\frac{cs+v+\rho}{c^2} < 0$, because $s < -\frac{v+\rho}{c} < 0$ would imply that the sum of score differentials is negative, supporting the weak null. In our implementation, we use the first upper bound in (A.18), which can be computed efficiently and get substantially smaller than 1 when $v \gg 0$.

Proof of Proposition A.3. For simplicity, we assume $c = 1$. The proof is analogous for any $c > 0$.

Recall that $\psi_E(\lambda) = -\log(1-\lambda) - \lambda$ for $\lambda \in [0,1)$. For any $\rho > 0$,

$$\begin{aligned} m(s,v) &= C(\rho) \int_0^1 \exp\{\lambda s - \psi_E(\lambda) v\} \cdot (1-\lambda)^{\rho-1} e^{-\rho(1-\lambda)} d\lambda \\ &= C(\rho) \int_0^1 e^{\lambda(s+v)} (1-\lambda)^v \cdot (1-\lambda)^{\rho-1} e^{-\rho(1-\lambda)} d\lambda \\ &= C(\rho) \int_0^1 (1-\lambda)^{v+\rho-1} e^{\lambda(s+v)-\rho(1-\lambda)} d\lambda \\ &= C(\rho) \left(\int_0^1 (1-\lambda)^{v+\rho-1} e^{-(s+v+\rho)(1-\lambda)} d\lambda \right) e^{s+v}, \end{aligned} \tag{A.19}$$

where in the last equality we used

$$\lambda(s+v) - \rho(1-\lambda) = (s+v) - (1-\lambda)(s+v) - (1-\lambda)\rho = -(s+v+\rho)(1-\lambda) + (s+v).$$

Now, let $a = v + \rho$ and $z = s + v + \rho$, and note that $a > 0$.

Case 1: $z = s + v + \rho > 0$**.** Using the change-of-variable formula $u = (s+v+\rho)(1-\lambda) = z(1-\lambda)$, we have that

$$\begin{aligned} m(s,v) &= C(\rho) \left(\int_z^0 \left(\frac{u}{z}\right)^{a-1} e^{-u} \frac{du}{-z} \right) e^{s+v} \\ &= C(\rho) \cdot \frac{1}{z^a} \left(\int_0^z u^{a-1} e^{-u} du \right) e^{s+v} \qquad \text{(A.20)} \\ &= C(\rho) \frac{\Gamma(a)\underline{\gamma}(a,z)}{z^a} e^{s+v}, \qquad \text{(A.21)} \end{aligned}$$

where we use the fact that the integral in (A.20) corresponds to the numerator of the lower incomplete gamma function $P(a,z)$ in (A.16). The expression (A.21) can be computed in closed-form.

Case 2: $z = s+v+\rho < 0$. Using the change-of-variable formula $u = -(s+v+\rho)(1-\lambda) = -z(1-\lambda)$, we obtain

$$\begin{aligned} m(s,v) &= C(\rho)\left(\int_{-z}^{0}\left(\frac{u}{-z}\right)^{a-1} e^{u}\frac{du}{z}\right)e^{s+v} \\ &= C(\rho)\cdot\frac{1}{(-z)^a}\left(\int_0^{-z} u^{a-1}e^u du\right)e^{s+v} \\ &= C(\rho)\cdot\frac{1}{|z|^a}\left(\int_0^{|z|} u^{a-1}e^u du\right)e^{s+v}. \end{aligned} \tag{A.22}$$

Although the integral in (A.22) is no longer a regularized lower incomplete gamma function, we can still show that $m(s,v)$ is upper-bounded by 1. Since $e^u \le e^{|z|} = e^{-z}$ for $u \le |z|$, we have that

$$\begin{aligned} m(s,v) &\le C(\rho)\cdot\frac{1}{|z|^a}\left(\int_0^{|z|} u^{a-1}du\right)e^{-z}\cdot e^{s+v} \\ &= C(\rho)\cdot\frac{1}{|z|^a}\left(\int_0^{|z|} u^{a-1}du\right)e^{-\rho} \qquad \text{(A.23)} \\ &= C(\rho)\cdot\frac{1}{|z|^a}\left(\frac{u^a}{a}\right)\Big|_0^{|z|} e^{-\rho} \\ &= \frac{C(\rho)e^{-\rho}}{v+\rho}, \qquad \text{(A.24)} \end{aligned}$$

where in (A.23) we used $-z + (s+v) = -(s+v+\rho) + (s+v) = -\rho$, and in (A.24) we substituted in $a = v+\rho$. We can further bound this value, using the fact that $v > 0$ and substituting back in $C(\rho)$:

$$\begin{aligned} m(s,v) \le \frac{C(\rho)e^{-\rho}}{v+\rho} &\le \frac{C(\rho)e^{-\rho}}{\rho} \\ &= \rho^{\rho-1}e^{-\rho}\cdot\left(\int_0^{\rho} u^{\rho-1}e^{-u}du\right)^{-1} \\ &\le \rho^{\rho-1}e^{-\rho}\cdot\left(e^{-\rho}\int_0^{\rho} u^{\rho-1}du\right)^{-1} \qquad \text{(A.25)} \\ &= \rho^{\rho-1}\cdot\left[\left(\frac{u^\rho}{\rho}\right)\Big|_0^{\rho}\right]^{-1} \\ &= 1, \qquad \text{(A.26)} \end{aligned}$$

where in (A.25) we used the fact that $e^{-\rho} \le e^{-u}$ for $u \in [0,\rho]$. □

A.2.2 The Polynomial Stitching Boundary

The *polynomial stitched boundary* (Theorem 1, Howard et al. (2021)) provides a fully closed-form (without any gamma functions) alternative to the aforementioned gamma-exponential mixture boundary. It is constructed by finding a smooth analytical upper bound on a sequence of linear uniform bounds across different timesteps. The boundary asymptotically grows with $O(\sqrt{v \log\log v})$ rate, matching the form of the law of the iterated logarithm (LIL). For example, a 95% EB CS for Δ_t (Theorem 3.2) using the polynomial stitching boundary is given as follows (assuming $|\hat{\delta}_i| \leq 1,\ \forall i$):

$$\hat{\Delta}_t \pm 2 \cdot \frac{1.7\sqrt{\left(\hat{V}_t \vee 1\right)\left(\log\log\left(2\left(\hat{V}_t \vee 1\right)\right) + 3.8\right)} + 3.4 \log\log\left(2\left(\hat{V}_t \vee 1\right)\right) + 13}{t} \tag{A.27}$$

where $\hat{V}_t$ is the intrinsic time.

The polynomial stitched boundary can be applied to both Theorems 3.1 and 3.2 by setting $\hat{V}_t = t$ and $\hat{V}_t = \sum_{i=1}^{t} (\hat{\delta}_i - \gamma_i)^2$ respectively. Previous work showed that the polynomial stitched boundary is a sub-gamma uniform boundary (Theorem 1, Howard et al. (2021)), which is also a "universal" sub-ψ uniform boundary for any CGF-like function ψ (Proposition 1, Howard et al. (2020)). We omit a full restatement of Howard et al. (2021)'s Theorem 1, which establishes the validity of the polynomial stitching boundary, but rather, we list its three hyperparameters for practical use:

- $v_{\text{opt}} > 0$ determines the value of the intrinsic time at which the boundary is tightest;
- $s > 1$ controls how the crossing probability is distributed over intrinsic time;
- $\eta > 1$ controls the geometric spacing of the intrinsic time.

Throughout this work, we fix $s = 1.4$ and $\eta = 2$, as recommended by the original paper, and only adjust v_{opt}, which serves the analogous role as the hyperparameter of the same name for the gamma-exponential boundary in Section A.2.1.

Although the stitching boundary is computed in closed form and matches the LIL rate, it is usually not as tight as the CM boundary in practice, and thus we use the CM boundary as our default in all of our main experiments.

A.3 Asymptotic CSs for Sequential Forecast Comparison

In their recent work, Waudby-Smith et al. (2021) introduce a new class of time-uniform CSs called asymptotic CSs, which trade the nonasymptotic guarantee of a standard CS for applicability to a wider variety of scenarios, e.g., estimating the average treatment effect in causal inference (for which a nonasymptotic CS is not known). Formally, a sequence of confidence intervals $(\hat{\theta}_t \pm R_t^{\mathsf{A}})_{t=1}^{\infty}$ is a $(1-\alpha)$-*asymptotic CS (AsympCS)* for $(\theta_t)_{t=1}^{\infty}$ if there exists a nonasymptotic $(1-\alpha)$-CS $(\hat{\theta}_t \pm R_t^{\mathsf{NA}})_{t=1}^{\infty}$ for $(\theta_t)_{t=1}^{\infty}$, such that

$$R_t^{\mathsf{NA}} / R_t^{\mathsf{A}} \xrightarrow{a.s.} 1. \tag{A.28}$$

Furthermore, the AsympCS has an *approximation rate* of $r(t)$ if $R_t^{\mathsf{NA}} - R_t^{\mathsf{A}} = O_{a.s.}(r(t))$. Definition (A.28) says that, as $t \to \infty$, the AsympCS is an "arbitrarily precise approximation" of the nonasymptotic CS, and it can be viewed as approximately satisfying the time-uniform coverage property when t is large.

Waudby-Smith et al. (2021) describes an asymptotic CS for time-varying means that can be applied to our setting of estimating $(\Delta_t)_{t=1}^{\infty}$ under Lyapunov CLT-type conditions. For the sake of completeness, we include the (simplified) assumptions and the resulting closed form of the asymptotic CS, adapted to our setting and notations.

Let $\sigma_t^2 = \mathbb{E}_{t-1}[(\hat{\delta}_t - \delta_t)^2]$ denote the conditional variance, $V_t = \sum_{i=1}^{t} \sigma_i^2$ be the cumulative conditional variance, and $\tilde{\sigma}_t^2 = t^{-1} V_t$ be the average. Let $\hat{\sigma}_t^2$ be any estimator of σ_t^2, such as $\hat{\sigma}_t^2 = t^{-1} \sum_{i=1}^{t} (\hat{\delta}_i - \hat{\Delta}_{i-1})^2$. (Notice that, in the setting of Theorem 3.2, $\hat{\sigma}_t^2 = t^{-1} \hat{V}_t$ with γ_i set to $\hat{\Delta}_{i-1}$.) Now, we assume the following:

(a) $\tilde{\sigma}_t^2 \xrightarrow{a.s.} \sigma_*^2$ for some $\sigma_*^2 > 0$;

(b) there exists $q > 2$ such that the q^{th} moments of $\hat{\delta}_t$ is uniformly bounded (a.s.) for all $t \geq 1$; and

(c) $\hat{\sigma}_t^2 / \tilde{\sigma}_t^2 \xrightarrow{a.s.} 1$.

As noted in the paper, these conditions can be substantially more general than either sub-Gaussianity or boundedness. Given these assumptions, we know by Theorem 2.3 of Waudby-Smith et al. (2021)

that, for any $\rho > 0$ and any $\alpha \in (0, 1)$,

$$C_t^{\mathsf{A}} := \left(\hat{\Delta}_t \pm \sqrt{\frac{2(t\hat{\sigma}_t^2\rho^2 + 1)}{t^2\rho^2} \log\left(\frac{\sqrt{t\hat{\sigma}_t^2\rho^2 + 1}}{\alpha} \right)} \right) \tag{A.29}$$

forms a $(1-\alpha)$-AsympCS for $(\Delta_t)_{t=1}^{\infty}$ with an approximation rate of $o(\sqrt{V_t \log V_t}/t)$. $\rho > 0$ is a hyperparameter that affects the relative tightness of the CS across time, analogous to the hyperparameter ρ in Section A.2. In our experiments, we follow Waudby-Smith et al. (2021) (Equation 74) and use the choice that approximately optimizes the width at a pre-specified time $t^* \geq 1$:

$$\rho(t^*) = \sqrt{\frac{2\log(1/\alpha) + \log(1 + 2\log(1/\alpha))}{t^*}}. \tag{A.30}$$

Unless specified otherwise, t^* is chosen to be 100 in our experiments.

As illustrated in Figures 3.3 and 3.4, the AsympCS is typically tighter than the EB CS (Theorem 3.2) for smaller values of t, and as t grows large the widths of the two CSs become close to one another.

A.4 Comparing Relative Forecasting Skills Using the Winkler Score

In a typical forecast comparison scenario, we are often interested in comparing a newly developed forecasting algorithm (say, p) with an existing baseline (say, q). For example, a company that already deploys a daily forecasting algorithm may want to A/B test if its newly developed method is at least as good as the existing one. In such settings, we may be interested in the *relative* improvement of a forecaster over a baseline, and early work by Murphy (1988) and Winkler (1994) propose using normalized scoring rules that better reflect the relative "skill" of the new forecaster.

In this section, we show how our main results can be extended in a unique way to construct time-uniform CSs and e-processes for the *average Winkler score* (Winkler, 1994), which is a normalized version of the average score differentials between probability forecasts on binary outcomes. Interestingly, these results yield SAVI approaches that are valid *without* a boundedness or sub-Gaussianity assumption on the underlying scoring rule, and instead they are valid whenever the scoring rule is

proper (Gneiting and Raftery, 2007). The Winkler score is particularly useful when comparing probability forecasters based on the logarithmic score, which is a strictly proper but unbounded score, as we showcased in Section 3.5.2. We remark that Lai et al. (2011) first showed the asymptotic normality of the average Winkler score. In contrast to their work, the methods we develop here are nonasymptotic and anytime-valid, depending only on the natural upper bound (of 1) on the Winkler score; we also allow the baseline forecaster to be nonconstant.

Formally, we first define the *(pointwise) Winkler score* $w(p, q, y)$ with a base scoring rule S as follows:

$$w(p, q, y) := \frac{S(p, y) - S(q, y)}{S(p, \mathbb{1}(p > q)) - S(q, \mathbb{1}(p > q))}, \quad p, q \in (0, 1),\ y \in \{0, 1\}, \tag{A.31}$$

where we set $0/0 := 0$. We note that (A.31) is equivalent to the increment in the e-process of (Henzi and Ziegel, 2022) (details in Section A.8.1), and thus we can interpret Henzi and Ziegel (2022)'s e-process for the strong null as betting directly proportional to the relative forecasting skill between the forecasters. We also define the *expected (pointwise) Winkler score* as

$$w(p, q; r) := \mathbb{E}_{y\sim r}[w(p, q, y)] = \frac{\mathbb{E}_{y\sim r}[S(p, y)] - \mathbb{E}_{y\sim r}[S(q, y)]}{S(p, \mathbb{1}(p > q)) - S(q, \mathbb{1}(p > q))}, \tag{A.32}$$

for $p, q \in (0, 1)$ and $r \in [0, 1]$. As before, $y \sim r$ denotes $y \sim \text{Bernoulli}(r)$ (conditional on p and q). It follows directly from (A.32) that, given a constant forecaster $q \in (0, 1)$, $S_q^w(p, y) = w(p, q, y)$ itself is a (strictly) proper scoring rule if S is (strictly) proper (Winkler, 1994). The score is also standardized in the sense that, if q is the "least skillful" calibrated forecaster, i.e., the constant, historical-average forecaster (*climatology* in weather forecasting), and p is another well-calibrated forecaster, then the expected Winkler score $w(p, q; r)$ is zero (minimum) when $p = q$ and one (maximum) when $p \in \{0, 1\}$. On the other hand, the empirical Winkler score $w(p, q, y)$ can take negative values, which would suggest that p is worse than q on forecasting the outcome y under S.

In the following lemma, we summarize the characteristics of the Winkler score that are useful for both its interpretation and the proofs that will follow shortly.

Lemma A.1 (Winkler (1994)). *Let S be a proper scoring rule. Then, for any $p, q \in (0,1)$ and $y \in \{0,1\}$,*

$$w(p,q,y) = \begin{cases} 1 & \text{if } y = \mathbb{1}(p > q); \\ \leq 0 & \text{otherwise.} \end{cases} \tag{A.33}$$

In the case that $y \neq \mathbb{1}(p > q)$, the denominator is non-negative and the numerator is non-positive.

See Winkler (1994, 1977) for a proof. Lemma A.1 establishes that p gets a positive score of 1 if it is at least as good as q, but otherwise, it does not get a positive score. Two implications are: (i) the Winkler score is bounded from above by 1, and (ii) when we take the average of pointwise Winkler scores over t forecasts and outcomes, we can read off the sign of the average to tell whether p has better or worse forecasting skills than q.

Returning to the sequential setup in Game 3.1, we now treat the pointwise Winkler scores between $(p_t)_{t=1}^\infty$ and $(q_t)_{t=1}^\infty$ as the analogs of pointwise score differentials from Section 3.4. Because $(p_t)_{t=1}^\infty$ and $(q_t)_{t=1}^\infty$ are predictable w.r.t. $\mathfrak{G}$, we replace the expectation in (A.32) with the conditional expectation w.r.t. $\mathcal{G}_{t-1}$. Then, for each t, we can define the *(expected) average Winkler score* up to t:

$$W_t := \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}_{t-1}[w(p_i, q_i, y_i)], \quad t \geq 1. \tag{A.34}$$

This is the time-varying sequence of parameters that we seek to estimate; we also analogously define the *weak Winkler (WW)* null

$$\mathcal{H}_0^{\text{ww},\geq}(p,q) : W_t \geq 0, \quad \forall t \geq 1. \tag{A.35}$$

For this null, the sign is the opposite of (3.14): we assert that p is at least as good as q as our null, and rejecting $\mathcal{H}_0^{\text{ww},\geq}(p,q)$ would mean that p is decidedly worse than q on average up to some time t. Note also that we slightly generalize the average score from Winkler (1994)'s to allow the baseline forecaster to be any predictable $(0,1)$-valued forecaster $(q_t)_{t=1}^\infty$.

We are now ready to present our main result. In the following, we denote the (empirical) pointwise Winkler scores as $\hat{w}_i = w(p_i, q_i, y_i)$ for each i and their average over time as $\hat{W}_t := \frac{1}{t}\sum_{i=1}^t w(p_i, q_i, y_i)$.

Proposition A.4 (Sequential inference on the average Winkler score). *Suppose that S is a proper scoring rule and that $p_i, q_i \in (0,1)$ for each $i \geq 1$. Let $(\underline{\gamma}_i)_{i=1}^\infty$ be a $[-1,\infty)$-valued predictable process*

and let $\hat{V}_t = \sum_{i=1}^{t} (\hat{w}_i - \underline{\gamma}_i)^2$.

1. *(One-sided EB CS for* $(W_t)_{t=1}^{\infty}$*.) For each* $\alpha \in (0,1)$*, the sequence of intervals* $(C_t^{\mathsf{EB}})_{t=1}^{\infty}$ *defined as*

$$C_t^{\mathsf{EB}} := \left(-\infty, \hat{W}_t + t^{-1} u_\alpha(\hat{V}_t)\right) \cap (-\infty, 1] \tag{A.36}$$

is a $(1-\alpha)$*-CS for* $(W_t)_{t=1}^{\infty}$*, for any sub-exponential uniform boundary* u_α *with crossing probability* α *and scale* 2*.*

2. *(Sub-exponential e-process for* $\mathcal{H}_0^{\mathsf{ww},\geq}$*.) For each* $\lambda \in [0, 1/2)$*, the process* $(E_t(\lambda))_{t=0}^{\infty}$ *defined as* $E_0(\lambda) = 1$ *and*

$$E_t(\lambda) := \exp\left\{-\lambda \hat{W}_t - \psi_{E,2}(\lambda)\hat{V}_t\right\} \tag{A.37}$$

is an e-process for $\mathcal{H}_0^{\mathsf{ww},\geq} : W_t \geq 0, \ \forall t$*, and so is the mixture process* $E_t^{\mathsf{mix}} := \int E_t(\lambda) dF(\lambda)$ *for any distribution* F *on* $[0, 1/c)$*.*

The proof is a direct application of Proposition A.1, using the upper bound of 1 on the empirical pointwise Winkler scores. Because the Winkler score is unbounded from below, the standard machinery only readily provides the upper confidence bound for $(W_t)_{t=1}^{\infty}$. Thus, we derive a one-sided CS in (A.36) that tells us the certainty to which we know W_t is away from 1. The sub-exponential e-process in (A.37) corresponds to this upper confidence bound and measures the evidence against the null that p is at least as good as q. From the sequential testing point-of-view, either a large value in the e-process or a small value of the upper confidence bound suggests that p *under*performs q; conversely, either a small value in the e-process or a value close to 1 for the upper confidence bound (i.e., a vacuous CS) tells us that there is no such evidence. Note that, to satisfy the constraint on the predictable process $(\gamma_i)_{i=1}^{\infty}$ to be bounded from below by -1, we can choose as default the running average as in Theorem 3.2, but cap it from below at -1, i.e., $\gamma_i = -1 \vee \hat{W}_{i-1}$.

Proof of Proposition A.4. We first use Lemma A.1 to obtain an upper bound of 1 on the pointwise empirical Winkler scores, $w_i = w(p_i, q_i, y_i)$. Then, the rest of the proof follows similarly from the proofs of Proposition A.1 as well as Theorem 3.2 and Theorem 3.3.

Specifically, define the process $(L_t(\lambda))_{t=0}^{\infty}$ as $L_0(\lambda) = 1$ and

$$L_t(\lambda) := \exp\left\{\lambda\left(-\hat{W}_t + W_t\right) - \psi_{E,2}(\lambda)\hat{V}_t\right\}, \tag{A.38}$$

which is a test supermartingale an w.r.t. $\mathfrak{G}$ for each $\lambda \in [0, 1/2)$ by Proposition A.1 and Lemma A.1. By definition, the process $(t(W_t - \hat{W}_t))_{t=0}^{\infty}$ is sub-exponential with scale 2 (i.e., sub-$\psi_{E,2}$) having the variance process $(\hat{V}_t)_{t=0}^{\infty}$. The results then follow analogously to Theorems 3.2 and 3.3. □

We close with the note that, if the main goal is rather to tightly estimate $(W_t)_{t=1}^{\infty}$ from both sides or to test the null $\mathcal{H}_0^{\mathsf{ww},\leq} : W_t \leq 0, \ \forall t$, then there is a way to use either the sub-Gaussianity or the boundedness assumption on scoring rules (rather than propriety) and apply any of our main Theorems; the proof would be analogous for each application. The caveat with the Winkler score is that it is unbounded from below even when using a bounded base scoring rule, such as the Brier score, because the lower bound depends on how close q can get to 0 or 1. If $q_t = q \in (0, 1)$ is the climatology forecaster, then this is not an issue and the two-sided approach can also be useful. We summarize the analogs of Theorem 3.2 and Theorem 3.3 for the average Winkler score as a corollary.

Corollary A.1 (Two-sided sequential inference on the average Winkler score.)**.** *Suppose there exists some $c > 0$ such that $\hat{w}_i \geq 1 - c$ for any $i \geq 1$. Let $(\gamma_i)_{i=1}^{\infty}$ be a $[1 - c, 1]$-valued predictable process and let $\hat{V}_t = \sum_{i=1}^{t}(\hat{w}_i - \gamma_i)^2$. Then,*

1. *(Two-sided EB CS for $(W_t)_{t=1}^{\infty}$.) For each $\alpha \in (0, 1)$, the sequence of intervals $(C_t^{\mathsf{EB}})_{t=1}^{\infty}$ defined as*

$$C_t^{\mathsf{EB}} := \left(\hat{W}_t \pm t^{-1}u_{\alpha/2}(\hat{V}_t)\right) \cap (-\infty, 1] \tag{A.39}$$

   is a $(1 - \alpha)$-CS for $(W_t)_{t=1}^{\infty}$, for any sub-exponential uniform boundary $u_{\alpha/2}$ with crossing probability $\alpha/2$ and scale c.

2. *(Sub-exponential e-process for $\mathcal{H}_0^{\mathsf{ww},\leq}$.) For each $\lambda \in [0, 1/c)$, the process $(E_t(\lambda))_{t=0}^{\infty}$ defined as $E_0(\lambda) = 1$ and*

$$E_t(\lambda) := \exp\left\{\lambda\hat{W}_t - \psi_{E,c}(\lambda)\hat{V}_t\right\} \tag{A.40}$$

   is an e-process for $\mathcal{H}_0^{\mathsf{ww},\leq} : W_t \leq 0, \ \forall t$, and so is the mixture process $E_t^{\mathsf{mix}} := \int E_t(\lambda)dF(\lambda)$ for any distribution F on $[0, 1/c)$.

The value of c may depend on both the choice of S and how close q_i can get to either 0 or 1. For example, if S is the Brier score and $q_i \in [q_0, 1-q_0]$ for some constant $q_0 \in (0,1)$, then $c = 2/q_0$.

A.5 Comparing Lagged Forecasts

Given an integer lag $h \geq 1$, if p_i and q_i were lag-h forecasts made at round i for the eventual outcome y_{i+h-1}, then we would be interested in the following time-varying parameter:

$$\Delta_t^{(h)} := \frac{1}{t-h+1} \sum_{i=1}^{t-h+1} \mathbb{E}_{i-1}\left[S(p_i, y_{i+h-1}) - S(q_i, y_{i+h-1})\right], \quad \forall t \geq h. \tag{A.41}$$

For each $t \geq h$, we take the average up to the $(t-h+1)$th round, because the forecasts made beyond that round can only be evaluated after the tth round. The conditional expectation is taken in such a way that the forecasters (p_i and q_i) are evaluated based on the information they had at the time of forecasting ($\mathcal{G}_{i-1}$) and not the one right before the outcome is realized ($\mathcal{G}_{i+h-1}$).

The case of $h = 1$ corresponds to the setting we considered in Section 3.4, but extending the construction to the case of $h > 1$ is not straightforward. For example, the sequence $(E_t(\lambda))_{t=0}^{\infty}$ defined analogously to the one in Theorem 3.3 would *not* be an e-process w.r.t. the game filtration $\mathfrak{G}$, let alone a process, because the tth term would include future outcomes that are not realized at time t. Rather, the process $(E_t(\lambda))_{t=0}^{\infty}$ now only satisfies the weaker property that $\mathbb{E}_{t-h}[E_t] \leq 1$ for all (non-stopping) times $t \geq h$ under $\mathcal{H}_0$. In their recent work, Arnold et al. (2021) refer to such processes as *sequential e-values for $\mathcal{H}_0$ at lag h* and propose to combine h subsequences of the original process that are each test supermartingales w.r.t. different sub-filtrations of $\mathfrak{G}$.

Although lag-h sequential e-values are not e-processes themselves, the recent preprints of Arnold et al. (2021); Henzi and Ziegel (2022) show that there is a workaround to turn them into an e-process possessing anytime-validity. Here, we adapt their approach and develop e- and p-processes for weaker nulls similar to the weak null in the lag-1 case; developing a tight CS for estimating $\Delta_t^{(h)}$ remains an open problem.

To proceed, we define two weak nulls related to the sequence of parameters $(\Delta_t^{(h)})_{t=h}^{\infty}$. The first is

a straightforward generalization of the lag-1 weak null (3.14) to any $h \geq 1$:

$$\mathcal{H}_0^{\mathsf{w}}(p, q; h) : \Delta_t^{(h)} \leq 0, \quad \forall t \geq h. \tag{A.42}$$

This recovers $\mathcal{H}_0^{\mathsf{w}}(p, q)$ when $h = 1$. We refer to (A.42) as the *lag-h weak null* between p and q.

Because of the aforementioned challenge in the $h > 1$ case, we also define a null hypothesis for which we can derive a more powerful e-process. The *lag-h period-wise (PW) weak null*, which we denote as $\mathcal{H}_0^{\mathsf{pw}}(p, q; h)$, asserts that the weak null holds at every hth step for all periods $k \in \{1, \dots, h\}$, making it (slightly) stronger than the weak null but weaker than the strong null.

Formally, define the index set

$$I_t^{[k]} = \left\{ k + 1 + hs : s = 0, 1, \dots, \left\lfloor \frac{t-k}{h} \right\rfloor - 1 \right\}, \tag{A.43}$$

which includes every hth round of the game starting at $k+1$ up to (at most) $t-h+1$. (For $t < h+k$, $I_t^{[k]} = \emptyset$.) Now, for each $k = 1, \dots, h$, we define $\Delta_t^{[k]} := \frac{1}{t-h+1} \sum_{i \in I_t^{[k]}} \delta_i$, so that $\sum_{k=1}^{h} \Delta_t^{[k]} = \Delta_t^{(h)}$. Then, the lag-$h$ PW weak null is defined as

$$\mathcal{H}_0^{\mathsf{pw}}(p, q; h) : \Delta_t^{[k]} \leq 0, \quad \forall t \geq h, \ \forall k = 1, \dots, h. \tag{A.44}$$

It is clear from their definitions that the following inclusion relationships hold between the three null hypotheses:

$$\mathcal{H}_0^{\mathsf{w}}(h) \supseteq \mathcal{H}_0^{\mathsf{pw}}(h) \supseteq \mathcal{H}_0^{\mathsf{s}}(h) \tag{A.45}$$

for any $h \geq 1$. When h is a small integer (say, 5 or 10) and t grows large, the lag-h PW weak null is still much weaker than the lag-h strong null.

Having defined the two nulls, we first present an e-process and a p-process for the lag-h PW null (A.44). Because we cannot straightforwardly derive an e-process for $h > 1$, we start with a p-process constructed using the lag-h sequential e-values and then use a p-to-e calibrator (Shafer et al., 2011) to obtain an e-process that remains valid at arbitrary stopping times. An analogous proposition for (A.42) is shown later and relies on similar proof techniques.

Let $\hat{\delta}_i^{(h)} = S(p_i, y_{i+h-1}) - S(q_i, y_{i+h-1})$ be the empirical pointwise score differential for lag-h fore-

casts. Note that $\delta_i^{(h)} = \mathbb{E}_{i-1}[\hat{\delta}_i^{(h)}]$. In addition, we say that a function $f : [0,1] \to [0,\infty)$ is a *p-to-e calibrator* if it is non-increasing and satisfies $\int_0^1 f(u)du = 1$.

Proposition A.5 (Sequential inference for $\mathcal{H}_0^{\mathsf{pw}}(h)$)**.** *Suppose that* $|\hat{\delta}_i^{(h)}| \le \frac{c}{2}$ *for all* $i \ge 1$*, for some* $c \in (0,\infty)$*. Let* $(\gamma_i)_{i=1}^{\infty}$ *be a* $[-\frac{c}{2}, \frac{c}{2}]$*-valued predictable process w.r.t.* $\mathfrak{G}$*. Also, for each* $k \in \{1,\dots,h\}$ *and* $\lambda \in [0, 1/c)$*, define*

$$E_t^{[k]}(\lambda) = \prod_{i \in I_t^{[k]}} \exp\left\{\lambda\hat{\delta}_i^{(h)} - \psi_{E,c}(\lambda)\left(\hat{\delta}_i^{(h)} - \gamma_i\right)^2\right\}, \quad \forall t \ge 0, \tag{A.46}$$

where $\prod_{i\in\emptyset}(\cdot) = 1$*. Then, for each* $\lambda \in [0, 1/c)$*, the following statements are true:*

1. *(Averaged sequential e-values.) The process*

$$\bar{E}_t^{\mathsf{pw}}(\lambda) := \frac{1}{h}\sum_{k=1}^{h} E_t^{[k]}(\lambda), \quad \forall t \ge 0, \tag{A.47}$$

   is adapted w.r.t. $\mathfrak{G}$ *and satisfies* $\mathbb{E}_P[\bar{E}_{\tau+h-1}^{\mathsf{pw}}(\lambda)] \le 1$ *for any* $\mathfrak{G}$*-stopping time* τ *and any* $P \in \mathcal{H}_0^{\mathsf{pw}}(p,q;h)$*.*

2. *(P-process.) The process* $(\mathsf{p}_t^{\mathsf{pw}})_{t=1}^{\infty}$ *defined by*

$$\mathsf{p}_t^{\mathsf{pw}} := \frac{he\log h}{\sum_{k=1}^{h}\left(1/\mathsf{p}_t^{[k]}\right)}, \quad \textit{where} \quad \mathsf{p}_t^{[k]} := 1 \wedge \left(1/\sup_{i\le t} E_i^{[k]}(\lambda)\right), \quad \forall t \ge 0, \tag{A.48}$$

   is a p-process for $\mathcal{H}_0^{\mathsf{pw}}(p,q;h)$ *w.r.t.* $\mathfrak{G}$*.*

3. *(Calibrated e-process.) Let* $f : [0,1] \to [0,\infty)$ *be any p-to-e calibrator. Then, the process* $(E_t^{\mathsf{pw}})_{t=0}^{\infty}$ *defined by* $E_0^{\mathsf{pw}} = 1$ *and*

$$E_t^{\mathsf{pw}} := f(\mathsf{p}_t^{\mathsf{pw}}), \quad \forall t \ge 1 \tag{A.49}$$

   is an e-process for $\mathcal{H}_0^{\mathsf{pw}}(p,q;h)$ *w.r.t.* $\mathfrak{G}$*.*

The structure of the index set ensures that $E_t^{[k]}(\lambda)$ for each k is adapted and non-increasing under the null. For example, with lag-3 forecasts, $E_t^{[k]}(\lambda)$ for each k is computed using each of the subsequences $(1,4,7,\dots)$, $(2,5,8,\dots)$, and $(3,6,9,\dots)$. As for the choice of a p-to-e calibrator f, we follow

Vovk and Wang (2021); Ramdas et al. (2022b) and use (as our default)

$$f(p) = \frac{1 - p + p \log p}{p(\log p)^2}, \quad p \in [0, 1]. \tag{A.50}$$

In words, sequential e-values are expected to be at most 1 at time $\tau + h - 1$, where τ is any stopping time w.r.t. $\mathfrak{G}$. In contrast, the p-process directly yields a valid sequential test without such a condition, and it can also be calibrated to yield an e-process.

Proof of Proposition A.5. Our goal is to derive a p-process for $\mathcal{H}_0^{\mathsf{pw}}(h)$ based on ideas from the proofs of Proposition 3.4 in Arnold et al. (2021) and from the validity of their proposed sequential test, and then to calibrate it into an e-process (Shafer et al., 2011; Ramdas et al., 2022b).

Sub-filtrations $\mathfrak{G}^{[k]}$ and processes $L_t^{[k]}$. Recall that $\mathfrak{G} = (\mathcal{G}_t)_{t=0}^{\infty}$, and define the $\mathfrak{G}^{[1]}, \dots, \mathfrak{G}^{[h]}$ as follows: for each $k = 1, \dots, h$,

$$\mathfrak{G}^{[k]} := \left(\mathcal{G}_t^{[k]}\right)_{t=0}^{\infty}, \quad \text{where} \quad \mathcal{G}_t^{[k]} := \mathcal{G}_{\left\lfloor \frac{t-k}{h} \right\rfloor h + k}. \tag{A.51}$$

Because $\left\lfloor \frac{t-k}{h} \right\rfloor h + k \leq \left(\frac{t-k}{h}\right) h + k \leq t$, we have $\mathcal{G}_t^{[k]} \subseteq \mathcal{G}_t \ \forall t$, i.e., $\mathfrak{G}^{[k]}$ is a sub-filtration of $\mathfrak{G}$ for each k. (Each $\mathcal{G}^{[k]}$ only updates its filtration every h steps.)

In the following, we fix $\lambda \in [0, 1/c)$ and omit any dependence on it for notational convenience. For each $k = 1, \dots, h$, define the process $(L_t^{[k]})_{t=0}^{\infty}$ as follows: $L_0^{[k]} := 1$ and, for each $t \geq 1$,

$$L_t^{[k]} := \prod_{i \in I_t^{[k]}} l_{i-1}(y_{i+h-1}), \tag{A.52}$$

where $\prod_{i \in \emptyset}(\cdot) = 1$ and

$$l_{i-1}(y_{i+h-1}) := \exp\left\{\lambda\left(\hat{\delta}_i^{(h)} - \delta_i^{(h)}\right) - \psi_{E,c}(\lambda)\left(\hat{\delta}_i^{(h)} - \gamma_i\right)^2\right\}. \tag{A.53}$$

(We index (A.53) by $i - 1$, because it only consists of $\mathcal{G}_{i-1}$-measurable terms aside from y_{i+h-1}. For example, $\delta_i^{(h)} = \mathbb{E}_{i-1}[\hat{\delta}_i^{(h)}]$ is $\mathcal{G}_{i-1}$-measurable.) Then, each $(L_t^{[k]})_{t=0}^{\infty}$ is an adapted process w.r.t. $\mathfrak{G}$, because the last index of $I_t^{[k]}$ is at most $t - h + 1$, and the outcome corresponding to that index is y_t,

which is $\mathcal{G}_t$-measurable.

$(L_t^{[k]})_{t=0}^{\infty}$ is a test supermartingale w.r.t. $\mathfrak{G}^{[k]}$ for each k. Recall that $\mathbb{E}[\hat{\delta}_i^{(h)} \mid \mathcal{G}_{i-1}] = \delta_i^{(h)}$ by definition. Since the score differentials are bounded by assumption, the proof of Proposition A.1 (with y_i replaced with y_{i+h-1} in the proof) implies that

$$\mathbb{E}\left[l_{i-1}(y_{i+h-1}) \mid \mathcal{G}_{i-1}\right] \leq 1 \quad \forall i \geq h. \tag{A.54}$$

Now, if $t < h$ or $\left\lfloor \frac{t-k}{h} \right\rfloor \neq \frac{t-k}{h}$ (i.e., not an integer), then $I_t^{[k]} = I_{t-1}^{[k]}$ by construction, so $L_t^{[k]} = L_{t-1}^{[k]}$. On the other hand, if $t \geq h$ and $\left\lfloor \frac{t-k}{h} \right\rfloor = \frac{t-k}{h}$, then algebra shows that $L_t^{[k]} = L_{t-1}^{[k]} \cdot l_{t-h}(y_t)$, and also that $\mathcal{G}_{t-1}^{[k]} = \mathcal{G}_{\left\lfloor \frac{(t-1)-k}{h} \right\rfloor h+k} = \mathcal{G}_{\left(\frac{t-k}{h}-1\right)h+k} = \mathcal{G}_{t-h}$. Thus,

$$\mathbb{E}\left[L_t^{[k]} \mid \mathcal{G}_{t-1}^{[k]}\right] = L_{t-1}^{[k]} \cdot \mathbb{E}\left[l_{t-h}(y_t) \mid \mathcal{G}_{t-h}\right] \leq L_{t-1}^{[k]}. \tag{A.55}$$

The above algebra also shows that each multiplicative increment of $L_t^{[k]}$ is either constant (1) or $\mathfrak{G}_t^{[k]}$-measurable. Therefore, $(L_t^{[k]})_{t=0}^{\infty}$ is a test supermartingale w.r.t. $\mathfrak{G}^{[k]}$.

$(\bar{E}_t^{\mathsf{pw}})_{t=0}^{\infty}$ is a sequential e-value of lag h for $\mathcal{H}_0^{\mathsf{pw}}$ (w.r.t. $\mathfrak{G}$). Under any $P \in \mathcal{H}_0^{\mathsf{pw}}(p, q; h)$, we know that

$$\Delta_t^{[k]} = \sum_{i \in I_t^{[k]}} \delta_i^{(h)} \leq 0, \quad \forall t \geq h. \tag{A.56}$$

We thus have, P-almost surely,

$$E_t^{[k]} = \prod_{i \in I_t^{[k]}} \exp\left\{\lambda \hat{\delta}_i^{(h)} - \psi_{E,c}(\lambda)\left(\hat{\delta}_i^{(h)} - \gamma_i\right)^2\right\} \tag{A.57}$$

$$\leq \exp\left\{-\sum_{i \in I_t^{[k]}} \delta_i^{(h)}\right\} \cdot \prod_{i \in I_t^{[k]}} \exp\left\{\lambda \hat{\delta}_i^{(h)} - \psi_{E,c}(\lambda)\left(\hat{\delta}_i^{(h)} - \gamma_i\right)^2\right\} = L_t^{[k]}, \quad \forall t \geq h. \tag{A.58}$$

In other words, under any $P \in \mathcal{H}_0^{\mathsf{w}}(p, q; h)$, $E_t^{[k]}$ is upper-bounded by $L_t^{[k]}$ for each k, where $(L_t^{[k]})_{t=0}^{\infty}$ is a test supermartingale w.r.t. $\mathfrak{G}^{[k]}$. By the supermartingale optional stopping theorem (e.g., Theorem

4.8.4, Durrett (2019)), we thus have that, for any stopping time $\tau^{[k]}$ w.r.t. $\mathfrak{G}^{[k]}$,

$$\mathbb{E}_P\left[E^{[k]}_{\tau^{[k]}}\right] \le 1, \tag{A.59}$$

under any $P \in \mathcal{H}_0^{\mathsf{w}}(p, q; h)$.

Finally, the construction (A.51) implies that, for any stopping time τ w.r.t. $\mathfrak{G}$, the mapping $\tau \mapsto \tau^{[k]}$ defined by

$$\tau^{[k]} := \left(\left\lfloor \frac{\tau - k - 1}{h} \right\rfloor + 1\right) h + k \tag{A.60}$$

gives a stopping time w.r.t. $\mathfrak{G}^{[k]}$ (Henzi and Ziegel, 2022), where $\tau^{[k]} \in \{\tau, \tau+1, \dots, \tau+(h-1)\}$. Therefore, for any stopping time τ w.r.t. $\mathfrak{G}$,

$$\mathbb{E}_P[\bar{E}_{\tau+h-1}] \le \frac{1}{h}\sum_{k=1}^{h} \mathbb{E}_P\left[E^{[k]}_{\tau^{[k]}}\right] \le 1, \tag{A.61}$$

for any $P \in \mathcal{H}_0^{\mathsf{w}}(p, q; h)$.

$(\mathsf{p}^{\mathsf{pw}}_t)_{t=0}^{\infty}$ is a p-process for $\mathcal{H}_0^{\mathsf{pw}}$. The key idea here is to first use the fact that $L^{[k]}_t$ is a test supermartingale w.r.t. $\mathfrak{G}^{[k]}$ that upper-bounds $E^{[k]}_t$, for each $k \in \{1, \dots, h\}$, and then use the time-uniform equivalence lemma for probabilities (Ramdas et al., 2020), along with a p-merging function (Vovk and Wang, 2021), to obtain a combined p-process.

First, define the following process for each $k = 1, \dots, h$:

$$\mathsf{q}^{[k]}_t := 1 \wedge \left(1/\sup_{i \le t} L^{[k]}_i\right), \quad \forall t \ge 1. \tag{A.62}$$

The process involves the running supremum of $(L^{[k]}_t)_{t=0}^{\infty}$, which is a test supermartingale w.r.t. $\mathfrak{G}^{[k]}$ as we showed earlier. In particular, (A.58) implies that $\mathsf{p}^{[k]}_t \ge \mathsf{q}^{[k]}_t$ for all t and k under $P \in \mathcal{H}_0^{\mathsf{pw}}$.

Applying Ville (1939)'s inequality to $(L^{[k]}_t)_{t=0}^{\infty}$, for any P,

$$P\left(\exists t \ge 1 : \mathsf{q}^{[k]}_t \le \alpha\right) = P\left(\sup_{t \ge 1} L^{[k]}_i \ge \frac{1}{\alpha}\right) \le \alpha, \quad \forall \alpha \in (0, 1). \tag{A.63}$$

Then, under any $P \in \mathcal{H}_0^{\mathsf{pw}}$, the fact that $\mathsf{p}_t^{[k]} \geq \mathsf{q}_t^{[k]}$ under P implies

$$P\left(\exists t \geq 1 : \mathsf{p}_t^{[k]} \leq \alpha\right) \leq \alpha, \quad \forall \alpha \in (0,1). \tag{A.64}$$

Now, following an earlier proof in (A.53) where we showed that $(L_t^{[k]})_{t=0}^{\infty}$ is an adapted process w.r.t. the game filtration $\mathfrak{G}$, we can analogously show that $(E_t^{[k]})_{t=0}^{\infty}$ is also an adapted process w.r.t. $\mathfrak{G}$, and so is $(\mathsf{p}_t^{[k]})_{t=0}^{\infty}$ by its definition. Then, by Lemma 2 of Ramdas et al. (2020), (i) $\Rightarrow$ (iii), equation (A.64) implies that

$$P\left(\mathsf{p}_\tau^{[k]} \leq \alpha\right) \leq \alpha, \quad \forall \alpha \in (0,1), \tag{A.65}$$

for any stopping time τ w.r.t. $\mathfrak{G}$ and $P \in \mathcal{H}_0^{\mathsf{pw}}(h)$. In other words, $(\mathsf{p}_t^{[k]})_{t=1}^{\infty}$ is a p-process for $\mathcal{H}_0^{\mathsf{pw}}(h)$ w.r.t. $\mathfrak{G}$, for each $k \in \{1, \ldots, h\}$.

Finally, we can merge the p-processes $(\mathsf{p}_t^{[k]})_{t=1}^{\infty}$ at any $\mathfrak{G}$-stopping times. For any $\mathfrak{G}$-stopping time τ, using the harmonic average p-merging function by Vovk and Wang (2021) combined with (A.65) gives, for any $P \in \mathcal{H}_0^{\mathsf{pw}}$,

$$P\left(\mathsf{p}_\tau^{\mathsf{pw}} \leq \alpha\right) \leq \alpha, \quad \forall \alpha \in (0,1). \tag{A.66}$$

$(E_t^{\mathsf{pw}})_{t=0}^{\infty}$ is an e-process for $\mathcal{H}_0^{\mathsf{pw}}$. This follows directly from the validity of a p-to-e calibrator for p-processes (e.g., Proposition 12, Ramdas et al. (2020)). □

The statements and proofs for the weak null $\mathcal{H}_0^{\mathsf{w}}(h)$ are completely analogous, except that instead of taking averages across the h sub-processes we have to take the minimum/maximum for e-/p-processes, because the weak null only implies that there exists some k for which $\Delta_t^{[k]} \leq 0$.

Proposition A.6 (Sequential inference for $\mathcal{H}_0^{\mathsf{w}}(h)$)**.** *Assume the same setup as Proposition A.5. Then, for each $\lambda \in [0, 1/c)$, the following statements are true:*

1. *(Minimum sequential e-values.) The process*

$$\bar{E}_t^{\mathsf{w}}(\lambda) := \min_{k=1,\ldots,h} E_t^{[k]}(\lambda) \tag{A.67}$$

satisfies $\mathbb{E}_P[\bar{E}_{\tau+h-1}^{\mathsf{pw}}(\lambda)] \leq 1$ for any $\mathfrak{G}$-stopping time τ and any $P \in \mathcal{H}_0^{\mathsf{w}}(p, q; h)$.

2. *(P-process.) The process* $(\mathsf{p}_t^{\mathsf{w}})_{t=1}^{\infty}$ *defined by*

$$\mathsf{p}_t^{\mathsf{w}} := \max_{k=1,\dots,h} \mathsf{p}_t^{[k]}, \quad \textit{where} \quad \mathsf{p}_t^{[k]} := 1 \wedge \left(1/\sup_{i \le t} E_i^{[k]}(\lambda)\right), \tag{A.68}$$

is an p-process for $\mathcal{H}_0^{\mathsf{w}}(p, q; h)$ *w.r.t.* $\mathfrak{G}$.

3. *(Calibrated e-process.) Let* $f : [0,1] \to [0,\infty)$ *be any p-to-e calibrator. Then, the process* $(E_t^{\mathsf{w}})_{t=0}^{\infty}$ *defined by* $E_0^{\mathsf{w}} = 1$ *and*

$$E_t^{\mathsf{w}} := f(\mathsf{p}_t^{\mathsf{w}}), \quad \forall t \ge 1 \tag{A.69}$$

is an e-process for $\mathcal{H}_0^{\mathsf{w}}(p, q; h)$ *w.r.t.* $\mathfrak{G}$.

The methods described in Propositions A.5 and A.6 both provide valid options for sequentially comparing lag-h forecasters. While E_t^{pw} may involve a seemingly less intuitive null hypothesis, it upper-bounds E_t^{w}, and it can grow more quickly when either null is false. Rejecting $\mathcal{H}_0^{\mathsf{pw}}(p, q; h)$ implies that there exists some $k \in \{1, \dots, h\}$ such that $\Delta_t^{[k]} > 0$ for some t. For example, if $h = 2$, then it implies p outperforms q on average on either odd or even days. A scenario in which rejecting $\mathcal{H}_0^{\mathsf{pw}}(h)$ would clearly not imply $\mathcal{H}_0^{\mathsf{w}}(h)$ is when (coincidentally) there is seasonality of period exactly h in the game — e.g., when comparing 7-day forecasts for a sequence of outcomes that have a different distribution every weekend, E_t^{w} and E_t^{pw} may differ significantly. A simple way to mitigate this issue is to simply monitor both e-processes (depending on the use case).

In Table A.1, we list the sequential e-values for $\mathcal{H}_0^{\mathsf{w}}$ (Proposition A.6), $\mathcal{H}_0^{\mathsf{pw}}$ (Proposition A.5), and $\mathcal{H}_0^{\mathsf{s}}$ (Henzi and Ziegel (2022); denoted as $\bar{E}^{\mathsf{s}}$), for the weather comparison tasks in Section 3.5.3 with lags $h = 1, \dots, 5$. As in Henzi and Ziegel (2022), no stopping is applied in any of the sequential e-values. As shown, while $\bar{E}^{\mathsf{w}}$ tends to be overly conservative, $\bar{E}^{\mathsf{pw}}$ remains relatively powerful despite testing a substantially weaker null than the strong null (for $\bar{E}^{\mathsf{s}}$). Across different locations and lags, $\bar{E}^{\mathsf{s}}$ is generally large (≥ 20) whenever $\bar{E}^{\mathsf{pw}}$ is large, and this is explained by the inclusion relationship between the nulls in (A.45). The comparison of HCLR against HCLR_ in Zurich is the only case where $\bar{E}^{\mathsf{pw}}$ exceeds $\bar{E}^{\mathsf{s}}$. In this case, the e-values drawn over time (similar to Figure 3.5) show that there are multiple time periods (2012-2013 and 2014-2015) during which both $\bar{E}^{\mathsf{s}}$ and $\bar{E}^{\mathsf{pw}}$ decrease substantially, and it is possible that the choice of the hyperparameter or the variance-adaptivity of our e-values affects how quickly they "rebound" after such sharp decreases.

Location	Lag	HCLR/IDR			IDR/HCLR_			HCLR/HCLR_		
		$\bar{E}^{\mathrm{w}}$	$\bar{E}^{\mathrm{pw}}$	$\bar{E}^{\mathrm{s}}$	$\bar{E}^{\mathrm{w}}$	$\bar{E}^{\mathrm{pw}}$	$\bar{E}^{\mathrm{s}}$	$\bar{E}^{\mathrm{w}}$	$\bar{E}^{\mathrm{pw}}$	$\bar{E}^{\mathrm{s}}$
Brussels	1	0.012	0.012	0.000	> 100	> 100	> 100	1.083	1.083	> 100
	2	0.021	0.033	0.000	0.196	1.659	> 100	0.510	1.196	> 100
	3	0.049	0.060	0.006	0.060	0.121	1.786	0.698	2.289	> 100
	4	0.053	1.032	22.811	0.018	0.042	0.000	0.114	1.855	> 100
	5	0.145	0.714	> 100	0.021	0.034	0.000	0.254	19.411	> 100
Frankfurt	1	0.034	0.034	0.000	1.284	1.284	> 100	> 100	> 100	> 100
	2	0.022	0.029	0.000	1.573	7.223	> 100	1.537	69.508	> 100
	3	0.022	0.041	0.000	0.311	3.814	> 100	0.836	> 100	> 100
	4	0.047	0.214	0.361	0.033	0.090	0.122	0.163	27.920	> 100
	5	0.037	0.334	2.468	0.023	0.104	0.001	0.173	1.781	> 100
London	1	0.041	0.041	0.029	0.277	0.277	1.351	0.285	0.285	2.845
	2	0.038	0.038	0.021	0.289	0.321	2.002	0.164	0.200	5.178
	3	0.037	0.061	0.185	0.087	0.367	0.203	0.141	0.241	9.613
	4	0.077	0.121	1.751	0.051	0.108	0.018	0.077	1.714	8.428
	5	0.070	0.208	4.949	0.032	0.066	0.002	0.113	0.279	1.427
Zurich	1	0.034	0.034	0.003	6.670	6.670	25.692	> 100	> 100	61.747
	2	0.054	0.061	0.012	0.328	0.415	19.229	2.195	> 100	74.745
	3	0.066	0.487	1.079	0.037	0.197	0.661	1.877	7.311	94.613
	4	0.091	1.553	30.478	0.023	0.066	0.004	0.210	54.131	47.069
	5	0.082	8.436	> 100	0.026	0.053	0.000	0.192	3.964	40.648

Table A.1: Lag-h sequential e-values between pairs of statistical postprocessing methods for ensemble weather forecasts across different locations and lags, where T is the last time step (January 01, 2017). $\bar{E}^{\mathrm{w}}$, $\bar{E}^{\mathrm{pw}}$, and $\bar{E}^{\mathrm{s}}$ indicate the lag-h sequential e-values for the lag-h weak, period-wise weak, and strong nulls, respectively. All procedures use the Brier score as the scoring rule. "p/q" indicates the null that "p is no better than q." Generally speaking, $\bar{E}^{\mathrm{w}}$ is the most conservative, while $\bar{E}^{\mathrm{pw}}$ can be powerful against its relatively weak null (compared to the strong null for $\bar{E}^{\mathrm{s}}$).

We close with the note that the choice of how aggressively one can bet, either via the choice of the hyperparameter in the mixture distribution F for $\bar{E}^{\mathrm{w}}$ and $\bar{E}^{\mathrm{pw}}$ (cf. Section 3.4.4) or the alternative probability π_1 for E^{s}, directly affects the power of these e-values. Developing powerful strategies for choosing F in the lagged scenario remains a problem deserving of future investigation.

A.6 Inference for Predictable Subsequences and Bounds

Martingale theory tells us that we can substitute each variable in the exponential supermartingale (3.8) with any predictable terms, similar to $(\gamma_i)_{i=1}^{\infty}$ in Theorem 3.2. In doing so, we must make sure that the resulting test supermartingale leads to estimating/testing an appropriate quantity of interest. Here,

we illustrate two useful extensions involving this general technique.

A.6.1 Inference for Predictable Subsequences

Suppose that each round of our forecast comparison game (Game 3.1) happens daily, but we are only interested in comparing the forecasters on weekdays, on every other day, or more interestingly, on days after some specific event happens (e.g., days following market crashes). To formalize this, we introduce a predictable $\{0,1\}$-valued process $\xi := (\xi_t)_{t=1}^{\infty}$ and then estimate/test the average score differential *only* at times when $\xi_t = 1$. The resulting parameter of interest is expressed as follows:

$$\Delta_t(\xi_{1:t}) := \frac{\sum_{i=1}^{t} \xi_i \delta_i}{\sum_{i=1}^{t} \xi_i} = \frac{1}{\sum_{i=1}^{t} \xi_i} \sum_{i=1}^{t} \xi_i \mathbb{E}_{i-1}\left[S(p_i, y_i) - S(q_i, y_i)\right], \tag{A.70}$$

where $\delta_i = \mathbb{E}_{i-1}[\hat{\delta}_i] = \mathbb{E}_{i-1}[S(p_i, y_i) - S(q_i, y_i)]$ and $\xi_{1:t} = (\xi_1, \dots, \xi_t)$. $\Delta_t(\xi_{1:t})$ measures the time-varying average score differential *only* for times when $\xi_i = 1$. Henzi and Ziegel (2022) introduce an analogous extension to testing the strong null (3.17), where the predictable condition $\xi_t = \mathbb{1}\left(\max\{p_t, q_t\} \geq \frac{1}{2}\right)$ is used to compare extreme precipitation forecasts.

Because the conditions are predictable, we have the property that $\mathbb{E}_{i-1}[\xi_i \hat{\delta}_i] = \xi_i \mathbb{E}_{i-1}[\hat{\delta}_i] = \xi_i \delta_i$, from which the proofs of Theorem 3.1 (assuming sub-Gaussianity), as well as Theorem 3.2 and Theorem 3.3 (assuming boundedness), straightforwardly follow. For example, for each $\lambda \in [0, 1/c)$, consider

$$L_t(\lambda; \xi_{1:t}) := \prod_{i:\xi_i=1} \exp\left\{\lambda(\hat{\delta}_i - \delta_i) - \psi_E(\lambda)(\hat{\delta}_i - \gamma_i)^2\right\} \tag{A.71}$$

$$= \prod_{i=1}^{t} \left[(1 - \xi_i) + \xi_i \exp\left\{\lambda(\hat{\delta}_i - \delta_i) - \psi_E(\lambda)(\hat{\delta}_i - \gamma_i)^2\right\}\right]. \tag{A.72}$$

Then, under the same conditions as Proposition A.1, $L_t(\lambda; \xi_{1:t})$ is a test supermartingale w.r.t. $\mathfrak{G}$:

$$\mathbb{E}_{t-1}[L_t(\lambda; \xi_{1:t})] = L_{t-1}(\lambda; \xi_{1:t-1})\left[(1 - \xi_i) + \xi_i \mathbb{E}_{t-1} \exp\left\{\lambda(\hat{\delta}_i - \delta_i) - \psi_E(\lambda)(\hat{\delta}_i - \gamma_i)^2\right\}\right] \tag{A.73}$$

$$\leq L_{t-1}(\lambda; \xi_{1:t-1}), \tag{A.74}$$

for each $t \geq 1$. We used the predictability of $(\xi_t)_{t=1}^{\infty}$ in (A.73) and the boundedness condition (see

proof of Proposition A.1) in (A.74). Applying this to the proof of Theorem 3.2 shows that we can construct an EB CS for $(\Delta_t(\xi_{1:t}))_{t=1}^{\infty}$.

Similarly, we can also derive the corresponding sub-exponential e-process for the null $\mathcal{H}_0^{\mathrm{w}}(\xi) : \Delta_t(\xi_{1:t}) \leq 0, \ \forall t$. This e-process is given by

$$E_t(\lambda; \xi_{1:t}) := \prod_{i:\xi_i=1} \exp\left\{\lambda\hat{\delta}_i - \psi_E(\lambda)(\hat{\delta}_i - \gamma_i)^2\right\}, \tag{A.75}$$

for any $\lambda \in [0, 1/c)$. This is an e-process because, under $\mathcal{H}_0^{\mathrm{w}}(\xi)$, we have that $\exp(-\lambda \sum_{i=1}^{t} \xi_i\delta_i) = \prod_{i:c_i=1} \exp(-\lambda\delta_i) \geq 1$, and thus

$$E_t(\lambda; \xi_{1:t}) \leq \prod_{i:\xi_i=1} \exp\left\{\lambda(\hat{\delta}_i - \delta_i) - \psi_E(\lambda)(\hat{\delta}_i - \gamma_i)^2\right\} = L_t(\lambda; \xi_{1:t}). \tag{A.76}$$

Since $E_t(\lambda; \xi_{1:t})$ is upper-bounded by the test supermartingale $L_t(\lambda; \xi_{1:t})$ for all t under $\mathcal{H}_0^{\mathrm{w}}(\xi)$, it follows that $E_t(\lambda; \xi_{1:t})$ is an e-process for $\mathcal{H}_0^{\mathrm{w}}(\xi)$ (Ramdas et al., 2020).

In summary, both the CS and the e-process remain valid under predictable conditions.

A.6.2 Inference Under Predictable Bounds

For Theorems 3.2 and 3.3, we require that the pointwise score differentials are bounded by some fixed constant, i.e., $|\hat{\delta}_i| \leq \frac{c}{2}$ for all i, for some $c \in (0, \infty)$. In practice, this may be restrictive when the value of c is not known a priori or its range shifts drastically over time. One way to mitigate this issue is to have a predictable bound $(c_i)_{i=1}^{\infty}$ at each round, such that

$$\left|\hat{\delta}_i\right| \leq \frac{c_i}{2}, \tag{A.77}$$

for $i \geq 1$, instead of having a uniform bound over all rounds. Predictable bounds can also be useful in cases where one can guess how bad/good the forecasts can be before each new round begins.

Here, we show that we can extend both Theorem 3.2 and Theorem 3.3 to work for predictably bounded score differentials. This result depends on the following facts about the exponential CGF-like function, $\psi_{E,c}(\lambda)$, *as a function of its scale* c. Below, we take $1/0 = \infty$.

Lemma A.2. *For each* $\lambda \geq 0$*, the function* $f_\lambda(c) := \psi_{E,c}(\lambda) = c^{-2}[-c\lambda - \log(1 - c\lambda)]$ *is non-decreasing*

and convex on $c \in (0, 1/\lambda)$. Furthermore, f_λ is strictly increasing and strongly convex on $c \in (0, 1/\lambda)$ if and only if $\lambda > 0$.

Proof. Since $f_\lambda(c)$ is twice differentiable w.r.t. c, it suffices to show that $f'_\lambda(c) \geq 0$ and $f''_\lambda(c) \geq 0$ for all c, and also that $f'_\lambda(c) > 0$ and $f''_\lambda(c) > 0$ for all c if and only if $\lambda > 0$.

Given that $0 \leq c\lambda < 1$, we utilize the Taylor series of $x \longmapsto -\log(1-x)$ at $x = 0$:

$$-\log(1 - c\lambda) = \sum_{t=1}^{\infty} \frac{(c\lambda)^t}{t} = c\lambda + \frac{c^2\lambda^2}{2} + \frac{c^3\lambda^3}{3} + \cdots, \tag{A.78}$$

which converges (absolutely). It then follows that

$$f_\lambda(c) = \frac{-c\lambda - \log(1 - c\lambda)}{c^2} = \frac{\lambda^2}{2} + \frac{c\lambda^3}{3} + \cdots = \lambda^2 \sum_{t=0}^{\infty} \frac{(c\lambda)^t}{t+2}. \tag{A.79}$$

Taking first derivatives term-by-term,

$$f'_\lambda(c) = \lambda^2 \sum_{t=1}^{\infty} \frac{t\lambda^t c^{t-1}}{t+2}. \tag{A.80}$$

Given that $c > 0$, we have that $f'_\lambda(c) \geq 0$ for any $\lambda \geq 0$. Furthermore, we have that $f'_\lambda(c) > 0$ for $\lambda > 0$ and $f'_\lambda(c) = 0$ for $\lambda = 0$.

Similarly, taking second derivatives term-by-term,

$$f''_\lambda(c) = \lambda^2 \sum_{t=2}^{\infty} \frac{t(t-1)\lambda^t c^{t-2}}{t+2}. \tag{A.81}$$

Given that $c > 0$, we have that $f''_\lambda(c) \geq 0$ for any $\lambda \geq 0$. Furthermore, we have that $f''_\lambda(c) > 0$ for $\lambda > 0$ and $f''_\lambda(c) = 0$ for $\lambda = 0$. □

In Figure A.1, we plot $\psi_{E,c}(\lambda)$ as a function of c, illustrating that it is indeed strictly increasing and strongly convex for different values of $\lambda > 0$, and we also show that $\psi_{E,1}$ as a function of λ approximates $\psi_{N,1}(\lambda) = \lambda^2/2$ as $\lambda \to 0^+$.

Now, we derive an e-process that involves predictable bounds and is upper-bounded by a test supermartingale that uses a uniform bound. Let c_0 be a (possibly infinite) constant such that $c_i \leq c_0$ for all i, and let $\hat{v}_i = (\hat{\delta}_i - \gamma_i)^2$ where $(\gamma_i)_{i=1}^{\infty}$ is any predictable sequence as in Theorems 3.2 and 3.3.

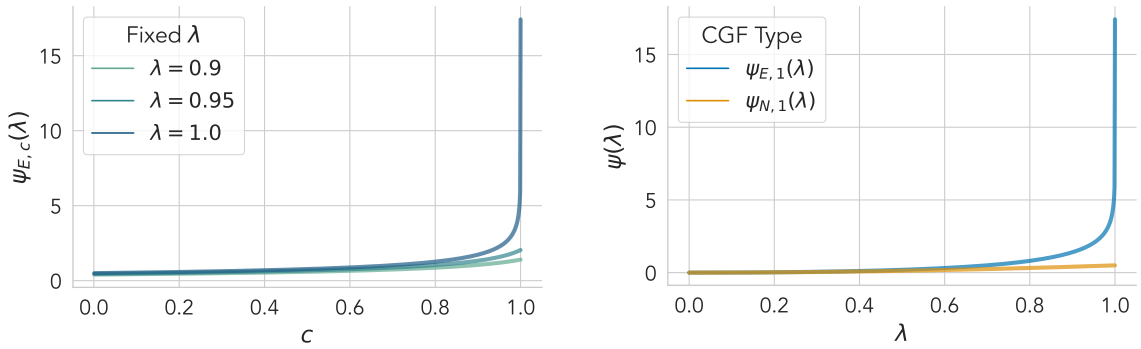


Figure A.1: *Left:* Plots of the exponential CGF-like function $f_\lambda(c) = \psi_{E,c}(\lambda)$ against $c \in (0, 1/\lambda)$, for fixed λ values of 0.9, 0.95, and 1.0. For each $\lambda \geq 0$, $f_\lambda(c)$ is strictly increasing and strongly convex on $c \in (0, 1/\lambda)$. *Right:* Comparing $\psi_{E,1}(\lambda)$, as a function of $\lambda \in [0, 1)$, with the Gaussian CGF $\psi_{N,1}(\lambda) = \lambda^2/2$.

Now, for each $\lambda \in [0, 1/c_0)$ (as before, we set $1/\infty = 0$ and $[0, 0) = \{0\}$), define the following processes: $\underset{\sim}{L}_0(\lambda) = L_0(\lambda) = 1$, and for $t \geq 1$,

$$\underset{\sim}{L}_t(\lambda) := \prod_{i=1}^{t} \exp\left\{\lambda\left(\hat{\delta}_i - \delta_i\right) - \psi_{E,c_0}(\lambda)\left(\hat{\delta}_i - \gamma_i\right)^2\right\}; \tag{A.82}$$

$$L_t(\lambda) := \prod_{i=1}^{t} \exp\left\{\lambda\left(\hat{\delta}_i - \delta_i\right) - \psi_{E,c_i}(\lambda)\left(\hat{\delta}_i - \gamma_i\right)^2\right\}. \tag{A.83}$$

(If $c_0 = \infty$, then ψ_{E,c_0} is not well-defined, so set $\underset{\sim}{L}_t(\lambda) = 1$ for all $t \geq 1$.)

Proposition A.7. *Suppose that $|\hat{\delta}_i| \leq \frac{c_i}{2}$, where $(c_i)_{i=1}^{\infty}$ is a strictly positive predictable sequence. Also, let $\hat{V}_t = \sum_{i=1}^{t}(\hat{\delta}_i - \gamma_i)^2$, where $(\gamma_i)_{i=1}^{\infty}$ is any $\left[-\frac{c_i}{2}, \frac{c_i}{2}\right]$-valued predictable sequence. Then, for each $\lambda \in [0, 1/c_0)$, the following statements are true:*

1. *$\underset{\sim}{L}_t(\lambda) \leq L_t(\lambda)$ for all $t \geq 1$;*
2. *The process $(L_t(\lambda))_{t=0}^{\infty}$ is a test supermartingale w.r.t. $\mathfrak{G}$;*
3. (A predictably-bounded e-process.) *The process $(E_t(\lambda))_{t=0}^{\infty}$, defined as $E_0(\lambda) = 1$ and*

$$E_t(\lambda) := \prod_{i=1}^{t} \exp\left\{\lambda\hat{\delta}_i - \psi_{E,c_i}(\lambda)\left(\hat{\delta}_i - \gamma_i\right)^2\right\}, \ \forall t \geq 1, \tag{A.84}$$

is an e-process for $\mathcal{H}_0^{\mathrm{w}}(p, q) : \Delta_t \leq 0, \ \forall t \geq 1$.

Proof. 1. Using the fact that $c_i \leq c_0$ for each i and that $\psi_{E,c}(\lambda)$ is non-decreasing in c by Lemma A.2,

we obtain

$$\underset{\sim}{L}_t(\lambda) = \exp\left\{\lambda S_t - \psi_{E,c_0}(\lambda)\hat{V}_t\right\} \leq L_t(\lambda). \tag{A.85}$$

2. If $c_0 = \infty$, then we must have $\lambda = 0$, so $(L_t(\lambda))_{t=0}^{\infty}$ always takes the value 1 and is a (trivial) test supermartingale. Otherwise, Proposition A.2 directly implies that $(L_t(\lambda))_{t=0}^{\infty}$ is a test supermartingale w.r.t. $\mathfrak{G}$.
3. Because $(c_i)_{i=1}^{\infty}$ is predictable w.r.t. $\mathfrak{G}$, the process $(E_t(\lambda))_{t=0}^{\infty}$ is adapted w.r.t. $\mathfrak{G}$. Then, $E_t(\lambda) \leq L_t(\lambda)$ (P-a.s.) for all t under any $P \in \mathcal{H}_0^{\mathrm{w}}(p,q)$, as in the proof of Theorem 3.3, and thus the result follows by Corollary 22 of Ramdas et al. (2020).

□

Note that, if a constant bound $c_0 = c > 0$ were known *a priori*, then $\underset{\sim}{L}_t(\lambda)$ coincides with the exponential test supermartingale in Equation (3.8). The e-process (A.84) can be more powerful than using the analogous $(\underset{\sim}{E}_t(\lambda))_{t=0}^{\infty}$ involving c_0 in some cases, although taking the mixture over λ (Section 3.4.3) may not yield a closed form.

A.7 Generalizations To Other Outcome and Forecast Types

In principle, the game-theoretic approach we describe in Section 3.4.1 can straightforwardly generalize beyond the case of probability forecasts on dichotomous events. We briefly discuss two such generalizations and to what extent our methods are applicable in each case.

The first is to the case of C-categorical outcomes, for $C \geq 2$. We can start with the game-theoretic setup (Game 3.1) and parameterize the outcome space using C-dimensional length-1 binary vectors, i.e., $\mathcal{Y} = \{\mathbf{e}_c\}_{c=1}^{C}$ where $\mathbf{e}_c = [\mathbb{1}(i = c)]_{i=1}^{C}$, and the set of forecasts as the C-dimensional probability simplex, i.e., $\mathcal{P} = \Delta^{C-1} = \{\mathbf{p} \in [0,1]^C : \sum_{c=1}^{C} p^{(c)} = 1\}$. Reality also makes its choices from Δ^{C-1}. Note that, if $C = 2$, we can recover the binary case via the mapping $\mathbf{p} = (1-p, p)$, for $p \in [0,1]$. Then, by choosing any bounded scoring rule for categorical outcomes, we can straightforwardly apply Theorems 3.2 and 3.3 to obtain CSs and e/p-processes (respectively) on the average score differentials. The C-dimensional Brier score, defined as $S(\mathbf{p}, \mathbf{y}) = 1 - \|\mathbf{p} - \mathbf{y}\|_2^2$, is bounded within $[0,1]$; the spherical and zero-one scores can be defined analogously (Gneiting and Raftery, 2007) and are similarly

bounded. We note that using the normalized Winkler score to utilize unbounded scores, as in Section A.4, is not straightforward.

The next extension is to the case of continuous outcomes. In this case, we can once again start with the game-theoretic setup (Game 3.1) and parameterize the outcome space as $\mathcal{Y} \subseteq \mathbb{R}^d$ for some $d \geq 1$. At each round t, Reality now chooses an arbitrary distribution r_t on $\mathcal{Y}$, from which y_t is sampled. Depending on the specific forecasting task, the forecasters may either predict (i) certain functional(s) of the outcome distribution, denoted as $\Gamma(P)$ for each $P \in \mathcal{P}$, or (ii) the CDF (or density) itself. As an example for (i), each forecaster may predict a level-α (e.g., 95%) prediction interval (l_t, u_t), in which case the statistician can use the α-interval score (Dunsmore, 1968):

$$S_\alpha((l,u),y) = -(u-l) - (2/\alpha)(l-y)\mathbb{1}(y<l) - (2/\alpha)(y-u)\mathbb{1}(y>u), \tag{A.86}$$

for $(l,u) \subseteq \mathcal{Y}$ and $y \in \mathcal{Y}$. As an example for (ii), each forecaster may predict a (Borel-measurable) CDF F_t for y_t, in which case the statistician can use the continuously ranked probability score (CRPS) (Matheson and Winkler, 1976):

$$S(F,y) = -\int_{-\infty}^{\infty} (F(x) - \mathbb{1}(x \geq y))^2 dx = \mathbb{E}_{Y,Y' \sim F}\left[|Y - Y'|\right] - \mathbb{E}_{Y \sim F}\left[|Y - y|\right], \tag{A.87}$$

for any CDF F and outcome $y \in \mathcal{Y}$. In either case, our main results (Theorems 3.2 and 3.3) are applicable when the associated score differentials are bounded. Specifically, we can allow the choices of $\mathcal{Y}$, $\mathcal{P}$, and S such that $\mathcal{P} \subseteq \mathcal{P}^{(c)}$, where

$$\mathcal{P}^{(c)} = \{p \in \Delta(\mathcal{Y}) : |S(p,y) - S(q,y)| \leq c/2, \ \forall q \in \Delta(\mathcal{Y})\}, \tag{A.88}$$

for some $c \in (0, \infty)$. For instance, if $\mathcal{Y} = [0,1]$, then our main theorems can be used to compare mean, quantile, or interval forecasts on $\mathcal{Y}$, using the corresponding scoring rule in each case (Gneiting, 2011). If (A.88) is restrictive for the use case, then one may consider using predictable bounds (Section A.6.2) or the asymptotic CS (Section A.3). Deriving a fully general anytime-valid procedure for unbounded domains and scoring rules remains an open problem.

In Table A.2, we summarize these extensions based on the different choices of the outcome space

Outcome Type	**Categorical**	**Continuous**	
Domain	$\mathcal{Y} = \{\mathbf{e}_c\}_{c=1}^{C}$	$\mathcal{Y} \subseteq \mathbb{R}^d$	
Reality's Choice	$r_t \in \Delta^{C-1}$	$r_t \in \Delta(\mathcal{Y})$ (arbitrary distribution)	
Forecast Type	**Probability**	**Functional**	**Distribution**
Domain	$\mathcal{P} = \Delta^{C-1}$	$\Gamma(\mathcal{P})$	$\mathcal{P} \subseteq \Delta(\mathcal{Y})$
Forecast Examples	any C-dim. probability	mean, prediction interval	CDF
Score Examples	Brier, spherical, 0-1, log scores	quadratic, interval scores	CRPS
Thms. 3.2 & 3.3 apply	if $\mathcal{P} \subseteq \mathcal{P}^{(c)}$ for some $c \in (0, \infty)$		

Table A.2: Different specifications of Game 3.1 based on the outcome space and the forecast type, and the types of scoring rules that can be used in each case. In principle, the game-theoretic setup in Section 3.4.1 can straightforwardly extend to these settings; our main approaches (Theorems 3.2 and 3.3) extend to cases where the score differentials are bounded.

$\mathcal{Y}$ and the forecast type $\mathcal{P}$ within Game 3.1.

A.8 Comparison with Other Forecast Comparison Methods

A.8.1 Methodological Comparison with Henzi and Ziegel (2022)

The biggest difference between our approach and Henzi and Ziegel (2022)'s (HZ) is in the difference between the strong and weak nulls, as described in the main text. Here, we summarize other methodological differences that are worth noting for practical use cases. HZ focus on sequentially comparing forecasts on dichotomous events using consistent scoring functions (Gneiting, 2011), which straightforwardly induce proper scoring rules, and they develop e-processes of the form

$$E_t^{\mathrm{HZ}}(\lambda_1, \dots, \lambda_t) = \prod_{i=1}^{t} \left(1 + \lambda_i \tilde{\delta}_i\right), \quad \text{where} \quad \tilde{\delta}_i = \frac{S(p_i, y_i) - S(q_i, y_i)}{|S(p_i, \mathbb{1}(p_i \geq q_i)) - S(q_i, \mathbb{1}(p_i \geq q_i))|}, \tag{A.89}$$

for a $[0,1]$-valued predictable sequence $(\lambda_t)_{t=1}^{\infty}$ and a *negatively oriented* scoring function S. The form of $\tilde{\delta}_i$ is exactly that of the Winkler score: by Lemma A.1 and reversing the orientation of S, we see that $\tilde{\delta}_i = -w(p_i, q_i, y_i)$, and thus HZ's e-process can be interpreted as betting on the relative forecasting skill as determined by the pointwise empirical Winkler score (A.31). In this sense, our e-process for the weak Winkler null in Proposition A.4 is a weak-null counterpart of HZ's e-process.

In terms of the specific form of the e-process, (A.89) is an example of a *product* form e-process, contrasting with our *exponential* form variant. The two forms of e-processes are both found the lit-

erature, such as the product form in Waudby-Smith and Ramdas (2023) and the exponential form in Howard et al. (2021) for estimating bounded means. Also, while the e-process we derive in (3.20) explicitly shows its variance-adaptive property and further utilizes the method of mixtures (Robbins, 1970), HZ's e-process seeks to optimize its power by optimizing the growth rate of the e-process in the worst case (GROW) (Grünwald et al., 2019) under a chosen alternative (typically set to a convex combination of p_t and q_t).

In terms of use cases, the CSs perform estimation and thus provide information as to exactly *how much* one forecaster is outperforming the other. The methods in our work are agnostic to the different types of outcomes (Section A.7), so they can, e.g., be applied to forecasts on categorical outcomes with $C > 2$ categories and to forecasts on bounded continuous outcomes. HZ's approach is applicable to any consistent scoring functions (Gneiting, 2011) on binary outcomes and can also test for forecast dominance w.r.t. all consistent scoring functions.

A.8.2 Comparison with DM and GW Tests

As we highlighted in Section 3.2, the key difference between our work and existing forecast comparison methods, such as Diebold and Mariano (1995); Giacomini and White (2006); Lai et al. (2011); Ehm and Krüger (2018), is whether they have an anytime-valid guarantee. Here, we present additional experiments to illustrate that (i) the DM and GW tests are *not* valid at arbitrary stopping times, like most other classical tests including Lai et al. (2011), and (ii) anytime-valid methods need not require larger sample sizes than DM and GW tests for high power.

To recap, the DM test of *unconditional* predictive ability that tests

$$\mathcal{H}_0^{\mathsf{DM}} : \mathbb{E}[\hat{\delta}_n] = 0, \quad \forall n \geq 1, \tag{A.90}$$

where the scoring rule is assumed to depend only on the forecast error, e.g., $S(p_n, y_n) = 1-(p_n-y_n)^2$. By the DM assumption, the loss differentials are assumed to be covariance stationary, implying that $\mathbb{E}[\hat{\delta}_n] = \delta$ for some fixed δ at any n. Given the (stationary) autocovariance function $\gamma(k)$ for score differentials and a consistent estimator $\hat{f}(0)$ of its spectrum at frequency zero, the DM test uses the asymptotic normality under $\mathcal{H}_0^{\mathsf{DM}}$ given by $\sqrt{n}(\hat{\Delta}_n - \mu)/\sqrt{2\pi\hat{f}(0)} \rightsquigarrow N(0,1)$.

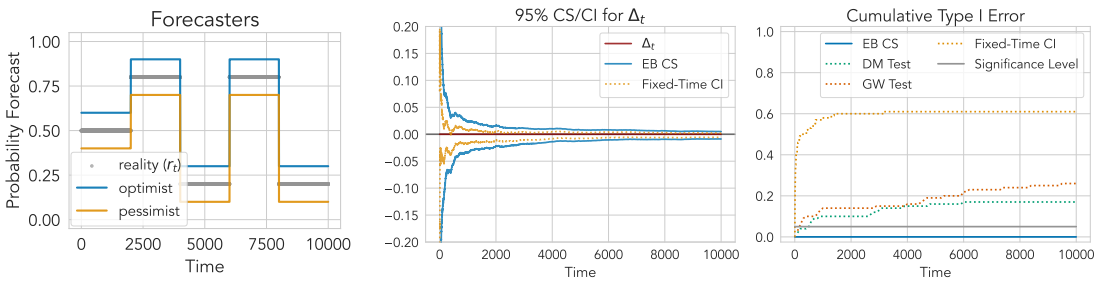


Figure A.2: *Left:* Two forecasters, denoted as `optimist` (blue) and `pessimist` (orange), on a simulated reality sequence (gray). There is no performance gap between the two in Brier score. *Middle:* The true average score differentials $(\Delta_t)_{t=1}^T$ (dark red) along with the 95% EB CS (blue) and the fixed-time CI (yellow). *Right:* Comparing the cumulative type I error rate for the EB CS (blue), the DM test of unconditional predictive ability (orange), the GW test of conditional predictive ability (brown), and Lai et al. (2011)'s asymptotic CIs (yellow). All tests are for one-sided nulls of the form "optimist performs no better than the pessimist." Unlike the EB CS, all classical fixed-time methods, including DM and GW tests, incur a cumulative miscoverage/false decision rate higher than $\alpha = 0.05$.

The GW test, on the other hand, is a test of *conditional* predictive ability that tests

$$\mathcal{H}_0^{\text{GW}} : \mathbb{E}_{n-1}[\hat{\delta}_{m,n}] = 0, \quad \forall n \geq 1. \tag{A.91}$$

Here, m is the maximum window size that each forecaster can look back to, meaning that the test now depends on the forecasting model. The GW assumption allows for nonstationarity, although the test statistic involves weights that depend on mixing assumptions (Lai et al., 2011).

First, we consider a simplistic setting in which $\Delta_t = 0$ for each time t and both the DM and GW assumptions are met. We compare two forecasters, named `optimist` (p_t) and `pessimist` (q_t), that are equally apart from Reality (r_t) in their forecasts (Figure A.2, left). For all methods, we test their form of the null that "the `optimist` is no better than the `pessimist`" under the Brier score. As expected, both the EB CS (Theorem 3.2) and the fixed-time CI (Lai et al., 2011) to quickly shrink to zero (Figure A.2, middle), and also neither the DM nor GW test falsely rejects the null at $T = 10,000$.

Now, we can also compute the cumulative type I error rate, which for p-values (p_t) is given by $\alpha_t = P(\exists i \leq t : \mathrm{p}_i \leq \alpha)$. For CS/CIs (C_t), this is equivalent in this case to the cumulative miscoverage rate $\alpha_t = P(\exists i \leq t : 0 \notin C_i)$ that we used earlier in Section 3.5.1, because $\Delta_t = 0$ under any $P \in \mathcal{H}_0$. The quantity is estimated over a repeated sampling of the data under P. We expect that an anytime-valid procedure satisfies $\alpha_t \leq \alpha$ for any t by definition, whereas classical fixed-time tests such as the

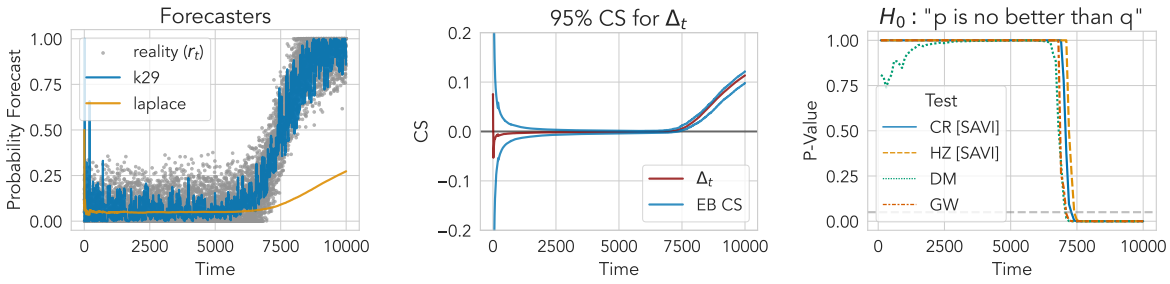


Figure A.3: *Left:* Two forecasters, `k29` (blue) and `laplace` (orange), on a simulated reality sequence (gray) that induces a changepoint in the loss differentials later in the time horizon. *Middle:* The 95% EB CS for $(\Delta_t)_{t=1}^T$ using the Brier score. Δ_t stays zero initially but trends positive later. *Right:* P-values for the null "`k29` is no better than `laplace`" at each sample size t. CR (ours; blue) and HZ (yellow) are anytime-valid (SAVI), whereas DM (green) and GW (orange) are not. When Δ_t quickly trends positive ($t \approx 7300$), all p-values shrink to zero, and neither CR nor HZ requires substantially many extra samples to get to zero compared to DM and GW.

DM and GW tests do not. As shown Figure A.2 (right), the cumulative type I errors of both the DM and GW tests exceed the significance level of $\alpha = 0.05$ after roughly 100 and 1000 steps, respectively, and they continue to trend upward in log-scale. This confirms that the p-values obtained by DM or GW tests, much like the fixed-time CI, are overconfident under continuous monitoring and thus at data-dependent stopping times, even when their assumptions are met. In other words, the DM and GW tests, along with fixed-time CIs, do not have an anytime-valid guarantee.

Next, we show that the anytime-validity of SAVI methods (CSs, e-processes, and p-processes), do not necessarily require larger sample sizes than the classical tests. We compare two forecasters, `k29` with a 3-degree polynomial kernel (p_t) and `laplace` (q_t), whose average and pointwise score differentials stay close to zero for a while ($t \leq 7000$) until a sharp changepoint in the data is introduced and Δ_t trends positive afterwards (Figure A.3, left). Note that this invalidates the covariance stationarity assumption of the DM test. The EB CS for Δ_t is drawn in the middle plot of Figure A.3, which shows that the CS uniformly covers the time-varying average as expected.

To illustrate that SAVI approaches do not necessarily require larger sample sizes for "detecting" this changepoint, we compare SAVI and non-SAVI p-values for the null that "`k29` is no better than `laplace`" under the Brier score. First, we plot the p-process $\mathrm{p}_t = 1/\sup_{i \leq t} E_i$, where $(E_t)_{t=0}^\infty$ is the sub-exponential e-process (3.20) that corresponds to the LCB of the CS. This is denoted in the right plot of Figure A.3 (denoted as "CR"). We also plot the p-process constructed from Henzi and Ziegel (2022)'s

e-process $(E_t^{\mathsf{HZ}})_{t=0}^{\infty}$ via the same mapping, i.e., $\mathfrak{p}_t^{\mathsf{HZ}} = 1/\sup_{i \le t} E_i^{\mathsf{HZ}}$. As shown in the plot, when compared against the DM and GW p-values, both our and HZ's p-processes shrink to zero nearly as quickly, indicating that they require comparable amounts of data to reject the null when Δ_t trends positive.

A.9 Additional Experiment Details and Results

A.9.1 Additional Details & Results from Numerical Simulations

Data Generation The reality sequence $(r_t)_{t=1}^{T}$ is specifically chosen to be non-IID and contain sharp changepoints, as drawn with gray dots in Figure 3.2:

$$r_t = [0.8 \cdot \theta_t + 0.2 \cdot (1 - \theta_t)] + \epsilon_t,$$

where

$$\theta_t = \begin{cases} 0.5 & \text{for } t \in [1, 2000] \\ 1 & \text{for } t \in [2001, 4000] \\ 0 & \text{for } t \in [4001, 6000] \\ 1 & \text{for } t \in [6001, 8000] \\ 0 & \text{for } t \in [8001, 10000] \end{cases}$$

and $\epsilon_t \sim \mathcal{N}(0, 0.1^2)$ is an independent Gaussian noise for each t.

All Pairwise Comparisons in Numerical Simulations In Figure A.4, we plot the 95% EB, Hoeffding-style, and asymptotic CSs for all pairwise comparisons between the constant baseline (`constant_0.5`), the Laplace forecaster (`laplace`), and the K29 forecasters with the 3-degree polynomial kernel and the Gaussian RBF kernel with bandwidth 0.01 (`k29_poly3` and `k29_rbf0.01`, respectively). The Brier score is used. Across all pairwise comparisons, both CSs uniformly cover the true score differentials across all times, regardless of whether the score differentials contain sharp changepoints and contain specific trends.

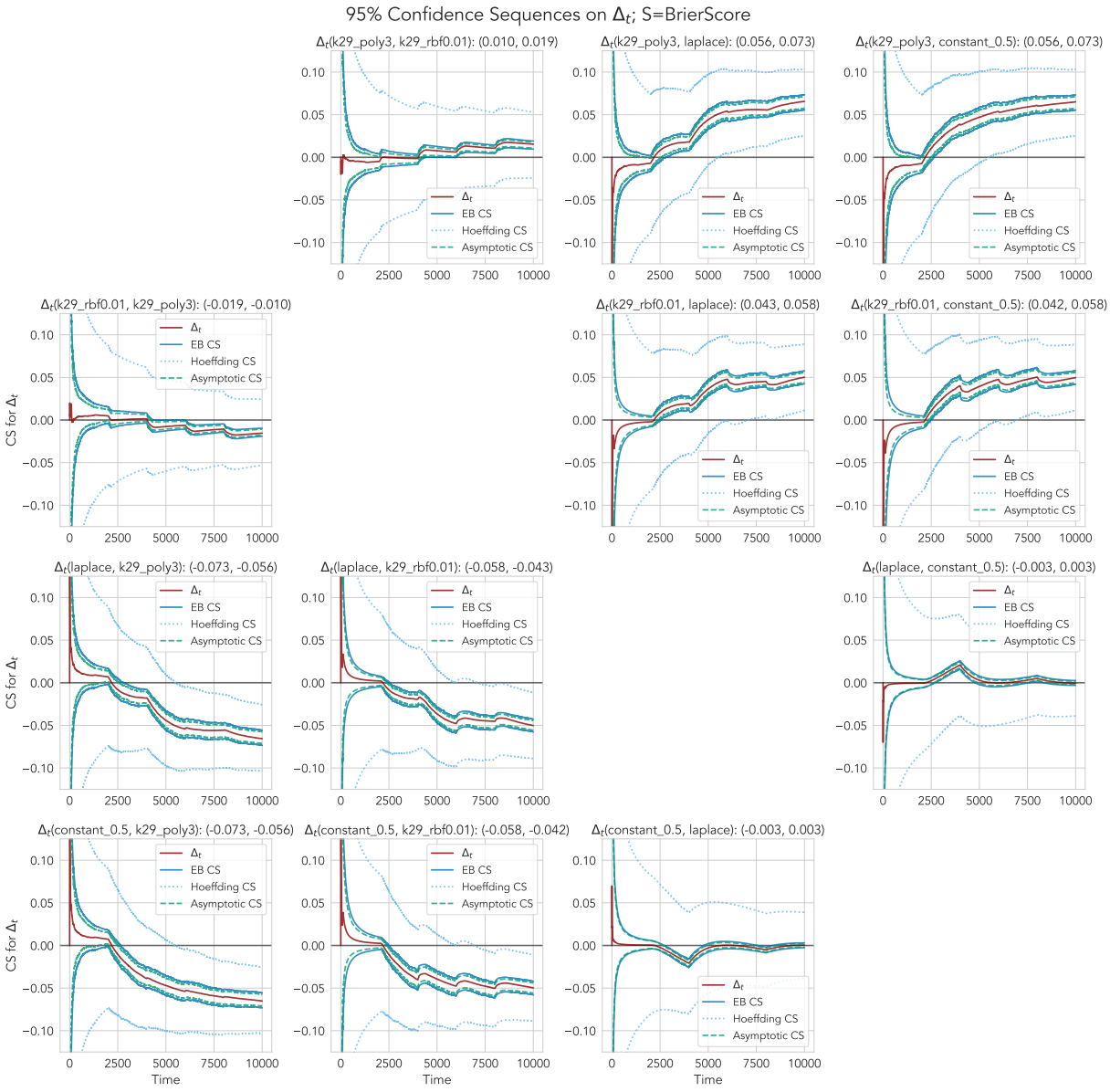


Figure A.4: 95% EB (blue), Hoeffding-style (skyblue), and asymptotic (green) CSs on Δ_t between four different forecasters (`k29_poly3`, `k29_rbf0.01`, `laplace`, and `constant_0.5`) plotted in Figure 3.2. Scoring rule is the Brier score, and positive values of Δ_t indicate that the first forecaster is better than the second. In all comparisons, both CSs cover Δ_t uniformly, and the width of the EB CS approaches that of the asymptotic CS as time grows large.

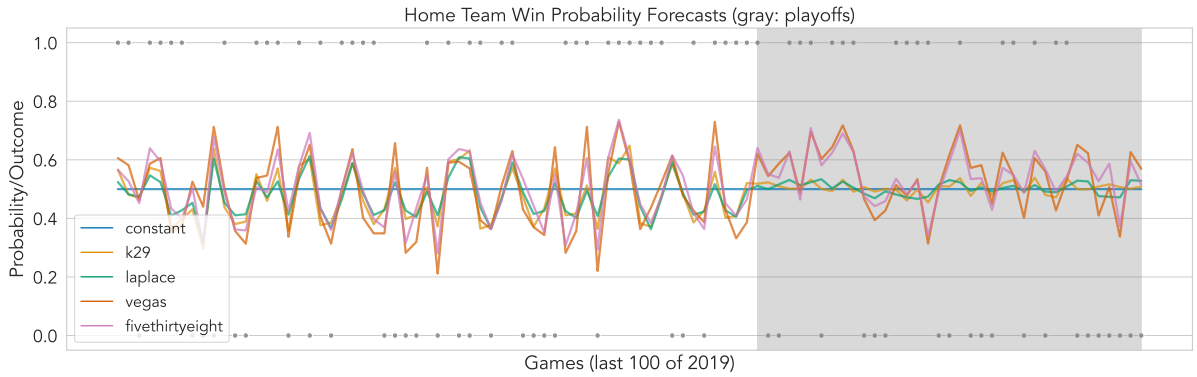


Figure A.5: Various forecasters on the last 100 MLB games played in 2019 (including regular season and postseason). FiveThirtyEight and Vegas forecasts are publicly available forecasts online; Laplace and K29 forecasts are made using historical outcomes as data without external information. *Note that the forecasts are computed using data from a 10-year window (2010 to 2019), but we only show the last 100 games here for visualization purposes.* The shaded region highlights the playoff games.

A.9.2 Additional Details & Results from the MLB Experiment

For all MLB-related experiments, we choose $v_{\text{opt}} = 100$, given the longer time horizon considered (compared to other experiments in this work).

Details on the MLB Forecasters Here, we describe the five Major League Baseball (MLB) forecasters that are compared in Section 3.5.2. Figure A.5 illustrate their forecasts on the last 100 games.

- `538`: Game-by-game probability forecasts on every MLB game since 1871, available at `https://data.fivethirtyeight.com/#mlb-elo`. According to the methodology report at `https://fivethirtyeight.com/features/how-our-mlb-predictions-work/`, the probabilities are calculated using an ELO-based rating system for each team, and game-specific adjustments are made for the starting pitcher as well as other external factors (travel, rest, home field advantage, etc.). Before each new season, team ratings are reverted to the mean by one-third and combined with preseason projections from other sources (Baseball Prospectus's PECOTA, FanGraphs' depth charts, and Clay Davenport's predictions).
- `vegas`: Pre-game closing odds made on each game by online sports bettors, as reported by `https://Vegas-Odds.com`. (Download source: `https://sports-statistics.com/sports-data/mlb-historical-odds-scores-datasets/`.) The betting odds are given in the American for-

mat, so each odds o is converted to its implied probability p via $p = \mathbb{1}(o \geq 0)\frac{100}{100+o} + \mathbb{1}(o < 0)\frac{-o}{100-o}$. Then, for each matchup, the pair of implied probabilities for each team is rescaled to sum to 1. For example, given a matchup between team A and team B with betting odds $o_A = -140$ and $o_B = +120$, the implied probabilities are $\tilde{p}_A = 0.58$ and $\tilde{p}_B = 0.45$, and the rescaled probabilities are $p_A = 0.56$ and $p_B = 0.44$.

- `constant`: a constant baseline predicting $p_t = 0.5$ for each t.
- `laplace`: A seasonally adjusted Laplace algorithm, representing the season win percentage for each team. Mathematically, it is given by $p_t = \frac{k_t + c_t}{n_t + 1}$, where k_t is the number of wins so far in the season, n_t is the number of games played in this season, and $c_t \in [0, 1]$ is a baseline that represents the final probability forecast from the previous season, reverted to the mean by one-third. For example, if the previous season ended after round t_0, then $k_t = \sum_{i=t_0}^{t-1} \mathbb{1}(y_i = 1)$, $n_t = t - t_0$, and $c_t = \frac{2}{3} \cdot p_{t_0} + \frac{1}{3} \cdot \frac{1}{2}$ (with $c_0 = \frac{1}{2}$). The final probability forecast for a game between two teams is rescaled to sum to 1.
- `k29`: The K29 algorithm applied to each team, using the Gaussian kernel with bandwidth 0.1, computed using data from the current season only. The final probability forecast for a game between two teams is rescaled to sum to 1.

All Pairwise Comparisons of MLB Forecasters Figure A.6 includes all pairwise comparisons between the five MLB forecasters considered in our experiment. See main text from Section 3.5.2 for further details.

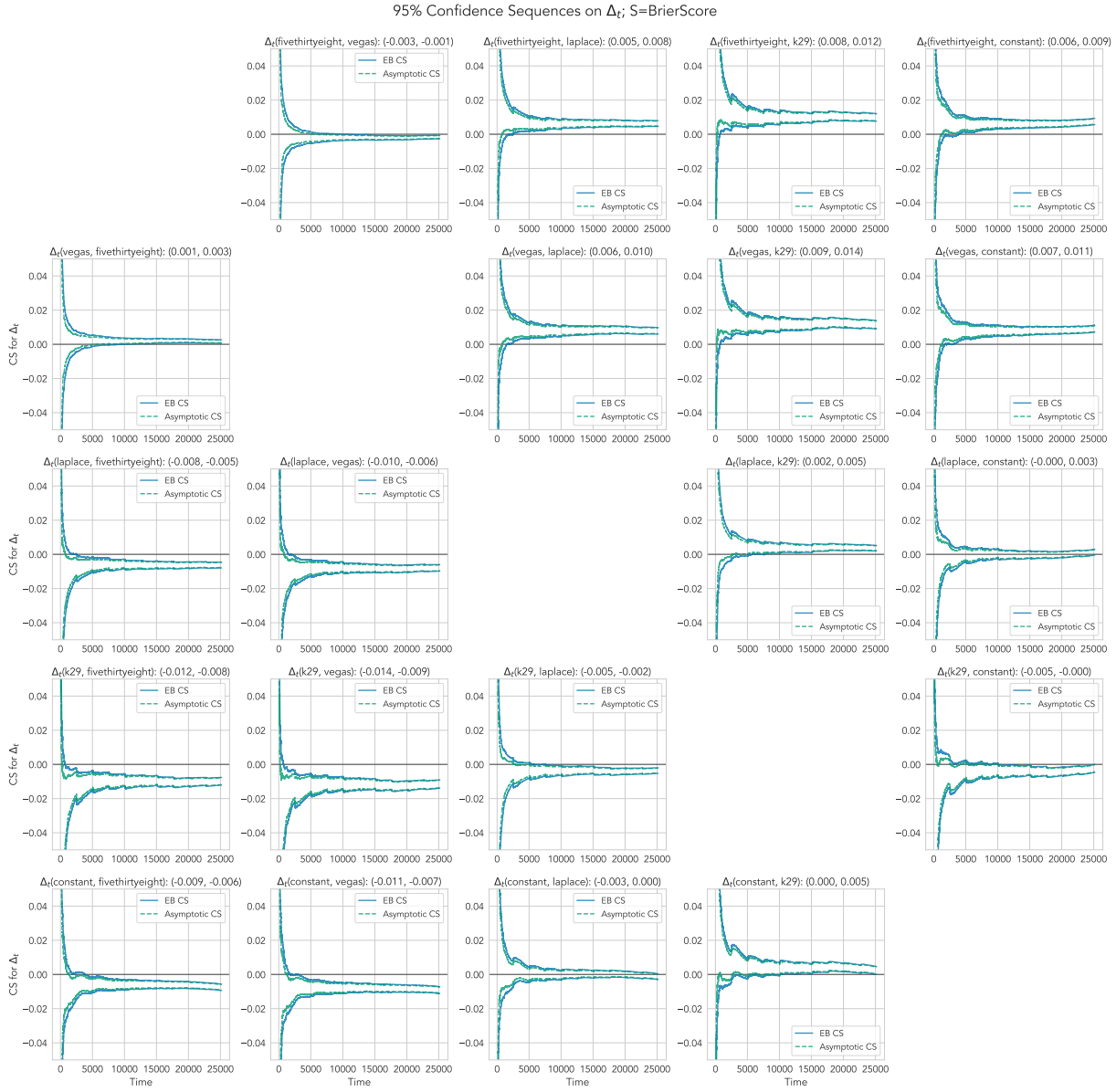


Figure A.6: Comparing MLB win probability forecasts from 2010 to 2019, using the EB and Hoeffding-style CSs at significance level $\alpha = 0.05$. $T = 25,165$ corresponds to the final game of the 2019 World Series. Note that the horizontal axis is drawn in log-scale. The Brier score is used. We find that, over time, the five forecasters are found to achieve significantly different predictive performance from each other (except `laplace` and `constant`), with the `vegas` forecaster achieving the best performance, followed by `fivethirtyeight`, `laplace` $\approx$ `constant`, and `k29`. The title of each subplot includes the 95% EB CS at $T = 25,165$.

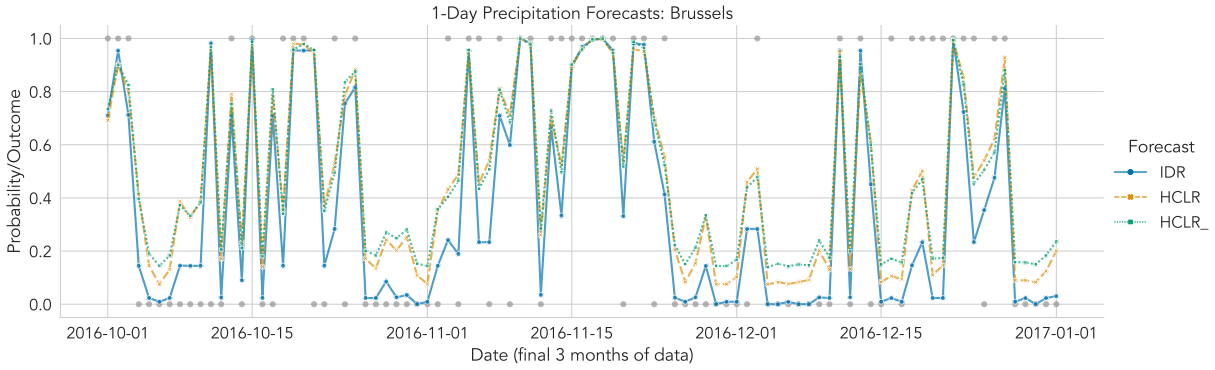


Figure A.7: Comparing three statistical postprocessing methods (IDR, HCLR, HCLR_) for 1-day ensemble weather forecasts on the Probability of Precipitation (PoP). The binary outcome is drawn as gray dots. *For visualization purposes, we plot the data and the forecasts only for the final 3 months (October 01, 2016 to January 01, 2017) and at one airport location (Brussels).*

A.9.3 Additional Details & Results from the Weather Experiment

The setup closely follows the comparison experiment by Henzi and Ziegel (2022), who compare statistical postprocessing methods for predicting the Probability of Precipitation (PoP) using the ensemble forecast data from the European Centre for Medium-Range Weather Forecasts (ECMWF; Molteni et al. (1996)). The dataset includes the observed 24-hour precipitation from January 06, 2007 to January 01, 2017 at four airport locations (Brussels, Frankfurt, London Heathrow, and Zurich), and for each location and date it also includes 1- to 5-day ensemble forecasts, consisting of a higher resolution forecast, 50 perturbed ensemble forecasts at a lower resolution, and a control run for the perturbed forecasts. They consider three statistical postprocessing methods in their experiments: isotonic distributional regression (IDR; Henzi et al. (2021)), heteroscedastic censored logistic regression (HCLR; Messner et al. (2014)), and a variant of HCLR without its scale parameter (HCLR_). Each method is applied to the first half of the data, separately for each airport location and lag $h = 1, \dots, 5$, and the second-half data is used to make sequential comparisons of the postprocessing methods. Note that each location has a different number of observations: 3,406 for Brussels, 3,617 for Frankfurt, 2,256 for London, and 3,241 for Frankfurt. See Section 5 in Henzi et al. (2021) and Section 5.1 in Henzi and Ziegel (2022) for further details about the dataset and the postprocessing methods.

In Figure A.7, we plot the three forecasters (1-day) on the probability of precipitation (PoP) for the final year (2016-2017) in Brussels.

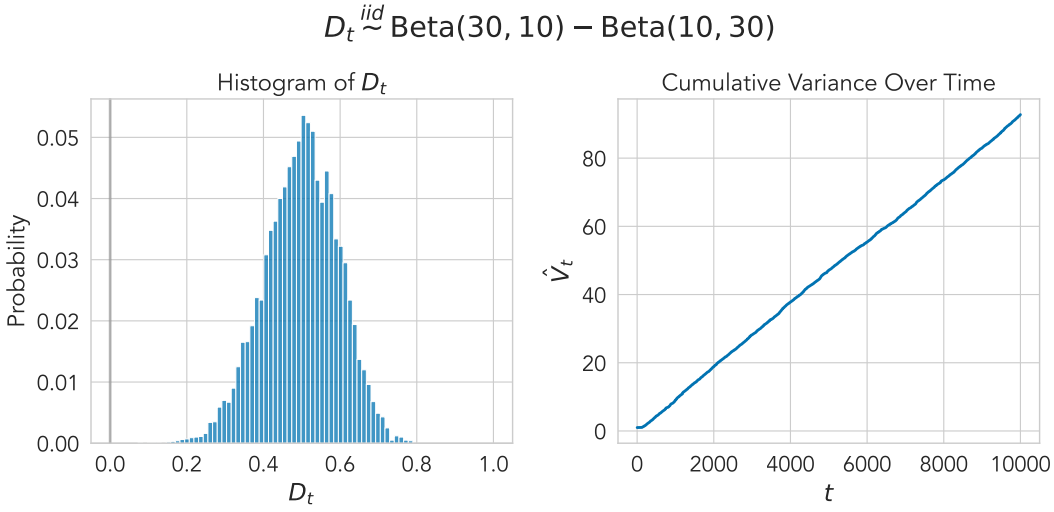


Figure A.8: (Left) Histogram of $\delta_i \overset{\text{IID}}{\sim} \text{Beta}(30, 10) - \text{Beta}(10, 30)$ for $i = 1, \ldots, 10,000$. (Right) Plot of the cumulative variance (intrinsic time) $\hat{V}_t = \sum_{i=1}^{t} (\hat{\delta}_i - \hat{\Delta}_{i-1})^2$, where $\hat{\Delta}_{i-1} = \sum_{j=1}^{i-1} \hat{\delta}_j$. Note that the horizontal axis t is drawn in log-scale. Also note that the hyperparameter v_{opt} determines the intrinsic time $\hat{V}_t$ at which the uniform boundary is the tightest.

A.9.4 Comparing CS Widths on IID Mean Differentials

The uniform boundaries we use in our CSs come with hyperparameter(s) that one can choose to optimize the CS widths at specific intrinsic times (i.e., values that the non-decreasing sequence $(\hat{V}_t)_{t=1}^{\infty}$ can take). As explained in Section A.2, this choice can be thought of as an additional fine-tuning step and is secondary to choosing the type of uniform boundary. Nevertheless, since it is a hyperparameter, we seek to find a reasonable default that can be used for typical scenarios of forecast comparison without an a priori knowledge of how large the intrinsic time can get.

To achieve this, we compare the widths of various time-uniform CSs for the mean differentials between two independent and identically distributed (IID) random variables. The main reason for using IID data is so that we can compare the width of our CSs with other CSs developed in previous work (Howard et al., 2021; Waudby-Smith and Ramdas, 2023; Waudby-Smith et al., 2021).

We compare both the Hoeffding-style CS (Theorem 3.1 and the empirical-Bernstein (EB) CS (Theorem 3.2) using both the conjugate-mixture (CM) (Section 3.4.3) and the polynomial stitching (Section A.2.2) uniform boundaries. We also include Hoeffding-style and EB CSs using the predictable-mixture (PM) boundary (Waudby-Smith and Ramdas, 2023), which is an efficient alternative to the conjugate-mixture boundary that can be used specifically for bounded IID means. We also include

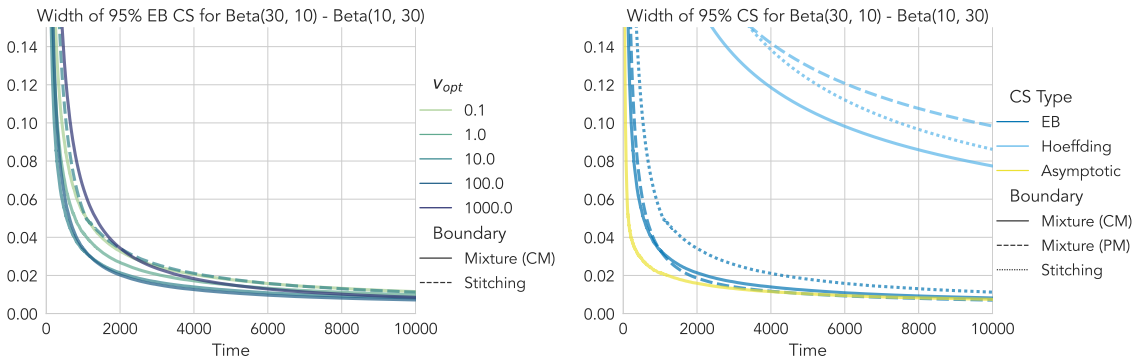


Figure A.9: *Left:* Hyperparameter tuning for the width of the conjugate-mixture EB CS by adjusting the optimal intrinsic time parameter v_{opt}. The choices $v_{\text{opt}} = 10$ and $v_{\text{opt}} = 100$ give the smallest widths overall, with the former being tighter early on and the latter later on. The width of stitching EB CS with $v_{\text{opt}} = 10$, drawn as a point of comparison, is wider than the mixture EB CS with $v_{\text{opt}} = 10$. *Right:* Comparing the widths of CS variants, including the conjugate- (CM) and predictable- (PM) mixture boundaries for EB/Hoeffding CSs, and also the asymptotic CS. Overall, the asymptotic CS is the tightest, although the mixture EB CSs get close; the stitching EB CS is slightly wider than the mixture variants, and all Hoeffding variants are considerably wider than the rest in this case.

the asymptotic CS (Waudby-Smith et al., 2021) that we described in Section A.3. As for the data, we use the difference between two IID Beta random variables, as a proxy for score differentials between two sets of forecasts: for $i = 1, \dots, 10,000$,

$$\delta_i \overset{\text{IID}}{\sim} \text{Beta}(30, 10) - \text{Beta}(10, 30). \tag{A.92}$$

Note that $-1 \leq \delta_i \leq 1$ a.s. and that $\mathbb{E}[\delta_i] = \frac{30}{30+10} - \frac{10}{10+30} = \frac{1}{2}$. Figure A.8 illustrates the data sampled according to (A.92) (left) as well as the intrinsic time $\hat{V}_t = \sum_{i=1}^{t} (\hat{\delta}_i - \hat{\Delta}_{i-1})^2$, where $\hat{\Delta}_{i-1} = \sum_{j=1}^{i-1} \hat{\delta}_j$, over log-scaled time (right).

Recall from Section A.2 that v_{opt} denotes the hyperparameter that specifies the intrinsic time at which the CS width is optimized. In Figure A.9 (left), we present the hyperparameter tuning results of several conjugate-mixture EB CSs with respect to its optimal intrinsic time hyperparameter v_{opt}. As a point of comparison, we include the Hoeffding-style CS and also the EB CS with the polynomial stitching bound, using $v_{\text{opt}} = 10$. Comparing the values of $v_{\text{opt}} \in \{0.1, 1, 10, 100, 1000\}$ for the mixture boundary, we find that the EB CS is generally the tightest across time with $v_{\text{opt}} = 10$ and $v_{\text{opt}} = 100$. Based on these results, we use the mixture EB CS with $v_{\text{opt}} = 10$ for our other experiments throughout the paper, unless specified otherwise.

In Figure A.9 (right), we now plot the widths of the 95% CS variants, optimized for the intrinsic time $v_{\text{opt}} = 10$ when applicable. We compare both the Hoeffding-style and EB CSs, the asymptotic CS (Section A.3); for the Hoeffding-style and EB CSs, we also include the predictable-mixture (PM) uniform boundary (Waudby-Smith and Ramdas, 2023), which is an efficient alternative to the conjugate-mixture boundary that can be used for IID means.

Generally speaking, we observe that the CSs are the tightest for the asymptotic CS, followed by the EB CS variants and the Hoeffding CS variants. This is consistent with our intuition, as the EB CS additionally makes use of the estimated variance to achieve smaller widths than the Hoeffding CS, and the asymptotic CS is the ideal "limit" of EB CS in terms of width. Among the EB CS variants, the conjugate-mixture variant is tighter towards the beginning ($t < 10^3$) while the predictable mixture becomes slightly tighter afterward, and the stitching CS is not as tight as the other two. This is also as expected, as both mixture CSs have similar widths (up to differences determined by the choice of hyperparameters) (Waudby-Smith and Ramdas, 2023) and the stitching CS tends to be looser in practice (Howard et al., 2021). This trend is also analogous for the Hoeffding CS variants, although the stitching variant does become tighter for larger t in this case.

Appendix B

Supplementary Materials for "Counterfactually Comparing Abstaining Classifiers"

B.1 Further Discussion

B.1.1 Additional Motivating Examples for the Counterfactual Score

Here, we include three additional examples that motivate the counterfactual score. These illustrate cases in which either (a) the missing predictions are utilized in a failure mode (Examples B.1 and B.2) or (b) the missing predictions are relevant to the evaluator's future uses (Examples B.2 and B.3).

Example B.1 (Inattentive driver in a self-driving car)**.** Consider an ML classifier in a semi-autonomous vehicle system that predicts a label (the weather, time of day, etc.) given the available sensory inputs. The predicted label is then used by the sequential decision making agent. In principle, when facing a high-uncertainty input, the classifier can abstain from a prediction and alert the driver to take back control. Yet, in reality, we would still greatly prefer a system that can make a safe decision in case the driver is inattentive[1] at the time, and cannot take back control. But how do we evaluate what a system would have done in situations where it decided to abstain from making a prediction?

[1]Driver inattention is a real issue in semi-autonomous vehicles; studies have shown that the lack of active involvement correlates with both driver fatigue and tardy reactions to take-over requests (Vogelpohl et al., 2019).

Example B.2 (Comparing ML radiologist assistants)**.** Consider a hospital that is evaluating third-party radiology APIs that can assist with its diagnosis system. An API will either give a prediction or abstain from making one, and then a human radiologist will examine the input on which the classifier chose to abstain (Raghu et al., 2019). The APIs can also abstain non-deterministically to improve upon their performance (Kalai and Kanade, 2021). Importantly, the hospital is wary that there are inputs for which the professional would also abstain or have cognitive biases against (Busby et al., 2018; Madras et al., 2018). Thus, it would need to occasionally rely on the classifier's predictions even on examples that it chose to abstain. If these "hidden predictions" are not readily available from the third-party providers (e.g., require extra costs), how can the hospital compare their services?

Example B.3 (Evaluating an abstaining classifier's internal biases)**.** Suppose that an independent agency is auditing an ML-based recidivism prediction system[2] that has been deployed for a certain amount of time. Given the high stakes of misclassification, the system is trained to occasionally (and randomly) abstain from making a prediction, such that the rejected cases can be examined by human judges. The auditing agency is interested in checking whether the ML classifier possesses internal biases against certain demographic groups, and in particular, it wants to estimate the classifier's accuracy on each demographic group *had it not abstained on any input*. While the agency has access to the system's past predictions and abstentions, it does not have access to the underlying predictive model or its abstention mechanism (i.e., black-box). How can the agency evaluate the system's biases while accounting for the missing predictions due to abstentions?

B.1.2 An Equivalent Formulation via the Potential Outcomes Framework

There are other equivalent ways to formulate our setup (Section 4.2.2) using variants of the potential outcomes framework. First, we can define a (potentially observed) prediction $f(X;R)$, which equals $f(X)$ if $R=0$ and $*$ if $R=1$, where the symbol $*$ indicates an abstention (the same notation is used in Rubin (1976)'s missing data framework). The score S is then $\mathrm{s}(f(X),Y)$ if $R=0$ and $*$ otherwise.

Alternatively, we can explicitly invoke Rubin (1974)'s potential outcomes framework to write $S(0) \leftarrow \mathrm{s}(f(X),Y)$ and $S(1) \leftarrow *$, where $S(r)$ refers to the score of the abstaining classifier when $R=r$

[2]Algorithmic approaches to recidivism prediction, such as COMPAS, have been both increasingly popular and highly controversial.

for each $r \in \{0,1\}$. We do not use this notation in our main chapter because $S(1)$ is not meaningful in our case.

B.1.3 Comparison with Condessa et al. (2017)'s Score

To better understand the counterfactual score $\psi = \mathbb{E}[S]$, we can contrast it with Condessa et al. (2017)'s notion of the 'classification quality score' θ. Assuming $S \in [0,1]$, the classification quality score is decomposed as follows:

$$\theta := \mathbb{E}[S \mid R=0]\,\mathbb{P}(R=0) + \mathbb{E}[1-S \mid R=1]\,\mathbb{P}(R=1). \tag{B.1}$$

In contrast, note that the counterfactual score is decomposed into

$$\psi = \mathbb{E}[S \mid R=0]\,\mathbb{P}(R=0) + \mathbb{E}[S \mid R=1]\,\mathbb{P}(R=1). \tag{B.2}$$

Thus, our target quantity ψ is large if the classifier is good on all inputs (abstentions or not), while θ is large if the classifier is good on points it predicts on but poor on points it abstains on. However, much like ψ, the challenge of estimating θ is driven entirely by the $\mathbb{E}[S|R=1]$ term, as the remaining terms are directly observed.

We note that estimates of ψ also yield estimates of θ, since $\theta + \psi$ is an observable quantity that can be straightforwardly estimated. Subtracting an estimate of ψ from the sum gives an estimate of θ.

B.1.4 The Plug-in and Inverse Propensity Weighting Estimators

The uniqueness of efficient influence functions tells us that the DR estimator outperforms two intuitive yet suboptimal estimators in an asymptotic and locally minimax sense. The first is the *plug-in estimator*, which is derived directly from the identified target $\psi = \mathbb{E}[\mu_0(X)]$ in Ppn. 4.2:

$$\hat{\psi}_{\mathsf{pi}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_0(X_i), \tag{B.3}$$

where $\hat{\mu}_0$ is any estimate of the regression function $\mu_0(x) = \mathbb{E}[S \mid R=0, X=x]$. The quality of this simple estimator directly depends on the estimation quality of $\hat{\mu}_0$ for μ_0, and in a nonparametric

setting, the estimator can suffer from the statistical curse of dimensionality. Another point of concern is that it makes no use of the missingness patterns.

The second is *inverse probability weighting (IPW)* estimator (Horvitz and Thompson, 1952; Rosenbaum, 1995):

$$\hat{\psi}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1-R_i}{1-\hat{\pi}(X_i)}S_i, \tag{B.4}$$

where $\hat{\pi}$ is an estimate of the abstention mechanism $\pi(x) = \mathbb{P}(R = 1 \mid X = x)$. If $\hat{\pi}$ consistently estimates π, the IPW estimator is unbiased; yet, it has the opposite problem to the plug-in estimator as it does not model the conditional score μ_0 at all.

B.2 Proofs

B.2.1 Proof of Proposition 4.1

Since (X, Y) is independent of the training data $\mathcal{D}_{\text{train}}$ for (f, π), and because ξ is an independent source of randomness, we can treat the functions f and π as fixed. Then, by definition, $S = \mathsf{s}(f(X), Y)$ is a deterministic function of (X, Y) and $R = \mathsf{r}(\pi(X), \xi)$ is a deterministic function of X and ξ. This means that the condition $S \perp\!\!\!\perp R \mid X$ is equivalent to saying that $Y \perp\!\!\!\perp \xi \mid X$. Given that ξ is independent of (X, Y), the latter condition follows.

B.2.2 Proof of Proposition 4.2

Positivity (Assumption 4.2) ensures that the conditional expectation $\mu_0(X) = \mathbb{E}[S \mid R = 0, X]$ is well-defined. Then,

$$\mathbb{E}[\mu_0(X)] = \mathbb{E}[\mathbb{E}[S \mid R = 0, X]] \overset{(\text{MAR})}{=} \mathbb{E}[\mathbb{E}[S \mid X]] = \mathbb{E}[S] = \psi, \tag{B.5}$$

where the second inequality follows from the MAR condition (Assumption 4.1), i.e., $S \perp\!\!\!\perp R \mid X$.

B.2.3 Proof Sketch of Theorem 4.1

We follow the relevant notations and derivations from Kennedy (2022). Denote $\mathbb{P}\{f\} = \mathbb{E}_{\mathbb{P}}[f(Z)]$ and $\mathbb{P}_n\{f\} = n^{-1}\sum_{i=1}^{n} f(Z_i)$ where $Z_i \overset{iid}{\sim} \mathbb{P}$. We use the *centered* influence function for $\psi(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[\mu_0(X)]$

(upon identification), defined as follows:

$$\mathtt{IF}_{\mathbb{P}}(x, r, s) := \left[\frac{1-r}{1-\pi(x)}\left(s - \mu_0(x)\right) + \mu_0(x)\right] - \psi(\mathbb{P}). \tag{B.6}$$

Here, $\mathtt{IF}_{\mathbb{P}}$ depends on $\mathbb{P}$, which determines π and μ_0. Analogously, we let $\hat{\mathbb{P}}$ denote the distribution of abstentions and score outcomes involving estimators $\hat{\pi}$ and $\hat{\mu}_0$ (in place of π and μ_0), and let $\mathtt{IF}_{\hat{\mathbb{P}}}$ and $\psi(\hat{\mathbb{P}})$ denote the corresponding influence function and target functional, respectively, defined using $\hat{\pi}$ and $\hat{\mu}_0$. Also, note that an uncentered version is shown in the main text for ease of explanation; the resulting variance does not change due to this centering. Using these definitions, we proceed with the proof in two steps.

Step 1: Showing that IF (B.6) **is the efficient influence function for** ψ**.** To show that IF is indeed the unique efficient influence function for ψ, we show that $\mathbb{P}\{\mathtt{IF}_{\mathbb{P}}\} = 0$ and that its bias term is second-order. The uniqueness and asymptotic efficiency of this EIF in a nonparametric setting, in general, is well-known (e.g., van der Vaart (2002)). First, observe that

$$\begin{aligned}
\mathbb{P}\{\mathtt{IF}_{\mathbb{P}}\} &= \mathbb{E}_{\mathbb{P}}\left[\frac{1-R}{1-\pi(X)}\left(S - \mu_0(X)\right) + \mu_0(X)\right] - \psi(\mathbb{P}) && \text{(B.7)}\\
&= \mathbb{E}_{\mathbb{P}}\left[\frac{\mathbb{E}\left[(1-R)(S-\mu_0(X)) \mid X\right]}{1-\pi(X)}\right] && \text{(B.8)}\\
&\overset{\text{(a)}}{=} 0, && \text{(B.9)}
\end{aligned}$$

where (a) follows from the fact that

$$\mathbb{E}\left[(1-R)S \mid X\right] = \pi(X)\cdot 0 + (1-\pi(X))\mathbb{E}\left[S \mid R=0, X\right] = (1-\pi(X))\mu_0(X). \tag{B.10}$$

Furthermore, for any distributions $\hat{\mathbb{P}}$ and $\mathbb{P}$, the bias term is given by

$$R_2(\hat{\mathbb{P}}, \mathbb{P}) = \psi(\hat{\mathbb{P}}) - \psi(\mathbb{P}) + \mathbb{P}\left\{\mathsf{IF}_{\hat{\mathbb{P}}}\right\} \tag{B.11}$$

$$= \psi(\hat{\mathbb{P}}) - \psi(\mathbb{P}) + \mathbb{E}_{\mathbb{P}}\left[\frac{1-R}{1-\hat{\pi}(X)}\left(S - \hat{\mu}_0(X)\right) + \hat{\mu}_0(X)\right] - \psi(\hat{\mathbb{P}}) \tag{B.12}$$

$$= \mathbb{E}_{\mathbb{P}}\left[\frac{1-R}{1-\hat{\pi}(X)}\left(S - \hat{\mu}_0(X)\right) + \hat{\mu}_0(X) - \mu_0(X)\right] \tag{B.13}$$

$$\overset{\text{(IE,a)}}{=} \mathbb{E}_{\mathbb{P}}\left[\frac{1-\pi(X)}{1-\hat{\pi}(X)}\left(\mu_0(X) - \hat{\mu}_0(X)\right) - \left(\mu_0(X) - \hat{\mu}_0(X)\right)\right] \tag{B.14}$$

$$= \mathbb{E}_{\mathbb{P}}\left[\frac{\left(\hat{\pi}(X) - \pi(X)\right)\left(\mu_0(X) - \hat{\mu}_0(X)\right)}{1-\hat{\pi}(X)}\right] \tag{B.15}$$

$$\leq \frac{1}{\epsilon} \cdot \left\|\hat{\pi} - \pi\right\|_{L_2(\mathbb{P})} \left\|\hat{\mu}_0 - \mu_0\right\|_{L_2(\mathbb{P})}. \tag{B.16}$$

This is a second-order product term in the difference of $\hat{\mathbb{P}}$ and $\mathbb{P}$, showing that IF is an influence function for $\mathbb{P}$.

Step 2: Showing the asymptotic normality of $\sqrt{n}(\hat{\psi}_{\mathsf{dr}} - \psi)$. To derive the explicit form of the limiting distribution, denote $\hat{\mathsf{IF}} = \mathsf{IF}_{\hat{\mathbb{P}}}$, and observe that the DR estimator is a "one-step" bias-corrected estimator (Bickel, 1975), given by $\hat{\psi}_{\mathsf{dr}} = \mathbb{P}_n\left\{\hat{\mathsf{IF}}\right\} + \psi(\hat{\mathbb{P}})$. Then, we have the following three-term decomposition:

$$\hat{\psi}_{\mathsf{dr}} - \psi = \mathbb{P}_n\left\{\hat{\mathsf{IF}}\right\} + \psi(\hat{\mathbb{P}}) - \psi(\mathbb{P}) \tag{B.17}$$

$$= \left(\mathbb{P}_n - \mathbb{P}\right)\left\{\hat{\mathsf{IF}}\right\} + R_2(\hat{\mathbb{P}}, \mathbb{P}) \tag{B.18}$$

$$= \left(\mathbb{P}_n - \mathbb{P}\right)\{\mathsf{IF}\} + \left(\mathbb{P}_n - \mathbb{P}\right)\left\{\hat{\mathsf{IF}} - \mathsf{IF}\right\} + R_2(\hat{\mathbb{P}}, \mathbb{P}). \tag{B.19}$$

The first term, which is a sample average term, has the desired limiting distribution by the central limit theorem:

$$\sqrt{n} \cdot \left(\mathbb{P}_n - \mathbb{P}\right)\{\mathsf{IF}\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\mathsf{IF}(Z_i) - \mathbb{E}_{\mathbb{P}}\left[\mathsf{IF}(Z)\right]\right] \rightsquigarrow \mathcal{N}\left(0, \mathsf{Var}_{\mathbb{P}}\left(\mathsf{IF}\right)\right). \tag{B.20}$$

Then, by Slutsky's theorem, it suffices to show that the other two terms are of order $o_{\mathbb{P}}(1/\sqrt{n})$. The third term, $R_2(\hat{\mathbb{P}}, \mathbb{P})$, is precisely the second-order bias term we derived in (B.16), and it is $o_{\mathbb{P}}(1/\sqrt{n})$

by the DR assumption (4.3).

The second term, called the empirical process term, can be shown to be of order $o_{\mathbb{P}}(1/\sqrt{n})$ when using cross-fitting to estimate $\hat{\mathbb{P}}$. Specifically, the sample splitting procedure guarantees that $\hat{\mathbb{P}} \perp\!\!\!\perp \mathbb{P}_n$ (where $\mathbb{P}_n$ now refers to the held-out fold in each step of cross-fitting), which is enough to show that

$$(\mathbb{P}_n - \mathbb{P})\{\hat{\mathsf{IF}} - \mathsf{IF}\} = O_{\mathbb{P}}\left(\frac{\|\hat{\mathsf{IF}} - \mathsf{IF}\|_{L^2(\mathbb{P})}}{\sqrt{n}}\right). \tag{B.21}$$

Since $\|\hat{\mathsf{IF}} - \mathsf{IF}\|_{L^2(\mathbb{P})} = o_{\mathbb{P}}(1)$ by assumption, the term itself is of order $o_{\mathbb{P}}(1/\sqrt{n})$ as desired. The loss of sample efficiency due to a single sample splitting can be recovered by the cross-fitting procedure. See, e.g., Lemma 1 and Proposition 1 of Kennedy (2022) for details.

B.2.4 Proof of Theorem 4.2

Given that $\mathsf{IF}_{\mathbb{P}}^{\mathsf{AB}} = \mathsf{IF}_{\mathbb{P}}^{\mathsf{A}} - \mathsf{IF}_{\mathbb{P}}^{\mathsf{B}}$, it is immediate that it is an influence function for $\Delta^{\mathsf{AB}} = \psi^{\mathsf{A}} - \psi^{\mathsf{B}}$ because $\mathbb{P}\{\mathsf{IF}_{\mathbb{P}}^{\mathsf{AB}}\} = \mathbb{P}\{\mathsf{IF}_{\mathbb{P}}^{\mathsf{A}}\} - \mathbb{P}\{\mathsf{IF}_{\mathbb{P}}^{\mathsf{B}}\} = 0$ and

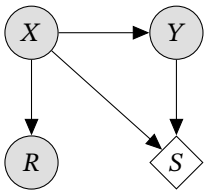
$$R_2(\hat{\mathbb{P}}, \mathbb{P}) \leq \frac{1}{\epsilon} \cdot \left(\|\hat{\pi}_{\mathsf{A}} - \pi_{\mathsf{A}}\|_{L_2(\mathbb{P})} \left\|\hat{\mu}_{0,\mathsf{A}} - \mu_{0,\mathsf{A}}\right\|_{L_2(\mathbb{P})} + \|\hat{\pi}_{\mathsf{B}} - \pi_{\mathsf{B}}\|_{L_2(\mathbb{P})} \left\|\hat{\mu}_{0,\mathsf{B}} - \mu_{0,\mathsf{B}}\right\|_{L_2(\mathbb{P})} \right). \tag{B.22}$$

The limiting distribution can also be derived analogously, where the upper bound in (B.22) reveals the additive form of the DR assumption (4.5).

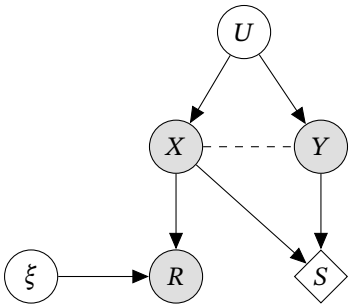
B.3 Illustration of the MAR Condition via Causal Graphs

Intuitively, the MAR condition is satisfied as long as the evaluation label is unknown to either classifier, simply because the classifier cannot access the actual score $S = \mathrm{s}(f(X), Y)$, which is a function of the true label Y, in making its abstention decision. This already implies $P(R = 1|S, X) = P(R = 1|X)$. We can further elucidate how the causal relationships between the random variables in our setup, and highlight how the MAR condition is generally satisfied, via graphical representations of the evaluation setup. The comparison case is an analogous extension to two abstaining classifiers.

Assuming that the abstaining classifier (f, π) does not depend on the evaluation output label Y, we (the evaluator) can treat both functions as fixed given the input X. We can then illustrate the MAR



(a) Simple DAG representation of our setup, assuming $X \to Y$.



(b) A more general graph that allows arbitrary relationships between X and Y as well as the classifier's internal randomness/bias (ξ).

Figure B.1: Two graphical representations of the random variables involved in our evaluation framework from §4.2, assuming that the true label Y is independent of the abstaining classifier (f, π). Shaded variables are observed by the evaluator; the score $S = \mathrm{s}(f(X), Y)$ is *partially* observed by the evaluator (depicted as a diamond node). In plot B.1a, assuming $X \to Y$, the simple DAG illustrates that S and R are d-separated given X. In plot B.1b, we further allow arbitrary relationships between X and Y, including $X \to Y$, $Y \to X$, and $U \to (X, Y)$ for some unobserved confounder U. The classifier's decision to abstain, R, is also allowed to additionally depend on some internal randomness and bias ξ that is independent of the evaluation data. Accounting for these generalizations, S and R are still d-separated given X, irrespective of the causal direction (if any) between X and Y.

condition via two causal graphs. First, suppose $X \to Y$ (for the sake of simplification). Then, we have the relationships $X \to Y, X \to R$ (Bernoulli with probability $\pi(X)$), and $(X, Y) \to S$ (deterministic via f and s). In the resulting graph, shown in Figure B.1a, the variables S and R are d-separated (Pearl, 2000) given X, i.e., $S \perp\!\!\!\perp R \mid X$. Note that S is partially observed and thus drawn as a diamond node, but it does not affect the conditional independence relationship. An alternative representation is possible via missingness graphs (Mohan et al., 2013), which would give us the same conclusion.

Next, we can remove the assumption on the relationship $X \to Y$, and allow any possible relationship between X and Y: $X \to Y$ (causal), $Y \to X$ (anticausal), or $U \to (X, Y)$, where U is an unobserved confounder to the prediction task. This is depicted as a dashed line between X and Y, along with a possible presence of U, in Figure B.1b. We can further allow the abstaining classifier to utilize some internal randomness or bias ξ, which is independent of the randomness in evaluation data, for its decision to abstain R. In the resulting graph, shown in Figure B.1b, none of the generalizations change the fact that S and R are d-separated given X, i.e., the MAR condition is satisfied.

Finally, as mentioned in the main text, the MAR condition can be violated when the evaluation data is not independent of the training data. For example, if the true label Y is used by the abstaining

classifier during its training to inform its abstention decision, then this would correspond to a graph in which there is an additional edge from Y to R, as the abstention function π now depends on Y. Then, S and R are no longer d-separated because there is a now connecting path via Y (common cause).

B.4 Positivity and Policy

Our identification results in §4.2.2 impose a requirement of positivity (Assumption 4.2) on the abstaining classifier (f, π), i.e., a demand that for some $\epsilon > 0$, the essential supremum of $\pi(x)$ is smaller than $1 - \epsilon$. This requirement is necessary: intuitively, if no feedback about the behaviour of f is available in a region, it is impossible (without further strong assumptions about the global structure of f) to determine the behaviour of the score in this region. Operationally, this is seen quite directly in the validity of the confidence intervals inferred from data (Figure B.5). Of course, the parameter ϵ also plays a quantitative role: the higher the ϵ, the better the validity and widths of our CIs. In other words, our ability to identify decays gracefully with ϵ, with complete inability if $\pi(x) = 1$ in a region of large mass.

While necessary, this positivity requirement is at odds with the practical deployment of client-facing abstaining classifiers. Indeed, there are two major reasons to implement an abstaining mechanism in such scenarios. In a positive sense, abstentions signal that the use of the underlying classifier f is inappropriate in a particular domain. However, in a negative sense, abstentions can also be employed in order to artificially limit a vendor's liability when their predictions (and the actions driven by the same) are incorrect. A pertinent example is the recent investigation of the Tesla autopilot by the NHTSA (2022) which found that in 16 incidents, the autopilot would deactivate and hand-off control to the driver at the very last seconds before a crash, thus artificially inflating the safety metrics of the system.

Part of the impetus behind studying a metric such as the counterfactual score is precisely to identify such behaviours before unsafe incidents bring them to light. Nevertheless, if vendors can stymie this investigation simply by ensuring that abstention is accompanied by a very high $\pi(x)$, then the method is not particularly useful.

This technical impasse begs for a policy-level treatment: through regulatory action, the executive may ensure that vendors supply evaluators (whether government agencies or independent reviewers) with abstaining classifiers that reveal the counterfactual decision of f at least an ϵ-fraction of the times when the decision is to abstain, where ϵ is set by mutual agreement of the stakeholders. Note that it is not enough to just supply evaluators with the predictions of f (although this would solve our particular problem formulation), since it is important to understand its behaviour in the context of when the abstaining classifier actually tends to reject points (i.e., it is equally important for evaluators and users to understand $\mathbb{E}[S \mid R = 1]$, which of course is estimable under our setup).

B.5 Confidence Sequences for Anytime-Valid Counterfactual Score Estimation

The nonparametric efficiency result of Theorem 4.1 yields an optimal inference procedure (either a hypothesis test or a confidence interval) for evaluating and comparing abstaining classifiers at a fixed sample size. Here, we go one step further and utilize a *confidence sequence (CS)* (Darling and Robbins, 1967; Howard et al., 2021), which is a sequence of confidence intervals whose validity holds uniformly over all sample sizes. This *time-uniform* property allows the evaluator to continuously monitor the result as more data is collected over time. The time-uniform property also implies *anytime-validity* (Johari et al., 2022; Grünwald et al., 2019), which allows the evaluator to run the experiment without pre-specifying the size of the evaluation set and compute the CIs as more data is collected. This implies that anytime-valid methods avoid the issue of inflated miscoverage rates coming from "data peeking." See Ramdas et al. (2022a) for an introduction.

Formally, for any $\alpha \in (0, 1)$, a $(1-\alpha)$-level (non-asymptotic) CS $(C_t)_{t\geq 1}$ for a parameter $\theta \in \mathbb{R}$ is a sequence of confidence intervals (CI) such that

$$\mathbb{P}(\forall t \geq 1 : \theta \in C_t) \geq 1 - \alpha. \tag{B.23}$$

Importantly, a CS contrasts with a fixed-time CI, whose guarantee no longer remains valid at stopping times: a CI only satisfies $\mathbb{P}(\theta \in C_t) \geq 1 - \alpha$ for a fixed sample size t.

Here, we describe how we can perform the proposed counterfactual comparison of abstaining classifiers using a variant of a CS that is asymptotic and readily applicable to causal estimands (Waudby-Smith et al., 2021). An (two-sided) $(1-\alpha)$-*asymptotic CS (AsympCS)* $(\tilde{C}_t)_{t\geq 1}$ for a parameter $\theta \in \mathbb{R}$ is a sequence of intervals, $\tilde{C}_t = (\hat{\theta}_t \pm \tilde{B}_t)$, for which there exists a non-asymptotic CS $(C_t)_{t\geq 1}$ for θ of the form $C_t = (\hat{\theta}_t \pm B_t)$ that satisfies

$$B_t / \tilde{B}_t \xrightarrow{\text{a.s.}} 1. \tag{B.24}$$

The AsympCS has an *approximation rate* of r_t if $\tilde{B}_t - B_t = O(r_t)$ almost surely.

Intuitively, an AsympCS is an arbitrarily precise approximation of a non-asymptotic CS. Because no known non-asymptotic CS exists for counterfactual quantities such as the ATE, AsympCS has been derived as an (only) viable alternative. Waudby-Smith et al. (2021) further leverage the (previously described) nonparametric efficiency theory and doubly robust estimation to derive an AsympCS for the ATE in randomized experiments and observational studies; we apply their theory to estimating the counterfactual scores and their differences. The resulting AsympCS is asymptotically time-uniform and anytime-valid, and its width scales similarly, up to logarithmic factors, to a fixed-time CI derived directly from Theorem 4.1.

Now we describe our main theorem for anytime-valid and counterfactual evaluation of an abstaining classifier. We consider evaluating the classifier on an i.i.d. test set that is continuously collected over time; let n be the (data-dependent) sample size with which inference is performed. As before, the comparison problem reduces to evaluating each abstaining classifier and taking their difference. We suppose that the nuisance functions $\hat{\pi}$ and $\hat{\mu}_0$ are learned via cross-fitting, and these are used to compute the EIF estimate (4.2). Now we can formally state an asymptotic CS for $\psi = \mathbb{E}[S]$ (4.1) that is anytime-valid and doubly robust. In the below, the $o(\cdot)$ notation refers to almost sure convergence.

Theorem B.1 (Anytime-valid DR estimation of the counterfactual score)**.** *Suppose that $\hat{\mu}_0$ and $\hat{\pi}$ consistently estimates μ_0 and π in $L_2(\mathbb{P})$, respectively, at a product rate of $o(\sqrt{\log\log n/n})$:*

$$\|\hat{\mu}_0 - \mu_0\|_{L_2(\mathbb{P})} \|\hat{\pi} - \pi\|_{L_2(\mathbb{P})} = o(\sqrt{\log\log n/n}). \tag{B.25}$$

Also, suppose that $\|\hat{\mathsf{IF}} - \mathsf{IF}\|_{L_2(\mathbb{P})} = o(1)$ and that IF has at least four finite moments.

Then, under Assumption 4.1 and 4.2, for any choice of $\rho > 0$,

$$\hat{\psi}_{\mathrm{dr}} \pm \sqrt{\widehat{\mathrm{Var}}_n\left(\widehat{\mathrm{IF}}\right)} \cdot \sqrt{\frac{2n\rho^2+1}{n^2\rho^2} \log\left(\frac{\sqrt{n\rho^2+1}}{\alpha}\right)} \tag{B.26}$$

forms a $(1-\alpha)$-AsympCS for ψ with an approximation rate of $\sqrt{\log\log n/n}$.

This result is an adaptation of Theorems 2.2 and 3.2 in Waudby-Smith et al. (2021) to our setup. The assumptions on $\hat{\pi}$ and $\hat{\mu}_0$ are analogous to the double robustness assumptions (4.3) in Theorem 4.1, as they require the same product rate up to logarithmic factors. Here, ρ is a free parameter that can be chosen to optimize the CS width (see Appendix C.3 of Waudby-Smith et al. (2021) for details).

Compared to the fixed-size CI of (4.1), whose width shrinks at a $O(1/\sqrt{n})$ rate, the width of the AsympCS in (B.26) shrinks at a $O(\sqrt{\log n/n})$ rate. This means that, in terms of the CI width, the extra cost of ensuring anytime-validity is logarithmic in n. In practice, the AsympCS may be wider than the CI from Theorem 4.1; nevertheless, the AsympCS may be preferred in scenarios where the evaluation/comparison is performed on continuously collected data. Another potential benefit of the AsympCS is the extension to settings with sequential and time-varying evaluation tasks (e.g., involving time-series forecasters that abstain). We leave the formalization of the time-varying setup as future work.

Finally, to apply Theorem B.1 to a comparison setting, we can construct two $(1-\alpha/2)$-AsympCSs, $C_n^{\mathsf{A}} = (L_n^{\mathsf{A}}, U_n^{\mathsf{A}})$ and $C_n^{\mathsf{B}} = (L_n^{\mathsf{B}}, U_n^{\mathsf{B}})$ for ψ^{A} and B respectively, and then combine them into one $(1-\alpha)$-AsympCS for $\Delta^{\mathsf{AB}} = \psi^{\mathsf{A}} - \psi^{\mathsf{B}}$ via $C_n^{\mathsf{AB}} = (L_n^{\mathsf{A}} - U_n^{\mathsf{B}}, U_n^{\mathsf{A}} - L_n^{\mathsf{B}})$.

B.6 Additional Experiments and Details

B.6.1 Details on the Simulated Data and Abstaining Classifiers

The evaluation set is generated as follows: $(X_{0i}, X_{1i}) \sim \mathsf{Unif}[0,1]$, $E_i \sim \mathsf{Ber}(0.15)$, and $Y_i = \mathbb{1}(X_{0i} + X_{1i} \geq 1)$ if $E_i = 0$ and $Y_i = \mathbb{1}(X_{0i} + X_{1i} < 1)$ otherwise (label noise). Classifier A uses a logistic regression model with the optimal linear decision boundary, i.e., $f^{\mathsf{A}}(x_0, x_1) = \sigma(x_0 + x_1 - 1)$, where $\sigma(u) = 1/(1+\exp(-u))$, achieving an accuracy of 0.85 by design. Classifier B, on the other hand,

has a (suboptimal) curved boundary: $f^{\mathsf{B}}(x_0, x_1) = 0 \vee (\frac{1}{2}(x_0^2 + x_1^2) + \frac{1}{10}) \wedge 1$. Classifier A is thus "oracle" logistic regression model with the same decision boundary, achieving an empirical score of 0.86 before abstentions; classifier B is a biased model that achieves an empirical score of 0.74 before abstentions.

For both classifiers, $\epsilon = 0.2$ determines the coefficient for positivity, and they are designed to abstain more frequently near their decision boundaries. For classifier A, $\pi^{\mathsf{A}}(x) = 1 - \epsilon$ if the distance from x to its boundary is less than δ, and $\pi^{\mathsf{A}}(x) = \epsilon$ otherwise; for classifier B, we use 0.8δ as the threshold, resulting in less abstentions than A. In some sense, this is a setting where ϵ-positivity is "minimally" satisfied because the abstention rate is always either ϵ or $1 - \epsilon$, and not in between, in all regions of the input space. If, say, the abstention rate was 0.5 in most parts but ϵ in a small region, the positivity level would still be ϵ but the estimation would in general be easier. Thus, this example can be viewed as a more challenging case than a standard causal inference setup with small regions of ϵ-positivity.

Figure B.2 shows both the predictions (blue circles: 0, green triangles: 1) and the abstention decisions (orange x's: predictions) for each classifier. Each classifier has a high chance of abstaining near its boundary (shaded orange region) and a low chance otherwise, meaning that abstentions are *not* spread out uniformly (MAR but not MCAR). In particular, classifier B hides many of its misclassifications as abstentions, leading to its high selective score ($\mathsf{Sel}^{\mathsf{B}} = 0.81$) relative to its counterfactual score ($\psi^{\mathsf{B}} = 0.74$).

The nuisance functions $\hat{\pi}$ and $\hat{\mu}_0$ for each classifier A and B are learned via 2-fold cross-fitting. In each case, we cap extreme propensity predictions by $\hat{\pi}^{\mathsf{A}}$ and $\hat{\pi}^{\mathsf{B}}$ are capped at $1 - \epsilon$.

On a 128-core CPU machine, using parallel processing, the entire compute time it took to produce Table 4.2 was approximately 5 minutes.

B.6.2 Power Analysis

To examine the efficiency of the DR estimator, we now analyze the power of the statistical test for $H_0 : \Delta^{\mathsf{AB}} = 0$ vs. $H_1 : \Delta^{\mathsf{AB}} \neq 0$ by inverting the DR CI. For different values of the sample size and the underlying performance gap, we compute the rejection rate of the statistical test across 1,000 runs. As before, the classifier A represents the oracle classifier that has the optimal decision boundary, which

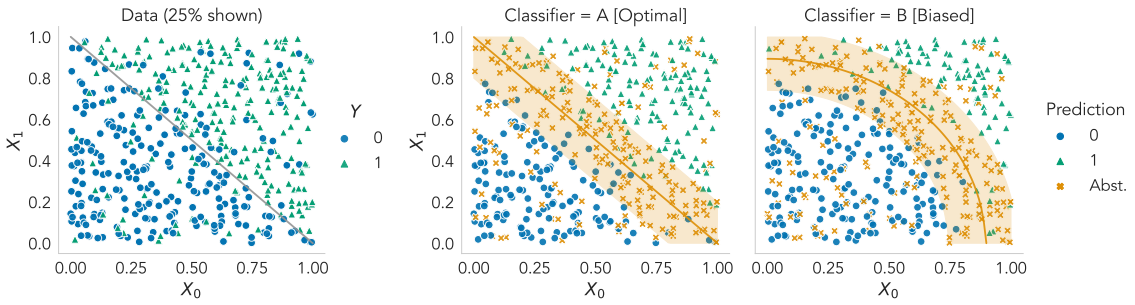


Figure B.2: A simulated example where we compare two hypothetical abstaining classifiers. The left plot shows a binary classification dataset (25% shown) in which the true decision boundary is linear. The two plots on the right show both the predictions (blue circles for 0; green triangles for 1) and the abstentions (orange x's) of two classifiers: A, which has the optimal linear boundary, and B, which has the biased nonlinear boundary. Both classifiers abstain w.p. $1 - \epsilon$ in the shaded (orange) region near the decision boundary and w.p. ϵ outside the region. For both classifiers, ϵ is set to 0.2 (positivity is satisfied). Because the abstention mechanism of either classifier is determined by the input, it is not uniformly spread out across the input domain (MAR). As a result, the difference in *selective scores*, i.e., $\mathbb{E}[S^{\mathsf{A}} \mid R^{\mathsf{A}} = 0] - \mathbb{E}[S^{\mathsf{B}} \mid R^{\mathsf{B}} = 0] \approx 0.044$, is substantially smaller than the difference in the *counterfactual scores*, i.e., $\Delta^{\mathsf{AB}} = \mathbb{E}[S^{\mathsf{A}} - S^{\mathsf{B}}] \approx 0.116$. Our 95% DR CI for Δ^{AB} the yields $(0.077, 0.145)$, using $n = 2,000$.

is linear, but the classifier B now uses a linear decision boundary that is shifted from the optimal one by a fixed amount, thereby shifting Δ^{AB} away from zero. As such, B performs increasingly worse as Δ^{AB} increases.

To increasingly vary the counterfactual score difference between two classifiers, we set A as the same classifier as in §B.6.1 and set B to use the (optimal) linear decision boundary of A shifted diagonally by a fixed amount μ. Specifically, $f^{\mathsf{B}}(x_0, x_1) = \sigma(x_0 + x_1 - (1 + \mu))$. An example with $\mu = 0.2$ is shown in Figure B.3. While Δ^{AB} is not strictly a linear function of μ, it is gradually increasing as μ increases, as shown in Table B.1. Aside from this difference, both classifiers use the same abstention mechanism as classifier A from the previous experiment, and the data generating process is also identical to the previous experiment.

Figure B.4 plots the rejection rates of the level-α statistical test, for $\alpha = 0.05$, against different values of Δ^{AB} (0 to 0.27) for various sample sizes ($n = 400, 800, 1600, 3200$). Here, we plot the miscoverage rate as a function of the resulting values of Δ^{AB} directly. We use the super learner to learn the nuisance functions. Overall, we see that as n or Δ^{AB} increases, the power of the statistical test quickly approaches 1, implying that the test can consistently detect a gap in counterfactual scores if either the sample size or the difference gets large.

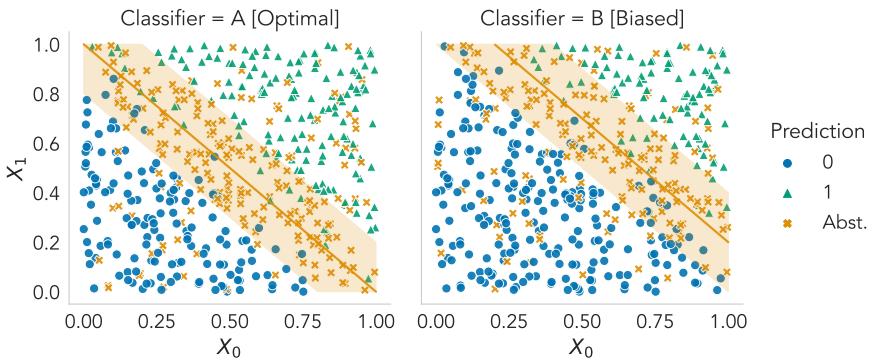


Figure B.3: A simulated example for the power experiment in which $\Delta^{\mathsf{AB}} = 0.123$. The evaluation data is the same as the one in Figure B.2. For B, the decision boundary of A is shifted diagonally upwards by $\mu = 0.2$; in the power experiment, we experiment with various values of μ (and thus Δ^{AB}).

Δ^{AB}	0.0	0.045	0.069	0.088	0.123	0.152	0.180	0.181	0.219	0.248	0.271
μ	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50

Table B.1: The relationship between Δ^{AB} and μ, the distance between the linear decision boundaries of A and B, in the power experiment of Section B.6.2.

On a 128-core CPU machine, using parallel processing, the entire compute time it took to produce Table B.4 was approximately 88 minutes.

B.6.3 Details on the CIFAR-100 Experiment

The abstaining classifiers compared in the experiments are variants of the VGG-16 CNN model with batch normalization (Simonyan and Zisserman, 2015). Specifically, the feature representation layers are obtained from a model[3] trained on the training set of the CIFAR-100 dataset and are fixed during evaluation. Using half ($n = 5,000$) of the validation set, we train a L2-regularized softmax output layer and its softmax response (SR) for the abstention mechanism. The comparison is done on the other half ($n = 5,000$) of the validation set. This version of the VGG-16 features and the softmax layer is used for all scenarios, with different abstention mechanisms described in the main text, except for the last comparison, where we compare this softmax layer with VGG-16's original 3-layer output model (2 hidden layers of size 512).

[3] https://github.com/chenyaofo/pytorch-cifar-models

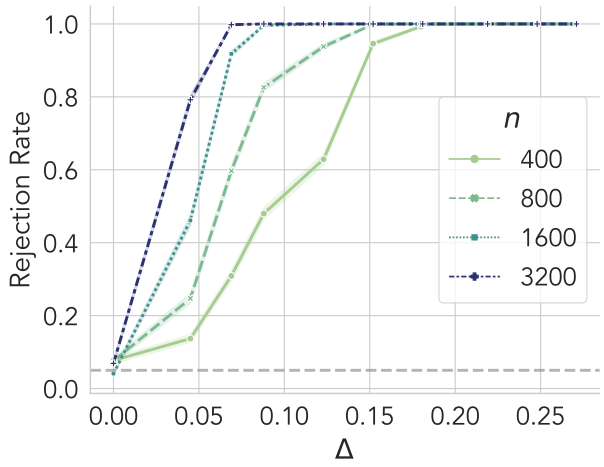


Figure B.4: Power of the statistical test for $H_0 : \Delta^{AB} = 0$ derived by our 95% DR CIs, plotted for different values of n (sample size) and Δ^{AB}, which varies based on the distance between the (linear) decision boundaries of A and B. Mean rejection rates of H_0 over 1,000 simulations are shown, with 1 standard error as shaded error bars. As either n or Δ^{AB} grows large, the power approaches 1.

The nuisance functions, $\hat{\pi}$ and $\hat{\mu}_0$ for each classifier in each scenario, also utilize the pre-trained representations of the VGG-16 layer, but their output layers (both L2-regularized linear models) are trained separately via cross-fitting.

The pre-trained VGG-16 features on the CIFAR-100 validation set were first obtained using a single NVIDIA A100 GPU, taking approximately 20 seconds. On a 128-core CPU machine, using parallel processing, the rest of the computation to produce Table 4.3 took less than 10 seconds (note that there are no repeated runs in this experiment).

B.6.4 Sensitivity to Different Positivity Levels

Here, we examine how the DR estimator is affected by the level of positivity, i.e., ϵ in (4.2). As discussed in the main chapter, positivity violations make it infeasible to properly identify and estimate causal estimands. In practice, we expect the DR estimator to remain valid up until ϵ becomes smaller than a certain (small) number. To validate this, we use the same setting from our first experiment (Section 4.4.1; Appendix B.6.1) but vary the level of positivity from $\epsilon = 0.5$ (MCAR) to $\epsilon = 0.1$ (positivity near-violation).

Figure B.5 plots the miscoverage rate of the DR estimator, averaged over 1,000 repeated simulations, using the three nuisance learner choices we used in Section 4.4.1. The result confirms that

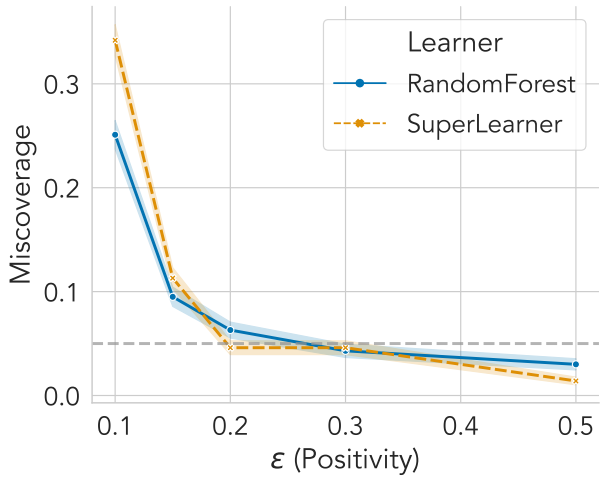


Figure B.5: Miscoverage rates of 95% doubly robust CIs by varying the level of ϵ (positivity), plotted for different nuisance function learners. Each point is the mean over 1,000 repeated simulations; shaded error bars represent 1 standard error.

the DR estimator, when using either the random forest or the super learner, retains validity as long as $\epsilon \geq 0.2$, in this particular case; as ϵ shrinks to below 0.2, the miscoverage rates start to go above the significance level. This confirms that there is a (problem-dependent) level of positivity we must expect for the DR estimator to work; otherwise, we do not expect the counterfactual target to be a meaningfully identifiable quantity in the first place.

On a 128-core CPU machine, using parallel processing, the entire compute time it took to produce Figure B.5 was approximately 12 minutes.

Bibliography

Abernethy, J. D. and Frongillo, R. M. (2012). A characterization of scoring rules for linear properties. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *PMLR*, pages 27.1–27.13. 3.2, 4

Agarwal, S. and Deshpande, A. (2022). On the power of randomization in fair classification and representation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1542–1551. 4.2.2

Arnold, S., Henzi, A., and Ziegel, J. F. (2021). Sequentially valid tests for forecast calibration. *arXiv preprint arXiv:2109.11761*. 3.6, A.5, A.5

Balsubramani, A. and Ramdas, A. (2016). Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 42–51. 2

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973. 4.1, 4.3.1

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org. `http://www.fairmlbook.org`. 4.2.2

Bauer, H. (2001). *Measure and Integration Theory*. De Gruyter. 4

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650. 4.3.1

Bickel, P. J. (1975). One-step Huber estimates in the linear model. *Journal of the American Statistical Association*, 70(350):428–434. B.2.3

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore. 4.1

Breiman, L. (1996). Stacked regressions. *Machine Learning*, 24:49–64. 4.3.1

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32. 4.3.1

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3. 3.2, 3.3.2, 3.3.2, 4.2

Busby, L. P., Courtier, J. L., and Glastonbury, C. M. (2018). Bias in radiology: The how and why of misses and misinterpretations. *Radiographics*, 38(1):236–247. B.2

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press. 2

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68. 4.1, 4.3.1

Choe, Y. J., Balakrishnan, S., Singh, A., Vettel, J. M., and Verstynen, T. (2018). Local white matter architecture defines functional brain dynamics. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 595–602. IEEE. 1.3

Choe, Y. J., Gangrade, A., and Ramdas, A. (2023). Counterfactually comparing abstaining classifiers. *arXiv preprint arXiv:2305.10564*. 1.3, 4

Choe, Y. J. and Ramdas, A. (2021). Comparing sequential forecasters. *arXiv preprint arXiv:2110.00115*. 1.3, 3

Chow, C.-K. (1957). An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254. 4.1

Chow, C.-K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46. 4.1, 4.2.2

Condessa, F., Bioucas-Dias, J., and Kovačević, J. (2017). Performance measures for classification systems with rejection. *Pattern Recognition*, 63:437–450. 4.1, 4.2.1, B.1.3

Cover, T. M. (1974). Universal gambling schemes and the complexity measures of kolmogorov and chaitin. *Technical Report, no. 12*. 2

Cover, T. M. (1991). Universal portfolios. *Mathematical Finance*, 1(1):1–29. 2

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199. 4.5

Darling, D. A. and Robbins, H. (1967). Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68. 2, 2.3, 3.1, 3.2, 3.4, 3.4.3, B.5

Darling, D. A. and Robbins, H. (1968). Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences*, 61(3):804–809. 2

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290. 3.2, 3.4.1

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72(2):169–183. 3.2

Dawid, A. P. and Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli*, pages 125–162. 2, 2

de Finetti, B. (1970). *Theory of Probability: A Critical Introductory Treatment*. New York: John Wiley. 2

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22. 3.2, 3.4

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3). 1.2, 3.2, 3.5.1, 3.6, A.8.2

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923. 1.2

Ding, P. and Li, F. (2018). Causal inference: A missing data perspective. *Statistical Science*, 33(2). 4.1

Dunsmore, I. (1968). A Bayesian approach to calibration. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2):396–405. A.7

Durrett, R. (2019). *Probability: Theory and examples*, volume 49. Cambridge University Press. 2.2, A.5

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles: Consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 505–562. 3.2

Ehm, W. and Krüger, F. (2018). Forecast dominance testing via sign randomization. *Electronic Journal of Statistics*, 12(2):3758–3793. 3.2, 3, A.8.2

El-Yaniv, R. and Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5). 4.1, 4.1

Fan, X., Grama, I., and Liu, Q. (2015). Exponential inequalities for martingales with applications. *Electronic Journal of Probability*, 20:1–22. A.1.1

Fisher, A. and Kennedy, E. H. (2021). Visually communicating and teaching intuition for influence functions. *The American Statistician*, 75(2):162–172. 4.3.1

Frongillo, R. M. and Kash, I. A. (2021). General truthfulness characterizations via convex analysis. *Games and Economic Behavior*, 130:636–662. 3.2

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30. 4.4.2

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6):1545–1578. 1.2, 3.2, 3.4.2, 3.5.1, A.8.2

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762. 3.2, 3.4, A.7, A.8.1, A.8.1

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268. 3.3.2

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378. 3.2, 3.3.2, 3.4, A.4, A.7

Good, I. (1971). Comment on "Measuring information and uncertainty" by Robert J. Buehler. *Foundations of Statistical Inference*, pages 337–339. 3.3.2

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(1):107–114. 3.2, 3.3.2

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2017). On fairness, diversity and randomness in algorithmic decision making. *arXiv preprint arXiv:1706.10208*. 4.2.2

Grünwald, P., de Heide, R., and Koolen, W. (2019). Safe testing. *arXiv preprint arXiv:1906.07801*. 2, 2.2, 2.3, 3.2, 3.4.4, 3.4.4, A.8.1, B.5

Grünwald, P. D. and Dawid, A. P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *the Annals of Statistics*, 32(4):1367–1433. 3.2

Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., and Davis, J. (2021). Machine learning with a reject option: A survey. *arXiv preprint arXiv:2107.11277*. 4.1

Henzi, A. and Ziegel, J. F. (2022). Valid sequential inference on probability forecast performance. *Biometrika*, 109(3):647–663. 3.2, 3.4.2, 3.4.4, 3.4.4, 3.5.3, 3.5, 3.5.3, 3.6, A.4, A.5, A.5, A.5, A.6.1, A.8.1, A.8.2, A.9.3

Henzi, A., Ziegel, J. F., and Gneiting, T. (2021). Isotonic distributional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(5):963–993. 3.5.3, A.9.3

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30. 2.2, 3.4.3, 3.4.3

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685. 4.3.1, B.1.4

Howard, S. R. and Ramdas, A. (2022). Sequential estimation of quantiles with applications to a/b testing and best-arm identification. *Bernoulli*, 28(3):1704–1728. 3.2

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2020). Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317. 2, 3.2, 3.2, 3.4.3, 3.4.3, 3.4.3, 3.4.3, A.1.1, A.2.2

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055 – 1080. 1.2, 1.1, 1.2, 2, 2.2, 2.2, 2.3, 2.3, 2.3, 3.1, 3.1, 3.2, 3.4, 3.4.3, 3.4.3, 3.4.3, 3.4.3, 3.4.3, 3.4.4, 3.4.4, 3.4.4, 3.5.1, A.2.1,

A.2.2, A.2.2, A.8.1, A.9.4, A.9.4, B.5

Jamieson, K. and Jain, L. (2018). A bandit approach to multiple testing with false discovery control. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3664–3674. 3.2

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR. 3.2

Johari, R., Koomen, P., Pekelis, L., and Walsh, D. (2022). Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, 70(3):1806–1821. 2, 2.2, 3.1, 3.2, 3.4.4, B.5

Ju, C., Schwab, J., and van der Laan, M. J. (2019). On adaptive propensity score truncation in causal inference. *Statistical Methods in Medical Research*, 28(6):1741–1760. 4.5

Jun, K.-S. and Orabona, F. (2019). Parameter-free online convex optimization with sub-exponential noise. In *Conference on Learning Theory*, pages 1802–1823. PMLR. 2

Kalai, A. and Kanade, V. (2021). Towards optimally abstaining from prediction with ood test examples. *Advances in Neural Information Processing Systems*, 34:12774–12785. 4.2.2, B.2

Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, 35(4):917–926. 2

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*. 4.1, 4.3, 4.3.1, B.2.3, B.2.3

Krichevsky, R. and Trofimov, V. (1981). The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207. 2

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*. 4.4.2

Lai, T. L. (1976a). Boundary crossing probabilities for sample sums and confidence sequences. *The Annals of Probability*, 4(2):299–312. 3.2

Lai, T. L. (1976b). On confidence sequences. *The Annals of Statistics*, 4(2):265–280. 3.1

Lai, T. L., Gross, S. T., and Shen, D. B. (2011). Evaluating probability forecasts. *The Annals of Statistics*, 39(5):2356–2382. 1.2, 3.2, 3.2, 3.4, 3.4.2, 3.4.2, 3.4.2, 3.5.1, 3.5.2, A.4, A.8.2, A.2, A.8.2, A.8.2

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. 4.3.1

Léger, M., Chatton, A., Le Borgne, F., Pirracchio, R., Lasocki, S., and Foucher, Y. (2022). Causal inference in case of near-violation of positivity: Comparison of methods. *Biometrical Journal*, 64(8):1389–1403. 4.5

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks.* Holden-day. 3

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons. 4.1

Madras, D., Pitassi, T., and Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems*, 31. B.2

Matheson, J. E. and Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, 22(10):1087–1096. A.7

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655. 3.2

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157. 1.2

Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A. (2014). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, 142(8):3003–3014. 3.5.3, A.9.3

Mohan, K., Pearl, J., and Tian, J. (2013). Graphical models for inference with missing data. *Advances in Neural Information Processing Systems*, 26. B.3

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529):73–119. A.9.3

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12):2417–2424. A.4

NHTSA (2022). INOA-EA22002-3184. Technical report, National Highway Traffic Safety Administration, U.S. Department of Transportation. Report on opening of an engineering analysis regarding

Tesla, Inc. products by the Office of Defects Investigation. Accessed on 26th Jan, 2023. B.4

Orabona, F. and Pál, D. (2016). Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29. 2

Ovcharov, E. Y. (2018). Proper scoring rules and Bregman divergence. *Bernoulli*, 24(1):53–79. 3.2

Pearl, J. (2000). Models, reasoning and inference. *Cambridge University Press*, 19(2). 4.1, B.3

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and Van Der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21(1):31–54. 4.5

Podkopaev, A., Blöbaum, P., Kasiviswanathan, S. P., and Ramdas, A. (2023). Sequential kernelized independence testing. *International Conference on Machine Learning*. 2.2

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., and Mullainathan, S. (2019). The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*. B.2

Rakhlin, A. and Sridharan, K. (2017). On equivalence of martingale tail bounds and deterministic regret inequalities. In *Conference on Learning Theory*, pages 1704–1722. PMLR. 2

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2022a). Game-theoretic statistics and safe anytime-valid inference. *arXiv preprint arXiv:2210.01948*. 2, 2, 2.2, 3.4.1, B.5

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*. 2, 2.2, 3.4.4, 3.4.4, 3.4.4, 3.4.4, A.1.3, A.5, A.5, A.5, A.6.1, 3

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2022b). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109. 2, 2.2, 2.2, 2.2, 2.2, 6, 2.2, 3.1, 3.2, 3.2, 3.4.4, 3.4.4, 3.4.4, A.5, A.5

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409. 2, 3.2, 3.4.3, 3.4.3, 3.4.3, 3.4.3, A.8.1

Robbins, H. and Siegmund, D. (1970). Boundary crossing probabilities for the Wiener process and sample sums. *The Annals of Mathematical Statistics*, 41(5):1410–1429. 2, 3.2

Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, 2:335–421. 4.3.1

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866. 4.1, 4.1, 4.3.1

Rosenbaum, P. R. (1995). *Observational studies*. Springer. 3, 4.3.1, 4.4.1, B.1.4

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688. 1.2, 4.1, 4.2, B.1.2

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592. 4.1, 4.1, 4.2, 4.5, B.1.2

Ruf, J., Larsson, M., Koolen, W. M., and Ramdas, A. (2022). A composite generalization of ville's martingale theorem. *arXiv preprint arXiv:2203.04485*. 2, 2

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801. 3.2

Schervish, M. J. (1989). A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856 – 1879. 3.2, 3.4

Schreuder, N. and Chzhen, E. (2021). Classification with abstention but without disparities. In *Uncertainty in Artificial Intelligence*, pages 1227–1236. PMLR. 4.2.2

Seillier-Moiseiwitsch, F. and Dawid, A. (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*, 88(421):355–359. 3.4.2

Shaer, S., Maman, G., and Romano, Y. (2023). Model-X sequential testing for conditional independence via testing by betting. In *International Conference on Artificial Intelligence and Statistics*, pages 2054–2086. PMLR. 2.2

Shafer, G. (2021). 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. 2, 2.1, 3.2, 3.4.1

Shafer, G., Shen, A., Vereshchagin, N., and Vovk, V. (2011). Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101. 2, A.5, A.5

Shafer, G. and Vovk, V. (2005). *Probability and finance: It's only a game!*, volume 491. Wiley. 2, 2

Shafer, G. and Vovk, V. (2019). *Game-theoretic foundations for probability and finance*, volume 455. Wiley. 2, 2, 2, 2.1, 3.1, 3.4.4

Shekhar, S. and Ramdas, A. (2021). Nonparametric two-sample testing by betting. *arXiv preprint arXiv:2112.09162*. 2.2

Shi, C., Blei, D., and Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32. 4.4.2

Shpitser, I., Mohan, K., and Pearl, J. (2015). Missing data as a causal and probabilistic problem. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 802–811. 4.1

Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*. 4.4.2, B.6.3

ter Schure, J. and Grünwald, P. (2022). ALL-IN meta-analysis: Breathing life into living systematic reviews. *F1000Research*, 11. 2.2

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288. 4.3.1

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer. 4.1

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1). 4.3.1

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer. 4.1

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge University Press. 4.1, 4.1

van der Vaart, A. W. (2002). Semiparametric statistics. In *Lectures on Probability Theory and Statistics (Saint-Flour, 1999)*, pages 331–457. Springer. 4.1, 4.3.1, 4.3.1, B.2.3

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., and Atencia, A. (2021). Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bulletin of the American Meteorological Society*,

102(3):E681–E699. 3.5.3

Ville, J. (1939). *Étude critique de la notion de collectif*. Gauthier-Villars. 2, 2, 2.3, 3.2, A.5

Vogelpohl, T., Kühn, M., Hummel, T., and Vollrath, M. (2019). Asleep at the automated wheel—sleepiness and fatigue during highly automated driving. *Accident Analysis & Prevention*, 126:70–84. 1

Vovk, V., Takemura, A., and Shafer, G. (2005). Defensive forecasting. In *International Workshop on Artificial Intelligence and Statistics*, pages 365–372. PMLR. 3.5.1

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754. 2, 2.2, 2.3, 3.2, 3.4.4, 3.4.4, A.5, A.5, A.5

Waggoner, B. (2021). Linear functions to the extended reals. *arXiv preprint arXiv:2102.09552*. 3.2

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*. 2

Wang, R. and Ramdas, A. (2022). False Discovery Rate Control with E-values. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 84(3):822–852. 2.2

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890. 2

Waudby-Smith, I., Arbour, D., Sinha, R., Kennedy, E. H., and Ramdas, A. (2021). Time-uniform central limit theory, asymptotic confidence sequences, and anytime-valid causal inference. *arXiv preprint arXiv:2103.06476*. 1.1, 1.2, 3.5.1, 3.5.1, A.3, A.3, A.3, A.3, A.9.4, B.5, B.5, B.5

Waudby-Smith, I. and Ramdas, A. (2023). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1.1, 1.2, 2, 2.1, 3.2, A.8.1, A.9.4, A.9.4

Waudby-Smith, I., Wu, L., Ramdas, A., Karampatziakis, N., and Mineiro, P. (2022). Anytime-valid off-policy inference for contextual bandits. *arXiv preprint arXiv:2210.10768*. A.1.1

Winkler, R. L. (1977). Rewarding expertise in probability assessment. In *Decision Making and Change in Human Affairs*, pages 127–140. Springer. A.4

Winkler, R. L. (1994). Evaluating probabilities: Asymmetric scoring rules. *Management Science*, 40(11):1395–1405. 3.4.3, 3.5.2, A.1.3, A.4, A.4, A.1, A.4, A.4

Winkler, R. L., Munoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., Lindley, D. V., Murphy, A. H., Oliver, R. M., and Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60. 3.2, 3.3.2, 3.4

Xu, Z., Wang, R., and Ramdas, A. (2021). A unified framework for bandit multiple testing. *Advances in Neural Information Processing Systems*, 34:16833–16845. 2.2

Yen, Y.-M. and Yen, T.-J. (2021). Testing forecast accuracy of expectiles and quantiles with the extremal consistent loss functions. *International Journal of Forecasting*, 37(2):733–758. 3.2

Zhang, X.-Y., Xie, G.-S., Li, X., Mei, T., and Liu, C.-L. (2023). A survey on learning to reject. *Proceedings of the IEEE*, 111(2):185–215. 4.1

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In van der Laan, M. J. and Rose, S., editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 459–474. Springer. 4.3.1

Ziegel, J. F., Krüger, F., Jordan, A., and Fasciati, F. (2020). Robust forecast evaluation of expected shortfall. *Journal of Financial Econometrics*, 18(1):95–120. 3.2