Carnegie Mellon University
Dietrich College of Humanities and Social Sciences
School of Computer Science
Dissertation

Submitted in Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy

Title: Uncertainty Quantification under Distribution Shifts

Presented by: Aleksandr Podkopaev

Accepted by: Department of Statistics, Department of Machine Learning

Readers:

_____
Aaditya Ramdas, Advisor

_____
Alessandro Rinaldo

_____
Zachary Chase Lipton

_____
Rina Foygel Barber (University of Chicago)

_____
Shiva Kasiviswanathan (Amazon)

Approved by the Committee on Graduate Degrees:

_____
Richard Scheines, Dean Date

CARNEGIE MELLON UNIVERSITY

Uncertainty Quantification under Distribution Shifts

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

DOCTOR OF PHILOSOPHY
IN
STATISTICS AND MACHINE LEARNING
BY

ALEKSANDR PODKOPAEV
DEPARTMENT OF STATISTICS
DEPARTMENT OF MACHINE LEARNING
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PA 15213

Carnegie Mellon University

JUNE 2023

Acknowledgements

Almost five years ago, I began a transformative journey as a PhD student at Carnegie Mellon University. Reflecting upon that period of my life, I recognize how lucky I have been to join a great department where a sense of community prevails: I have been fortunate to be surrounded by exceptional individuals who left an incredible mark on my professional and personal maturation, making my overall PhD experience truly remarkable.

The words of my deepest appreciation go to the faculty and staff at Carnegie Mellon University who are committed to fostering a unique environment where young researchers, like me, can only thrive. In addition to providing excellent background knowledge for conducting research, they were always available and willing to help in any possible way. I want to thank Rob Kass for helping me with the first research step during my first two years in the department.

It is a very challenging task to choose the right words to convey my gratitude to my advisor, Aaditya Ramdas. His dedicated mentorship, thorough expertise, and boundless enthusiasm had a significant impact on shaping me as a researcher. I am grateful for his kindness and patience while I was learning the invaluable skill of formalizing important research questions and tackling challenging problems. However, his guidance over the years extends far beyond pure academic pursuits. I will cherish in memory our group meetings and occasional hangouts.

I am very fortunate to have remarkable people as my thesis committee members: Alessandro Rinaldo, Zachary Chase Lipton, Rina Foygel Barber, and Shiva Kasiviswanathan, each of whom represents an excellent research role model. The breadth and depth of their knowledge across various disciplines served as an exceptional source of inspiration. Their insightful questions helped me to refine my ideas and communicate my thoughts more efficiently.

My fellow students and friends — Tudor, Ian, Alec, YJ, Nick, Matteo, Mike, Nil-Jana, Mikaela, Jinjin, Tim, Kayla, Anni, Maya, Neil, Luca, Chirag, Niccolò, and many others — are very talented researchers and just extraordinary people who were always there to offer support with any matters. Our interaction helped me go through the most challenging times, and I am very happy that our life paths crossed.

I dedicate this thesis to my family, whose permanent support has formed the basis of this journey. Regardless of the audacity of my ideas, they have always stood by my side, sharing pure love and encouragement. Their lasting faith in my abilities is what made this PhD journey possible. The highest words of appreciation go to my partner, Kimberly, whose presence in my life simply makes every day more colorful: having you along my side helps me navigate through the most uncertain periods of life.

Abstract

In the realm of scientific discoveries and practical applications, reliable application of statistical methodologies necessitates a thorough examination of potential failure scenarios. One major concern is related to the robustness of deployed methods to changes in data distribution. A classical assumption that collected data consists of observations drawn independently from the same unknown distribution (referred to as the i.i.d. assumption) is frequently violated in real-world scenarios. Therefore, it becomes essential to design statistical methods that are either inherently robust to or capable of effectively handling violations of conventional assumptions.

The first part of this thesis is devoted to topics in sequential testing — a complementary approach to traditional batch testing. Unlike batch testing where the sample size is specified before collecting data, sequential tests process data online and update inference on the fly. Specifically, we consider two closely related problems of sequential nonparametric two-sample and independence testing, which have extensive applications in various sub-fields of machine learning and statistics, often involving high-dimensional observation spaces, such as images or text. One major drawback of batch nonparametric two-sample and independence tests is that in general composite nonparametric settings, even if the null hypothesis is false, it is not possible to determine beforehand collecting how much data is sufficient to reject the null. If an analyst strongly believes that the null is false but specified sample size that was too small, then nothing can rescue the situation as the error budget is fully utilized. Conversely, excessive data collection followed by batch testing, is highly sub-optimal from several standpoints, including memory and computation usage. To address these limitations, we develop consistent sequential tests for both problems and justify their excellent empirical performance.

In addition, we consider the problem of detecting harmful distribution shifts. In practical settings, the assumption that the test data, observed during model deployment, are independent of and identically distributed as the data used for training is often violated. Therefore, it is essential to augment a learned model with a set of tools that raise alerts whenever critical changes occur. Naive testing for the presence of distribution shifts is not fully practical as it fails to account for the *malignancy* of a shift. Raising unnecessary alarms in benign scenarios can lead to delays and a substantial increase in deployment costs. In this work, we define a *harmful* shift as the one characterized by a significant drop in model performance according to pre-defined metrics and develop sequential tests to detect the presence of such harmful distribution shifts.

The second part of this thesis is devoted to topics in predictive uncertainty quantification. For a test point, classification models usually output a set of scores between zero and one, and a natural intention is to interpret those in a frequentist way (as probabilities of belonging to each of the classes). However, without additional (strong) assumptions, such interpretation fails to hold true. The discrepancy between the forecasts and long-run label frequencies is called model *miscalibration*. As an alternative way of communicating uncertainty, set-valued prediction returns a set of labels for classification or an interval/collection of intervals for regression problems. Amongst various tools for performing set-valued prediction, conformal prediction has become popular due to its reliable reflection of uncertainty under minimal assumptions.

One problem that is being considered is that of distribution-free posthoc recalibration in the context of binary classification. We establish a connection between calibration and alternative methods for quantifying predictive uncertainty and use it to derive an impossibility result for distribution-free recalibration via popular scaling-based recalibration methods. In the separate project, we consider assumption-light ways of quantifying predictive uncertainty in the presence of label shift when at the deployment stage class label proportions change (common in medical settings). We analyze strategies for handling label shift without labeled data from the target domain.

Contents

List of Figures

Chapter 1

Introduction

Reliable application of statistical methodologies in practice requires a careful analysis of potential failure scenarios. One major concern is the robustness of deployed tools in the presence of changes in data distribution. A classical assumption that a collected dataset consists of observations drawn independently from the same unknown distribution (referred to as the i.i.d. assumption) is often violated in real-world scenarios. Therefore, it becomes essential to design statistical methods that are inherently robust to or are capable of effectively handling violations of conventional assumptions.

In predictive settings, the i.i.d. assumption allows making rigorous theoretical claims about the expected performance of trained predictive models on previously unseen data. However, in real-world scenarios, a deployed machine learning model often makes predictions on data sampled from the *target* distribution (denoted as Q) which differs from the *source* distribution (denoted as P) that generated the training data. Moreover, the distribution of test data may also drift over time. This phenomenon is referred to as *dataset shift*; see the book by Quionero-Candela et al. (2009). Two commonly studied types of dataset shift include *covariate shift* (Shimodaira, 2000), where $Q(X) \neq P(X)$ but $Q(Y \mid X) = P(Y \mid X)$, and *label shift* (Saerens et al., 2002), where $Q(Y) \neq P(Y)$ but $Q(X \mid Y) = P(X \mid Y)$.

While posing structural assumptions about the nature of a present shift allows reasoning about the expected behavior on the target domain, even with access to only unlabeled data, relying on covariate/label shift assumptions suffers from a major drawback: such (unverifiable) assumptions may often be unrealistic in practice, and distribution shifts that occur in practice are generally more complex. While the label shift assumption may be sensible in medical diagnosis, the prevalence of certain diseases in the population: $P(Y)$, and the corresponding symptoms: $P(X|Y = y)$, might both change over time (during epidemics or due to potential mutations), thus violating the underlying assumptions.

In experimental settings, especially when observations are collected over time, the i.i.d. assumption can frequently be violated. For example, evolving user behavior or system dynamics can introduce non-statitionarities and temporal dependencies in the data. Deploying classical statistical tests in such settings thus leads to invalid inferential claims.

In this thesis, we study violations of the i.i.d. assumption from the perspectives of detection and adaptation. Chapter 2 is devoted to topics in sequential experimentation. Unlike traditional batch tests, where the sample size is specified before collecting data, sequential tests process data online. Sequential tests are particularly well-suited for large-scale experimentation in the tech industry where the experiments may not have a fixed stopping rule. Moreover, adaptive choices frequently often made by data scientists may invalidate traditional approaches, and thus require the development of new tools with rigorous statistical guarantees.

Two specific problems we consider include nonparametric two-sample and independence testing. In the former, given observations from two distributions: P and Q, the goal is to test the null hypothesis that the distributions are the same ($H_0 : P = Q$) against the alternative that they are not ($H_1 : P \neq Q$). In the latter, given paired observations drawn from an unknown joint distribution P_{XY}, the goal is to test the null that the random variables are independent ($H_0 : P_{XY} = P_X \times P_Y$) against the alternative that they are not ($H_1 : P_{XY} \neq P_X \times P_Y$). Both problems have broad applications in many sub-fields of machine learning and statistics, often involving high-dimensional and structured observation spaces, such as images or text.

While there exists a large body of literature on batch nonparametric two-sample and independence testing, related methods suffer from a common limitation. Specifically, in general composite nonparametric settings, even if the null hypothesis is false, it is never known beforehand collecting how much data will be "enough" to reject the null. If an analyst strongly believes that the null hypothesis is false but specified the sample size which happened to be too small, then nothing can rescue the situation as the error budget: α, is fully utilized. Conversely, excessive data collection, followed by batch testing, is highly sub-optimal from several standpoints, including memory and computation usage. Our sequential two-sample and independence tests address those limitations and demonstrate excellent empirical performance.

While our new methods are interesting even under the i.i.d. setting, they also address a separate limitation of the related batch approaches. Batch two-sample and independence testing are usually conducted by computing a permutation p-value for some chosen dependence measure, with an implicit assumption that the distribution of the data does not change over time. Even under mild changes in distribution, this approach is no longer valid, and an inflated false alarm rate is often observed empirically. In contrast, our tests better handle non-stationarity in the data distribution: they remain provably valid and powerful even if data distribution drifts.

In addition, we consider the problem of detecting harmful distribution shifts. In practical settings, the assumption that the test data, observed during model deployment, are independent of and identically distributed as the data used for training is often violated. To ensure the trustworthiness of a machine learning system, it is essential to augment a learned model with a set of tools that raise alerts whenever critical changes occur. Naive testing for the presence of distribution shifts is not fully practical as it fails to account for the *malignancy* of a shift. Raising unnecessary alarms

in benign scenarios can lead to delays and a substantial increase in deployment costs. In this work, we define a *harmful* shift as the one characterized by a significant drop in model performance according to pre-defined metrics and develop sequential tests to detect the presence of such harmful distribution shifts.

Chapter 3 is devoted to topics in predictive uncertainty quantification. For a test point, common classification models output a set of scores between zero and one, and a natural intention is to interpret those in a frequentist way (as probabilities of belonging to each of the classes). However, without additional (strong) assumptions, such interpretation fails to hold true. The discrepancy between the forecasts and long-run label frequencies is called model *miscalibration*. As an alternative way of communicating uncertainty, set-valued prediction returns a set of labels for classification or an interval/collection of intervals for regression problems. Amongst various tools for performing set-valued prediction, conformal prediction has recently become popular due to its reliable reflection of uncertainty under minimal assumptions.

One problem that is being considered is that of distribution-free posthoc recalibration in the context of binary classification. We establish a connection between calibration and alternative methods for quantifying predictive uncertainty and use it to derive an impossibility result for distribution-free recalibration via popular scaling-based recalibration methods. In the separate project, we consider assumption-light ways of quantifying predictive uncertainty in the presence of label shift when at the deployment stage class label proportions change (common in medical settings). We analyze strategies for handling label shift without access to labeled data from the target domain.

Contributions. In this thesis, we develop several new statistical methods that rigorously quantify uncertainty in the presence of distribution shifts and justify their practical relevance across a wide range of synthetic and real settings. The rest of this documents is organized as follows:

- Part I is devoted to topics in sequential experimentation. In Chapter 2, we analyze kernelized approaches for sequential nonparametric independence testing. In Chapter 3, we analyze predictive approaches for sequential nonparametric two-sample and independence testing. In Chapter 4, we develop tests for harmful distribution shifts. These results are based on (Podkopaev and Ramdas, 2022; Podkopaev et al., 2023; Podkopaev and Ramdas, 2023).

- Part II is devoted to topics in predictive uncertainty quantification. In Chapter 5, we consider distribution-free posthoc recalibration in the context of binary classification. In Chapter 6, we analyze predictive uncertainty quantification under label shift. These results are based on (Gupta et al., 2020; Podkopaev and Ramdas, 2021).

Part I

Safe, Anytime-Valid Inference

Chapter 2

Sequential Kernelized Independence Testing

2.1 Introduction

Independence testing is a fundamental statistical problem that has also been studied within information theory and machine learning. Given paired observations (X, Y) sampled from some (unknown) joint distribution P_{XY}, the goal is to test the null hypothesis that X and Y are independent. The literature on independence testing is vast as there is no unique way to measure dependence, and different measures give rise to different tests. Traditional measures of dependence, such as Pearson's r, Spearman's ρ, and Kendall's τ, are limited to the case of univariate random variables. Kernel tests (Jordan and Bach, 2001; Gretton et al., 2005c,a) are amongst the most prominent modern tools for nonparametric independence testing that work for general $\mathcal{X}, \mathcal{Y}$ spaces.

In the literature, heavy emphasis has been placed on *batch* testing when one has access to a sample whose size is specified in advance. However, even if random variables are dependent, the sample size that suffices to detect dependence is never known a priori. If the results of a test are promising yet non-conclusive (e.g., a p-value is slightly larger than a chosen significance level), one may want to collect more data and re-conduct the study. This is not allowed by traditional batch tests. We focus on sequential tests that allow peeking at observed data to decide whether to stop and reject the null or to continue collecting data.

Problem Setup. Suppose that one observes a stream of data: $(Z_t)_{t\geq 1}$, where $Z_t = (X_t, Y_t) \overset{\text{iid}}{\sim} P_{XY}$. We design sequential tests for the following pair of hypotheses:

$$H_0 : Z_t \overset{\text{iid}}{\sim} P_{XY},\ t \geq 1 \text{ and } P_{XY} = P_X \times P_Y, \tag{2.1a}$$

$$H_1 : Z_t \overset{\text{iid}}{\sim} P_{XY},\ t \geq 1 \text{ and } P_{XY} \neq P_X \times P_Y. \tag{2.1b}$$

Following the framework of "tests of power one" (Darling and Robbins, 1968), we define a level-α sequential test as a mapping $\Phi : \cup_{t=1}^{\infty}(\mathcal{X} \times \mathcal{Y})^t \to \{0, 1\}$ that satisfies

$$\mathbb{P}_{H_0}\left(\exists t \geq 1 : \Phi(Z_1, \ldots, Z_t) = 1\right) \leq \alpha.$$

As is standard, 0 stands for "do not reject the null yet" and 1 stands for "reject the null and stop". Defining the stopping time $\tau := \inf\{t \geq 1 : \Phi(Z_1, \ldots, Z_t) = 1\}$ as the first time that the test outputs 1, a sequential test must satisfy

$$\mathbb{P}_{H_0}\left(\tau < \infty\right) \leq \alpha.$$

We work in a very general composite nonparametric setting: H_0 and H_1 consist of huge classes of distributions (discrete/continuous) for which there may not be a common reference measure, making it impossible to define densities and thus ruling out likelihood-ratio based methods.

Our Contributions. Following the principle of testing by betting, we design consistent sequential nonparametric independence tests. Our bets are inspired by popular kernelized dependence measures: Hilbert-Schmidt independence criterion (HSIC) (Gretton et al., 2005a), constrained covariance criterion (COCO) (Gretton et al., 2005c), and kernelized canonical correlation (KCC) (Jordan and Bach, 2001). We provide theoretical guarantees on *time-uniform* type I error control for these tests — the type I error is controlled even if the test is continuously monitored and adaptively stopped — and further establish consistency and asymptotic rates for our sequential HSIC under the i.i.d. setting. Our tests also remain valid even if the underlying distribution changes over time. Additionally, while the initial construction of our tests requires bounded kernels, we also develop variants based on symmetry-based betting that overcome this requirement. This strategy can be readily used with a linear kernel to construct a sequential linear correlation test. We justify the practicality of our tests through a detailed empirical study.

In the remainder of this section, we mention some related work on this topic. We start by highlighting two major shortcomings of existing tests that our new tests overcome.

(i) Limitations of Corrected Batch tests and Reduction to Two-sample Testing. Batch tests (without corrections for multiple testing) have an inflated false alarm rate under continuous monitoring (see Appendix A.1.1). Naïve Bonferroni corrections restore type I error control but generally result in tests with low power. This motivates a direct design of sequential tests (not by correcting batch tests). It is tempting to reduce sequential independence testing to sequential two-sample testing, for which a powerful solution has been recently designed (Shekhar and Ramdas, 2021). This can be achieved by splitting a single data stream into two and permuting the X data in one of the streams (see Appendix A.1.2). Still, the splitting results in inefficient use of data and thus low power, compared to our new direct approach (Figure 2.1).

(ii) Time-varying Independence Testing: Beyond the i.i.d. Setting. A common practice of using a permutation p-value for batch independence testing requires (X, Y)-pairs to be i.i.d. (more generally, exchangeable). If data

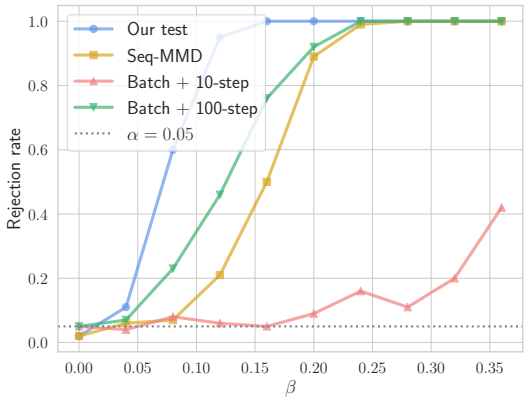


Figure 2.1: Valid sequential independence tests for: $Y_t = X_t\beta + \varepsilon_t$, $X_t, \varepsilon_t \sim \mathcal{N}(0,1)$. Batch + n-step is batch HSIC with Bonferroni correction applied every n steps (allowing monitoring only at those steps). Seq-MMD refers to the reduction to two-sample testing (Appendix A.1.2). Our test outperforms other tests.

distribution drifts, the resulting test is no longer valid, and even under mild changes, an inflated false alarm rate is observed empirically. Our tests handle more general non-stationary settings. For a stream of independent data: $(Z_t)_{t\geq 1}$, where $Z_t \sim P_{XY}^{(t)}$, consider the following pair of hypotheses:

$$H_0 : P_{XY}^{(t)} = P_X^{(t)} \times P_Y^{(t)}, \ \forall t, \tag{2.2a}$$

$$H_1 : \exists t' : P_{XY}^{(t')} \neq P_X^{(t')} \times P_Y^{(t')}. \tag{2.2b}$$

Suppose that under H_0 in (2.2a), it holds that either $P_X^{(t-1)} = P_X^{(t)}$ or $P_Y^{(t-1)} = P_Y^{(t)}$ for each $t \geq 1$, meaning that either the distribution of X may change or that of Y may change, but not both simultaneously. In this case, our tests control the type I error, whereas batch independence tests fail to.

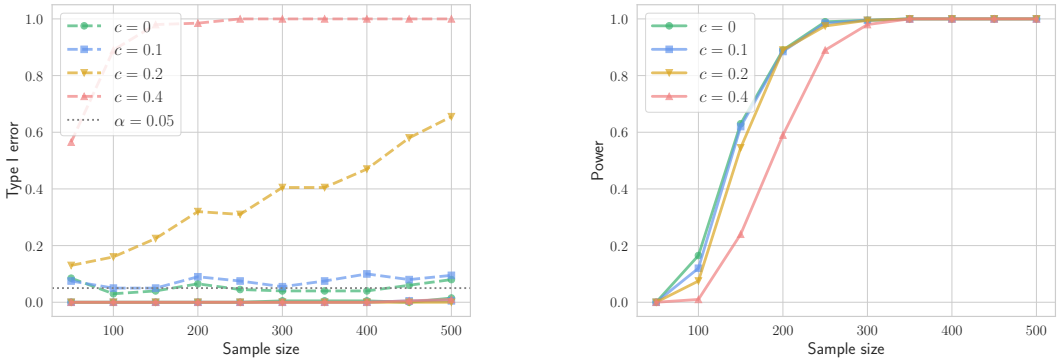
Example 1. *Let $((W_t, V_t))_{t\geq 1}$ be a sequence of i.i.d. jointly Gaussian random variables with zero mean and covariance matrix with ones on the diagonal and ρ off the diagonal. For $t = 1, 2, \ldots$ and $i \in \{0, 1\}$, consider the following stream:*

$$\begin{cases} X_{2t-i} = 2c\sin(t) + W_{2t-1}, \\ Y_{2t-i} = 3c\sin(t) + V_{2t-1}, \end{cases} \tag{2.3}$$

Setting $\rho = 0$ falls into the null case (2.2a)*, whereas any $\rho \neq 0$ implies dependence as per* (2.2b)*. Visually, it is hard to distinguish between H_0 and H_1: the drift makes data seem dependent (see Appendix A.5.1). In Figure 2.2a, we show that our test controls type I error, whereas batch test fails*[*].

Related Work. In addition to the aforementioned papers on batch independence testing, our work is also related to methods for "safe, anytime-valid inference", e.g., confidence sequences (Waudby-Smith and Ramdas, 2023, and

[*]This is also related to Yule's nonsense correlation (Yule, 1926; Ernst et al., 2017), which would not pose a problem for our method.



(a) The null is true: $W \perp\!\!\!\perp V$ in Example 1.

(b) The alternative is true with $\rho = 1/2$.

Figure 2.2: (Batch) HSIC: dashed lines, SKIT: solid lines. Under distribution drift (2.3), SKIT controls type I error under H_0 and has high power under H_1. Batch HSIC fails to control type I error under H_0 (hence we do not plot its power).

references therein) and e-processes (Grünwald et al., 2020; Ramdas et al., 2022). Sequential nonparametric two-sample tests of Balsubramani and Ramdas (2016), based on linear-time test statistics and empirical Bernstein inequality for random walks, are amongst the first results in this area. While such tests are valid in the same sense as ours, betting-based tests are much better empirically (Shekhar and Ramdas, 2021).

The roots of the principle of testing by betting can be traced back to Ville's 1939 doctoral thesis (Ville, 1939) and was recently popularized by Shafer (2021). The latter work considered it mainly in the context of parametric and simple hypotheses, far from our setting. The most closely related works to the current paper are (Shekhar and Ramdas, 2021; Shaer et al., 2023; Grünwald et al., 2023) which also handle composite and nonparametric hypotheses. Shekhar and Ramdas (2021) use testing by betting to design sequential nonparametric two-sample tests, including a state-of-the-art sequential kernel maximum mean discrepancy test. Two recent works by Shaer et al. (2023); Grünwald et al. (2023), developed in parallel to the current paper, extend these ideas to the setting of sequential conditional independence tests ($H_0 : X \perp\!\!\!\perp Y \mid Z$) under the model-X assumption, i.e., when the distribution $X \mid Z$ is assumed to be known. Our methods are very different from the aforementioned papers because when $Z = \emptyset$, the model-X assumption reduces to assuming P_X is known, which we of course avoid. The current paper can be seen as extending the ideas from (Shekhar and Ramdas, 2021) to nonparametric independence testing.

2.2 Sequential Kernel Independence Test

We begin with a brief summary of the principle of testing by betting (Shafer, 2021; Shafer and Vovk, 2019). Suppose that one observes a sequence of random variables $(Z_t)_{t\geq 1}$, where $Z_t \in \mathcal{Z}$. A player begins with initial capital $\mathcal{K}_0 = 1$. At round t of the game, she selects a payoff function $f_t : \mathcal{Z} \to [-1, \infty)$ that satisfies $\mathbb{E}_{Z\sim P_Z}[f_t(Z) \mid \mathcal{F}_{t-1}] = 0$ for all $P_Z \in H_0$, where $\mathcal{F}_{t-1} = \sigma(Z_1, \ldots, Z_{t-1})$, and bets a fraction of her wealth $\lambda_t \mathcal{K}_{t-1}$ for an $\mathcal{F}_{t-1}$-measurable

$\lambda_t \in [0,1]$. Once Z_t is revealed, her wealth is updated as

$$\mathcal{K}_t = \mathcal{K}_{t-1}(1 + \lambda_t f_t(Z_t)). \tag{2.4}$$

A level-α sequential test is obtained using the following stopping rule: $\Phi(Z_1, \ldots, Z_t) = \mathbb{1}\{\mathcal{K}_t \geq 1/\alpha\}$, i.e., the null is rejected once the player's capital exceeds $1/\alpha$. If the null is true, imposed constraints on sequences of payoffs $(f_t)_{t\geq 1}$ and betting fractions $(\lambda_t)_{t\geq 1}$ prevent the player from making money. Formally, the wealth process $(\mathcal{K}_t)_{t\geq 0}$ is a nonnegative martingale. The validity of the resulting test then follows from Ville's inequality (Ville, 1939).

To ensure that the resulting test has power under the alternative, payoffs and betting fractions have to be chosen carefully. Inspired by sequential two-sample tests of Shekhar and Ramdas (2021), our construction relies on dependence measures: $m(P_{XY}; \mathcal{C})$, which admit a variational representation:

$$\sup_{c\in\mathcal{C}} \left[\mathbb{E}_{P_{XY}} c(X,Y) - \mathbb{E}_{P_X \times P_Y} c(X,Y)\right], \tag{2.5}$$

for some class $\mathcal{C}$ of bounded functions $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The supremum above is often achieved at some $c^* \in \mathcal{C}$, and in this case, c^* is called the "witness function". In what follows, we use sufficiently rich functional classes $\mathcal{C}$ for which the following characteristic condition holds:

$$\begin{cases} m(P_{XY}; \mathcal{C}) = 0, & \text{under } H_0, \\ m(P_{XY}; \mathcal{C}) > 0, & \text{under } H_1, \end{cases} \tag{2.6}$$

for H_0 and H_1 defined in (2.1). To proceed, we bet on *pairs* of points from P_{XY}. Swapping Y-components in a pair of points from P_{XY}: Z_{2t-1} and Z_{2t}, gives two points from $P_X \times P_Y$: $\tilde{Z}_{2t-1} = (X_{2t-1}, Y_{2t})$ and $\tilde{Z}_{2t} = (X_{2t}, Y_{2t-1})$. We consider payoffs $f(Z_{2t-1}, Z_{2t})$ of the form:

$$s \cdot \Big((c(Z_{2t-1}) + c(Z_{2t})) - (c(\tilde{Z}_{2t-1}) - c(\tilde{Z}_{2t})) \Big), \tag{2.7}$$

where the scaling factor $s > 0$ ensures that $f(z, z') \in [-1, 1]$ for any $z, z' \in \mathcal{X} \times \mathcal{Y}$. When the witness function c^* is used in the above, we denote the resulting function as the "oracle payoff" f^*. Let the oracle wealth process $(\mathcal{K}^*_t)_{t\geq 0}$ be defined by using f^* along with the betting fraction

$$\lambda^\star = \frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{\mathbb{E}\left[f^\star(Z_1, Z_2)\right] + \mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]}. \tag{2.8}$$

We have the following result regarding the above quantities, whose proof is presented in Appendix A.2.2.

Theorem 2.1. *Let $\mathcal{C}$ denote a class of functions $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ for measuring dependence as per* (2.5)*.*

1. *Under* H_0 *in* (2.1a) *and* (2.2a)*, any payoff* f *of the form* (2.7) *satisfies* $\mathbb{E}_{H_0}[f(Z_1, Z_2)] = 0$.

2. *Suppose that* $\mathcal{C}$ *satisfies* (2.6)*. Under* H_1 *in* (2.1b)*, the oracle payoff* f^* *based on the witness function* c^* *satisfies* $\mathbb{E}_{H_1}[f^\star(Z_1, Z_2)] > 0$*. Further, for* $\lambda^\star$ *defined in* (2.8)*, it holds that* $\mathbb{E}_{H_1}[\log(1 + \lambda^\star f^\star(Z_1, Z_2)] > 0$*. Hence,* $\mathcal{K}_t^\star \xrightarrow{\text{a.s.}} +\infty$*, which implies that the oracle test is consistent:* $\mathbb{P}_{H_1}(\tau^\star < \infty) = 1$*, where* $\tau^\star = \inf\{t \geq 1 : \mathcal{K}_t^\star \geq 1/\alpha\}$.

Remark 1. *While the betting fraction* (2.8) *suffices to guarantee the consistency of the corresponding test, the fastest growth rate of the wealth process is ensured by considering*

$$\lambda_{\mathrm{K}}^\star \in \operatorname*{arg\,max}_{\lambda \in (0,1)} \mathbb{E}\left[\log(1 + \lambda f^\star(Z_1, Z_2)\right].$$

Overshooting *with the betting fraction may, however, result in the wealth tending to zero almost surely.*

Example 2. *Consider a sequence* $(W_t)_{t\geq 1}$*, where*

$$W_t = \begin{cases} 1, & \text{with probability } 3/5, \\ -1, & \text{with probability } 2/5. \end{cases}$$

In this case, we have $\lambda_{\mathrm{K}}^\star = 1/5$ *and* $\mathbb{E}[\log(1 + \lambda^\star W_t)] > 0$*, implying that* $\mathcal{K}_t \xrightarrow{\text{a.s.}} +\infty$*. On the other hand, it is easy to check that for* $\tilde{\lambda} = 2\lambda_{\mathrm{K}}^\star$ *we have:* $\mathbb{E}[\log(1+\tilde{\lambda}W_t)] < 0$*. As a consequence, for the wealth process* $\mathcal{K}_t$ *corresponding to the pair* $(f^*, \tilde{\lambda})$ *it holds that* $\mathcal{K}_t \xrightarrow{\text{a.s.}} 0$.

To construct a practical test, we select an appropriate class $\mathcal{C}$ for which the condition (2.6) holds and replace the oracle $f^\star$ and $\lambda^\star$ with predictable estimates $(f_t)_{t\geq 1}$ and $(\lambda_t)_{t\geq 1}$, meaning that those are computed using data observed prior to a given round of the game. We begin with a particular dependence measure, namely HSIC (Gretton et al., 2005a), and defer extensions to other measures to Section 2.3.

HSIC-based Sequential Kernel Independence Test (SKIT). Let $\mathcal{G}$ be a separable RKHS[†] with positive-definite kernel $k(\cdot, \cdot)$ and feature map $\varphi(\cdot)$ on $\mathcal{X}$. Let $\mathcal{H}$ be a separable RKHS with positive-definite kernel $l(\cdot, \cdot)$ and feature map $\psi(\cdot)$ on $\mathcal{Y}$.

Assumption 1. *Suppose that:*

(A1) Kernels k *and* l *are nonnegative and bounded by one:* $\sup_{x\in\mathcal{X}} k(x, x) \leq 1$ *and* $\sup_{y\in\mathcal{Y}} l(y, y) \leq 1$.

(A2) The product kernel $k \otimes l : (\mathcal{X} \times \mathcal{Y})^2 \to \mathbb{R}$*, defined as* $(k \otimes l)((x, y), (x', y')) := k(x, x')l(y, y')$*, is a characteristic kernel on the joint domain.*

[†]Recall that an RKHS is a Hilbert space $\mathcal{G}$ of real-valued functions over $\mathcal{X}$, for which the evaluation functional $\delta_x : \mathcal{G} \to \mathbb{R}$, which maps $g \in \mathcal{G}$ to $g(x)$, is a continuous map, and this fact must hold for every $x \in \mathcal{X}$. Each RKHS is associated with a unique positive-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which can be viewed as a generalized inner product on $\mathcal{X}$. We refer the reader to (Muandet et al., 2017) for an extensive recent survey of kernel methods.

Assumption (A1) is used to justify that the mean embeddings introduced later are well-defined elements of RKHSs, and the particular bounds are used to simplify the constants. Assumption (A2) is a sufficient condition for the characteristic condition (2.6) to hold (Fukumizu et al., 2007b), and we use it to argue about the consistency of our test. Under mild assumptions, it can be further relaxed to characteristic property of the kernels on the respective domains (Gretton, 2015). We note that the most common kernels on $\mathbb{R}^d$: Gaussian (RBF) and Laplace, satisfy both (A1) and (A2). Define mean embeddings of the joint and marginal distributions:

$$\begin{aligned} \mu_{XY} &:= \mathbb{E}_{P_{XY}} \left[\varphi(X) \otimes \psi(Y) \right], \\ \mu_X := \mathbb{E}_{P_X} \left[\varphi(X) \right], &\quad \mu_Y := \mathbb{E}_{P_Y} \left[\psi(Y) \right]. \end{aligned} \tag{2.9}$$

The cross-covariance operator $C_{XY} : \mathcal{H} \to \mathcal{G}$ associated with the joint measure P_{XY} is defined as

$$C_{XY} := \mu_{XY} - \mu_X \otimes \mu_Y,$$

where $\otimes$ is the outer product operation. This operator generalizes the covariance matrix. *Hilbert-Schmidt independence criterion* (HSIC) is a criterion defined as Hilbert-Schmidt norm, a generalization of Frobenius norm for matrices, of the cross-covariance operator (Gretton et al., 2005a):

$$\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}) := \|C_{XY}\|^2_{\mathrm{HS}}. \tag{2.10}$$

HSIC is simply the squared kernel maximum mean discrepancy (MMD) between mean embeddings of P_{XY} and $P_X \times P_Y$ in the *product RKHS* $\mathcal{G} \otimes \mathcal{H}$ on $\mathcal{X} \times \mathcal{Y}$, defined by a product kernel $k \otimes l$. We can rewrite (2.10) as

$$\left(\sup_{\substack{g \in \mathcal{G} \otimes \mathcal{H} \\ \|g\|_{\mathcal{G} \otimes \mathcal{H}} \leq 1}} \mathbb{E}_{P_{XY}} \left[g(X, Y) \right] - \mathbb{E}_{P_X \times P_Y} \left[g(X', Y') \right] \right)^2,$$

which matches the form (2.5). The witness function for HSIC admits a closed form (see Appendix A.4):

$$g^\star = \frac{\mu_{XY} - \mu_X \otimes \mu_Y}{\|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G} \otimes \mathcal{H}}}, \tag{2.11}$$

where μ_{XY}, μ_X and μ_Y are defined in (2.9). The oracle payoff based on HSIC: $f^\star(Z_{2t-1}, Z_{2t})$, is given by

$$\frac{1}{2} \left(g^\star(Z_{2t-1}) + g^\star(Z_{2t}) - g^\star(\tilde{Z}_{2t-1}) - g^\star(\tilde{Z}_{2t}) \right), \tag{2.12}$$

which has the form (2.7) with $s = 1/2$. To construct the test, we use estimators $(f_t)_{t\geq 1}$ of the oracle payoff $f^\star$ obtained by replacing $g^\star$ in (2.12) with the plug-in estimator:

$$\widehat{g}_t = \frac{\widehat{\mu}_{XY} - \widehat{\mu}_X \otimes \widehat{\mu}_Y}{\|\widehat{\mu}_{XY} - \widehat{\mu}_X \otimes \widehat{\mu}_Y\|_{\mathcal{G}\otimes\mathcal{H}}}, \tag{2.13}$$

where $\widehat{\mu}_{XY}, \widehat{\mu}_X, \widehat{\mu}_Y$ denote the empirical mean embeddings (plug-in estimators of (2.9)) computed at round t as[‡]

$$\begin{aligned} \widehat{\mu}_{XY} &= \tfrac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i) \otimes \psi(Y_i), \\ \widehat{\mu}_X = \tfrac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i), &\quad \widehat{\mu}_Y = \tfrac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \psi(Y_i). \end{aligned} \tag{2.14}$$

Note that in (2.13) the witness function is defined as an operator. We clarify this point in Appendix A.4. To select betting fractions, we follow Cutkosky and Orabona (2018) who state the problem of choosing the optimal betting fraction for coin betting as an online optimization problem with exp-concave losses and propose a strategy based on online Newton step (ONS) (Hazan et al., 2007) as a solution. ONS betting fractions are inexpensive to compute while being supported by strong theoretical guarantees. We also consider other strategies for selecting betting fractions and defer a detailed discussion to Appendix A.3. We conclude with formal guarantees on time-uniform type I error control and consistency of HSIC-based SKIT. In fact, we establish a stronger result: we show that the wealth process grows exponentially and characterize the rate of the growth of wealth in terms of the true HSIC score. The proof is deferred to Appendix A.2.2.

Algorithm 1 Online Newton step (ONS) strategy for selecting betting fractions

Input: sequence of payoffs $(f_t(Z_{2t-1}, Z_{2t}))_{t\geq 1}$, $\lambda_1^{\mathrm{ONS}} = 0$, $a_0 = 1$.
for $t = 1, 2, \ldots$ **do**
  Observe $f_t(Z_{2t-1}, Z_{2t})$;
  Set $z_t = f_t(Z_{2t-1}, Z_{2t})/(1 - \lambda_t^{\mathrm{ONS}} f_t(Z_{2t-1}, Z_{2t}))$;
  Set $a_t = a_{t-1} + z_t^2$;
  Set $\lambda_{t+1}^{\mathrm{ONS}} := \frac{1}{2} \wedge \left(0 \vee \left(\lambda_t^{\mathrm{ONS}} - \frac{2}{2-\log 3} \cdot \frac{z_t}{a_t}\right)\right)$;

Theorem 2.2. *Suppose that Assumption 1 is satisfied. The following claims hold for HSIC-based SKIT (Algorithm 2):*

1. *Suppose that H_0 in* (2.1a) *or* (2.2a) *is true. Then SKIT ever stops with probability at most α: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

[‡]At round t, evaluating HSIC-based payoff requires a number of operations that is linear in t (see Appendix A.6.2). Thus after T steps, we have expended a total of $O(T^2)$ computation, the same as batch HSIC. However, our test threshold is $1/\alpha$, but batch HSIC requires permutations to determine the right threshold, requiring recomputing HSIC hundreds of times. Thus, our test is actually more computationally feasible than batch HSIC.

Algorithm 2 HSIC-based SKIT

Input: significance level $\alpha \in (0,1)$, data stream $(Z_i)_{i\geq 1}$, where $Z_i = (X_i, Y_i) \sim P_{XY}$, $\lambda_1^{\mathrm{ONS}} = 0$.
for $t = 1, 2, \ldots$ **do**
  Use $Z_1, \ldots, Z_{2(t-1)}$ to compute $\widehat{g}_t$ as in (2.13);
  Compute HSIC payoff $f_t(Z_{2t-1}, Z_{2t})$;
  Update the wealth process $\mathcal{K}_t$ as in (2.4);
  if $\mathcal{K}_t \geq 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Compute $\lambda_{t+1}^{\mathrm{ONS}}$ (Algorithm 1);

2. *Suppose that H_1 in* (2.1b) *is true. Then it holds that $\mathcal{K}_t \overset{\text{a.s.}}{\longrightarrow} +\infty$ and thus SKIT is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$. Further, the wealth grows exponentially, and the corresponding growth rate satisfies*

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} \overset{\text{a.s.}}{\geq} \frac{\mathbb{E}[f^\star(Z_1,Z_2)]}{4} \cdot \left(\frac{\mathbb{E}[f^\star(Z_1,Z_2)]}{\mathbb{E}[(f^\star(Z_1,Z_2))^2]} \wedge 1 \right), \tag{2.15}$$

where $f^\star$ is the oracle payoff defined in (2.12).

Since $\mathbb{E}\left[f^\star(Z_1,Z_2)\right] = \sqrt{\mathrm{HSIC}(P_{XY};\mathcal{G},\mathcal{H})}$ and $\mathbb{E}\left[(f^\star(Z_1,Z_2))^2\right] \leq 1$, Theorem 2.2 implies that:

$$\liminf_{t\to\infty} \left(\tfrac{1}{t} \log \mathcal{K}_t \right) \overset{\text{a.s.}}{\geq} \tfrac{1}{4} \cdot \mathrm{HSIC}(P_{XY};\mathcal{G},\mathcal{H}).$$

However, the lower bound (2.15) is never worse. In particular, if the variance of the oracle payoffs: $\sigma^2 = \mathbb{V}\left[f^\star(Z_1,Z_2)\right]$, is small, meaning that $\sigma^2 \leq \mathbb{E}\left[f^\star(Z_1,Z_2)\right](1-\mathbb{E}\left[f^\star(Z_1,Z_2)\right])$, we get a faster rate: $\sqrt{\mathrm{HSIC}(P_{XY};\mathcal{G},\mathcal{H})}/4$, reminiscent of an empirical Bernstein type adaptation. Up to some small constants, we show that this is the best possible exponent that adapts automatically between the low- and high-variance regimes. We do this by considering the oracle test, i.e., assuming that the oracle HSIC payoff $f^\star$ is known. Amongst the betting fractions that are constrained to lie in $[-0.5, 0.5]$, like ONS bets, the optimal growth rate is ensured by taking

$$\lambda^\star = \underset{\lambda\in[-0.5,0.5]}{\arg\max}\ \mathbb{E}\left[\log(1+\lambda f^\star(Z_1,Z_2))\right]. \tag{2.16}$$

We have the following result about the growth rate of the oracle test, whose proof is deferred to Appendix A.2.2.

Proposition 1. *The optimal log-wealth $S^\star := \mathbb{E}\left[\log(1+\lambda^\star f^\star(Z_1,Z_2))\right]$ — that can be achieved by an oracle betting scheme* (2.16) *which knows $f^\star$ from* (2.12) *and the underlying distribution — satisfies:*

$$S^\star \leq \frac{\mathbb{E}\left[f^\star(Z_1,Z_2)\right]}{2} \left(\frac{8\mathbb{E}\left[f^\star(Z_1,Z_2)\right]}{3\mathbb{E}\left[(f^\star(Z_1,Z_2))^2\right]} \wedge 1 \right). \tag{2.17}$$

Remark 2 (Minibatching)**.** *While our test processes the data stream in pairs, it is possible to use larger batches of points from P_{XY}. For a batch size is $b \geq 2$, at round t, the bet is placed on $\{(X_{b(t-1)+1}, Y_{b(t-1)+1}), \ldots, (X_{bt}, Y_{bt})\}$. In this case, the empirical mean embeddings are computed analogous to* (2.14) *but using $\{(X_i, Y_i)\}_{i \leq b(t-1)}$. We defer the details to Appendix A.4. Such payoff function satisfies the necessary conditions for the wealth process to be a nonnegative martingale, and hence, the resulting sequential test has time-uniform type I error control. The same argument as in the proof of Theorem 2.2 can be used to show that the resulting test is consistent. The main downside of minibatching is that monitoring of the test (and hence, optional stopping) is allowed only after processing every b points from P_{XY}.*

Distribution Drift. As discussed in Section 2.1, batch independence tests have an inflated false alarm rate even under mild changes in distribution. In contrast, SKIT remains valid even when the data distribution drifts over time. For a stream of independent points, we claimed that our test controls the type I error as long as only one of the marginal distributions changes at each round. In Appendix A.4, we provide an example that shows that this assumption is necessary for the validity of our tests. Our tests can also be used to test instantaneous independence between two streams. Formally, define $\mathcal{D}_t := \{(X_i, Y_i)\}_{i \leq 2t}$ and consider:

$$H_0 : \forall t,\ X_{2t-1} \perp\!\!\!\perp Y_{2t-1} \mid \mathcal{D}_{t-1} \text{ and } X_{2t} \perp\!\!\!\perp Y_{2t} \mid \mathcal{D}_{t-1}, \tag{2.18a}$$

$$H_1 : \exists t' : X_{2t'-1} \not\perp\!\!\!\perp Y_{2t'-1} \mid \mathcal{D}_{t-1} \text{ or } X_{2t'} \not\perp\!\!\!\perp Y_{2t'} \mid \mathcal{D}_{t-1}. \tag{2.18b}$$

Assumption 2. *Suppose that under H_0 in* (2.18a)*, it also holds that:*

(a) The cross-links between X and Y streams are not allowed, meaning that for all $t \geq 1$,

$$\begin{aligned} Y_t &\perp\!\!\!\perp X_{t-1} \mid Y_{t-1}, \{(X_i, Y_i)\}_{i \leq t-2}, \\ X_t &\perp\!\!\!\perp Y_{t-1} \mid X_{t-1}, \{(X_i, Y_i)\}_{i \leq t-2}. \end{aligned} \tag{2.19}$$

(b) For all $t \geq 1$, either (X_t, X_{t-1}) or (Y_t, Y_{t-1}) are exchangeable conditional on $\{(X_i, Y_i)\}_{i \leq t-2}$.

In the above, (a) relaxes the independence assumption within each pair, and (b) generalizes the assumption about allowed changes in the marginal distributions of X and Y. Under the above setting, we deduce that our test retains type-1 error control, and the proof is deferred to Appendix A.2.2.

Theorem 2.3. *Suppose that H_0 in* (2.18a) *is true. Further, assume that Assumption 2 holds. Then HSIC-based SKIT (Algorithm 2) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Chwialkowski and Gretton (2014) considered a related (at a high level) problem of testing instantaneous independence between a pair of time series. Similar to distribution drift, HSIC fails to test independence between innovations in time series since naively permuting one series destroys the underlying structure. Chwialkowski and

Gretton (2014) used a subset of permutations — rotations by circular shifting (allowed by their assumption of strict stationarity) of one series for preserving the structure — to design a p-value and used the assumption of mixing (decreasing memory of a process) to justify the asymptotic validity. The setting we consider is very different, and we make no assumptions of mixing or stationarity anywhere. Related works on independence testing for time series also include (Chwialkowski et al., 2014; Besserve et al., 2013). In the next section, we extend the methodology to other dependence measures.

2.3 Alternative Dependence Measures

Let $\mathcal{C}_1$ and $\mathcal{C}_2$ denote some classes of bounded functions $c_1 : \mathcal{X} \to \mathbb{R}$ and $c_2 : \mathcal{Y} \to \mathbb{R}$ respectively. For a class $\mathcal{C}$ of functions $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ that factorize into the product: $c(x, y) = c_1(x)c_2(y)$ for some $c_1 \in \mathcal{C}_1$ and $c_2 \in \mathcal{C}_2$, the general form of dependence measures (2.5) reduces to

$$m(P_{XY}; \mathcal{C}_1, \mathcal{C}_2) = \sup_{c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2} \operatorname{Cov}\left(c_1(X), c_2(Y)\right).$$

Next, we develop SKITs based on two dependence measures of this form: COCO and KCC. While the corresponding witness functions do not admit a closed form, efficient algorithms for computing the plug-in estimates are available.

Witness Functions for COCO. *Constrained covariance* (COCO) is a criterion for measuring dependence based on covariance between smooth functions of random variables:

$$\sup_{\substack{g,h: \\ \|g\|_{\mathcal{G}} \leq 1, \\ \|h\|_{\mathcal{H}} \leq 1}} \operatorname{Cov}\left(g(X), h(Y)\right) = \sup_{\substack{g,h: \\ \|g\|_{\mathcal{G}} \leq 1, \\ \|h\|_{\mathcal{H}} \leq 1}} \langle h, C_{XY} g \rangle_{\mathcal{H}}, \tag{2.20}$$

where the supremum is taken over the unit balls in the respective RKHSs (Gretton et al., 2005c,b). At round t, we are interested in empirical witness functions computed from $\{(X_i, Y_i)\}_{i \leq 2(t-1)}$. The key observation is that maximizing the objective function in (2.20) with the plug-in estimator of the cross-covariance operator requires considering only functions in $\mathcal{G}$ and $\mathcal{H}$ that lie in the span of the data:

$$\begin{aligned} \widehat{g}_t &= \sum_{i=1}^{2(t-1)} \alpha_i \left(\varphi(X_i) - \frac{1}{2(t-1)} \sum_{j=1}^{2(t-1)} \varphi(X_j) \right), \\ \widehat{h}_t &= \sum_{i=1}^{2(t-1)} \beta_i \left(\psi(Y_i) - \frac{1}{2(t-1)} \sum_{j=1}^{2(t-1)} \psi(Y_j) \right). \end{aligned} \tag{2.21}$$

Coefficients α and β that solve the maximization problem (2.20) define the leading eigenvector of the following generalized eigenvalue problem (see Appendix A.4):

$$\begin{pmatrix} 0 & \frac{1}{2(t-1)}\tilde{K}\tilde{L} \\ \frac{1}{2(t-1)}\tilde{L}\tilde{K} & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \gamma \begin{pmatrix} \tilde{K} & 0 \\ 0 & \tilde{L} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \tag{2.22}$$

where $\tilde{K} = HKH$, $\tilde{L} = HLH$, and $H = \mathbf{I}_{2(t-1)} - (1/(2(t-1))\mathbf{1}\mathbf{1}^\top$ is centering projection matrix. Computing the leading eigenvector for (2.22) is computationally demanding for moderately large t. A common practice is to use low-rank approximations of K and L with fast-decaying spectrum (Jordan and Bach, 2001). We present an approach based on incomplete Cholesky decomposition in Appendix A.6.1.

Witness Functions for KCC. *Kernelized canonical correlation* (KCC) relies on the regularized correlation between smooth functions of random variables:

$$\sup_{\substack{g\in\mathcal{G},\\ h\in\mathcal{H}}} \frac{\operatorname{Cov}\left(g(X), h(Y)\right)}{\sqrt{\mathbb{V}\left[g(X)\right] + \kappa_1 \left\|g\right\|_{\mathcal{G}}^2} \cdot \sqrt{\mathbb{V}\left[h(Y)\right] + \kappa_2 \left\|h\right\|_{\mathcal{H}}^2}}, \tag{2.23}$$

where regularization is necessary for obtaining meaningful estimates of correlation (Jordan and Bach, 2001; Fukumizu et al., 2007a). Witness functions for KCC have the same form as for COCO (2.21), but α and β define the leading eigenvector of a modified problem (Appendix A.4).

SKIT based on COCO or KCC. Given a pair of the witness functions $g^\star$ and $h^\star$ for COCO (or KCC) criterion, the corresponding oracle payoff: $f^\star(Z_{2t-1}, Z_{2t})$, is given by

$$\frac{1}{2}\left(g^\star(X_{2t}) - g^\star(X_{2t-1})\right)\left(h^\star(Y_{2t}) - h^\star(Y_{2t-1})\right), \tag{2.24}$$

which has the form (2.7) with $s = 1/2$. To construct the test, we rely on estimates $(f_t)_{t\geq 1}$ of the oracle payoff $f^\star$ obtained by using $\widehat{g}_t$ and $\widehat{h}_t$, defined in (2.21), in (2.24). We assume that α and β in (2.22) are normalized: $\alpha^\top \tilde{K}\alpha = 1$ and $\beta^\top \tilde{L}\beta = 1$. We conclude with a guarantee on time-uniform false alarm rate control of SKITs based on COCO (Algorithm 3), whose proof is deferred to Appendix A.2.3.

Algorithm 3 SKIT based on COCO (or KCC)

Input: significance level $\alpha \in (0,1)$, data stream $(Z_i)_{i\geq 1}$, where $Z_i = (X_i, Y_i) \sim P_{XY}$, $\lambda_1^{\mathrm{ONS}} = 0$.
for $t = 1, 2, \ldots$ **do**
  Use $Z_1, \ldots, Z_{2(t-1)}$ to compute $\widehat{g}_t$ and $\widehat{h}_t$ as in (2.21);
  Compute COCO payoff $f_t(Z_{2t-1}, Z_{2t})$;
  Update the wealth process $\mathcal{K}_t$ as in (2.4);
  if $\mathcal{K}_t \geq 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Compute $\lambda_{t+1}^{\mathrm{ONS}}$ (Algorithm 1);

Theorem 2.4. *Suppose that (A1) in Assumption 1 is satisfied. Then, under H_0 in* (2.1a) *and* (2.18a)*, COCO/KCC-based SKIT (Algorithm 3) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Remark 3. *The above result does not contain a claim regarding the consistency of the corresponding tests. If (A2) in Assumption 1 holds, the same argument as in the proof of Theorem 2.2 can be used to deduce that SKITs based on the* oracle *payoffs (with oracle witness functions $g^\star$ and $h^\star$) are consistent. In contrast to HSIC, for which the oracle witness function is closed-form and the respective plug-in estimator is amenable for the analysis, to argue about the consistency of SKITs based on COCO/KCC, it is necessary to place additional assumptions, especially since low-rank approximations of kernel matrices are involved. We note that a sufficient condition for consistency is that the payoffs are positive on average: $\liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i}) \overset{\text{a.s.}}{>} 0$.*

Synthetic Experiments. To compare SKITs based on HSIC, COCO, and KCC payoffs, we use RBF kernel with hyperparameters taken to be inversely proportional to the second moment of the underlying variables; we observed no substantial difference when such selection is data-driven (median heuristic). We consider settings where the complexity of a task is controlled through a single univariate parameter:

(a) *Gaussian model.* For for $t \geq 1$, we consider $Y_t = X_t\beta + \varepsilon_t$, where $X_t, \varepsilon_t \sim \mathcal{N}(0,1)$. We have that $\beta \neq 0$ implies nonzero linear correlation (hence dependence). We consider 20 values for β, spaced uniformly in [0,0.3], and use $\lambda_X = 1/4$ and $\lambda_Y = 1/(4(1+\beta^2))$ as kernel hyperparameters.

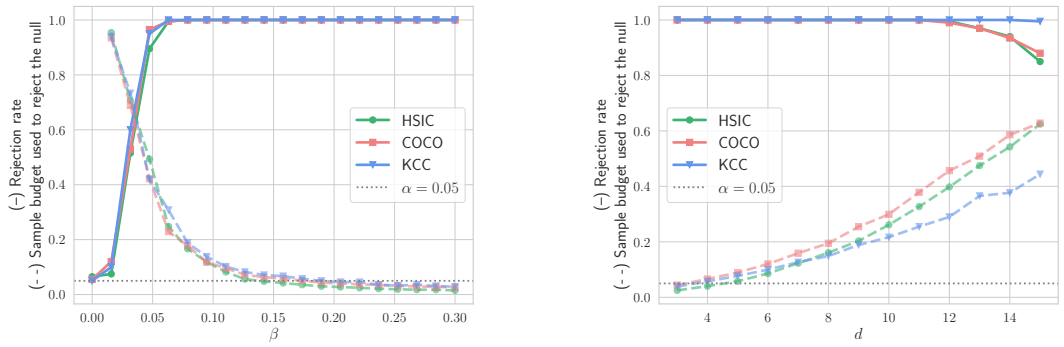
(b) *Spherical model.* We generate a sequence of dependent but linearly uncorrelated random variables by taking $(X_t, Y_t) = ((U_t)_{(1)}, (U_t)_{(2)})$, where $U_t \overset{\text{iid}}{\sim} \text{Unif}(\mathbb{S}^d)$, for $t \geq 1$. $\mathbb{S}^d$ denotes a unit sphere in $\mathbb{R}^d$ and $u_{(i)}$ is the i-th coordinate of u. We consider $d \in \{3, \ldots, 15\}$, and use $\lambda_X = \lambda_Y = d/4$ as kernel hyperparameters.

We stop monitoring after observing 20000 points from P_{XY} (if SKIT does not stop by that time, we retain the null) and aggregate the results over 200 runs for each value of β and d. In Figure 2.3, we confirm that SKITs control the type I error and adapt to the complexity of a task. In settings with a very low signal-to-noise ratio (small β or large d), SKIT's power drops, but in such cases, both sequential and batch independence tests inevitably require a lot of data to reject the null. We defer additional experiments to Appendix A.5.4.

2.4 Symmetry-based Betting Strategies

In this section, we develop a betting strategy that relies on symmetry properties, whose advantage is that it overcomes the kernel boundedness assumption that underlined the SKIT construction. For example, using this betting strategy with a linear kernel: $k(x,y) = l(x,y) = \langle x, y\rangle$ readily implies a valid sequential linear correlation test. Consider

$$W_t = \widehat{g}_t(Z_{2t-1}) + \widehat{g}_t(Z_{2t}) - \widehat{g}_t(\tilde{Z}_{2t-1}) - \widehat{g}_t(\tilde{Z}_{2t}), \tag{2.25}$$



(a) Gaussian model.

(b) Spherical model.

Figure 2.3: Rejection rate and scaled sample size used to reject the null hypothesis for synthetic data. Inspecting the rejection rate for $\beta = 0$ (independence holds) confirms that the type I error is controlled. Further, we confirm that SKITs are adaptive to the complexity (smaller β and larger d correspond to harder settings).

where $\widehat{g}_t = \widehat{\mu}_{XY} - \widehat{\mu}_X \otimes \widehat{\mu}_Y$ is the *unnormalized* plug-in witness function computed from $\{Z_i\}_{i \leq 2(t-1)}$. Symmetry-based betting strategies rely on the following key fact.

Proposition 2. *Under any distribution in H_0, W_t is symmetric around zero, conditional on $\mathcal{F}_{t-1}$.*

By construction, we expect the sign and magnitude of W_t to be positively correlated under the alternative. We consider three payoff functions that aim to exploit this fact.

1. *Composition with an odd function.* This approach is based on the idea from sequential symmetry testing (Ramdas et al., 2020) that composition with an odd function of a symmetric around zero random variable is mean-zero. Absent knowledge regarding the scale of considered random variables, it is natural to standardize $\{W_i\}_{i \geq 1}$ in a predictable way. We consider

$$f_t^{\text{odd}}(W_t) = \tanh\left(W_t / N_{t-1}\right), \tag{2.26}$$

where $N_t = Q_{0.9}(\{|W_i|\}_{i \leq t}) - Q_{0.1}(\{|W_i|\}_{i \leq t})$, and $Q_\alpha(\{|W_i|\}_{i \leq t})$ is the α-quantile of the empirical distribution of the absolute values of $\{W_i\}_{i \leq t}$. (The choices of 0.1 and 0.9 are heuristic, and can be replaced by other constants without violating the validity of the test.) The composition approach has demonstrated promising empirical performance for the betting-based two-sample tests of Shekhar and Ramdas (2021) and conditional independence tests of Shaer et al. (2023).

2. *Rank-based approach.* Inspired by sequential signed-rank test of symmetry around zero of Reynolds Jr. (1975), we consider the following payoff function:

$$f_t^{\text{rank}}(W_t) = \text{sign}(W_t) \cdot \frac{\text{rk}(|W_t|)}{t}, \tag{2.27}$$

where $\mathrm{rk}(|W_t|) = \sum_{i=1}^{t} \mathbb{1}\{|W_i| \leq |W_t|\}$.

3. *Predictive approach.* At round t, we fit a probabilistic predictor $p_t : \mathbb{R}_+ \to [0,1]$, e.g., logistic regression, using $\{|W_i|, \mathrm{sign}[W_i]\}_{i \leq t-1}$ as feature-label pairs. We consider the following payoff function:

$$f_t^{\mathrm{pred}}(W_t) = (2p_t(|W_t|) - 1)_+ \cdot (1 - 2\ell_t(W_t)), \tag{2.28}$$

where $(\cdot)_+ = \max\{\cdot, 0\}$ and $\ell_t(|W_t|, \mathrm{sign}[W_t])$ is the misclassification loss of the predictor p_t.

In the next result, whose proof is deferred to Appendix A.2.4, we show that symmetry-based SKITs are valid.

Algorithm 4 SKIT with symmetry-based betting

Input: significance level $\alpha \in (0,1)$, data stream $(Z_i)_{i \geq 1}$, where $Z_i = (X_i, Y_i) \sim P_{XY}$, $\lambda_1^{\mathrm{ONS}} = 0$.
for $t = 1, 2, \ldots$ **do**
  Observe Z_{2t-1}, Z_{2t} and compute W_t as in (2.25);
  Compute payoff $f_t^{\mathrm{odd}}(W_t)$ as in (2.26);
  Update the wealth process $\mathcal{K}_t$ as in (2.4);
  if $\mathcal{K}_t \geq 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Compute $\lambda_{t+1}^{\mathrm{ONS}}$ (Algorithm 1);

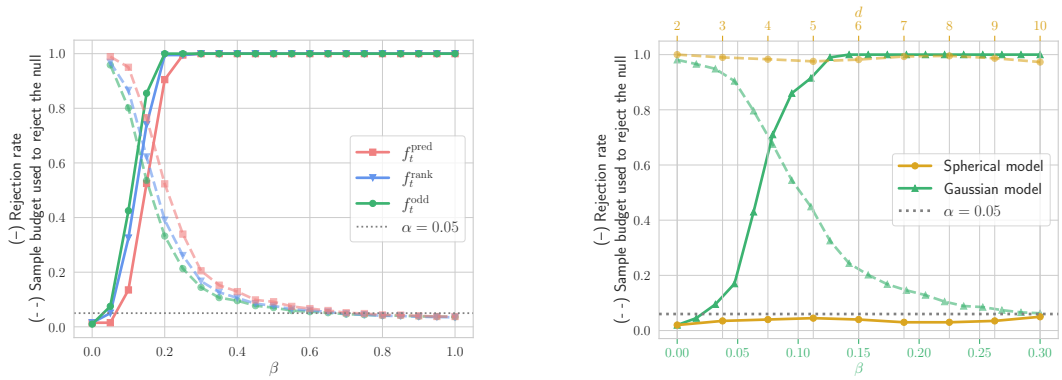
Theorem 2.5. *Under H_0 in* (2.1a) *and* (2.18a)*, the symmetry-based SKIT (Algorithm 4) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Synthetic Experiments. To compare the symmetry-based payoffs, we consider the Gaussian model along with aGRAPA betting fractions. For visualization purposes, we complete monitoring after observing 2000 points from the joint distribution. In Figure 2.4a, we observe that the resulting SKITs demonstrate similar performance. In Figure 2.4b, we demonstrate that SKIT with a linear kernel has high power under the Gaussian model, whereas its false alarm rate does not exceed α under the spherical model. Additional synthetic experiments can be found in Appendix A.5.3.

Real Data Experiment. We analyze average daily temperatures[§] in four European cities: London, Amsterdam, Zurich, and Nice, from January 1, 2017, to May 31, 2022. The processes underlying temperature formation are complex and act both on macro (e.g., solar phase) and micro (e.g., local winds) levels. While average daily temperatures in selected cities share similar cyclic patterns, one may still expect the temperature fluctuations occurring in nearby locations to be dependent. We use SKIT for testing instantaneous independence (as per (2.18)) between fluctuations (assuming that the conditions that underlie our test hold).

We run SKIT with the rank-based payoff and ONS betting fractions for each pair of cities using $6/\alpha$ as a rejection threshold (accounting for multiple testing). We select the kernel hyperparameters via the median heuristic using recordings for the first 20 days. In Figures 2.5, we illustrate that SKIT supports our conjecture that temperature fluctuations are dependent in nearby locations. We also run this experiment for four cities in South Africa (see

[§]data source: https://www.wunderground.com



(a) (b)

Figure 2.4: (a) SKITs with symmetry-based payoffs have high power under the Gaussian model. (b) SKIT with linear kernel has high power under the Gaussian model (X and Y are linearly correlated for $\beta \neq 0$), and its false alarm rate is controlled under the spherical model (X and Y are linearly uncorrelated but dependent).

Appendix A.5.5). In addition, we analyze the performance of SKIT on MNIST data; the details are deferred to Appendix A.5.6.

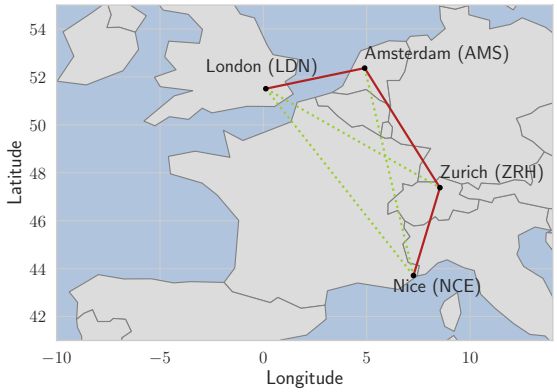


Figure 2.5: Solid lines connect cities for which the null is rejected. SKIT supports the conjecture regarding dependent temperature fluctuations in nearby locations.

2.5 Conclusion

A key advantage of sequential tests is that they can be continuously monitored, allowing an analyst to adaptively decide whether to stop and reject the null hypothesis or to continue collecting data, without inflating the false positive rate. In this paper, we design consistent sequential kernel independence tests (SKITs) following the principle of testing by betting. SKITs are also valid beyond the i.i.d. setting, allowing the data distribution to drift over time. Experiments on synthetic and real data confirm the power of SKITs.

Chapter 3

Sequential Predictive Two-Sample and Independence Testing

3.1 Introduction

We consider two closely-related problems of nonparametric two-sample and independence testing. In the former, given observations from two distributions P and Q, the goal is to test the null hypothesis that the distributions are the same: $H_0 : P = Q$, against the alternative that they are not: $H_1 : P \neq Q$. In the latter, given observations from some joint distribution P_{XY}, the goal is to test the null hypothesis that the random variables are independent: $H_0 : P_{XY} = P_X \times P_Y$, against the alternative that they are not: $H_1 : P_{XY} \neq P_X \times P_Y$. Kernel tests, such as kernel-MMD (Gretton et al., 2012) for two-sample and HSIC (Gretton et al., 2005a) for independence testing, are amongst the most popular methods for solving these tasks which work well on data from simple distributions. However, their performance is sensitive to the choice of a kernel and respective parameters, like bandwidth, and applying such tests requires additional effort. Further, selecting kernels for structured data, like images, is a nontrivial task. Lastly, kernel tests suffer from decaying power in high dimensions (Ramdas et al., 2015).

Predictive two-sample and independence tests (2STs and ITs respectively) aim to address such limitations of kernelized approaches. The idea of using classifiers for two-sample testing dates back to Friedman (2004) who proposed using the output scores as a dimension reduction method. More recent works focused on the direct evaluation of a learned model for testing. In an initial arXiv 2016 preprint, Kim et al. (2021) proposed and analyzed predictive 2STs based on sample-splitting, namely testing whether the accuracy of a model trained on the first split of data and estimated on the second split is significantly better than chance. The authors established the consistency of asymptotic and exact tests in high-dimensional settings and provided rates for the case of Gaussian distributions. Inspired by this work, Lopez-Paz and Oquab (2017) soon after demonstrated that empirically predictive 2STs often outperform

state-of-the-art 2STs, such as kernel-MMD. Recently, Hediger et al. (2022) proposed a related test that utilizes out-of-bag predictions for bagging-based classifiers, such as random forests. To incorporate measures of model confidence, many authors have also explored using test statistics that are based on the output scores instead of the binary class predictions (Kim et al., 2019; Liu et al., 2020; Cheng and Cloninger, 2022; Kübler et al., 2022).

The focus of the above works is on *batch* tests which are calibrated to have a fixed false positive rate (say, 5%) if the sample size is specified in advance. In contrast, we focus on the setting of sequentially released data. Our tests allow on-the-fly decision-making: an analyst can use observed data to decide whether to stop and reject the null or to collect more data, without inflating the false alarm rate.

Problem Setup. First, we define the problems of sequential two-sample and independence testing.

Definition 1 (Sequential two-sample testing)**.** *Suppose that we observe a stream of i.i.d. observations* $((Z_t, W_t))_{t\geq 1}$*, where* $W_t \sim \mathrm{Rademacher}(1/2)$*, the distribution of* $Z_t \mid W_t = +1$ *is denoted* P*, and that of* $Z_t \mid W_t = -1$ *is denoted* Q*. The goal is to design a sequential test for*

$$H_0 : P = Q, \tag{3.1a}$$

$$H_1 : P \neq Q. \tag{3.1b}$$

Definition 2 (Sequential independence testing)**.** *Suppose that we observe that a stream of observations:* $((X_t, Y_t))_{t\geq 1}$*, where* $(X_t, Y_t) \sim P_{XY}$ *for* $t \geq 1$*. The goal is to design a sequential test for*

$$H_0 : (X_t, Y_t) \sim P_{XY} \text{ and } P_{XY} = P_X \times P_Y, \tag{3.2a}$$

$$H_1 : (X_t, Y_t) \sim P_{XY} \text{ and } P_{XY} \neq P_X \times P_Y. \tag{3.2b}$$

We operate in the framework of "power-one tests" (Darling and Robbins, 1968) and define a level-α sequential test as a mapping $\Phi : \cup_{t=1}^{\infty} \mathcal{Z}^t \to \{0, 1\}$ that satisfies: $\mathbb{P}_{H_0}(\exists t \geq 1 : \Phi(Z_1, \ldots, Z_t) = 1) \leq \alpha$. We refer to such notion of type I error control as *time-uniform*. Here, 0 stands for "do not reject the null yet" and 1 stands for "reject the null and stop". Defining the stopping time as the first time that the test outputs 1: $\tau := \inf\{t \geq 1 : \Phi(Z_1, \ldots, Z_t) = 1\}$, a sequential test must satisfy

$$\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha. \tag{3.3}$$

We aim to design *consistent* tests which are guaranteed to stop if the alternative happens to be true:

$$\mathbb{P}_{H_1}(\tau < \infty) = 1. \tag{3.4}$$

Our construction follows the principle of testing by betting (Shafer, 2021). The most closely related work is that of "nonparametric 2ST by betting" of Shekhar and Ramdas (2021), which later inspired several follow-up works, including sequential (marginal) kernelized independence tests of Podkopaev et al. (2023), and the sequential conditional independence tests under the model-X assumption of Grünwald et al. (2023) and Shaer et al. (2023). We

extend the line of work of Shekhar and Ramdas (2021) and of Podkopaev et al. (2023), studying predictive approaches in detail.

Sequential predictive 2STs have been studied by Lhéritier and Cazals (2018, 2019), but in practice, those tests were found to be inferior to the ones developed by Shekhar and Ramdas (2021). Recently, Pandeva et al. (2022) proposed a related test that handles the case of the unknown class proportions using ideas from (Wasserman et al., 2020). As we shall see, our tests are closely related to (Lhéritier and Cazals, 2018, 2019; Pandeva et al., 2022), but are consistent under much milder assumptions.

Sequential Nonparametric Two-Sample and Independence Testing by Betting. Suppose that one observes a sequence of random variables $(Z_t)_{t\geq 1}$, where $Z_t \in \mathcal{Z}$. The principle of testing by betting (Shafer and Vovk, 2019; Shafer, 2021) can be described as follows. A player starts the game with initial capital $\mathcal{K}_0 = 1$. At round t, she selects a payoff function $f_t : \mathcal{Z} \to [-1, \infty)$ that satisfies $\mathbb{E}_{Z\sim P_Z}\left[f_t(Z) \mid \mathcal{F}_{t-1}\right] = 0$ for all $P_Z \in H_0$, where $\mathcal{F}_{t-1} = \sigma(Z_1, \ldots, Z_{t-1})$, and bets a fraction of her wealth $\lambda_t \mathcal{K}_{t-1}$ for an $\mathcal{F}_{t-1}$-measurable $\lambda_t \in [0, 1]$. Once Z_t is revealed, her wealth is updated as

$$\mathcal{K}_t = \mathcal{K}_{t-1} + \lambda_t \mathcal{K}_{t-1} f_t(Z_t) = \mathcal{K}_{t-1}\left(1 + \lambda_t f_t(Z_t)\right). \tag{3.5}$$

The wealth of a player measures evidence against the null hypothesis, and if a player can make money in such game, we reject the null. For testing H_0 at level $\alpha \in (0, 1)$, we use the stopping rule:

$$\tau = \inf\left\{t \geq 1 : \mathcal{K}_t \geq 1/\alpha\right\}. \tag{3.6}$$

The validity of the test follows from Ville's inequality (Ville, 1939), a time-uniform generalization of Markov's inequality, since $(\mathcal{K}_t)_{t\geq 0}$ is a nonnegative martingale under any $P_Z \in H_0$. To ensure high power, one has to choose $(f_t)_{t\geq 1}$ and $(\lambda_t)_{t\geq 1}$ to guarantee the growth of the wealth if the alternative is true. In the context of two-sample and independence testing, Shekhar and Ramdas (2021) and Podkopaev et al. (2023) recently proposed effective betting strategies based on kernelized measures of statistical distance and dependence respectively which admit a variational representation. In a nutshell, datapoints observed prior to a given round are used to estimate the *witness* function — one that best highlights the discrepancy between P and Q for two-sample (or between P_{XY} and $P_X \times P_Y$ for independence) testing — and a bet is formed as an estimator of a chosen measure of distance (or dependence). In contrast, our bets are based on evaluating the performance of a sequentially learned predictor that distinguishes between instances from distributions of interest.

Remark 4. *In practical settings, an analyst may not be able to continue collecting data forever and may adaptively stop the experiment before the wealth exceeds* $1/\alpha$*. In such case, one may use a different threshold for rejecting the null at a stopping time* τ*, namely* U/α*, where* U *is a (stochastically larger than) uniform random variable on* $[0, 1]$

drawn independently from $(\mathcal{F}_t)_{t\geq 0}$. This choice strictly improves the power of the test without violating the validity; see (Ramdas and Manole, 2023).

Contributions. We develop sequential predictive two-sample (Section 3.2) and independence (Section 3.3) tests. We establish sufficient conditions for consistency of our tests and relate those to evaluation metrics of the underlying models. We conduct an extensive empirical study on synthetic and real data, justifying the superiority of our tests over the kernelized ones on structured data.

3.2 Classification-based Two-Sample Testing

Let $\mathcal{G} : \mathcal{Z} \to [-1, 1]$ denote a class of predictors used to distinguish between instances from P (labeled as $+1$) and Q (labeled as -1)*. We assume that: (a) if $g \in \mathcal{G}$, then $-g \in \mathcal{G}$, (b) if $g \in \mathcal{G}$ and $s \in [0, 1]$, then $sg \in \mathcal{G}$, and (c) predictions are based on $\operatorname{sign}[g(\cdot)]$, and if $g(z) = 0$, then z is assigned to the positive class. Two natural evaluation metrics of a predictor $g \in \mathcal{G}$ include the misclassification and the squared risks:

$$R_{\mathrm{m}}(g) := \mathbb{P}\left(W \cdot \operatorname{sign}\left[g\left(Z\right)\right] < 0\right), \quad R_{\mathrm{s}}(g) := \mathbb{E}\left[\left(g(Z) - W\right)^2\right], \tag{3.7}$$

which give rise to the following measures of distance between P and Q, namely

$$d_{\mathrm{m}}(P, Q) := \sup_{g\in\mathcal{G}} \left(\tfrac{1}{2} - R_{\mathrm{m}}(g)\right), \quad d_{\mathrm{s}}(P, Q) := \sup_{g\in\mathcal{G}} \left(1 - R_{\mathrm{s}}(g)\right). \tag{3.8}$$

It is easy to check that $d_{\mathrm{m}}(P, Q) \in [0, 1/2]$ and $d_{\mathrm{m}}(P, Q) \in [0, 1]$. The upper bounds hold due to the non-negativity of the risks and the lower bounds follow by considering $g : g(z) = 0, \forall z \in \mathcal{Z}$. Note that the misclassification risk is invariant to rescaling ($R_{\mathrm{m}}(sg) = R_{\mathrm{m}}(g)$, $\forall s \in (0, 1]$), whereas the squared risk is not, and rescaling any g to optimize the squared risk provides better contrast between P and Q. In the next result, whose proof is deferred to Appendix B.4.3, we present an important relationship between the squared risk of a rescaled predictor and its expected margin: $\mathbb{E}\left[W \cdot g(Z)\right]$.

Proposition 3. *Fix an arbitrary predictor $g \in \mathcal{G}$. The following claims hold:*

1. *For the misclassification risk, we have that:*

$$\sup_{s\in[0,1]} \left(\tfrac{1}{2} - R_{\mathrm{m}}(sg)\right) = \left(\tfrac{1}{2} - R_{\mathrm{m}}(g)\right) \vee 0 = \left(\tfrac{1}{2} \cdot \mathbb{E}\left[W \cdot \operatorname{sign}\left[g(Z)\right]\right]\right) \vee 0. \tag{3.9}$$

*Similar argument can be applied to general scoring-based classifiers: $g : \mathcal{Z} \to \mathbb{R}$, e.g., SVMs, by considering $\tilde{\mathcal{G}} = \{\tilde{g} : \tilde{g}(z) = \tanh(s \cdot g(z)),\ g \in \mathcal{G},\ s > 0\}$, where the constant $s > 0$ corrects the scale of the scores.

2. *For the squared risk, we have that:*

$$\sup_{s\in[0,1]} (1 - R_{\mathrm{s}}(sg)) \geq (\mathbb{E}\left[W \cdot g(Z)\right] \vee 0) \cdot \left(\frac{\mathbb{E}\left[W \cdot g(Z)\right]}{\mathbb{E}\left[g^2(Z)\right]} \wedge 1\right) \tag{3.10}$$

Further, $d_{\mathrm{s}}(P, Q) > 0$ *if and only if there exists* $g \in \mathcal{G}$ *such that* $\mathbb{E}\left[W \cdot g(Z)\right] > 0$.

Consider an arbitrary predictor $g \in \mathcal{G}$. Note that under the null H_0 in (3.1a), the misclassification risk $R_{\mathrm{m}}(g)$ does not depend on g, being equal to 1/2, whereas the squared risk $R_{\mathrm{s}}(g)$ does. In contrast, the lower bound (3.10) no longer depends on g under the null H_0, being equal to 0.

Oracle Test. It is a known fact that the minimizer of either the misclassification or the squared risk is $g^{\mathrm{Bayes}}(z) = 2\eta(z) - 1$, where $\eta(z) = \mathbb{P}\left(W = +1 \mid Z = z\right)$. Since g^{Bayes} may not belong to $\mathcal{G}$, we consider $g_\star \in \mathcal{G}$, which minimizes either the misclassification or the squared risk over predictors in $\mathcal{G}$, and omit superscripts for brevity. To design payoff functions, we follow Proposition 3 and consider

$$f^{\mathrm{m}}_\star(Z_t, W_t) = W_t \cdot \operatorname{sign}\left[g_\star(Z_t)\right] \in \{-1, 1\}, \tag{3.11a}$$

$$f^{\mathrm{s}}_\star(Z_t, W_t) = W_t \cdot g_\star(Z_t) \in [-1, 1]. \tag{3.11b}$$

Let the *oracle* wealth processes based on misclassification and squared risks $(\mathcal{K}^{\mathrm{m},\star}_t)_{t\geq 0}$ and $(\mathcal{K}^{\mathrm{s},\star}_t)_{t\geq 0}$ be defined by using the payoff functions (3.11a) and (3.11b) respectively, along with a predictable sequence of betting fractions $(\lambda_t)_{t\geq 1}$ selected via online Newton step (ONS) strategy (Hazan et al., 2007) (Algorithm 5), which has been studied in the context of coin-betting by Cutkosky and Orabona (2018). If a constant betting fraction is used throughout: $\lambda_t = \lambda$, $\forall t$, then

$$\mathbb{E}\left[\tfrac{1}{t}\log \mathcal{K}^{i,\star}_t\right] = \mathbb{E}\left[\log(1 + \lambda f^i_\star(Z, W))\right], \quad i \in \{\mathrm{m}, \mathrm{s}\}. \tag{3.12}$$

To illustrate the tightness of our results, we consider the optimal constant betting fractions which maximize the log-wealth (3.12) and are constrained to lie in $[-0.5, 0.5]$, like ONS bets:

$$\lambda^i_\star = \operatorname*{arg\,max}_{\lambda\in[-0.5,0.5]} \mathbb{E}\left[\log(1 + \lambda f^i_\star(Z, W))\right], \quad i \in \{\mathrm{m}, \mathrm{s}\}. \tag{3.13}$$

Algorithm 5 Online Newton step (ONS) strategy for selecting betting fractions

Input: sequence of payoffs $(f_t)_{t\geq 1}$, $\lambda^{\mathrm{ONS}}_1 = 0$, $a_0 = 1$.
for $t = 1, 2, \ldots$ **do**
  Observe $f_t \in [-1, 1]$;
  Set $z_t := f_t/(1 - \lambda^{\mathrm{ONS}}_t)$;
  Set $a_t := a_{t-1} + z^2_t$;
  Set $\lambda^{\mathrm{ONS}}_{t+1} := \frac{1}{2} \wedge \left(0 \vee \left(\lambda^{\mathrm{ONS}}_t - \frac{2}{2-\log 3} \cdot \frac{z_t}{a_t}\right)\right)$;

We have the following result for the oracle tests, whose proof is deferred to Appendix B.4.3.

Theorem 3.1. *The following claims hold:*

1. *Suppose that H_0 in* (3.1a) *is true. Then the oracle sequential test based on either $(\mathcal{K}_t^{\mathrm{m},\star})_{t\geq 0}$ or $(\mathcal{K}_t^{\mathrm{s},\star})_{t\geq 0}$ ever stops with probability at most α: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*
2. *Suppose that H_1 in* (3.1b) *is true. Then:*

   (a) *The growth rate of the oracle wealth process $(\mathcal{K}_t^{\mathrm{m},\star})_{t\geq 0}$ satisfies:*

$$\liminf_{t\to\infty} \left(\tfrac{1}{t}\log \mathcal{K}_t^{\mathrm{m},\star}\right) \overset{\text{a.s.}}{\geq} \left(\tfrac{1}{2} - R_{\mathrm{m}}(g_\star)\right)^2. \tag{3.14}$$

   If $R_{\mathrm{m}}(g_\star) < 1/2$, then the test based on $(\mathcal{K}_t^{\mathrm{m},\star})_{t\geq 0}$ is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$. Further, the optimal growth rate achieved by $\lambda_\star^{\mathrm{m}}$ in (3.13) *satisfies:*

$$\mathbb{E}\left[\log(1 + \lambda_\star^{\mathrm{m}} f_\star^{\mathrm{m}}(Z, W))\right] \leq \left(\tfrac{16}{3}\cdot\left(\tfrac{1}{2} - R_{\mathrm{m}}(g_\star)\right)^2 \wedge \left(\tfrac{1}{2} - R_{\mathrm{m}}(g_\star)\right)\right). \tag{3.15}$$

   (b) *The growth rate of the oracle wealth process $(\mathcal{K}_t^{\mathrm{s},\star})_{t\geq 0}$ satisfies:*

$$\liminf_{t\to\infty} \left(\tfrac{1}{t}\log \mathcal{K}_t^{\mathrm{s},\star}\right) \overset{\text{a.s.}}{\geq} \tfrac{1}{4}\cdot\mathbb{E}\left[W \cdot g_\star(Z)\right]. \tag{3.16}$$

   If $\mathbb{E}[W \cdot g_\star(Z)] > 0$, then the test based on $(\mathcal{K}_t^{\mathrm{s},\star})_{t\geq 0}$ is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$. Further, the optimal growth rate achieved by $\lambda_\star^{\mathrm{s}}$ in (3.13) *satisfies:*

$$\mathbb{E}\left[\log(1 + \lambda_\star^{\mathrm{s}} f_\star^{\mathrm{s}}(Z, W))\right] \leq \tfrac{1}{2}\cdot\mathbb{E}\left[W \cdot g_\star(Z)\right]. \tag{3.17}$$

Theorem 3.1 precisely characterizes the properties of the oracle wealth processes and relates those to interpretable metrics of predictive performance. Further, the proof of Theorem 3.1 highlights a direct impact of the variance of the payoffs on the wealth growth rate, and hence the power of the resulting sequential tests (as the null is rejected once the wealth exceeds $1/\alpha$).

The second moment of the payoffs based on the misclassification risk (3.11a) is equal to one, resulting in a *slow* growth: the bound (3.14) is proportional to *squared* deviation of the misclassification risk from one half. The bound (3.15) shows that the growth rate with the ONS strategy matches, up to constants, that of the oracle betting fraction. Note that the second term in (3.15) characterizes the growth rate if $R_{\mathrm{m}}(g_\star) < 5/16$ (low Bayes risk). In this regime, the growth rate of our test is at least $(3/16)\cdot(1/2 - R_{\mathrm{m}}(g_\star))$ which is close to the optimal rate. The second moment of the payoffs based on the squared risk is more insightful. First, we present a result for the case when the oracle predictor $g_\star$ in (3.11b) is replaced by an arbitrary $g \in \mathcal{G}$. The proof is deferred to Appendix B.4.3.

Corollary 3.1.1. *Consider an arbitrary $g \in \mathcal{G}$ with nonnegative expected margin: $\mathbb{E}[W \cdot g(Z)] \geq 0$. Then the growth rate of the corresponding wealth process $(\mathcal{K}_t^{\mathrm{s}})_{t\geq 0}$ satisfies:*

$$\liminf_{t\to\infty} \left(\tfrac{1}{t}\log \mathcal{K}_t^{\mathrm{s}}\right) \overset{\text{a.s.}}{\geq} \tfrac{1}{4}\Big(\sup_{s\in[0,1]} (1 - R_{\mathrm{s}}(sg)) \wedge \mathbb{E}[W \cdot g(Z)]\Big) \tag{3.18a}$$

$$\geq \tfrac{1}{4}\left(\mathbb{E}[W \cdot g(Z)]\right)^2, \tag{3.18b}$$

and the optimal growth rate achieved by $\lambda_\star^{\mathrm{s}}$ in (3.13) *satisfies:*

$$\mathbb{E}[\log(1 + \lambda_\star^{\mathrm{s}} f^{\mathrm{s}}(Z, W))] \leq \Big(\tfrac{4}{3} \cdot \sup_{s\in[0,1]} (1 - R_{\mathrm{s}}(sg))\Big) \wedge \left(\tfrac{1}{2} \cdot \mathbb{E}[W \cdot g(Z)]\right). \tag{3.19}$$

Corollary 3.1.1 states that for an arbitrary $g \in \mathcal{G}$, the growth rate is lower bounded by the minimum of the expected margin and the (optimized) squared risk of such predictor. While the latter term is always smaller for the optimal $g_\star$, this may not hold for an arbitrary $g \in \mathcal{G}$. The lower bound (3.18b), which follows from Proposition 3, is always worse than that for $g_\star$ (the expected margin is squared). The upper bound (3.19) shows that the growth rate with the ONS strategy matches, up to constants, that of the optimal constant betting fraction. Before presenting a practical sequential 2ST, we provide two important remarks that further contextualize the current work in the literature.

Remark 5. *In practice, we learn a predictor sequentially and have to choose a learning algorithm. Note that* (3.18a) *suggests that direct margin maximization may hurt the power of the resulting 2ST: the squared risk is sensitive to miscalibrated and overconfident predictors. Kübler et al. (2022) made a similar conjecture in the context of batch two-sample testing. To optimize the power, the authors suggested minimizing the cross-entropy or the squared loss and related such approach to maximizing the signal-to-noise ratio, a heuristic approach that was proposed earlier by Sutherland et al. (2017)*[†].

Remark 6. *Suppose that $g^{\mathrm{Bayes}} \in \mathcal{G}$ and consider the payoff function based on the squared risk* (3.11b)*. At round t, the wealth of a player $\mathcal{K}_{t-1}$ is multiplied by*

$$\begin{aligned} 1 + \lambda_t \cdot W_t \cdot g^{\mathrm{Bayes}}(Z_t) &= (1 - \lambda_t) \cdot 1 + \lambda_t \cdot \left(1 + W_t \cdot g^{\mathrm{Bayes}}(Z_t)\right) \\ &= (1 - \lambda_t) \cdot 1 + \lambda_t \cdot \frac{(\eta(Z_t))^{\mathbb{1}\{W_t=1\}} (1 - \eta(Z_t))^{\mathbb{1}\{W_t=-1\}}}{\left(\frac{1}{2}\right)^{\mathbb{1}\{W_t=1\}} \left(\frac{1}{2}\right)^{\mathbb{1}\{W_t=-1\}}}, \end{aligned} \tag{3.20}$$

and hence, the betting fractions interpolate between the regimes of not betting and betting using a likelihood ratio. From this standpoint, 2STs of Lhéritier and Cazals (2018, 2019); Pandeva et al. (2022) set $\lambda_t = 1$, $\forall t$, and use only the second term for updating the wealth despite the fact that the true likelihood ratio is unknown. An argument about the consistency of such test hence requires imposing strong assumptions about a sequence of predictors $(g_t)_{t\geq 1}$ (Lhéritier

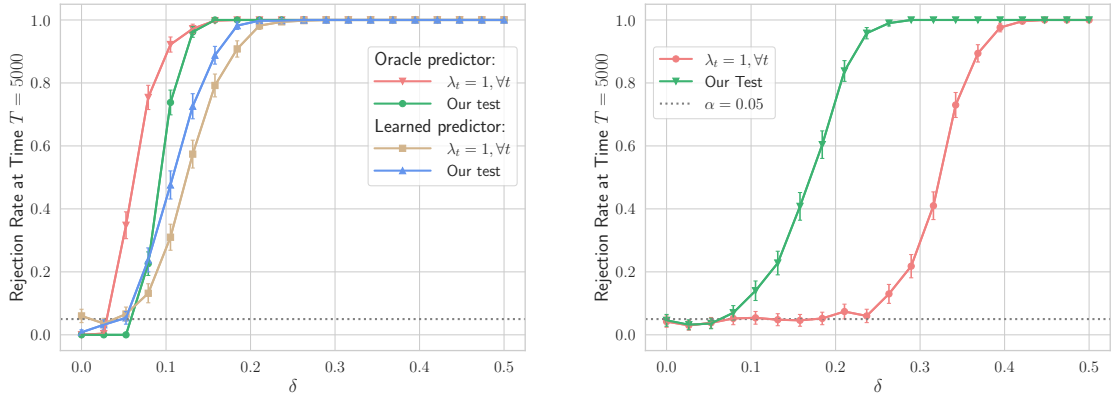
[†]Standard CLT does not apply directly when the conditioning set grows; see (Kim and Ramdas, 2020).

and Cazals, 2018, 2019). Our test differs in a critical way: we use a sequence of betting fractions, $(\lambda_t)_{t\geq 1}$, which adapts to the quality of the underlying predictors, yielding a consistent test under much weaker assumptions.

Example 3. *Consider $P = \mathcal{N}(0,1)$ and $Q = \mathcal{N}(\delta, 1)$ for 20 values of δ, equally spaced in $[0, 0.5]$. For a given δ, the Bayes-optimal predictor is*

$$g^{\text{Bayes}}(z) = \frac{\varphi(z;0,1) - \varphi(z;\delta,1)}{\varphi(z;0,1) + \varphi(z;\delta,1)} \in [-1,1], \tag{3.21}$$

where $\varphi(z;\mu,\sigma^2)$ denotes the density of $\mathcal{N}(\mu,\sigma^2)$ evaluated at z. In Figure 3.1a, we compare tests that use (a) the Bayes-optimal predictor, (b) a predictor constructed with the plug-in estimates of the means and variances. While in the former case betting using a likelihood ratio ($\lambda_t = 1$, $\forall t$) is indeed optimal, our test with an adaptive sequence $(\lambda_t)_{t\geq 1}$ is superior when a predictor is learned. The difference becomes even more drastic in Figure 3.1b where a (regularized) k-NN predictor is used.



(a) Parametric model.

(b) Nonparametric model (k-NN).

Figure 3.1: Comparison between our 2ST with adaptive betting fractions and the likelihood ratio test for Example 3. While the likelihood ratio test is better if the Bayes-optimal predictor is used, our test is superior if a predictor is learned. The results are aggregated over 500 runs for each value of δ.

Practical Test. Let $\mathcal{A}_{\text{c}} : (\cup_{t\geq 1}(\mathcal{Z}\times\{-1,+1\})^t)\times\mathcal{G} \to \mathcal{G}$ denote a learning algorithm which maps a training dataset of any size and previously used classifier, to an updated predictor. For example, $\mathcal{A}_{\text{c}}$ may apply a single gradient descent step using the most recent observation to update a model. We start with $\mathcal{D}_0 = \emptyset$ and $g_1 \in \mathcal{G} : g_1(z) = 0$, for any $z \in \mathcal{Z}$. At round t, we use one of the payoffs:

$$f_t^{\text{m}}(Z_t, W_t) = W_t \cdot \text{sign}\left[g_t(Z_t)\right] \in \{-1,1\}, \tag{3.22a}$$

$$f_t^{\text{s}}(Z_t, W_t) = W_t \cdot g_t(Z_t) \in [-1,1]. \tag{3.22b}$$

After (Z_t, W_t) is used for betting, we update a training dataset: $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(Z_t, W_t)\}$, and an existing predictor: $g_{t+1} = \mathcal{A}_{\text{c}}(\mathcal{D}_t, g_t)$. We summarize our sequential classification-based 2ST (Seq-C-2ST) in Algorithm 6. While we

do not need any assumptions to confirm the type I error control, we place some mild assumptions on the learning algorithm $\mathcal{A}_{\mathrm{c}}$ to argue about the consistency.

Algorithm 6 Sequential classification-based 2ST (Seq-C-2ST)

Input: level $\alpha \in (0,1)$, data stream $((Z_t, W_t))_{t\geq 1}$, $g_1(z) \equiv 0$, $\mathcal{A}_{\mathrm{c}}$, $\mathcal{D}_0 = \emptyset$, $\lambda_1^{\mathrm{ONS}} = 0$.
for $t = 1, 2, \ldots$ **do**
  Evaluate the payoff $f_t^{\mathrm{s}}(Z_t, W_t)$ as in (3.22a);
  Using λ_t^{ONS}, update the wealth process $\mathcal{K}_t^{\mathrm{s}}$ as per (3.5);
  if $\mathcal{K}_t^{\mathrm{s}} \geq 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Update the training dataset: $\mathcal{D}_t := \mathcal{D}_{t-1} \cup \{(Z_t, W_t)\}$;
    Update predictor: $g_{t+1} = \mathcal{A}_{\mathrm{c}}(\mathcal{D}_t, g_t)$;
    Compute $\lambda_{t+1}^{\mathrm{ONS}}$ (Algorithm 5) using $f_t^{\mathrm{s}}(Z_t, W_t)$;

Assumption 3 (R_{m}-learnability)**.** *Suppose that H_1 in* (3.1b) *is true. An algorithm $\mathcal{A}_{\mathrm{c}}$ is such that the resulting sequence $(g_t)_{t\geq 1}$ satisfies:* $\limsup_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \mathbb{1}\{W_i \cdot \operatorname{sign}[g_i(Z_i)] < 0\} \overset{\text{a.s.}}{<} 1/2$.

Assumption 4 (R_{s}-learnability)**.** *Suppose that H_1 in* (3.1b) *is true. An algorithm $\mathcal{A}_{\mathrm{c}}$ is such that the resulting sequence $(g_t)_{t\geq 1}$ satisfies:* $\limsup_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} (g_i(Z_i) - W_i)^2 \overset{\text{a.s}}{<} 1$.

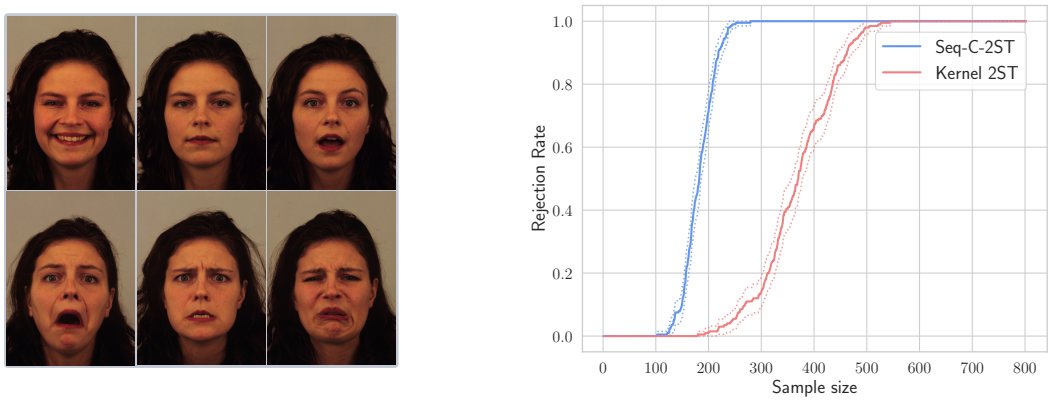
In words, the above assumptions state that a sequence of predictors $(g_t)_{t\geq 1}$ is better than a chance predictor on average. We conclude with the following result, whose proof is deferred to Appendix B.4.3.

Theorem 3.2. *The following claims hold for Seq-C-2ST (Algorithm 6):*

1. *If H_0 in* (3.1a) *is true, the test ever stops with probability at most α:* $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.
2. *Suppose that H_1 in* (3.1b) *is true. Then:*
   (a) *Under Assumption 3, the test with the payoff* (3.22a) *is consistent:* $\mathbb{P}_{H_1}(\tau < \infty) = 1$.
   (b) *Under Assumption 4, the test with the payoff* (3.22b) *is consistent:* $\mathbb{P}_{H_1}(\tau < \infty) = 1$.

Real Data Experiment. To compare sequential classification-based and kernelized 2STs, we consider Karolinska Directed Emotional Faces dataset (KDEF) (Lundqvist et al., 1998) that contains images of actors and actresses expressing different emotions: afraid (AF), angry (AN), disgusted (DI), happy (HA), neutral (HE), sad (SA), and surprised (SU). Following earlier works (Lopez-Paz and Oquab, 2017; Jitkrittum et al., 2016), we focus on straight profile only and assign HA, NE, SU emotions to the positive class (instances from P), and AF, AN, DI emotions to the negative class (instances from Q); see Figure 3.2a. We remove corrupted images and obtain a dataset containing 802 images with six different emotions. The original images (562×762 pixels) are cropped to exclude the background, resized to 64×64 pixels and converted to grayscale.

For Seq-C-2ST, we use a small CNN as an underlying model and defer details about the architecture and training to Appendix B.5.1. As a reference kernel-based 2ST, we use the sequential MMD test of Shekhar and Ramdas (2021) and



(a) (b)

Figure 3.2: (a) Examples of instances from P (top row) and Q (bottom row) for KDEF dataset. (b) Rejection rates for our test (Seq-C-2ST) and the sequential kernelized 2ST. While both tests achieve perfect power with enough data, our test is superior to the kernelized approach, requiring fewer observations to do so. The results are averaged over 200 random orderings of the data.

adapt it to the setting where at each round either an observation from P or that from Q is revealed; see Appendix B.5.1 for details. In Figure 3.2b, we illustrate that while both tests achieve perfect power after processing sufficiently many observations, our Seq-C-2ST requires fewer observations to do so.

3.3 Classification-based Independence Testing

Sequential Classification-based Independence Test (Seq-C-IT). Under the setting of Definition 2, a single point from P_{XY} is revealed at each round. Following (Podkopaev et al., 2023), we bet on two points from P_{XY} (labeled as $+1$) and utilize external randomization to produce instances from $P_X \times P_Y$ (labeled as -1). Let $\mathcal{A}_{\mathrm{c}}^{\mathrm{IT}} : (\cup_{t\geq 1}((\mathcal{X}\times\mathcal{Y})\times\{-1,+1\})^t)\times\mathcal{G}\to\mathcal{G}$ denote a learning algorithm which maps a training dataset of any size and previously used classifier, to an updated predictor. We start with $\mathcal{D}_0=\emptyset$ and $g_1 : g_1(x,y)=0, \forall (x,y)\in\mathcal{X}\times\mathcal{Y}$. We use derandomized versions of the payoffs (3.22), e.g., instead of (3.22b), we use

$$\begin{aligned} f_t^{\mathrm{s}}((X_{2t-1},Y_{2t-1}),(X_{2t},Y_{2t})) = \tfrac{1}{4}\left(g_t(X_{2t-1},Y_{2t-1})+g_t(X_{2t},Y_{2t})\right) \\ -\tfrac{1}{4}\left(g_t(X_{2t-1},Y_{2t})+g_t(X_{2t},Y_{2t-1})\right). \end{aligned} \tag{3.23}$$

After $(X_{2t-1},Y_{2t-1}),(X_{2t},Y_{2t})$ have been used for betting, we update a training dataset:

$$\mathcal{D}_t = \mathcal{D}_{t-1}\cup\{((X_{2t-1},Y_{2t-1}),+1),((X_{2t},Y_{2t}),+1),((X_{2t-1},Y_{2t}),-1),((X_{2t},Y_{2t-1}),-1)\},$$

and an existing predictor: $g_{t+1} = \mathcal{A}_{\mathrm{c}}^{\mathrm{IT}}(\mathcal{D}_t, g_t)$. Seq-C-IT inherits the time-uniform type I error control and the consistency guarantees of Theorem 3.2, and we omit details for brevity.

Synthetic Experiments. In our evaluation, we first consider synthetic datasets where the complexity of the independence testing setup is characterized by a single univariate parameter. We set the monitoring horizon to $T = 5000$ points from P_{XY}, and for each parameter value, we aggregate the results over 200 runs. In particular, we use the following synthetic settings:

1. *Spherical model.* Let $(U_t)_{t\geq 1}$ be a sequence of random vectors on a unit sphere in $\mathbb{R}^d$: $U_t \overset{\text{iid}}{\sim} \text{Unif}(\mathbb{S}^d)$, and let $u_{(i)}$ denote the i-th coordinate of u. For $t \geq 1$, we take

$$(X_t, Y_t) = ((U_t)_{(1)}, (U_t)_{(2)}).$$

   We consider $d \in \{3, \ldots, 10\}$, where larger d defines a harder setup.

2. *Hard-to-detect-dependence (HTDD) model.* We sample $((X_t, Y_t))_{t\geq 1}$ from

$$p(x, y) = \tfrac{1}{4\pi^2}\left(1 + \sin(wx)\sin(wy)\right) \cdot \mathbb{1}\left\{(x, y) \in [-\pi, \pi]^2\right\}. \tag{3.24}$$

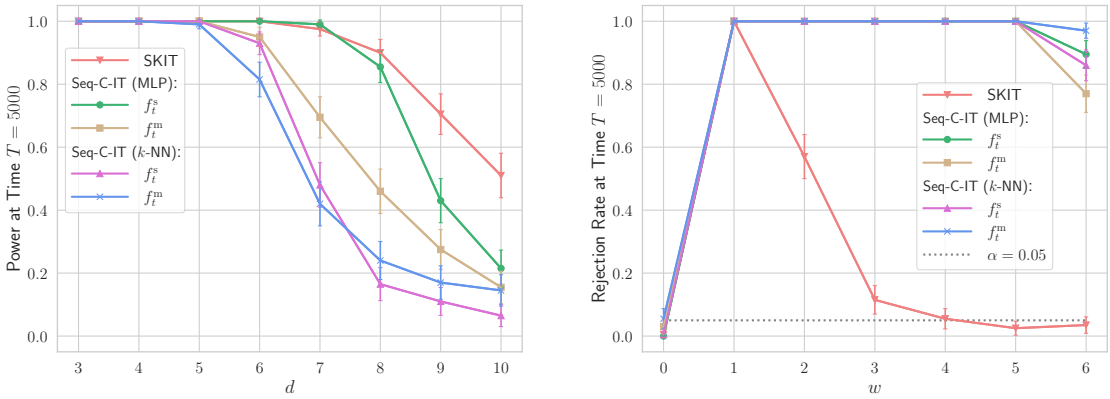
   We consider $w \in \{0, \ldots, 6\}$, where H_0 is true (random variables are independent) if and only if $w = 0$. For $w > 0$, $\text{Corr}\,(X, Y) \approx 1/w^2$, and the setup is harder for larger w.

For the comparison, we use two predictive models to construct Seq-C-ITs:

1. Let $\mathcal{N}_t(z) := \mathcal{N}(z, \mathcal{D}_{t-1}, k_t)$ define the set of k_t closest points in $\mathcal{D}_{t-1}$ to a query point $z := (x, y)$. We consider a *regularized* k-NN predictor: $\widehat{g}_t(z) = \frac{1}{k_t+1}\sum_{(Z,W)\in\mathcal{N}_t(z)} W$. We select the number of neighbors using the square-root rule: $k_t = \sqrt{|\mathcal{D}_{t-1}|} = \sqrt{4(t-1)}$.
2. We use a multilayer perceptron (MLP) with three hidden layers and 128, 64 and 32 neurons respectively and the parameters learned using an incremental training scheme.

We use the HSIC-based sequential kernelized independence test (SKIT) (Podkopaev et al., 2023) as a reference test and defer details, such as MLP training scheme and SKIT hyperparameters, to Appendix B.5.1. In Figure 3.3, we observe that SKIT outperforms Seq-C-ITs under the spherical model (with no localized dependence structure), whereas, under the structured HTDD model, Seq-C-ITs, is superior. Further, inspecting Figure 3.3b at $w = 0$ confirms that all tests control the type I error. We refer the reader to Appendix B.5.2 for additional experiments on synthetic data with localized dependence where Seq-C-ITs are superior. In Appendix B.5.2, we also provide the results for the average *stopping times* of our tests: we empirically confirm that our tests are adaptive to the complexity of a problem at hand: they stop earlier on easy tasks and later on harder ones.

Real Data Experiment. We compare two independence tests on MNIST image dataset (LeCun et al., 1998). To simulate the null setting, we sample pairs of random images from the entire dataset, and to simulate the alternative, we sample pairs of random images depicting the same digit (Figure 3.4a). For Seq-C-IT, we use MLP with the same

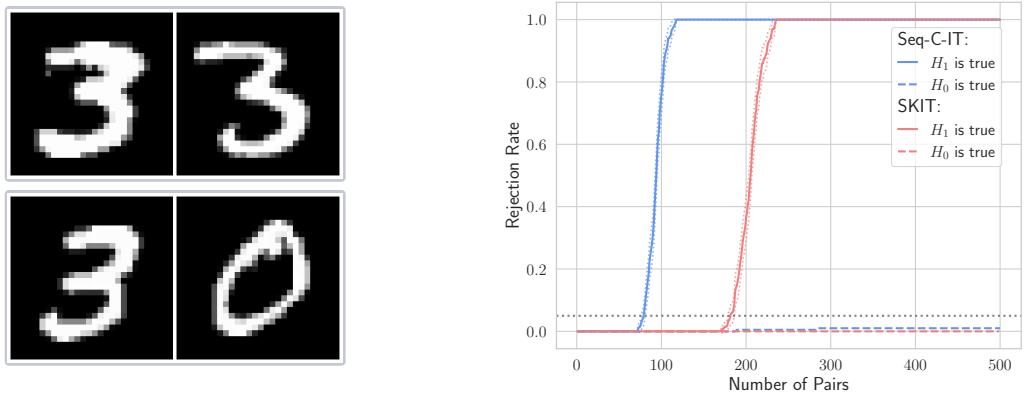


(a) Spherical model.

(b) HTDD model.

Figure 3.3: Power of different sequential independence tests on synthetic data from Section 3.3. Under the spherical model (no localized dependence), SKIT is better than Seq-C-ITs. Under the (structured) HTDD model, SKIT is inferior to sequential predictive independence tests.

architecture as for simulations on synthetic data. For SKIT, we use the median heuristic with 20 points from P_{XY} to compute kernel hyperparameters. In Figure 3.4b, we show that while both tests control the type I error under H_0, SKIT is inferior to Seq-C-IT under H_1, requiring twice as much data to achieve perfect power.



(a)

(b)

Figure 3.4: (a) Instances from the P_{XY} (top row) and $P_X \times P_Y$ (bottom row) for MNIST dataset. (b) While both independence tests control the type I error under H_0, Seq-C-IT outperforms SKIT under H_1, rejecting the null much sooner. The results are aggregated over 200 runs.

3.4 Conclusion

While kernel methods are state-of-the-art for nonparametric two-sample and independence testing, their performance often deteriorates on complex data, e.g., high-dimensional data with localized dependence. In such settings, prediction-based tests are often much more effective. In this work, we developed sequential predictive two-sample and

independence tests following the principle of testing by betting. Our tests control the type I error despite continuously monitoring the data and are consistent under weak and tractable assumptions. Further, our tests provably adapt to the complexity of a problem at hand: they stop earlier on easy tasks and later on harder ones. An additional advantage of our tests is that an analyst may modify the design choices, e.g., model architecture, on-the-fly. Through experiments on synthetic and real data, we confirm that our tests are competitive to kernel-based ones overall and outperform those under structured settings.

We refer the reader to the Appendix for additional results that were not included in the main paper:

1. In Appendix B.1, we complement classification-based ITs with a regression-based approach. Regression-based ITs represent an alternative to the classification-based approach in settings where a data stream $((X_t, Y_t))_{t\geq 1}$ may be processed directly as feature-response pairs.

2. In Section 3.2, we considered the case of balanced classes, meaning that at each round, an instance from either P or Q is observed with equal chance. In Appendix B.2, we extend the methodology to a more general case of two-sample testing with unknown class proportions.

3. Batch two-sample and independence tests rely on either a cutoff computed using the asymptotic null distribution of a chosen test statistic (when it is tractable) or a permutation p-value, and if the distribution drifts, both approaches fail to provide the type I error control. In contrast, Seq-C-2ST and Seq-C-IT remain valid beyond the i.i.d. setting by construction (analogous to tests developed by Shekhar and Ramdas (2021); Podkopaev et al. (2023)), and we refer the reader to Appendix B.3 for more details.

Chapter 4

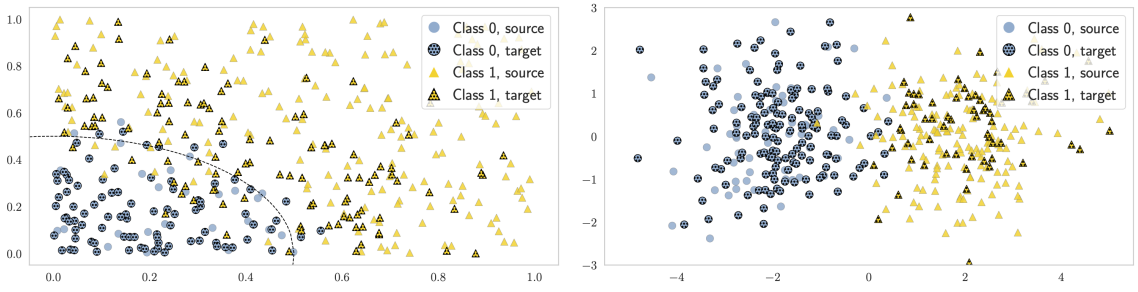
Tracking the Risk of a Deployed Model and Detecting Harmful Distribution Shifts

4.1 Introduction

Developing a machine learning system usually involves data splitting where one of the labeled folds is used to assess its generalization properties. Under the assumption that the incoming test instances (target) are sampled independently from the same underlying distribution as the training data (source), estimators of various performance metrics, such as accuracy or calibration, are accurate. However, a model deployed in the real world inevitably encounters variability in the input distribution, a phenomenon referred to as *dataset shift; see the book by Quionero-Candela et al. (2009). Commonly studied settings* include covariate shift (Shimodaira, 2000) and label shift (Saerens et al., 2002). While testing whether a distribution shift is present has been studied both in offline (Rabanser et al., 2019; Gretton et al., 2012; Hu and Lei, 2020) and online (Vovk et al., 2005; Vovk, 2020a,b) settings, a natural question is whether an intervention is required once there is evidence that a shift has occurred.

A trustworthy machine learning system has to be supplemented with a set of tools designed to raise alarms whenever critical changes to the environment take place. Vovk et al. (2021) propose retraining once an i.i.d. assumption becomes violated and design corresponding online testing protocols. However, naively testing for the presence of distribution shift is not fully practical since it does not take into account the *malignancy* of a shift (Rabanser et al., 2019). To elaborate, users are typically interested in how a model performs according to certain prespecified metrics. In *benign* scenarios, distribution shifts could be present but may not significantly affect model performance. Raising unnecessary alarms might then lead to delays and a substantial increase in the cost of model deployment. The recent approach by Vovk et al. (2021) based on conformal test martingales is highly dependent on the choice of conformity

score. In general, the methodology raises an alarm whenever a deviation from i.i.d. is detected, which does not necessarily imply that the deviation is harmful (see Appendix C.1.2).



(a) (b)

Figure 4.1: Samples from the source and the target (hatched) distributions under benign (a) covariate and (b) label shifts. (a) $X \sim \mathrm{Unif}([0,1] \times [0,1])$ on the source and $X_i \sim \mathrm{Beta}(1,2)$, $i = 1,2$ on the target. Labels satisfy: $\mathbb{P}(Y = 1 \mid X = x) = \mathbb{P}\left(x_1^2 + x_2^2 + \varepsilon \geq 1/4\right)$ where $\varepsilon \sim \mathcal{N}(0, 0.01)$. (b) Marginal probability of class 1 changes from $\pi_1^S = 0.7$ on the source to $\pi_1^T = 0.3$ on the target. Covariates satisfy: $X \mid Y = y \sim \mathcal{N}(\mu_y, I_2)$, where $\mu_0 = (-2, 0)^\top$, $\mu_1 = (2, 0)^\top$. In both cases, a model which separates well data from the source will generalize well to the target.

In some cases, it is possible to handle structured shifts in a post-hoc fashion without performing expensive actions, such as model retraining. One example arises within the context of distribution-free uncertainty quantification where the goal is to supplement predictions of a model with a measure of uncertainty valid under minimal assumptions. Recent works (Tibshirani et al., 2019; Gupta et al., 2020; Podkopaev and Ramdas, 2021) show how to adapt related procedures for handling covariate and label shifts without labeled data from the target. However, both aforementioned shifts impose restrictive assumptions on the possible changes in the underlying probability distribution, assuming either that $P(X)$ changes but $P(Y|X)$ stays unchanged (covariate shift assumption), or that $P(Y)$ changes but $P(X|Y)$ stays unchanged (label shift assumption).

Thinking of distribution shifts only in terms of covariate or label shifts has two drawbacks: such (unverifiable) assumptions often may be unrealistic, and even if they were plausible, such shifts may be benign and thus could be ignored. To elaborate on the first point, it is evident that while distribution shifts constantly occur in practice, they may generally have a more complex nature. In medical diagnosis, $P(Y)$ and $P(X|Y = y)$ could describe the prevalence of certain diseases in the population and symptoms corresponding to disease y. One might reasonably expect not only the former to change over time (say during flu season or epidemics) but also the latter (due to potential mutations and partially-effective drugs/vaccines), thus violating both the covariate and label shift assumptions. Regarding the second point, a model capable of separating classes sufficiently well on the source distribution can sometimes generalize well to the target. We illustrate such benign covariate and label shifts on Figures 4.1a and 4.1b respectively. We argue that the critical distinction—from the point of view on raising alarms—should be built between harmful and benign shifts, and not between covariate and label shifts.

A related question is whether labeled data from one distribution can be used for training a model in a way that it generalizes well to another distribution where it is hard to obtain labeled examples. Importance-weighted risk minimization can yield models with good generalization properties on the target domain, but the corresponding statistical guarantees typically become vacuous if the importance weights are unbounded, which happens when the source distribution's support fails to cover the target support. Adversarial training schemes (Ganin et al., 2016; Wu et al., 2019) for deep learning models often yield models with reasonable performance on some types of distribution shifts, but some unanticipated shifts could still degrade performance. This paper does not deal with how to train a model if a particular type of shift is anticipated; it answers the question of when one should consider retraining (or re-evaluating) a currently deployed model.

We argue for triggering a warning once the non-regularities in the data generating distribution lead to a statistically significant increase in a user-specified risk metric. We design tools for nonparametric sequential testing for an unfavorable change in a chosen risk function of any black-box model. The procedure can be deployed in settings where (some) true target labels can be obtained, immediately after prediction or in a delayed fashion.

During the preparation of our paper, we noticed a very recent preprint (Kamulete, 2022) on broadly the same topic. While they also advise against testing naively for the presence of a shift, their approach is different from ours as (a) it is based on measuring the malignancy of a shift based on outlier scores (while we suggest measuring malignancy via a drop in test accuracy or another prespecified loss), and arguably more importantly (b) their procedure is non-sequential (it is designed to be performed once, e.g, at the end of the year, and cannot be continuously monitored, but ours is designed to flag alarms at any moment when a harmful shift is detected). Adapting fixed-time testing to the sequential settings requires performing corrections for multiple testing: if the correction is not performed, the procedure is no longer valid, but if naively performed, the procedure becomes too conservative, due to the dependence among the tests being ignored (see Appendix C.1.1). We reduce the testing problem to performing sequential estimation that allows us to accumulate evidence over time, without throwing away any data or the necessity of performing explicit corrections for multiple testing; these are implicitly handled efficiently by the martingale methods that underpin the sequential estimation procedures (Howard et al., 2021; Waudby-Smith and Ramdas, 2023).

In summary, the main contributions of this work are:

1. We deviate from the literature on detecting covariate or label shifts and instead focus on differentiating harmful and benign shifts. We pose the latter problem as a nonparametric sequential hypothesis test, and we differentiate between malignant and benign shifts by measuring changes in a user-specified risk metric.

2. We utilize recent progress in sequential estimation to develop tests that provably control the false alarm rate despite the multiple testing issues caused by continuously monitoring the deployed model (Section 4.2.2), and without constraining the form of allowed distribution shifts. For example, we do not require the target data to itself be i.i.d.; our methods are provably valid even if the target distribution is itself shifting or drifting over time.

3. We evaluate the framework on both simulated (Section 4.3.1) and real data (Section 4.3.2), illustrating its promising empirical performance. In addition to traditional losses, we also study several generalizations of the Brier score (Brier, 1950) to multiclass classification.

4.2 Sequential Testing for a Significant Risk Increase

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the covariate and label spaces respectively. Consider predictors $f : \mathcal{X} \to \mathcal{Y}$. Let $\ell(\cdot, \cdot)$ be the loss function chosen to be monitored, with $R(f) := \mathbb{E}\left[\ell(f(X), Y)\right]$ denoting the corresponding expected loss, called the risk of f.

We assume from here onwards that ℓ is bounded, which is the only restriction made. This is needed to quantify how far the empirical target risk is from the true target risk without making assumptions on $P(X, Y)$. This is not a restriction of the current paper only, but appears broadly in statistical learning theory (or the study of concentration inequalities): if one picks an unbounded loss function—say the logarithmic loss—which can take on infinite values, then it is impossible to say how far the true and empirical risks are without further assumptions, for example that the distribution of $P(Y|X)$ is light tailed. We prefer not to make such assumptions in this paper, keeping it as assumption-light as possible and thus generally applicable in various domains. The restriction to bounded losses is not heavy since examples abound in the machine learning literature; see some examples in Appendix C.2.

Remark 7. *For classification, sometimes one does not predict a single label, but a distribution over labels. In that case the range of f would be $\Delta^{|\mathcal{Y}|}$. This poses no issue, and it is common to use bounded loss functions and risks such as the Brier score, as exemplified in Section 4.3. On a different note, for regression, the loss (like squared error) is bounded only if the observations are. This is reasonable in some contexts (predicting rain or snow) but possibly not in others (financial losses).*

4.2.1 Casting the Detection of Risk Increase as a Sequential Hypothesis Test

We aim to trigger a warning whenever the risk on the target domain exceeds the risk on the source by a non-negligible amount specified in advance. For example, alerting could happen once it is possible to conclude with certain confidence that the accuracy has decreased by 10%. Shifts that lead to a decrease or an insignificant increase in the risk are then treated as benign. Formally, we aim to construct a sequential test for the following pair of hypotheses:

$$H_0 : \quad R_T(f) \leq R_S(f) + \varepsilon_{\text{tol}}, \quad \text{vs.} \quad H_1 : \quad R_T(f) > R_S(f) + \varepsilon_{\text{tol}}, \tag{4.1}$$

where $\varepsilon_{\text{tol}} \geq 0$ is an acceptable tolerance level, and $R_S(f)$ and $R_T(f)$ stand for the risk of f on the source and target domains respectively. Assume that one observes a sequence of data points $Z_1, Z_2, \ldots$. At each time point t, a sequential test takes the first t elements of this sequence and output either a 0 (continue) or 1 (reject the null and

stop). The resulting sequence of 0s and 1s satisfies the property that if the null H_0 is true, then the probability that the test ever outputs a 1 and stops (false alarm) is at most δ. In our context, this means that if a distribution shift is benign, then with high probability, the test will never output a 1 and stop, and thus runs forever. Formally, a level-δ sequential test Φ defined as a mapping $\bigcup_{n=1}^{\infty} \mathcal{Z}^n \to \{0,1\}$ must satisfy: $\mathbb{P}_{H_0}(\exists t \geq 1 : \Phi(Z_1, ..., Z_t) = 1) \leq \delta$. Note that the sequential nature of a test is critical here as we aim to develop a framework capable of continuously updating inference as data from the target is collected, making it suitable for many practical scenarios. We distinguish between two important settings when we assume that the target data either satisfies (a) an i.i.d. assumption (under the same or different distribution as the source) or (b) only an independence assumption. While the i.i.d. assumption may be arguably reasonable on the source, it is usually less realistic on the target, since in practice, one may expect the distribution to drift slowly in a non-i.i.d. fashion instead of shifting sharply but staying i.i.d.. Under setting (b), the quantity of interest on the target domain is the *running risk*:

$$R^{(t)}(f) = \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[\ell\left(f(X_i'), Y_i'\right)\right], \quad t \geq 1,$$

where the expected value is taken with respect to the joint distribution of (X_i', Y_i'), possibly different for each test point i. The goal transforms into designing a test for the following pair of hypotheses:

$$H_0: \quad R_T^{(t)}(f) \leq R_S(f) + \varepsilon_{\text{tol}},\ \forall t \geq 1, \quad \text{vs.} \quad H_1: \quad \exists t^\star \geq 1 : R_T^{(t^\star)}(f) > R_S(f) + \varepsilon_{\text{tol}}. \tag{4.2}$$

When considering other notions of risk beyond the misclassification error, one could also be interested in relative changes in the risk, and thus a sequential test for the following pair of hypotheses:

$$H_0': \quad R_T(f) \leq (1 + \varepsilon_{\text{tol}}) R_S(f), \quad \text{vs.} \quad H_1': \quad R_T(f) > (1 + \varepsilon_{\text{tol}}) R_S(f). \tag{4.3}$$

The proposed framework handles all of the aforementioned settings as we discuss next. The most classical approach for sequential testing is the sequential probability ratio test (SPRT) due to Wald (1945). However, it can only be applied, even for a point null and a point alternative, when the relevant underlying distributions are known. While extensions of the SPRT exist to the composite null and alternative (our setting above), these also require knowledge of the distributions of the test statistics (e.g., empirical risk) under the null and alternative, and being able to maximize the likelihood. Clearly, we make no distributional assumptions and so we require a nonparametric approach. We perform sequential testing via the dual problem of nonparametric sequential estimation, a problem for which there has been much recent progress to draw from.

4.2.2 Sequential Testing via Sequential Estimation

When addressing a particular prediction problem, the true risk on neither the source nor the target domains is known. Performance of a model on the source domain is usually assessed through a labeled holdout source sample of a fixed size n_S: $\{(X_i, Y_i)\}_{i=1}^{n_S}$. We can write:

$$R_S(f) + \varepsilon_{\text{tol}} = \widehat{R}_S(f) + \Big(R_S(f) - \widehat{R}_S(f)\Big) + \varepsilon_{\text{tol}},$$

where $\widehat{R}_S(f) := \left(\sum_{i=1}^{n_S} \ell(f(X_i), Y_i)\right)/n_S$. For any fixed tolerance level $\delta_S \in (0, 1)$, classic concentration results can be used to obtain an upper confidence bound $\varepsilon_{\text{appr}}$ on the difference $R_S(f) - \widehat{R}_S(f)$, and thus to conclude that with probability at least $1 - \delta_S$:

$$R_S(f) + \varepsilon_{\text{tol}} \leq \widehat{U}_S(f) + \varepsilon_{\text{tol}}, \quad \text{where} \quad \widehat{U}_S(f) = \widehat{R}_S(f) + \varepsilon_{\text{appr}}. \tag{4.4}$$

For example, by Hoeffding's inequality, $n_S = O(1/\varepsilon^2_{\text{appr}})$ points suffice for the above guarantee, but that bound can be quite loose when the individual losses $\ell(f(X_i), Y_i)$ have low variance. In such settings, recent variance-adaptive confidence bounds (Waudby-Smith and Ramdas, 2023; Howard et al., 2021) are tighter. It translates to an increase in the power of the framework, allowing for detecting harmful shifts much earlier, while still controlling the false alarm rate at a prespecified level.

In contrast, the estimator of the target risk has to be updated as losses on test instances are observed. While the classic concentration results require specifying in advance the size of a sample used for estimation, time-uniform confidence sequences retain validity under adaptive data collection settings. For any chosen $\delta_T \in (0, 1)$, those yield a time-uniform lower confidence bound on $R_T(f)$:

$$\mathbb{P}\left(\exists t \geq 1 : R_T(f) < \widehat{L}_T^{(t)}(f)\right) \leq \delta_T,$$

where $\widehat{L}_T^{(t)}(f)$ is the bound constructed after processing t test points. We typically set $\delta_S = \delta_T = \delta/2$, where δ refers to the desired type I error. Under the independence assumption ((4.2)), the form of the drift in the distribution of the data is allowed to change with time. From the technical perspective, the difference with the i.i.d. setting is given by the applicability of particular concentration results. While the betting-based approach of Waudby-Smith and Ramdas (2023) necessitates assuming that random variables share a common mean, proceeding with the conjugate-mixture empirical-Bernstein bounds (Howard et al., 2021) allows us to lift the common-mean assumption and handle a time-varying mean. We summarize the testing protocol in Algorithm 7 and refer the reader to Appendix C.5 for a review of the concentration results used in this work. One can easily adapt the framework to proceed with a fixed, absolute threshold on the risk, rather than a relative threshold $R_S(f) + \varepsilon_{\text{tol}}$, e.g., we can raise a warning once accuracy drops below 80%, rather than 5% below the training accuracy.

Algorithm 7 Sequential testing for an absolute increase in the risk.

Input: Predictor f, loss ℓ, tolerance level ε_{tol}, sample from the source $\{(X_i, Y_i)\}_{i=1}^{n_S}$.

1: **procedure**
2: Compute the upper confidence bound on the source risk $\widehat{U}_S(f)$;
3: **for** $t = 1, 2, \ldots$ **do**
4: Compute the lower confidence bound on the target risk $\widehat{L}_T^{(t)}(f)$;
5: **if** $\widehat{L}_T^{(t)}(f) > \widehat{U}_S(f) + \varepsilon_{\text{tol}}$ **then**
6: Reject H_0 ((4.1)) and fire off a warning.

Testing for a relative increase in the risk is performed by replacing the line 5 in the Algorithm 7 by the condition $\widehat{L}_T^{(t)}(f) > (1 + \varepsilon_{\text{tol}})\widehat{U}_S(f)$. For both cases, the proposed test controls type I error as formally stated next. The proof is presented in Appendix C.4.

Proposition 4. *Fix any $\delta \in (0, 1)$. Let $\delta_S, \delta_T \in (0, 1)$ be chosen in a way such that $\delta_S + \delta_T = \delta$. Let $\widehat{L}_T^{(t)}(f)$ define a time-uniform lower confidence bound on $R_T^{(t)}(f)$ at level δ_T after processing t data points ($t \geq 1$), and let $\widehat{U}_S(f)$ define an upper confidence bound on $R_S(f)$ at level δ_S. Then:*

$$\begin{cases} \mathbb{P}_{H_0}\left(\exists t \geq 1 : \widehat{L}_T^{(t)}(f) > \widehat{U}_S(f) + \varepsilon_{tol}\right) \leq \delta, \\ \mathbb{P}_{H_0'}\left(\exists t \geq 1 : \widehat{L}_T^{(t)}(f) > (1 + \varepsilon_{tol})\widehat{U}_S(f)\right) \leq \delta, \end{cases} \tag{4.5}$$

that is, the procedure described in Algorithm 7 controls the type I error for testing the hypotheses H_0 ((4.1) and (4.2)) and H_0' ((4.3)).

Remark 8. *Both the testing protocol (Algorithm 7) and the corresponding guarantee (Proposition 4) are stated in a form which requires the lower bound on the target risk to be recomputed after processing each test point. More generally, test data could be processed in minibatches of size $m \geq 1$.*

Remark 9. *Type I error guarantee in* (4.5) *holds under continuous monitoring. This goes beyond standard fixed-time guarantees, for which type I error is controlled only when the sample size is fixed in advance, and not under continuous monitoring. Define a stopping time of a sequential test in Algorithm 7:*

$$N(\delta) := \inf\left\{t \geq 1 : \widehat{L}_T^{(t)}(f) > \widehat{U}_S(f) + \varepsilon_{tol}\right\}.$$

Then the guarantee in (4.5) *can be restated as: $\mathbb{P}_{H_0}(N(\delta) < \infty) \leq \delta$, that is the probability of ever raising a false alarm is at most δ.*

From sequential testing to changepoint detection. A valid sequential test can be transformed into a changepoint detection procedure with certain guarantees (Lorden, 1971). The key characteristics of changepoint detection procedures are *average run length* (ARL), or average time to a false alarm, and *average detection delay* (ADD).

One way to convert a sequential test into a detection procedure is by running a separate test starting at each time point $t = 1, 2, \ldots$, and claiming a change whenever the first one of the tests rejects the null. Subsequently, these tests yield a sequence of stopping variables $N_1(\delta), N_2(\delta), \ldots$ The corresponding stopping time is defined as:

$$N^\star(\delta) := \inf_{k=1,2,\ldots} \left(N_k(\delta) + (k-1)\right).$$

Lorden (1971) established a lower bound on the (worst-case) ARL of such changepoint detection procedure of the form: $\mathbb{E}_{H_0}\left[N^\star(\delta)\right] \geq 1/\delta$. The (worst-case) average detection delay is defined as:

$$\overline{\mathbb{E}}_1 N(\delta) = \sup_{m \geq 1} \operatorname{ess\,sup} \mathbb{E}_m \left[(N(\delta) - (m-1))_+ \mid (X'_1, Y'_1), \ldots, (X'_{m-1}, Y'_{m-1})\right],$$

where $\mathbb{E}_m$ denotes expectation under P_m, the distribution of a sequence $(X'_1, Y'_1), (X'_1, Y'_1), \ldots$ under which (X'_m, Y'_m) is the first term from a shifted distribution.

4.3 Experiments

In Section 4.3.1, we analyze the performance of the testing procedure on a collection of simulated datasets. First, we consider settings where the i.i.d. assumption on the target holds, and then relax it to the independence assumption. In Section 4.3.2, we evaluate the framework on real data. We consider classification problems with different metrics of interest including misclassification loss, several versions of the Brier score and miscoverage loss for set-valued predictors. Due to space limitations, we refer the reader to Appendix C.2 for a detailed review of the considered loss functions.

4.3.1 Simulated Data

Tracking the risk under the i.i.d. assumption. Here we induce label shift on the target domain and emulate a setting where it noticeably harms the accuracy of a predictor by modifying the setup from Section 4.1 through updating the class centers to $\mu_0 = (-1, 0)^\top$ and $\mu_1 = (1, 0)^\top$, making the classes largely overlap. The (oracle) Bayes-optimal predictor on the source domain is:

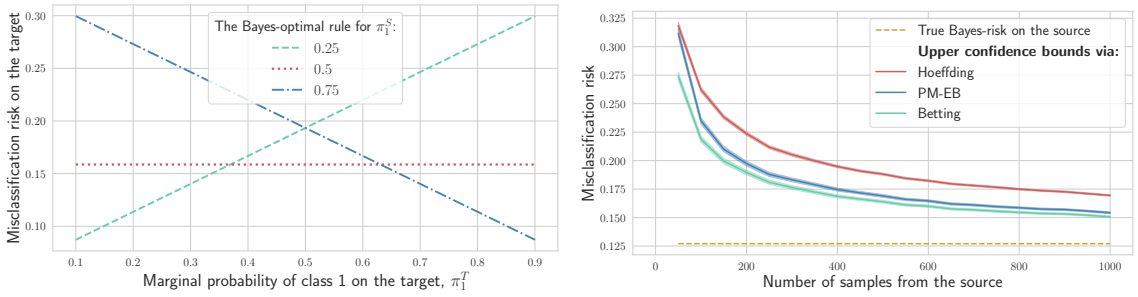
$$f^\star(x) = \frac{\pi_1^S \cdot \varphi(x; \mu_1, I_2)}{\pi_0^S \cdot \varphi(x; \mu_0, I_2) + \pi_1^S \cdot \varphi(x; \mu_1, I_2)}, \tag{4.6}$$

where $\varphi(x; \mu_i, I_2)$ denotes the probability density function of a Gaussian random vector with mean μ_i, $i \in \{0, 1\}$ and an identity covariate matrix. Let ℓ be the 0-1 loss, and thus the misclassification risk $R_T(f^*)$ on the target is:

$$\mathbb{P}_T\left(f^\star(X) \neq Y\right) = \mathbb{P}\left(X^\top(\mu_1 - \mu_0) < \log\left(\frac{\pi_0^S}{\pi_1^S}\right) + \frac{1}{2}\left(\|\mu_1\|_2^2 - \|\mu_0\|_2^2\right) \mid Y = 1\right) \cdot \pi_1^T$$
$$+ \mathbb{P}\left(X^\top(\mu_1 - \mu_0) \geq \log\left(\frac{\pi_0^S}{\pi_1^S}\right) + \frac{1}{2}\left(\|\mu_1\|_2^2 - \|\mu_0\|_2^2\right) \mid Y = 0\right) \cdot \pi_0^T.$$

For three values of π_1^S, the source marginal probability of class 1, we illustrate how label shift affects the misclassification risk of the Bayes-optimal predictor on Figure 4.2a, noting that it is linear in π_1^T. Importantly, whether label shift hurts or helps depends on the value of π_1^S.

We fix $\pi_1^S = 0.25$ and use the corresponding Bayes-optimal rule. On Figure 4.2b, we compare upper confidence bounds on the source risk due to different concentration results, against the size of the source holdout set. Variance-adaptive upper confidence bounds—predictably-mixed empirical-Bernstein (PM-EB) and betting-based (see Appendix C.5)—are much tighter than the non-adaptive Hoeffding's bound. Going forward, we use a source holdout set of 1000 points to compute upper confidence bound on the source risk, where $\varepsilon_{\text{appr}}$ from (4.4) is around 0.025.

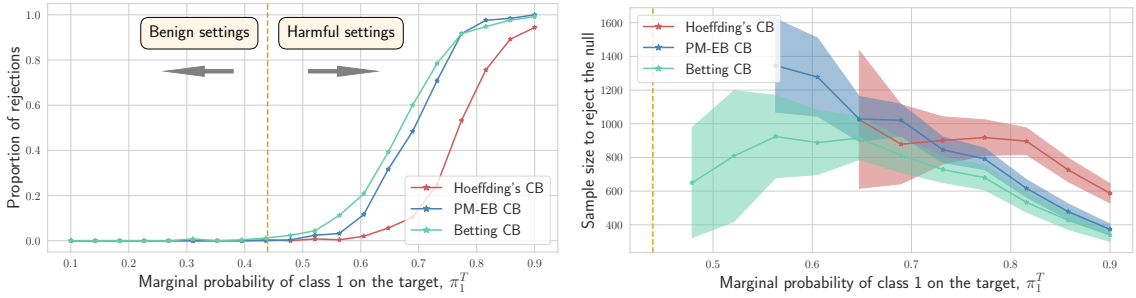


(a) (b)

Figure 4.2: (a) The misclassification risk on the target of the Bayes-optimal predictors for three values of π_1^S. Notice that label shift does not necessarily lead to an increase in the risk. (b) Upper confidence bounds $\widehat{U}_S(f)$ on the misclassification risk on the source obtained via several possible concentration results. For each sample size, the results are aggregated over 1000 random data draws. The variance-adaptive confidence bounds (predictably-mixed empirical-Bernstein and the betting-based one) are much tighter when compared against the non-adaptive one.

Next, we fix $\varepsilon_{\text{tol}} = 0.05$, i.e., we treat a 5% drop in accuracy as significant. For 20 values of π_1^T, evenly spaced in the interval $[0.1, 0.9]$, we sample 40 batches of 50 data points from the target distribution. On Figure 4.3a, we track the proportion of null rejections after repeating the process 250 times. Note that here stronger label shift hurts accuracy more. On Figure 4.3b, we illustrate average size of a sample from the target needed to reject the null. The results confirm that tighter bounds yield better detection procedures, with the most powerful test utilizing the betting-based bounds (Waudby-Smith and Ramdas, 2023). A similar analysis for the Brier score (Brier, 1950) as a target metric

is presented in Appendix C.6. We further study the performance of the framework under the covariate shift setting (Appendix C.8).



(a) (b)

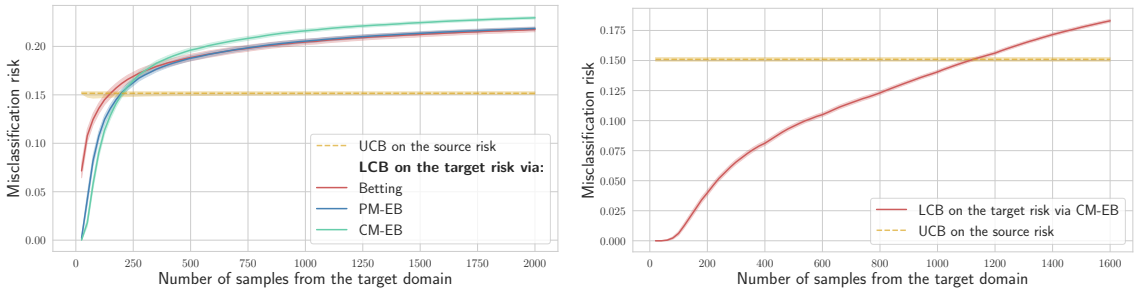
Figure 4.3: (a) Proportion of null rejections when testing for an increase in the misclassification risk after processing 2000 samples from a shifted distribution. The vertical dashed yellow line separates null (benign) and alternative (harmful) settings. Testing procedures that rely on variance-adaptive confidence bounds (CBs) have more power. (b) Average sample size from the target that was needed to reject the null. Tighter concentration results allow to raise an alarm after processing less samples.

Tracking beyond the i.i.d. setting (distribution *drift*). Here we consider testing for an increase in the running risk ((4.2)). First, we fix $\pi_1^T = 0.75$ and keep the data generation pipeline as before, that is, the target data are still sampled in an i.i.d. fashion. We compare lower confidence bounds on the target risk studied before against the conjugate-mixture empirical-Bernstein (CM-EB) bound (Howard et al., 2021). We note that this bound on the running mean of the random variables does not have a closed-form and has to be computed numerically. On Figure 4.4a, we illustrate that the lower confidence bounds based on betting are generally tighter only for a small number of samples. Similar results hold for the Brier score as a target metric (see Appendix C.6.1).

Next, we lift the i.i.d. assumption by modifying the data generation pipeline: starting with $\pi_1^T = 0.25$, we increase π_1^T by 0.1 after sampling each 200 instances, until it reaches the value 0.85. It makes CM-EB the only valid lower bound on the running risk on the target domain. The results of running the framework for this setting are presented on Figure 4.4b.

4.3.2 Real Data

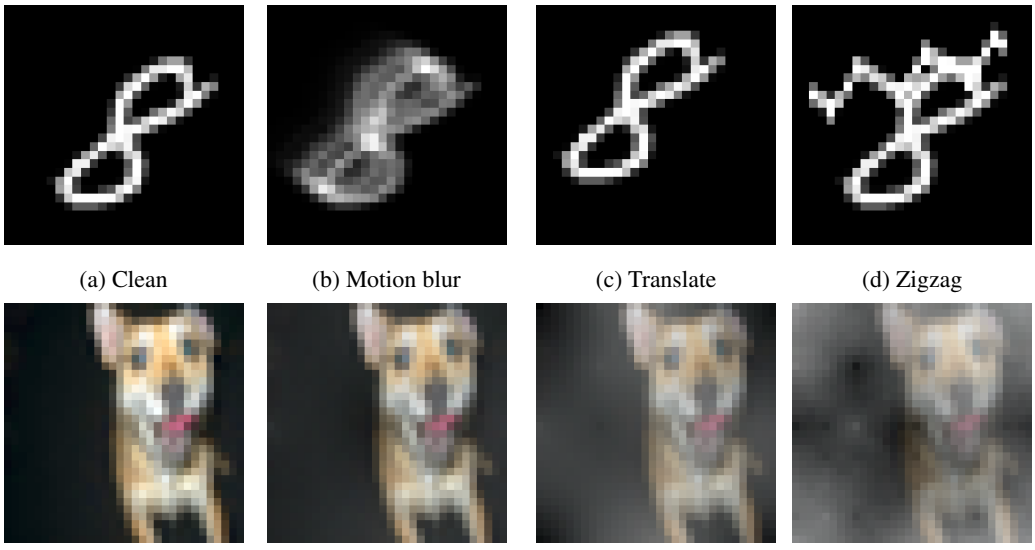
In deep learning, out-of-distribution robustness is often assessed based on a model performance gap between the original data (used for training) and data to which various perturbations are applied. We focus on two image classification datasets with induced corruptions: MNIST-C (Mu and Gilmer, 2019) and CIFAR-10-C (Krizhevsky, 2009; Hendrycks and Dietterich, 2019). We illustrate an example of a clean MNIST image on Figure 4.5a along with its corrupted versions after applying *motion blur*, blur along a random direction (Figure 4.5b), *translate*, affine transformation along a random direction (Figure 4.5c), and *zigzag*, randomly oriented zigzag over an image



(a) (b)

Figure 4.4: (a) Different lower confidence bounds (LCB) on the target risk under the i.i.d. assumption. Betting-based LCB is only tighter than conjugate-mixture empirical-Bernstein (CM-EB) for a small number of samples. (b) Under distribution drift, only CM-EB performs estimation of the running risk. The resulting test consistently detects a harmful increase in the running risk.

(Figure 4.5d). For CIFAR-10-C, we consider corrupting original images by applying the *fog* effect with 3 levels of severity as illustrated on the bottom row of Figure 4.5. For both cases, clean or corrupted test samples are passed as input to networks trained on clean data. While corruptions are visible to the human eye, one might still hope that they will not significantly hurt classification performance. We use the betting-based confidence bounds on the source and conjugate-mixture empirical-Bernstein confidence bounds on the target domain.



(a) Clean (b) Motion blur (c) Translate (d) Zigzag

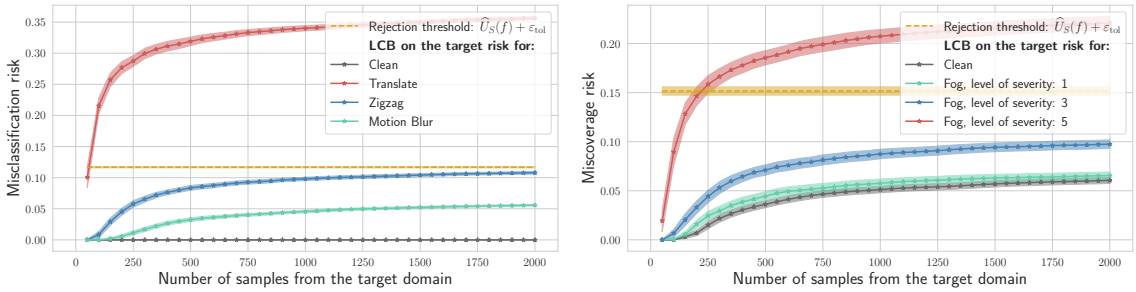
(e) Clean (f) Fog, severity: 1 (g) Fog, severity: 3 (h) Fog, severity: 5

Figure 4.5: Examples of MNIST-C ((a)–(d)) and CIFAR-10-C ((e)–(h)) images.

Tracking the risk of a point predictor on MNIST-C dataset. We train a shallow CNN on clean MNIST data and run the framework testing whether the misclassification risk increases by 10%, feeding the network with data in batches of 50 points either from the original or shifted distributions. Details regarding the network architecture and

the training process are given in Appendix C.7. On Figure 4.6a, we illustrate the results after running the procedure 50 times for each of the settings. The horizontal dashed line defines the rejection threshold that has been computed using the source data: once the lower bound on the target risk (solid lines) exceeds this value, the null hypothesis is rejected. When passing clean MNIST data as input, we do not observe degrading performance. Further, applying different corruptions leads to both benign and harmful shifts. While to the human eye the translate effect is arguably the least harmful one, it is the most harmful to the performance of a network. Such observation is also consistent with findings of Mu and Gilmer (2019) who observe that if trained on clean MNIST data without any data augmentation, CNNs tend to fail under this corruption. We validate this observation by retraining the network several times in Appendix C.7.

Tracking the risk of a set-valued predictor on CIFAR-10-C dataset. For high-consequence settings, building accurate models only can be insufficient as it is crucial to quantify uncertainty in predictions. One way to proceed is to output a set of candidate labels for each point as a prediction. The goal could be to cover the correct label of a test point with high probability (Vovk et al., 2005) or control other notions of risk (Bates et al., 2021). We follow Bates et al. (2021) who design a procedure that uses a holdout set for tuning the parameters of a wrapper built on top of the original model which, under the i.i.d. assumption, is guaranteed to have low risk with high probability (see Appendix C.7 for details). Below, we build a wrapper around a ResNet-32 model that controls the miscoverage risk ((C.8)) at level 0.1 with probability at least 0.95. For each run, CIFAR-10 test set is split at random into three folds used for: (a) learning a wrapper (1000 points), (b) estimating upper confidence bound on the miscoverage risk on the source (1000 points), and (c) evaluation purposes on either clean or corrupted images. We take $\varepsilon_{\text{tol}} = 0.05$, that is, 5% drop in coverage is treated as significant. Figure 4.6b illustrates that only the most intense level of fog is consistently harmful to coverage. We also consider setting a lower prescribed miscoverage level (0.05) for the set-valued predictor (see Appendix C.7). When larger prediction sets are produced, adding fog to images becomes less harmful.



(a) (b)

Figure 4.6: (a) Performance of the testing framework on MNIST-C dataset. Only the translation effect is consistently harmful to the classification performance of a CNN trained on clean data. (b) Performance of the testing framework on CIFAR-10-C dataset. Only the most severe version of the fog lead to a significant degradation in performance measured by a decrease in coverage of a set-valued predictor trained on top of a model trained on clean data.

4.4 Conclusion

An important component of building reliable machine learning systems is making them alarm a user when potentially unsafe behavior is observed, instead of allowing them to fail silently. Ideally, a warning should be displayed when critical changes affecting model performance are present, e.g., a significant degradation of the target performance metrics, like accuracy or calibration. In this work, we considered one particular failure scenario of deployed models—presence of distribution shifts. Relying solely on point estimators of the performance metrics ignores uncertainty in the evaluation, and thus fails to represent a theoretically grounded approach. We developed a set of tools for deciding whether the performance of a model on the test data becomes significantly worse than the performance on the training data in a data-adaptive way. The proposed framework based on performing sequential estimation requires observing true labels for test data (possibly, in a delayed fashion). Across various types of distribution shifts considered in this work, it demonstrated promising empirical performance for differentiating between harmful and benign ones.

Part II

Assumption-Light Predictive Uncertainty Quantification

Chapter 5

Distribution-Free Binary Classification: Prediction Sets, Confidence Intervals and Calibration

5.1 Introduction

Let $\mathcal{X}$ and $\mathcal{Y} = \{0, 1\}$ denote the feature and label spaces for binary classification. Consider a predictor $f : \mathcal{X} \to \mathcal{Z}$ that produces a prediction in some space $\mathcal{Z}$. If $\mathcal{Z} = \{0, 1\}$, f corresponds to a point prediction for the class label, but often class predictions are based on a 'scoring function'. Examples are, $\mathcal{Z} = \mathbb{R}$ for SVMs, and $\mathcal{Z} = [0, 1]$ for logistic regression, random forests with class probabilities, or deep models with a softmax top layer. In such cases, a higher value of $f(X)$ is often interpreted as higher belief that $Y = 1$. In particular, if $\mathcal{Z} = [0, 1]$, it is tempting to interpret $f(X)$ as a probability, and hope that

$$f(X) \approx \mathbb{P}(Y = 1 \mid X). \tag{5.1}$$

However, such hope is unfounded, and in general (5.1) will be far from true without strong distributional assumptions, which may not hold in practice. Valid uncertainty estimates that are related to (5.1) can be provided, but ML models do not satisfy these out of the box. This paper discusses three notions of uncertainty quantification: calibration, prediction sets (PS) and confidence intervals (CI), defined next. A function $f : \mathcal{X} \to [0, 1]$ is said to be (perfectly) calibrated if

$$\mathbb{E}\left[Y \mid f(X) = a\right] = a \qquad \text{a.s. for all } a \text{ in the range of } f. \tag{5.2}$$

Define $calL \equiv \{\{0\}, \{1\}, \{0,1\}, \emptyset\}$ and fix $\alpha \in (0,1)$. A function $S : \mathcal{X} \to \mathcal{L}$ is a $(1-\alpha)$-PS if

$$\mathbb{P}(Y \in S(X)) \geq 1 - \alpha. \tag{5.3}$$

Finally, let $\mathcal{I}$ denote the set of all subintervals of $[0,1]$. A function $C : \mathcal{X} \to \mathcal{I}$ is a $(1-\alpha)$-CI if

$$\mathbb{P}(\mathbb{E}[Y \mid X] \in C(X)) \geq 1 - \alpha. \tag{5.4}$$

All three notions are 'natural' in their own sense, but also different at first sight. We show that they are in fact tightly connected (see Figure 5.1), and focus on the implications of this result for calibration. Our analysis is in the distribution-free setting, that is, we are concerned with understanding what kinds of valid uncertainty quantification is possible without distributional assumptions on the data.

Our work primarily extends the ideas of Vovk et al. (2005, Section 5) and Barber (2020). We also discuss Platt scaling (Platt, 1999), binning (Zadrozny and Elkan, 2001) and the recent work of Vaicenavicius et al. (2019). Other related work is cited as needed, and further discussed in Section 5.5. All proofs appear ordered in the Appendix.

Notation: Let P denote any distribution over $\mathcal{X} \times \mathcal{Y}$. In practice, the available labeled data is often split randomly into the *training set* and the *calibration set*. Typically, we use n to denote the number of calibration data points, so $\{(X_i, Y_i)\}_{i \in [n]}$ is the calibration data, where we use the shorthand $[a] := \{1, 2, \ldots a\}$. A prototypical test point is denoted (X_{n+1}, Y_{n+1}). All data are drawn i.i.d. from P, denoted succinctly as $\{(X_i, Y_i)\}_{i \in [n+1]} \sim P^{n+1}$. As above, random variables are denoted in upper case. The learner observes realized values of all random variables (X_i, Y_i), except Y_{n+1}. (All sets and functions are implicitly assumed to be measurable.)

5.2 Calibration, Confidence Intervals and Prediction Sets

Calibration captures the intuition of (5.1) but is a weaker requirement, and was first studied in the meteorological literature for assessing probabilistic rain forecasts (Brier, 1950; Sanders, 1963; Murphy and Epstein, 1967; Dawid, 1982). Murphy and Epstein (1967) described the ideal notion of calibration, called *perfect calibration* (5.2), which has also been referred to as *calibration in the small* (Vovk and Petej, 2014), or sometimes simply as *calibration* (Guo et al., 2017; Vaicenavicius et al., 2019; Dawid, 1982). The types of functions that can achieve perfect calibration can be succinctly captured as follows.

Proposition 5. *A function $f : \mathcal{X} \to [0,1]$ is perfectly calibrated if and only if there exists a space $\mathcal{Z}$ and a function $g : \mathcal{X} \to \mathcal{Z}$, such that*

$$f(x) = \mathbb{E}[Y \mid g(X) = g(x)] \quad \textit{almost surely } P_X. \tag{5.5}$$

(If parsing (5.5) is tricky: to evaluate f at x, first set $g(x) \equiv z$, then calculate $\mathbb{E}[Y \mid g(X) = z]$.) Vaicenavicius et al. Vaicenavicius et al. (2019) stated and gave a short proof for the 'only if' direction. While the other direction is also

straightforward, together they lead to an appealingly simple and complete characterization. The proof of Proposition 5 is in Appendix D.1.

It is helpful to consider two extreme cases of Proposition 5. First, setting g to be the identity function yields that the Bayes classifier $\mathbb{E}[Y|X]$ is perfectly calibrated. Second, setting $g(\cdot)$ to any constant implies that $\mathbb{E}[Y]$ is also a perfect calibrator. Naturally, we cannot hope to estimate the Bayes classifier without assumptions, but even the simplest calibrator $\mathbb{E}[Y]$ can only be approximated in finite samples. Since Proposition 5 states that calibration is possible iff the RHS of (5.5) is known exactly for some g, perfect calibration is impossible in practice. Thus we resort to satisfying the requirement (5.2) approximately, which is implicitly the goal of many empirical calibration techniques.

Definition 3 (Approximate calibration)**.** *A predictor $f : \mathcal{X} \to [0,1]$ is (ε, α)-approximately calibrated for some $\alpha \in (0,1)$ and a function $\varepsilon : [0,1] \to [0,1]$ if with probability at least $1-\alpha$, we have*

$$|\mathbb{E}[Y|f(X)] - f(X)| \leq \varepsilon(f(X)). \tag{5.6}$$

Note that when the definition is applied to a test point (X_{n+1}, Y_{n+1}), there may be two sources of randomness in $\mathbb{E}[Y_{n+1} \mid f(X_{n+1})]$: the randomness in the test point, as well as randomness in f—the latter may be statistical randomness via learning on the training data, or algorithmic randomness used to train f. There can also be randomness in ε. All probabilities and expectations in this paper should be viewed through this lens. In practice, calibration is often achieved via a post-processing step. Hence, with increasing amount of the calibration data, one might hope that ε in Definition 3 vanishes to 0. We formalize this below.

Definition 4 (Asymptotic calibration)**.** *A sequence of predictors $\{f_n\}_{n\in\mathbb{N}}$ from $\mathcal{X} \to [0,1]$ is asymptotically calibrated at level $\alpha \in (0,1)$ if there exists a sequence of functions $\{\varepsilon_n\}_{n\in\mathbb{N}}$ such that f_n is (ε_n, α)-approximately calibrated for every n, and $\varepsilon_n(f_n(X_{n+1})) = o_P(1)$.*

We will show that the notions of approximate and asymptotic calibration are related to prediction sets (5.3) and confidence intervals (5.4). PSs and CIs are only 'informative' if the sets or intervals produced by them are small: confidence intervals are measured by their length (denoted as $|C(\cdot)|$), and prediction sets are measured by their diameter ($\text{diam}(S(\cdot)) := |\text{convex hull}(S(\cdot))|$). Observe that for binary classification, the diameter of a PS is either 0 or 1.

For a given distribution, one might expect prediction sets to have a larger diameter than the length of the confidence intervals, since we want to cover the actual value of Y_{n+1} and not its (conditional) expectation. As an example, if $\mathbb{E}[Y|X=x] = 0.5$ for every x, then the shortest possible confidence interval is $(0.5, 0.5]$ whose diameter is 0. However, a valid $(1-\alpha)$-PS has no choice but to output $\{0,1\}$ for at least $(1-2\alpha)$ fraction of the points (and a random guess for the other 2α fraction), and thus must have expected diameter $\geq 1-2\alpha$ even in the limit of infinite data.

Recently, Barber (2020) built on an earlier result of Vovk et al. (2005) to show that if an algorithm provides an interval C which is a $(1-\alpha)$-CI for all product distributions P^{n+1} (of the training data and test-point), then $S := C \cap \{0,1\}$ is also a $(1-\alpha)$-PS whenever P is a nonatomic distribution. An immediate implication is that $C(\cdot)$ must always contain one of the end-points 0 or 1 with probability $1-\alpha$. Since this implication holds for all distributions P, including the one with $\mathbb{E}[Y|X] \equiv 0.5$ discussed above, it implies that distribution-free CIs must necessarily be wide, and in particular their length cannot shrink to 0 as $n \to \infty$. This can be treated as an impossibility result for the existence of (distribution-free) informative CIs.

One way to circumvent these impossibilities is to consider CIs for functions with 'lower resolution' than $\mathbb{E}[Y|X]$. To this end, we introduce a notion of a CI or PS 'with respect to f' (w.r.t.f). As we discuss in Section 5.3 (and Section 5.3.1 in particular), these notions are connected to calibration.

Definition 5 (CI or PS w.r.t. f)**.** *A function $C : \mathcal{Z} \to \mathcal{I}$ is a $(1-\alpha)$-CI with respect to $f : \mathcal{X} \to \mathcal{Z}$ if*

$$\mathbb{P}(\mathbb{E}[Y \mid f(X)] \in C(f(X))) \geq 1-\alpha. \tag{5.7}$$

Analogously, a function $S : \mathcal{Z} \to \mathcal{L}$ is a $(1-\alpha)$-PS with respect to $f : \mathcal{X} \to \mathcal{Z}$ if

$$\mathbb{P}(Y \in S(f(X))) \geq 1-\alpha. \tag{5.8}$$

When instantiated for a test point (X_{n+1}, Y_{n+1}), the probability in definitions (5.7) and (5.8) is not only over the test point, but also over the randomness in the pair (f, C) or (f, S), which are usually learned on labeled data. In order to produce PSs and CIs, one typically fixes a function f learned on an independent split of the labeled data, and considers learning a C or S that provides guarantees (5.7) and (5.8). For example, S can be produced using inductive conformal techniques Proedrou et al. (2002); Papadopoulos et al. (2002). In this case, C or S would be random as well; to make this explicit, we often denote C or S as $\widehat{C}_n$ or $\widehat{S}_n$.

5.3 Relating Notions of Distribution-free Uncertainty Quantification

As preluded to above, we consider a standard setting for valid distribution-free uncertainty quantification where the 'training' data is used to learn a scoring function $f : \mathcal{X} \to \mathcal{Z}$ and then held-out data 'calibration' data is used to estimate uncertainty. We establish that in this setting, the notions of calibration, PSs and CIs are closely related. Figure 5.1 summarizes this section's takeaway message. Here, and in the rest of the section, if P is the distribution of data, then we denote the distribution of the random variable $Z = f(X)$ as $P_{f(X)}$.

In Section 5.3.1, we show that if an algorithm provides a CI, it can be used to provide a calibration guarantee and vice-versa (Theorem 5.1). This result is true even if the CI and calibration guarantees are not assumption-free. Section 5.3.2 shows that for all distributions P such that $P_{f(X)}$ is nonatomic, if an algorithm constructs a distribution-free CI with respect to f, then it can be used to construct a distribution-free PS with respect to f (Theorem 5.2). This

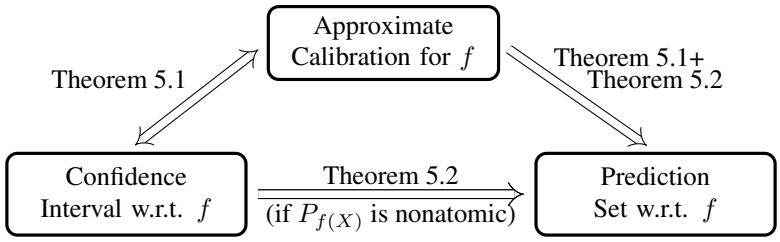


Figure 5.1: Relationship between notions of distribution-free uncertainty quantification.

result might seem surprising since one typically expects the length of CIs to shrink to 0 in the limit of infinite data, whereas PSs have a fixed distribution-dependent lower bound on their diameter. Connecting our results, we infer the key impossibility result for asymptotic calibration in Section 5.3.3 (Theorem 5.3). Informally, our result shows that for a large class of standard scoring functions f (such as logistic regression, deep networks with a final softmax layer, SVMs), it is impossible to achieve distribution-free asymptotic calibration without a 'discretization' step. Parametric schemes such as Platt scaling (Platt, 1999) do not perform such discretization and thus cannot lead to distribution-free calibration. To complement this lower bound, we provide calibration guarantees for one possible discretization step (histogram binning) in Section 5.4.

5.3.1 Relating Calibration and Confidence Intervals

Given a predictor f that is (ε, α)-approximately calibrated, there is a trivial way to construct a function C that is a $(1-\alpha)$-CI: for $x \in \mathcal{X}$,

$$\underbrace{|\mathbb{E}\left[Y \mid f(x)\right] - f(x)| \leq \varepsilon(f(x))}_{\text{calibration}} \implies \underbrace{\mathbb{E}\left[Y \mid f(x)\right] \in C(f(x))}_{\text{CI w.r.t. } f} := [f(x) - \varepsilon(f(x)), f(x) + \varepsilon(f(x))]. \tag{5.9}$$

On the other hand, given C that is a $(1-\alpha)$-CI with respect to f, define for $z \in \text{Range}(f)$ the left-endpoint, right-endpoint and midpoint functions respectively:

$$u_C(z) := \sup\left\{g : g \in C(z)\right\},\ l_C(z) := \inf\left\{g : g \in C(z)\right\},\ m_C(z) := (u_C(z) + l_C(z))/2. \tag{5.10}$$

Consider the midpoint $m_C(f(x))$ as a 'corrected' prediction for $x \in \mathcal{X}$:

$$\widetilde{f}(x) := m_C(f(x)),\ x \in \mathcal{X}, \tag{5.11}$$

and let $\varepsilon(\cdot) = \sup_{z \in \text{Range}(f)}\left\{|C(z)|/2\right\}$ be the function returning the largest interval radius. Then $\widetilde{f}$ is (ε, α)-approximately calibrated for a non-trivial ε. These claims are formalized next.

Theorem 5.1. *Fix any $\alpha \in (0,1)$. Let $f : \mathcal{X} \to [0,1]$ be a predictor that is (ε, α)-approximately calibrated for some function ε. Then the function C in* (5.9) *is a $(1-\alpha)$-CI with respect to f.*

Conversely, fix a scoring function $f : \mathcal{X} \to \mathcal{Z}$. If C is a $(1-\alpha)$-CI with respect to f, then the predictor $\widetilde{f}$ in (5.11) *is (ε, α)-approximately calibrated for $\varepsilon(\cdot) = \sup_{z \in Range(f)} \{|C(z)|/2\}$.*

The proof is in Appendix D.2. An important implication of Theorem 5.1 is that having a sequence of predictors that is asymptotically calibrated yields a sequence of confidence intervals with vanishing length as $n \to \infty$. This is formalized in the following corollary, also proved in Appendix D.2.

Corollary 5.1.1. *Fix any $\alpha \in (0,1)$. If a sequence of predictors $\{f_n\}_{n \in \mathbb{N}}$ is asymptotically calibrated at level α, then construction* (5.9) *yields a sequence of functions $\{C_n\}_{n \in \mathbb{N}}$ such that each C_n is a $(1-\alpha)$-CI with respect to f_n and $|C_n(f_n(X_{n+1}))| = o_P(1)$.*

Next, we show that for a large class of scoring functions, CIs and PSs are also related in the distribution-free setting. This connection along with Corollary 5.2.1 (below) leads to an impossibility result for distribution-free asymptotic calibration for certain functions f (Theorem 5.3 in Section 5.3.3).

5.3.2 Relating Distribution-free Confidence Intervals and Prediction Sets

Suppose a function satisfies a CI guarantee with respect to f no matter what the data-generating distribution P is. We show that such a function would also provide a PS guarantee for all P such that $P_{f(X)}$ is nonatomic. To write our theorem, we define the 'discretize' function to transform a confidence interval C to a prediction set: $\text{disc}(C) := C \cap \{0,1\} \subseteq \mathcal{L}$. In the following theorem, the CI and PS guarantees provided (per equations (5.7) and (5.8)) are to be understood as marginal over both the calibration and test-data. To make this explicit, we denote the CI function as $\widehat{C}_n$.

Theorem 5.2. *Fix $f : \mathcal{X} \to \mathcal{Z}$ and $\alpha \in (0,1)$. If $\widehat{C}_n$ is a $(1-\alpha)$-CI with respect to f for all distributions P, then disc$(\widehat{C}_n)$ is a $(1-\alpha)$-PS with respect to f for all distributions P for which $P_{f(X)}$ is nonatomic.*

The proof is in Appendix D.2. It adapts the proof of Barber (2020, Theorem 1). Their result connects the notions of CI and PS, but not with respect to f (like in equations (5.3), (5.4)). By adapting the result for CIs and PSs with respect to f, and using Theorem 5.1, we are able to relate CIs and PSs to calibration and use this to prove an impossibility result for asymptotic calibration. This is done in the proof of Theorem 5.3 in the Section 5.3.3. A corollary of Theorem 5.2 that is used in Theorem 5.3 (but is also important on its own) is stated next.

Corollary 5.2.1. *Fix $f : \mathcal{X} \to \mathcal{Z}$ and $\alpha \in (0,1)$. If $\widehat{C}_n$ is a $(1-\alpha)$-CI with respect to f for all P, and there exists a P such that $P_{f(X)}$ is nonatomic, then we can construct a distribution Q such that*

$$\mathbb{E}_{Q^{n+1}} |\widehat{C}_n(f(X_{n+1}))| \geq 0.5 - \alpha.$$

The proof is in Appendix D.2. For a given f, the bound in the corollary needs existence of P such that $P_{f(X)}$ is nonatomic. These f are characterized in the discussion after Corollary 5.2.2 (Section 5.3.3), and formally in the proof

of Theorem 5.3. One expects the length of a confidence interval to vanish as $n \to \infty$. Corollary 5.2.1 shows that this is impossible in a distribution-free manner for certain f.

5.3.3 Necessary Condition for Distribution-Free Asymptotic Calibration

The characterization of calibration in Proposition 5 shows that a function f is a calibrated probabilistic classifier if and only if it takes the form (5.5) for some function g, and in particular f is calibrated by defining $g = f$. Observe that for the purposes of calibration, the actual values taken by f are only as informative as the *partition* of $\mathcal{X}$ provided by its level sets. Denote this partition as $\{\mathcal{X}_z\}_{z\in\mathcal{Z}}$, where $\mathcal{X}_z = \{x \in \mathcal{X} : f(x) = z\}$. Then we may equivalently rewrite (5.5) as identifying values $\{f_z\}_{z\in\mathcal{Z}}$ where $f_z = P(Y_{n+1} = 1 \mid X_{n+1} \in \mathcal{X}_z)$. This allows us to re-characterize calibration as follows.

Corollary 5.2.2 (to Proposition 5)**.** *Any calibrated classifier f is characterized by a partition of $\mathcal{X}$ into subsets $\{\mathcal{X}_z\}_{z\in\mathcal{Z}}$ and corresponding conditional probabilities $\{f_z\}_{z\in\mathcal{Z}}$ for some index set $\mathcal{Z}$.*

Corollary 5.1.1 shows that asymptotic calibration allows construction of CIs whose lengths vanish asymptotically. Corollary 5.2.1 shows however that asymptotically vanishing CIs are impossible (without distributional assumptions) for f if there exists a distribution P such that $P_{f(X)}$ is nonatomic. Consequently asymptotic calibration is also impossible for such f. If $\mathcal{Z}$ is countable, then by the axioms of probability, $\sum_{z\in\mathcal{Z}} \mathbb{P}(X \in \mathcal{X}_z) = \mathbb{P}(X \in \mathcal{X}) = 1$, and so $\mathbb{P}(X \in \mathcal{X}_z) \neq 0$ for at least some z. Thus $P_{f(X)}$ cannot be nonatomic for any P. On the other hand, if $\mathcal{Z}$ is uncountable we can show that there always exists a P such that $P_{f(X)}$ is nonatomic. Hence distribution-free asymptotic calibration is impossible for such f. This argument is formalized in the following theorem. In the statement, we used $\mathcal{X}^{(f)}$ to denote the partition that a function f induces on $\mathcal{X}$, and we use $|\mathcal{X}^{(f)}|$ to denote its cardinality (which may be infinite). Also $\aleph_0$ denotes the largest cardinality of a countable set, which corresponds to the cardinality of $\mathbb{N}$. The proof of the following theorem is in Appendix D.2.

Theorem 5.3. *Let $\alpha \in (0, 0.5)$ be a fixed threshold. If a sequence of scoring functions $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated at level α for every distribution P then*

$$\limsup_{n\to\infty} |\mathcal{X}^{(f_n)}| \leq \aleph_0.$$

In words, the cardinality of the partition induced by f_n must be at most countable for large enough n. The following phrasing is convenient: f is said to lead to a *fine partition* of $\mathcal{X}$ if $|\mathcal{X}^{(f)}| > \aleph_0$. Then, for the purposes of distribution-free asymptotic calibration, Theorem 5.3 necessitates us to consider f that do not lead to fine partitions. Popular scoring functions such as logistic regression, deep neural-nets with softmax output and SVMs lead to continuous f that induce fine partitions of $\mathcal{X}$ and thus cannot be asymptotically calibrated without distributional assumptions.

This impossibility result can be extended to many parametric calibration schemes that 'recalibrate' an existing f through a wrapper $h_n : \mathcal{Z} \to [0, 1]$ learned on the calibration data, with the goal that $h_n \circ f$ is nearly calibrated:

$\mathbb{E}[Y \mid h_n(f(X))] \approx h_n(f(X))$. For instance, consider methods like Platt scaling (Platt, 1999), temperature scaling (Guo et al., 2017) and beta calibration (Kull et al., 2017). Each of these methods learns a continuous and monotonic* (hence bijective) wrapper h_n, and thus $\mathbb{E}[Y \mid h_n(f(X))] = \mathbb{E}[Y \mid f(X)]$. If h_n is a good calibrator, we would have $\mathbb{E}[Y \mid f(X)] \approx h_n(f(X))$. One way to formalize this is to consider whether an interval around $h_n(f(X))$ is a CI for $\mathbb{E}[Y \mid f(X)]$. In other words — does there exist a function $\varepsilon_n : [0,1] \to [0,1]$ such that for every distribution P,

$$\widetilde{C}_n(f(X)) := [h_n(f(X)) - \varepsilon_n(h_n(f(X))), h_n(f(X)) + \varepsilon_n(h_n(f(X)))]$$

is a $(1-\alpha)$-CI with respect to f and $\varepsilon_n(h_n(f(X))) = o_P(1)$? Theorem 5.3 shows that this is impossible if f leads to a fine partition of $\mathcal{X}$, irrespective of the properties of h_n. Thus the aforementioned parametric calibration methods cannot lead to asymptotic calibration in general (that is, without further distributional assumptions). It is likely that the implications of our results also apply to other continuous parametric methods that are not necessarily monotonic, as well as calibration schemes that directly aim to learn a calibrated predictor instead of post-hoc calibration or recalibration.

A well-known calibration method that does not produce a fine partition of $\mathcal{X}$ is histogram binning (Zadrozny and Elkan, 2001). In Section 5.4, we analyze histogram binning and show that any scoring function can be 'binned' to achieve distribution-free calibration. We explicitly quantify the finite-sample approximate calibration guarantees that automatically also lead to asymptotic calibration. We also discuss calibration in the online setting and calibration under covariate shift.

5.4 Achieving Distribution-free Approximate Calibration

In Section 5.4.1, we prove a distribution-free approximate calibration guarantee given a fixed partitioning of the feature space into finitely many sets. This calibration guarantee also leads to asymptotic calibration. In Section 5.4.2, we discuss a natural method for obtaining such a partition using sample-splitting, called histogram binning. Histogram binning inherits the bound in Section 5.4.1. This shows that binning schemes lead to distribution-free approximate calibration. In Section 5.4.3 and 5.4.4 we discuss extensions of this scheme to adaptive sampling and covariate shift respectively.

5.4.1 Distribution-free Calibration Given a Fixed Sample-space Partition

Suppose we have a fixed partition of $\mathcal{X}$ into B regions $\{\mathcal{X}_b\}_{b\in[B]}$, and let $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$ be the expected label probability in region $\mathcal{X}_b$. Denote the partition-identity function as $\mathcal{B} : \mathcal{X} \to [B]$ where $\mathcal{B}(x) = b$ if and only if $x \in \mathcal{X}_b$.

*This assumes that the parameters satisfy natural constraints as discussed in the original papers: $a, b \geq 0$ for beta scaling with at least one of them nonzero, $A < 0$ for Platt scaling and $T > 0$ for temperature scaling.

Given a calibration set $\{(X_i, Y_i)\}_{i\in[n]}$, let $\widehat{s}_b := |\{i \in [n] : \mathcal{B}(X_i) = b\}|$ be the number of points from the calibration set that belong to region $\mathcal{X}_b$. In this subsection, we assume that $\widehat{s}_b \geq 1$ (in Section 5.4.2 we show that the partition can be constructed to ensure that $\widehat{s}_b$ is $\Omega(n/B)$ with high probability). Define

$$\widehat{\pi}_b := \frac{1}{\widehat{s}_b} \sum_{i:\mathcal{B}(X_i)=b} Y_i \qquad \text{and} \qquad \widehat{V}_b := \frac{1}{\widehat{s}_b} \sum_{i:\mathcal{B}(X_i)=b} (Y_i - \widehat{\pi}_b)^2 \tag{5.12}$$

as the empirical average and variance of the Y values in a partition. We now deploy an empirical Bernstein bound Audibert et al. (2007) to produce a confidence interval for π_b.

Theorem 5.4. *For any $\alpha \in (0, 1)$, with probability at least $1 - \alpha$,*

$$|\pi_b - \widehat{\pi}_b| \leq \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3\ln(3B/\alpha)}{\widehat{s}_b}, \quad \textit{simultaneously for all } b \in [B].$$

The theorem is proved in Appendix D.3. Using the crude deterministic bound $\widehat{V}_b \leq 1$ we get that the length of the confidence interval for partition b is $O(1/\sqrt{\widehat{s}_b})$. However, if for some b, $\mathcal{X}_b$ is highly informative or homogeneous in the sense that π_b is close to 0 or 1, we expect $\widehat{V}_b \ll 1$. In this case, Theorem 5.4 *adapts* and provides an $O(1/\widehat{s}_b)$ length interval for π_b. Let $b^\star = \arg\min_{b\in[B]} \widehat{s}_b$ denote the index of the region with the minimum number of calibration examples.

Corollary 5.4.1. *For $\alpha \in (0, 1)$, the function $f_n(x) := \widehat{\pi}_{\mathcal{B}(x)}$ is (ε, α)-approximately calibrated with*

$$\varepsilon(\cdot) = \sqrt{\frac{\widehat{V}_{b^\star} \ln(3B/\alpha)}{2\widehat{s}_{b^\star}}} + \frac{3\ln(3B/\alpha)}{2\widehat{s}_{b^\star}}.$$

Thus, $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated at level α.

The proof is in Appendix D.3. Thus, any finite partition of $\mathcal{X}$ can be used for asymptotic calibration. However, the finite sample guarantee of Corollary 5.4.1 can be unsatisfactory if the sample-space partition is chosen poorly, since it might lead to small $\widehat{s}_{b^\star}$. In Section 5.4.2, we present a data-dependent partitioning scheme that provably guarantees that $\widehat{s}_{b^\star}$ scales as $\Omega(n/B)$ with high probability.

5.4.2 Identifying a Data-dependent Partition using Sample Splitting

Here, we describe ways of constructing the partition $\{\mathcal{X}_b\}_{b\in[B]}$ through histogram binning Zadrozny and Elkan (2001). Binning uses a sample splitting strategy, where the partition is learned on the first part and $\{\widehat{\pi}_b\}_{b\in[B]}$ are estimated on the second part. Formally, the labeled data is split at random into the training set $\mathcal{D}_{\text{tr}}$ and calibration set $\mathcal{D}_{\text{cal}}$. Then $\mathcal{D}_{\text{tr}}$ is used to train an underlying scoring classifier $g : \mathcal{X} \to [0, 1]$ (in general the range of the classifier could be any

interval of $\mathbb{R}$ but for simplicity we describe it for $[0,1]$). The classifier g usually does not satisfy a valid calibration guarantee out-of-the-box but can be calibrated using binning as follows.

A *binning scheme* $\mathcal{B}$ is any partition of $[0,1]$ into B non-overlapping intervals $I_1, \ldots, I_B$, such that $\bigcup_{b\in[B]} I_b = [0,1]$ and $I_b \cap I_{b'} = \emptyset$ for $b \neq b'$. $\mathcal{B}$ and g induce a partition of $\mathcal{X}$ as follows:

$$\mathcal{X}_b = \{x \in \mathcal{X} : \ g(x) \in I_b\}\,, \ b \in [B]. \tag{5.13}$$

The simplest binning scheme corresponds to *fixed-width binning*. In this case, bins have the form

$$I_i = \left[\frac{i-1}{B}, \frac{i}{B}\right), i = 1, \ldots, B-1 \ \text{ and } \ I_B = \left[\frac{B-1}{B}, 1\right].$$

However, fixed-width binning suffers from the drawback that there may exist bins with very few calibration points (low $\widehat{s}_b$), while other bins may get many calibration points. For bins with low $\widehat{s}_b$, the $\widehat{\pi}_b$ estimates cannot be guaranteed to be well calibrated, since the bound of Theorem 5.4 could be large. To remedy this, we consider *uniform-mass binning*, which aims to guarantee that each region $\mathcal{X}_b$ contains approximately equal number of data points from the calibration set. This is done by estimating the empirical quantiles of $g(X)$. First, the calibration set $\mathcal{D}_{\text{cal}}$ is randomly split into two parts, $\mathcal{D}^1_{\text{cal}}$ and $\mathcal{D}^2_{\text{cal}}$. Then $\widehat{q}_j$ is simply defined as the (j/B)-th quantile of the empirical distribution of the values $\{g(X_i), i \in \mathcal{D}^1_{\text{cal}}\}$ for $j \in [B-1]$. Consequently, the bins are defined as:

$$I_1 = [0, \widehat{q}_1)\,, I_i = [\widehat{q}_{i-1}, \widehat{q}_i]\,, i = 2, \ldots, B-1 \ \text{ and } \ I_B = (\widehat{q}_{B-1}, 1]\,.$$

Next, only $\mathcal{D}^2_{\text{cal}}$ is used for calibrating the underlying classifier. Kumar et al. (2019) showed that uniform-mass binning provably controls the number of calibration samples that fall into each bin (see Appendix D.5.2). Building on their result, we show the following guarantee for $\widehat{s}_{b^\star} = \min_{b\in[B]} \widehat{s}_b$.

Theorem 5.5. *There exists a universal constant c such that if $\left|\mathcal{D}^1_{cal}\right| \geq cB\ln(2B/\alpha)$, then with probability at least $1-\alpha$,*

$$\widehat{s}_{b^\star} \geq \left|\mathcal{D}^2_{cal}\right|/2B - \sqrt{|\mathcal{D}^2_{cal}|\ln(2B/\alpha)/2},$$

Thus even if $|\mathcal{D}^1_{cal}|$ does not grow with n, as long as $|\mathcal{D}^2_{cal}| = \Omega(n)$, uniform-mass binning is approximately calibrated at level $(\widetilde{O}(\sqrt{B\ln(1/\alpha)/n}), \alpha)$, and hence also asymptotically calibrated for any $\alpha \in (0,1)$.

The proof is in Appendix D.3. In words, if we use a small number of points (independent of n) for uniform-mass binning, and the rest to estimate bin probabilities, we achieve (approximate/asymptotic) distribution-free calibration.

5.4.3 Distribution-free Calibration in the Online Setting

So far, we have considered the batch setting with a fixed calibration set of size n. However, often a practitioner might want to query additional calibration data until a desired confidence level is achieved. This is called the *online* or *adaptive* setting. In this case, the results of Section 5.4 are no longer valid since the number of calibration samples is unknown a priori and may even be dependent on the data. In order to quantify uncertainty in the online setting, we use *time-uniform* concentration bounds Howard et al. (2021, 2020); these hold simultaneously for all possible values of the calibration set size $n \in \mathbb{N}$.

Fix a partition of $\mathcal{X}$, $\{\mathcal{X}_b\}_{bn\in[B]}$. For some value of n, let the calibration data be given as $\mathcal{D}_{\text{cal}}^{(n)}$. We use the superscript notation to emphasize the dependence on the current size of the calibration set. Let $\{(X_i^b, Y_i^b)\}_{i\in[\widehat{s}_b^{(n)}]}$ be examples from the calibration set that fall into the partition $\mathcal{X}_b$, where $\widehat{s}_b^{(n)} := |\{i \in [n] : \mathcal{B}(X_i) = b\}|$ is the total number of points that are mapped to $\mathcal{X}_b$. Let the empirical label average and cumulative (unnormalized) empirical variance be denoted as

$$\overline{Y}_t^b = \frac{1}{t}\sum_{i=1}^{t} Y_i^b, \qquad \widehat{V}_b^+ = 1 \vee \sum_{i=1}^{\widehat{s}_b^{(n)}} \left(Y_i^b - \overline{Y}_{i-1}^b\right)^2. \tag{5.14}$$

Note the normalization difference between $\widehat{V}_b^+$ and $\widehat{V}^b$ used in the batch setting. The following theorem constructs confidence intervals for $\{\pi_b\}_{b\in[B]}$ that are valid uniformly for any value of n.

Theorem 5.6. *For any $\alpha \in (0,1)$, with probability at least $1-\alpha$,*

$$|\pi_b - \widehat{\pi}_b| \le \frac{7\sqrt{\widehat{V}_b^+ \ln\left(1 + \ln \widehat{V}_b^+\right)} + 5.3 \ln\left(\frac{6.3B}{\alpha}\right)}{\widehat{s}_b^{(n)}}, \quad \textit{simultaneously for all } b \in [B] \textit{ and all } n \in \mathbb{N}. \tag{5.15}$$

Thus $\widehat{\pi}_b$ is asymptotically calibrated at any level $\alpha \in (0,1)$.

The proof is in Appendix D.3. Due to the crude bound: $\widehat{V}_b^+ \le \widehat{s}_b^{(n)}$, we can see that the width of confidence intervals roughly scales as $O(\sqrt{\ln(1+\ln \widehat{s}_b^{(n)})/\widehat{s}_b^{(n)}})$. In comparison to the batch setting, only a small price is paid for not knowing beforehand how many examples will be used for calibration.

5.4.4 Calibration under Covariate Shift

Here, we briefly consider the problem of calibration under covariate shift Shimodaira (2000). In this setting, calibration data $\{(X_i, Y_i)\}_{i\in[n]} \sim P^n$ is from a 'source' distribution P, while the test point is from a shifted 'target' distribution $(X_{n+1}, Y_{n+1}) \sim \widetilde{P} = \widetilde{P}_X \times P_{Y|X}$, meaning that the 'shift' occurs only in the covariate distribution while $P_{Y|X}$ does not change. We assume the likelihood ratio (LR)

$$w : \mathcal{X} \to \mathbb{R}; \quad w(x) := \mathrm{d}\widetilde{P}_X(x)/\mathrm{d}P_X(x)$$

is well-defined. The following is unambiguous: *if w is arbitrarily ill-behaved and unknown, the covariate shift problem is hopeless, and one should not expect any distribution-free guarantees*. Nevertheless, one can still make nontrivial claims using a 'modular' approach towards assumptions:

Condition (A): $w(x)$ is known exactly and is bounded.

Condition (B): an asymptotically consistent estimator $\widehat{w}(x)$ for $w(x)$ can be constructed.

We show the following: under Condition (A), a weighted estimator using w delivers approximate and asymptotic distribution-free calibration; under Condition (B), weighting with a plug-in estimator for w continues to deliver asymptotic distribution-free calibration. It is clear that Condition (B) will always require distributional assumptions: asymptotic consistency is nontrivial for ill-behaved w. Nevertheless, the above two-step approach makes it clear where the burden of assumptions lie: not with calibration step, but with the w estimation step. Estimation of w is a well studied problem in the covariate-shift literature and there is some understanding of what assumptions are needed to accomplish it, but there has been less work on recognizing the resulting implications for calibration. Luckily, many practical methods exist for estimating w given unlabeled samples from $\widetilde{P}_X$ (Bickel et al., 2007; Huang et al., 2007; Kanamori et al., 2009). In summary, if Condition (B) is possible, then distribution-free calibration is realizable, and if Condition (B) is not met (even with infinite samples), then it implies that w is probably very ill-behaved, and so distribution-free calibration is also likely to be impossible.

For a fixed partition $\{\mathcal{X}_b\}_{b\in[B]}$, one can use the labeled data from the source distribution to estimate $\mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]$ (unlike $\mathbb{E}_P[Y \mid X \in \mathcal{X}_b]$ as before), given oracle access to w:

$$\breve{\pi}_b^{(w)} := \frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}. \tag{5.16}$$

As preluded to earlier, assume that

$$\text{for all } x \in \mathcal{X},\ L \le w(x) \le U \text{ for some } 0 < L \le 1 \le U < \infty. \tag{5.17}$$

The 'standard' i.i.d. assumption on the test point equivalently assumes w is known and $L = U = 1$. We now present our first claim: $\breve{\pi}_b^{(w)}$ satisfies a distribution-free approximate calibration guarantee. To show the result, we assume that the sample-space partition was constructed via uniform-mass binning (on the source domain) with sufficiently many points, as required by Theorem 5.5. This guarantees that all regions satisfy $|\{i : \mathcal{B}(X_i) = b\}| = \Omega(n/B)$ with high probability.

Theorem 5.7. *Assume w is known and bounded* (5.17)*. Then for an explicit universal constant $c > 0$, with probability at least $1 - \alpha$,*

$$\left|\breve{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]\right| \le c\left(\frac{U}{L}\right)^2 \sqrt{\frac{B\ln(6B/\alpha)}{2n}}, \quad \textit{simultaneously for all } b \in [B],$$

as long as $n \geq c(U/L)^2 B \ln^2(6B/\alpha)$. *Thus* $\breve{\pi}_b^{(w)}$ *is asymptotically calibrated at any level* $\alpha \in (0,1)$.

The proof is in Appendix D.4. Theorem 5.7 establishes distribution-free calibration under Condition (A). For Condition (B), using k *unlabeled* samples from the source and target domains, assume that we construct an estimator $\widehat{w}_k$ of w that is consistent, meaning

$$\sup_{x \in \mathcal{X}} |\widehat{w}_k(x) - w(x)| \xrightarrow{P} 0. \tag{5.18}$$

We now define an estimator $\breve{\pi}_b^{(\widehat{w}_k)}$ by plugging in $\widehat{w}_k$ for w in the right hand side of (5.16):

$$\breve{\pi}_b^{(\widehat{w}_k)} := \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}.$$

Proposition 6. *If* $\widehat{w}_k$ *is consistent* (5.18)*, then* $\breve{\pi}_b^{(\widehat{w}_k)}$ *is asymptotically calibrated at any level* $\alpha \in (0,1)$.

In Appendix D.4, we illustrate through preliminary simulations that w can be estimated using unlabeled data from the target distribution, and consequently approximate calibration can be achieved on the target domain. Recently, Park et al. (2020) also considered calibration under covariate shift through importance weighting, but they do not show validity guarantees in the same sense as Theorem 5.7. For real-valued regression, distribution-free prediction sets under covariate shift were constructed using conformal prediction Tibshirani et al. (2019) under Condition (A), and is thus a precursor to our modular approach.

5.5 Other Related Work

The problem of assessing the calibration of binary classifiers was first studied in the meteorological and statistics literature (Brier, 1950; Sanders, 1963; Murphy and Epstein, 1967; Murphy, 1972a,b, 1973; Dawid, 1982; DeGroot and Fienberg, 1983; Bröcker, 2012; Ferro and Fricker, 2012); we refer the reader to the review by Dawid (2014) for more details. These works resulted in two common ways of measuring calibration: reliability diagrams (DeGroot and Fienberg, 1983) and estimates of the squared expected calibration error (ECE) Sanders (1963): $\mathbb{E}(f(X) - \mathbb{E}[Y \mid f(X)])^2$. Squared ECE can easily be generalized to multiclass settings and some related notions such as absolute deviation ECE and top-label ECE have also been considered, for instance Guo et al. (2017); Naeini et al. (2015). ECE is typically estimated through binning, which provably leads to underestimation of ECE for calibrators with continuous output (Vaicenavicius et al., 2019; Kumar et al., 2019). Certain methods have been proposed to estimate ECE without binning (Zhang et al., 2020; Widmann et al., 2019), but they require distributional assumptions for provability.

While these papers have focused on the difficulty of *estimating* calibration error, ours is the first formal impossibility result for *achieving* calibration for many commonly used calibration schemes. In particular, Kumar

et al. (2019, Theorem 4.1) show that the scaling-binning procedure achieves calibration error close to the best within a fixed, regular, injective parametric class. However, as discussed in Section 5.3.3 (after Theorem 5.3), we show that the best predictor in an injective parametric class itself cannot have a distribution-free guarantee. In summary, our results show not only that (some form of) binning is necessary for distribution-free calibration (Theorem 5.3), but also sufficient (Corollary 5.4.1).

Apart from classical methods for calibration (Platt, 1999; Zadrozny and Elkan, 2001, 2002; Niculescu-Mizil and Caruana, 2005), some new methods have been proposed recently in the ML literature, primarily for calibration of deep neural networks Lakshminarayanan et al. (2017); Guo et al. (2017); Kumar et al. (2018); Tran et al. (2019); Seo et al. (2019); Kuleshov et al. (2018); Kendall and Gal (2017); Wenger et al. (2020); Milios et al. (2018). These calibration methods perform well in practice but do not have distribution-free guarantees.

Calibration has natural applications in numerous sensitive domains where uncertainty estimation is desirable (healthcare, finance, forecasting). Recently, calibrated classifiers have been used as a part of the pipeline for anomaly detection Hendrycks et al. (2019); Lee et al. (2018) and label shift estimation Saerens et al. (2002); Alexandari et al. (2020); Garg et al. (2020).

5.6 Conclusion

We analyze calibration for binary classification problems from the standpoint of robustness to distributional assumptions. By connecting calibration to other ways of quantifying uncertainty, we establish that popular parametric scaling methods cannot provide provable informative calibration guarantees in the distribution-free setting. In contrast, we showed that a standard nonparametric method – histogram binning – satisfies approximate and asymptotic calibration guarantees without distributional assumptions. We also establish guarantees for the cases of streaming data and covariate shift.

Chapter 6

Distribution-Free Uncertainty Quantification for Classification under Label Shift

6.1 Introduction

It is common in classification to assume access to labeled data $\{(X_i, Y_i)\}_{i=1}^n$ where $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y} = \{1, \ldots, K\}$ denote the covariates, or features, and the labels respectively, and the pairs (X_i, Y_i), $i = 1, \ldots, n$ are sampled i.i.d. from some unknown joint distribution P over $\mathcal{X} \times \mathcal{Y}$. Such dataset is used to learn a predictor f, a mapping from $\mathcal{X}$ to rankings or distributions over $\mathcal{Y}$, by optimizing some loss/risk. However, accurate point prediction alone can be insufficient in certain applications, e.g., medical diagnosis, where trustworthy deployment of a model requires a valid measure of uncertainty associated with corresponding predictions.

Common prediction models are mappings of the form $f : \mathcal{X} \to \Delta_K$, where Δ_K refers to the probability simplex in $\mathbb{R}^K$, and a prediction on a new (test) point $X \in \mathcal{X}$ is performed by picking the top-ranked class according to $f(X)$. One hopes that the output vector $f(X)$ reflects the true conditional probabilities of classes given the observed input, but this won't be true without additional distributional and modeling assumptions, that are typically strong and unverifiable in practice. In this work, we focus on two categories of post-processing procedures — calibration via post-hoc binning and conformal prediction — that use held-out data (referred to as *calibration* dataset) and a trained model to construct a corresponding *wrapper* that provably quantifies predictive uncertainty when no distributional assumptions are made about the data generating mechanism. (This generality comes at a certain price which we discuss further.)

We work in the context of *distribution-free* uncertainty quantification and, in particular, focus on producing prediction sets (Section 6.2) and calibrated probabilities (Section 6.3), which are complementary approaches for

classifier UQ. While the former aims to produce a set of labels that contains the truth with high probability, the latter aims to amend the output of a probabilistic predictor so that it has a rigorous frequentist interpretation. It is useful to view the task through the lens of how actionable the corresponding notion is in a given setup. For example, in a binary classification setup with only 4 possible prediction sets $\{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$, if we were to observe prediction sets $\{1, 2\}$ for large fraction of data points, one might end up quite disappointed. Thus, calibration could be a better way of quantifying uncertainty in the binary case. However, mathematical guarantees on calibration degrade with growing number of classes, but the aforementioned prediction sets become an attractive option with more labels. To summarize, neither of two notions provide a complete answer to the question of UQ for classification on their own, but together they represent two of the more principled distribution-free approaches towards UQ that are practically efficient and theoretically grounded.

In real-world applications, the *target* distribution (generating test data) might not be the same as the *source* distribution (generating training data) which can both hurt a model's generalization and lead to violation of the assumptions under which even assumption-lean UQ is valid. As meaningful reasoning about uncertainty on the target domain is hopeless without any additional information about the type of distribution shift, one may hope that it is possible to make simplifying assumptions which would allow us to perform appropriate corrections and construct procedures with non-trivial guarantees. Let P, Q stand for the source and target distributions defined on $\mathcal{X} \times \mathcal{Y}$, with p, q being the PDFs or PMFs associated with P and Q respectively. Two common assumptions about the type of shift include *covariate shift* (Shimodaira, 2000): $q(x) \neq p(x)$ but $q(y \mid x) = p(y \mid x)$, and *label shift* (Saerens et al., 2002): $q(y) \neq p(y)$ but $q(x \mid y) = p(x \mid y)$. Both assumptions allow for a tractable interpretation when viewing the data generating process as a causal or anti-causal model respectively. For example, label shift is a reasonable assumption in medical applications where diseases (Y) cause symptoms (X): it is intuitive that some sort of correction might be required when a predictor trained in ordinary conditions is deployed during extreme ones, e.g., during a pandemic.

Classic approaches for handling the aforementioned shifts make an assumption that the target support is contained in the source support, so that the covariate or label likelihood ratios (or *importance weights*) $q(x)/p(x)$ or $q(y)/p(y)$ are well-defined. In applications, true weights are never known exactly, so the construction of consistent estimators has received a lot of attention in the ML community. For label shift dominant approaches that are still computationally feasible in modern high-dimensional regimes, and that perform estimation using labeled data only from the source distribution, include: (a) Black Box Shift Estimation (BBSE) (Lipton et al., 2018) and related Regularized Learning under Label Shift (RLLS) (Azizzadenesheli et al., 2019), (b) Maximum Likelihood Label Shift (MLLS) and its variants (Saerens et al., 2002; Alexandari et al., 2020).

Within the context of distribution-free UQ, covariate shift has recently received attention. Focusing on regression, Tibshirani et al. (2019) generalize construction of conformal prediction intervals to handle the case of known covariate likelihood ratio, and empirically demonstrate that the modified procedure works reasonably well with a plug-in estimator for the importance weights. For binary classification, Gupta et al. (2020) propose a way of calibrating probabilistic predictors under covariate shift, and quantify miscalibration of the resulting estimator.

In this work, we close an existing gap for quantifying predictive uncertainty under label shift. Building on recent results about distribution-free calibration and (split-)conformal prediction, we adapt both to handling label shift through an appropriate form of reweighting. While typical application of those frameworks requires labeled data from the target to provide guarantees, we show that under reasonable assumptions one can still reason about uncertainty on the target even if only unlabeled data is available. In contrast to covariate shift where we observe X and need the covariate likelihood ratio of X to reweight, under label shift we observe X but need the likelihood ratio of Y to reweight. We also consider an alternative way of addressing label shift by performing label-conditional conformal classification (Vovk et al., 2005, 2016; Sadinle et al., 2019; Guan and Tibshirani, 2022).

6.2 Conformal Classification

We begin with the notion of prediction sets as a way of quantifying predictive uncertainty. Formally, we wish to construct an uncertainty set function $C : \mathcal{X} \to 2^{\mathcal{Y}}$, such that for a new (test) data point we can guarantee that:

$$\mathbb{P}\left(Y_{n+1} \in C(X_{n+1})\right) \geq 1 - \alpha. \tag{6.1}$$

Conformal prediction (Vovk et al., 2005) has received attention recently both in regression (Lei et al., 2018; Romano et al., 2019; Barber et al., 2021) and classification (Cauchois et al., 2021; Romano et al., 2020; Angelopoulos et al., 2021) settings. It does not require making any distributional assumptions, which comes at the price of provably providing only *marginal* guarantees as stated in (6.1) which should be contrasted with possibly the ultimate goal of obtaining prediction sets with guarantees conditional on a given input.

Since conditional guarantees often require making restrictive and unverifiable assumptions, we instead focus on procedures that might provably provide marginal coverage guarantees but still tend to demonstrate good conditional coverage empirically. Being flexible, conformal prediction allows to proceed with both probabilistic and scoring classifiers. Within this framework, one usually defines a non-conformity score, a higher value of which on a given data point indicates that it is more 'atypical'. For example, even if a classifier outputs only the ranking of predicted classes, a rank of the true class defines a valid non-conformity score. Keeping in mind that our techniques extend to other types of classifiers, we nevertheless focus on probabilistic predictors in this work which are also dominant in modern machine learning.

6.2.1 Exchangeable Conformal

Consider a sequence of candidate nested prediction sets $\{\mathcal{F}_\tau(x)\}_{\tau \in \mathcal{T}}$: $\mathcal{F}_{\tau_1}(x) \subseteq \mathcal{F}_{\tau_2}(x) \subseteq \mathcal{Y}$ for any $\tau_1 \leq \tau_2 \in \mathcal{T}$, with $\mathcal{F}_{\inf \mathcal{T}} = \emptyset$ and $\mathcal{F}_{\sup \mathcal{T}} = \mathcal{Y}$ (Gupta et al., 2022). For any point $(x, y) \in \mathcal{X} \times \mathcal{Y}$ define

$$r(x, y) := \inf \left\{\tau \in \mathcal{T} : y \in \mathcal{F}_\tau(x)\right\}, \tag{6.2}$$

as the smallest radius of the set in a sequence $\{\mathcal{F}_\tau(x)\}_{\tau\in\mathcal{T}}$ that captures y. Within split-conformal framework, available dataset is split at random into two parts: the first is used to construct a nested sequence and the second is used to select the smallest $\tau^\star$ that guarantees validity.

If the true class-posterior distribution $\pi_y(x) = \mathbb{P}[Y = y \mid X = x]$ is known, the optimal prediction set for any $x \in \mathcal{X}$ with conditional coverage guarantee is based on the corresponding density level sets (Vovk et al., 2005; Lei et al., 2013; Gupta et al., 2022; Sadinle et al., 2019): one should pick the largest $\tau_\alpha(x)$ and include all labels with probabilities $\pi_y(x)$ exceeding $\tau_\alpha(x)$ so that the corresponding total probability mass is at least $1-\alpha$. When ties are present, such procedure can yield conservative sets, e.g., if for some $x \in \mathcal{X}$ all classes are equally probable in a 10-class problem, then $\tau_\alpha(x) = 0.1$ and the proposed set would simply be $\mathcal{Y}$. For the discussion that follows we assume that there are no ties or that they are broken as formally discussed in Appendix E.2.1. Then, to construct the optimal prediction set, one should start with an empty one and keep including labels as long as the total probability mass of labels included before is less than $1-\alpha$. Formally,

$$C_\alpha^{\text{oracle}}(x) := \{y \in \mathcal{Y} : \rho_y(x;\pi) < 1-\alpha\}, \tag{6.3}$$
$$\text{where} \quad \rho_y(x;\pi) := \sum_{y'=1}^{K} \pi_{y'}(x)\mathbb{1}\{\pi_{y'}(x) > \pi_y(x)\}$$

is the total probability mass of labels that are more likely than $y \in \mathcal{Y}$. Notice that for any $x \in \mathcal{X}$ and the corresponding most likely label $y^\star$ it holds that $\rho_{y^\star}(x;\pi) = 0$. When an estimator $\widehat{\pi}$ of the true conditional distribution is used, split-conformal framework provides a way of updating the threshold $1-\alpha$ in (6.3) in order to retain coverage guarantees. However, naive conformalization of the nested sequence suggested by the form (6.3) yields prediction sets with correct marginal coverage but typically inferior conditional coverage in practice. Due to that reason and a desire of consistency, i.e., recovering the oracle prediction sets from the conformal ones in the limit, we instead use a randomized version of (6.3) defined as

$$\widetilde{C}_\alpha^{\text{oracle}}(x) = \{y : \rho_y(x;\pi) + u \cdot \pi_y(x) \le 1-\alpha\}, \tag{6.4}$$

where u is a realization of $\text{Unif}([0,1])$, sampled independently of anything else (Vovk et al., 2005; Romano et al., 2020). Note that replacing strict inequality by a non-strict does not expand the prediction set as equality happens with zero probability and that induced randomization can result in exclusion only of a single label from the set $C_\alpha^{\text{oracle}}(x)$. The form of the optimal prediction sets (6.4) suggests to consider the following nested sequence:

$$\mathcal{F}_\tau(x, u; \widehat{\pi}) = \{y \in \mathcal{Y} : \rho_y(x;\widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \le \tau\}, \tag{6.5}$$

for $\tau \in \mathcal{T} = [0, 1]$. Then for any triple (X, Y, U) the corresponding radius (6.2), or score, is given by

$$\begin{aligned} r(X, Y, U; \widehat{\pi}) &= \inf \{\tau \in \mathcal{T} : \rho_Y(X; \widehat{\pi}) + U \cdot \widehat{\pi}_Y(X) \leq \tau\} \\ &= \rho_Y(X; \widehat{\pi}) + U \cdot \widehat{\pi}_Y(X). \end{aligned} \tag{6.6}$$

Adapting to label shift can be performed with other non-conformity scores proposed recently for conformal classification (Cauchois et al., 2021; Angelopoulos et al., 2021), and we further discuss the subtleties behind our choice in Appendix E.2.2. Assume that the dataset is split at random into two parts: training $\{(X_i, Y_i)\}_{i \in \mathcal{I}_1}$ and calibration $\{(X_i, Y_i)\}_{i \in \mathcal{I}_2}$, where for simplicity the calibration data points are indexed as $\mathcal{I}_2 = \{1, \ldots, n\}$. When the data are exchangeable, the non-conformity scores $r_i = r(X_i, Y_i, U_i; \widehat{\pi}) \in [0, 1]$, $i \in \mathcal{I}_2 \cup \{n + 1\}$ are exchangeable as well, which in turn implies that the prediction set

$$\begin{aligned} \mathcal{F}_{\tau^\star}(x, u; \widehat{\pi}) &= \{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leq \tau^\star\}, \\ \tau^\star &= Q_{1-\alpha}\left(\{r_i\}_{i \in \mathcal{I}_2} \cup \{1\}\right), \end{aligned} \tag{6.7}$$

does attain the right coverage guarantee*. This is a classic result in conformal prediction and represents a simple fact about quantiles of exchangeable random variables, stated next for completeness.

Theorem 6.1. *If $\{(X_i, Y_i)\}_{i=1}^{n+1}$ are exchangeable, then:*

$$\mathbb{P}(Y_{n+1} \in \mathcal{F}_{\tau^\star}(X_{n+1}, U_{n+1}; \widehat{\pi}) \mid \{(X_i, Y_i)\}_{i \in \mathcal{I}_1}) \geq 1 - \alpha.$$

Further, if the non-conformity scores are almost surely distinct, then the above probability is upper bounded by $1 - \alpha + 1/(n + 1)$.

The proof is given in Appendix E.2.3. Notice that the randomized sequence (6.5) might yield empty, and thus non-actionable prediction sets, which is the consequence of deploying randomization only. Substituting the condition in (6.7) with $\mathbb{1}\{\rho_y(x; \widehat{\pi}) > 0\} \cdot (\rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x))$ ensures that the prediction set always includes the most likely label. Such a construction trivially inherits the coverage guarantee stated in Theorem 6.1, and we refer the reader to Appendix E.2.2 for further details.

6.2.2 Label-shifted Conformal

To illustrate the necessity of accounting for label shift we consider the following toy classification task with 3 classes $\mathcal{Y} = \{1, 2, 3\}$ where class proportions are given as $p = (0.1, 0.6, 0.3)$ and $q = (0.3, 0.2, 0.5)$, and for each data point the covariates are sampled according to $X \mid Y = y \sim \mathcal{N}(\mu_y, \Sigma)$ where $\mu_1 = (-2; 0)^\top$, $\mu_2 = (2; 0)^\top$, $\mu_3 =$

*$Q_\beta(F) := \inf\{z : F(z) \geq \beta\}$ is β-quantile of a distribution F. For a multiset $\{z_1, \ldots, z_m\}$ we write $Q_\beta(\{z_1, \ldots, z_m\}) := Q_\beta\left(\frac{1}{m}\sum_{i=1}^m \delta_{z_i}\right)$, where δ_a is a point-mass distribution at a, to denote quantiles of the corresponding empirical distribution.

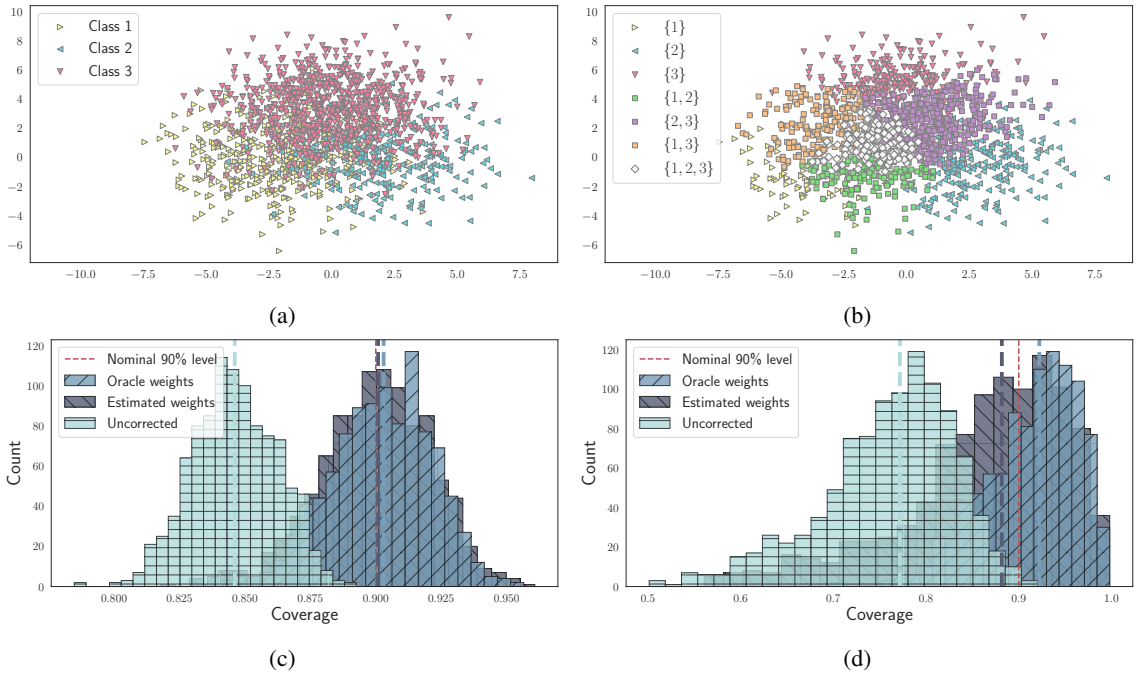


Figure 6.1: (a) Test data sample for the toy simulation in Section 6.2.2. (b) Corresponding conformal prediction sets when label shift is accounted for with oracle importance weights. (c) Empirical coverage on shifted data for the toy simulation in Section 6.2.2. (d): Empirical coverage on the `wine quality` dataset. Dashed vertical lines describe the median coverage values, which are significantly worse when label shift is not accounted for, while using estimated weights mimics the oracle reasonably well.

$\left(0; 2\sqrt{3}\right)^\top$, $\Sigma = \mathrm{diag}(4, 4)$. First, we perform the standard routine for constructing split-conformal prediction sets for a single draw of data from the source and target distributions using the Bayes-optimal rule as an underlying predictor. We illustrate a single draw of the test data on Figure 6.1a and the resulting prediction sets on Figure 6.1b. Next, we repeat the simulation 1000 times and track empirical coverage on the test set. Results on Figure 6.1c demonstrate the necessity of correcting for label shift as the classic conformal prediction sets introduced in Section 6.2.1 fail to achieve the correct marginal coverage.

Assume that the true likelihood ratios $w(y) = q(y)/p(y)$ are known for all $y \in \mathcal{Y}$. In order to obtain provably valid prediction sets, we consider instead:

$$\begin{aligned}
\mathcal{F}^{(w)}_{\tau^\star}(x, u; \widehat{\pi}) &= \left\{ y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u \cdot \widehat{\pi}_y(x) \leq \tau^\star_w(y) \right\}, \\
\tau^\star_w(y) &= Q_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{p}^w_i(y)\delta_{r_i} + \tilde{p}^w_{n+1}(y)\delta_1 \right), \qquad (6.8) \\
\text{where} \quad \tilde{p}^w_i(y) &= \frac{w(Y_i)}{\sum_{j=1}^{n} w(Y_j) + w(y)}, \quad i = 1, \ldots, n, \\
\tilde{p}^w_{n+1}(y) &= \frac{w(y)}{\sum_{j=1}^{n} w(Y_j) + w(y)}. \qquad (6.9)
\end{aligned}$$

In addition to the fact that the empirical distribution used to calibrate the threshold in (6.8) is different from the one used in exchangeable setting (6.7), notice that the thresholds themselves now vary depending on the class label. The formal guarantee for the prediction set (6.8) is stated next.

Theorem 6.2. *For any $\alpha \in (0,1)$, if the true likelihood ratios $w(y) = q(y)/p(y)$ are known for all $y \in \mathcal{Y}$, it holds that*

$$\mathbb{P}(Y_{n+1} \in \mathcal{F}_{\tau^\star}^{(w)}\left(X_{n+1}, U_{n+1}; \widehat{\pi}\right) \,|\, \{(X_i, Y_i)\}_{i \in \mathcal{I}_1}) \geq 1 - \alpha.$$

The proof is given in Appendix E.2.3. It relies on the concept of *weighted exchangeability* introduced by Tibshirani et al. (2019) to handle covariate shift in regression, and we adapt those ideas here to correct for label shift in classification. Returning to the example considered in the beginning of this section, Figure 6.1c illustrates that calibrating the threshold τ as in (6.8) with either oracle or estimated importance weights allows to achieve the target marginal coverage. Here we use BBSE (Lipton et al., 2018) to estimate the importance weights; more details are provided in Appendix E.1.

Next, we perform a similar experiment with the `wine quality` dataset (Cortez et al., 2009). We refer the reader to Appendix E.2.4 for details regarding data pre-processing and modeling steps. The source and target class proportions are taken to be $p = (0.1, 0.4, 0.5)$ and $q = (0.4, 0.5, 0.1)$ and the data are resampled accordingly. Using a shallow multilayer perceptron as an underlying predictor and BBSE for importance weights estimation, at each iteration we repeat the routine for random splits of the original dataset and compare empirical coverage for different conformal prediction sets. Marginal coverage results given in Figure 6.1d support the idea that both shift-corrected conformal prediction sets demonstrate superior coverage performance compared with uncorrected ones. While conformal sets with oracle importance weights closely match the nominal coverage level, sets that proceed with estimated ones have a slightly downgraded performance. Arising basically due to an imperfect classification model and an imperfect importance weight estimation procedure, it highlights an important issue we discuss next.

While (weighted) exchangeability arguments yield a coverage guarantee in case of known importance weights, in practice one only has access to a corresponding estimator. Dominant methods, which we briefly touch upon in Appendix E.1, estimate importance weights using a separate labeled dataset from the source distribution and unlabeled dataset from the target. Under reasonable assumptions, such as identifiability and boundedness of the true importance weights, these estimators are known to be consistent as the size of both samples grows. For succinctness, we write $k = |\mathcal{D}_{\text{est}}|$ to denote the *total* size of the datasets used for constructing an estimator $\widehat{w}_k$ of the importance weights w.

Corollary 6.2.1. *Fix $\alpha \in (0,1)$. Assume that $\widehat{w}_k$ is a consistent estimator of w. Further, assume that for the true w and all $y \in \mathcal{Y}$, the discrete distribution in* (6.8) *does not have a jump at level $1-\alpha$. Then:*

$$\lim_{k \to \infty} \mathbb{P}\left(Y_{n+1} \in \mathcal{F}_{\tau^\star}^{(\widehat{w}_k)}\left(X_{n+1}, U_{n+1}; \widehat{\pi}\right)\right) \geq 1 - \alpha.$$

The proof is given in Appendix E.2.3. To demonstrate why presence of a jump might cause problems, consider a simplified example. Let $Z \sim \mathrm{Ber}(p)$ for which the quantile corresponding to any given level α is given by

$$Q_\alpha\left((1-p)\cdot\delta_0 + p\cdot\delta_1\right) = \mathbb{1}\left\{p > 1-\alpha\right\},$$

Assume that we are given a sample of coin tosses $Z_1, \ldots, Z_n$ with the same bias parameter p. Even though the sample average $\overline{Z}_n$ is a consistent estimator of p, it nonetheless does not imply that the corresponding plug-in quantile estimator is consistent as the continuous mapping theorem cannot be invoked due to a discontinuity at $p = 1-\alpha$. Indeed, let

$$\widehat{q}_n := Q_\alpha\left(\left(1-\overline{Z}_n\right)\cdot\delta_0 + \overline{Z}_n\cdot\delta_1\right) = \mathbb{1}\left\{\overline{Z}_n > 1-\alpha\right\},$$

and observe that $\widehat{q}_n \sim \mathrm{Ber}\left(\mathbb{P}\left(\overline{Z}_n > 1-\alpha\right)\right)$. Then by the normal approximation it follows that:

$$\mathbb{P}\left(\overline{Z}_n > 1-\alpha\right) \approx 1 - \Phi\left(\sqrt{n}\frac{(1-\alpha)-p}{\sqrt{p(1-p)}}\right).$$

If $p > 1-\alpha$, we can conclude that $\widehat{q}_n$ converges in probability to 1, and thus the estimator is consistent (similarly for $p < 1-\alpha$). In case of equality, $\widehat{q}_n$ converges to $\mathrm{Ber}(1/2)$, and thus the estimator will not be consistent. Still, for a more general setting of the distribution defined in (6.8) it is reasonable to expect the assumption regarding absence of jumps to be satisfied as also confirmed by our conducted empirical study.

Label-conditional conformal prediction. Observing multiple points sharing the same label in a dataset makes it possible to apply the split-conformal framework in a way that makes the resulting prediction sets inherently robust to label shift (Vovk et al., 2005, 2016; Sadinle et al., 2019; Guan and Tibshirani, 2022). Assume that a set of significance levels for each class $\{\alpha_y\}_{y\in\mathcal{Y}}$ has been chosen (e.g., $\alpha_y = \alpha$ for all y). By further splitting the calibration set $\mathcal{I}_2$ into $|\mathcal{Y}| = K$ groups depending on the corresponding labels, $\mathcal{I}_{2,y} := \{i \in \mathcal{I}_2 : Y_i = y\}$, one can consider prediction sets of the following form:

$$\begin{aligned}\mathcal{F}^c_{\tau^\star_c}(x, u; \widehat{\pi}) &= \left\{y \in \mathcal{Y} : \rho_y(x; \widehat{\pi}) + u\cdot\widehat{\pi}_y(x) \le \tau^\star_c(y)\right\},\\ \tau^\star_c(y) &= Q_{1-\alpha_y}\left(\left\{r_i\right\}_{i\in\mathcal{I}_{2,y}} \cup \{1\}\right).\end{aligned} \tag{6.10}$$

In other words, we separately apply split-conformal prediction framework for each label; this is like performing a separate hypothesis test for each label to determine whether there is sufficient evidence to exclude the label from the prediction set. To elaborate, the label shift assumption states that conditional distribution of X given $Y = y$ for all $y \in \mathcal{Y}$ does not change between source and target distributions. Thus for a test point (X_{n+1}, Y_{n+1}) the corresponding non-conformity score $r(X_{n+1}, Y_{n+1}, U_{n+1}; \widehat{\pi})$ together with $\{r_i\}_{i\in\mathcal{I}_{2,Y_{n+1}}}$ forms a collection of exchangeable random

variables, which implies label-conditional validity, that is:

$$\mathbb{P}\left(Y_{n+1} \notin \mathcal{F}^c_{\tau^\star_c}\left(X_{n+1}, U_{n+1}; \widehat{\pi}\right) \mid Y_{n+1} = y\right) \leq \alpha_y,$$

for all $y \in \mathcal{Y}$. When $\alpha_y = \alpha$ for all y, one can marginalize over y using *any* distribution (shifted or not), to yield $\mathbb{P}\left(Y_{n+1} \notin \mathcal{F}^c_{\tau^\star_c}\left(X_{n+1}, U_{n+1}; \widehat{\pi}\right)\right) \leq \alpha$. Thus, the label-conditional conformal framework yields a stronger guarantee than the standard (marginal) conformal and, it is automatically robust to changes in class proportions, retaining validity under label shift. The price to pay for the stronger conditional guarantee is larger prediction sets: for example, when the classes are not well-separated, label-conditional conformal can be expected to yield larger prediction sets; see Appendix E.2.5 for a careful empirical study. It should also be noted that the label-conditional conformal framework requires splitting available calibration data into K parts that could result in large losses of statistical efficiency when the number of classes K is large. On the other hand, such construction allows to tackle label shift in a way that does not require importance weights estimation, and thus get exact finite-sample guarantee instead of asymptotic one established in Corollary 6.2.1. Thus, we view the label-conditional conformal framework as a complementary approach, perhaps worth utilizing when the amount of calibration data is larger relative to the number of labels.

6.3 Calibration

While prediction sets describe a construction on top of the output of a predictor, calibration quantifies whether the output itself admits a rigorous frequentist interpretation. In contrast to the binary setting where there is usually no confusion about a definition of a calibrated predictor, there is one in the multiclass setting. First, we state a definition of a canonically calibrated predictor.

Definition 6 (Calibration)**.** *A probabilistic predictor $f : \mathcal{X} \to \Delta_K$ is said to be calibrated if*

$$\mathbb{P}\left(Y = y \mid f(X)\right) = f_y(X), \quad y \in \mathcal{Y},$$

where $f_y(x)$ denotes the y-th coordinate of $f(x)$.

Observe that canonical calibration requires the whole output vector to reflect the true conditional probabilities. Two extreme examples of canonically calibrated predictors include: (a) f^{Marg}: $f^{\text{Marg}}_y(x) = p(y)$, (b) f^{Bayes}: $f^{\text{Bayes}}_y(x) = \pi_y(x)$. In words, the former predictor outputs marginal probabilities of classes and the latter outputs the true class-posterior probabilities. In terms of classification efficiency, however, the first one is useless, while the second minimizes the classification risk, or the probability of incorrectly classifying a new point. Minimizing classification risk with respect to zero-one loss is computationally infeasible, and thus one refers instead to minimizing so-called *surrogate* losses, e.g., cross-entropy loss, with possibly added regularization terms. As a result, one obtains prediction models that are not calibrated out-of-the-box without making strong distributional and modeling

assumptions, and thus aims to achieve it by performing post-processing using held-out data. While this topic has attracted a lot attention from practitioners recently, less results have been established on the theoretical side providing formal guarantees for common procedures that target improving model's calibration. Recognized approaches include Platt scaling (Platt, 1999), temperature scaling (Guo et al., 2017), histogram binning (Zadrozny and Elkan, 2001), isotonic regression (Zadrozny and Elkan, 2002) and others.

Model miscalibration is usually assessed using either reliability curves or related one-dimensional summary statistics. It is known that popular metrics, such as Expected Calibration Error (ECE), are not reliable since plug-in estimates can be biased if binning, or discretization, of the output of the resulting model is not performed (Kumar et al., 2019; Vaicenavicius et al., 2019). Gupta et al. (2020) establish the necessity of binning for obtaining distribution-free calibration guarantees in a binary classification setup. Binning represents coarsening of the sample space and is defined as the partitioning of the probability simplex into non-overlapping bins: $\Delta_K = B_1 \cup \cdots \cup B_M$, $B_i \cap B_j = \emptyset$, $i \neq j$. Then a predictor f induces a partition of the sample space:

$$\mathcal{X}_m := \{x \in \mathcal{X} : f(x) \in B_m\}, \quad m \in \mathcal{M} := \{1, \ldots, M\}.$$

Since provable guarantees for canonical calibration require binning of the probability simplex, it is clear that the task becomes prohibitive with growing number of classes as each bin has to be supplied with sufficiently many data points during the calibration step for the resulting guarantees to be meaningful. One solution is given by either referring to other notions of UQ, such as the aforementioned prediction sets, or by relaxing the notion of calibration in the multiclass setting. One of well-known relaxations is class-wise, or marginal, calibration (Zadrozny and Elkan, 2002; Vaicenavicius et al., 2019; Kull et al., 2019).

Definition 7 (Class-wise calibration)**.** *A probabilistic predictor $f : \mathcal{X} \to \Delta_K$ is said to be class-wise calibrated if*

$$\mathbb{P}(Y = y \mid f_y(X)) = f_y(X), \quad y \in \mathcal{Y}. \tag{6.11}$$

Vaicenavicius et al. (2019) illustrate the difference with the canonical calibration through useful examples. In the binary setting, the two notions are equivalent with class-wise calibration being a weaker requirement for larger number of classes. It is achieved by reducing the original multiclass problem to K one-vs-all binary problems with the standard post-processing routine applied consequently to each one. We focus on canonical calibration for multiclass problems as per Definition 6 and explicitly mention important implications for the binary setting, and thus marginal calibration.

6.3.1 Calibration for i.i.d. Data

First, we assume that the binning scheme has been chosen and use $g : \mathcal{X} \to \mathcal{M}$ to denote the bin-mapping function: $g(x) = m$ if and only if $f(x) \in B_m$. The calibration set $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$ is used for estimating

$$\pi_{y,m}^P := \mathbb{P}\left(Y = y \mid f(X) \in B_m\right), \quad y \in \mathcal{Y}, \tag{6.12}$$

for all bins $m \in \mathcal{M}$. The superscript here highlights that probabilities correspond to the source distribution P and the notation will become convenient when we talk about label shift setting. With finite data one can only estimate (6.12) with quantifiable measures of error, and thus provably satisfy the calibration requirement only approximately:

$$\mathbb{P}\left(Y = y \mid \widehat{\pi}_{y,g(X)}^P\right) \approx \widehat{\pi}_{y,g(X)}^P. \tag{6.13}$$

Let $N_m = |\{(X_i, Y_i) \in \mathcal{D}_{\text{cal}} : f(X_i) \in B_m\}|$ denote the number of calibration points that fall into bin $m \in \mathcal{M}$. Note that $\{N_m\}_{m\in\mathcal{M}}$ are random and satisfy $\sum_{m=1}^M N_m = n$. Empirical frequencies of class labels $y \in \mathcal{Y}$ in each bin $m \in \mathcal{M}$:

$$\widehat{\pi}_{y,m}^P := \frac{1}{N_m} \sum_{i=1}^n \mathbb{1}\left\{Y_i = y, f(X_i) \in B_m\right\}, \tag{6.14}$$

are natural candidates to satisfy the approximate calibration condition (6.13). For convenience, let $\pi_m^P := (\pi_{1,m}^P, \ldots, \pi_{K,m}^P)^\top$ denote a vector with coordinates representing bin-conditional class probabilities and let $h : \mathcal{X} \to \Delta_K$ denote the *recalibrated* predictor, i.e., the function that maps any feature vector to the corresponding vector of *calibrated* probability estimates: $h(x) = \widehat{\pi}_{g(x)}$.

Theorem 6.3. *Fix $\alpha \in (0, 1)$. With probability at least $1 - \alpha$, $\left\|\widehat{\pi}_m^P - \pi_m^P\right\|_1 \leq \varepsilon_m$, simultaneously for all $m \in \mathcal{M}$, where*

$$\varepsilon_m := \frac{2}{\sqrt{N_m}} \sqrt{\frac{1}{2} \ln\left(\frac{M 2^K}{\alpha}\right)}.$$

As a consequence, with probability at least $1 - \alpha$,

$$\sum_{y=1}^K \left|\mathbb{P}\left(Y = y \mid h(X) = z\right) - z_y\right| \leq \max_{m\in\mathcal{M}} \varepsilon_m,$$

simultaneously for all z in the range of h.

The proof is given in Appendix E.3.1. In words, Theorem 6.3 states that as long as the least-populated bin contains sufficiently many points, the output of the recalibrated predictor will approximately satisfy condition (6.13). The first part of Theorem 6.3 justifies use of empirical frequencies in place of unknown population quantities using the language of the confidence intervals. In the binary setting, the fact that it yields the desired calibration guarantee, has

been formally established by Gupta et al. (2020), and the second part of the theorem states a corresponding result for canonical calibration in the multiclass setting.

A natural question is whether one can guarantee that each bin is supplied with a sufficient number of calibration data points in order to obtain meaningful bounds. We note that in the binary setting, one way to provably spread the calibration data evenly across bins is uniform-mass, or equal frequency, binning (Kumar et al., 2019; Gupta et al., 2020; Gupta and Ramdas, 2021).

6.3.2 Label-shifted Calibration

For illustrating the necessity of accounting for label shift we consider the following binary classification problem: $\mathcal{Y} = \{0, 1\}$ with class probabilities given as $p(0) = p(1) = 1/2$ and $q(0) = 0.2$, $q(1) = 0.8$, i.e., while on the source domain classes are equally balanced, on the target class 1 becomes dominant. For each data point, conditionally on the corresponding label, the covariates are sampled according to $X \mid Y = y \sim \mathcal{N}(\mu_y, \Sigma)$, where

$$\mu_0 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{pmatrix}.$$

Similarly to the toy example from Section 6.2.2, here the class-posterior probabilities, and thus the Bayes-optimal rules have a closed form for both source and target domains. Not only do they minimize the probability of misclassifying a new point from the corresponding domain but also they are calibrated[†]. For the source distribution a perfect probabilistic predictor is given by

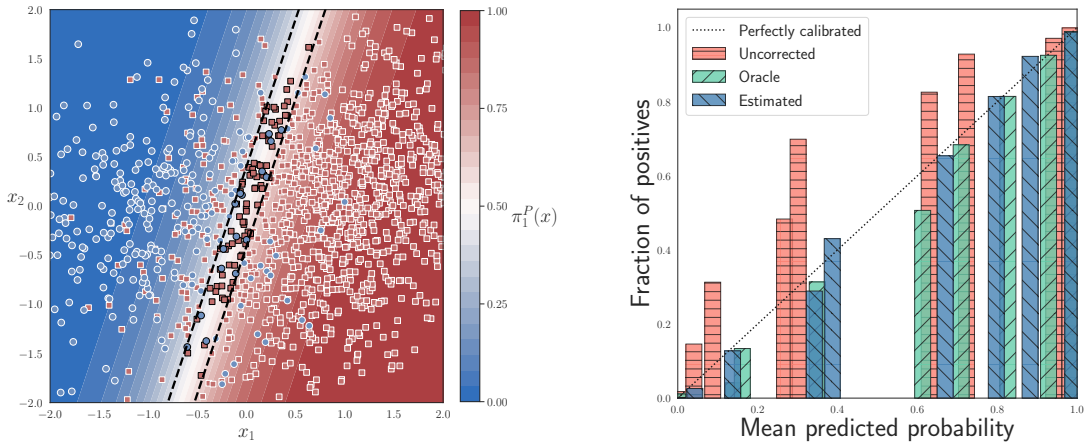
$$\pi_1^P(x) = \frac{p(1) \cdot \varphi(x; \mu_1, \Sigma)}{p(0) \cdot \varphi(x; \mu_0, \Sigma) + p(1) \cdot \varphi(x; \mu_1, \Sigma)}, \tag{6.15}$$

where $\varphi(x; \mu_i, \Sigma)$, $i = 0, 1$ denotes the PDF of a Gaussian random vector with the corresponding parameters. As illustrated on Figure 6.2a, even though the Bayes-optimal rule is calibrated on the source, a correction is required to obtain a calibrated classifier under label shift. We sample points from the target distribution and highlight those that fall inside the area $S = \left\{x \in \mathbb{R}^2 : \pi_1^P(x) \in [0.4; 0.6]\right\}$ with boundary given by the black dashed lines. When the shift is present, predictor (6.15) is no longer calibrated, since otherwise one should expect roughly half of the test data points inside S to be labeled as class 1 (red squares) and half as class 0 (blue circles), which clearly does not happen.

If both the true class-posterior distribution $\pi_y^P(x)$ and the true label likelihood ratios w are known, then the form of the adjustment of the probabilistic classifier under label shift is a simple implication of the Bayes rule (Saerens et al., 2002):

$$\pi_y^Q(x) = \frac{w(y) \cdot \pi_y^P(x)}{\sum_{k=1}^K w(k) \cdot \pi_k^P(x)}. \tag{6.16}$$

[†]Recall that in the binary setting, canonical and class-wise calibration are equivalent.



(a) (b)

Figure 6.2: (a) Sampled points from the target distribution plotted against the true source class-posterior probabilities. (b) Reliability curves for Fisher's LDA calibrated via binning with/without taking label shift into account. The deviation of uncorrected probabilities from the diagonal line (perfect calibration) reflects the need to correct for label shift; recalibration based on estimated weights is almost identical to using oracle weights, both of which result in near-perfect calibration.

While in the oracle setting predictor (6.16) is indeed calibrated on the target, in practice neither $\pi_y^P(x)$ nor w are known. Using corresponding plug-in estimators in (6.16) would guarantee calibration of the resulting predictor only asymptotically and under restricting modeling assumptions, and thus to obtain the distribution-free guarantees the output of the original predictor has to be discretized, or binned as in the i.i.d. setting. Relationship (6.16) does clearly continue to hold as formally stated next.

Proposition 7. *Under label shift, for any class label $y \in \mathcal{Y}$ and any bin B_m, $m \in \mathcal{M}$ it holds that:*

$$\pi_{y,m}^Q = \frac{w(y) \cdot \pi_{y,m}^P}{\sum_{k=1}^K w(k) \cdot \pi_{k,m}^P}.$$

In Section 6.3.1 we justified the use of empirical frequencies of class labels $\left\{\widehat{\pi}_m^P\right\}_{m \in \mathcal{M}}$ for achieving canonical calibration of a predictor on the source domain and, as it has been noted in Section 6.2.2, there are estimators of the importance weights which are known to be provably consistent under reasonable assumptions. Thus, with an estimator $\widehat{w}$ at hand, Proposition 7 suggests an appropriate correction to provably obtain asymptotically calibrated predictors on the target:

$$\widehat{\pi}_{y,m}^{(\widehat{w})} = \frac{\widehat{w}(y) \cdot \widehat{\pi}_{y,m}^P}{\sum_{k=1}^K \widehat{w}(k) \cdot \widehat{\pi}_{k,m}^P}, \quad y \in \mathcal{Y}, \tag{6.17}$$

for all bins $m \in \mathcal{M}$. Theorem 6.3 quantifies the error when the empirical label frequencies are used as estimators for the true unknown bin-conditional class probabilities on the source domain. However, different bounds on ε_m could be available depending on chosen binning scheme, and thus we instead quantify how this estimation error on the source

domain translates into the estimation error on the target for the cases when the importance weights are known and when they are rather estimated. As we shall see, the performance depends on the ratio of the largest to the smallest nonzero importance weight. Define the *condition number*:

$$\kappa := \frac{\sup_k w(k)}{\inf_{k:w(k)\neq 0} w(k)},$$

with $\kappa = 1$ corresponding to label shift not being present. Next, we quantify the miscalibration of the predictor (6.17).

Theorem 6.4. *Let $\widehat{w}$ be an estimator of w and let $\widehat{\pi}_{y,m}^{(\widehat{w})}$ denote the reweighted empirical frequencies* (6.17) *for all labels $y \in \mathcal{Y}$ and bins $m \in \mathcal{M}$. For any bin $m \in \mathcal{M}$, it holds that:*

$$\left\| \widehat{\pi}_m^{(\widehat{w})} - \pi_m^Q \right\|_1 \leq \underbrace{2\kappa \cdot \left\| \widehat{\pi}_m^P - \pi_m^P \right\|_1}_{(a)} + \underbrace{\frac{2 \left\| \widehat{w} - w \right\|_\infty}{\inf_{l:w(l)\neq 0} w(l)}}_{(b)}. \tag{6.18}$$

The proof is given in Appendix E.3.1. In words, the calibration error on the target decomposes into two terms where (a) is controlled by the calibration error on the source and (b) is controlled by the importance weights estimation error. Further, under reasonable assumptions common procedures, such as BBSE and RLLS, construct estimators of the importance weights which are not only known to be consistent but also have quantifiable error (Lipton et al., 2018; Azizzadenesheli et al., 2019). Similarly, any proper binning scheme that provably controls number of calibration points in each bin, e.g., uniform-mass binning in the binary setting (Kumar et al., 2019), yields finite-sample guarantees for the calibration error on the source (Gupta et al., 2020). Thus, finite-sample guarantees for the miscalibration of the resulting predictor on the target domain trivially follow by virtue of Theorem 6.4 via invoking simple probabilistic arguments.

Within the same binary classification setup from the beginning of Section 6.3.2, we also compare calibration via uniform-mass binning with and without accounting for label shift but this time we use Fisher's LDA as an underlying classifier, which differs from the Bayes-optimal rule by using estimators of the corresponding means and covariance matrices in (6.15). Results illustrated on Figure 6.2b via the reliability curves indicate that shift-corrected binning with either true or estimated importance weights yields a calibrated predictor on the target domain while uncorrected fails to do so as expected. To complete the empirical study, Appendix E.3.2 further examines calibration with and without accounting for label shift on the `wine quality` dataset from Section 6.2.2.

6.4 Discussion

For safety-critical applications model's prediction must be supported with a proper measure of uncertainty. As various ad-hoc procedures provide valid inference only under assumptions that are either unrealistic or unverifiable, it is essential to understand whether non-trivial guarantees can be obtained in an assumption-lean manner. Guided by this

principle, we analyzed distribution-free uncertainty quantification for classification via two complementary notions: prediction sets and calibration.

We focused on a less studied — but still highly relevant to real-world scenarios — setting of label shift. While it is evident that label shift does hurt model's calibration, the corresponding impact on prediction sets is less obvious. In the extreme example of almost perfectly separable data, prediction sets are usually expected to contain the most likely label only, and thus coverage is not expected to suffer much no matter how the class proportions change for the test data. Still, as we illustrated, in less idealized settings, a correction for label shift is necessary. By adapting conformal prediction sets and calibration via binning to label shift, we close an existing gap for distribution-free uncertainty quantification under two standard ways of generalizing beyond the classic i.i.d. setting. Importantly, those adaptations do not require labeled data from the target domain which can be useful in applications where the labeling process is expensive. We note that handling label shift should be expected to be an easier task rather than handling another common setting — covariate shift — as the latter typically involves estimating a high-dimensional, and usually continuous, likelihood ratio.

With theoretical results available for calibration in the binary setting, and thus class-wise (coordinatewise) calibration in a more general multiclass setting, establishing meaningful guarantees for "full" canonical calibration in the latter setting remains an intriguing future research direction. One particular example is related to the question of the importance weights estimation under label shift. While approaches based on confusion matrices, e.g., BBSE and RLLS, provably yield consistent estimators under relatively mild assumptions, alternative approaches, such as MLLS with preceding ad-hoc calibration on the source domain, tend to perform better empirically (Alexandari et al., 2020). Theoretical foundations for MLLS developed recently by Garg et al. (2020) require the underlying predictor to be canonically calibrated which is itself, unfortunately, hard to guarantee provably which creates a (somewhat circular) gap between theory and practice.

Bibliography

Alexandari, A., Kundaje, A., and Shrikumar, A. (2020). Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International Conference on Machine Learning*. 61, 63, 76, 192

Angelopoulos, A. N., Bates, S., Malik, J., and Jordan, M. I. (2021). Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*. 64, 66, 193

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2007). Tuning bandit algorithms in stochastic environments. In *International Conference on Algorithmic Learning Theory*. 56, 189

Azizzadenesheli, K., Liu, A., Yang, F., and Anandkumar, A. (2019). Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations*. 63, 75, 191

Balsubramani, A. and Ramdas, A. (2016). Sequential nonparametric testing with the law of the iterated logarithm. In *Uncertainty in Artificial Intelligence*. 8

Barber, R. F. (2020). Is distribution-free inference possible for binary regression? *Electronic Journal of Statistics*. 49, 51, 53

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, J. R. (2021). Predictive inference with the jackknife+. *The Annals of Statistics*. 64

Bates, S., Angelopoulos, A. N., Lei, L., Malik, J., and Jordan, M. I. (2021). Distribution-free, risk-controlling prediction sets. *Journal of the ACM*. 45, 157, 167

Berrett, T. B. and Samworth, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika*. 148

Besserve, M., Logothetis, N. K., and Schölkopf, B. (2013). Statistical analysis of coupled time series with kernel cross-spectral density operators. In *Advances in Neural Information Processing Systems*. 15

Bickel, S., Brückner, M., and Scheffer, T. (2007). Discriminative learning for differing training and test distributions. In *International Conference on Machine Learning*. 59

Breiman, L. (1962). Optimal gambling systems for favorable games. *Berkeley Symposium on Mathematical Statistics and Probability*. 104

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*. 37, 42, 49, 60, 156

Bröcker, J. (2012). Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate dynamics*. 60

Cauchois, M., Gupta, S., and Duchi, J. C. (2021). Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*. 64, 66

Cheng, X. and Cloninger, A. (2022). Classification logit two-sample testing by neural networks for differentiating near manifold densities. *IEEE Transactions on Information Theory*. 22

Chwialkowski, K. and Gretton, A. (2014). A kernel independence test for random processes. In *International Conference on Machine Learning*. 14

Chwialkowski, K. P., Sejdinovic, D., and Gretton, A. (2014). A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems*. 15

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*. 68, 199, 206

Cutkosky, A. and Orabona, F. (2018). Black-box reductions for parameter-free online learning in banach spaces. In *Conference on Learning Theory*. 12, 25, 100, 131

Darling, D. A. and Robbins, H. E. (1968). Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences*. 6, 22

Dawid, A. P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*. 49, 60

Dawid, A. P. (2014). Probability forecasting. *Wiley StatsRef: Statistics Reference Online*. 60

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*. 60

Ernst, P. A., Shepp, L. A., and Wyner, A. J. (2017). Yule's "nonsense correlation" solved! *The Annals of Statistics*. 7

Fan, X., Grama, I., and Liu, Q. (2015). Exponential inequalities for martingales with applications. *Electronic Journal of Probability*. 98

Ferro, C. A. and Fricker, T. E. (2012). A bias-corrected decomposition of the Brier score. *Quarterly Journal of the Royal Meteorological Society*. 60

Friedman, J. H. (2004). On multivariate goodness-of-fit and two-sample testing. Technical report, Stanford University. 21

Fukumizu, K., Bach, F. R., and Gretton, A. (2007a). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*. 16

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007b). Kernel measures of conditional dependence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems*. 11

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*. 36

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. (2020). A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*. 61, 76, 192

Gretton, A. (2015). A simpler condition for consistency of a kernel independence test. *arXiv preprint: 1501.06103*. 11

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*. 21, 34

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005a). Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*. 5, 6, 10, 11, 21, 92

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005b). Kernel methods for measuring independence. *Journal of Machine Learning Research*. 15, 103

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. (2005c). Kernel constrained covariance for dependence measurement. In *Tenth International Workshop on Artificial Intelligence and Statistics*. 5, 6, 15, 103

Grünwald, P., Henzi, A., and Lardy, T. (2023). Anytime-valid tests of conditional independence under model-X. *Journal of the American Statistical Association*. 8, 22

Grünwald, P., de Heide, R., and Koolen, W. M. (2020). Safe testing. In *Information Theory and Applications Workshop*. 8

Guan, L. and Tibshirani, R. (2022). Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society (Series B)*. 64, 69, 199

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*. 49, 55, 60, 61, 71

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. K. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*. 64, 65

Gupta, C., Podkopaev, A., and Ramdas, A. (2020). Distribution-free binary classification: prediction sets, confidence intervals and calibration. In *Advances in Neural Information Processing Systems*. 3, 35, 63, 71, 73, 75

Gupta, C. and Ramdas, A. (2021). Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*. 73

Gupta, C. and Ramdas, A. (2022). Top-label calibration and multiclass-to-binary reductions. In *International Conference on Learning Represention*s. 157

Hazan, E., Agarwal, A., and Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*. 12, 25

Hediger, S., Michel, L., and Näf, J. (2022). On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis*. 22

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*. 43

Hendrycks, D., Mazeika, M., and Dietterich, T. (2019). Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*. 61

Hoeffding, W. (1961). The strong law of large numbers for U–statistics. Technical report, University of North Carolina. 95

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2020). Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*. 58, 179

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*. 36, 39, 43, 58, 161, 163, 164, 179, 189

Hu, X. and Lei, J. (2020). A distribution-free test of covariate shift using conformal prediction. *arXiv: 2010.07147*. 34

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*. 59

Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. (2016). Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems*. 29

Jordan, M. I. and Bach, F. R. (2001). Kernel independent component analysis. *Journal of Machine Learning Research*. 5, 6, 16, 103

Kamulete, V. M. (2022). Test for non-negligible adverse shifts. In *Uncertainty in Artificial Intelligence*. 36

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*. 59, 186

Kelly, J. L. (1956). A new interpretation of information rate. *The Bell System Technical Journal*. 104

Kendall, A. and Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*. 61

Kim, I., Lee, A. B., and Lei, J. (2019). Global and local two-sample tests via regression. *Electronic Journal of Statistics*. 22

Kim, I. and Ramdas, A. (2020). Dimension-agnostic inference using cross U-statistics. *Bernoulli*. 27

Kim, I., Ramdas, A., Singh, A., and Wasserman, L. (2021). Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*. 21

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto. 43

Kübler, J. M., Stimper, V., Buchholz, S., Muandet, K., and Schölkopf, B. (2022). AutoML two-sample test. In *Advances in Neural Information Processing Systems*. 22, 27

Kuleshov, V., Fenner, N., and Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*. 61

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*. 71

Kull, M., Silva Filho, T. M., and Flach, P. (2017). Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*. 55

Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*. 57, 60, 71, 73, 75, 190

Kumar, A., Sarawagi, S., and Jain, U. (2018). Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*. 61

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*. 61

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. 31, 113

Lee, K., Lee, H., Lee, K., and Shin, J. (2018). Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*. 61

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*. 64

Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*. 65

Lhéritier, A. and Cazals, F. (2018). A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*. 23, 27

Lhéritier, A. and Cazals, F. (2019). Low-complexity nonparametric bayesian online prediction with universal guarantees. In *Advances in Neural Information Processing Systems*. 23, 27, 28

Lipton, Z. C., Wang, Y., and Smola, A. J. (2018). Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*. 63, 68, 75, 191, 192, 199

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. (2020). Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*. 22

Lopez-Paz, D. and Oquab, M. (2017). Revisiting classifier two-sample tests. In *International Conference on Learning Representations*. 21, 29

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*. 40, 41

Lundqvist, D., Flykt, A., and Öhman, A. (1998). The karolinska directed emotional faces - KDEF. Technical report, Karolinska Institutet. 29

Milios, D., Camoriano, R., Michiardi, P., Rosasco, L., and Filippone, M. (2018). Dirichlet-based gaussian processes for large-scale calibrated classification. In *Advances in Neural Information Processing Systems*. 61

Mu, N. and Gilmer, J. (2019). MNIST-C: A robustness benchmark for computer vision. *arXiv preprint:1906.02337*. 43, 45

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*. 10

Murphy, A. H. (1972a). Scalar and vector partitions of the probability score: Part i. two-state situation. *Journal of Applied Meteorology*. 60

Murphy, A. H. (1972b). Scalar and vector partitions of the probability score: Part ii. n-state situation. *Journal of Applied Meteorology*. 60

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of applied Meteorology*. 60

Murphy, A. H. and Epstein, E. S. (1967). Verification of probabilistic predictions: A brief review. *Journal of Applied Meteorology*. 49, 60

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*. 60

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*. 61

Pandeva, T., Bakker, T., Naesseth, C. A., and Forré, P. (2022). E-valuating classifier two-sample tests. *arXiv preprint: 2210.13027*. 23, 27, 128

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *European Conference on Machine Learning*. 51

Park, S., Bastani, O., Weimer, J., and Lee, I. (2020). Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*. 60

Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*. 49, 52, 55, 61, 71

Podkopaev, A., Blöbaum, P., Kasiviswanathan, S. P., and Ramdas, A. (2023). Sequential kernelized independence testing. In *International Conference on Machine Learning*. 3, 22, 23, 30, 31, 33, 122, 129, 148

Podkopaev, A. and Ramdas, A. (2021). Distribution free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*. 3, 35, 153

Podkopaev, A. and Ramdas, A. (2022). Tracking the risk of a deployed model and detecting harmful distribution shifts. In *International Conference on Learning Representations*. 3

Podkopaev, A. and Ramdas, A. (2023). Sequential predictive two-sample and independence testing. *arXiv preprint: 2305.00143*. 3

Proedrou, K., Nouretdinov, I., Vovk, V., and Gammerman, A. (2002). Transductive confidence machines for pattern recognition. In *European Conference on Machine Learning*. 51

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset Shift in Machine Learning*. The MIT Press. 1, 34

Rabanser, S., Günnemann, S., and Lipton, Z. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*. 34

Ramdas, A., Jakkam Reddi, S., Poczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. *AAAI Conference on Artificial Intelligence*. 21

Ramdas, A. and Manole, T. (2023). Randomized and exchangeable improvements of Markov's, Chebyshev's and Chernoff's inequalities. *arXiv preprint: 2304.02611*. 24

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint: 2009.03167*. 18, 122

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*. 8

Reynolds Jr., M. R. (1975). A sequential signed-rank test for symmetry. *The Annals of Statistics*. 18

Romano, Y., Patterson, E., and Candès, E. (2019). Conformalized quantile regression. In *Advances in Neural Information Processing Systems*. 64

Romano, Y., Sesia, M., and Candès, E. J. (2020). Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*. 64, 65, 153, 193, 194, 195

Sadinle, M., Lei, J., and Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*. 64, 65, 69, 199

Saerens, M., Latinne, P., and Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*. 1, 34, 61, 63, 73, 192

Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*. 49, 60

Seo, S., Seo, P. H., and Han, B. (2019). Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *IEEE Conference on Computer Vision and Pattern Recognition*. 61

Shaer, S., Maman, G., and Romano, Y. (2023). Model-free sequential testing for conditional independence via testing by betting. In *Conference on Artificial Intelligence and Statistics*. 8, 18, 22, 122

Shafer, G. (2021). Testing by betting: a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 8, 22, 23

Shafer, G. and Vovk, V. (2019). *Game-Theoretic Foundations for Probability and Finance*. Wiley Series in Probability and Statistics. Wiley. 8, 23

Shekhar, S. and Ramdas, A. (2021). Nonparametric two-sample testing by betting. *arXiv preprint: 2112.09162*. 6, 8, 9, 18, 22, 23, 29, 33, 89, 90, 121, 122, 129, 145

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*. 1, 34, 58, 63

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*. 27

Tibshirani, R. J., Barber, R. F., Candès, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*. 35, 60, 63, 68, 195, 206, 208

Tran, G.-L., Bonilla, E. V., Cunningham, J., Michiardi, P., and Filippone, M. (2019). Calibrating deep convolutional gaussian processes. In *International Conference on Artificial Intelligence and Statistics*. 61

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. B. (2019). Evaluating model calibration in classification. In *International Conference on Artificial Intelligence and Statistics*. 49, 60, 71, 171

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer. 206

Ville, J. (1939). *Étude critique de la notion de collectif*. Thèses de l'entre-deux-guerres. Gauthier-Villars. 8, 9, 23, 92, 131

Vovk, V. (2020a). Testing for concept shift online. *arXiv: 2012.14246*. 34, 152

Vovk, V. (2020b). Testing randomness online. *Statistical Science*. 34, 152

Vovk, V., Fedorova, V., Nouretdinov, I., and Gammerman, A. (2016). Criteria of efficiency for conformal prediction. In *Symposium on Conformal and Probabilistic Prediction with Applications*. 64, 69, 199

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer. 34, 45, 49, 51, 64, 65, 69, 154, 157, 193, 199

Vovk, V. and Petej, I. (2014). Venn-Abers predictors. In *Conference on Uncertainty in Artificial Intelligence*. 49

Vovk, V., Petej, I., Nouretdinov, I., Ahlberg, E., Carlsson, L., and Gammerman, A. (2021). Retrain or not retrain: conformal test martingales for change-point detection. In *Symposium on Conformal and Probabilistic Prediction with Applications*. 34, 152, 153, 155, 156

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*. 38

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*. 23, 127

Waudby-Smith, I. and Ramdas, A. (2023). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society, Series B (Methodology), with discussion*. 7, 36, 39, 42, 105, 161, 162

Wenger, J., Kjellström, H., and Triebel, R. (2020). Non-parametric calibration for classification. In *International Conference on Artificial Intelligence and Statistics*. 61

Widmann, D., Lindsten, F., and Zachariah, D. (2019). Calibration tests in multi-class classification: a unifying framework. In *Advances in Neural Information Processing Systems*. 60

Wu, Y., Winston, E., Kaushik, D., and Lipton, Z. (2019). Domain adaptation with asymmetrically-relaxed distribution alignment. In *International Conference on Machine Learning*. 36

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series. *Journal of the Royal Statistical Society*. 7

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*. 49, 55, 56, 61, 71

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *International Conference on Knowledge Discovery and Data Mining*. 61, 71

Zhang, J., Kailkhura, B., and Han, T. (2020). Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*. 60

Appendix

Appendix A

Additional Results for Chapter 2

A.1 Independence Testing for Streaming Data

In Section A.1.1, we describe a permutation-based approach for conducting batch HSIC and show that continuous monitoring of batch HSIC (without corrections for multiple testing) leads to an inflated false alarm rate. In Section A.1.2, we introduce the sequential two-sample testing (2ST) problem and describe a reduction of sequential independence testing to sequential 2ST. In Section A.1.3, we compare our test to HSIC in the batch setting.

A.1.1 Failure of Batch HSIC under Continuous Monitoring

To conduct independence testing using batch HSIC, we use permutation p-value (with $M = 1000$ random permutations): $P = \frac{1}{M+1}(1 + \sum_{m=1}^{M} \mathbb{1}\{T_m \geq T\})$, where T_m is the value of HS-norm computed from the m-th permutation and T is HS-norm value on the original data. In other words, suppose that we are given a sample $Z_1, \ldots, Z_t$, where $Z_i = (X_i, Y_i)$. Let $\mathcal{S}_t$ denote the set of all permutations of t indices and let $\sigma \sim \mathrm{Unif}(\mathcal{S}_t)$ be a random permutation of indices. Then:

$$
\begin{aligned}
(X_1, Y_1), \ldots, (X_t, Y_t) &\Longrightarrow T = \widehat{\mathrm{HSIC}}_b\left((X_1, Y_1), \ldots, (X_t, Y_t)\right) \\
(X_1, Y_{\sigma_m(1)}), \ldots, (X_t, Y_{\sigma_m(t)}), &\Longrightarrow T_m = \widehat{\mathrm{HSIC}}_b\left((X_1, Y_{\sigma_m(1)}), \ldots, (X_t, Y_{\sigma_m(t)})\right), \quad m \in \{1, \ldots, M\},
\end{aligned}
$$

where we use a biased estimator of HSIC:

$$
\widehat{\mathrm{HSIC}}_b = \frac{1}{t^2}\sum_{i,j} K_{ij}L_{ij} + \frac{1}{t^4}\sum_{i,j,q,r} K_{ij}L_{qr} - \frac{2}{t^3}\sum_{i,j,q} K_{ij}L_{iq} = \frac{1}{t^2}\mathrm{tr}KHLH.
$$

For brevity, we use $K_{ij} = k(X_i, X_j)$, $L_{ij} = l(Y_i, Y_j)$ for $i, j \in \{1, \ldots, t\}$. Next, we study batch HSIC under (a) *fixed-time* and (b) *continuous* monitoring. We consider a simple case when X and Y are independent standard Gaussian random variables. We consider (re)conducting a test at 12 different sample sizes: $t \in \{50, 100, \ldots, 600\}$:

(a) Under fixed-time monitoring, for each value of t, we sample a sequence $Z_1, \ldots, Z_t$ (100 times for each t) and conduct batch-HSIC test. The goal is to confirm that batch-HSIC controls type I error by tracking the standard miscoverage rate.

(b) Under continuous monitoring, we sample new datapoints and re-conduct the test. We illustrate inflated type I error by tracking the *cumulative miscoverage rate*, that is, the fraction of times, the test falsely rejects the independence null.

The results are presented in Figure A.1. For Bonferroni correction, we decompose the error budget as: $\alpha = \sum_{i=1}^{\infty} \frac{\alpha}{i(i+1)}$, that is, for t-th test we use threshold $\alpha_t = \alpha/(t(t+1))$ for testing.

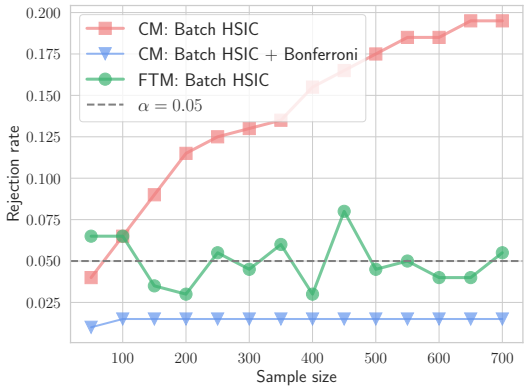


Figure A.1: Inflated false alarm rate of batch HSIC under continuous monitoring (CM, red line with squares) for the case when X and Y are independent standard Gaussian random variables. Bonferroni correction (CM, blue line with triangles) restores type I error control. As expected, type I error is controlled at a specified level under fixed-time monitoring (FTM, green line with circles).

A.1.2 Sequential Independence Testing via Sequential Two-Sample Testing

First, we introduce the sequential two-sample testing problem. Suppose that we observe a stream of data: $(\tilde{X}_1, \tilde{Y}_1), (\tilde{X}_2, \tilde{Y}_2), \ldots$, where $(\tilde{X}_t, \tilde{Y}_t) \overset{\text{iid}}{\sim} P \times Q$. Two-sample testing refers to testing:

$$H_0 : (\tilde{X}_t, \tilde{Y}_t) \overset{\text{iid}}{\sim} P \times Q \text{ and } P = Q, \quad \text{vs.} \quad H_1 : (\tilde{X}_t, \tilde{Y}_t) \overset{\text{iid}}{\sim} P \times Q \text{ and } P \neq Q.$$

In Figure 2.1, we compared our test against the approach based on the reduction of independence testing to two-sample testing. We used the sequential two-sample kernel MMD test of Shekhar and Ramdas (2021) with the product kernel $\tilde{K}$ (that is, a product of Gaussian kernels) and the same set of hyperparameters as for our test for a fair

comparison. To reduce sequential independence testing to any off-the-shelf sequential two-sample testing procedure, we convert the original sequence of points from P_{XY} to a sequence of i.i.d. $(\tilde{X}_t, \tilde{Y}_t)$-pairs, where $\tilde{X}_t \sim P_{XY}$ and $\tilde{Y}_t \sim P_X \times P_Y$ respectively; see Figure A.2a. At t-th round, we randomly choose one point as $\tilde{X}_t$, e.g., (X_1, Y_1) for the first triple. Next, we obtain $\tilde{Y}_t$ by randomly matching X and Y from two other pairs, e.g., (X_2, Y_3) or (X_3, Y_2) for the first triple. In fact, the betting-based sequential two-sample test of (Shekhar and Ramdas, 2021) allows removing the effect of randomization (i.e., throwing away one observation in each triple), by averaging payoffs evaluated on $(\tilde{X}_t, \tilde{Y}_t^{(1)})$ and $(\tilde{X}_t, \tilde{Y}_t^{(2)})$. Other approaches — which do not require throwing data away — are also available (Figures A.2b) but those only yield an i.i.d. sequence only under the null.

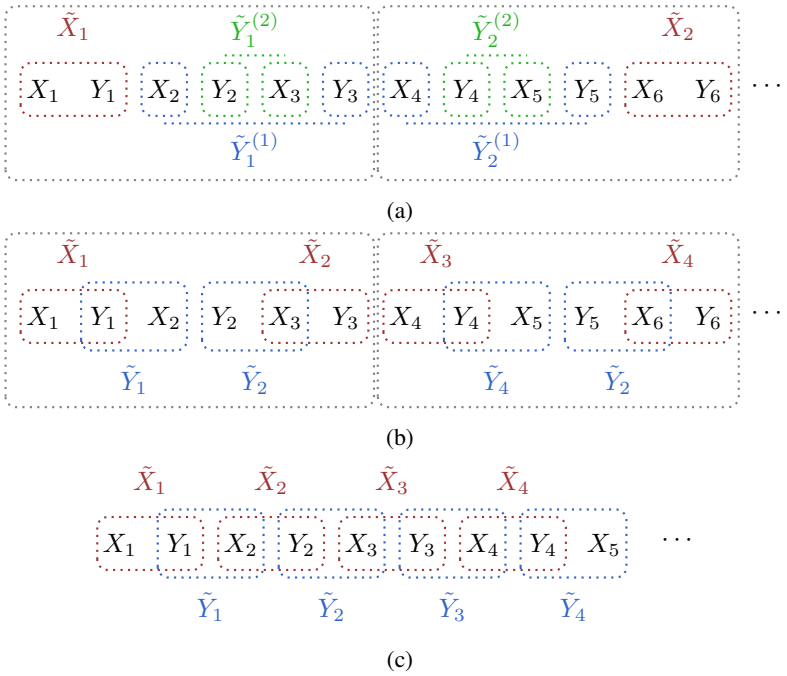


Figure A.2: Reducing sequential independence testing to sequential two-sample testing. Processing as per (a) results in a sequence of i.i.d. observations both under the null and under the alternative (making the results about power valid). Processing data as per (b) gives an i.i.d. sequence only under the null. Reduction (b) is very similar to reduction (c). However, the latter makes $\tilde{X}_i$, $i \geq 2$, dependent on the past, and thus can not be used directly for considered sequential two-sample tests.

Additional Details of the Simulation Presented in Figure 2.1. We consider the Gaussian model: $Y_t = X_t\beta + \varepsilon_t$, where $X_t, \varepsilon_t \sim \mathcal{N}(0,1)$, $t \geq 1$. We consider 10 values of β: $\beta \in \{0, 0.04, \ldots, 0.36\}$, and for each β we repeat the simulation 100 times. In this simulation, we compare three approaches for testing independence (valid under continuous monitoring):

1. **HSIC-based SKIT** proposed in this work (Algorithm 2);

2. **Batch HSIC** adapted to continuous monitoring via Bonferroni correction. We allow monitoring after processing every n, $n \in \{10, 100\}$, new points from P_{XY}, that is, the permutation p-value (computed over 2500 randomly sampled permutations) is compared against rejection thresholds: $\alpha_n = \alpha/(n(n+1))$, $n = 1, 2, \ldots$

3. **Sequential independence testing via reduction sequential 2ST** as described above.

We use RBF kernel with the same set of kernel hyperparameters for all testing procedures: $\lambda_X = 1/4$, $\lambda_Y = 1/(4(1+\beta^2))$.

A.1.3 Comparison in the Batch Setting

Sequential tests are complementary to batch tests and are not intended to replace them, and hence comparing the two on equal footing is hard. To highlight this, consider two simple scenarios. If we have 2000 data points, and HSIC fails to reject, there is not much we can do to rescue the situation. But if SKIT fails to reject, an analyst can collect more data and continue testing, retaining type I error control. In contrast, with 2 million points, HSIC will take forever to run, especially due to permutations. But if the alternative is true and the signal is strong, then SKIT may reject within 200 samples and stop. In short, the ability of SKIT to continue collecting and analyzing data is helpful for hard problems, and the ability of SKIT to stop early is helpful for easy problems. There is no easy sense in which one can compare them apples to apples and there is no sense in which batch HSIC uniformly dominates SKIT or vice versa. In a real setting, if an analyst has a strong hunch that the null is false and has the ability to collect data and run HSIC, the question is how much data should be collected? The answer depends on the underlying data distribution, which is of course unknown. With SKIT, data can be collected and analyzed sequentially. Theorem 2.2 implies that SKIT will stop early on easy problems and later on harder problems, all without knowing anything about the problem in advance. If however, one has a fixed batch of data, no chance to collect more, and no computational constraints, then running HSIC makes more sense.

To illustrate that batch HSIC can be superior to SKIT, we compare tests on a dataset with a prespecified sample size (500 observations from the Gaussian model) and track the empirical rejection rates of two tests. In Figure A.3, we show that HSIC actually has higher power than SKIT. However, for $\beta = 0.1$ (where all tests have low power), Figure 2.3a shows that collecting just a bit more data (which is allowed) is needed for SKIT to reach perfect power. We also added a third method (D-SKIT) which removes the effect of the ordering of random variables under the assumptions that $\{(X_i, Y_i)\}_{i=1}^n$ are independent draws from P_{XY}. Let $\{\sigma_b\}_{b=1}^B$ define B random permutations of n indices, and let $\mathcal{K}_n^b$ denote the wealth after betting on a sequence ordered according to σ_b. For each b, $\mathcal{K}_n^b$ has expectation at most one, and hence (by linearity of expectation and Markov's inequality) $\mathbb{1}\left\{\frac{1}{B}\sum_{i=1}^B \mathcal{K}_n^b \geq 1/\alpha\right\}$ is a valid level-α batch test. This test is a bit more stable: it improves SKIT's power on moderate-complexity setups at the cost of a slight power loss on more extreme ones.

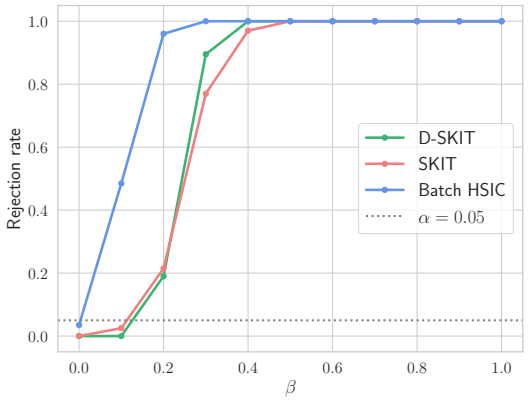


Figure A.3: Comparison of SKIT and HSIC under Gaussian model in the batch setting. Non-surprisingly, batch HSIC performs best. D-SKIT improves over SKIT's power on moderate-complexity setups at the cost of a slight power loss on more extreme ones.

A.2 Proofs

Section A.2.1 contains auxiliary results needed to prove the results presented in this paper. In Section A.2.2, we prove the results from Section 2.2. In Secton A.2.3, we prove the results from Section 2.3.

A.2.1 Auxiliary Results

Theorem A.1 (Ville's inequality (Ville, 1939))**.** *Suppose that* $(\mathcal{M}_t)_{t\geq 0}$ *is a nonnegative supermartingale process adapted to a filtration* $\{\mathcal{F}_t : t \geq 0\}$*. Then, for any* $a > 0$ *it holds that:*

$$\mathbb{P}\left(\exists t \geq 1 : \mathcal{M}_t \geq a\right) \leq \frac{\mathbb{E}\left[\mathcal{M}_0\right]}{a}.$$

Theorem A.2 (Theorem 3 in (Gretton et al., 2005a))**.** *Assume that* k *and* l *are bounded almost everywhere by 1, and are nonnegative. Then for* $n > 1$ *and any* $\delta \in (0, 1)$*, it holds with probability at least* $1 - \delta$ *that:*

$$\left|\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}) - \widehat{\mathrm{HSIC}}_b(P_{XY}; \mathcal{G}, \mathcal{H})\right| \leq \sqrt{\frac{\log(6/\delta)}{\alpha^2 n}} + \frac{C}{n},$$

where $\alpha^2 > 0.24$ *and* C *are some absolute constants.*

A.2.2 Proofs for Section 2.2

In Section A.2.2, we prove several intermediate results. The proofs of the main results are deferred to Section A.2.2.

Supporting Lemmas

Before we state the first result, recall the definition of the empirical mean embeddings computed from the first $2(t-1)$ datapoints:

$$\begin{aligned} \widehat{\mu}_{XY}^{(t)} &= \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i) \otimes \psi(Y_i), \\ \widehat{\mu}_X^{(t)} = \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \varphi(X_i), &\quad \widehat{\mu}_Y^{(t)} = \frac{1}{2(t-1)} \sum_{i=1}^{2(t-1)} \psi(Y_i), \end{aligned} \tag{A.1}$$

where we highlight the dependence on the number of processed datapoints. We have the following result.

Lemma A.2.1. *For the empirical* (A.1) *and population* (2.9) *mean embeddings, it holds that:*

$$\left\| \widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)} \right\|_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} \left\| \mu_{XY} - \mu_X \otimes \mu_Y \right\|_{\mathcal{G}\otimes\mathcal{H}}. \tag{A.2}$$

Proof. We have

$$\begin{aligned} \left\| \mu_{XY} - \mu_X \otimes \mu_Y \right\|_{\mathcal{G}\otimes\mathcal{H}}^2 &= \mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H}), \\ \left\| \widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)} \right\|_{\mathcal{G}\otimes\mathcal{H}}^2 &= \widehat{\mathrm{HSIC}}_b^{(t)}(P_{XY}; \mathcal{G}, \mathcal{H}), \end{aligned}$$

where the latter is a biased estimator of HSIC, computed from $2(t-1)$ datapoints from P_{XY}. From Theorem A.2 and the Borel-Cantelli lemma, it follows that:

$$\left\| \widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)} \right\|_{\mathcal{G}\otimes\mathcal{H}}^2 \xrightarrow{\text{a.s.}} \left\| \mu_{XY} - \mu_X \otimes \mu_Y \right\|_{\mathcal{G}\otimes\mathcal{H}}^2.$$

The result then follows from the continuous mapping theorem. □

Lemma A.2.2. *Suppose that H_1 in* (2.1b) *is true. Then for the oracle* (2.11) *and plug-in* (2.13) *witness functions, it holds that:*

$$\langle \widehat{g}_t, g^\star \rangle_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} 1. \tag{A.3}$$

As a consequence, $\left\| \widehat{g}_t - g^\star \right\|_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} 0$.

Proof. Suppose that the alternative in (2.1b) happens to be true. Then since k and l are characteristic kernels, it follows that:

$$\left\| \mu_{XY} - \mu_X \otimes \mu_Y \right\|_{\mathcal{G}\otimes\mathcal{H}} > 0.$$

We aim to show that:

$$\left\langle \frac{\widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)}}{\left\| \widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)} \right\|_{\mathcal{G}\otimes\mathcal{H}}}, \frac{\mu_{XY} - \mu_X \otimes \mu_Y}{\left\| \mu_{XY} - \mu_X \otimes \mu_Y \right\|_{\mathcal{G}\otimes\mathcal{H}}} \right\rangle_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} 1.$$

From Lemma A.2.1, we know that: $\left\|\widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)}\right\|_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} \|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G}\otimes\mathcal{H}}$. Hence it suffices to show that

$$\left\langle \widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)}, \mu_{XY} - \mu_X \otimes \mu_Y \right\rangle_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} \|\mu_{XY} - \mu_X \otimes \mu_Y\|^2_{\mathcal{G}\otimes\mathcal{H}}\,. \tag{A.4}$$

Recall that: $\mu_{XY} - \mu_X \otimes \mu_Y = \mathbb{E}\left[\varphi(\tilde{X}) \otimes \psi(\tilde{Y})\right] - \mathbb{E}\left[\varphi(\tilde{X})\right] \otimes \mathbb{E}\left[\psi(\tilde{Y})\right]$. We have:

$$\widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)} = \left(1 - \frac{1}{2(t-1)}\right)\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} \varphi(X_i)\otimes\psi(Y_i) - \frac{1}{4(t-1)^2 - 2(t-1)} \sum_{\substack{j,k=1:\\ j\neq k}}^{2(t-1)} \varphi(X_j)\otimes\psi(Y_k)\right).$$

Further, it holds that:

$$\begin{aligned}
&\left\langle \widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)}, \mu_{XY} - \mu_X \otimes \mu_Y \right\rangle_{\mathcal{G}\otimes\mathcal{H}} \\
= \quad & \left(1 - \frac{1}{2(t-1)}\right)\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\langle\varphi(\tilde{X}),\varphi(X_i)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(Y_i)\rangle_{\mathcal{H}}\right]\right) \\
- \quad & \left(1 - \frac{1}{2(t-1)}\right)\left(\frac{1}{4(t-1)^2 - 2(t-1)} \sum_{\substack{j,k=1:\\ j\neq k}}^{2(t-1)} \mathbb{E}_{\tilde{X}}\left[\langle\varphi(\tilde{X}),\varphi(X_j)\rangle_{\mathcal{G}}\right]\mathbb{E}_{\tilde{Y}}\left[\langle\psi(\tilde{Y}),\psi(Y_k)\rangle_{\mathcal{H}}\right]\right),
\end{aligned}$$

For any $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have:

$$\begin{aligned}
\left|\mathbb{E}_{\tilde{X},\tilde{Y}}\left[\langle\varphi(\tilde{X}),\varphi(x)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(y)\rangle_{\mathcal{H}}\right]\right| &\leq \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\left|\langle\varphi(\tilde{X}),\varphi(x)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(y)\rangle_{\mathcal{H}}\right|\right] \\
&\leq \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\sqrt{k(\tilde{X},\tilde{X})k(x,x)l(\tilde{Y},\tilde{Y})k(y,y)}\right] \\
&\leq 1,
\end{aligned}$$

and similarly, for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ it holds that:

$$\left|\mathbb{E}_{\tilde{X}}\left[\langle\varphi(\tilde{X}),\varphi(x)\rangle_{\mathcal{G}}\right]\mathbb{E}_{\tilde{Y}}\left[\langle\psi(\tilde{Y}),\psi(y)\rangle_{\mathcal{H}}\right]\right| \leq 1.$$

Hence, by the SLLN, it follows that ($(X, Y), (\tilde{X}, \tilde{Y}) \overset{\text{iid}}{\sim} P_{XY}$):

$$\begin{aligned}
\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} \mathbb{E}_{\tilde{X},\tilde{Y}}\left[\langle\varphi(\tilde{X}),\varphi(X_i)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(Y_i)\rangle_{\mathcal{H}}\right] \xrightarrow{\text{a.s.}} \quad & \mathbb{E}_{X,Y,\tilde{X},\tilde{Y}}\left[\langle\varphi(\tilde{X}),\varphi(X)\rangle_{\mathcal{G}}\langle\psi(\tilde{Y}),\psi(Y)\rangle_{\mathcal{H}}\right] \\
&= \langle\mu_{XY},\mu_{XY}\rangle_{\mathcal{G}\otimes\mathcal{H}}\,.
\end{aligned}$$

Similarly, by the SLLN for U-statistics with bounded kernel (Hoeffding, 1961), it follows that:

$$\begin{aligned}
&\frac{1}{4(t-1)^2-2(t-1)}\sum_{\substack{j,k=1:\\ j\neq k}}^{2(t-1)}\mathbb{E}_{\tilde{X}}\left[\langle\varphi(\tilde{X}),\varphi(X_j)\rangle_{\mathcal{G}}\right]\mathbb{E}_{\tilde{Y}}\left[\langle\psi(\tilde{Y}),\psi(Y_k)\rangle_{\mathcal{H}}\right]\\
\xrightarrow{\text{a.s.}}\;&\mathbb{E}_{X,\tilde{X}}\left[\langle\varphi(\tilde{X}),\varphi(X)\rangle_{\mathcal{G}}\right]\mathbb{E}_{Y,\tilde{Y}}\left[\langle\psi(\tilde{Y}),\psi(Y)\rangle_{\mathcal{H}}\right]\\
=\;&\langle\mu_X\otimes\mu_Y,\mu_X\otimes\mu_Y\rangle_{\mathcal{G}\otimes\mathcal{H}}\,.
\end{aligned}$$

Hence, we deduce that:

$$\begin{aligned}
\left\langle\widehat{\mu}_{XY}^{(t)}-\widehat{\mu}_X^{(t)}\otimes\widehat{\mu}_Y^{(t)},\mu_{XY}-\mu_X\otimes\mu_Y\right\rangle_{\mathcal{G}\otimes\mathcal{H}}&\xrightarrow{\text{a.s.}}\langle\mu_{XY},\mu_{XY}\rangle_{\mathcal{G}\otimes\mathcal{H}}-\langle\mu_X\otimes\mu_Y,\mu_X\otimes\mu_Y\rangle_{\mathcal{G}\otimes\mathcal{H}}\\
&=\langle\mu_{XY}-\mu_X\otimes\mu_Y,\mu_{XY}-\mu_X\otimes\mu_Y\rangle_{\mathcal{G}\otimes\mathcal{H}}\\
&=\|\mu_{XY}-\mu_X\otimes\mu_Y\|^2_{\mathcal{G}\otimes\mathcal{H}}\,.
\end{aligned}$$

Recalling (A.4), the proof of (A.3) is complete. To establish the consequence, simply note that:

$$\|\widehat{g}_t-g^\star\|_{\mathcal{G}\otimes\mathcal{H}}=\sqrt{2\left(1-\langle\widehat{g}_t,g^\star\rangle_{\mathcal{G}\otimes\mathcal{H}}\right)},$$

and hence the result follows. □

Lemma A.2.3. *Suppose that $(x_t)_{t\geq1}$ is a sequence of numbers such that $\lim_{t\to\infty}x_t=x$. Then the corresponding sequence of partial averages also converges to x, that is, $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n x_t=x$. This also implies that if $(X_t)_{t\geq1}$ is a sequence of random variables such that $X_t\xrightarrow{\text{a.s.}}X$, then $(\sum_{t=1}^n X_t)/n\xrightarrow{\text{a.s.}}X$.*

Proof. Fix any $\varepsilon>0$. Since $(x_t)_{t\geq1}$ is converging, then $\exists M>0$:

$$|x_t-x|\leq M,\quad\forall t\geq1.$$

Now, let n_0 be such that $|x_t-x|\leq\varepsilon/2$ for all $n>n_0$. Further, choose any $n_1>n_0$: $Mn_0/n_1\leq\varepsilon/2$. Hence, for any $\tilde{n}>n_1$, it holds that:

$$\begin{aligned}
\left|\frac{1}{\tilde{n}}\sum_{t=1}^{\tilde{n}}x_t-x\right|&\leq\left|\frac{1}{\tilde{n}}\sum_{t=1}^{n_0}x_t-x\right|+\left|\frac{1}{\tilde{n}}\sum_{t=n_0+1}^{\tilde{n}}x_t-x\right|\\
&\leq\frac{1}{\tilde{n}}\sum_{t=1}^{n_0}|x_t-x|+\frac{1}{\tilde{n}}\sum_{t=n_0+1}^{\tilde{n}}|x_t-x|\\
&\leq\frac{n_0}{\tilde{n}}M+\frac{\tilde{n}-n_0}{\tilde{n}}\frac{\varepsilon}{2}\;\leq\;\frac{\varepsilon}{2}+\frac{\varepsilon}{2}\;=\;\varepsilon,
\end{aligned}$$

which implies the result. □

Before we state the next result, recall that HSIC-based payoffs are based on the predictable estimates $\{\widehat{g}_i\}_{i\geq 1}$ of the oracle witness function $g^\star$ and have the following form:

$$\begin{aligned} f_i(Z_{2i-1}, Z_{2i}) &= \frac{1}{2}\left(\widehat{g}_i(Z_{2i-1}) + \widehat{g}_i(Z_{2i})\right) - \frac{1}{2}\left(\widehat{g}_i(\tilde{Z}_{2i-1}) + \widehat{g}_i(\tilde{Z}_{2i})\right), \quad i \geq 1. \\ f^\star(Z_{2i-1}, Z_{2i}) &= \frac{1}{2}\left(g^\star(Z_{2i-1}) + g^\star(Z_{2i})\right) - \frac{1}{2}\left(g^\star(\tilde{Z}_{2i-1}) + g^\star(\tilde{Z}_{2i})\right). \end{aligned} \tag{A.5}$$

Lemma A.2.4. *Suppose that H_1 in* (2.1b) *is true. Then it holds that:*

$$\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i}) \xrightarrow{\text{a.s.}} \mathbb{E}\left[f^\star(Z_1, Z_2)\right], \tag{A.6}$$

$$\frac{1}{t}\sum_{i=1}^{t} \left(f_i(Z_{2i-1}, Z_{2i})\right)^2 \xrightarrow{\text{a.s.}} \mathbb{E}\left[\left(f^\star(Z_1, Z_2)\right)^2\right]. \tag{A.7}$$

Proof. We start by proving (A.6). Note that:

$$\begin{aligned} f_i(Z_{2i-1}, Z_{2i}) &= \frac{1}{2}\left(\widehat{g}_i(Z_{2i-1}) + \widehat{g}_t(Z_{2i})\right) - \frac{1}{2}\left(\widehat{g}_i(\tilde{Z}_{2i-1}) + \widehat{g}_i(\tilde{Z}_{2i})\right) \\ &= \frac{1}{2}\left\langle \widehat{g}_i, (\varphi(X_{2i}) - \varphi(X_{2i-1})) \otimes (\psi(Y_{2i}) - \psi(Y_{2i-1}))\right\rangle_{\mathcal{G}\otimes\mathcal{H}}. \end{aligned}$$

Next, observe that:

$$\begin{aligned} \left|\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i}) - \mathbb{E}\left[f^\star(Z_1, Z_2)\right]\right| \leq{}& \left|\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i}) - \frac{1}{t}\sum_{i=1}^{t} f^\star(Z_{2i-1}, Z_{2i})\right| \\ &+ \underbrace{\left|\frac{1}{t}\sum_{i=1}^{t} f^\star(Z_{2i-1}, Z_{2i}) - \mathbb{E}\left[f^\star(Z_1, Z_2)\right]\right|}_{\xrightarrow{\text{a.s.}} 0}, \end{aligned}$$

where the second term converges almost surely to 0 by the SLLN. For the first term, we have that:

$$\left|\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i}) - \frac{1}{t}\sum_{i=1}^{t} f^\star(Z_{2i-1}, Z_{2i})\right| \leq \frac{1}{t}\sum_{i=1}^{t} \left|f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i})\right|.$$

Finally, note that:

$$\begin{aligned} \left|f_i(Z_{2i-1}, Z_{2i}) - f^\star(Z_{2i-1}, Z_{2i})\right| &= \frac{1}{2}\left|\left\langle \widehat{g}_i - g^\star, (\varphi(X_{2i}) - \varphi(X_{2i-1})) \otimes (\psi(Y_{2i}) - \psi(Y_{2i-1}))\right\rangle_{\mathcal{G}\otimes\mathcal{H}}\right| \\ &\leq \left\|\widehat{g}_i - g^\star\right\|_{\mathcal{G}\otimes\mathcal{H}} \xrightarrow{\text{a.s.}} 0, \end{aligned} \tag{A.8}$$

where the convergence result is due to Lemma A.2.2. The result (A.6) then follows after invoking Lemma A.2.3. Next, we prove (A.7). Note that:

$$\begin{aligned}\frac{1}{t}\sum_{i=1}^{t}(f_i(Z_{2i-1},Z_{2i}))^2 &= \frac{1}{t}\sum_{i=1}^{t}(f_i(Z_{2i-1},Z_{2i})-f^\star(Z_{2i-1},Z_{2i})+f^\star(Z_{2i-1},Z_{2i}))^2\\ &= \underbrace{\frac{1}{t}\sum_{i=1}^{t}(f_i(Z_{2i-1},Z_{2i})-f^\star(Z_{2i-1},Z_{2i}))^2}_{\xrightarrow{\text{a.s.}}0}\\ &+\frac{2}{t}\sum_{i=1}^{t}(f^\star(Z_{2i-1},Z_{2i}))(f_i(Z_{2i-1},Z_{2i})-f^\star(Z_{2i-1},Z_{2i}))\\ &+\underbrace{\frac{1}{t}\sum_{i=1}^{t}(f^\star(Z_{2i-1},Z_{2i}))^2}_{\xrightarrow{\text{a.s.}}\mathbb{E}[(f^\star(Z_1,Z_2))^2]},\end{aligned}$$

where the first convergence result is due to (A.8) and Lemma A.2.3 and the second convergence result is due to the SLLN. Using (A.8) and Lemma A.2.3, we deduce that:

$$\left|\frac{2}{t}\sum_{i=1}^{t}(f^\star(Z_{2i-1},Z_{2i}))(f_i(Z_{2i-1},Z_{2i})-f^\star(Z_{2i-1},Z_{2i}))\right| \le 2\cdot\frac{1}{t}\sum_{i=1}^{t}|f_i(Z_{2i-1},Z_{2i})-f^\star(Z_{2i-1},Z_{2i})| \xrightarrow{\text{a.s.}} 0,$$

and hence we conclude that the convergence (A.7) holds.

□

Main Results

Theorem 2.1. *Let $\mathcal{C}$ denote a class of functions $c : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ for measuring dependence as per* (2.5)*.*

1. *Under H_0 in* (2.1a) *and* (2.2a)*, any payoff f of the form* (2.7) *satisfies $\mathbb{E}_{H_0}[f(Z_1,Z_2)]=0$.*
2. *Suppose that $\mathcal{C}$ satisfies* (2.6)*. Under H_1 in* (2.1b)*, the oracle payoff f^* based on the witness function c^* satisfies $\mathbb{E}_{H_1}[f^\star(Z_1,Z_2)]>0$. Further, for $\lambda^\star$ defined in* (2.8)*, it holds that $\mathbb{E}_{H_1}[\log(1+\lambda^\star f^\star(Z_1,Z_2)]>0$. Hence, $\mathcal{K}_t^\star \xrightarrow{\text{a.s.}} +\infty$, which implies that the oracle test is consistent: $\mathbb{P}_{H_1}(\tau^\star<\infty)=1$, where $\tau^\star = \inf\{t\ge 1:\mathcal{K}_t^\star\ge 1/\alpha\}$.*

Proof. 1. Under H_0 in (2.1a), we have that:

$$(X_{2t-1},Y_{2t-1}) \overset{d}{=} (X_{2t},Y_{2t}) \overset{d}{=} (X_{2t-1},Y_{2t}) \overset{d}{=} (X_{2t},Y_{2t-1}),$$

and hence, the first part of the Proposition trivially follows from the linearity of expectation. Under distribution drift, we use that at least one of the marginal distributions does not change at each round. For example, suppose

that at round t, it holds that: $P_X^{2t-1} = P_X^{2t}$. For the stream of independent observations, we have: $X_{2t} \perp\!\!\!\perp Y_{2t-1}$ and $X_{2t-1} \perp\!\!\!\perp Y_{2t}$. Further, under the H_0 in (2.2a), it holds that: $X_{2t-1} \perp\!\!\!\perp Y_{2t-1}$ and $X_{2t} \perp\!\!\!\perp Y_{2t}$. Hence, we have:

$$(X_{2t-1}, Y_{2t-1}) \stackrel{d}{=} (X_{2t}, Y_{2t-1}) \quad \text{and} \quad (X_{2t-1}, Y_{2t}) \stackrel{d}{=} (X_{2t}, Y_{2t}),$$

and hence, we get the result using linearity of expectation.

2. Under the i.i.d. setting, we have

$$\mathbb{E}\left[f^\star(Z_{2t-1}, Z_{2t}) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}\left[f^\star(Z_1, Z_2)\right] = s \cdot m(P_{XY}; \mathcal{C}),$$

and hence the result follows from the fact that the functional class $\mathcal{C}$ satisfies the characteristic condition (2.6).

3. Let $W := f^\star(Z_1, Z_2)$, and consider $\mathbb{E}_{H_1}\left[\log(1 + \lambda W)\right]$. We know that $\mathbb{E}_{H_1}\left[W\right] > 0$. We use the following inequality (Fan et al., 2015, Equation (4.12)): for any $y \geq -1$ and $\lambda \in [0, 1)$, it holds:

$$\log(1 + \lambda y) \geq \lambda y + y^2 \left(\log(1 - \lambda) + \lambda\right)$$

Hence

$$\mathbb{E}\left[\log(1 + \lambda W)\right] \geq \lambda \mathbb{E}\left[W\right] + \mathbb{E}\left[W^2\right] \left(\log(1 - \lambda) + \lambda\right).$$

Finally, using that $\log(1 - x) + x \geq -x^2/(2(1 - x))$ for $x \in [0, 1)$, we get:

$$\mathbb{E}_{H_1}\left[\log(1 + \lambda^\star W)\right] \geq \frac{(\mathbb{E}_{H_1}\left[W\right])^2/2}{\mathbb{E}_{H_1}\left[W\right] + \mathbb{E}_{H_1}\left[W^2\right]} > 0,$$

where recall that:

$$\lambda^\star = \frac{\mathbb{E}\left[W\right]}{\mathbb{E}\left[W\right] + \mathbb{E}\left[W^2\right]} \in (0, 1).$$

The wealth process corresponding to the oracle test satisfies:

$$\mathcal{K}_t = \prod_{i=1}^{t} (1 + \lambda^\star f^\star(Z_{2i-1}, Z_{2i})) = \exp\left(t \cdot \frac{1}{t} \sum_{i=1}^{t} \log(1 + \lambda^\star f^\star(Z_{2i-1}, Z_{2i}))\right).$$

By the Strong Law of Large Numbers (SLLN), we have:

$$\frac{1}{t} \sum_{i=1}^{t} \log(1 + \lambda^\star f^\star(Z_{2i-1}, Z_{2i})) \stackrel{\text{a.s.}}{\longrightarrow} \mathbb{E}\left[\log(1 + \lambda^\star W)\right] > 0.$$

Hence, we get that $\mathcal{K}_t \stackrel{\text{a.s.}}{\longrightarrow} +\infty$, and hence, the oracle test is consistent.

□

Theorem 2.2. *Suppose that Assumption 1 is satisfied. The following claims hold for HSIC-based SKIT (Algorithm 2):*

1. *Suppose that H_0 in* (2.1a) *or* (2.2a) *is true. Then SKIT ever stops with probability at most α: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*
2. *Suppose that H_1 in* (2.1b) *is true. Then it holds that $\mathcal{K}_t \overset{\text{a.s.}}{\longrightarrow} +\infty$ and thus SKIT is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$. Further, the wealth grows exponentially, and the corresponding growth rate satisfies*

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} \overset{\text{a.s.}}{\geq} \frac{\mathbb{E}[f^\star(Z_1,Z_2)]}{4} \cdot \left(\frac{\mathbb{E}[f^\star(Z_1,Z_2)]}{\mathbb{E}[(f^\star(Z_1,Z_2))^2]} \wedge 1 \right), \tag{2.15}$$

where $f^\star$ is the oracle payoff defined in (2.12)*.*

Remark 10. *While it will be clear from the proof that the i.i.d. assumption is sufficient but not necessary for asymptotic power one, the more relaxed sufficient conditions are slightly technical to state and thus omitted.*

Proof. 1. First, let us show that the predictable estimates of the oracle payoff function are bounded when the scaling factor $s = 1/2$ is used. Recall that:

$$\begin{aligned} f_t((x',y'),(x,y)) &= \frac{1}{2}\left(\widehat{g}_t(x',y') - \widehat{g}_t(x',y) + \widehat{g}_t(x,y) - \widehat{g}_t(x,y')\right) \\ &= \frac{1}{2}\left\langle \widehat{g}_t, \varphi(x')\otimes\psi(y') - \varphi(x')\otimes\psi(y) + \varphi(x)\otimes\psi(y) - \varphi(x)\otimes\psi(y')\right\rangle_{\mathcal{G}\otimes\mathcal{H}} \\ &= \frac{1}{2}\left\langle \widehat{g}_t, (\varphi(x') - \varphi(x))\otimes(\psi(y') - \psi(y))\right\rangle_{\mathcal{G}\otimes\mathcal{H}}. \end{aligned} \tag{A.9}$$

Note that:

$$\begin{aligned} |f_t((x',y'),(x,y))| &\leq \frac{1}{2}\left\|\widehat{g}_t\right\|_{\mathcal{G}\otimes\mathcal{H}} \left\|(\varphi(x') - \varphi(x))\otimes(\psi(y') - \psi(y))\right\|_{\mathcal{G}\otimes\mathcal{H}} \\ &\leq \frac{1}{2}\left\|(\varphi(x') - \varphi(x))\otimes(\psi(y') - \psi(y))\right\|_{\mathcal{G}\otimes\mathcal{H}} \\ &= \frac{1}{2}\left\|\varphi(x') - \varphi(x)\right\|_{\mathcal{G}} \cdot \left\|\psi(y') - \psi(y)\right\|_{\mathcal{H}} \\ &= \frac{1}{2}\sqrt{2(1-k(x',x))}\cdot\sqrt{2(1-l(y',y))} \\ &= 1. \end{aligned}$$

and hence, $f_t((x',y'),(x,y)) \leq [-1,1]$. Next, we show that constructed payoff function yields a fair bet. Indeed, we have that:

$$\mathbb{E}\left[f_t(Z_{2t-1},Z_{2t}) \mid \mathcal{F}_{t-1}\right] = \left\langle \widehat{g}_t, \mu_{XY} - \mu_X\otimes\mu_Y\right\rangle_{\mathcal{G}\otimes\mathcal{H}},$$

and in particular, the above implies that $\mathbb{E}_{H_0}\left[f_t(Z_{2t-1},Z_{2t}) \mid \mathcal{F}_{t-1}\right] = 0$ for H_0 in (2.1a). For H_0 in (2.2a), it is easy to see that the result holds using the form (A.9). We use that $X_{2t-1} \perp\!\!\!\perp Y_{2t-1}$, $X_{2t} \perp\!\!\!\perp Y_{2t}$, $X_{2t} \perp\!\!\!\perp Y_{2t-1}$, $X_{2t-1} \perp\!\!\!\perp Y_{2t}$, and the fact that at least one of the marginal distributions does not change.

Next, we show that for all strategies for selecting betting fractions that are considered in this work, the resulting wealth process is a nonnegative martingale. In case aGRAPA/ONS strategies are used, the resulting wealth

process is clearly a nonnegative martingale since betting fractions are predictable. The mixed wealth process $\left(\mathcal{K}_t^{\text{mixed}}\right)_{t\geq 1}$ is a nonnegative martingale under the null H_0, and hence

$$\begin{aligned}
\mathbb{E}_{H_0}\left[\mathcal{K}_t^{\text{mixed}} \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}\left[\int_0^1 \mathcal{K}_{t-1}^{\lambda}(1+\lambda f_t(Z_{2t-1}, Z_{2t}))\nu(\lambda)d\lambda \mid \mathcal{F}_{t-1}\right] \\
&= \int_0^1 \mathcal{K}_{t-1}^{\lambda}\mathbb{E}_{H_0}\left[1+\lambda f_t(Z_{2t-1}, Z_{2t}) \mid \mathcal{F}_{t-1}\right]\nu(\lambda)d\lambda \\
&= \int_0^1 \mathcal{K}_{t-1}^{\lambda}\nu(\lambda)d\lambda \\
&= \mathcal{K}_{t-1}^{\text{mixed}},
\end{aligned}$$

where the interchange of conditional expectation and integration is justified by the conditional monotone convergence theorem. The assertion of the Theorem then follows directly from Ville's inequality (Proposition A.1) when $a = 1/\alpha$.

2. Next, we establish the consistency of HSIC-based SKIT with ONS betting strategy. Under the ONS betting strategy, for any sequence of outcomes $(f_i)_{i\geq 1}$, $f_i \in [-1, 1]$, $i \geq 1$, it holds that (see the proof of Theorem 1 in (Cutkosky and Orabona, 2018)):

$$\log \mathcal{K}_t(\lambda_0) - \log \mathcal{K}_t = O\left(\log\left(\sum_{i=1}^t f_i^2\right)\right), \tag{A.10}$$

where $\mathcal{K}_t(\lambda_0)$ is the wealth of any constant betting strategy $\lambda_0 \in [-1/2, 1/2]$ and $\mathcal{K}_t$ is the wealth corresponding to the ONS betting strategy. It follows that the wealth process corresponding to the ONS betting strategy satisfies

$$\frac{\log \mathcal{K}_t}{t} \geq \frac{\log \mathcal{K}_t(\lambda_0)}{t} - C \cdot \frac{\log t}{t}, \tag{A.11}$$

for some absolute constant $C > 0$. Next, let us consider:

$$\lambda_0 = \frac{1}{2}\left(\left(\frac{\sum_{i=1}^t f_i}{\sum_{i=1}^t f_i^2} \wedge 1\right) \vee 0\right).$$

We obtain:

$$\begin{aligned}
\frac{\log \mathcal{K}_t(\lambda_0)}{t} &= \frac{1}{t}\sum_{i=1}^t \log(1+\lambda_0 f_i) \\
&\overset{\text{(a)}}{\geq} \frac{1}{t}\sum_{i=1}^t (\lambda_0 f_i - \lambda_0^2 f_i^2) \\
&= \left(\frac{\frac{1}{t}\sum_{i=1}^t f_i}{4} \vee 0\right) \cdot \left(\frac{\frac{1}{t}\sum_{i=1}^t f_i}{\frac{1}{t}\sum_{i=1}^t f_i^2} \wedge 1\right),
\end{aligned} \tag{A.12}$$

where in (a) we used* that $\log(1+x) \geq x - x^2$ for $x \in [-1/2, 1/2]$. From Lemma A.2.4, it follows for $f_i = f_i(Z_{2i-1}, Z_{2i})$ that:

$$\frac{\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i})}{4} \cdot \left(\frac{\frac{1}{t}\sum_{i=1}^{t} f_i(Z_{2i-1}, Z_{2i})}{\frac{1}{t}\sum_{i=1}^{t} (f_i(Z_{2i-1}, Z_{2i}))^2} \wedge 1 \right) \xrightarrow{\text{a.s.}} \frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{4} \cdot \left(\frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]} \wedge 1 \right). \tag{A.13}$$

Further, note that:

$$\mathbb{E}\left[f^\star(Z_1, Z_2)\right] = \|\mu_{XY} - \mu_X \otimes \mu_Y\|_{\mathcal{G}\otimes\mathcal{H}} = \sqrt{\mathrm{HSIC}(P_{XY}; \mathcal{G}, \mathcal{H})},$$

which is positive if the H_1 is true. Hence, using (A.11), we deduce that the growth rate of the ONS wealth process satisfies

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} \geq \frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{4} \cdot \left(\frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]} \wedge 1 \right). \tag{A.14}$$

We conclude that the test is consistent, that is, if H_1 is true, then $\mathbb{P}(\tau < \infty) = 1$.

□

Proposition 1. *The optimal log-wealth $S^\star := \mathbb{E}\left[\log(1 + \lambda^\star f^\star(Z_1, Z_2))\right]$ — that can be achieved by an oracle betting scheme* (2.16) *which knows $f^\star$ from* (2.12) *and the underlying distribution — satisfies:*

$$S^\star \leq \frac{\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{2} \left(\frac{8\mathbb{E}\left[f^\star(Z_1, Z_2)\right]}{3\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]} \wedge 1 \right). \tag{2.17}$$

Proof. We start by establishing the upper bound in (2.17). The fact that $S^\star \leq \mathbb{E}\left[f^\star(Z_1, Z_2)\right]/2$ trivially follows from $\mathbb{E}\left[\log(1 + \lambda f^\star(Z_1, Z_2))\right] \leq \lambda \mathbb{E}\left[f^\star(Z_1, Z_2)\right] \leq \mathbb{E}\left[f^\star(Z_1, Z_2)\right]/2$. Since for any $x \in [-0.5, 0.5]$, it holds that: $\log(1+x) \leq x - 3x^2/8$, we know that:

$$S^\star \leq \max_{\lambda \in [-0.5, 0.5]} \left(\lambda \mathbb{E}\left[f^\star(Z_1, Z_2)\right] - \frac{3}{8}\lambda^2 \mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right] \right), \tag{A.15}$$

and by solving the maximization problem, we get the upper bound:

$$S^\star \leq \frac{2}{3} \frac{\left(\mathbb{E}\left[f^\star(Z_1, Z_2)\right]\right)^2}{\mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right]}, \tag{A.16}$$

assuming $\left(\mathbb{E}\left[f^\star(Z_1, Z_2)\right]\right)^2 / \mathbb{E}\left[(f^\star(Z_1, Z_2))^2\right] \leq 3/8$. On the other hand, it always holds that: $S^\star \leq \mathbb{E}\left[f^\star(Z_1, Z_2)\right]/2$. To obtain the claimed bound, we multiply the RHS of (A.16) by two, which completes the proof of (2.17). □

*A slightly better constant for the growth rate (0.3 in place of 1/4) can be obtained by using the inequality: $\log(1+x) \geq x - \frac{5}{6}x^2$, that holds $\forall x \in [-0.5, 0.5]$.

Theorem 2.3. *Suppose that H_0 in* (2.18a) *is true. Further, assume that Assumption 2 holds. Then HSIC-based SKIT (Algorithm 2) satisfies:* $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.

Proof. Recall that at round t, the payoff function has form:

$$f_t((X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})) = \frac{1}{2}\langle \widehat{g}_t, (\varphi(X_{2t}) - \varphi(X_{2t-1})) \otimes (\psi(Y_{2t}) - \psi(Y_{2t-1}))\rangle_{\mathcal{G}\otimes\mathcal{H}}.$$

Let $\mathcal{D}_t = \{(X_i, Y_i)\}_{i \leq 2(t-1)}$. To establish validity, we need to show that under H_0 in (2.18a),

$$\mathbb{E}\left[f_t((X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})) \mid \mathcal{D}_t\right] = 0, \tag{A.17}$$

and hence it suffices to show that:

$$\mathbb{E}\left[(\varphi(X_{2t}) - \varphi(X_{2t-1})) \otimes (\psi(Y_{2t}) - \psi(Y_{2t-1})) \mid \mathcal{D}_t\right] = 0.$$

Due to independence under the null H_0, we have:

$$\begin{aligned}
\mathbb{E}\left[\varphi(X_{2t-1}) \otimes \psi(Y_{2t-1}) \mid \mathcal{D}_t\right] &= \mathbb{E}\left[\varphi(X_{2t-1}) \mid \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right] =: \mu_X^{2t-1} \otimes \mu_Y^{2t-1}, \\
\mathbb{E}\left[\varphi(X_{2t}) \otimes \psi(Y_{2t}) \mid \mathcal{D}_t\right] &= \mathbb{E}\left[\varphi(X_{2t}) \mid \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t}) \mid \mathcal{D}_t\right] =: \mu_X^{2t} \otimes \mu_Y^{2t},
\end{aligned}$$

Consider one of the cross-terms $\varphi(X_{2t}) \otimes \psi(Y_{2t-1})$. We have the following:

$$\begin{aligned}
\mathbb{E}\left[\varphi(X_{2t}) \otimes \psi(Y_{2t-1}) \mid \mathcal{D}_t\right] &\overset{a}{=} \mathbb{E}\left[\mathbb{E}\left[\varphi(X_{2t}) \otimes \psi(Y_{2t-1}) \mid X_{2t-1}, \mathcal{D}_t\right] \mid \mathcal{D}_t\right] \\
&\overset{b}{=} \mathbb{E}\left[\mathbb{E}\left[\varphi(X_{2t}) \mid X_{2t-1}, \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid X_{2t-1}, \mathcal{D}_t\right] \mid \mathcal{D}_t\right] \\
&\overset{c}{=} \mathbb{E}\left[\mathbb{E}\left[\varphi(X_{2t}) \mid X_{2t-1}, \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right] \mid \mathcal{D}_t\right] \\
&\overset{d}{=} \mathbb{E}\left[\mathbb{E}\left[\varphi(X_{2t}) \mid X_{2t-1}, \mathcal{D}_t\right] \mid \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right] \\
&\overset{e}{=} \mathbb{E}\left[\varphi(X_{2t}) \mid \mathcal{D}_t\right] \otimes \mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right] \\
&\overset{f}{=} \mu_X^{2t} \otimes \mu_Y^{2t-1}.
\end{aligned}$$

In the above, (a) uses the law of iterated expectations and conditioning on X_{2t-1}, (b) uses the assumption (2.19) about conditional independence, (c) uses the independence null assumption (2.1a), (d) uses that $\mathbb{E}\left[\psi(Y_{2t-1}) \mid \mathcal{D}_t\right]$ is $\sigma(\mathcal{D}_t)$-measurable, (e) uses the law of iterated expectations, and (f) uses the definitions of the mean embeddings of conditional distributions. An analogous argument can be used to deduce:

$$\mathbb{E}\left[\varphi(X_{2t-1}) \otimes \psi(Y_{2t}) \mid \mathcal{D}_t\right] = \mu_X^{2t-1} \otimes \mu_Y^{2t}.$$

We get that:

$$\begin{aligned}\mathbb{E}\left[(\varphi(X_{2t}) - \varphi(X_{2t-1})) \otimes (\psi(Y_{2t}) - \psi(Y_{2t-1})) \mid \mathcal{D}_t\right] &= \mu_X^{2t-1} \otimes \mu_Y^{2t-1} + \mu_X^{2t} \otimes \mu_Y^{2t} - \mu_X^{2t-1} \otimes \mu_Y^{2t} - \mu_X^{2t} \otimes \mu_Y^{2t-1} \\ &= (\mu_X^{2t} - \mu_X^{2t-1}) \otimes (\mu_Y^{2t} - \mu_Y^{2t-1}),\end{aligned}$$

and hence, if either (X_{2t-1}, X_{2t}) or (Y_{2t-1}, X_{2t}) are exchangeable conditional on $\mathcal{D}_t$, it follows that either $\mu_X^{2t} = \mu_X^{2t-1}$ or $\mu_Y^{2t} = \mu_Y^{2t-1}$ respectively. This, in turn, implies that (A.17) holds, and hence, the result follows.

□

A.2.3 Proofs for Section 2.3

Theorem 2.4. *Suppose that (A1) in Assumption 1 is satisfied. Then, under H_0 in* (2.1a) *and* (2.18a)*, COCO/KCC-based SKIT (Algorithm 3) satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

Proof. It suffices to show that the proposed payoff functions are bounded. The rest of the proof follows will follow the same steps as the proof of Theorem 2.2 (for a stream of independent observations) or Theorem 2.3 (for time-varying independence null), and we omit the details. Note that:

$$\begin{aligned}\left|\widehat{h}_t(y') - \widehat{h}_t(y)\right| &= \left|\langle \widehat{h}_t, \psi(y')\rangle_{\mathcal{H}} - \langle \widehat{h}_t, \psi(y)\rangle_{\mathcal{H}}\right| \\ &= \left|\langle \widehat{h}_t, \psi(y') - \psi(y)\rangle_{\mathcal{H}}\right| \\ &\leq \left\|\widehat{h}_t\right\|_{\mathcal{H}} \|\psi(y') - \psi(y)\|_{\mathcal{H}} \\ &\leq \|\psi(y') - \psi(y)\|_{\mathcal{H}} \\ &= \sqrt{2(1 - l(y, y'))} \\ &\leq \sqrt{2},\end{aligned}$$

where we used that $\left\|\widehat{h}_t\right\|_{\mathcal{H}} \leq 1$ due to normalization. Analogous bound holds for $|\widehat{g}_t(x') - \widehat{g}_t(x)|$. We conclude that any predictable estimate of the oracle payoff function for COCO (or KCC) satisfies

$$|f_t((x', y'), (x, y))| \leq 1,$$

as proposed. The fact that the payoff function is fair trivially follows from the definition. Regarding the existence of the oracle payoff, whose mean is positive under H_1 in (2.1b), note that if k and l are characteristic kernels, then COCO and KCC satisfy the characteristic condition (2.6); see Jordan and Bach (2001); Gretton et al. (2005c,b). Hence, the result follows from Theorem 2.1. This completes the proof.

□

A.2.4 Proofs for Section 2.4

Theorem 2.5. *Under H_0 in* (2.1a) *and* (2.18a)*, the symmetry-based SKIT (Algorithm 4) satisfies:* $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.

Proof. For any $t \geq 1$, we have that the payoffs defined in (2.26), (2.27), and (2.28) are bounded: $f_t(w) \in [-1, 1]$, $\forall w \in \mathbb{R}$. Due to Proposition 2, we know that, under the null, W_t is a random variable that is symmetric around zero (conditional on $\mathcal{F}_{t-1}$). Hence, for the composition approach, it trivially follows that $\mathbb{E}_{H_0}\left[f_t^{\text{odd}}(W_t) \mid \mathcal{F}_{t-1}\right] = 0$ since a composition with an odd function is used. For the rank and predictive approaches, we use the fact that, under the null, $\text{sign}(W_t) \perp\!\!\!\perp |W_t| \mid \mathcal{F}_{t-1}$. Since, $\mathbb{E}_{H_0}\left[\text{sign}(W_t) \mid \mathcal{F}_{t-1}\right] = 0$, it then follows that $\mathbb{E}_{H_0}\left[f_t^{\text{rank}}(W_t) \mid \mathcal{F}_{t-1}\right] = 0$. Using that $\text{sign}(W_t) \perp\!\!\!\perp |W_t| \mid \mathcal{F}_{t-1}$ and by conditioning on the sign of W_t, we get:

$$\mathbb{E}_{H_0}\left[\ell_t(W_t) \mid \mathcal{F}_{t-1}\right] = \frac{1}{2}\mathbb{P}_{H_0}\left(p_t(|W_t|) \geq 1/2\right) + \frac{1}{2}\mathbb{P}_{H_0}\left(p_t(|W_t|) < 1/2\right) = \frac{1}{2}.$$

Hence $\mathbb{E}_{H_0}\left[1 - 2\ell_t(W_t) \mid \mathcal{F}_{t-1}\right] = 0$. The rest of the proof regarding the validity of the symmetry-based SKITs follows the same steps as the proof of Theorem 2.2, and we omit the details.

□

A.3 Selecting Betting Fractions

As alluded to in Remark 1, sticking to a single fixed betting fraction, $\lambda_t = \lambda \in [0, 1]$, $t \geq 1$, may result in a wealth process that either has a sub-optimal growth rate under the alternative or tends to zero almost surely (see Figure A.4). *Mixing* over different betting fractions is a simple approach that often works well in practice. Given a fine grid of values: $\Lambda = \left\{\lambda^{(1)}, \ldots, \lambda^{(J)}\right\}$, e.g., uniformly spaced values on the unit interval, consider

$$\mathcal{K}_t^{\text{mixed}} = \frac{1}{|\Lambda|} \sum_{\lambda^{(j)} \in \Lambda} \mathcal{K}_t(\lambda^{(j)}), \tag{A.18}$$

where $\left(\mathcal{K}_t(\lambda^{(j)})\right)_{t \geq 0}$ is a wealth process corresponding to a constant-betting strategy with betting fraction $\lambda^{(j)}$ [†].

While mixing often works well in practice, it introduces additional tuning hyperparameters, e.g., grid size. We consider two compelling approaches for the selection of betting fractions in a predictable way, meaning that λ_t depends only on $\{(X_i, Y_i)\}_{i \leq 2(t-1)}$. In addition to the ONS strategy (Algorithm 1), we also consider *aGRAPA* strategy (Algorithm 8). The idea that effective betting strategies are ones that maximize a gambler's expected log capital dates back to early works of Kelly (1956) and Breiman (1962). Assuming that the same betting fraction is used, the log

[†]Practically, it is advisable to start with a coarse grid (small J) at small t and occasionally add another grid point, so that the grid becomes finer over time. Whenever a grid point is added, it is like adding another stock to a portfolio, and the wealth must be appropriately redistributed; we omit the details for brevity.

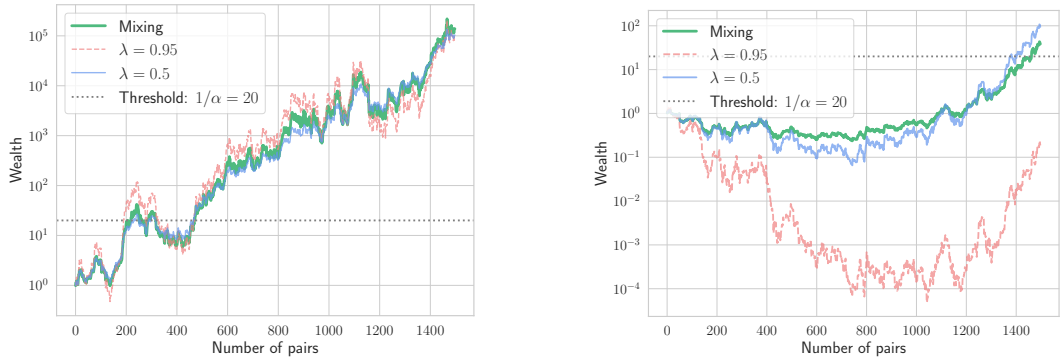


Figure A.4: SKIT with HSIC payoff function on two particular realizations of streams of dependent data: $Y_t = 0.1 \cdot X_t + \varepsilon_t$, $X_t, \varepsilon_t \sim \mathcal{N}(0,1)$. For both cases, we consider a mixed wealth process for $\Lambda = \{0.05, 0.1, \ldots, 0.95\}$. We observe that the mixed wealth process follows closely the best of constant-betting strategies with $\lambda \in \{0.5, 0.95\}$.

capital after round $(t-1)$ is

$$\log \mathcal{K}_{t-1}(\lambda) = \sum_{i=1}^{t-1} \log\left(1 + \lambda f_i(Z_{2i-1}, Z_{2i})\right).$$

Algorithm 8 aGRAPA strategy for selecting betting fractions

Input: sequence of payoffs $(f_t(Z_{2t-1}, Z_{2t}))_{t\geq 1}$, $\lambda_1^{\mathrm{aGRAPA}} = 0$, $\mu_0^{(1)} = 0$, $\mu_0^{(2)} = 1$, $c = 0.9$.
for $t = 1, 2, \ldots$ **do**
  Set $\mu_t^{(1)} = \mu_{t-1}^{(1)} + f_t(Z_{2t-1}, Z_{2t})$;
  Set $\mu_t^{(2)} = \mu_{t-1}^{(2)} + (f_t(Z_{2t-1}, Z_{2t}))^2$;
  Set $\lambda_{t+1}^{\mathrm{aGRAPA}} = c \wedge \left(0 \vee \left(\mu_t^{(1)}/\mu_t^{(2)}\right)\right)$;

Following Waudby-Smith and Ramdas (2023), we set the derivative to zero and use Taylor's expansion to get

$$\lambda_t^{\mathrm{aGRAPA}} = \left(\left(\frac{\sum_{i=1}^{t-1} f_i(Z_{2i-1}, Z_{2i})}{\sum_{i=1}^{t-1} \left(f_i(Z_{2i-1}, Z_{2i})\right)^2}\right) \vee 0\right) \wedge c.$$

Truncation at zero is inspired by the fact that $\mathbb{E}_{H_1}\left[f^\star(Z_{2t-1}, Z_{2t}) \mid \mathcal{F}_{t-1}\right] > 0$, whereas truncation at $c \in (0,1]$ (e.g., $c = 0.9$) is necessary to guarantee that the wealth process is indeed nonnegative.

A.4 Omitted Details for Sections 2.2 and 2.3

In this section, we complement the material presented in the main paper by deriving the forms of the witness functions for the dependence criteria considered in this work.

Oracle Witness Function for HSIC. Let us derive the form of the oracle witness function for HSIC. Note that:

$$\begin{aligned}
&\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \left[\mathbb{E}_{P_{XY}}\left[g(X,Y)\right] - \mathbb{E}_{P_X\times P_Y}\left[g(X',Y')\right]\right] \\
= &\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \left[\mathbb{E}_{P_{XY}}\left[\langle g, \varphi(X)\otimes\psi(Y)\rangle_{\mathcal{G}\otimes\mathcal{H}}\right] - \mathbb{E}_{P_X\times P_Y}\left[\langle g, \varphi(X')\otimes\psi(Y')\rangle_{\mathcal{G}\otimes\mathcal{H}}\right]\right] \\
= &\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \left[\langle g, \mathbb{E}_{P_{XY}}\left[\varphi(X)\otimes\psi(Y)\right]\rangle_{\mathcal{G}\otimes\mathcal{H}} - \langle g, \mathbb{E}_{P_X\times P_Y}\left[\varphi(X')\otimes\psi(Y')\right]\rangle_{\mathcal{G}\otimes\mathcal{H}}\right] \\
= &\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \left[\langle g, \mu_{XY}\rangle_{\mathcal{G}} - \langle g, \mu_X\otimes\mu_Y\rangle_{\mathcal{G}\otimes\mathcal{H}}\right] \\
= &\sup_{g:\|g\|_{\mathcal{G}\otimes\mathcal{H}}\leq 1} \langle g, \mu_{XY} - \mu_X\otimes\mu_Y\rangle_{\mathcal{G}\otimes\mathcal{H}},
\end{aligned}$$

from which it is easy to derive the oracle witness function for HSIC.

Remark 11. *Note that in* (2.13) *the witness function is defined as an operator:* $\widehat{g}_t : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$. *To clarify, for any* $z=(x,y)\in\mathcal{X}\times\mathcal{Y}$, *we have*

$$(\widehat{\mu}_{XY} - \widehat{\mu}_X\otimes\widehat{\mu}_Y)(z) = \frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} k(X_i,x)l(Y_i,y) - \left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} k(X_i,x)\right)\cdot\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)} l(Y_i,y)\right),$$

and the denominator in (2.13) *can be expressed in terms of kernel matrices* $K,L\in\mathbb{R}^{2(t-1)\times 2(t-1)}$ *with entries* $K_{ij}=k(X_i,X_j)$, $L_{ij}=l(Y_i,Y_j)$, $i,j\in\{1,\ldots,2(t-1)\}$, *as:*

$$\|\widehat{\mu}_{XY} - \widehat{\mu}_X\otimes\widehat{\mu}_Y\|_{\mathcal{G}\otimes\mathcal{H}} = \frac{1}{2(t-1)}\sqrt{\mathrm{tr}(KHLH)},$$

where $H = \boldsymbol{I}_{2(t-1)} - (1/(2(t-1)))\boldsymbol{1}\boldsymbol{1}^\top$ *is the centering projection matrix.*

Remark 12. *While the empirical witness functions for COCO/KCC* (2.21) *are defined as operators, we use those as functions in the definition of the corresponding payoff function. To clarify, for any* $x\in\mathcal{X}$ *and* $y\in\mathcal{Y}$, *we have*

$$\begin{aligned}
\widehat{g}_t(x) &= \sum_{i=1}^{2(t-1)}\alpha_i\left(k(X_i,x) - \frac{1}{2(t-1)}\sum_{j=1}^{2(t-1)} k(X_j,x)\right), \\
\widehat{h}_t(y) &= \sum_{i=1}^{2(t-1)}\beta_i\left(l(Y_i,y) - \frac{1}{2(t-1)}\sum_{j=1}^{2(t-1)} l(Y_j,y)\right).
\end{aligned}$$

Minibatched Payoff Function for HSIC. The minibatched payoff function at round t has the following form:

$$f_t(Z_{b(t-1)+1},\ldots,Z_{bt}) = \frac{1}{b}\sum_{i=1}^{b}\widehat{g}_t(X_{b(t-1)+i},Y_{b(t-1)+i}) - \frac{1}{b(b-1)}\sum_{\substack{i,j=1\\ i\neq j}}^{b}\widehat{g}_t(X_{b(t-1)+i},Y_{b(t-1)+j}).$$

Note that:

$$\begin{aligned} f_t(Z_{b(t-1)+1}, \ldots, Z_{bt}) &= \frac{1}{b}\sum_{i=1}^{b} \langle \widehat{g}_t, \varphi(X_{b(t-1)+i}) \otimes \psi(Y_{b(t-1)+i}) \rangle_{\mathcal{G}\otimes\mathcal{H}} \\ &\quad - \frac{1}{b(b-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{b} \langle \widehat{g}_t, \varphi(X_{b(t-1)+i}) \otimes \psi(Y_{b(t-1)+j}) \rangle_{\mathcal{G}\otimes\mathcal{H}} \\ &= \left\langle \widehat{g}_t, \frac{1}{2b(b-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{b} \big(\varphi(X_{b(t-1)+i}) - \varphi(X_{b(t-1)+j})\big) \otimes \big(\psi(Y_{b(t-1)+i}) - \psi(Y_{b(t-1)+j})\big) \right\rangle_{\mathcal{G}\otimes\mathcal{H}}. \end{aligned}$$

Let $\mathcal{F}'_{t-1} = \sigma(\{(X_i, Y_i)\}_{i \leq b(t-1)})$. We have that:

$$\mathbb{E}\left[f_t(Z_{b(t-1)+1}, \ldots, Z_{bt}) \mid \mathcal{F}'_{t-1}\right] = \langle \widehat{g}_t, \mu_{XY} - \mu_X \otimes \mu_Y \rangle_{\mathcal{G}\otimes\mathcal{H}},$$

and in particular, $\mathbb{E}_{H_0}\left[f_t(Z_{b(t-1)+1}, \ldots, Z_{bt}) \mid \mathcal{F}'_{t-1}\right] = 0$ if the null H_0 in (2.1a) is true. It suffices to show that the payoff is bounded. Since $\|\widehat{g}_t\|_{\mathcal{G}\otimes\mathcal{H}} = 1$, we can easily deduce that:

$$\begin{aligned} \left|f_t(Z_{b(t-1)+1}, \ldots, Z_{bt})\right| &\leq \frac{1}{2b(b-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{b} \left\| \big(\varphi(X_{b(t-1)+i}) - \varphi(X_{b(t-1)+j})\big) \otimes \big(\psi(Y_{b(t-1)+i}) - \psi(Y_{b(t-1)+j})\big) \right\|_{\mathcal{G}\otimes\mathcal{H}} \\ &= \frac{1}{2b(b-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{b} \left\|\varphi(X_{b(t-1)+i}) - \varphi(X_{b(t-1)+j})\right\|_{\mathcal{G}} \left\|\psi(Y_{b(t-1)+i}) - \psi(Y_{b(t-1)+j})\right\|_{\mathcal{H}} \\ &= \frac{1}{2b(b-1)} \sum_{\substack{i,j=1 \\ i\neq j}}^{b} \sqrt{2(1 - k(X_{b(t-1)+i}, X_{b(t-1)+j}))}\sqrt{2(1 - l(Y_{b(t-1)+i}, Y_{b(t-1)+j}))} \\ &\leq 1. \end{aligned}$$

Hence, we conclude that the wealth process constructed using a minibatched version of the payoff function is also a nonnegative martingale.

Example 4. *For $t \geq 1$, consider*

$$(X_t, Y_t) = \left(\frac{V_t + 1 - 1/t}{2}, \frac{V'_t + 1 - 1/t}{2}\right),$$

where $V_t, V'_t \overset{\text{iid}}{\sim} \text{Ber}(1/2)$. Note that $\mathcal{X} = \mathcal{Y} \subseteq [0, 1]$, which means that a pair of linear kernels, $k(x, x') = xx'$ and $l(y, y') = yy'$ are nonnegative and bounded by one on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Note that for a linear kernel,

$$\widehat{g}_t(x, y) = \widehat{g}_t \cdot x \cdot y.$$

Hence,

$$f_t((X_{2t-1},Y_{2t-1}),(X_{2t},Y_{2t})) = \frac{\widehat{g}_t}{2}\left(X_{2t}-X_{2t-1}\right)\left(Y_{2t}-Y_{2t-1}\right)$$
$$= \frac{\widehat{g}_t}{8}\left(V_{2t}-V_{2t-1}+\frac{1}{2t(2t-1)}\right)\left(V'_{2t}-V'_{2t-1}+\frac{1}{2t(2t-1)}\right).$$

In particular, $\mathbb{E}\left[f_t((X_{2t-1},Y_{2t-1}),(X_{2t},Y_{2t}))\mid\mathcal{F}_{t-1}\right]\neq 0$*, implying that the wealth process* $(\mathcal{K}_t)_{t\geq 0}$ *is no longer a nonnegative martingale.*

Witness Functions for COCO. Let Φ and Ψ be a pair of matrices whose columns represent embeddings of $X_1,\ldots,X_{2(t-1)}$ and $Y_1,\ldots,Y_{2(t-1)}$, that is, $\varphi(X_i)=k(X_i,\cdot)$ and $\psi(Y_i)=l(Y_i,\cdot)$ for $i=1,\ldots,2(t-1)$. Recall that

$$\widehat{g}=\sum_{i=1}^{2(t-1)}\alpha_i\left(\varphi(X_i)-\frac{1}{2(t-1)}\sum_{j=1}^{2(t-1)}\varphi(X_j)\right)=\Phi H\alpha,$$
$$\widehat{h}=\sum_{i=1}^{2(t-1)}\beta_i\left(\psi(Y_i)-\frac{1}{2(t-1)}\sum_{j=1}^{2(t-1)}\psi(Y_j)\right)=\Psi H\beta,$$

where $H=\mathbf{I}_{2(t-1)}-\frac{1}{2(t-1)}11^\top$ is the centering projection matrix. We have

$$\langle h,\widehat{C}_{XY}g\rangle_{\mathcal{H}}=\frac{1}{2(t-1)}(\alpha^\top H\Phi^\top)(\Phi H\Psi^\top)(\Psi H\beta)=\frac{1}{2(t-1)}\alpha^\top HKHLH\beta=\frac{1}{2(t-1)}\alpha^\top\tilde{K}\tilde{L}\beta,$$
$$\|g\|_{\mathcal{G}}^2=\alpha^\top\tilde{K}\alpha,$$
$$\|h\|_{\mathcal{H}}^2=\beta^\top\tilde{L}\beta,$$

where $\tilde{K}:=HKH$ and $\tilde{L}:=HLH$ are centered kernel matrices. Hence, the maximization problem in (2.20) can be expressed as:

$$\begin{aligned}\max_{\alpha,\beta}\quad & \frac{1}{2(t-1)}\alpha^\top\tilde{K}\tilde{L}\beta\\ \text{subject to}\quad & \alpha^\top\tilde{K}\alpha=1,\quad \beta^\top\tilde{L}\beta=1.\end{aligned}\tag{A.19}$$

After introducing Lagrange multipliers, it can then be shown that α and β, which solve (A.19), exactly correspond to the generalized eigenvalue problem (2.22).

Witness Functions for KCC. Introduce empirical covariance operators:

$$\widehat{C}_X=\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)}\varphi(X_i)\otimes\varphi(X_i)-\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)}\varphi(X_i)\right)\otimes\left(\frac{1}{2(t-1)}\sum_{i=1}^{n}\varphi(X_i)\right)=\frac{1}{2(t-1)}\Phi H\Phi^\top,$$
$$\widehat{C}_Y=\frac{1}{n}\sum_{i=1}^{2(t-1)}\psi(Y_i)\otimes\psi(Y_i)-\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)}\psi(Y_i)\right)\otimes\left(\frac{1}{2(t-1)}\sum_{i=1}^{2(t-1)}\psi(Y_i)\right)=\frac{1}{2(t-1)}\Psi H\Psi^\top.$$

Then the empirical variance terms can be expressed as:

$$\widehat{\mathbb{V}}\left[g(X)\right] = \langle g, \widehat{C}_X g\rangle_{\mathcal{G}} = \frac{1}{2(t-1)}(\alpha^\top H\Phi^\top)(\Phi H\Phi^\top)(\Phi H\alpha) = \frac{1}{2(t-1)}\alpha^\top \tilde{K}^2\alpha,$$
$$\widehat{\mathbb{V}}\left[h(Y)\right] = \langle h, \widehat{C}_Y h\rangle_{\mathcal{H}} = \frac{1}{2(t-1)}(\beta^\top H\Psi^\top)(\Psi H\Psi^\top)(\Psi H\beta) = \frac{1}{2(t-1)}\beta^\top \tilde{L}^2\beta.$$

Thus, an empirical estimator of the kernel canonical correlation (2.23) can be obtained by solving:

$$\begin{aligned} \max_{\alpha,\beta} \quad & \frac{1}{2(t-1)}\alpha^\top \tilde{K}\tilde{L}\beta \\ \text{subject to} \quad & \frac{1}{2(t-1)}\alpha^\top \tilde{K}^2\alpha + \kappa_1\alpha^\top \tilde{K}\alpha = 1, \\ & \frac{1}{2(t-1)}\beta^\top \tilde{L}^2\beta + \kappa_2\beta^\top \tilde{L}\beta = 1. \end{aligned}$$

After introducing Lagrange multipliers, it can then be shown that α and β, which solve (2.23), correspond to the generalized eigenvalue problem:

$$\begin{pmatrix} 0 & \frac{1}{2(t-1)}\tilde{K}\tilde{L} \\ \frac{1}{2(t-1)}\tilde{L}\tilde{K} & 0 \end{pmatrix}\begin{pmatrix}\alpha \\ \beta\end{pmatrix} = \gamma \begin{pmatrix} \kappa_1\tilde{K} + \frac{1}{2(t-1)}\tilde{K}^2 & 0 \\ 0 & \kappa_2\tilde{L} + \frac{1}{2(t-1)}\tilde{L}^2 \end{pmatrix}\begin{pmatrix}\alpha \\ \beta\end{pmatrix},$$

A.5 Additional Simulations

This section contains: (a) additional experiments on synthetic dataset and (b) data visualizations of the datasets used in this paper.

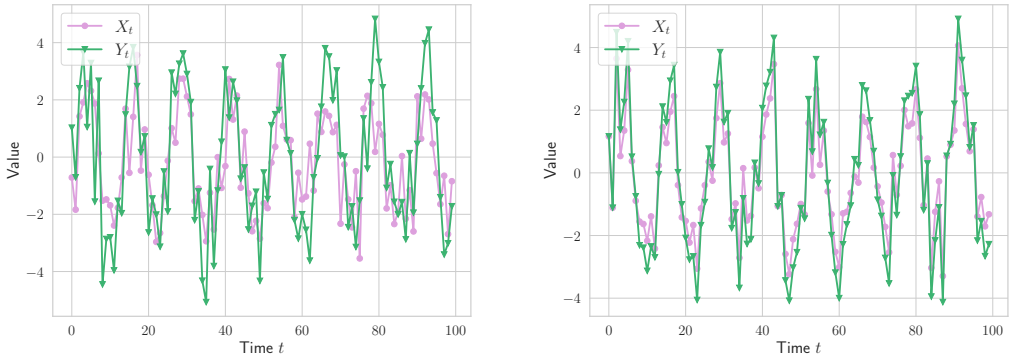
A.5.1 Test of Instantaneous Dependence

In Figure A.5, we demonstrate it is hard to visually tell the difference between independence and dependence under distribution drift setting (2.2). See Example 1 for details.

A.5.2 Distribution Drift

In this section, we consider the linear Gaussian model with an underlying distribution drift:

$$Y_t = X_t\beta_t + \varepsilon_t, \quad X_t, \varepsilon_t \sim \mathcal{N}(0,1), \quad t \geq 1,$$



(a) Independent noise.

(b) Dependent noise.

Figure A.5: Sample of independent (subplot (a)) and dependent ($\rho = 0.5$, subplot(b)) data according to (2.3). The purpose of visualizing raw data is to demonstrate that dependence is hard to detect visually, and dependence refers to more than temporal correlation which may be present due to cyclical trends.

that is, in contrast to the Gaussian linear model (Section 2.3), β_t changes over time. We gradually increase it from $\beta_t = 0$ to $\beta_t = 0.1$ in increments of 0.02, that is:

$$\underbrace{\beta_0, \ldots, \beta_{b-1}}_{=0}, \underbrace{\beta_b, \ldots, \beta_{2b-1}}_{=0.02}, \ldots, \underbrace{\beta_{5b-1}, \ldots}_{=0.1}$$

and, starting with β_{5b}, we keep it equal to 0.1. We consider $b \in \{100, 200, 400\}$ as possible block sizes. Note that there is a transition from independence (first b datapoints in a stream) to dependence. In Figure A.6, we show that our test performs well under the distribution drift setting and consistently detects dependence.

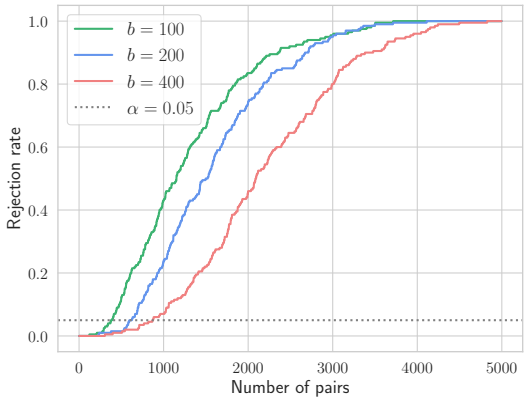
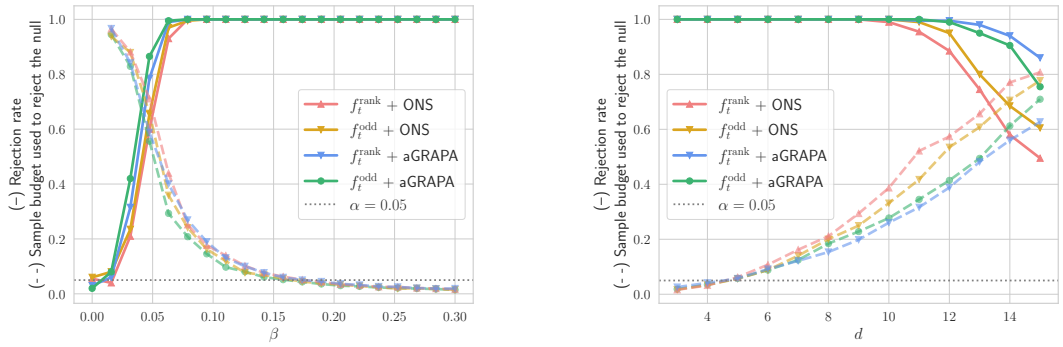


Figure A.6: Rejection rate of sequential independence test under distribution drift setting. Focusing on the non-i.i.d. time-varying setting, we confirm that our test has high power under the alternative.

A.5.3 Symmetry-based Payoff Functions

In this section, we complement the comparison presented in Section 2.4 between the rank- and composition-based betting strategies (since those require minimal tuning) used with ONS or aGRAPA criteria for selecting betting fractions. We also increase the monitoring horizon to 20000 datapoints. In Figure A.7a, we consider the Gaussian linear model, but in contrast to the setting considered in Section 2.4, we focus on harder testing settings by considering $\beta \in [0, 0.3]$. In Figure A.7b, we compare composition- and rank-based approaches when data are sampled from the spherical model. In both cases, composition and rank-based approaches are similar; none of the payoffs uniformly dominates the other. We also observe that selecting betting fractions via aGRAPA criterion tends to result in a bit more powerful testing procedure.



(a) (b)

Figure A.7: (a) Comparison of symmetry-based betting strategies under the Gaussian model. The betting strategy based on composition with an odd function performs only slightly better than the rank-based strategy. (b) SKIT with composition- and rank-based betting strategies under the spherical model. None of the betting strategies uniformly dominates the other. aGRAPA criterion for selecting betting fractions tends to result in a bit more powerful testing procedure.

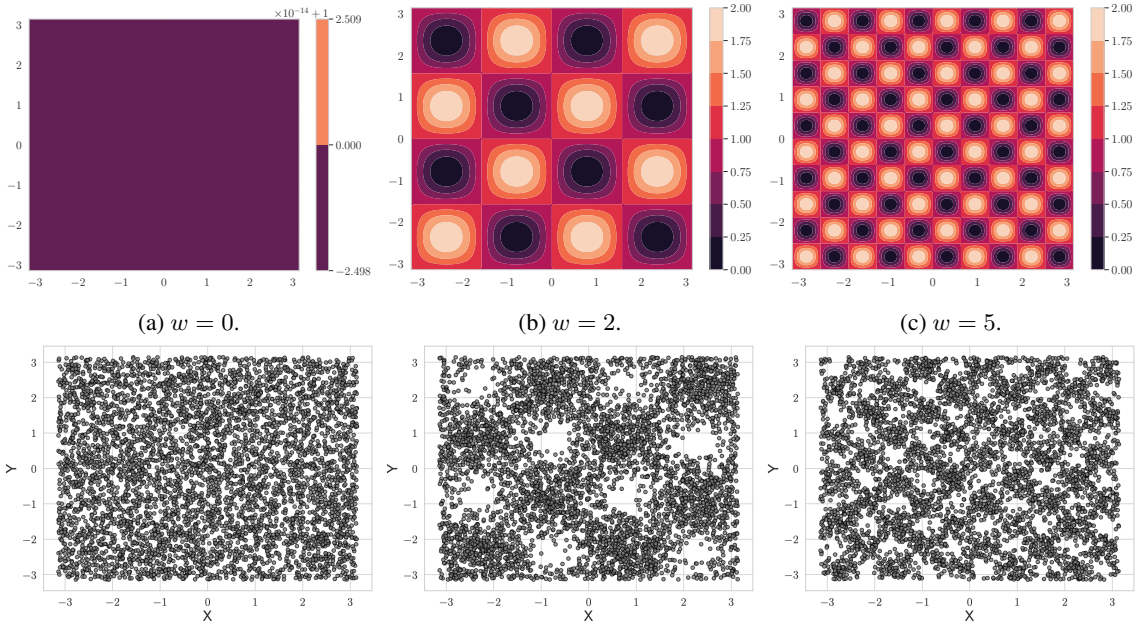
A.5.4 Hard-to-detect Dependence

Hard-to-detect dependence. Consider the joint density $p(x, y)$ of the form:

$$\frac{1}{4\pi^2}\left(1 + \sin(wx)\sin(wy)\right) \cdot \mathbb{1}\left\{(x, y) \in [-\pi, \pi]^2\right\}. \tag{A.20}$$

With the null case corresponding to $w = 0$, the testing problem becomes harder with growing w. In Figure A.8, we illustrate the densities and a data sample for the hard-to-detect setting (A.20).

We use $\lambda_X = \lambda_Y = 3/(4\pi^2)$ as RBF kernel hyperparameters. For visualization purposes, we stop monitoring after observing 20000 datapoints from P_{XY}, and if a SKIT does not reject H_0 by that time, we assume that the null is retained. The results are aggregated over 200 runs for each value of w. In Figure A.9, where the null case corresponds



(a) $w = 0$. (b) $w = 2$. (c) $w = 5$.

(d) $w = 0$. (e) $w = 2$. (f) $w = 5$.

Figure A.8: Visualization of the densities (top) and a dataset of size 5000 (bottom) sampled from the corresponding distribution.

to $w = 0$, we confirm that SKITs have time-uniform type I error control. The average rejection rate starts to drop for $w \geq 3$, meaning that observing 20000 points from P_{XY} does not suffice to detect dependence.

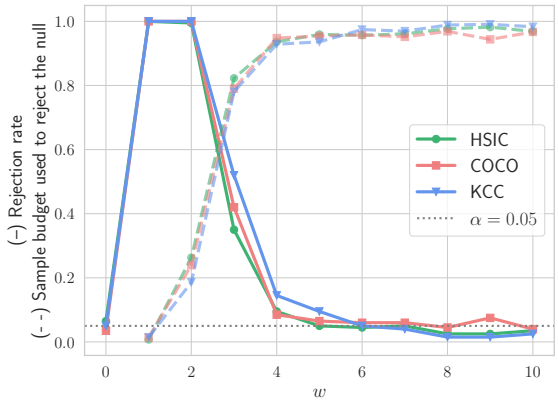


Figure A.9: Rejection rate (solid) and fraction of samples used before the null hypothesis was rejected (dashed) for hard-to-detect dependence model. By inspecting the rejection rate for $w = 0$ (independence holds), we confirm that the type I error is controlled. Further, SKIT is adaptive to the complexity of a problem (larger w corresponds to a harder setting).

A.5.5 Additional Results for Real Data

In Figure A.10, we illustrate that the average daily temperature in selected cities share similar seasonal patterns. We repeat the same experiment as in Section 2.4, but for four cities in South Africa: Cape Town (CT), Port Elizabeth (PE), Durban (DRN), and Bloemfontein (BFN). In Figures A.10d and A.10e, we illustrate the resulting wealth processes for each pair of cities and for each region. Finally, we illustrate the pairs of cities for which the null has been rejected in Figure A.10c.

A.5.6 Experiment with MNIST data

In this section, we analyze the performance of SKIT on high-dimensional real data. This experiment is based on MNIST dataset (LeCun et al., 1998) where pairs of digits are observed at each step; under the null one sees digits (a, b) where a and b are uniformly randomly chosen, but under the alternative one sees (a, a'), i.e., two different images of the same digit. To estimate kernel hyperparameters, we deploy the median heuristic using 20 pairs of images.

We illustrate the results in Figure A.11. Under the null, our test does not reject more often than the required 5%, but its power increases with sample size under the alternative, reaching power one after processing ≈ 500 pairs of digits (points from P_{XY}) on average.

A.6 Scaling Sequential Testing Procedures

Updating the wealth process at each round requires evaluating the payoff function at a new pair of observations (and hence computing the witness function corresponding to a chosen dependence criterion). In this section, we provide details about the ways of reducing the computational complexity of this step, which are necessary to scale the proposed sequential testing frameworks to moderately large sample sizes. Note that the proposed implementation of COCO allows updating kernel hyperparameters on the fly. In contrast, linear-time updates for HSIC require fixing kernel hyperparameters in advance.

A.6.1 Incomplete/Pivoted Cholesky Decomposition for COCO and KCC

Suppose that we want to evaluate COCO payoff function on the next pair of points $(X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})$. In order to do so, we need to compute $g_{1,t}$ and $g_{2,t}$, that is solve the generalized eigenvalue problem. Note that solving generalized eigenvalue problem at each iteration could be computationally prohibitive. One simple way is to use a random subsample of datapoints when performing witness function estimation, e.g., once the sample size n exceeds n_s, e.g., $n_s = 25$, we randomly subsample (without replacement) a sample of size n_s to estimate witness functions. Alternatively, a common approach is to reduce computational burden through incomplete Cholesky decomposition. The idea is to use the fact that kernel matrices tend to demonstrate rapid spectrum decay, and thus

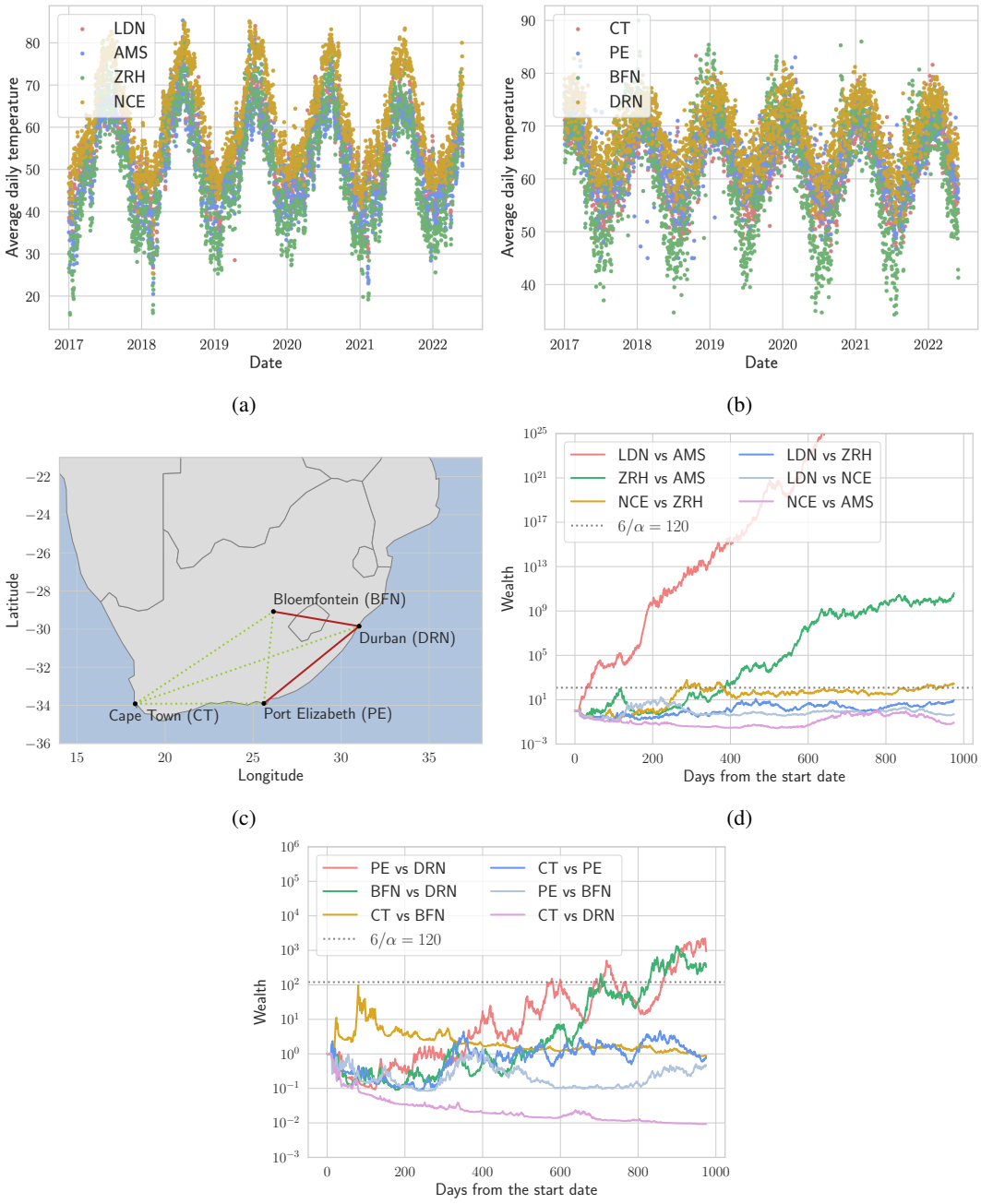


Figure A.10: Temperatures for selected cities in Europe (subplot (a)) and South Africa (subplot (b)) share similar seasonal patterns. Map (subplot (c)) where solid red lines connect those cities for which the null is rejected. SKIT supports our conjecture about dependent temperature fluctuations for geographically close cities. For completeness, we also plot wealth processes for SKIT used on weather data for Europe (subplot (d)) and South Africa (subplot (e)).

low-rank approximations can be used to scale the procedures. Suppose that $K \approx G_1 G_1^T$ and $L \approx G_2 G_2^T$ where G_i's are lower triangular matrices of size $n \times M$ (M depends on the preset approximation error level). After computing

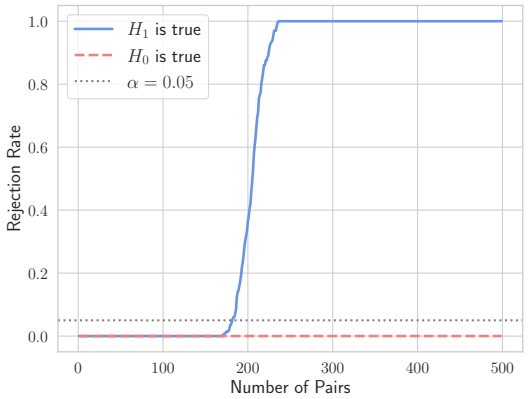


Figure A.11: Rejection rate for SKIT on MNIST data. Under the null (red dashed line), our test does not reject more often than the required 5%, but its power increases with sample size under the alternative (blue solid line). Each pair corresponds to two points from P_{XY}, and hence, SKIT reaches power one after processing ≈ 500 pairs of images on average.

Cholesky decomposition, we center both matrices via left multiplication by H and compute SVDs of HG_1 and HG_2, that is, $HG_1 = U_1\Lambda_1 V_1^\top$ and $HG_2 = U_2\Lambda_2 V_2^\top$. We have:

$$\tilde{K} \approx U_1\Lambda_1^2 U_1^\top, \quad \tilde{L} \approx U_2\Lambda_2^2 U_2^\top.$$

Our goal is to find the largest eigenvalue/eigenvector pair for $Ax = \gamma Bx$ for a PD matrix B. Since:

$$Ax = \gamma Bx \iff B^{-1/2}AB^{-1/2}(B^{1/2}x) = \gamma(B^{1/2}x),$$

it suffices to leading eigenvalue/eigenvector pair for:

$$B^{-1/2}AB^{-1/2}y = \gamma y.$$

Then $x = B^{-1/2}y$ is a generalized eigenvector for the initial problem.

COCO. For COCO, we have:

$$B = \begin{pmatrix} \tilde{K} & 0 \\ 0 & \tilde{L} \end{pmatrix} \approx \begin{pmatrix} U_1\Lambda_1^2 U_1^\top & 0 \\ 0 & U_2\Lambda_2^2 U_2^\top \end{pmatrix} = \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} \Lambda_1^2 & 0 \\ 0 & \Lambda_2^2 \end{pmatrix} \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}^\top$$

$$\implies B^{-1/2} \approx \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & \Lambda_2^{-1} \end{pmatrix} \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}^\top =: \mathcal{U}\Lambda^{-1}\mathcal{U}^\top.$$

We also have:

$$A \approx \begin{pmatrix} 0 & \frac{1}{n} U_1 \Lambda_1^2 U_1^\top U_2 \Lambda_2^2 U_2^\top \\ \frac{1}{n} U_2 \Lambda_2^2 U_2^\top U_1 \Lambda_1^2 U_1^\top & 0 \end{pmatrix}$$
$$= \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{n} \Lambda_1^2 U_1^\top U_2 \Lambda_2^2 \\ \frac{1}{n} \Lambda_2^2 U_2^\top U_1 \Lambda_1^2 & 0 \end{pmatrix} \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}^\top.$$

Thus we have:

$$B^{-1/2} A B^{-1/2} \approx \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{n} \Lambda_1 U_1^\top U_2 \Lambda_2 \\ \frac{1}{n} \Lambda_2 U_2^\top U_1 \Lambda_1 & 0 \end{pmatrix} \begin{pmatrix} U_1 & 0 \\ 0 & U_2 \end{pmatrix}^\top.$$

Hence, we only need to compute the leading eigenvector (say, z^*) for:

$$\begin{pmatrix} 0 & \frac{1}{n} \Lambda_1 U_1^\top U_2 \Lambda_2 \\ \frac{1}{n} \Lambda_2 U_2^\top U_1 \Lambda_1 & 0 \end{pmatrix} \in \mathbb{R}^{(M_1+M_2)\times(M_1+M_2)}.$$

It implies that the leading eigenvector for $B^{-1/2} A B^{-1/2}$ is then $\mathcal{U} z^*$, and the solution for the generalized eigenvalue problem is given by:

$$\mathcal{U} \Lambda^{-1} z^* = \begin{pmatrix} U_1 \Lambda_1^{-1} z_1^* \\ U_2 \Lambda_2^{-1} z_2^* \end{pmatrix} =: \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}.$$

Next, we need to normalize this vector of coefficients appropriately, i.e., we need to guarantee that $\left\| \tilde{K}^{1/2} \alpha \right\|_2 = 1$ and $\left\| \tilde{L}^{1/2} \beta \right\|_2 = 1$, and thus re-normalizing naively is quadratic in n. Instead, note that in order to compute incomplete Cholesky decomposition, we choose a tolerance parameter δ so that: $\left\| PKP^\top - G_1 G_1^\top \right\|_* = \left\| K - G_1 G_1^\top \right\|_* \leq \delta$ (nuclear norm). Let $\Delta = K - G_1 G_1^\top$. We know that:

$$\alpha^\top \tilde{K} \alpha = \alpha^\top HKH\alpha = \alpha^\top H(\Delta + G_1 G_1^\top) H\alpha = \alpha^\top H \Delta H \alpha + \alpha^\top H G_1 G_1^\top H \alpha$$

First, note that $\alpha^\top H \Delta H \alpha \leq \delta \|H\alpha\|_2^2$. Next,

$$G_1^\top H = V_1 \Lambda_1 U_1^\top.$$

Given an initial vector of parameters α_0 and β_0, vectors of coefficients can be normalized in linear time using

$$\alpha = \frac{\alpha_0}{\sqrt{\left\| G_1^\top H \alpha_0 \right\|_2^2 + \delta \left\| H \alpha_0 \right\|_2^2}} = \frac{U_1 \Lambda_1^{-1} z_1^*}{\sqrt{\|V_1 z_1^*\|_2^2 + \delta \left\| H \alpha_0 \right\|_2^2}} = \frac{U_1 \Lambda_1^{-1} z_1^*}{\sqrt{\|z_1^*\|_2^2 + \delta \left\| H \alpha_0 \right\|_2^2}},$$
$$\beta = \frac{\beta_0}{\sqrt{\left\| G_2^\top H \beta_0 \right\|_2^2 + \delta \left\| H \beta_0 \right\|_2^2}} = \frac{U_2 \Lambda_2^{-1} z_2^*}{\sqrt{\|V_2 z_2^*\|_2^2 + \delta \left\| H \beta_0 \right\|_2^2}} = \frac{U_2 \Lambda_2^{-1} z_2^*}{\sqrt{\|z_2^*\|_2^2 + \delta \left\| H \beta_0 \right\|_2^2}}.$$

For small δ, we essentially normalize by $\alpha_0^\top \tilde{K} \alpha_0$ and $\beta_0^\top \tilde{L} \beta_0$ as expected. It also makes sense to use $\delta = n \cdot \delta_0$. Still, re-estimating the witness functions after processing $2t$, $t \geq 1$ points is computationally intensive. In contrast to HSIC, for which there are no clear benefits of skipping certain estimation steps, for COCO we estimate the witness functions after processing $2t^2$, $t \geq 1$ points.

KCC. For KCC, we have:

$$\begin{aligned}
B &= \begin{pmatrix} \kappa_1 \tilde{K} + \frac{1}{n}\tilde{K}^2 & 0 \\ 0 & \kappa_2 \tilde{L} + \frac{1}{n}\tilde{L}^2 \end{pmatrix} \\
&\approx \begin{pmatrix} \kappa_1 U_1 \Lambda_1^2 U_1^\top + \frac{1}{n} U_1 \Lambda_1^4 U_1^\top & 0 \\ 0 & \kappa_2 U_2 \Lambda_2^2 U_2^\top + \frac{1}{n} U_2 \Lambda_2^4 U_2^\top \end{pmatrix} \\
&= \begin{pmatrix} U_1 \Lambda_1^2 \left(\kappa_1 \mathbf{I}_n + \Lambda_1^2 \frac{1}{n}\right) U_1^\top & 0 \\ 0 & U_2 \Lambda_2^2 \left(\kappa_2 \mathbf{I}_n + \Lambda_2^2 \frac{1}{n}\right) U_2^\top \end{pmatrix} \\
&= \mathcal{U} \begin{pmatrix} \Lambda_1^2 \left(\kappa_1 \mathbf{I}_n + \Lambda_1^2 \frac{1}{n}\right) & 0 \\ 0 & \Lambda_2^2 \left(\kappa_2 \mathbf{I}_n + \Lambda_2^2 \frac{1}{n}\right) \end{pmatrix} \mathcal{U}^\top,
\end{aligned}$$

which implies that:

$$B^{-1/2} \approx \mathcal{U} \begin{pmatrix} \left(\kappa_1 \mathbf{I}_n + \Lambda_1^2 \frac{1}{n}\right)^{-1/2} \Lambda_1^{-1} & 0 \\ 0 & \left(\kappa_2 \mathbf{I}_n + \Lambda_2^2 \frac{1}{n}\right)^{-1/2} \Lambda_2^{-1} \end{pmatrix} \mathcal{U}^\top.$$

Recall that:

$$A \approx \mathcal{U} \begin{pmatrix} 0 & \frac{1}{n}\Lambda_1^2 U_1^\top U_2 \Lambda_2^2 \\ \frac{1}{n}\Lambda_2^2 U_2^\top U_1 \Lambda_1^2 & 0 \end{pmatrix} \mathcal{U}^\top.$$

Thus,

$$\begin{aligned}
B^{-1/2} A B^{-1/2} &\approx \mathcal{U} \begin{pmatrix} 0 & M_* \\ M_*^\top & 0 \end{pmatrix} \mathcal{U}^\top, \\
\text{where} \quad M_* &= \frac{1}{n} \left(\kappa_1 \mathbf{I}_n + \Lambda_1^2 \frac{1}{n}\right)^{-1/2} \Lambda_1 U_1^\top U_2 \Lambda_2 \left(\kappa_2 \mathbf{I}_n + \Lambda_2^2 \frac{1}{n}\right)^{-1/2}.
\end{aligned}$$

Equivalently,

$$M_* = \frac{1}{n} \rho_{\kappa_1}(\Lambda_1) \Lambda_1 U_1^\top U_2 \Lambda_2 \rho_{\kappa_2}(\Lambda_2), \quad \text{where} \quad \rho_\kappa(x) = \frac{1}{\sqrt{x^2/n + \kappa}}.$$

Hence, we only need to compute the leading eigenvector (say, z^*) for:

$$\begin{pmatrix} 0 & M_* \\ M_*^\top & 0 \end{pmatrix}$$

It implies that the leading eigenvector for $B^{-1/2}AB^{-1/2}$ is then $\mathcal{U}z^*$. For the initial generalized eigenvalue problem, an approximate solution (due to using low-rank approximations of kernel matrices) is given by:

$$B^{-1/2}\mathcal{U}z^* = \begin{pmatrix} U_1\rho_{\kappa_1}(\Lambda_1)\Lambda_1^{-1}z_1^* \\ U_2\rho_{\kappa_2}(\Lambda_2)\Lambda_2^{-1}z_2^* \end{pmatrix} = \begin{pmatrix} U_1\Lambda_1^{-1}\rho_{\kappa_1}(\Lambda_1)z_1^* \\ U_2\Lambda_2^{-1}\rho_{\kappa_2}(\Lambda_2)z_2^* \end{pmatrix} =: \begin{pmatrix} \alpha_0 \\ \beta_0 \end{pmatrix}.$$

Next, we need to normalize this vector of coefficients appropriately, i.e., we need to guarantee that $\left\|\tilde{K}^{1/2}\alpha\right\|_2 = 1$ and $\left\|\tilde{L}^{1/2}\beta\right\|_2 = 1$, and thus re-normalizing naively is quadratic in n. Instead, note that in order to compute incomplete Cholesky decomposition, we choose a tolerance parameter δ so that: $\left\|PKP^\top - G_1G_1^\top\right\|_* = \left\|K - G_1G_1^\top\right\|_* \leq \delta$ (nuclear norm). Let $\Delta = K - G_1G_1^\top$. We know that:

$$\alpha^\top\tilde{K}\alpha = \alpha^\top HKH\alpha = \alpha^\top H(\Delta + G_1G_1^\top)H\alpha = \alpha^\top H\Delta H\alpha + \alpha^\top HG_1G_1^\top H\alpha$$

First, note that $\alpha^\top H\Delta H\alpha \leq \delta\|H\alpha\|_2^2$. Next,

$$G_1^\top H = V_1\Lambda_1U_1^\top.$$

Given an initial vector of parameters α_0 and β_0, vectors of coefficients can be normalized in linear time using

$$\alpha = \frac{\alpha_0}{\sqrt{\left\|G_1^\top H\alpha_0\right\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}} = \frac{U_1\rho_{\kappa_1}(\Lambda_1)\Lambda_1^{-1}z_1^*}{\sqrt{\|V_1\rho_{\kappa_1}(\Lambda_1)z_1^*\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}} = \frac{U_1\rho_{\kappa_1}(\Lambda_1)\Lambda_1^{-1}z_1^*}{\sqrt{\|\rho_{\kappa_1}(\Lambda_1)z_1^*\|_2^2 + \delta\left\|H\alpha_0\right\|_2^2}},$$
$$\beta = \frac{\beta_0}{\sqrt{\left\|G_2^\top H\beta_0\right\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}} = \frac{U_2\rho_{\kappa_2}(\Lambda_2)\Lambda_2^{-1}z_2^*}{\sqrt{\|V_2\rho_{\kappa_2}(\Lambda_2)z_2^*\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}} = \frac{U_2\rho_{\kappa_2}(\Lambda_2)\Lambda_2^{-1}z_2^*}{\sqrt{\|\rho_{\kappa_2}(\Lambda_2)z_2^*\|_2^2 + \delta\left\|H\beta_0\right\|_2^2}}.$$

A.6.2 Linear-time Updates of the HSIC Payoff Function

Suppose that we want to evaluate HSIC payoff function on the next pair of points $(X_{2t+1}, Y_{2t+1}), (X_{2t+2}, Y_{2t+2})$. In order to do so, we need to compute: $\widehat{g}_t(X_{2t+2}, Y_{2t+2})$. It is clear that the computational of evaluating $\widehat{\mu}_{XY}(x, y)$ and $(\widehat{\mu}_X \otimes \widehat{\mu}_Y)(x, y)$ on a given pair (x, y) is linear in t. However, we also need to compute the normalization constant:

$$\|\widehat{\mu}_{XY} - \widehat{\mu}_X \otimes \widehat{\mu}_Y\|_{\mathcal{G}\otimes\mathcal{H}}. \tag{A.21}$$

Recall that:

$$\left\|\widehat{\mu}_{XY}^{(t)} - \widehat{\mu}_X^{(t)} \otimes \widehat{\mu}_Y^{(t)}\right\|_{\mathcal{G}\otimes\mathcal{H}}^2 = \frac{1}{(2t)^2}\mathrm{tr}K^{(t)}H^{(t)}L^{(t)}H^{(t)},$$

where $K^{(t)}$ and $L^{(t)}$ are kernel matrices corresponding to the first $2t$ pairs, $H^{(t)} := \mathbf{I}_{2t} - \frac{1}{2t}\mathbf{1}_{2t}\mathbf{1}_{2t}^\top$. Instead of computing the normalization constant naively, we next establish a more efficient way of computing (A.21) in time

linear in t by caching certain values. Introduce:

$$\Delta_1^{(t)} = \sum_{i,j=1}^{2t} K_{ij} L_{ij} = \mathrm{tr} K^{(t)} L^{(t)},$$
$$\Delta_2^{(t)} = \sum_{i,j}^{2t} K_{ij} = \mathbf{1}_{2t}^\top K^{(t)} \mathbf{1}_{2t},$$
$$\Delta_3^{(t)} = \sum_{i,j}^{2t} L_{ij} = \mathbf{1}_{2t}^\top L^{(t)} \mathbf{1}_{2t},$$
$$\Delta_4^{(t)} = \sum_{i=1}^{2t} \sum_{j,q=1}^{2t} K_{ij} L_{iq} = \mathbf{1}_{2t}^\top K^{(t)} L^{(t)} \mathbf{1}_{2t}.$$

We have:

$$\left\| \widehat{\mu}_{XY}^{(t+1)} - \widehat{\mu}_X^{(t+1)} \otimes \widehat{\mu}_Y^{(t+1)} \right\|_{\mathcal{G}\otimes\mathcal{H}}^2 = \frac{1}{(2t+2)^2} \Delta_1^{(t+1)} + \frac{1}{(2t+2)^4} \Delta_2^{(t+1)} \cdot \Delta_3^{(t)} - \frac{2}{(2t+2)^3} \Delta_4^{(t+1)}.$$

Next, we show how to speed up computations via caching certain intermediate values. Kernel matrices have the following structure:

$$K^{(t+1)} = \begin{pmatrix} K^{(t)} & K_{\cdot,2t+1} & K_{\cdot,2t+2} \\ K_{\cdot,2t+1}^\top & K_{2t+1,2t+1} & K_{2t+1,2t+2} \\ K_{\cdot,2t+2}^\top & K_{2t+2,2t+1} & K_{2t+2,2t+2} \end{pmatrix}, \quad L^{(t+1)} = \begin{pmatrix} L^{(t)} & L_{\cdot,2t+1} & L_{\cdot,2t+2} \\ L_{\cdot,2t+1}^\top & L_{2t+1,2t+1} & L_{2t+1,2t+2} \\ L_{\cdot,2t+2}^\top & L_{2t+2,2t+1} & L_{2t+2,2t+2} \end{pmatrix},$$

where $K_{\cdot,2t+1}, K_{\cdot,2t+2}, L_{\cdot,2t+1}, L_{\cdot,2t+2} \in \mathbb{R}^{2t}$ contain kernel function evaluations:

$$K_{\cdot,m} = \begin{pmatrix} k(X_1, X_m) \\ \vdots \\ k(X_{2t}, X_m) \end{pmatrix}, \quad L_{\cdot,m} = \begin{pmatrix} l(Y_1, Y_m) \\ \vdots \\ l(Y_{2t}, Y_m) \end{pmatrix}, \quad m \in \{2t+1, 2t+2\}.$$

First, it is easy to derive that:

$$\begin{aligned} \mathrm{tr} K^{(t+1)} L^{(t+1)} = {} & \mathrm{tr} K^{(t)} L^{(t)} + 2(L_{\cdot,2t+1}^\top K_{\cdot,2t+1}) + 2(L_{\cdot,2t+2}^\top K_{\cdot,2t+2}) + \\ & + K_{2t+1,2t+1} L_{2t+1,2t+1} + K_{2t+2,2t+2} L_{2t+2,2t+2} \\ & + K_{2t+1,2t+2} L_{2t+2,2t+1} + K_{2t+2,2t+1} L_{2t+1,2t+2}. \end{aligned}$$

Thus, if the value $\mathrm{tr}K^{(t)}L^{(t)}$ is cached, then $\mathrm{tr}K^{(t+1)}L^{(t+1)}$ can be computed in linear time. Note that:

$$K^{(t+1)}\mathbf{1}_{2t+2} = \begin{pmatrix} K^{(t)}\mathbf{1}_{2t} + k_{\cdot,2t+1} + k_{\cdot,2t+2} \\ K_{\cdot,2t+1}^{\top}\mathbf{1}_{2t} + K_{2t+1,2t+1} + K_{2t+1,2t+2} \\ K_{\cdot,2t+2}^{\top}\mathbf{1}_{2t} + K_{2t+2,2t+1} + K_{2t+2,2t+2} \end{pmatrix},$$

which can be computed in linear time if $K^{(t)}\mathbf{1}_{2t}$ is stored (similar result holds for $L^{(t+1)}\mathbf{1}_{2t+2}$). It thus follows that $\mathbf{1}_{2t+2}^{\top}K^{(t+1)}\mathbf{1}_{2t+2}$, $\mathbf{1}_{2t+2}^{\top}L^{(t+1)}\mathbf{1}_{2t+2}$ and $\mathbf{1}_{2t+2}^{\top}K^{(t+1)}L^{(t+1)}\mathbf{1}_{2t+2}$ can all be computed in linear time. To sum up, we need to cache $\mathrm{tr}K^{(t)}L^{(t)}$, $K^{(t)}\mathbf{1}_{2t}$, $L^{(t)}\mathbf{1}_{2t}$ to compute the normalization constant in linear time.

Appendix B

Additional Results for Chapter 3

B.1 Regression-based Independence Testing

Regression-based independence tests represent an alternative to classification-based approaches in settings where a data stream $((X_t, Y_t))_{t\geq 1}$ may be processed directly as feature-response pairs. Suppose that one selects a functional class $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$ for performing such prediction task, and let ℓ denote a loss function that evaluates the quality of predictions. For example, if $(Y_t)_{t\geq 1}$ is a sequence of univariate random variables, one can use the squared loss: $\ell(g(x), y) = (g(x) - y)^2$, or the absolute loss: $\ell(g(x), y) = |g(x) - y|$.

Such tests rely on the following idea: if the alternative H_1 in (3.2b) is true and a sequence of sequentially updated predictors $(g_t)_{t\geq 1}$ has nontrivial predictive power, then the losses on random instances drawn from the joint distribution P_{XY} are expected to be less on average than the losses on random instances from $P_X \times P_Y$. For the t-th pair of points from P_{XY}, we can label the losses of g_t on all possible (X, Y)-pairs as

$$\begin{aligned} L_{2t-1} &= \ell\left(g_t(X_{2t-1}), Y_{2t-1}\right), \quad & L_{2t} &= \ell\left(g_t(X_{2t}), Y_{2t}\right), \\ L'_{2t-1} &= \ell\left(g_t(X_{2t-1}), Y_{2t}\right), \quad & L'_{2t} &= \ell\left(g_t(X_{2t}), Y_{2t-1}\right). \end{aligned} \tag{B.1}$$

One can view this problem as sequential two-sample testing under distribution drift (due to incremental learning of $(g_t)_{t\geq 1}$). Hence, one may use either Seq-C-2ST from Section 3.2 or sequential kernelized 2ST of Shekhar and Ramdas (2021) on the resulting sequence of the losses on observations from P_{XY} and $P_X \times P_Y$. In what follows, we analyze a direct approach where testing is performed by comparing the losses on instances drawn from the two distributions. A critical difference with a construction of Seq-C-2ST is that to design a valid betting strategy one has to ensure that the payoff functions are lower bounded by negative one.

B.1.1 Proxy Regression-based Independence Test

To avoid cases when some expected values are not well-defined, we assume for simplicity that $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$ for som $d \geq 1$: $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_2 \leq B_1\right\}$ for some $B_1 > 0$. Similarly, we assume that $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$: $\mathcal{Y} = \{y \in \mathbb{R} : |y| \leq B_2\}$ for some $B_2 > 0$. We note that the construction of the regression-based IT will not require explicit knowledge of constants B_1 and B_2. First, we consider a setting where an instance either from the joint distribution or an instance from the product of the marginal distributions is observed at each round.

Definition 8 (Proxy Setting)**.** *Suppose that we observe a stream of i.i.d. observations* $((X_t, Y_t, W_t))_{t\geq 1}$*, where* $W_t \sim \mathrm{Rademacher}(1/2)$*, the distribution of* $(X_t, Y_t) \mid W_t = +1$ *is* $P_X \times P_Y$*, and that of* $(X_t, Y_t) \mid W_t = -1$ *is* P_{XY}*. The goal is to design a test for the following pair of hypotheses:*

$$H_0 : P_{XY} = P_X \times P_Y, \tag{B.2a}$$

$$H_1 : P_{XY} \neq P_X \times P_Y. \tag{B.2b}$$

Oracle Proxy Sequential Regression-based IT. To construct an oracle test, we assume having access to the oracle predictor $g_\star : \mathcal{X} \to \mathcal{Y}$, e.g., the minimizer of the squared risk is $g_\star(x) = \mathbb{E}\left[Y \mid X = x\right]$. Formalizing the above intuition, we use $\mathbb{E}\left[W\ell(g_\star(X), Y)\right]$ as a natural way for measuring dependence between X and Y. To enforce boundedness of the payoff functions, we use ideas of the tests for symmetry from (Ramdas et al., 2020; Shekhar and Ramdas, 2021; Podkopaev et al., 2023; Shaer et al., 2023), namely we use a composition with an odd function:

$$f^{\mathrm{r}}_\star(X_t, Y_t, W_t) = \tanh\left(s_\star \cdot W_t \cdot \ell(g_\star(X_t), Y_t)\right) \in [-1, 1], \tag{B.3}$$

where $s_\star > 0$ is an appropriately selected scaling factor*. Since under H_0 in (B.2a), $s_\star \cdot W_t \cdot \ell(g_\star(X_t), Y_t)$ is a random variable that is symmetric around zero, it follows that $\mathbb{E}[f^{\mathrm{r}}_\star(X_t, Y_t, W_t)] = 0$, and, using the argument analogous to the proof of Theorem 3.1, we can easily deduce that a sequential IT based on $f^{\mathrm{r}}_\star$ controls the type I error control. The scaling factor $s_\star$ is selected in a way that guarantees that, if H_1 in (B.2b) is true and if $\mathbb{E}\left[W\ell(g_\star(X), Y)\right] > 0$, then $\mathbb{E}\left[f^{\mathrm{r}}_\star(X, Y, W)\right] > 0$, which is a sufficient condition for consistency of the oracle test. In particular, we show that it suffices to consider:

$$s_\star = \sqrt{\frac{2\mu_\star}{\nu_\star}}, \tag{B.4a}$$

$$\text{where} \qquad \mu_\star = \mathbb{E}\left[W\ell(g_\star(X), Y)\right], \tag{B.4b}$$

$$\nu_\star = \mathbb{E}\left[(1 + W)\left(\ell(g_\star(X), Y)\right)^3\right]. \tag{B.4c}$$

Without loss of generality, we assume that $\nu_\star$ is bounded away from zero (which is a very mild assumption since $\nu_\star$ essentially corresponds to a cubic risk of $g_\star$ on data drawn from the product of the marginal distributions $P_X \times$

*We note that rescaling is important for arguing about consistency and not the type I error control.

P_Y). Let the *oracle* regression-based wealth process $\left(\mathcal{K}_t^{\mathrm{r},\star}\right)_{t\geq 0}$ be defined by using the payoff function (B.3) with a scaling factor defined in (B.4a), along with a predictable sequence of betting fractions $(\lambda_t)_{t\geq 1}$ selected via the ONS strategy (Algorithm 5). We have the following result about the oracle regression-based IT, whose proof is deferred to Appendix B.4.4.

Theorem B.1. *The following claims hold for the oracle sequential regression-based IT based on* $\left(\mathcal{K}_t^{\mathrm{r},\star}\right)_{t\geq 0}$*:*

1. *Suppose that* H_0 *in* (B.2a) *is true. Then the test ever stops with probability at most* α*:* $\mathbb{P}_{H_1}(\tau < \infty) \leq \alpha$.
2. *Suppose that* H_1 *in* (B.2b) *is true. Further, suppose that:* $\mathbb{E}\left[W\ell(g_\star(X), Y)\right] > 0$*. Then the test is consistent:* $\mathbb{P}_{H_1}(\tau < \infty) = 1$.

Practical Proxy Sequential Regression-based IT. To construct a practical test, we use a sequence of predictors $(g_t)_{t\geq 1}$ that are updated sequentially as more data are observed. We write $\mathcal{A}_{\mathrm{r}} : (\cup_{t\geq 1}(\mathcal{X}\times\mathcal{Y})^t) \times \mathcal{G} \to \mathcal{G}$ to denote a chosen regressor learning algorithm which maps a training dataset of any size and previously used predictor, to an updated predictor. We start with $\mathcal{D}_0 = \emptyset$ and some initial guess $g_1 \in \mathcal{G}$. At round t, we use the payoff function:

$$f_t^{\mathrm{r}}(X_t, Y_t, W_t) = \tanh\left(s_t \cdot W_t \cdot \ell(g_t(X_t), Y_t)\right). \tag{B.5}$$

where a sequence of predictable scaling factors $(s_t)_{t\geq 1}$ is defined as follows: we set $s_0 = 0$ and define:

$$s_t = \sqrt{\frac{2\mu_t}{\nu_t}}, \tag{B.6a}$$

$$\text{where} \qquad \mu_t = \left(\frac{1}{t-1}\sum_{i=1}^{t-1} W_i \cdot \ell(g_i(X_i), Y_i)\right) \vee 0, \tag{B.6b}$$

$$\nu_t = \frac{1}{t-1}\sum_{i=1}^{t-1}(1+W_i)\cdot\left(\ell(g_i(X_i), Y_i)\right)^3. \tag{B.6c}$$

After (X_t, Y_t, W_t) has been used for betting, we update a training dataset: $\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(X_t, Y_t, W_t)\}$, and an existing predictor: $g_{t+1} = \mathcal{A}_{\mathrm{r}}(\mathcal{D}_t, g_t)$. We summarize this practical sequential 2ST in Algorithm 9.

For simplicity, we consider a class of functions $\mathcal{G} := \{g_\theta : \mathcal{X} \to \mathcal{Y},\ \theta \in \Theta\}$ for some parameter set Θ which we assume to be a subset of a metric space. In this case, a sequence of predictors $(g_t)_{t\geq 1}$ is associated with the corresponding sequence of parameters $(\theta_t)_{t\geq 1}$: for $t \geq 1$, $g_t(\cdot) = g(\cdot;\theta_t)$ for some $\theta_t \in \Theta$. To argue about the consistency of the resulting test, we make two assumptions.

Assumption 5 (Smoothness)**.** *We assume that:*

- *Predictors in* $\mathcal{G}$ *are* L_1*-Lipschitz smooth:*

$$\sup_{x\in\mathcal{X}} |g(x;\theta) - g(x;\theta')| \leq L_1 \left\|\theta - \theta'\right\|, \quad \forall \theta, \theta' \in \Theta. \tag{B.7}$$

Algorithm 9 Proxy Sequential Regression-based IT

Input: significance level $\alpha \in (0,1)$, data stream $((X_t, Y_t, W_t))_{t\geq 1}$, $g_1(z) \equiv 0$, $\mathcal{A}_{\mathrm{r}}$, $\mathcal{D}_0 = \emptyset$, $\lambda_1^{\mathrm{ONS}} = 0$, $s_1 = 0$.
for $t = 1, 2, \ldots$ **do**
  Evaluate the payoff $f_t^{\mathrm{r}}(X_t, Y_t, W_t)$ as in (B.5);
  Using λ_t^{ONS}, update the wealth process $\mathcal{K}_t^{\mathrm{r}}$ as in (3.5);
  if $\mathcal{K}_t^{\mathrm{r}} \geq 1/\alpha$ **then**
    Reject H_0 and stop;
  else
    Update the training dataset: $\mathcal{D}_t := \mathcal{D}_{t-1} \cup \{(X_t, Y_t)\}$;
    Update predictor: $g_{t+1} = \mathcal{A}_{\mathrm{r}}(\mathcal{D}_t, g_t)$;
    Compute s_{t+1} as in (B.6a);
    Compute $\lambda_{t+1}^{\mathrm{ONS}}$ (Algorithm 5) using $f_t^{\mathrm{r}}(X_t, Y_t, W_t)$;

- *The loss function ℓ is L_2-Lipschitz smooth:*

$$\sup_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} |\ell(g(x;\theta), y) - \ell(g(x;\theta'), y)| \leq L_2 \sup_{x\in\mathcal{X}} |g(x;\theta) - g(x;\theta')|, \quad \forall \theta, \theta' \in \Theta. \tag{B.8}$$

In words, Assumption (B.7) states that the outputs of predictors, whose parameters are close, will also be close. Assumption (B.8) states that that the losses of two predictors, whose outputs are close, will also be close. For example, if $\mathcal{G}$ is a class of linear predictors: $g_\theta(x) = \theta^\top x$, $x \in \mathcal{X}$, then Assumption 5 will be trivially satisfied for the squared and the absolute losses if $\mathcal{X}$ and $\mathcal{Y}$ are bounded. Note that we do not need an explicit knowledge of L_1 or L_2 for designing a test. Second, we make a *learnability* assumption about algorithm $\mathcal{A}_{\mathrm{r}}$.

Assumption 6 (Learnability)**.** *Suppose that H_1 in* (B.2b) *is true. We assume that the regressor learning algorithm $\mathcal{A}_{\mathrm{r}}$ is such that for the resulting sequence of parameters $(\theta_t)_{t\geq 1}$, it holds that $\theta_t \overset{\text{a.s.}}{\to} \theta_\star$, where $\theta_\star$ is a random variable taking values in Θ and $\mathbb{E}\left[W\ell(g(X;\theta_\star), Y) \mid \theta_\star\right] \overset{\text{a.s.}}{>} 0$, where $(X, Y, W) \perp\!\!\!\perp \theta_\star$.*

We conclude with the following result for the practical proxy sequential regression-based IT, whose proof is deferred to Appendix B.4.4.

Theorem B.2. *The following claims hold for the proxy sequential regression-based IT (Algorithm 9):*

1. *Suppose that H_0 in* (B.2a) *is true. Then the test ever stops with probability at most α: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*
2. *Suppose that H_1 in* (B.2b) *is true. Further, suppose that Assumptions 5 and 6 are satisfied. Then the test is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$.*

Sequential Regression-based Independence Test (Seq-R-IT). Next, we instantiate this test for the sequential independence testing setting (as per Definition 2) where we observe sequence $((X_t, Y_t))_{t\geq 1}$, where $(X_t, Y_t) \overset{\text{iid}}{\sim} P_{XY}$, $t \geq 1$. Analogous to Section 3.3, we bet on the outcome of two observations drawn from the joint distribution P_{XY}.

To proceed, we derandomize the payoff function (B.5) and consider

$$f_t^{\mathrm{r}}((X_{2t-1},Y_{2t-1}),(X_{2t},Y_{2t})) = \frac{1}{4}\left(\tanh\left(s_t\cdot\ell\left(g_t(X_{2t-1}),Y_{2t}\right)\right)+\tanh\left(s_t\cdot\ell\left(g_t(X_{2t}),Y_{2t-1}\right)\right)\right) - \frac{1}{4}\left(\tanh\left(s_t\cdot\ell\left(g_t(X_{2t}),Y_{2t}\right)\right)-\tanh\left(s_t\cdot\ell\left(g_t(X_{2t-1}),Y_{2t-1}\right)\right)\right). \tag{B.9}$$

After betting on the outcome of the t-th pair of observations from P_{XY}, we update a training dataset:

$$\mathcal{D}_t = \mathcal{D}_{t-1}\cup\{(X_{2t-1},Y_{2t-1}),(X_{2t},Y_{2t})\},$$

and a predictive model: $\widehat{g}_{t+1}=\mathcal{A}_{\mathrm{r}}(\mathcal{D}_t,\widehat{g}_t)$.

B.1.2 Synthetic Experiments

To evaluate the performance of Seq-R-IT, we consider the *Gaussian linear model*. Let $(X_t)_{t\geq 1}$ and $(\varepsilon_t)_{t\geq 1}$ denote two independent sequences of i.i.d. standard Gaussian random variables. For $t\geq 1$, we take

$$(X_t,Y_t)=(X_t,X_t\beta+\varepsilon_t),$$

where $\beta\neq 0$ implies nonzero linear correlation (hence dependence). We consider 20 values of β equally spaced in $[0,1/2]$. For the comparison, we use:

1. *Seq-R-IT with ridge regression.* We use ridge regression as an underlying model: $\widehat{g}_t(x)=\beta_0^{(t)}+x\beta_1^{(t)}$, where

$$(\beta_0^{(t)},\beta_1^{(t)})=\operatorname*{arg\,min}_{\beta_0,\beta_1}\sum_{i=1}^{2(t-1)}\left(Y_i-X_i\beta_1-\beta_0\right)^2+\lambda\beta_1^2.$$

2. *Seq-C-IT with QDA.* Note that $P_{XY}=\mathcal{N}(\mu,\Sigma^+)$ and $P_X\times P_Y=\mathcal{N}(\mu,\Sigma^-)$, where

$$\mu=\begin{pmatrix}0\\0\end{pmatrix},\quad \Sigma^+=\begin{pmatrix}1&\beta\\\beta&1+\beta^2\end{pmatrix},\quad \Sigma^-=\begin{pmatrix}1&0\\0&1+\beta^2\end{pmatrix}.$$

   For this problem, an oracle predictor which minimizes the misclassification risk is

$$g^\star(x,y)=\frac{\varphi((x,y);\mu^+,\Sigma^+)-\varphi((x,y);\mu^-,\Sigma^-)}{\varphi((x,y);\mu^-,\Sigma^-)+\varphi((x,y);\mu^+,\Sigma^+)}\in[-1,1], \tag{B.10}$$

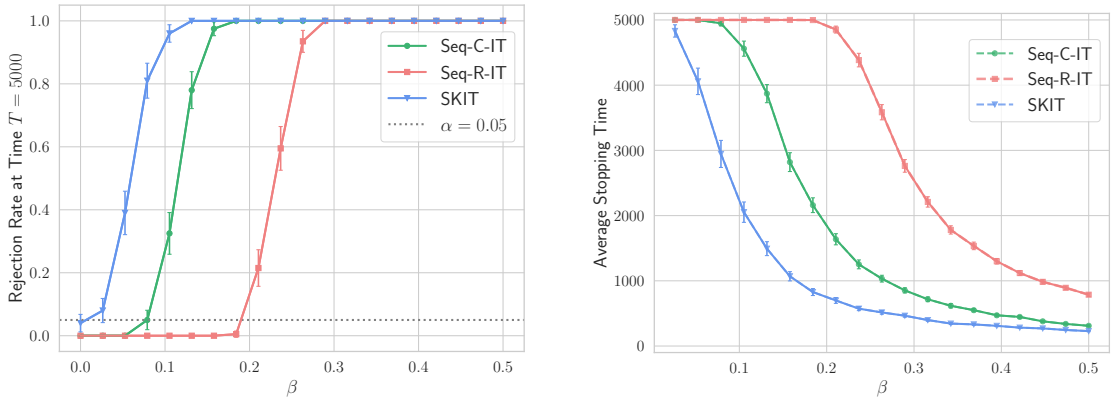
   where $\varphi((x,y);\mu,\Sigma)$ denotes the density of the Gaussian distribution $\mathcal{N}(\mu,\Sigma)$ evaluated at (x,y). Recall that $\mathcal{D}_{t-1}=\{(Z_i,+1)\}_{i\leq 2(t-1)}\cup\{(Z_i',-1)\}_{i\leq 2(t-1)}$ denotes the training dataset that is available at round t for training a predictor $\widehat{g}_t:\mathcal{X}\times\mathcal{Y}\to[-1,1]$. We deploy Seq-C-IT with an estimator $\widehat{g}_t$ of (B.10), obtained by

using plug-in estimates of $\mu^+, \Sigma^+, \mu^-, \Sigma^-$, computed from $\mathcal{D}_{t-1}$:

$$\widehat{\mu}_t^+ = \frac{1}{2(t-1)} \sum_{Z \in \mathcal{D}_{t-1}^+} Z, \qquad \widehat{\Sigma}_t^+ = \left(\frac{1}{2(t-1)} \sum_{Z \in \mathcal{D}_{t-1}^+} ZZ^\top \right) - (\widehat{\mu}_t^+)(\widehat{\mu}_t^+)^\top,$$

and $\widehat{\mu}_t^-, \widehat{\Sigma}_t^-$ are computed similarly from $\mathcal{D}_t^-$.

In addition, we also include HSIC-based SKIT to the comparison and defer the details regarding kernel hyperparameters to Appendix B.5.1. We set the monitoring horizon to $T = 5000$ points from P_{XY} and aggregate the results over 200 sequences of observations for each value of β. We illustrate the result in Figure B.1: while Seq-R-IT has high power for large values of β, we observe its inferior performance against Seq-C-IT (and SKIT) under the harder settings. Improving regression-based betting strategies, e.g., designing better scaling factors that still yield a provably consistent test, is an open question for future research.



(a) (b)

Figure B.1: Comparison between Seq-R-IT, Seq-C-IT and HSIC-based SKIT under the Gaussian linear model. Inspecting Figure B.1a at $\beta = 0$ confirms that all tests control the type I error. Non-surprisingly, kernel-based SKIT performs better than predictive tests under this model (no localized dependence). We also observe that Seq-C-IT performs better than Seq-R-IT.

B.2 Two-sample Testing with Unbalanced Classes

In Section 3.2, we developed a sequential 2ST under the assumption at each round, an instance from either P or Q is revealed with equal probability. Such assumption was reasonable for designing Seq-C-IT, where external randomization produced two instances from P_{XY} and $P_X \times P_Y$ at each round. Next, we generalize our sequential 2ST to a more general setting of unbalanced classes.

Definition 9 (Sequential two-sample testing with unbalanced classes)**.** *Let $\pi \in (0, 1)$. Suppose that we observe a stream of i.i.d. observations $((Z_t, W_t))_{t \geq 1}$, where $W_t \sim \text{Rademacher}(\pi)$, the distribution of $Z_t \mid W_t = +1$ is*

denoted P, and that of $Z_t \mid W_t = -1$ is denoted Q. We set the goal of designing a sequential test for the following pair of hypotheses:

$$H_0 : P = Q, \tag{B.11a}$$

$$H_1 : P \neq Q. \tag{B.11b}$$

For what follows, we will focus on the payoff based on the squared risk due to its relationship to the likelihood-ratio-based test (Remark 6). In particular, after correcting the likelihood under the null in (3.20) to account for a general positive class proportion π, we can deduce that (see Appendix B.4.5):

$$(1-\lambda_t)\cdot 1 + \lambda_t \cdot \frac{(\eta_t(Z_t))^{\mathbb{1}\{W_t=1\}}\,(1-\eta_t(Z_t))^{\mathbb{1}\{W_t=0\}}}{(\pi)^{\mathbb{1}\{W_t=1\}}\,(1-\pi)^{\mathbb{1}\{W_t=0\}}} = 1 + \lambda_t \cdot \frac{W_t\,(g_t(Z_t)-(2\pi-1))}{1+W_t(2\pi-1)}, \tag{B.12}$$

where $\eta_t(z) = (g_t(z)+1)/2$, and hence, a natural payoff function for the case with unbalanced classes is

$$f_t^{\mathrm{u}}(Z_t, W_t) = \frac{W_t\,(g_t(Z_t)-(2\pi-1))}{1+W_t(2\pi-1)}. \tag{B.13}$$

Note that the payoff for the balanced case (3.22b) is recovered by setting $\pi = 1/2$. It is easy to check that (see Appendix B.4.5): (a) $f_t^{\mathrm{u}}(z,w) \geq -1$ for any $(z,w) \in \mathcal{Z}\times\{-1,1\}$, and (b) if H_0 in (B.11a) is true, then $\mathbb{E}_{H_0}\left[f_t^{\mathrm{u}}(Z_t,W_t) \mid \mathcal{F}_{t-1}\right] = 0$, where $\mathcal{F}_{t-1} = \sigma(\{(Z_i,W_i)\}_{i\leq t-1})$. This in turn implies that a wealth process that relies on the payoff function f_t^{u} in (B.13) is a nonnegative martingale, and hence, the corresponding sequential 2ST is valid. However, the positive class proportion π, needed to use the payoff function (B.13), is generally unknown beforehand. First, let us consider the case when $\lambda_t = 1$, $t \geq 1$. In this case, the wealth of a gambler that uses the payoff function (B.13) after round t is

$$\mathcal{K}_t = \frac{\prod_{i=1}^t (\eta_i(Z_i))^{\mathbb{1}\{W_i=1\}}\,(1-\eta_i(Z_i))^{\mathbb{1}\{W_i=0\}}}{\prod_{i=1}^t \pi^{\mathbb{1}\{W_i=1\}}\,(1-\pi)^{\mathbb{1}\{W_i=0\}}}. \tag{B.14}$$

Note that:

$$\widehat{\pi}_t := \frac{1}{t}\sum_{i=1}^t \mathbb{1}\{W_t=1\} = \operatorname*{arg\,max}_{\pi\in[0,1]} \left(\prod_{i=1}^t \pi^{\mathbb{1}\{W_i=1\}}\,(1-\pi)^{\mathbb{1}\{W_i=0\}}\right),$$

is the MLE for π computed from $\{W_i\}_{i\leq t}$. In particular, if we consider a process $(\tilde{\mathcal{K}}_t)_{t\geq 0}$, where

$$\tilde{\mathcal{K}}_t := \frac{\prod_{i=1}^t (\eta_i(Z_i))^{\mathbb{1}\{W_i=1\}}\,(1-\eta_i(Z_i))^{\mathbb{1}\{W_i=0\}}}{\prod_{i=1}^t (\widehat{\pi}_t)^{\mathbb{1}\{W_i=1\}}\,(1-\widehat{\pi}_t)^{\mathbb{1}\{W_i=0\}}}, \quad t\geq 1,$$

it follows that $\tilde{\mathcal{K}}_t \leq \mathcal{K}_t$, $\forall t \geq 1$, meaning that $(\tilde{\mathcal{K}}_t)_{t\geq 0}$ is a process that is upper bounded by a nonnegative martingale with initial value one. This in turn implies that a test based on $(\tilde{\mathcal{K}}_t)_{t\geq 0}$ is a valid level-α sequential 2ST for the case of unknown class proportions. This idea underlies the running MLE sequential likelihood ratio test of Wasserman et al.

(2020) and has been recently considered in the context of two-sample testing by Pandeva et al. (2022). In case of nontrivial betting fractions: $(\lambda_t)_{t\geq 1}$, representation of the wealth process (B.14) no longer holds, and to proceed, we modify the rules of the game and use minibatching. A bet is placed on every b (say, 5 or 10) observations, meaning that for a given minibatch size $b \geq 1$, at round t we bet on $\{(Z_{b(t-1)+i}, W_{b(t-1)+i})\}_{i\in\{1,\ldots,b\}}$. The MLE of π computed from the t-th minibatch is

$$\widehat{\pi}_t = \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} \mathbb{1}\{W_i = +1\}.$$

We consider a payoff function of the following form:

$$f_t^{\mathrm{u}}\left(\left\{(Z_{b(t-1)+i}, W_{b(t-1)+i})\right\}_{i\in\{1,\ldots,b\}}\right) = \prod_{i=b(t-1)+1}^{bt} \left(\frac{1 + W_i g_t(Z_i)}{1 + W_i(2\widehat{\pi}_t - 1)}\right) - 1. \tag{B.15}$$

In words, the above payoff essentially compares the performance of a predictor g_t, trained on $\{(Z_i, W_i)\}_{i\leq b(t-1)}$ and evaluated on the t-th minibatch, to that of a trivial baseline predictor to form a bet. In particular, setting $b = 1$ yields a valid, yet a powerless test. Indeed, we have $\widehat{\pi}_t = \mathbb{1}\{W_t = 1\} = (W_t + 1)/2$. In this case, the payoff (B.15) reduces to

$$\frac{W_t\left(g_t(Z_t) - (2\widehat{\pi}_t - 1)\right)}{1 + W_t(2\widehat{\pi}_t - 1)} = \frac{W_t g_t(Z_t) - 1}{2} \overset{\text{a.s.}}{\in} [-1, 0],$$

implying that the wealth can not grow even if the null is false. Define a wealth processes $(\mathcal{K}_t^{\mathrm{u}})_{t\geq 0}$ based on the payoff functions (B.15) along with a predictable sequence of betting fractions $(\lambda_t)_{t\geq 1}$ selected via ONS strategy (Algorithm 5). Let $\mathcal{F}_t = \sigma(\{(Z_i, W_i)\}_{i\leq bt})$ for $t \geq 1$, with $\mathcal{F}_0$ denoting a trivial sigma-algebra. We conclude with the following result, whose proof is deferred to Appendix B.4.5.

Theorem B.3. *Suppose that H_0 in* (B.11a) *is true. Then $(\mathcal{K}_t^{\mathrm{u}})_{t\geq 0}$ is a nonnegative supermartingale adapted to $(\mathcal{F}_t)_{t\geq 0}$. Hence, the sequential 2ST based on $(\mathcal{K}_t^{\mathrm{u}})_{t\geq 0}$ satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

B.3 Testing under Distribution Drift

First, we define the problem of two-sample testing when at each round instances from both distributions are observed.

Definition 10 (Sequential two-sample testing)**.** *Suppose that we observe that a stream of observations: $((X_t, Y_t))_{t\geq 1}$, where $(X_t, Y_t) \overset{\text{iid}}{\sim} P_X \times P_Y$ for $t \geq 1$. The goal is to design a sequential test for*

$$H_0 : (X_t, Y_t) \overset{\text{iid}}{\sim} P_X \times P_Y \text{ and } P_X = P_Y, \tag{B.16a}$$

$$H_1 : (X_t, Y_t) \overset{\text{iid}}{\sim} P_X \times P_Y \text{ and } P_X \neq P_Y. \tag{B.16b}$$

Under the two-sample testing setting (Definition 10), we label observations from P_Y as positive $(+1)$ and observations from P_X as negative (-1). We write $\mathcal{A}_{\mathrm{c}}^{\mathrm{2ST}} : (\cup_{t\geq 1}(\mathcal{X} \times \{-1, +1\})^t) \times \mathcal{G} \to \mathcal{G}$ to denote a chosen

learning algorithm which maps a training dataset of any size and previously used predictor, to an updated predictor. We start with $\mathcal{D}_0 = \emptyset$ and $g_1 : g_1(x) = 0, \forall x \in \mathcal{X}$. At round t, we bet using derandomized versions of the payoffs (3.22), namely

$$f_t^{\mathrm{m}}(X_t, Y_t) = \tfrac{1}{2}\left(\operatorname{sign}\left[g_t(Y_t)\right] - \operatorname{sign}\left[g_t(X_t)\right]\right), \tag{B.17a}$$

$$f_t^{\mathrm{s}}(X_t, Y_t) = \tfrac{1}{2}\left(g_t(Y_t) - g_t(X_t)\right). \tag{B.17b}$$

After (X_t, Y_t) has been used for betting, we update a training dataset and an existing predictor:

$$\mathcal{D}_t = \mathcal{D}_{t-1} \cup \{(Y_t, +1), (X_t, -1)\}, \quad g_{t+1} = \mathcal{A}_{\mathrm{c}}^{2\mathrm{ST}}(\mathcal{D}_t, g_t).$$

Testing under Distribution Drift. Batch two-sample and independence tests generally rely on either a cutoff computed using the asymptotic null distribution of a chosen test statistic (if tractable) or a permutation p-value. Both approaches require imposing i.i.d. (or exchangeability, for the latter option) assumption about the data distribution, and if the distribution drifts, both approaches fail to guarantee the type I error control. In contrast, Seq-C-2ST and Seq-C-IT remain valid beyond the i.i.d. setting by construction (analogous to tests developed in (Shekhar and Ramdas, 2021; Podkopaev et al., 2023)). First, we define the problems of sequential two-sample and independence testing under distribution drift.

Definition 11 (Sequential two-sample testing under distribution drift)**.** *Suppose that we observe that a stream of independent observations:* $((X_t, Y_t))_{t\geq 1}$*, where* $(X_t, Y_t) \sim P_X^{(t)} \times P_Y^{(t)}$*,* $t \geq 1$*. The goal is to design a sequential test for the following pair of hypotheses:*

$$H_0 : P_X^{(t)} = P_Y^{(t)}, \ \forall t, \tag{B.18a}$$

$$H_1 : \exists t' : P_X^{(t')} \neq P_Y^{(t')}. \tag{B.18b}$$

Definition 12 (Sequential independence testing under distribution drift)**.** *Suppose that we observe that a stream of independent observations from the joint distribution which drifts over time:* $((X_t, Y_t))_{t\geq 1}$*, where* $(X_t, Y_t) \sim P_{XY}^{(t)}$*. The goal is to design a sequential test for the following pair of hypotheses:*

$$H_0 : P_{XY}^{(t)} = P_X^{(t)} \times P_Y^{(t)}, \ \forall t, \tag{B.19a}$$

$$H_1 : \exists t' : P_{XY}^{(t')} \neq P_X^{(t')} \times P_Y^{(t')}. \tag{B.19b}$$

The superscripts highlight that, in contrast to the standard i.i.d. setting (Definitions 10 and 2), the underlying distributions may drift over time. For independence testing, we need to impose an additional assumption that enables reasoning about the type I error control of Seq-C-IT.

Assumption 7. *Consider the setting of independence testing under distribution drift (Definition 12). We assume that for each $t \geq 1$, it holds that either $P_X^{(t-1)} = P_X^{(t)}$ or $P_Y^{(t-1)} = P_Y^{(t)}$, meaning that at each step either the distribution of X changes or that of Y changes, but not both simultaneously*[†].

We have the following result about the type I error control of our tests under distribution drift.

Corollary B.3.1. *The following claims hold:*

1. *Suppose that H_0 in* (B.18a) *is true. Then Seq-C-2ST satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*
2. *Suppose that H_0 in* (B.19a) *is true. Further, suppose that Assumption 7 is satisfied. Then Seq-C-IT satisfies: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*

The above result follows from the fact the payoff functions underlying Seq-C-2ST (B.17) and Seq-C-IT (3.23) are valid under the more general null hypotheses (B.18a) and (B.19a) respectively. The rest of the proof of Corollary B.3.1 follows the same steps as that of Theorem 3.2, and we omit the details. We conclude with an example which shows that Assumption 7 is necessary for the type I error control.

Example 5. *Consider the following case when the null H_0 in* (B.19a) *is true, but Assumption 7 is not satisfied. We show that Seq-C-IT fails to control type I error (at any prespecified level $\alpha \in (0,1)$), and for simplicity, focus on the payoff function based on the squared risk* (3.23)*. Suppose that we observe a sequence of observations: $((X_t, Y_t))_{t\geq 1}$, where $(X_t, Y_t) = (t + W_t, t + V_t)$ and $W_t, V_t \overset{\text{iid}}{\sim} \text{Bern}(1/2)$. It suffices to show that there exists a sequence of predictors $(g_t)_{t\geq 1}$, for which*

$$\liminf_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} f_t^{\text{s}}((X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t})) \overset{\text{a.s.}}{>} 0. \tag{B.20}$$

If (B.20) *holds, then using the same argument as in the proof of Theorem 3.2, one can then deduce that $\mathbb{P}(\tau < \infty) = 1$. Consider the following sequence of predictors $(g_t)_{t\geq 1}$:*

$$g_t(x, y) = \left(\left(x - \left(2t - \tfrac{1}{2}\right)\right)\left(y - \left(2t - \tfrac{1}{2}\right)\right) \wedge 1\right) \vee -1.$$

We have:

$$\begin{aligned}
g_t(X_{2t}, Y_{2t}) &= \left(\left(W_{2t} + \tfrac{1}{2}\right)\left(V_{2t} + \tfrac{1}{2}\right) \wedge 1\right) \vee -1, \\
g_t(X_{2t-1}, Y_{2t-1}) &= \left(W_{2t-1} - \tfrac{1}{2}\right)\left(V_{2t-1} - \tfrac{1}{2}\right), \\
g_t(X_{2t}, Y_{2t-1}) &= \left(W_{2t} + \tfrac{1}{2}\right)\left(V_{2t-1} - \tfrac{1}{2}\right), \\
g_t(X_{2t-1}, Y_{2t}) &= \left(W_{2t-1} - \tfrac{1}{2}\right)\left(V_{2t} + \tfrac{1}{2}\right).
\end{aligned}$$

[†]Technically, a slightly weaker condition suffices — at odd t, the distribution can change arbitrarily, but at even t, either the distribution of X changes or that of Y changes but not both; however, this weaker condition is slightly less intuitive than the stated condition.

Simple calculation shows that:

$$\mathbb{E}\left[g_t(X_{2t}, Y_{2t})\right] = 11/16, \quad \mathbb{E}\left[g_t(X_{2t-1}, Y_{2t-1})\right] = \mathbb{E}\left[g_t(X_{2t}, Y_{2t-1})\right] = \mathbb{E}\left[g_t(X_{2t-1}, Y_{2t})\right] = 0$$

and hence, for all $t \geq 1$, it holds that $\mathbb{E}\left[f_t^s((X_{2t-1}, Y_{2t-1}), (X_{2t}, Y_{2t}))\right] = 11/64 > 0$. This in turn implies (B.20)*, and hence, we conclude that Seq-C-IT fails to control the type I error.*

B.4 Proofs

B.4.1 Auxiliary Results

Proposition 8 (Ville's inequality (Ville, 1939))**.** *Suppose that $(\mathcal{M}_t)_{t\geq 0}$ is a nonnegative supermartingale process adapted to a filtration $(\mathcal{F}_t)_{t\geq 0}$. Then, for any $a > 0$ it holds that:*

$$\mathbb{P}\left(\exists t \geq 1 : \mathcal{M}_t \geq a\right) \leq \frac{\mathbb{E}\left[\mathcal{M}_0\right]}{a}.$$

B.4.2 Supporting Lemmas

Lemma B.3.1. *Consider sequential two-sample testing setting (Definition 1). Suppose that a predictor $g \in \mathcal{G}$ satisfies $\mathbb{E}\left[f(Z, W)\right] > 0$, where $f(z, w) := wg(z)$.*

(a) Consider the wealth process $(\mathcal{K}_t)_{t\geq 0}$ based on f along with the ONS strategy for selecting betting fractions (Algorithm 5). Then we have the following lower bound on the growth rate of the wealth process:

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} \overset{\text{a.s.}}{\geq} \frac{1}{4}\left(\frac{\left(\mathbb{E}\left[f(Z, W)\right]\right)^2}{\mathbb{E}\left[f^2(Z, W)\right]} \wedge \mathbb{E}\left[f(Z, W)\right]\right). \tag{B.21}$$

(b) For $\lambda_\star = \arg\max_{\lambda\in[-0.5,0.5]} \mathbb{E}\left[\log(1 + \lambda f(Z, W))\right]$, it holds that:

$$\mathbb{E}\left[\log(1 + \lambda_\star f(Z, W))\right] \leq \frac{4}{3} \cdot \frac{\left(\mathbb{E}\left[f(Z, W)\right]\right)^2}{\mathbb{E}\left[\left(f(Z, W)\right)^2\right]} \wedge \frac{\mathbb{E}\left[f(Z, W)\right]}{2}. \tag{B.22}$$

An analogous result holds when the payoff function $f(z, w) := w \cdot \operatorname{sign}\left[g(z)\right]$ is used instead.

Proof. (a) Under the ONS betting strategy, for any sequence of outcomes $(f_t)_{t\geq 1}$, $f_t \in [-1, 1]$, it holds that (see the proof of Theorem 1 in (Cutkosky and Orabona, 2018)):

$$\log \mathcal{K}_t(\lambda_0) - \log \mathcal{K}_t = O\left(\log\left(\sum_{i=1}^{t} f_i^2\right)\right), \tag{B.23}$$

where $\mathcal{K}_t(\lambda_0)$ is the wealth of any constant betting strategy $\lambda_0 \in [-1/2, 1/2]$ and $\mathcal{K}_t$ is the wealth corresponding to the ONS betting strategy. Hence, it follows that

$$\frac{\log \mathcal{K}_t}{t} \geq \frac{\log \mathcal{K}_t(\lambda_0)}{t} - C \cdot \frac{\log t}{t}, \tag{B.24}$$

for some absolute constant $C > 0$. Next, consider

$$\lambda_0 = \frac{1}{2}\left(\left(\frac{\sum_{i=1}^t f_i}{\sum_{i=1}^t f_i^2} \wedge 1\right) \vee 0\right).$$

We obtain:

$$\begin{aligned}
\frac{\log \mathcal{K}_t(\lambda_0)}{t} &= \frac{1}{t}\sum_{i=1}^t \log(1 + \lambda_0 f_i) \\
&\overset{\text{(a)}}{\geq} \frac{1}{t}\sum_{i=1}^t (\lambda_0 f_i - \lambda_0^2 f_i^2) \\
&= \left(\frac{\frac{1}{t}\sum_{i=1}^t f_i}{4} \vee 0\right) \cdot \left(\frac{\frac{1}{t}\sum_{i=1}^t f_i}{\frac{1}{t}\sum_{i=1}^t f_i^2} \wedge 1\right),
\end{aligned} \tag{B.25}$$

where in (a) we used that $\log(1 + x) \geq x - x^2$ for $x \in [-1/2, 1/2]$. From (B.24), it then follows that:

$$\begin{aligned}
\liminf_{t\to\infty} \frac{\log \mathcal{K}_t}{t} &\overset{\text{a.s.}}{\geq} \left(\frac{\mathbb{E}\left[f(Z,W)\right]}{4} \vee 0\right) \cdot \left(\frac{\mathbb{E}\left[f(Z,W)\right]}{\mathbb{E}\left[f^2(Z,W)\right]} \wedge 1\right) \\
&= \frac{1}{4}\left(\frac{\left(\mathbb{E}\left[f(Z,W)\right]\right)^2}{\mathbb{E}\left[f^2(Z,W)\right]} \wedge \mathbb{E}\left[f(Z,W)\right]\right),
\end{aligned}$$

which completes the proof of the first assertion of the lemma.

(b) Since $\log(1 + x) \leq x - 3x^2/8$ for any $x \in [-0.5, 0.5]$, we know that:

$$\begin{aligned}
\mathbb{E}\left[\log\left(1 + \lambda_\star f(Z,W)\right)\right] &\leq \mathbb{E}\left[\lambda_\star f(Z,W) - \frac{3}{8}\left(\lambda_\star f(Z,W)\right)^2\right] \\
&\leq \max_{\lambda \in [-0.5, 0.5]} \left(\lambda \cdot \mathbb{E}\left[f(Z,W)\right] - \frac{3\lambda^2}{8} \cdot \mathbb{E}\left[\left(f(Z,W)\right)^2\right]\right).
\end{aligned}$$

The optimizer of the above is

$$\tilde{\lambda} = \frac{4\mathbb{E}\left[f(Z,W)\right]}{3\mathbb{E}\left[\left(f(Z,W)\right)^2\right]} \wedge \frac{1}{2}.$$

Hence, as long as $\mathbb{E}\left[f(Z,W)\right] \leq (3/8) \cdot \mathbb{E}\left[\left(f(Z,W)\right)^2\right]$, we have:

$$\mathbb{E}\left[\log\left(1 + \lambda_\star f(Z,W)\right)\right] \leq \frac{2}{3}\frac{\left(\mathbb{E}\left[f(Z,W)\right]\right)^2}{\mathbb{E}\left[\left(f(Z,W)\right)^2\right]}. \tag{B.26}$$

If however, $\mathbb{E}\left[f(Z,W)\right] > (3/8)\cdot\mathbb{E}\left[\left(f(Z,W)\right)^2\right]$, then we know that:

$$\mathbb{E}\left[\log\left(1+\lambda_\star f(Z,W)\right)\right] \leq \frac{\mathbb{E}\left[f(Z,W)\right]}{2}.$$

To bring it to a convenient form, we multiply the upper bound in (B.26) by two and get the bound (B.22), which completes the proof of the second assertion of the lemma.

□

B.4.3 Proofs for Section 3.2

Proposition 3. *Fix an arbitrary predictor $g \in \mathcal{G}$. The following claims hold:*

1. *For the misclassification risk, we have that:*

$$\sup_{s\in[0,1]}\left(\tfrac{1}{2}-R_{\mathrm{m}}(sg)\right) = \left(\tfrac{1}{2}-R_{\mathrm{m}}(g)\right)\vee 0 = \left(\tfrac{1}{2}\cdot\mathbb{E}\left[W\cdot\operatorname{sign}\left[g(Z)\right]\right]\right)\vee 0. \tag{3.9}$$

2. *For the squared risk, we have that:*

$$\sup_{s\in[0,1]}\left(1-R_{\mathrm{s}}(sg)\right) \geq \left(\mathbb{E}\left[W\cdot g(Z)\right]\vee 0\right)\cdot\left(\frac{\mathbb{E}\left[W\cdot g(Z)\right]}{\mathbb{E}\left[g^2(Z)\right]}\wedge 1\right) \tag{3.10}$$

Further, $d_{\mathrm{s}}(P,Q) > 0$ if and only if there exists $g\in\mathcal{G}$ such that $\mathbb{E}\left[W\cdot g(Z)\right] > 0$.

Proof. 1. The first equality in (3.9) follows from two facts: (a) for any $g\in\mathcal{G}$ and any $s\in(0,1]$, it holds that $R_{\mathrm{m}}(sg) = R_{\mathrm{m}}(g)$, (b) $R_{\mathrm{m}}(0) = 1/2$. The second equality easily follows from the following fact: $\operatorname{sign}\left[x\right]/2 = 1/2 - \mathbb{1}\left\{x<0\right\}$.

2. Consider an arbitrary predictor $g\in\mathcal{G}$. Let us consider all possible scenarios:

   (a) If $\mathbb{E}\left[W\cdot g(Z)\right]\leq 0$, then the RHS of (3.10) is zero. For the LHS of (3.10), we have that:

   $$\sup_{s\in[0,1]}\left(1-R_{\mathrm{s}}(sg)\right) \geq 1-R_{\mathrm{s}}(0) = 0,$$

   so the bound (3.10) holds.

   (b) Next, assume that $\mathbb{E}\left[W\cdot g(Z)\right] > 0$, then it is easy to derive that:

   $$s_\star := \operatorname*{arg\,max}_{s\in[0,1]}\left(1-R_{\mathrm{s}}(sg)\right) = \frac{\mathbb{E}\left[W\cdot g(Z)\right]}{\mathbb{E}\left[g^2(Z)\right]}\wedge 1. \tag{B.27}$$

A simple calculation shows that:

$$1 - R_{\mathrm{s}}(s_\star g) \geq \mathbb{E}\left[W \cdot g(Z)\right] \cdot \left(\frac{\mathbb{E}\left[W \cdot g(Z)\right]}{\mathbb{E}\left[g^2(Z)\right]} \wedge 1\right),$$

and hence, we conclude that the bound (3.10) holds.

To establish the second part of the statement, note that $d_{\mathrm{s}}(P, Q) > 0$ iff there is a predictor $g \in \mathcal{G}$ such that $R_{\mathrm{s}}(g) < 1$. For the squared risk, we have:

$$1 - R_{\mathrm{s}}(g) = 2\mathbb{E}\left[W \cdot g(Z)\right] - \mathbb{E}\left[g^2(Z)\right], \tag{B.28}$$

and hence, $R_{\mathrm{s}}(g) < 1$ trivially implies that $\mathbb{E}\left[W \cdot g(Z)\right] > 0$. The converse implication trivially follows from (3.10). Hence, the result follows.

□

Theorem 3.1. *The following claims hold:*

1. *Suppose that H_0 in* (3.1a) *is true. Then the oracle sequential test based on either $(\mathcal{K}_t^{\mathrm{m},\star})_{t\geq 0}$ or $(\mathcal{K}_t^{\mathrm{s},\star})_{t\geq 0}$ ever stops with probability at most α: $\mathbb{P}_{H_0}(\tau < \infty) \leq \alpha$.*
2. *Suppose that H_1 in* (3.1b) *is true. Then:*
   (a) *The growth rate of the oracle wealth process $(\mathcal{K}_t^{\mathrm{m},\star})_{t\geq 0}$ satisfies:*

$$\liminf_{t\to\infty}\left(\tfrac{1}{t}\log \mathcal{K}_t^{\mathrm{m},\star}\right) \overset{\text{a.s.}}{\geq} \left(\tfrac{1}{2} - R_{\mathrm{m}}\left(g_\star\right)\right)^2. \tag{3.14}$$

   If $R_{\mathrm{m}}\left(g_\star\right) < 1/2$, then the test based on $(\mathcal{K}_t^{\mathrm{m},\star})_{t\geq 0}$ is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$. Further, the optimal growth rate achieved by $\lambda_\star^{\mathrm{m}}$ in (3.13) *satisfies:*

$$\mathbb{E}\left[\log(1 + \lambda_\star^{\mathrm{m}} f_\star^{\mathrm{m}}(Z, W))\right] \leq \left(\tfrac{16}{3} \cdot \left(\tfrac{1}{2} - R_{\mathrm{m}}(g_\star)\right)^2 \wedge \left(\tfrac{1}{2} - R_{\mathrm{m}}(g_\star)\right)\right). \tag{3.15}$$

   (b) *The growth rate of the oracle wealth process $(\mathcal{K}_t^{\mathrm{s},\star})_{t\geq 0}$ satisfies:*

$$\liminf_{t\to\infty}\left(\tfrac{1}{t}\log \mathcal{K}_t^{\mathrm{s},\star}\right) \overset{\text{a.s.}}{\geq} \tfrac{1}{4} \cdot \mathbb{E}\left[W \cdot g_\star(Z)\right]. \tag{3.16}$$

   If $\mathbb{E}\left[W \cdot g_\star(Z)\right] > 0$, then the test based on $(\mathcal{K}_t^{\mathrm{s},\star})_{t\geq 0}$ is consistent: $\mathbb{P}_{H_1}(\tau < \infty) = 1$. Further, the optimal growth rate achieved by $\lambda_\star^{\mathrm{s}}$ in (3.13) *satisfies:*

$$\mathbb{E}\left[\log(1 + \lambda_\star^{\mathrm{s}} f_\star^{\mathrm{s}}(Z, W))\right] \leq \tfrac{1}{2} \cdot \mathbb{E}\left[W \cdot g_\star(Z)\right]. \tag{3.17}$$

Proof. 1. We trivially have that the payoff functions (3.11a) and (3.11b) are bounded: $\forall(z,w) \in \mathcal{Z} \times \{-1,1\}$, it holds that $f_\star^{\mathrm{m}}(z,w) \in [-1,1]$ and $f_\star^{\mathrm{s}}(z,w) \in [-1,1]$. Further, under the null H_0 in (3.1a), it trivially holds that $\mathbb{E}_{H_0}\left[f_\star^{\mathrm{m}}(Z_t,W_t) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{H_0}\left[f_\star^{\mathrm{s}}(Z_t,W_t) \mid \mathcal{F}_{t-1}\right] = 0$, where $\mathcal{F}_{t-1} = \sigma(\{(Z_i,W_i)\}_{i\le t-1})$. Since ONS betting fractions $\left(\lambda_t^{\mathrm{ONS}}\right)_{t\ge 1}$ are predictable, we conclude that the resulting wealth process is a nonnegative martingale. The assertion of the Theorem then follows directly from Ville's inequality (Proposition 8) when $a = 1/\alpha$.

2. Suppose that H_1 in (3.1b) is true. First, we prove the results for the lower bounds:

   (a) Consider the wealth process based on the misclassification risk $(\mathcal{K}_t^{\mathrm{m},\star})_{t\ge 0}$. Note that for all $t \ge 1$:

   $$\mathbb{E}\left[f_\star^{\mathrm{m}}(Z_t,W_t)\right] = 2\cdot\left(\frac{1}{2} - R_{\mathrm{m}}\left(g_\star\right)\right), \quad \left(f_\star^{\mathrm{m}}(Z_t,W_t)\right)^2 = 1.$$

   Since $\mathbb{E}\left[f_\star^{\mathrm{m}}(Z_t,W_t)\right] \in [0,1]$, we also have $\left(\mathbb{E}\left[f_\star^{\mathrm{m}}(Z_t,W_t)\right]\right)^2 \le \mathbb{E}\left[f_\star^{\mathrm{m}}(Z_t,W_t)\right]$. From the first part of Lemma B.3.1, it follows that:

   $$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t^{\mathrm{m},\star}}{t} \overset{\text{a.s.}}{\ge} \frac{1}{4}\left(\mathbb{E}\left[f_\star^{\mathrm{m}}(Z_t,W_t)\right]\right)^2 = \left(\frac{1}{2} - R_{\mathrm{m}}\left(g_\star\right)\right)^2.$$

   From the second part of Lemma B.3.1, and (B.22) in particular, it follows that:

   $$\mathbb{E}\left[\log\left(1+\lambda_\star^{\mathrm{m}} f_\star^{\mathrm{m}}(Z,W)\right)\right] \le \left(\frac{16}{3}\cdot\left(\frac{1}{2} - R_{\mathrm{m}}(g_\star)\right)^2 \wedge \left(\frac{1}{2} - R_{\mathrm{m}}(g_\star)\right)\right).$$

   The first term in the above is smaller or equal than the second one whenever $R_{\mathrm{m}}(g_\star) \ge 5/16$. We conclude that the assertion of the theorem is true.

   (b) Next, we consider the wealth process based on the squared error: $(\mathcal{K}_t^{\mathrm{s},\star})_{t\ge 0}$. Note that:

   $$\begin{aligned}\mathbb{E}\left[f_\star^{\mathrm{s}}(Z_t,W_t)\right] &= \mathbb{E}\left[W\cdot g_\star(Z)\right],\\ \mathbb{E}\left[\left(f_\star^{\mathrm{s}}(Z_t,W_t)\right)^2\right] &= \mathbb{E}\left[g_\star^2(Z)\right],\end{aligned}$$

   and hence from Lemma B.3.1, it follows that:

   $$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t^{\mathrm{s},\star}}{t} \overset{\text{a.s.}}{\ge} \frac{1}{4}\left(\frac{\left(\mathbb{E}\left[W\cdot g_\star(Z)\right]\right)^2}{\mathbb{E}\left[g_\star^2(Z)\right]} \wedge \mathbb{E}\left[W\cdot g_\star(Z)\right]\right). \tag{B.29}$$

   In the above, we assume that the following case is not possible: $g_\star(Z) \overset{\text{a.s.}}{=} 0$ (for such $g_\star$, the corresponding expected margin and the growth rate of the resulting wealth process are clearly zero, and will still be

highlighted in our resulting bound). Next, note that since $g_\star \in \arg\min_{g\in\mathcal{G}} R_{\mathrm{s}}(g)$, we have that:

$$1 - R_{\mathrm{s}}(g_\star) = \sup_{s\in[0,1]} \left(1 - R_{\mathrm{s}}(sg_\star)\right),$$

meaning that $g_\star$ can not be improved by scaling with $s < 1$. From Proposition 3, and (B.27) in particular, it follows that:

$$\frac{\mathbb{E}\left[W \cdot g_\star(Z)\right]}{\mathbb{E}\left[g_\star^2(Z)\right]} \geq 1, \tag{B.30}$$

and hence, the bound (B.29) reduces to

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t^{\mathrm{s},\star}}{t} \overset{\text{a.s.}}{\geq} \frac{\mathbb{E}\left[W \cdot g_\star(Z)\right]}{4}.$$

From the second part of Lemma B.3.1, it follows that:

$$\mathbb{E}\left[\log\left(1 + \lambda_\star^{\mathrm{s}} f_\star^{\mathrm{s}}(Z,W)\right)\right] \leq \frac{4}{3}\frac{\left(\mathbb{E}\left[W \cdot g_\star(Z)\right]\right)^2}{\mathbb{E}\left[\left(g_\star(Z)\right)^2\right]} \wedge \frac{\mathbb{E}\left[W \cdot g_\star(Z)\right]}{2}. \tag{B.31}$$

Next, we use that $g_\star$ satisfies (B.30), which implies that the second term in (B.31) is smaller, and hence,

$$\mathbb{E}\left[\log\left(1 + \lambda_\star^{\mathrm{s}} f_\star^{\mathrm{s}}(Z,W)\right)\right] \leq \frac{\mathbb{E}\left[W \cdot g_\star(Z)\right]}{2},$$

which concludes the proof of the second part of the theorem.

□

Corollary 3.1.1. *Consider an arbitrary $g \in \mathcal{G}$ with nonnegative expected margin: $\mathbb{E}\left[W \cdot g(Z)\right] \geq 0$. Then the growth rate of the corresponding wealth process $(\mathcal{K}_t^{\mathrm{s}})_{t\geq 0}$ satisfies:*

$$\liminf_{t\to\infty} \left(\tfrac{1}{t}\log \mathcal{K}_t^{\mathrm{s}}\right) \overset{\text{a.s.}}{\geq} \tfrac{1}{4}\Big(\sup_{s\in[0,1]} \left(1 - R_{\mathrm{s}}(sg)\right) \wedge \mathbb{E}\left[W \cdot g(Z)\right]\Big) \tag{3.18a}$$

$$\geq \tfrac{1}{4}\left(\mathbb{E}\left[W \cdot g(Z)\right]\right)^2, \tag{3.18b}$$

and the optimal growth rate achieved by $\lambda_\star^{\mathrm{s}}$ in (3.13) *satisfies:*

$$\mathbb{E}\left[\log(1 + \lambda_\star^{\mathrm{s}} f^{\mathrm{s}}(Z,W))\right] \leq \Big(\tfrac{4}{3} \cdot \sup_{s\in[0,1]} \left(1 - R_{\mathrm{s}}(sg)\right)\Big) \wedge \left(\tfrac{1}{2} \cdot \mathbb{E}\left[W \cdot g(Z)\right]\right). \tag{3.19}$$

Proof. Following the same argument as that of the proof of Theorem 3.1, we can deduce that:

$$\liminf_{t\to\infty} \frac{\log \mathcal{K}_t^{\mathrm{s}}}{t} \overset{\text{a.s.}}{\geq} \frac{1}{4}\left(\frac{\left(\mathbb{E}\left[W \cdot g(Z)\right]\right)^2}{\mathbb{E}\left[g^2(Z)\right]} \wedge \mathbb{E}\left[W \cdot g(Z)\right]\right). \tag{B.32}$$

Hence, it suffices to argue that the lower bound (B.32) is equivalent to (3.18a). Without loss of generality, we can assume that $\mathbb{E}\left[W \cdot g(Z)\right] \geq 0$, and further, the two lower bounds are equal if $\mathbb{E}\left[W \cdot g(Z)\right] = 0$. Hence, we consider the case when $\mathbb{E}\left[W \cdot g(Z)\right] > 0$. First, let us consider the case when

$$\frac{\mathbb{E}\left[W \cdot g(Z)\right]}{\mathbb{E}\left[g^2(Z)\right]} < 1. \tag{B.33}$$

Using (B.27), we get that:

$$\sup_{s \in [0,1]} \left(1 - R_{\mathrm{s}}\left(sg\right)\right) = \frac{\left(\mathbb{E}\left[W \cdot g(Z)\right]\right)^2}{\mathbb{E}\left[g^2(Z)\right]}, \tag{B.34}$$

and hence, two bounds coincide. For the upper bound (3.19), we use Lemma B.3.1, and the upper bound (B.22) in particular. Note that the first term in (B.22) is less than the second term whenever

$$\frac{\mathbb{E}\left[W \cdot g(Z)\right]}{\mathbb{E}\left[\left(g(Z)\right)^2\right]} \leq \frac{3}{8} < 1.$$

However, in this regime we also know that (B.34) holds, and hence the two bounds coincide. This completes the proof. □

Theorem 3.2. *The following claims hold for Seq-C-2ST (Algorithm 6):*

1. *If H_0 in* (3.1a) *is true, the test ever stops with probability at most α: $\mathbb{P}_{H_0}\left(\tau < \infty\right) \leq \alpha$.*
2. *Suppose that H_1 in* (3.1b) *is true. Then:*
   (a) *Under Assumption 3, the test with the payoff* (3.22a) *is consistent: $\mathbb{P}_{H_1}\left(\tau < \infty\right) = 1$.*
   (b) *Under Assumption 4, the test with the payoff* (3.22b) *is consistent: $\mathbb{P}_{H_1}\left(\tau < \infty\right) = 1$.*

Proof. 1. We trivially have that the payoff functions (3.22a) and (3.22b) are bounded: $\forall t \geq 1$ and $\forall (z, w) \in \mathcal{Z} \times \{-1, 1\}$, it holds that $f_t^{\mathrm{m}}(z, w) \in [-1, 1]$ and $f_t^{\mathrm{s}}(z, w) \in [-1, 1]$. Further, under the null H_0 in (3.1a), it trivially holds that $\mathbb{E}_{H_0}\left[f_t^{\mathrm{m}}(Z_t, W_t) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{H_0}\left[f_t^{\mathrm{s}}(Z_t, W_t) \mid \mathcal{F}_{t-1}\right] = 0$, where $\mathcal{F}_{t-1} = \sigma(\{(Z_i, W_i)\}_{i \leq t-1})$. Since ONS betting fractions $\left(\lambda_t^{\mathrm{ONS}}\right)_{t \geq 1}$ are predictable, we conclude that the resulting wealth process is a nonnegative martingale. The assertion of the Theorem then follows directly from Ville's inequality (Proposition 8) when $a = 1/\alpha$.

2. Note that if ONS strategy for selecting betting fractions is deployed, then (B.25) implies that the tests will be consistent as long as

$$\liminf_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} f_i \overset{\text{a.s.}}{>} 0, \tag{B.35}$$

where for $i \geq 1$, $f_i = f_i^{\mathrm{m}}(Z_i, W_i)$ and $f_i = f_i^{\mathrm{s}}(Z_i, W_i)$ for the payoffs based on the misclassification and the squared risks respectively.

(a) Recall that

$$f_i^{\mathrm{m}}(Z_i, W_i) = W_i \cdot \operatorname{sign}\left[g_i(Z_i)\right],$$

and Assumption 3 states that:

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \mathbb{1}\left\{W_i \cdot \operatorname{sign}\left[g_i(Z_i)\right] < 0\right\} \overset{\text{a.s.}}{<} \frac{1}{2}.$$

Since $\mathbb{1}\{x < 0\} = (1 - \operatorname{sign}[x])/2$, we get that:

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \left(\frac{1}{2} - \frac{W_i \cdot \operatorname{sign}\left[g_i(Z_i)\right]}{2}\right) \overset{\text{a.s.}}{<} \frac{1}{2},$$

which, after rearranging and multiplying by two, implies that:

$$\liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} W_i \cdot \operatorname{sign}\left[g_i(Z_i)\right] \overset{\text{a.s.}}{>} 0.$$

Hence, a sufficient condition for consistency (B.35) holds, and we conclude that the result is true.

(b) Recall that

$$f_i^{\mathrm{s}}(Z_i, W_i) = W_i \cdot g_i(Z_i),$$

and Assumption 4 states that:

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \left(g_i(Z_i) - W_i\right)^2 \overset{\text{a.s}}{<} 1,$$

which is equivalent to

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \left(g_i^2(Z_i) - 2W_i \cdot g_i(Z_i)\right) \overset{\text{a.s}}{<} 0.$$

It is easy to see that the above, in turn, implies that:

$$\liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} W_i \cdot g_i(Z_i) \overset{\text{a.s}}{>} 0.$$

Hence, a sufficient condition for consistency (B.35) holds, and we conclude that the result is true.

□

B.4.4 Proofs for Appendix B.1

Theorem B.1. *The following claims hold for the oracle sequential regression-based IT based on* $\left(\mathcal{K}_t^{\mathrm{r},\star}\right)_{t\geq 0}$*:*

1. *Suppose that* H_0 *in* (B.2a) *is true. Then the test ever stops with probability at most* α*:* $\mathbb{P}_{H_1}(\tau < \infty) \leq \alpha$*.*

2. *Suppose that H_1 in* (B.2b) *is true. Further, suppose that:* $\mathbb{E}\left[W\ell(g_\star(X),Y)\right] > 0$. *Then the test is consistent:* $\mathbb{P}_{H_1}\left(\tau < \infty\right) = 1$.

Proof. 1. We trivially have that the payoff function (B.3) is bounded: $\forall (x,y,w) \in \mathcal{X}\times\mathcal{Y}\times\{-1,1\}$, it holds that $f^{\mathrm{r}}_\star(x,y,w) \in [-1,1]$. Further, under the null H_0 in (B.2a), it trivially holds that $\mathbb{E}_{H_0}\left[f^{\mathrm{r}}_\star(X_t,Y_t,W_t) \mid \mathcal{F}_{t-1}\right] = 0$, where $\mathcal{F}_{t-1} = \sigma(\{(X_i,Y_i,W_i)\}_{i\le t-1})$. Since ONS betting fractions $\left(\lambda^{\mathrm{ONS}}_t\right)_{t\ge 1}$ are predictable, we conclude that the resulting wealth process is a nonnegative martingale. The assertion of the Theorem then follows directly from Ville's inequality (Proposition 8) when $a = 1/\alpha$.

2. Note that if ONS strategy for selecting betting fractions is deployed, then (B.25) implies that the tests will be consistent as long as

$$\liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} f^{\mathrm{r}}_\star(X_i,Y_i,W_i) \overset{\text{a.s.}}{>} 0. \tag{B.36}$$

Note that:

$$\frac{1}{t}\sum_{i=1}^{t} f^{\mathrm{r}}_\star(X_i,Y_i,W_i) = \frac{1}{t}\sum_{i=1}^{t} \tanh\left(s_\star \cdot W_i\ell(g_\star(X_i),Y_i)\right) \overset{\text{a.s.}}{\to} \mathbb{E}\left[\tanh\left(s_\star\cdot W\ell(g_\star(X),Y)\right)\right].$$

Note that for any $x\in\mathbb{R}: \tanh(x) \ge x - \frac{1}{3}\cdot\max\left\{x^3,0\right\}$. Hence, for any $s>0$, it holds that:

$$\begin{aligned}
\mathbb{E}\left[\tanh\left(s\cdot W\ell(g_\star(X),Y)\right)\right] &\ge s\mathbb{E}\left[W\ell(g_\star(X),Y)\right] - \frac{1}{3}\mathbb{E}\left[\max\left\{s^3\cdot W(\ell(g_\star(X),Y))^3,0\right\}\right] \\
&= s\mathbb{E}\left[W\ell(g_\star(X),Y)\right] - \frac{s^3}{3}\mathbb{E}\left[(\ell(g_\star(X),Y))^3\cdot\max\left\{W,0\right\}\right] \\
&= s\mathbb{E}\left[W\ell(g_\star(X),Y)\right] - \frac{s^3}{6}\mathbb{E}\left[(1+W)\cdot(\ell(g_\star(X),Y))^3\right],
\end{aligned} \tag{B.37}$$

where we used that $\max\{W,0\} = (W+1)/2$ since $W\in\{-1,1\}$. Maximizing the RHS of (B.37) over $s>0$ yields $s_\star$ defined in (B.4a). Hence,

$$\begin{aligned}
\mathbb{E}\left[\tanh\left(s_\star\cdot W\ell(g_\star(X),Y)\right)\right] &\ge s_\star\mathbb{E}\left[W\ell(g_\star(X),Y)\right] - \frac{s_\star^3}{6}\mathbb{E}\left[(1+W)\cdot(\ell(g_\star(X),Y))^3\right] \\
&= s_\star\left(\mathbb{E}\left[W\ell(g_\star(X),Y)\right] - \frac{s_\star^2}{6}\mathbb{E}\left[(1+W)\cdot(\ell(g_\star(X),Y))^3\right]\right) \\
&= s_\star\left(\mathbb{E}\left[W\ell(g_\star(X),Y)\right] - \frac{1}{3}\mathbb{E}\left[W\ell(g_\star(X),Y)\right]\right) \\
&= \frac{2s_\star}{3}\mathbb{E}\left[W\ell(g_\star(X),Y)\right] > 0.
\end{aligned}$$

Hence, we conclude that the oracle regression-based IT is consistent since the sufficient condition (B.38) holds. □

Theorem B.2. *The following claims hold for the proxy sequential regression-based IT (Algorithm 9):*

1. *Suppose that H_0 in* (B.2a) *is true. Then the test ever stops with probability at most α:* $\mathbb{P}_{H_0}\left(\tau<\infty\right) \le \alpha$.

2. *Suppose that H_1 in* (B.2b) *is true. Further, suppose that Assumptions 5 and 6 are satisfied. Then the test is consistent:* $\mathbb{P}_{H_1}(\tau < \infty) = 1$.

Proof. 1. We trivially have that the payoff function (B.5) is bounded: $\forall (x, y, w) \in \mathcal{X} \times \mathcal{Y} \times \{-1, 1\}$, it holds that $f_t^{\mathrm{r}}(x, y, w) \in [-1, 1]$. Further, under the null H_0 in (B.2a), it trivially holds that $\mathbb{E}_{H_0}\left[f_t^{\mathrm{r}}(X_t, Y_t, W_t) \mid \mathcal{F}_{t-1}\right] = 0$, where $\mathcal{F}_{t-1} = \sigma(\{(X_i, Y_i, W_i)\}_{i \leq t-1})$. Since ONS betting fractions $(\lambda_t^{\mathrm{ONS}})_{t \geq 1}$ are predictable, we conclude that the resulting wealth process is a nonnegative martingale. The assertion of the Theorem then follows directly from Ville's inequality (Proposition 8) with $a = 1/\alpha$.

2. Note that if ONS strategy for selecting betting fractions is deployed, then (B.25) implies that the tests will be consistent as long as

$$\liminf_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} f_t^{\mathrm{r}}(X_i, Y_i, W_i) \overset{\text{a.s.}}{>} 0. \tag{B.38}$$

   (a) **Step 1.** Consider a predictable sequence of scaling factors $(s_t)_{t \geq 1}$, defined in (B.6a), and the corresponding sequences $(\mu_t)_{t \geq 1}$ and $(\nu_t)_{t \geq 1}$, defined in (B.6b) and (B.6c) respectively. For $t \geq 1$, let $\mathcal{F}_t := \sigma(\{(X_i, Y_i, W_i)\}_{i \leq t})$. Since the losses are bounded, we have that:

$$\left(W_i \cdot \ell(g(X_i; \theta_i), Y_i) - \mathbb{E}\left[W_i \cdot \ell(g(X_i; \theta_i), Y_i) \mid \mathcal{F}_{i-1}\right]\right)_{i \geq 1},$$

   is a bounded martingale difference sequence (BMDS). By the Strong Law of Large Numbers for BMDS, it follows that:

$$\frac{1}{t} \sum_{i=1}^{t} \left(W_i \cdot \ell(g(X_i; \theta_i), Y_i) - \mathbb{E}\left[W_i \cdot \ell(g(X_i; \theta_i), Y_i) \mid \mathcal{F}_{i-1}\right]\right) \overset{\text{a.s.}}{\to} 0.$$

   Since $((X_t, Y_t, W_t))_{t \geq 1}$ is a sequence of i.i.d. observations, we can write

$$\frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[W_i \cdot \ell(g(X_i; \theta_i), Y_i) \mid \mathcal{F}_{i-1}\right] = \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[W \cdot \ell(g(X; \theta_i), Y) \mid \theta_i\right],$$

where $(X, Y, W) \perp\!\!\!\perp (\theta_t)_{t\geq 1}, \theta_\star$. Using Assumption 5, we get that:

$$\begin{aligned} & \left| \frac{1}{t} \sum_{i=1}^{t} \mathbb{E}\left[W \cdot \ell(g(X;\theta_i), Y) \mid \theta_i\right] - \mathbb{E}\left[W \cdot \ell(g(X;\theta_\star), Y) \mid \theta_\star\right] \right| \\ \leq & \frac{1}{t} \sum_{i=1}^{t} \sup_{\substack{x\in\mathcal{X} \\ y\in\mathcal{Y}}} |\ell(g(x;\theta_i), y) - \ell(g(x;\theta_\star), y)| \\ \leq & \frac{1}{t} \sum_{i=1}^{t} L_2 \sup_{x\in\mathcal{X}} |g(x;\theta_i) - g(x;\theta_\star)| \\ \leq & \frac{1}{t} \sum_{i=1}^{t} L_2 \cdot L_1 \cdot \|\theta_i - \theta_\star\| \overset{\text{a.s.}}{\to} 0, \end{aligned} \tag{B.39}$$

since $\|\theta_i - \theta_\star\| \overset{\text{a.s.}}{\to} 0$ by Assumption 6. In particular, this implies that $\mu_t \overset{\text{a.s.}}{\to} \mathbb{E}\left[W\ell(g(X;\theta_\star), Y) \mid \theta_\star\right]$. Similar argument can be used to show that $\nu_t \overset{\text{a.s.}}{\to} \mathbb{E}\left[(1+W) \cdot (\ell(g(X;\theta_\star), Y))^3 \mid \theta_\star\right]$, and hence,

$$s_t \overset{\text{a.s.}}{\to} \sqrt{\frac{2\mathbb{E}\left[W\ell(g(X;\theta_\star), Y) \mid \theta_\star\right]}{\mathbb{E}\left[(1+W) \cdot (\ell(g(X;\theta_\star), Y))^3 \mid \theta_\star\right]}} =: s_\star. \tag{B.40}$$

Note that $s_\star$ is a random variable which is positive (almost surely) by Assumption 6.

(b) **Step 2.** Recall that for any $x \in \mathbb{R}$: $\tanh(x) \geq x - \frac{1}{3} \cdot \max\left\{x^3, 0\right\}$ and that $\max\{W, 0\} = (W+1)/2$ since $W \in \{-1, 1\}$. We have:

$$\begin{aligned} \frac{1}{t} \sum_{i=1}^{t} f_i^{\mathrm{r}}(X_i, Y_i, W_i) &= \frac{1}{t} \sum_{i=1}^{t} \tanh\left(s_i \cdot W_i \ell(g(X_i;\theta_i), Y_i)\right) \\ &\geq \frac{1}{t} \sum_{i=1}^{t} \left(s_i \cdot W_i \cdot \ell(g(X_i;\theta_i), Y_i) - \frac{s_i^3}{6} \cdot (1+W_i) \cdot (\ell(g(X_i;\theta_i), Y_i))^3 \right). \end{aligned}$$

Note that θ_i and s_i are $\mathcal{F}_{i-1}$-measurable (see Step 1 for the definition of $\mathcal{F}_{i-1}$). Under a minor technical assumption that $(s_t)_{t\geq 1}$ is a sequence of bounded scaling factors (the lower bound is trivially zero and the upper bound also holds if ν_t are bounded away from zero almost surely which is reasonable given the

definition of ν_t), we can use analogous argument regarding a BMDS in Step 1 to deduce that:

$$\begin{aligned} & \liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} f_i^{\mathrm{r}}(X_i, Y_i, W_i) \\ \geq \quad & \liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \left(s_i \cdot \mathbb{E}\left[W \cdot \ell(g(X;\theta_i), Y) \mid \theta_i\right] - \frac{s_i^3}{6}\mathbb{E}\left[(1+W)\cdot(\ell(g(X;\theta_i), Y))^3 \mid \theta_i\right]\right). \end{aligned} \tag{B.41}$$

Using argument analogous to (B.39), we can show that:

$$\frac{1}{t}\sum_{i=1}^{t} \mathbb{E}\left[(1+W)\cdot(\ell(g(X;\theta_i), Y))^3 \mid \theta_i\right] \overset{\text{a.s.}}{\to} \mathbb{E}\left[(1+W)\cdot(\ell(g(X;\theta_\star), Y))^3 \mid \theta_\star\right]. \tag{B.42}$$

Combining (B.39), (B.40) and (B.42), we deduce that

$$\begin{aligned} & \frac{1}{t}\sum_{i=1}^{t} \left(s_i \cdot \mathbb{E}\left[W \cdot \ell(g(X;\theta_i), Y) \mid \theta_i\right] - \frac{s_i^3}{6}\mathbb{E}\left[(1+W)\cdot(\ell(g(X;\theta_i), Y))^3 \mid \theta_i\right]\right) \\ \overset{\text{a.s.}}{\to} \quad & s_\star \cdot \mathbb{E}\left[W \cdot \ell(g(X;\theta_\star), Y) \mid \theta_\star\right] - \frac{s_\star^3}{6} \cdot \mathbb{E}\left[(1+W)\cdot(\ell(g(X;\theta_\star), Y))^3 \mid \theta_\star\right] \\ = \quad & \frac{2s_\star}{3} \cdot \mathbb{E}\left[W \cdot \ell(g(X;\theta_\star), Y) \mid \theta_\star\right]. \end{aligned}$$

Hence, from (B.41) it follows that:

$$\liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} f_i^{\mathrm{r}}(X_i, Y_i, W_i) \geq \frac{2s_\star}{3} \cdot \mathbb{E}\left[W \cdot \ell(g(X;\theta_\star), Y) \mid \theta_\star\right],$$

where the RHS is a random variable which is positive almost surely. Hence, a sufficient condition for consistency (B.38) holds which concludes the proof.

□

B.4.5 Proofs for Appendix B.2

Two-Sample Testing with Unbalanced Classes. Note that ($g(z) = 2\eta(z) - 1$):

$$\begin{aligned}
& (1-\lambda_t)\cdot 1 + \lambda_t \cdot \frac{(\eta(Z_t))^{\mathbb{1}\{W_t=1\}}\,(1-\eta(Z_t))^{1-\mathbb{1}\{W_t=1\}}}{(\pi)^{\mathbb{1}\{W_t=1\}}\,(1-\pi)^{1-\mathbb{1}\{W_t=1\}}} \\
= \;& (1-\lambda_t)\cdot 1 + \lambda_t \cdot \frac{\left(\frac{1+g(Z_t)}{2}\right)^{\mathbb{1}\{W_t=1\}}\left(\frac{1-g(Z_t)}{2}\right)^{1-\mathbb{1}\{W_t=1\}}}{(\pi)^{\mathbb{1}\{W_t=1\}}\,(1-\pi)^{1-\mathbb{1}\{W_t=1\}}} \\
= \;& (1-\lambda_t)\cdot 1 + \frac{\lambda_t}{2} \cdot \frac{(1+g(Z_t))^{\mathbb{1}\{W_t=1\}}\,(1-g(Z_t))^{1-\mathbb{1}\{W_t=1\}}}{(\pi)^{\mathbb{1}\{W_t=1\}}\,(1-\pi)^{1-\mathbb{1}\{W_t=1\}}} \\
= \;& (1-\lambda_t)\cdot 1 + \frac{\lambda_t}{2} \cdot \frac{1+W_t g(Z_t)}{(\pi)^{\mathbb{1}\{W_t=1\}}\,(1-\pi)^{1-\mathbb{1}\{W_t=1\}}} \\
= \;& (1-\lambda_t)\cdot 1 + \frac{\lambda_t}{2} \cdot \frac{2}{1+W_t(2\pi-1)} \cdot (1+W_t g(Z_t)) \\
= \;& (1-\lambda_t)\cdot 1 + \frac{\lambda_t}{1+W_t(2\pi-1)} \cdot (1+W_t g(Z_t)) \\
= \;& 1 + \lambda_t \cdot \frac{W_t\,(g(Z_t)-(2\pi-1))}{1+W_t(2\pi-1)}.
\end{aligned}$$

Payoff for the Case of Unbalanced Classes (known π). To see that the payoff function (B.13) is lower bounded by negative one, note that:

$$\begin{aligned}
f_t^{\mathrm{u}}(z,1) &= \frac{g_t(z)-(2\pi-1)}{2\pi} \geq \frac{-1-(2\pi-1)}{2\pi} = -1, \\
f_t^{\mathrm{u}}(z,-1) &= \frac{-g_t(z)+(2\pi-1)}{2(1-\pi)} \geq \frac{-1+(2\pi-1)}{2(1-\pi)} = -1.
\end{aligned}$$

To see that such payoff is fair, note that:

$$\mathbb{E}_{H_0}\left[f_t^{\mathrm{u}}(Z_t,W_t) \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_P\left[\pi \cdot \frac{g_t(Z_t)-(2\pi-1)}{2\pi}\right] - \mathbb{E}_Q\left[(1-\pi)\cdot \frac{g_t(Z_t)-(2\pi-1)}{2(1-\pi)} \mid \mathcal{F}_{t-1}\right] = 0,$$

where $\mathcal{F}_{t-1} = \sigma\left(\{(Z_i,W_i)\}_{i\leq t-1}\right)$.

Theorem B.3. *Suppose that H_0 in* (B.11a) *is true. Then $(\mathcal{K}_t^{\mathrm{u}})_{t\geq 0}$ is a nonnegative supermartingale adapted to $(\mathcal{F}_t)_{t\geq 0}$. Hence, the sequential 2ST based on $(\mathcal{K}_t^{\mathrm{u}})_{t\geq 0}$ satisfies: $\mathbb{P}_{H_0}(\tau<\infty)\leq\alpha$.*

Proof. First, we show that $(\mathcal{K}_t^{\mathrm{u}})_{t\geq 0}$ is a nonnegative supermartingale. For any $t\geq 1$, the wealth $\mathcal{K}_{t-1}$ is multiplied at round t by

$$1+\lambda_t f_t^{\mathrm{u}}\left(\left\{(Z_{b(t-1)+i},W_{b(t-1)+i})\right\}_{i\in\{1,\ldots,b\}}\right) = (1-\lambda_t)\cdot 1 + \lambda_t \cdot \frac{\prod_{i=b(t-1)+1}^{bt}(1+W_i g_t(Z_i))}{\prod_{i=1}^{b}(1+W_i(2\widehat{\pi}_t-1))}.$$

Since $\lambda_t \in [0, 0.5]$, we conclude that the process $(\mathcal{K}_t^{\mathrm{u}})_{t\geq 0}$ is nonnegative. Next, note that since $\widehat{\pi}_t$ is the MLE of π computed from a t-th minibatch, it follows that:

$$\begin{aligned}1 + \lambda_t f_t^{\mathrm{u}}\left(\left\{(Z_{b(t-1)+i}, W_{b(t-1)+i})\right\}_{i\in\{1,\ldots,b\}}\right) &\leq (1-\lambda_t)\cdot 1 + \lambda_t \cdot \frac{\prod_{i=b(t-1)+1}^{bt}(1+W_i g_t(Z_i))}{\prod_{i=b(t-1)+1}^{bt}(1+W_i(2\pi-1))} \\ &= (1-\lambda_t)\cdot 1 + \lambda_t \cdot \prod_{i=b(t-1)+1}^{bt}\left(\frac{1+W_i g_t(Z_i)}{1+W_i(2\pi-1)}\right).\end{aligned}$$

Recall that $\mathcal{F}_{t-1} = \sigma\left(\{Z_i, W_i\}_{i\leq b(t-1)}\right)$. It suffices to show that if H_0 is true, then

$$\mathbb{E}_{H_0}\left[\prod_{i=b(t-1)+1}^{bt}\left(\frac{1+W_i g_t(Z_i)}{1+W_i(2\pi-1)}\right) \mid \mathcal{F}_{t-1}\right] = 1.$$

Note that the individual terms in the above product are independent conditional on $\mathcal{F}_{t-1}$. Hence,

$$\mathbb{E}_{H_0}\left[\prod_{i=b(t-1)+1}^{bt}\left(\frac{1+W_i g_t(Z_i)}{1+W_i(2\pi-1)}\right) \mid \mathcal{F}_{t-1}\right] = \prod_{i=b(t-1)+1}^{bt}\mathbb{E}_{H_0}\left[\frac{1+W_i g_t(Z_i)}{1+W_i(2\pi-1)} \mid \mathcal{F}_{t-1}\right].$$

For any $i \in \{b(t-1)+1, \ldots, bt\}$, it holds that:

$$\begin{aligned}\mathbb{E}_{H_0}\left[\frac{1+W_i g_t(Z_i)}{1+W_i(2\pi-1)} \mid \mathcal{F}_{t-1}\right] &= \mathbb{E}_{H_0}\left[\pi\cdot\frac{1+g_t(Z_i)}{1+(2\pi-1)} + (1-\pi)\cdot\frac{1-g_t(Z_i)}{1-(2\pi-1)} \mid \mathcal{F}_{t-1}\right] \\ &= \mathbb{E}_{H_0}\left[\frac{1+g_t(Z_i)}{2} + \frac{1-g_t(Z_i)}{2} \mid \mathcal{F}_{t-1}\right] \\ &= 1.\end{aligned}$$

Hence, we conclude that $(\mathcal{K}_t^{\mathrm{u}})_{t\geq 0}$ is a nonnegative supermartingale adapted to $(\mathcal{F}_t)_{t\geq 0}$. The time-uniform type I error control of the resulting test then follows from Ville's inequality (Proposition 8). □

B.5 Additional Experiments and Details

B.5.1 Modeling Details

CNN Architecture and Training. We use CNN with 4 convolutional layers (kernel size is taken to be 3×3) and 16, 32, 32, 64 filters respectively. Further, each convolutional layer is followed by max-pooling layer (2×2). After flattening, those layers are followed by 1 fully connected layer with 128 neurons. Dropout ($p = 0.5$) and early stopping (with patience equal to ten epochs and 20% of data used in the validation set) is used for regularization. ReLU activation functions are used in each layer. Adam optimizer is used for training the network. We start training

after processing twenty observations, and update the model parameters after processing every next ten observations. Maximum number of epochs is set to 25 for each training iteration. The batch size is set to 32.

Single-stream Sequential Kernelized 2ST. The construction of this test is the extension of 2ST of Shekhar and Ramdas (2021) to the case when at each round an observation only from a single distribution (P or Q) is revealed. Let $\mathcal{G}$ denote an RKHS with positive-definite kernel k and canonical feature map $\varphi(\cdot)$ defined on $\mathcal{Z}$. Recall that instances from P as labeled as $+1$ and instances from Q are labeled as -1 (characterized by W). The mean embeddings of P and Q are then defined as

$$\widehat{\mu}_P^{(t)} = \frac{1}{N_+(t)} \sum_{i=1}^{t} \varphi(Z_i) \cdot \mathbb{1}\{W_i = +1\},$$
$$\widehat{\mu}_Q^{(t)} = \frac{1}{N_-(t)} \sum_{i=1}^{t} \varphi(Z_i) \cdot \mathbb{1}\{W_i = -1\},$$

where $N_+(t) = |i \leq t : W_i = +1|$ and $N_-(t) = |i \leq t : W_i = -1|$. The corresponding payoff function is

$$f_t^{\mathrm{k}}(Z_{t+1}, W_{t+1}) = W_{t+1} \cdot \widehat{g}_t(Z_{t+1}),$$
$$\text{where} \quad \widehat{g}_t = \frac{\widehat{\mu}_P^{(t)} - \widehat{\mu}_Q^{(t)}}{\left\|\widehat{\mu}_P^{(t)} - \widehat{\mu}_Q^{(t)}\right\|_{\mathcal{G}}}.$$

To make the test computationally efficient, it is critical to update the normalization constant efficiently. Suppose that at round $t+1$, an instance from P is observed. In this case, $\widehat{\mu}_Q^{(t+1)} = \widehat{\mu}_Q^{(t)}$. Note that:

$$\begin{aligned}
\widehat{\mu}_P^{(t+1)} &= \frac{1}{N_+(t+1)} \sum_{i=1}^{t+1} \varphi(Z_i) \cdot \mathbb{1}\{W_i = +1\} \\
&= \frac{1}{N_+(t)+1} \sum_{i=1}^{t+1} \varphi(Z_i) \cdot \mathbb{1}\{W_i = +1\} \\
&= \frac{1}{N_+(t)+1} \varphi(Z_{t+1}) + \frac{1}{N_+(t)+1} \sum_{i=1}^{t} \varphi(Z_i) \cdot \mathbb{1}\{W_i = +1\} \\
&= \frac{1}{N_+(t)+1} \varphi(Z_{t+1}) + \frac{N_+(t)}{N_+(t)+1} \widehat{\mu}_P^{(t)}.
\end{aligned}$$

Hence, we have:

$$\begin{aligned}
\left\|\widehat{\mu}_P^{(t+1)} - \widehat{\mu}_Q^{(t+1)}\right\|_{\mathcal{G}}^2 &= \left\|\widehat{\mu}_P^{(t+1)} - \widehat{\mu}_Q^{(t)}\right\|_{\mathcal{G}}^2 \\
&= \left\|\widehat{\mu}_P^{(t+1)}\right\|_{\mathcal{G}}^2 - 2\left\langle \widehat{\mu}_P^{(t+1)}, \widehat{\mu}_Q^{(t)} \right\rangle_{\mathcal{G}} + \left\|\widehat{\mu}_Q^{(t)}\right\|_{\mathcal{G}}^2.
\end{aligned}$$

In particular,

$$\begin{aligned}
\left\langle \widehat{\mu}_P^{(t+1)}, \widehat{\mu}_Q^{(t)} \right\rangle_{\mathcal{G}} &= \left\langle \frac{1}{N_+(t)+1} \varphi(Z_{t+1}) + \frac{N_+(t)}{N_+(t)+1} \widehat{\mu}_P^{(t)}, \widehat{\mu}_Q^{(t)} \right\rangle_{\mathcal{G}} \\
&= \frac{1}{N_+(t)+1} \left\langle \varphi(Z_{t+1}), \widehat{\mu}_Q^{(t)} \right\rangle_{\mathcal{G}} + \frac{N_+(t)}{N_+(t)+1} \left\langle \widehat{\mu}_P^{(t)}, \widehat{\mu}_Q^{(t)} \right\rangle_{\mathcal{G}}.
\end{aligned}$$

Note that:

$$\left\langle \varphi(Z_{t+1}), \widehat{\mu}_Q^{(t)} \right\rangle_{\mathcal{G}} = \frac{1}{N_-(t)} \sum_{i=1}^{t} k(Z_{t+1}, Z_i) \cdot \mathbb{1}\left\{W_i = -1\right\}.$$

Next, we assume for simplicity that $k(x, x) = 1, \forall x$ which holds for RBF kernel. Observe that:

$$\begin{aligned}\left\|\widehat{\mu}_P^{(t+1)}\right\|_{\mathcal{G}}^2 &= \left\langle \widehat{\mu}_P^{(t+1)}, \widehat{\mu}_P^{(t+1)} \right\rangle_{\mathcal{G}} \\ &= \frac{1}{\left(N_+(t)+1\right)^2} + \frac{2N_+(t)}{\left(N_+(t)+1\right)^2} \left\langle \varphi(Z_{t+1}), \widehat{\mu}_P^{(t)} \right\rangle_{\mathcal{G}} + \frac{(N_+(t))^2}{\left(N_+(t)+1\right)^2} \left\|\widehat{\mu}_P^{(t)}\right\|_{\mathcal{G}}^2.\end{aligned}$$

By caching intermediate results, we can compute the normalization constant using linear in t number of kernel evaluations. We start betting once at least one instance is observed from both P and Q. For simulations, we use RBF kernel and the median heuristic with first 20 instances to compute the kernel hyperparameter.

MLP Training Scheme. We begin training after processing twenty datapoints from P_{XY} which gives a training dataset with 40 datapoints (due to randomization). When updating a model, we use previous parameters as initialization. We use the following update scheme: we start after next $n_0 = 10$ datapoints from P_{XY} are observed. Once n_0 becomes less than 1% of the size of the existing training dataset, we increase it by ten, that is, $n_t = n_{t-1} + 10$. When we fit the model, we set the maximum number of epochs to be 25 and use early stopping with patience of 3 epochs.

Kernel Hyperparameters for Synthetic Experiments. For SKIT, we use RBF kernels:

$$k(x, x') = \exp\left(-\lambda_X \left\|x - x'\right\|_2^2\right), \quad l(y, y') = \exp\left(-\lambda_Y \left\|y - y'\right\|_2^2\right).$$

For simulations on synthetic data, we take kernel hyperparameters to be inversely proportional to the second moment of the underlying variables (the median heuristic yields similar results):

$$\lambda_X = \frac{1}{2\mathbb{E}\left[\left\|X - X'\right\|_2^2\right]}, \quad \lambda_Y = \frac{1}{2\mathbb{E}\left[\left\|Y - Y'\right\|_2^2\right]}.$$

1. *Spherical model.* By symmetry, we have: $P_X = P_Y$, and hence we take $\lambda_X = \lambda_Y$. We have

$$\mathbb{E}\left[(X - X')^2\right] = 2\mathbb{E}\left[X^2\right] = \frac{2}{d}.$$

2. *HTDD model.* By symmetry, we have: $P_X = P_Y$, and hence we take $\lambda_X = \lambda_Y$. We have

$$\mathbb{E}\left[(X - X')^2\right] = 2\mathbb{E}\left[X^2\right] = \frac{2\pi^2}{3}.$$

3. *Sparse signal model.* We have

$$\mathbb{E}\left[\|X - X'\|_2^2\right] = 2\mathbb{E}\left[\|X\|_2^2\right] = 4d,$$
$$\mathbb{E}\left[\|Y - Y'\|_2^2\right] = 2\mathbb{E}\left[\|Y\|_2^2\right] = 2\mathrm{tr}(B_s B_s^\top + I_d) = 2(d + \sum_{i=1}^{d} \beta_i^2).$$

4. *Gaussian model.* We have

$$\mathbb{E}\left[(X - X')^2\right] = 2\mathbb{E}\left[X^2\right] = 2,$$
$$\mathbb{E}\left[(Y - Y')^2\right] = 2\mathbb{E}\left[Y^2\right] = 2(1 + \beta^2).$$

Ridge Regression. We use ridge regression as an underlying predictive model: $\widehat{g}_t(x) = \beta_0^{(t)} + x\beta_1^{(t)}$, where the coefficients are obtained by solving:

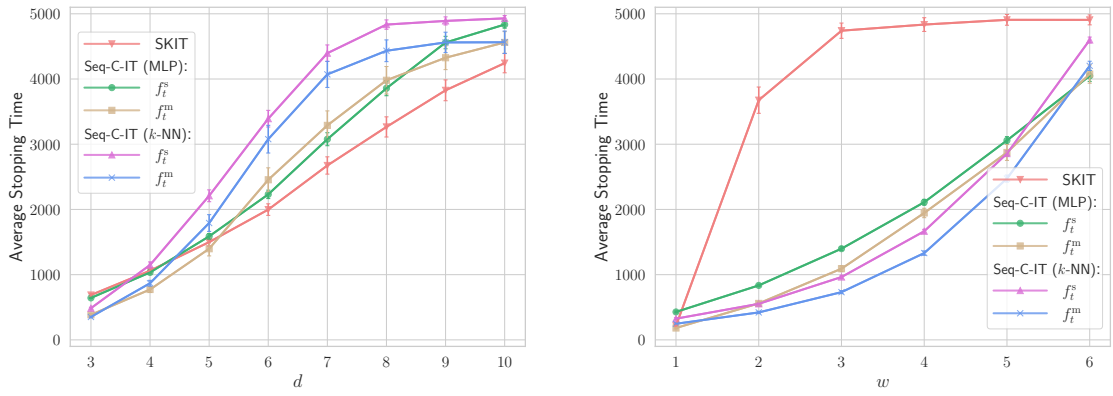
$$(\beta_0^{(t)}, \beta_1^{(t)}) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^{2(t-1)} (Y_i - X_i\beta_1 - \beta_0)^2 + \lambda\beta_1^2.$$

Let $\Gamma = \mathrm{diag}(0, 1)$. Let $\mathbf{X}_{t-1} \in \mathbb{R}^{2(t-1)\times 2}$ be such that $(\mathbf{X}_{t-1})_i = (1, X_i)$, $i \in [1, 2(t-1)]$. Finally, let $\mathbf{Y}_{t-1}$ be a vector of responses: $(\mathbf{Y}_{t-1})_i = Y_i$, $i \in [1, 2(t-1)]$. Then:

$$\beta^{(t)} = \arg\min_{\beta} \|\mathbf{Y}_{t-1} - \mathbf{X}_{t-1}\beta\|^2 + \lambda\beta^\top\Gamma\beta = \left(\mathbf{X}_{t-1}^\top\mathbf{X}_{t-1} + \lambda\Gamma\right)^{-1}(\mathbf{X}_{t-1}^\top\mathbf{Y}_{t-1}).$$

B.5.2 Additional Experiments for Seq-C-IT

In Figure B.2, we present average stopping times for ITs under the synthetic settings from Section 3.3. We confirm that all tests adapt to the complexity of a problem at hand, stopping earlier on easy tasks and later on harder ones. We



(a) Spherical model.

(b) HTDD model.

Figure B.2: Stopping times of ITs on synthetic data from Section 3.3. Subplot (a) shows that SKIT is only marginally better than Seq-C-IT (MLP) due to slightly better sample efficiency under the spherical model (no localized dependence). Under the structured HTDD model, SKIT is inferior to Seq-C-ITs.

also consider two additional synthetic examples where Seq-C-IT outperforms a kernelized approach:

1. *Sparse signal model.* Let $(X_t)_{t\geq 1}$ and $(\varepsilon_t)_{t\geq 1}$ be two independent sequences of standard Gaussian random vectors in $\mathbb{R}^d$: $X_t, \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \mathbf{I}_d)$, $t \geq 1$. We take

$$(X_t, Y_t) = (X_t, B_s X_t + \varepsilon_t),$$

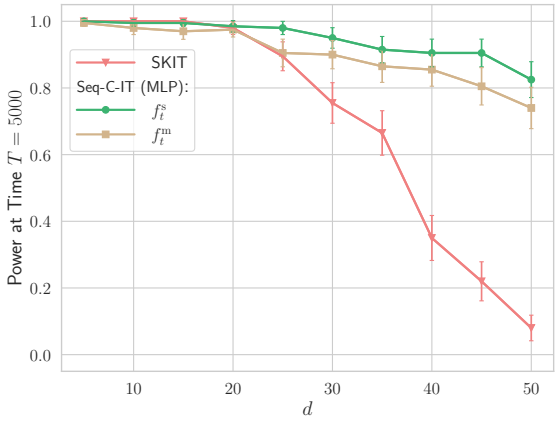
   where $B_s = \text{diag}(\beta_1, \ldots, \beta_d)$ and only $s = 5$ of $\{\beta_i\}_{i=1}^d$ are nonzero being sampled from $\text{Unif}([-0.5, 0.5])$. We consider $d \in \{5, \ldots, 50\}$.

2. *Nested circles model.* Let $(L_t)_{t\geq 1}$, $(\Theta_t)_{t\geq 1}$, $(\varepsilon_t^{(1)})_{t\geq 1}$, $(\varepsilon_t^{(2)})_{t\geq 1}$ denote sequences of random variables where $L \overset{\text{iid}}{\sim} \text{Unif}(1, \ldots, l)$ for some prespecified $l \in \mathbb{N}$, $\Theta_t \overset{\text{iid}}{\sim} \text{Unif}([0, 2\pi])$, and $\varepsilon_t^{(1)}, \varepsilon_t^{(2)} \overset{\text{iid}}{\sim} \mathcal{N}(0, (1/4)^2)$. For $t \geq 1$, we take

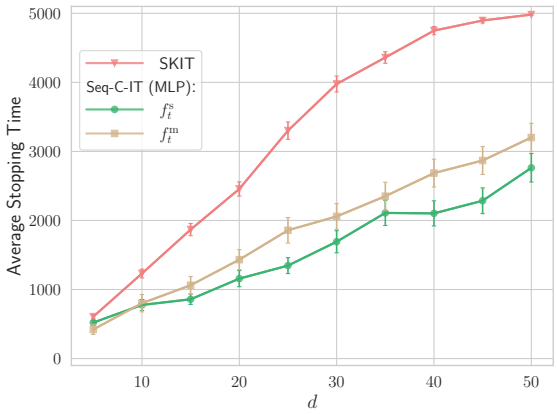
$$(X_t, Y_t) = (L_t \cos(\Theta_t) + \varepsilon_t^{(1)}, L_t \sin(\Theta_t) + \varepsilon_t^{(2)}). \tag{B.43}$$

   We consider $l \in \{1, \ldots, 10\}$.

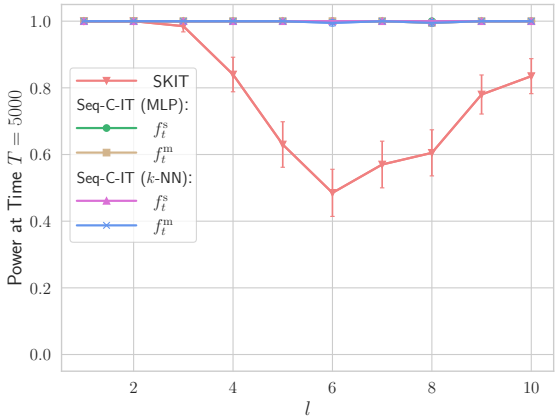
In Figure B.3, we show that Seq-C-ITs significantly outperform SKIT under these models. We note that the degrading performance of kernel-based tests under the nested circles model (B.43) has been also observed in earlier works (Berrett and Samworth, 2019; Podkopaev et al., 2023).



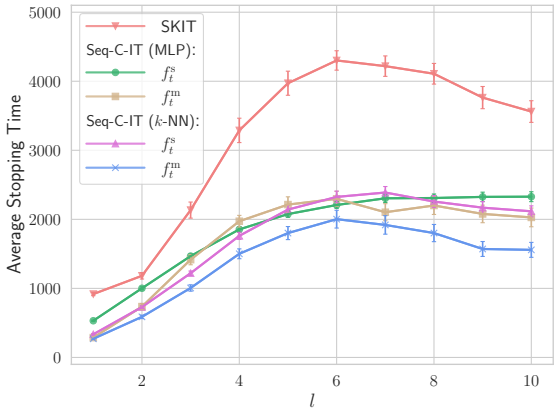
(a) Sparse signal model.



(b) Sparse signal model.



(c) Nested circles model.



(d) Nested circles model.

Figure B.3: Rejection rates (left column) and average stopping times (right column) of sequential ITs for synthetic datasets from Appendix B.5.2. In both cases, SKIT is inferior to Seq-C-ITs.

Appendix C

Additional Results for Chapter 4

C.1 Issues with Existing Tests for Distribution Shifts/Drifts

C.1.1 Non-sequential Tests Have Highly Inflated False Alarm Rates when Continuously Monitored

In this work, we propose a framework that utilizes confidence sequences (CSs), and thus allows for continuous monitoring of model performance. On the other hand, traditional (fixed-time) testing procedures are not valid under sequential settings, unless corrections for multiple testing are performed. First, we illustrate that deploying fixed-time detection procedures under sequential settings necessarily leads to raising false alarms. Then, we illustrate that naive corrections for multiple testing—without taking advantage of the dependence between the tests—lead to losses of power of the resulting procedure.

Deploying fixed-time tests without corrections for multiple testing. Under the i.i.d. setting, our framework reduces to testing whether the means corresponding to two unknown distributions are significantly different. Here we consider a simplified setting: assume that one observes a sequence $Z_1, Z_2, \ldots$ of bounded i.i.d. random variables, and the goal is to construct a lower confidence bound for the corresponding mean μ. In this case, a natural alternative to confidence sequences is a lower bound obtained by invoking the Central Limit Theorem:

$$\overline{Z}_t - z_\delta \cdot \frac{\widehat{\sigma}_t}{\sqrt{t}},$$

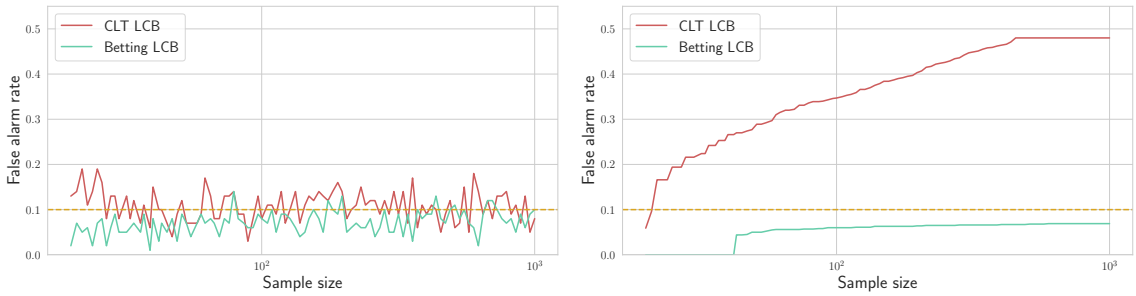
where z_δ is $(1-\delta)$-quantile of the standard Gaussian random variable and $\overline{Z}_t$, $\widehat{\sigma}_t$ denote the sample average and sample standard deviation respectively computed using first t instances $Z_1, \ldots, Z_t$. For the study below, we use the critical (significance) level $\delta = 0.1$. We sample the data as: $Z_t \sim \mathrm{Ber}(0.6)$, $t = 1, 2, \ldots$, and consider 100 possible

sample sizes, evenly spaced between 20 and 1000 on a logarithmic scale. Next, we compare the CLT lower bound with the betting-based one (which underlies the framework proposed in this work) under the following settings:

1. **Fixed-time monitoring.** For a given sample size t, we sample the sequence $Z_1, \ldots, Z_t$, compute the lower bounds and check whether the true mean is covered only once. For each value of the sample size, we resample data 100 times and record the miscoverage rate, that is, the fraction of times the true mean is miscovered.

2. **Continuous monitoring.** Here, the lower bound is recomputed once new data become available. We resample the whole data sequence $Z_1, \ldots, Z_{1000}$ 1000 times, and for each value of the sample size, we track the *cumulative* miscoverage rate, that is, the fraction of times the true mean has been miscovered *at some time* up to t.

Under fixed-time monitoring (Figure C.1a), the false alarm rate is controlled at prespecified level δ by both procedures. However, under continuous monitoring (Figure C.1b), deploying the CLT lower bound leads to raising false alarms. At the same time, the betting-based lower bound controls the false alarm rate under both types of monitoring.



(a) (b)

Figure C.1: False alarm rate for the CLT and betting-based lower confidence bound (LCB) under: (a) fixed-time monitoring and (b) continuous monitoring. Note that both bounds control the false alarm rate at a prespecified level $\delta = 0.1$ under fixed-time monitoring. However under continuous monitoring, the false alarm rate of the CLT bound quickly exceeds the critical level $\delta = 0.1$. At the same time, the betting LCB successfully controls the false alarm rate.

Deploying fixed-time tests with corrections for multiple testing. Next, we illustrate that adapting fixed-time tests to sequential settings via performing corrections for multiple testing comes at the price of significant power losses. Performing the Bonferroni correction requires splitting the available error budget δ among the times when testing is performed. In particular, we consider:

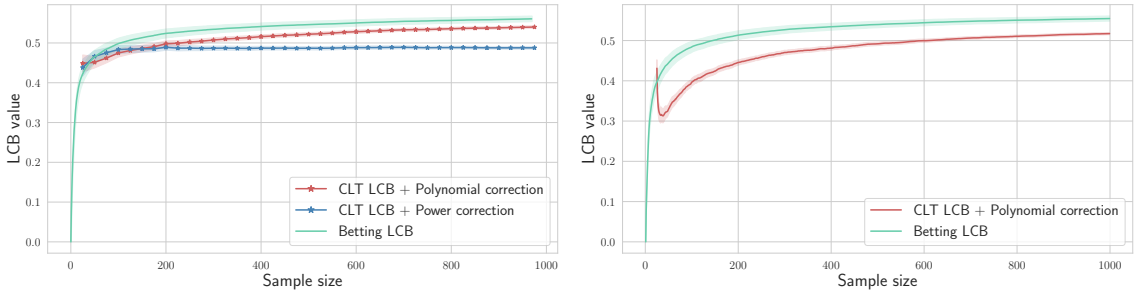
$$\begin{aligned} \textbf{Power correction:} \quad & \sum_{i=1}^{\infty} \frac{\delta}{2^i} = \delta, \\ \textbf{Polynomial correction:} \quad & \frac{6}{\pi^2} \sum_{i=1}^{\infty} \frac{\delta}{i^2} = \delta. \end{aligned} \tag{C.1}$$

Note that the second option is preferable as the terms in the sum decrease at a slower rate, thus allowing for a narrower sequence of intervals. Proceeding under the setup considered in the beginning of this section (data points are sampled from $\mathrm{Ber}(0.6)$), we consider two scenarios:

1. We recompute the CLT lower bound each time a batch of 25 samples is received and perform the Bonferroni correction (utilizing both ways of splitting the error budget described in (C.1)). We present the lower bounds on Figure C.2a (the results have been aggregated over 100 data draws). Observe that:

   - While the sequence of intervals shrinks in size with growing number of samples under the polynomial correction, this is not the case under the power correction.
   - Not only the betting confidence sequence is uniformly tighter than the CLT-based over all considered sample sizes, it also allows for monitoring at arbitrary stopping times. Note that the CLT lower bound allows for monitoring only at certain times (marked with stars on Figure C.2a).

2. For a fairer comparison with the betting-based bound, we also consider recomputing the CLT bound each time a new sample is received. Since utilizing the power correction quickly leads to numerical overflows, we utilize only the polynomial correction. We present the lower bounds on Figure C.2b (the results have been aggregated over 100 data draws). While now the CLT lower bound can be monitored at arbitrary stopping times, it is substantially lower (thus more conservative) than the betting-based.



(a) (b)

Figure C.2: Adapting the CLT lower bound to continuous monitoring via performing corrections for multiple testing: (a) each time a batch of 25 samples is received, (b) each time a new sample is received. Under both settings, the CLT-based lower bound is more conservative than the betting-based, which, in testing terminology, means that the resulting testing framework has less power.

C.1.2 Conformal Test Martingales may not Differentiate between Harmful and Benign Shifts

Testing the i.i.d. assumption online can be performed using conformal test martingales (Vovk et al., 2021; Vovk, 2020b,a). Below, we review building blocks underlying a conformal test martingale.

1. First, one has to pick a *conformity score*. Assigning a lower score to a sample indicates abnormal behavior. Vovk et al. (2021) consider the regression setting and, like us, use scores that depends on true labels: for a point (x_i, y_i), let $\widehat{y}_i$ denote the output of a predictor on input x_i. The authors propose a score of the form:

$$\alpha_i = -\left|y_i - \widehat{y}_i\right|. \tag{C.2}$$

Note that lower scores defined in (C.2) clearly reflect degrading performance of a predictor (possibly due to the presence of distribution drift). Under the classification setting, we propose to consider the following conformity score:

$$\alpha_i = \sum_{i=1}^{K} f_k(x_i) \cdot \mathbb{1}\left\{f_k(x_i) \leq f_{y_i}(x_i)\right\} = 1 - \sum_{i=1}^{K} f_k(x_i) \cdot \mathbb{1}\left\{f_k(x_i) > f_{y_i}(x_i)\right\}, \tag{C.3}$$

which is a rescaled estimated probability mass of all the labels that are more likely than the true one (here, we assume that predictor f outputs an element of $\Delta^{|\mathcal{Y}|}$). Rescaling in (C.3) is used to ensure that this score represent the *conformity* score, that is, the higher the value is, the better a given data point conforms. Note that if for a given data point, the true label happens to be top-ranked by a predictor f, then such point receives the largest conformity score equal to one. This score is inspired by recent works in conformal classification (Romano et al., 2020; Podkopaev and Ramdas, 2021).

2. After processing n data points, a *transducer* transforms a collection of conformity scores into a conformal p-value:

$$P_n = p\left(\{(x_i, y_i)\}_{i=1}^{n}, u\right) := \frac{\left|i \in \{1, \ldots, n\} : \alpha_i < \alpha_n\right| + u \cdot \left|i \in \{1, \ldots, n\} : \alpha_i = \alpha_n\right|}{n},$$

where $u \sim \mathrm{Unif}([0, 1])$. Conformal p-values are i.i.d. uniform on $[0, 1]$ when the data points are i.i.d. (or more generally, exchangeable; see (Vovk et al., 2021)). Note that the design of conformal p-value P_n ensures it takes small value when the conformity score α_n is small, that is, when abnormal behavior is being observed in a sequence.

3. A betting martingale is used to gamble again the null hypothesis that a sequence of random variables is distributed uniformly and independently on $[0, 1]$. Formally, a betting martingale is a measurable function $F : [0, 1]^\star \to [0, \infty]$ such that $F(\square) = 1$ ($\square$ defines an empty sequence and $Z^\star$ stands for the set of all finite sequences of elements of Z) and for each sequence $(u_1, \ldots, u_{n-1}) \in [0, 1]^{n-1}$ and any $n \geq 1$:

$$\int_0^1 F(u_1, \ldots, u_{n-1}, u) du = F(u_1, \ldots, u_{n-1}).$$

The simplest example is given by the product of simple bets:

$$F(u_1, \ldots, u_n) = \varepsilon^n \left(\prod_{i=1}^{n} u_i \right)^{1-\varepsilon}, \quad \varepsilon > 0, \tag{C.4}$$

but more sophisticated options are available (Vovk et al., 2005). For the simulations that follow, we use simple mixture martingale which is obtained by integrating (C.4) over $\varepsilon \in [0, 1]$.

Conformal test martingale S_n is obtained by plugging in the sequence of conformal p-values $P_1, \ldots, P_n$ into the betting martingale. The test starts with $S_0 = 1$ and it rejects at the first time n when S_n exceeds $1/\alpha$. They type I error control for this test is justified by Ville's inequality which states that for any nonnegative martingale (which S_n is, under the i.i.d. null), the entire process S_n stays below $1/\alpha$ with probability at least $1 - \alpha$. Mathematically:

$$\mathbb{P}\left(\exists n : S_n \geq 1/\alpha\right) \leq \alpha.$$

To study conformal test martingales, we consider the label shift setting described in Section 4.3.1. Recall that for this setting we know exactly when a shift in label proportions becomes harmful to misclassification risk of the Bayes-optimal rule on the source distribution (see Figures 4.2a and 4.3a). For the simulations that follow, we assume that the marginal probability of class 1 on the source is $\pi_1^S = 0.25$, and use the corresponding optimal rule.

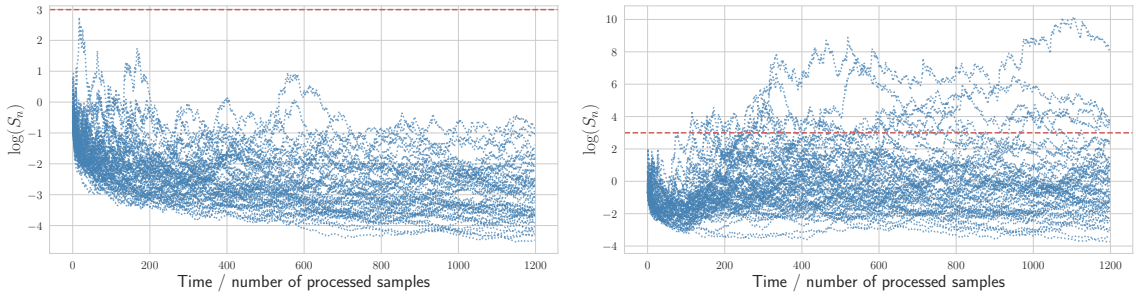
We analyze conformal test martingales under several distribution drift scenarios differing in their severity and rate, and start with the settings where a sharp shift is present.

1. ***Harmful* distribution *shift* with *cold start*.** Here, the data are sampled i.i.d. from a shifted distribution corresponding to $\pi_1^T = 0.75$. We illustrate 50 runs of the procedure on Figure C.3a. Recall that when the data are sampled i.i.d. the conformal p-values are i.i.d. uniform on $[0, 1]$. Under the (exchangeability) null, conformal test martingales are not growing, and thus are not able to detect that a present shift, even though it corresponds to a harmful setting.

2. ***Harmful* distribution *shift* with *warm start*.** For a fairer comparison, we also consider a warm start setting when the first 100 points are sampled i.i.d. from the source distribution ($\pi_1^T = 0.25$), followed by the data sampled i.i.d. from a shifted distribution ($\pi_1^T = 0.75$). We illustrate 50 runs of the procedure on Figure C.3b. In this case, conformal test martingales demonstrate better detection properties. However, a large fraction of conformal test martingales still is incapable of detecting a shift.

The simulations above illustrate that conformal test martingales are inferior to the framework proposed in this work whenever a sharp distribution shift happens in the early stage of a model deployment, even when such shift is harmful. Next, we consider several settings where instead of changing sharply, the distribution drifts gradually.

3. ***Slow* and *benign* distribution *drift*.** Starting with the marginal probability of class 1, $\pi_1^T = 0.1$, we keep *increasing* π_1^T by 0.05 each time a batch of 75 data points is sampled until it reaches the value 0.45. Recall from



(a) (b)

Figure C.3: 50 runs of conformal test martingales (blue dotted lines) under harmful distribution shift with: (a) cold start (shift happens in the beginning), (b) warm start (shift happens in an early stage of a model deployment). The horizontal red dashed line outlines to the rejection threshold due to Ville's inequality. Even though warm start improves detection properties, only a small fraction of conformal test martingales detects a shift that leads to more than 10% drop in classification accuracy.

Section 4.3.1 that those values of π_1^T correspond to a *benign* setting where the risk of the predictor on the target domain does not exceed substantially the source risk. We illustrate 50 runs of the procedure on Figure C.4a.

4. ***Slow* and *harmful* distribution *drift*.** Starting with the marginal probability of class 1 $\pi_1^T = 0.5$, we keep *increasing* π_1^T by 0.05 each time a batch of 75 data points is sampled until it reaches the value 0.85. Recall from Section 4.3.1 that those values of π_1^T correspond to a *harmful* setting where the risk of the predictor on the target domain is substantially larger than the source risk. We illustrate 50 runs of the procedure on Figure C.4b.

5. ***Sharp* and *harmful* distribution *drift*.** Starting with the marginal probability of class 1, $\pi_1^T = 0.1$, we keep *increasing* π_1^T by 0.2 each time a batch of 150 data points is sampled until it reaches the value 0.9. We illustrate 50 runs of the procedure on Figure C.4c.

The settings where the distribution drifts gradually illustrate several shortcomings of conformal test martingales:

- Conformal test martingales consistently detect *only* sharp distribution drifts. Recall from Section 4.3.1 that increasing π_1^T from 0.1 to 0.9 results in more than 20% accuracy drop. When a drift is slow (Figures C.4a and C.4b), conformal test martingales demonstrate much less power.

- Inspired by the ideas of Vovk et al. (2021) who assumed, like us, that (some) true data labels are observed, we designed a conformity score that reflects decrease in performance. On Figures C.4a and C.4b, conformal test martingales illustrate similar behavior but the corresponding settings are drastically different. Only one corresponds to a benign drift when the risk does not become significantly worse than the source risk. Thus, even though it is possible to make conformal test martingale reflect degrading performance, it is hard to incorporate evaluation of the malignancy of a change *in an interpretable way*, like a decrease of an important metric.

- Another problem of using conformal test martingales to be aware of is that after some time the corresponding values of the test martingale (larger implies more evidence for a shift) could start to decrease as the shifted

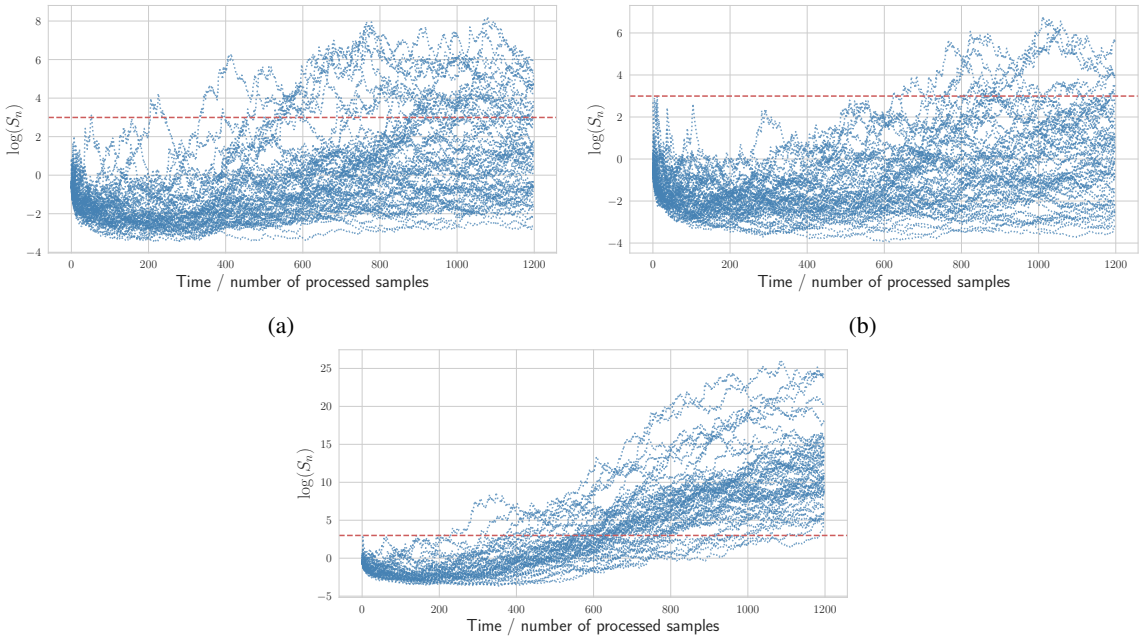


Figure C.4: 50 runs of conformal test martingales (blue dotted lines) under gradual distribution drifts: (a) slow and benign, (b) slow and harmful, (c) sharp and harmful. The horizontal red dashed line outlines to the rejection threshold due to Ville's inequality. Note that conformal test martingales consistently detect only sharp distribution drifts. Moreover, conformal test martingales illustrate similar behavior under (a) and (b) but the corresponding settings are drastically different.

distribution becomes the 'new normal' (Vovk et al., 2021). This is because they measure deviations from iid data, not degradations in performance from some benchmark (like source accuracy).

C.2 Loss Functions

For our simulations, we consider the following bounded losses. Below, let $\widehat{y}(x; f) := \arg\max_{k \in \mathcal{Y}} f_k(x)$ denote the label prediction of a model f on a given input $x \in \mathcal{X}$.

Multiclass losses. The most common example is arguably the misclassification loss and its generalization that allows for a label-dependent cost of a mistake:

$$\ell^{\text{mis}}(f(x), y) := \mathbb{1}\{\widehat{y}(x; f) \neq y\} \in \{0, 1\}, \quad \ell^{\text{w-mis}}(f(x), y) := \ell_y \cdot \mathbb{1}\{\widehat{y}(x; f) \neq y\} \in [0, L],$$

where $\{\ell_k\}_{k \in \mathcal{Y}}$ is a collection of per-class costs and $L = \max_{k \in \mathcal{Y}} \ell_k$. The loss $\ell^{\text{w-mis}}$ is more relevant to high-stakes decision settings and imbalanced classification. However, high accuracy alone can often be insufficient. The Brier score (squared error), introduced initially for the binary setting (Brier, 1950), is commonly employed to encourage

calibration of probabilistic classifiers. For multiclass problems, one could consider the mean-squared error of the whole output vector:

$$\ell^{\text{brier}}(f(x), y) := \frac{1}{2}\|f(x) - h(y)\|^2 \in [0, 1], \tag{C.5}$$

where $h : \mathcal{Y} \to \{0, 1\}^{|\mathcal{Y}|}$ is a one-hot label encoder: $h_{y'}(y) = \mathbb{1}\{y' = y\}$ for $y, y' \in \mathcal{Y}$. Top-label calibration (Gupta and Ramdas, 2022) restricts attention to the entry corresponding to the top-ranked label. A closely related loss function, which we call the *top-label Brier score*, is the following:

$$\ell^{\text{brier-top}}(f(x), y) := (f_{\widehat{y}(x;f)}(x) - \mathbb{1}\{\widehat{y}(x; f) = y\})^2 = (f_{\widehat{y}(x;f)}(x) - h_{\widehat{y}(x;f)}(y))^2 \in [0, 1]. \tag{C.6}$$

Alternatively, instead of the top-ranked label, one could focus only on the entry corresponding to the true class. It gives rise to another loss function which we call the *true-class* Brier score:

$$\ell^{\text{brier-true}}(f(x), y) := (f_y(x) - 1)^2 \in [0, 1]. \tag{C.7}$$

The loss functions ℓ^{brier}, $\ell^{\text{brier-top}}$, $\ell^{\text{brier-true}}$ trivially reduce to the same loss function in the binary setting. In Appendix C.3, we present a more detailed study of the Brier score in the multiclass setting with several illustrative examples.

Set-valued predictors. The proposed framework can be used for set-valued predictors that output a subset of $\mathcal{Y}$ as a prediction. Such predictors naturally arise in multilabel classification, where more than a single label can be the correct one, or as a result of post-processing point predictors. Post-processing could target covering the correct label of a test point with high probability (Vovk et al., 2005) or controlling other notions of risk (Bates et al., 2021) like the miscoverage loss:

$$\ell^{\text{miscov}}(y, S(x)) = \mathbb{1}\{y \notin S(x)\}, \tag{C.8}$$

where $S(x)$ denotes the output of a set-valued predictor on any given input $x \in \mathcal{X}$. When considering multilabel classification, relevant loss functions include the symmetric difference between the output and the set of true labels, false negative rate and false discovery rate.

C.3 Brier Score in the Multiclass Setting

This section contains derivations of decompositions stated in Section 4.2 and comparisons between introduced versions of the Brier score.

Brier score decompositions. Define a function $c : \mathcal{X} \to \Delta^{|\mathcal{Y}|}$, with entries $c_k(X) := \mathbb{P}(Y = k \mid f(X))$. In words, coordinates of $c(X)$ represent the true conditional probabilities of belonging to the corresponding classes given an output vector $f(X)$. Recall that $h : \mathcal{Y} \to \{0,1\}^{|\mathcal{Y}|}$ is a one-hot label encoder. The expected Brier score for the case when the whole output vector is considered (that is, the expected value of the loss defined in (C.5)) satisfies the following decomposition:

$$\begin{aligned}
2 \cdot R^{\text{brier}}(f) &= \mathbb{E}\|f(X) - h(Y)\|^2 \\
&= \mathbb{E}\|f(X) - c(X) + c(X) - h(Y)\|^2 \\
&\overset{(a)}{=} \mathbb{E}\|f(X) - c(X)\|^2 + \mathbb{E}\|c(X) - h(Y)\|^2 \\
&= \mathbb{E}\|f(X) - c(X)\|^2 + \mathbb{E}\|c(X) - \mathbb{E}\left[c(X)\right] + \mathbb{E}\left[c(X)\right] - h(Y)\|^2 \\
&\overset{(b)}{=} \underbrace{\mathbb{E}\|f(X) - c(X)\|^2}_{\text{calibration error}} - \underbrace{\mathbb{E}\|c(X) - \mathbb{E}\left[c(X)\right]\|^2}_{\text{sharpness}} + \underbrace{\mathbb{E}\|h(Y) - \mathbb{E}\left[h(Y)\right]\|^2}_{\text{intrinsic uncertainty}}.
\end{aligned}$$

Above, (a) follows by conditioning on $f(X)$ for the cross-term and recalling that $\mathbb{E}\left[h(Y) \mid f(X)\right] = c(X)$, (b) also follows by conditioning on $f(X)$ and noticing that $\mathbb{E}\left[h(Y)\right] = \mathbb{E}\left[c(X)\right]$. Now, recall that a predictor is (canonically) calibrated if $f(X) \overset{a.s.}{=} c(X)$, in which case the calibration error term is simply zero.

Next, we consider the top-label Brier score $\ell^{\text{brier-top}}$. Define $c^{\text{top}} : \mathcal{X} \to [0,1]$, as:

$$c^{\text{top}}(X) := \mathbb{P}\left(Y = \widehat{y}(X; f) \mid f_{\widehat{y}(X;f)}(X), \widehat{y}(X; f)\right),$$

or the fraction of correctly classified points among those that are predicted to belong to the same class and share the same confidence score as X. Following essentially the same argument as for the standard Brier score, we get that:

$$\begin{aligned}
&\mathbb{E}\left[\ell^{\text{brier-top}}(f(X), Y)\right] \\
&= \mathbb{E}\left(f_{\widehat{y}(X;f)}(X) - h_{\widehat{y}(X;f)}(Y)\right)^2 \\
&= \mathbb{E}\left(f_{\widehat{y}(X;f)}(X) - c^{\text{top}}(X) + c^{\text{top}}(X) - h_{\widehat{y}(X;f)}(Y)\right)^2 \\
&= \mathbb{E}\left(f_{\widehat{y}(X;f)}(X) - c^{\text{top}}(X)\right)^2 + \mathbb{E}\left(c^{\text{top}}(X) - h_{\widehat{y}(X;f)}(Y)\right)^2 \\
&= \underbrace{\mathbb{E}\left(f_{\widehat{y}(X;f)}(X) - c^{\text{top}}(X)\right)^2}_{\text{top-label calibration error}} - \underbrace{\mathbb{E}\left(c^{\text{top}}(X) - \mathbb{E}\left[c^{\text{top}}(X)\right]\right)^2}_{\text{top-label sharpness}} + \underbrace{\mathbb{V}\left(h_{\widehat{y}(X;f)}(Y)\right)}_{\substack{\text{variance of the} \\ \text{misclassification loss}}}.
\end{aligned}$$

Note that in contrast to the classic Brier score decomposition, the last term in this decomposition depends only on the top-class prediction of the underlying predictor, and thus on its accuracy.

Comparison of the scores in multiclass setting. Recall that the difference between three versions of the Brier score arises when one moves beyond the binary classification setting. We illustrate the difference by considering a 4-class classification problem where the data represent a balanced (that is all classes are equally likely) mixture of 4 Gaussians

with identity covariance matrix and mean vectors being the vertices of a 2-dimensional unit cube. One such sample is presented on Figure C.5a.

Next, we analyze locally the Brier scores when the Bayes-optimal rule is used as an underlying predictor, that is we split the area into small rectangles and estimate the mean score within each rectangle by a sample average. The results are presented on Figures C.5b, C.5c and C.5d. Note that the difference between the assigned scores is mostly observed for the points that lie at the intersection of 4 classes where the support of the corresponding output vectors is large.

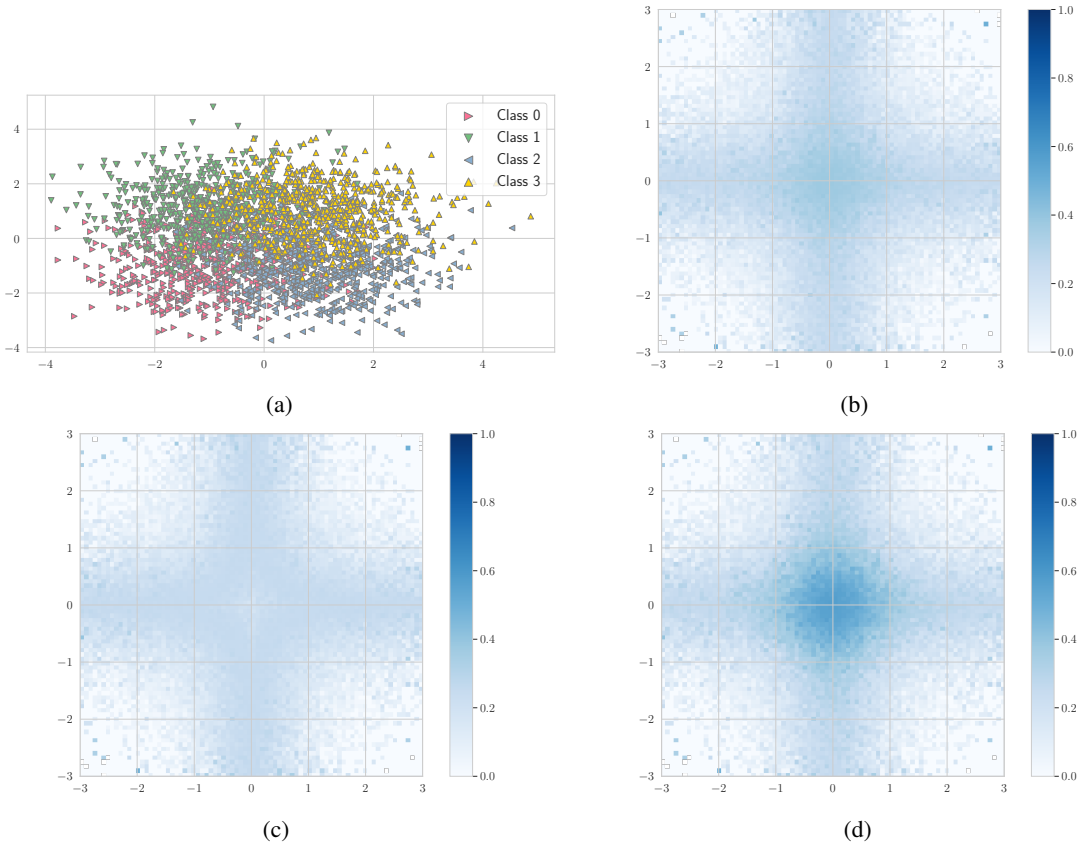
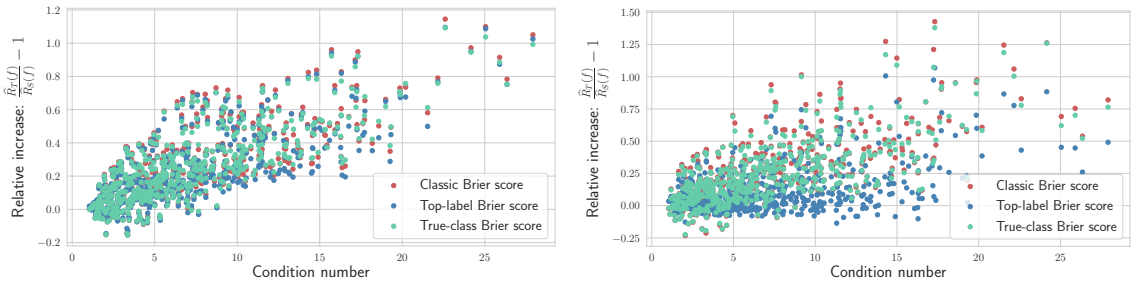


Figure C.5: (a) Visualization of 4-class classification problem with all classes being equally likely; (b) localized classic Brier score ℓ^{brier} ((C.5)); (c) localized top-label Brier score $\ell^{\text{brier-top}}$ ((C.6)); (d) localized true-class Brier score $\ell^{\text{brier-true}}$ ((C.7)).

Brier scores under label shift. Here we consider the case when label shift on the target domain is present. First, introduce the label likelihood ratios, also known as the importance weights, $w_y := \pi_y^T / \pi_y^S$, $y \in \mathcal{Y}$. For measuring the strength of the shift, we introduce the *condition number*: $\kappa = \sup_y w_y / \inf_{y: w_y \neq 0} w_y$. Note that when the shift is not present, the condition number $\kappa = 1$. To evaluate the sensitivity of the losses to the presence of label shift, we proceed as follows: first, the class proportions for both source and target domains are sampled from the Dirichlet distribution (to avoid extreme class proportion, we perform truncation at levels 0.15 and 0.85 and subsequent renormalization).

Then we use the Bayes-optimal rule for the source domain to perform predictions on the target and compute the corresponding losses. On Figure C.6, we illustrate relative increase in the average Brier scores plotted against the corresponding condition number when all data points are considered (Figure C.6a) and when attention is restricted to the area where classes highly intersect (Figure C.6b). In general, all three versions of the Brier score suffer similarly on average, but in the localized area where classes highly intersect, the top-label Brier score does not increase significantly under label shift.



(a) (b)

Figure C.6: (a) Relative increase for different versions of the Brier score in the multiclass setting under label shift; (b) Relative increase for different versions of the Brier score in the multiclass setting under label shift when attention is restricted to the area where classes highly intersect (cube with vertices at $(\pm 1/2, \pm 1/2)$). While in general, all three versions of the Brier score suffer similarly on average, in the localized area where classes highly intersect, the top-label Brier score does not increase significantly under label shift.

C.4 Proofs

Proof of Proposition 4. For brevity, we omit writing f for the source/target risks and the corresponding bound upper/lower confidence bounds. Starting with the absolute change and the null $H_0 : R_T - R_S \leq \varepsilon_{\text{tol}}$, we have:

$$\begin{aligned}
&\mathbb{P}_{H_0}\left(\exists t \geq 1 : \widehat{L}_T^{(t)} > \widehat{U}_S + \varepsilon_{\text{tol}}\right) \\
&= \mathbb{P}_{H_0}\left(\exists t \geq 1 : \left(\widehat{L}_T^{(t)} - R_T\right) - \left(\widehat{U}_S - R_S\right) > \varepsilon_{\text{tol}} - (R_T - R_S)\right) \\
&\leq \mathbb{P}_{H_0}\left(\exists t \geq 1 : \left(\widehat{L}_T^{(t)} - R_T\right) - \left(\widehat{U}_S - R_S\right) > 0\right).
\end{aligned}$$

Note that $\exists t \geq 1 : (\widehat{L}_T^{(t)} - R_T) - (\widehat{U}_S - R_S) > 0$ implies that either $\exists t \geq 1 : \widehat{L}_T^{(t)} - R_T > 0$ or $\widehat{U}_S - R_S < 0$. Thus, invoking union bound yields:

$$\begin{aligned}
&\mathbb{P}_{H_0}\left(\exists t \geq 1 : \left(\widehat{L}_T^{(t)} - R_T\right) - \left(\widehat{U}_S - R_S\right) > 0\right) \\
&\leq \mathbb{P}\left(\exists t \geq 1 : \widehat{L}_T^{(t)} - R_T > 0\right) + \mathbb{P}\left(\widehat{U}_S - R_S < 0\right) \\
&\leq \delta_T + \delta_S,
\end{aligned}$$

by construction and validity guarantees for $\widehat{L}_T^{(t)}$ and $\widehat{U}_S$. Similarly, considering the relative change, i.e., the null: $H_0' : R_T \leq (1+\varepsilon'_{\text{tol}})R_S$, we have:

$$\begin{aligned}
&\mathbb{P}_{H_0'}\left(\exists t \geq 1 : \widehat{L}_T^{(t)} > (1+\varepsilon'_{\text{tol}})\widehat{U}_S\right) \\
&= \mathbb{P}_{H_0'}\left(\exists t \geq 1 : \left(\widehat{L}_T^{(t)} - R_T\right) - (1+\varepsilon'_{\text{tol}})\left(\widehat{U}_S - R_S\right) > (1+\varepsilon'_{\text{tol}})R_S - R_T\right) \\
&\leq \mathbb{P}_{H_0'}\left(\exists t \geq 1 : \left(\widehat{L}_T^{(t)} - R_T\right) - (1+\varepsilon'_{\text{tol}})\left(\widehat{U}_S - R_S\right) > 0\right).
\end{aligned}$$

Similarly, note that $\exists t \geq 1 : (\widehat{L}_T^{(t)} - R_T) - (1+\varepsilon'_{\text{tol}})(\widehat{U}_S - R_S) > 0$ implies that either $\exists t \geq 1 : \widehat{L}_T^{(t)} - R_T > 0$ or $\widehat{U}_S - R_S < 0$. Thus, invoking union bound yields the desired result.

□

C.5 Primer on the Upper and Lower Confidence Bounds

This section contains the details for the concentration results used in this work. Results presented in this section are not new were developed in a series of recent works (Waudby-Smith and Ramdas, 2023; Howard et al., 2021). We follow the notation from Waudby-Smith and Ramdas (2023) for consistency and use the superscript (t) when referring to confidence sequences (CS) and (n) when referring to confidence intervals (CI).

Predictably-mixed Hoeffding's (PM-H) confidence sequence. Then the upper and lower endpoints of the predictably-mixed Hoeffding's (PM-H) confidence sequence are given by:

$$\begin{aligned}
L_{\text{PM-H}}^{(t)} &:= \left(\frac{\sum_{i=1}^t \lambda_i Z_i}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\delta) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right), \\
U_{\text{PM-H}}^{(t)} &:= \left(\frac{\sum_{i=1}^t \lambda_i Z_i}{\sum_{i=1}^t \lambda_i} + \frac{\log(1/\delta) + \sum_{i=1}^t \psi_H(\lambda_i)}{\sum_{i=1}^t \lambda_i}\right),
\end{aligned}$$

where $\psi_H(\lambda) := \lambda^2/8$ and $\lambda_1, \lambda_2, \ldots$ is a predictable mixture. We use a particular predictable mixture given by:

$$\lambda_t^{\text{PM-H}} := \sqrt{\frac{8\log(1/\delta)}{t\log(t+1)}} \wedge 1.$$

When approximating the risk on the source domain, one would typically have a holdout sample of a fixed size n, and so one could either use the classic upper limit of the Hoeffding's confidence interval, which is recovered by taking equal $\lambda_i = \lambda = \sqrt{8\log(1/\delta)/n}$, $i = 1, \ldots, n$, in which case the upper and lower limits simplify to:

$$L_{\text{H}}^{(n)} := \left(\frac{\sum_{i=1}^n Z_i}{n} - \sqrt{\frac{\log(1/\delta)}{2n}}\right), \quad U_{\text{H}}^{(n)} := \left(\frac{\sum_{i=1}^n Z_i}{n} + \sqrt{\frac{\log(1/\delta)}{2n}}\right),$$

or by considering running intersection of the predictably mixed Hoeffding's confidence sequence: $(\min_{t\le n} U^{(t)}_{\text{PM-H}}, \max_{t\le n} L^{(t)}_{\text{PM-H}})$.

Predictably-mixed empirical-Bernstein (PM-EB) confidence sequence. The upper and lower endpoints of the predictably-mixed empirical-Bernstein (PM-EB) confidence sequence are given by:

$$L^{(t)}_{\text{PM-EB}} := \frac{\sum_{i=1}^t \lambda_i Z_i}{\sum_{i=1}^t \lambda_i} - \frac{\log(1/\delta) + \sum_{i=1}^t v_i \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i},$$
$$U^{(t)}_{\text{PM-EB}} := \frac{\sum_{i=1}^t \lambda_i Z_i}{\sum_{i=1}^t \lambda_i} + \frac{\log(1/\delta) + \sum_{i=1}^t v_i \psi_E(\lambda_i)}{\sum_{i=1}^t \lambda_i},$$

where

$$v_i := 4\left(X_i - \widehat{\mu}_{i-1}\right)^2, \quad \text{and} \quad \psi_E(\lambda) := \left(-\log(1-\lambda) - \lambda\right)/4, \quad \text{for} \quad \lambda \in [0,1).$$

One particular choice of a predictable mixture $(\lambda^{\text{PM-EB}}_t)_{t=1}^\infty$ is given by:

$$\lambda^{\text{PM-EB}}_t := \sqrt{\frac{2\log(1/\delta)}{\widehat{\sigma}^2_{t-1} t \log(1+t)}} \wedge c, \quad \widehat{\sigma}^2_t := \frac{\frac{1}{4} + \sum_{i=1}^t (Z_i - \widehat{\mu}_i)^2}{t+1}, \quad \widehat{\mu}_t := \frac{\frac{1}{2} + \sum_{i=1}^t Z_i}{t+1},$$

for some $c \in (0,1)$. We use $c = 1/2$ and also set $\widehat{\mu}_0 = 1/2$, $\widehat{\sigma}_0 = 1/4$. If given a sample of a fixed size n, we consider running intersection along with the predictable sequence given by:

$$\lambda^{\text{PM-EB}}_t = \sqrt{\frac{2\log(1/\delta)}{n\widehat{\sigma}^2_{t-1}}} \wedge c, \quad t = 1, \ldots, n.$$

Betting-based confidence sequence. Tighter confidence intervals/sequences can be obtained by invoking tools from martingale analysis and deploying betting strategies for confidence intervals/sequences construction proposed in (Waudby-Smith and Ramdas, 2023). While those can not be computed in closed-form, empirically they tend to outperform previously considered confidence intervals/sequences. Recall that we are primarily interested in one-sided results, and for simplicity we discuss. For any $m \in [0,1]$, introduce a capital (wealth) process:

$$\mathcal{K}^{\pm}_t(m) := \prod_{i=1}^t \left(1 \pm \lambda^{\pm}_i(m) \cdot (Z_i - m)\right),$$

where $\left\{\lambda^+_t(m)\right\}_{t=1}^\infty$ and $\left\{\lambda^-_t(m)\right\}_{t=1}^\infty$ are $[0, 1/m]$-valued and $[0, 1/(1-m)]$-valued predictable sequences respectively. A particular example of such predictable sequences we use is given by:

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

where, for example, $c = 1/2$ or $3/4$ and $\dot{\lambda}_t^{\pm}$ do not depend on m. Such choice guarantees, in particular, that the resulting martingale is nonnegative. For example, the wealth process $\mathcal{K}_t^+(m)$ incorporates a belief that the true mean μ is larger than m and the converse belief is incorporated in $\mathcal{K}_t^-(m)$, that is, the wealth is expected to be accumulated under the corresponding belief (e.g., consider $m = 0$ and the corresponding $\mathcal{K}_t^+(0)$ with a high value, and $m = 1$ and the corresponding $\mathcal{K}_t^+(1)$ with a low value). Using that $\mathcal{K}_t^+(m)$ is non-increasing in m, i.e., $m_2 \geq m_1$, then $\mathcal{K}_t^+(m_2) \leq \mathcal{K}_t^+(m_1)$, we thus can use grid search (up to specified approximation error Δ_{grid}) to efficiently approximate $L_{\text{Bet}}^{(t)} = \inf B_t^+$, where

$$\mathcal{B}_t^+ := \left\{ m \in [0,1] : \mathcal{K}_t^+(m) < 1/\delta \right\},$$

that is, the collection of all m for which that the capital wasn't accumulated. Then we can consider $L_{\text{Bet}}^{(n)} = \max_{t \leq n} L_{\text{Bet}}^{(t)}$. When $m = \mu$ is considered, none of $\mathcal{K}_t^{\pm}(\mu)$ is expected to be large, since by Ville's inequality:

$$\mathbb{P}\left(\exists t \geq 1 : \mathcal{K}_t^+(\mu) \geq 1/\delta\right) \leq \delta,$$

and thus we know that with high probability the true mean is larger than $L_{\text{Bet}}^{(n)}$. That is,

$$\begin{aligned}\mathbb{P}\left(\mu < L_{\text{Bet}}^{(n)}\right) = \mathbb{P}\left(\mu < \max_{t \leq n} L_{\text{Bet}}^{(t)}\right) &= \mathbb{P}\left(\exists t \geq 1 : \mu < \inf B_t^+\right) \\ &= \mathbb{P}\left(\exists t \geq 1 : \mathcal{K}_t^+(\mu) \geq 1/\delta\right) \leq \delta.\end{aligned}$$

By a similar argument, we get that with high probability, the true mean is less than $U_{\text{Bet}}^{(n)} = \min_{t \leq n} \sup B_t^-$:

$$\begin{aligned}\mathbb{P}\left(\mu > U_{\text{Bet}}^{(n)}\right) = \mathbb{P}\left(\mu > \min_{t \leq n} \sup B_t^-\right) &= \mathbb{P}\left(\exists t \geq 1 : \mu > \sup B_t^-\right) \\ &= \mathbb{P}\left(\exists t \geq 1 : K_t^-(\mu) \geq 1/\delta\right) \leq \delta.\end{aligned}$$

Conjugate-mixture empirical-Bernstein (CM-EB) confidence sequence. Below, we present a shortened description of CM-EB and refer the reader to Howard et al. (2021) for more details. Assume that one observes a sequence of random variables Z_t, bounded in $[a, b]$ almost surely for all t, and the goal is to construct a confidence sequence for $\mu_t := t^{-1} \sum_{i=1}^t \mathbb{E}_{i-1} Z_i$, the average conditional expectation. Theorem 4 in Howard et al. (2021) states that for any $(\widehat{Z}_t)$, $[a, b]$-valued predictable sequence, and any u, the sub-exponential uniform boundary with crossing probability α for scale $c = b - a$, it holds that:

$$\mathbb{P}\left(\forall t \geq 1 : \left|\overline{Z}_t - \mu_t\right| < \frac{u\left(\sum_{i=1}^t (Z_i - \widehat{Z}_i)^2\right)}{t}\right) \geq 1 - 2\alpha,$$

where a reasonable choice for the predictable sequence $(\widehat{Z}_t)$ is given by $\widehat{Z}_t = (t-1)^{-1} \sum_{i=1}^{t-1} Z_i$.

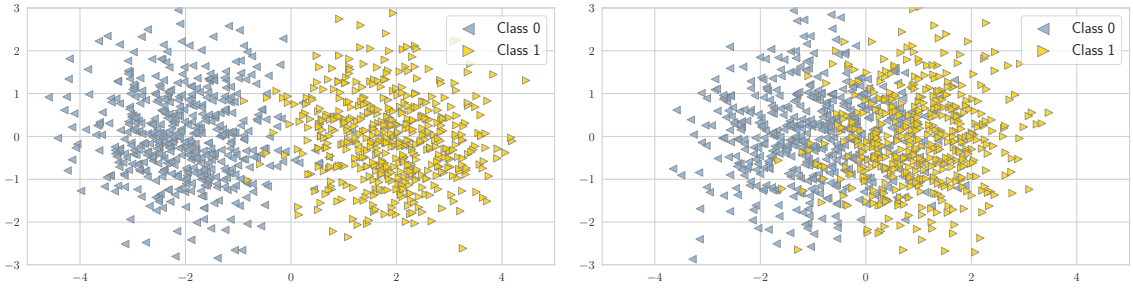
The key challenge which is addressed by conjugate mixtures is obtaining sublinear uniform boundaries that allows the radius of the confidence sequences to shrink to zero asymptotically. When a closed form of the confidence sequence is not required, the gamma-exponential mixture generally yields the tightest bounds. The procedure relies on the following *mixing* result (Lemma 2, Howard et al. (2021)) which states that for any $\alpha \in (0, 1)$ and any chosen probability distribution F on $[0, \lambda_{\max})$:

$$u_{\alpha}^{\mathrm{CM}}(v) := \sup \left\{ s :\in \mathbb{R} : \underbrace{\int \exp\left(\lambda s - \psi(\lambda) v\right) \mathrm{d}F(\lambda)}_{=:m(s,v)} < \frac{1}{\alpha} \right\},$$

yields a sub-ψ uniform boundary with crossing probability α. When the gamma distribution is used for mixing, $m(s, v)$ has a closed form given in [Proposition 9, Howard et al. (2021)]. Subsequently, the resulting *gamma-exponential mixture* boundary $u_{\alpha}^{\mathrm{CM}}(v)$ is computed by numerically solving the equation $m(s, v) = 1/\alpha$ in s. Howard et al. (2021) provide the packages for computing the corresponding confidence sequence.

C.6 Experiments on Simulated Data

Figure C.7 illustrates data samples for two settings where the presence of label shift is not expected to cause degradation in model performance (measured in terms of absolute increase in misclassification risk) for the first one ($\mu_0 = (-2, 0)^\top$, $\mu_1 = (2, 0)^\top$), but label shift may potentially degrade performance for the second ($\mu_0 = (-1, 0)^\top$, $\mu_1 = (1, 0)^\top$). For both cases, samples follow the same data generation pipeline as in Section 4.1 with only changes in class centers.



(a) (b)

Figure C.7: (a) Simulated dataset with well-separated classes. Presence of label shift presumably will not lead to a high absolute increase in the misclassification risk. (b) In contrast, when the classes are *not* well-separated, presence of label shift presumably might hurt the misclassification risk.

C.6.1 Brier Score as a Target Metric

Here we replicate the empirical study from Section 4.3.1 but use the Brier score as a target metric. Recall that all three multiclass versions of the Brier score discussed in this work reduce to the same loss in the binary setting. First, we compare upper confidence bounds for the Brier score computed by invoking different concentration results on Figure C.8. Similar to the misclassification risk, variance-adaptive confidence bounds exploit the low-variance structure and are tighter when compared against the non-adaptive one.

Next, we perform empirical analysis of the power of the testing framework. We take $\varepsilon_{\text{tol}} = 0.1$ which corresponds to testing for a 10% relative increase in the Brier score. We take $n_S = 1000$ data points from the source distribution to compute upper confidence bound on the source risk $\widehat{U}_S(f)$. Subsequently, we sample the data from the target distribution in batches of 50, with maximum number of samples from the target set to be 2000. On Figure C.8b, we present the proportion of cases when the null hypothesis is rejected out of 250 simulations performed for each candidate class 1 probability. On Figure C.8c, we illustrate average sample size from the target domain that was needed to reject the null hypothesis. When a stronger and more harmful label shift is present, less samples are required to reject the null, and moreover, the most powerful tests utilize upper/lower confidence bounds obtained via the betting approach. On Figure C.8d, we present the comparison of different time-uniform lower confidence bounds.

C.7 Experiments on Real Datasets

C.7.1 MNIST-C Simulation

Architecture and training. For MNIST-C dataset, we train a shallow CNN with two convolutional layers (each with 3×3 kernel matrices), each followed by max-pooling layers. Subsequently, the result is flattened and followed by a dropout layer ($p = 0.5$), a fully-connected layer with 128 neurons and an output layer. Note that the network is trained on original (clean) MNIST data, which is split split into two folds with 10% of data used for validation purposes. All images are scaled to $[0, 1]$ range before the training is performed.

Training multiple networks. To validate observations regarding shift malignancy from Section 4.3.2, we train 5 different networks (following the same training protocol) and report aggregated (over 25 random ordering of the data from the target) results on Figure C.9. The observation that applying translation to the MNIST images represents a harmful shift is consistent across all networks.

C.7.2 CIFAR-10-C Simulation

Architecture and training. The model underlying a set-valued predictor is a standard ResNet-32. It is trained for 50 epochs on the original (clean) CIFAR-10 dataset, without data augmentation, using 10% of data for validation

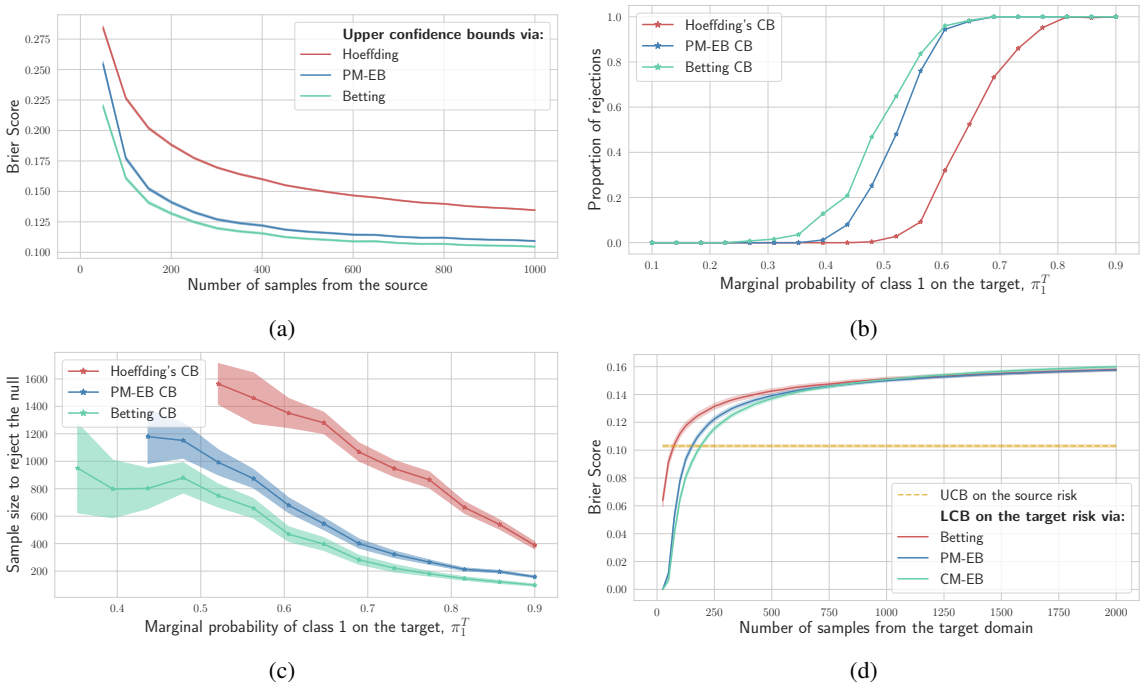


Figure C.8: (a) Upper confidence bounds $\widehat{U}_S(f)$ on the Brier score for the source domain. Similar to the misclassification risk, variance-adaptive confidence bounds are tighter when compared against the Hoeffding's one. For each fixed number of data points from the source domain used to compute $\widehat{U}_S(f)$, presented results are aggregated over 1000 random data draws. (b) Proportion of null rejections made by the procedure when testing for 10% relative increase of the Brier score. (c) Average sample size from the target distribution that was needed to reject the null. Invoking tighter concentration results allows to raise an alarm after processing less samples from the target domain. (d) Different lower/upper confidence bounds on the target/source domain for the Brier score.

purposes. All images are scaled to $[0, 1]$ range before the training is performed. The accuracy of the resulting network is $\approx 80.5\%$.

Transforming a point predictor into a set-valued one. To transform a point predictor into a set-valued one, we consider a sequence of candidate prediction sets $S_\lambda(x)$, parameterized by univariate parameter λ, with larger λ leading to larger prediction sets. Under the considered setting, the underlying predictor is an estimator of the true conditional probabilities $\pi_y(x) = \mathbb{P}(Y = y \mid X = x)$. Given a learnt predictor f, we can define

$$\rho_y(x; f) := \sum_{k=1}^{K} f_k(x) \cdot \mathbb{1}\{f_k(x) > f_y(x)\}$$

to represent estimated probability mass of the labels that are more likely than y. Subsequently, we can consider the following sequence of set-valued predictors:

$$S_\lambda(x) = \{y \in \mathcal{Y} : \rho_y(x; f) \leq \lambda\}, \quad \lambda \in \Lambda := [0, 1],$$

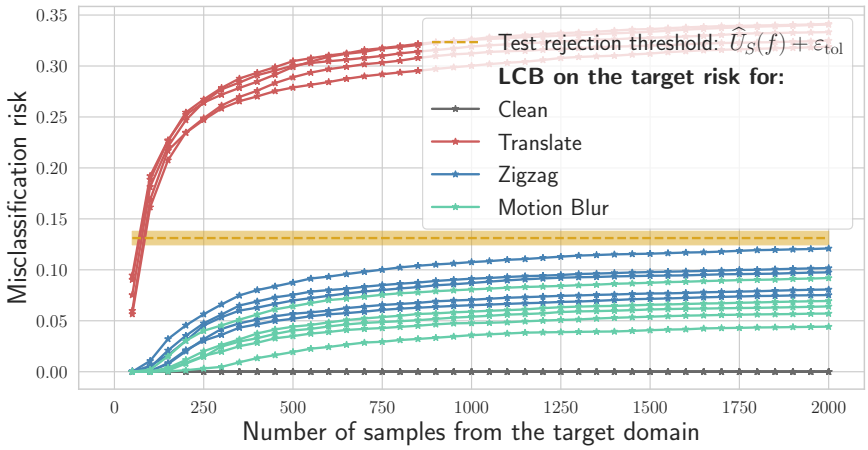


Figure C.9: Lines of the same color correspond to 5 different CNNs. For each network, the results aggregated over 25 random runs of the testing framework for randomly permuted test data. Applying translate effect is consistently harmful to the performance of CNNs trained on clean MNIST data. The bar around the yellow dashed line corresponds to 2 standard deviations.

that is the sequence is based on the estimated density level sets, starting by including the most likely labels according to the predictions of f. To tune the parameter λ, we follow Bates et al. (2021): we keep a labeled holdout *calibration set*, and use it to pick:

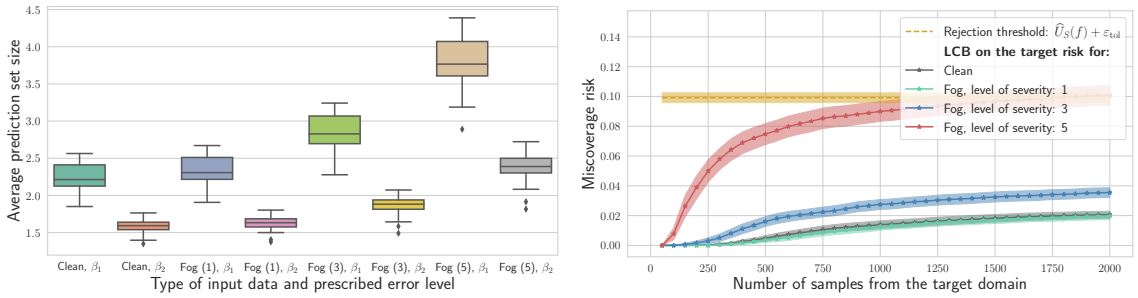
$$\widehat{\lambda} = \inf \left\{ \lambda \in \Lambda : \ \widehat{R}^{+}(\lambda') < \beta, \ \forall \lambda' > \lambda \right\},$$

where $\widehat{R}^{+}(\lambda')$ is an upper confidence bound for the risk function at level β. The resulting set-valued predictor is then (β, γ)-RCPS, that is,

$$\mathbb{P}\left(R(S_{\widehat{\lambda}}) \leq \beta\right) \geq 1 - \gamma,$$

under the i.i.d. assumption. More details and validity guarantees can be found in Bates et al. (2021).

Set-valued predictor when $\beta = 0.05$ is used as a prescribed error level. In contrast to $\beta = 0.1$ used in the main paper, we also consider decreasing β to 0.05, which in words, corresponds to increasing a desired coverage level of the resulting set-valued predictor. Figure C.10a compares average sizes of the prediction sets for two candidate values: $\beta_1 = 0.05$ and $\beta_2 = 0.1$, when the set-valued predictor is passes either clean CIFAR-10 data, or images to which fog corruption has been applied. As expected, decreasing β leads to larger prediction sets on average, with the size increasing when corrupted images are passes as input, that the size reflects uncertainty in prediction. In Figure C.10b, we observe that when we run the testing framework for the set-valued predictor corresponding to $\beta = 0.05$, only the most severe version of corruptions by adding fog is consistently marked as harmful, and thus raising an alarm. Similar to Section 4.3.2, we also use $\varepsilon_{\text{tol}} = 0.05$.



(a) (b)

Figure C.10: (a) Average size of prediction sets for $\beta_1 = 0.05$ and $\beta_2 = 0.1$ and different types of input data. First, lower β, corresponding to higher desired coverage, leads to larger prediction sets on average. Second, average size of the prediction sets increases when more corrupted images are passed as input, thus reflecting uncertainty in prediction. (b) Results of running the framework when $\beta_1 = 0.05$ is used to construct a wrapper. Observe that setting a lower prescribed error level β and thus enlarging resulting prediction sets partially mitigates the impact of corrupting images with the fog effect. However, the most severe form of such corruption still consistently leads to rejecting the null. The bars around dashed and solid lines correspond to 2 standard deviations.

C.8 Testing for Harmful Covariate Shift

In this section, we consider a case when covariate shift is present on the target domain, that is, when the marginal distribution $P(X)$ changes but $P(Y|X)$ does not. Consider the binary classification setting with accuracy being a target metric. It is known that the optimal decision rule for this case is the Bayes-optimal rule:

$$f^\star(x) = \mathbb{1}\left\{\mathbb{P}\left(Y = 1 \mid X = x\right) \geq 1/2\right\},$$

which minimizes the probability of misclassifying a new data point. Then one might expect that a change in the marginal distribution of X does not necessarily call for retraining if a learnt predictor f is 'close' to $f^\star$ (which itself could be a serious assumption to make). However, a change in the marginal distribution of X could indicate, in particular, that one should reconsider certain design choices made during training a model. To illustrate when it could be useful, we consider the following data generation pipeline:

1. Initially, each point is assigned either to the origin or to a circle of radius 1 centered at the origin with probability $1/2$.

2. For points assigned to the origin, the coordinates are sampled from multivariate Gaussian distribution with zero mean and rescaled identity covariance matrix: $\frac{1}{36} I_2$.

3. For points assigned to the circle, the coordinates are sampled from multivariate Gaussian distribution with the same covariance matrix but with the mean vector:

$$\begin{cases} \mu_x = \cos(\varphi), \\ \mu_y = \sin(\varphi), \end{cases}$$

where $\varphi^S \sim \mathrm{Unif}([-\pi/3, \pi/3])$ on the source domain and $\varphi^T \sim \mathrm{Unif}([0, 2\pi])$ on the target domain (see Figure C.11 for a visualization).

4. Then points are assigned the corresponding labels according to:

$$\mathbb{P}(Y = 1 \mid X = x) = \mathbb{1}\left\{x_1^2 + x_2^2 \geq \frac{1}{2}\right\}.$$

It is easy to see that a linear predictor, e.g., logistic regression, can achieve high accuracy if deployed on the data sampled from the source distribution. However, it will clearly fail to recover the true relationship between the features and responses. In this case, a change in the marginal distribution of X might indicate that updating a functional class could be necessary. On this data, we also run the proposed framework testing for a 10% increase in misclassification risk ($\varepsilon_{\text{tol}} = 0.1$). At each run, we use 200 points to train a logistic regression and 100 points to estimate the betting-based upper confidence bound on the source risk. On the target domain, we use the lower confidence bound due to conjugate-mixture empirical-Bernstein (CM-EB). The results presented on Figure C.11c (which have been aggregated over 100 random data draws) illustrate that the framework successfully detects a harmful shift, requiring only a small number of samples to do so.

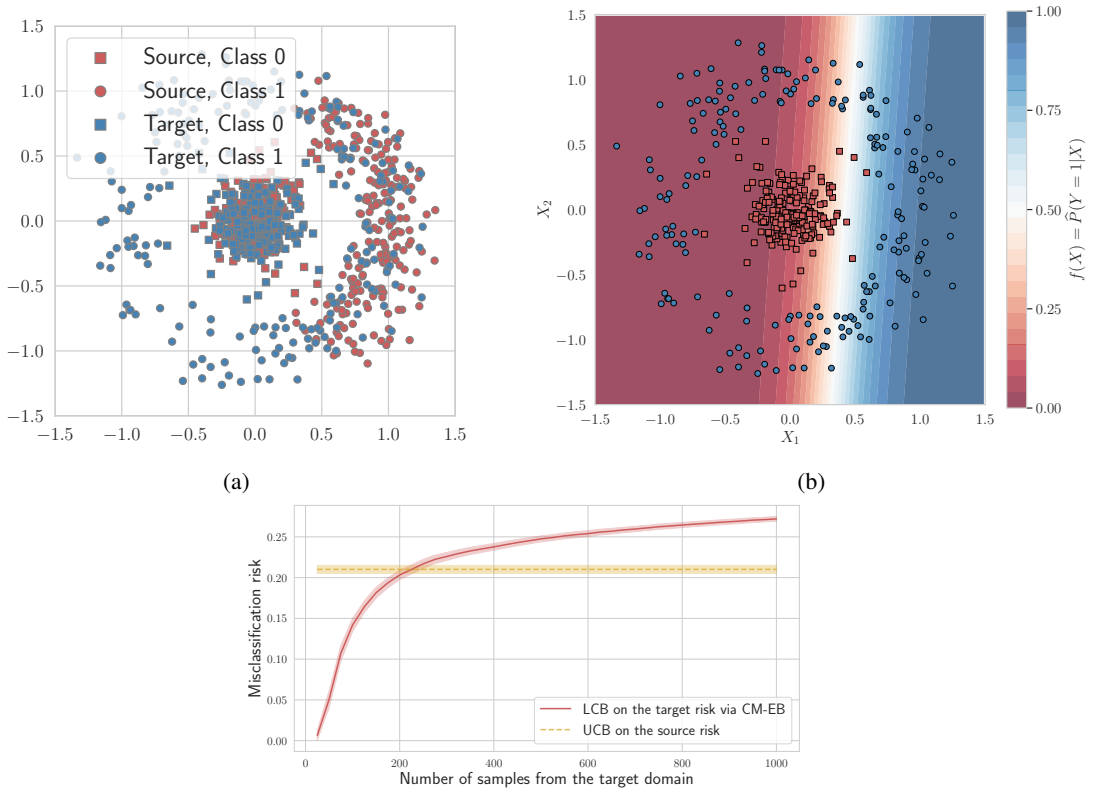


Figure C.11: (a) Data samples from the source (red) and target (blue) distributions for the covariate shift simulation. (b) Logistic regression predictor learnt on the source distribution plotted along with a data sample from the target distribution. While learnt predictor clearly has high accuracy on the source domain, it fails to approximate the true underlying data generating distribution. (c) Results of running the framework when testing for a 10% increase in the misclassification risk. The framework detects a harmful shift after processing only a small number of samples.

Appendix D

Additional Results for Chapter 5

The Appendix contains proofs of results in the main paper ordered as they appear. Auxiliary results needed for some of the proofs are stated in Appendix D.5.

D.1 Proof of Proposition 5

The 'if' part of the theorem is due to Vaicenavicius et al. (2019, Proposition 1); we reproduce it for completeness. Let $\sigma(g), \sigma(f)$ be the sub σ-algebras generated by g and f respectively. By definition of f, we know that f is $\sigma(g)$-measurable and, hence, $\sigma(f) \subseteq \sigma(g)$. We now have:

$$\begin{aligned}
\mathbb{E}\left[Y \mid f(X)\right] &= \mathbb{E}\left[\mathbb{E}\left[Y \mid g(X)\right] \mid f(X)\right] && \text{(by tower rule since } \sigma(f) \subseteq \sigma(g)\text{)} \\
&= \mathbb{E}\left[f(X) \mid f(X)\right] && \text{(by property (5.5))} \\
&= f(X).
\end{aligned}$$

The 'only if' part can be verified for $g = f$. Since f is perfectly calibrated,

$$\mathbb{E}\left[Y \mid f(X) = f(x)\right] = f(x),$$

almost surely P_X.

□

D.2 Proofs of results in Section 5.3

Proof of Theorem 5.1. Assume that one is given a predictor f that is (ε, α)-approximately calibrated. Then the assertion follows from the definition of (ε, α)-approximate calibration since:

$$|\mathbb{E}[Y \mid f(X)] - f(X)| \leq \varepsilon(f(X)) \implies \mathbb{E}[Y \mid f(X)] \in C(f(X)).$$

Now we show the proof in the other direction. Since ε is a constant-valued function that depends on C, let us denote its constant output as $\varepsilon_C := \varepsilon(\cdot) = \sup_{z \in \text{Range}(f)} \{|C(z)|/2\}$.

If m_C was injective, $\mathbb{E}[Y \mid m_C(f(X))] = \mathbb{E}[Y \mid f(X)]$ and thus if $\mathbb{E}[Y \mid f(X)] \in C(f(X))$ (which happens with probability at least $1 - \alpha$), we would have $\mathbb{E}[Y \mid m_C(f(X))] \in C(f(X))$ and so

$$|\mathbb{E}[Y \mid m_C(f(X))] - m_C(f(X)| \leq \sup_{z \in \text{Range}(f)} \{|C(z)|/2\} = \varepsilon_C.$$

This serves as an intuition for the proof in the general case, when m_C need not be injective. Note that,

$$\begin{aligned}
|\mathbb{E}[Y \mid m_C(f(X))] - m_C(f(X))| &= |\mathbb{E}[Y \mid m_C(f(X))] - \mathbb{E}[m_C(f(X)) \mid m_C(f(X))]| \\
&\overset{(1)}{=} |\mathbb{E}[\mathbb{E}[Y \mid f(X)] \mid m_C(f(X))] - \mathbb{E}[m_C(f(X)) \mid m_C(f(X))]| \\
&\overset{(2)}{=} |\mathbb{E}[\mathbb{E}[Y \mid f(X)] - m_C(f(X)) \mid m_C(f(X))]| \\
&\overset{(3)}{\leq} \mathbb{E}[|\mathbb{E}[Y \mid f(X)] - m_C(f(X))| \mid m_C(f(X))], \qquad \text{(D.1)}
\end{aligned}$$

where we use the tower rule in (1) (since m_C is a function of f), linearity of expectation in (2) and Jensen's inequality in (3). To be clear, the outermost expectation above is over $f(X)$ (conditioned on $m_C(f(X))$). Consider the event

$$A : \mathbb{E}[Y \mid f(X)] \in C(f(X)).$$

On A, by definition we have:

$$|\mathbb{E}[Y \mid f(X)] - m_C(f(X))| = \frac{u_C(f(X)) - l_C(f(X))}{2} \leq \sup_{z \in \text{Range}(f)} \left(\frac{|C(z)|}{2} \right) = \varepsilon_C.$$

By monotonicity property of conditional expectation, we also have that conditioned on A,

$$\mathbb{E}[|\mathbb{E}[Y \mid f(X)] - m_C(f(X))| \mid m_C(f(X))] \leq \mathbb{E}[\varepsilon_C \mid m_C(f(X))] = \varepsilon_C,$$

with probability 1. Thus by the relationship proved in the series of equations ending in (D.1), we have that conditioned on A, with probability 1,

$$|\mathbb{E}[Y \mid m_C(f(X))] - m_C(f(X))| \leq \varepsilon_C.$$

Since we are given that C is a $(1-\alpha)$-CI with respect to f, $\mathbb{P}(A) \geq 1-\alpha$. For any event B, it holds that $\mathbb{P}(B) \geq \mathbb{P}(B|A)\,\mathbb{P}(A)$. Setting

$$B : |\mathbb{E}[Y \mid m_C(f(X))] - m_C(f(X))| \leq \varepsilon_C,$$

we obtain:

$$\mathbb{P}(|\mathbb{E}[Y \mid m_C(f(X))] - m_C(f(X))| \leq \varepsilon_C) \geq 1-\alpha.$$

Thus, we conclude that $m_C(f(\cdot))$ is (ε, α)-approximately calibrated. □

Proof of Corollary 5.1.1. Let $\{f_n\}_{n\in\mathbb{N}}$ be asymptotically calibrated sequence with the corresponding sequence of functions $\{\varepsilon_n\}_{n\in\mathbb{N}}$ that satisfy $\varepsilon_n(f_n(X_{n+1})) = o_P(1)$. From Theorem 5.1, we can construct corresponding functions C_n that are $(1-\alpha)$-CI with respect to f_n and satisfy

$$|C_n(f_n(X_{n+1}))| = 2\varepsilon_n(f_n(X_{n+1})) = o_P(1).$$

This concludes the proof. □

Proof of Theorem 5.2. In the proof we write the test point as (X_{n+1}, Y_{n+1}). Suppose $\widehat{C}_n$ is a $(1-\alpha)$-CI with respect to f for all distributions P. We show that $\widehat{C}_n$ covers the label Y_{n+1} itself for distributions P such that $P_{f(X)}$ is nonatomic (and thus disc$(\widehat{C}_n)$ would also cover the labels).

Let P be any distribution such that $P_{f(X)}$ is nonatomic. Fix a set of $m \geq n+1$ samples from the distribution P denoted as $\mathcal{T} = \{(A^{(j)}, B^{(j)})\}_{j\in[m]}$. Given $\mathcal{T}$, consider a distribution Q corresponding to the following sampling procedure for $(X, Y) \sim Q$:

$$\begin{cases} \text{sample an index } j \text{ uniformly at random from } [m] \\ \text{set } (X, Y) = (A^{(j)}, B^{(j)}). \end{cases}$$

The distribution function for Q is given by

$$m^{-1} \sum_{j=1}^{m} \delta_{(A^{(j)}, B^{(j)})}.$$

where $\delta_{(a,b)}$ denotes the points mass at (a, b). Note that Q is only defined conditional on $\mathcal{T}$. Observe the following facts about Q:

- supp$(Q) = \{(A^{(j)}, B^{(j)})\}_{j\in[m]}$.

- Consider any $(x, y) \in \text{supp}(Q)$. Let $(x, y) = (A^{(j)}, B^{(j)})$ for some $j \in [m]$. Then

$$\begin{aligned}\mathbb{E}_Q\left[Y \mid f(X) = f(x)\right] &= \mathbb{E}_Q\left[Y \mid f(X) = f(A^{(j)})\right] \\ &\overset{\xi_1}{=} \mathbb{E}_Q\left[Y \mid X = A^{(j)}\right] \\ &\overset{\xi_2}{=} B^{(j)} = y.\end{aligned}$$

  Above ξ_1 holds since $P_{f(X)}$ is nonatomic so that the $f(X^{(i)})$'s are unique almost surely. Note that $P_{f(X)}$ is nonatomic only if P_X itself is nonatomic. Thus the $A^{(j)}$'s are unique almost surely, and ξ_2 follow. In other words, if $(X, Y) \sim Q$, then we have

$$Y = \mathbb{E}_Q\left[Y \mid f(X)\right]. \tag{D.2}$$

Suppose the data distribution was Q, that is $\{(X_i, Y_i)\}_{i\in[n+1]} \sim Q^{n+1}$. Define the event that the CI guarantee holds as

$$E_1 : \mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right] \in \widehat{C}_n(f(X_{n+1})), \tag{D.3}$$

and the event that the PS guarantee holds as

$$E_2 : Y_{n+1} \in \widehat{C}_n(f(X_{n+1})). \tag{D.4}$$

Then due to (D.2), the events are exactly the same under Q:

$$E_1 \overset{Q}{\equiv} E_2. \tag{D.5}$$

In particular, this means

$$\mathbb{P}_{Q^{n+1}}(\mathbb{E}_Q\left[Y_{n+1} \mid f(X_{n+1})\right] \in \widehat{C}_n(f(X_{n+1}))) = \mathbb{P}_{Q^{n+1}}(Y_{n+1} \in \widehat{C}_n(f(X_{n+1}))). \tag{D.6}$$

If $\widehat{C}_n$ is a distribution-free CI, then $\mathbb{P}_{Q^{n+1}}(E_1) \geq 1 - \alpha$ and thus $\mathbb{P}_{Q^{n+1}}(E_2) \geq 1 - \alpha$. This shows that for Q, $\text{disc}(\widehat{C}_n)$ is a $(1-\alpha)$-PI. Note that Q corresponds to sampling *with replacement* from a fixed set $\mathcal{T}$ where each element is drawn with respect to P. Although $Q \neq P$, we expect that as $m \to \infty$ (while n is fixed), Q and P coincide. This would prove the result for general P. To formalize this intuition, we describe a distribution which is close to Q but corresponds to sampling *without replacement* from $\mathcal{T}$ instead.

For this, now suppose that $\{(X_i, Y_i)\}_{i\in[n+1]} \sim R^{n+1}$ where R^{n+1} corresponds to sampling without replacement from $\mathcal{T}$. Formally, to draw from R^{n+1}, we first draw a surjective mapping $\lambda : [n+1] \to [m]$ as

$$\lambda \sim \text{Unif}\ (n\text{-sized ordered subsets of } [m]),$$

and set $(X_i, Y_i) = (A^{(\lambda(i))}, B^{(\lambda(i))})$ for $i \in [n+1]$.

First we quantify precisely the intuition that as $m \to \infty$, Q^{n+1} and R^{n+1} are essentially identical. Consider the event $T :=$ no index is repeated in Q^{n+1}. Let $\mathbb{P}(T) = \tau_m$ for some m and note that $\lim_{m\to\infty} \tau_m = 1$. Now consider any probability event E over $\{(X_i, Y_i)\}_{i\in[n+1]}$ (such as E_1 or E_2). We have

$$\begin{aligned}\mathbb{P}_{Q^{n+1}}(E) &= \mathbb{P}_{Q^{n+1}}(E|T) \cdot \mathbb{P}(T) + \mathbb{P}_{Q^{n+1}}(E|T^c) \cdot \mathbb{P}(T^c) \\ &\in [\mathbb{P}_{Q^{n+1}}(E|T) \cdot \mathbb{P}(T), \mathbb{P}_{Q^{n+1}}(E|T) \cdot \mathbb{P}(T) + \mathbb{P}(T^c)].\end{aligned}$$

Now observe that $\mathbb{P}_{Q^{n+1}}(E|T) = \mathbb{P}_{R^{n+1}}(E)$ to conclude

$$\mathbb{P}_{Q^{n+1}}(E) \in [\mathbb{P}_{R^{n+1}}(E) \cdot \mathbb{P}(T), \mathbb{P}_{R^{n+1}}(E) \cdot \mathbb{P}(T) + \mathbb{P}(T^c)].$$

Since $m \geq n+1$, $\mathbb{P}(T) \neq 0$ so we can invert the above and substitute $\tau_m = \mathbb{P}(T)$ to get

$$\mathbb{P}_{R^{n+1}}(E) \in \left[\tau_m^{-1}(\mathbb{P}_{Q^{n+1}}(E) - (1-\tau_m)),\ \tau_m^{-1}\mathbb{P}_{Q^{n+1}}(E)\right]. \tag{D.7}$$

Consider $E = E_2$ defined in equation (D.4). We showed that $\mathbb{P}_{Q^{n+1}}(E_2) \geq 1 - \alpha$. Thus from (D.7),

$$\mathbb{P}_{R^{n+1}}(E_2) \geq \tau_m^{-1}(1 - \alpha - (1 - \tau_m)).$$

The above is with respect to R^{n+1} which is conditional on a fixed draw $\mathcal{T}$. However since the right hand side is independent of $\mathcal{T}$, we can also include the randomness in $\mathcal{T}$ to say:

$$\mathbb{P}_{R^{n+1},\mathcal{T}}(E_2) \geq \tau_m^{-1}(1 - \alpha - (1 - \tau_m)). \tag{D.8}$$

Observe that if we consider the marginal distribution over R^{n+1} and $\mathcal{T}$ (that is we include the randomness in $\mathcal{T}$ as above), $\{(X_i, Y_i)\}_{i\in[n+1]} \overset{iid}{\sim} P$. This is not true if we do not marginalize over $\mathcal{T}$, in particular since the (X_i, Y_i)'s are not independent (due to sampling without replacement). Thus equation (D.8) can be restated as

$$\mathbb{P}_{P^{n+1}}(E_2) \geq \tau_m^{-1}(1 - \alpha - (1 - \tau_m)),$$

Since m can be set to any number and $\lim_{m\to\infty} \tau_m = 1$, we can indeed conclude

$$\mathbb{P}_{P^{n+1}}(E_2) \geq 1 - \alpha.$$

Recall that E_2 is the event that $Y_{n+1} \in \widehat{C}_n(X_{n+1})$; equivalently $Y_{n+1} \in \text{disc}\widehat{C}_n(X_{n+1})$. Thus $\text{disc}(\widehat{C}_n)$ provides a $(1-\alpha)$-PI for P such that $P_{f(X)}$ is nonatomic. □

Proof of Corollary 5.2.1. Let P be any distribution such that $P_{f(X)}$ is nonatomic. By Theorem 5.2, $\widehat{C}_n$ must provide both a prediction set and a confidence interval for P:

$$\mathbb{P}(\mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right] \in \widehat{C}_n(f(X_{n+1}))) \geq 1 - \alpha,$$

and

$$\mathbb{P}(Y_{n+1} \in \widehat{C}_n(f(X_{n+1}))) \geq 1 - \alpha.$$

Thus by a union bound

$$\mathbb{P}_{P^{n+1}}(\{Y_{n+1}, \mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right]\} \subseteq \widehat{C}_n(f(X_{n+1}))) \geq 1 - 2\alpha. \tag{D.9}$$

Now consider a distribution P such that $P_{f(X)}$ is nonatomic and $\mathbb{P}(Y = 1 \mid X) = 0.5$ a.s. P_X so that $\mathbb{E}\left[Y_{n+1} \mid f(X)\right] = 0.5$ a.s. $P_{f(X)}$. The inequality (D.9) is true for this P as well. If

$$\{Y_{n+1}, \mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right]\} \subseteq \widehat{C}_n(f(X_{n+1})),$$

then $|\widehat{C}_n(X_{n+1})| \geq |Y_{n+1} - \mathbb{E}\left[Y_{n+1} \mid f(X_{n+1})\right]| \geq 0.5$. Thus

$$\mathbb{P}_{P^{n+1}}(|\widehat{C}_n(f(X_{n+1}))| \geq 0.5) \geq 1 - 2\alpha.$$

Consequently we have

$$\begin{aligned}\mathbb{E}_{P^{n+1}}|\widehat{C}_n(f(X_{n+1}))| &\geq 0.5(1 - 2\alpha) \\ &= 0.5 - \alpha.\end{aligned}$$

This concludes the proof. □

Proof of Theorem 5.3 Suppose that $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated and satisfies

$$\limsup_{n\to\infty} \left|\mathcal{X}^{(f_n)}\right| > \aleph_0,$$

that is, for every $m \in \mathbb{N}$, there exists $n \geq m$ such that $\mathcal{X}^{(f_n)}$ is an uncountable set. We will show a contradiction using Corollary 5.2.1 for f_n and a certain C_n to be defined shortly.

First, we verify the condition of Corollary 5.2.1 for f_n if $\mathcal{X}^{(f_n)}$ is uncountable: we construct a distribution P such that $P_{(f_n(X))}$ is nonatomic. Let the range of f_n acting on $\mathcal{X}$ be denoted as $f_n(\mathcal{X})$, and for $z \in f_n(\mathcal{X})$ let the level set at value z be denoted as $\mathcal{X}_z^{(f_n)}$. Since the sets $\mathcal{X}^{(f_n)}$ are measurable, we can define $P(X)$ as follows:

$$P(f_n(X)) = \text{Unif}(f_n(\mathcal{X})); \quad P(X \mid f_n(X)) = \text{Unif}\left(\mathcal{X}_{f_n(X)}^{(f_n)}\right). \tag{D.10}$$

$P(X)$ along with any conditional probability function $P(Y \mid X)$ constitutes a valid probability distribution P. Further, from the construction, since $\mathcal{X}^{(f_n)}$ is uncountable, $P_{f_n(X)}$ is guaranteed to be nonatomic.

Next, since $\{f_n\}_{n\in\mathbb{N}}$ is asymptotically calibrated, by Corollary 5.1.1, one can construct a sequence of functions $\{C_n\}_{n\in\mathbb{N}}$ such that each C_n is a $(1-\alpha)$-CI with respect to f_n for any distribution Q, and

$$|C_n(f_n(X_{n+1}))| = o_Q(1).$$

Thus there exists a constant m such that for $n \geq m$ and any distribution Q,

$$\mathbb{E}_{Q^{n+1}} |C_n(f_n(X_{n+1}))| < 0.5 - \alpha. \tag{D.11}$$

However, since $\limsup_{n\to\infty} |\mathcal{X}^{(f_n)}| > \aleph_0$, there exists an $n \geq m$ such that $\mathcal{X}^{(f_n)}$ is uncountable. Hence the requirements of Corollary 5.2.1 are satisfied by $\widehat{C}_n$ and f_n: namely $\widehat{C}_n$ is a $(1-\alpha)$-CI with respect to f for all distributions P, and there exists a P such that $P_{f_n(X)}$ is nonatomic. Thus Corollary 5.2.1 yields that we can construct a distribution Q such that

$$\mathbb{E}_{Q^{n+1}} |C_n(f_n(X_{n+1}))| \geq 0.5 - \alpha,$$

which is a contradiction to (D.11). Hence our hypothesis that $\limsup_{n\to\infty} |\mathcal{X}^{(f_n)}| > \aleph_0$ must be false, concluding the proof. □

D.3 Proofs of Results in Section 5.4 (other than Section 5.4.4)

Proof of Theorem 5.4 Let $E_{\mathcal{B}(x)}$ the event that $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. On the event $E_{\mathcal{B}(x)}$, within each region $\mathcal{X}_b$, the number of point from the calibration set is known and the Y_i's in each bin represent independent Bernoulli random variables that share the same mean $\pi_b = \mathbb{E}[Y \mid X \in \mathcal{X}_b]$. Consider any fixed region $\mathcal{X}_b$, $b \in [B]$. Using Theorem D.3, we obtain that:

$$\mathbb{P}\left(|\pi_b - \widehat{\pi}_b| > \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3\ln(3B/\alpha)}{\widehat{s}_b} \;\middle|\; E_{\mathcal{B}(x)} \right) \leq \alpha/B.$$

Applying union bound across all regions of the sample-space partition, we get that:

$$\mathbb{P}\left(\forall b \in [B] :\ |\pi_b - \widehat{\pi}_b| \leq \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3\ln(3B/\alpha)}{\widehat{s}_b} \;\middle|\; E_{\mathcal{B}(x)} \right) \geq 1 - \alpha.$$

Because this is true for any $\mathcal{B}(x)$, we can marginalize to obtain the assertion of the theorem in unconditional form. □

Proof of Corollary 5.4.1 We show a calibration guarantee by using Theorem 5.1. Consider the scoring function as $\mathcal{B}$ with $\mathcal{Z} = [B]$. Then by Theorem 5.4, $C : [B] \to \mathcal{I}$ given by

$$C(b) = \left[\widehat{\pi}_b - \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3\ln(3B/\alpha)}{\widehat{s}_b}, \widehat{\pi}_b + \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3\ln(3B/\alpha)}{\widehat{s}_b} \right], \ b \in [B],$$

provides a $(1-\alpha)$-CI with respect to $\mathcal{B}$. Let $b^\star = \min_{b\in[B]} \widehat{s}_b$. To apply Theorem 5.4, we define

$$\varepsilon(\cdot) = \sup_{b\in[B]} |C(b)/2| = \sqrt{\frac{\widehat{V}_{b^\star} \ln(3B/\alpha)}{2\widehat{s}_{b^\star}}} + \frac{3\ln(3B/\alpha)}{2\widehat{s}_{b^\star}},$$

and the mid-point function m_C for C is given by $m_C(b) = \widehat{\pi}_b$. Applying Theorem 5.1 gives the first part of the result.

Next, suppose some bin b has $\mathbb{P}(\mathcal{B}(X) = b) = 0$. Then, a test point X_{n+1} almost surely does not belong to the bin, and the bin can be ignored for our calibration guarantee. Thus without loss of generality, suppose every $b \in [B]$ satisfies

$$\mathbb{P}(\mathcal{B}(X) = b) > 0.$$

Let $\min_{b\in[B]} \mathbb{P}(\mathcal{B}(X) = b) = \tau > 0$. Then for a fixed number of samples n, any particular bin b, and any constant $\alpha \in (0,1)$ we have by Hoeffding's inequality with probability $1 - \alpha/B$

$$\widehat{s}_b \geq n\tau - \sqrt{\frac{n\ln(B/\alpha)}{2}}.$$

Taking a union bound, we have with probability $1-\alpha$, simultaneously for every $b \in [B]$,

$$\widehat{s}_b \geq n\tau - \sqrt{\frac{n\ln(B/\alpha)}{2}} = \Omega(n),$$

and in particular $\widehat{s}_{b^\star} = \Omega(n)$ where $b^\star = \arg\min_{b\in[B]} \widehat{s}_b$. Thus by the first part of this corollary, f_n is ε_n calibrated where $\varepsilon_n = O(\sqrt{n^{-1}}) = o(1)$. This concludes the proof. □

Proof of Theorem 5.5 Denote $|\mathcal{D}^2_{cal}| = n$. Let $p_j = \mathbb{P}(g(X) \in I_j)$ be the true probability that a random point falls into partition $\mathcal{X}_j$. Assume c is such that we can use Lemma D.3.1 to guarantee that with probability at least $1 - \alpha/2$, uniform mass binning scheme is 2-well-balanced. Hence, with probability at least $1 - \alpha/2$:

$$\frac{1}{2B} \leq p_j \leq \frac{2}{B}, \ \forall j \in [B]. \tag{D.12}$$

Moreover, by Hoeffding's inequality we get that for any fixed region of sample-space partition, with probability at least $1 - \alpha/2B$,

$$\widehat{s}_j \geq np_j - \sqrt{\frac{n \ln(2B/\alpha)}{2}}. \tag{D.13}$$

Hence, by union bound across applied accross all regions and using (D.12), we get that with probability at least $1 - \alpha/2$:

$$\widehat{s}_{b^\star} \geq n/(2B) - \sqrt{\frac{n \ln(2B/\alpha)}{2}},$$

where the first term dominates asymptotically (for fixed B). Hence, we get that with probability at least $1 - \alpha$, $s_{b^\star} = \Omega\left(n/B\right)$. By invoking the result of Corollary 5.4.1 and observing that $\widehat{V}_b \leq 1$, we conclude that uniform mass binning is (ε, α) approximately calibrated with $\varepsilon(\cdot) = O(\sqrt{B \ln(B/\alpha)/n})$ as desired. This also leads to asymptotic calibration by Corollary 5.4.1. □

Proof of Theorem 5.6. The proof is based on the result for an empirical-Bernstein confidence sequences for bounded observations Howard et al. (2021). We condition on the event $E^{\infty}_{\mathcal{B}(x)}$ defined as $(\mathcal{B}(X_1), \mathcal{B}(X_1), \dots) = (\mathcal{B}(x_1), \mathcal{B}(x_2), \dots)$, that is the random variables denoting which partition the infinite stream of samples fall in (thus allowing our bound to hold for every possible value of n). On $E^{\infty}_{\mathcal{B}(x)}$, the label values within each partition of the sample-space partition represent independent Bernoulli random variable that share the same mean $\pi_b = \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right], b \in [B]$. Consequently, the bound obtained can be marginalized over $E^{\infty}_{\mathcal{B}(x)}$ to obtain the assertion of the theorem in unconditional form. Now we show the bound that applies conditionally on $E^{\infty}_{\mathcal{B}(x)}$.

Consider any fixed region of the sample-space partition $\mathcal{X}_b$ and corresponding points $\left\{\left(X_i^b, Y_i^b\right)\right\}_{i=1}^{\widehat{s}_b}$. Then $S_t = \left(\sum_{i=1}^t Y_i^b\right) - t\pi_b$ is a sub-exponential process with variance process:

$$\widehat{V}_t^+ = \sum_{i=1}^{t} \left(Y_i^b - \overline{Y}_{i-1}^b\right)^2.$$

Howard et al. (2020, Proposition 2) implies that S_t is also a sub-gamma process with variance process $\widehat{V}_t$ and the same scale $c = 1$. Since the theorem holds for any sub-exponential uniform boundary, we choose one based on analytical convenience. Recall definition of the polynomial stitching function

$$\mathcal{S}_\alpha(v) := \sqrt{k_1^2 v l(v) + k_2^2 c^2 l^2(v)} + k_2 c l(v), \quad \text{where} \begin{cases} l(v) := \ln h(\ln_\eta(v/m)) + \ln(l_0/\alpha), \\ k_1 := (\eta^{1/4} + \eta^{-1/4})/\sqrt{2}, \\ k_2 := (\sqrt{\eta} + 1)/\sqrt{2}. \end{cases}$$

where $l_0 = 1$ for the scalar case. Note that for $c > 0$ it holds that $\mathcal{S}_\alpha(v) \leq k_1\sqrt{v l(v)} + 2ck_2 l(v)$.

From Howard et al. (2021, Theorem 1), it follows that $u(v) = \mathcal{S}_\alpha(v \vee m)$ is a sub-gamma uniform boundary with scale c and crossing probability α. Applying Theorem D.2 with $h(k) \leftarrow (k+1)^s \zeta(s)$ where $\zeta(\cdot)$ is Riemann zeta

function and parameters $\eta \leftarrow e$, $s \leftarrow 1.4$, $c \leftarrow 1$, $m \leftarrow 1$ and $\alpha \leftarrow \alpha/(2B)$, yields that $k_2 \leq 1.88, k_1 \leq 1.46$ and $l(v) = 1.4 \cdot \ln\ln(ev) + \ln(2\zeta(1.4)B/\alpha)$. Since Theorem D.2 provides a bound that holds uniformly across time t, then it provides a guarantee for $t = \widehat{s}_b$, in particular. Hence, with probability at least $1 - \alpha/B$,

$$
\begin{aligned}
|\pi_b - \widehat{\pi}_b| &\leq \frac{1.46\sqrt{\widehat{V}_b^+ \cdot 1.4 \cdot \ln\ln\left(e\left(\widehat{V}_b^+ \vee 1\right)\right) + \ln(6.3B/\alpha)}}{\widehat{s}_b} + \frac{5.27 \cdot \ln\ln\left(e\left(\widehat{V}_b^+ \vee 1\right)\right) + 3.76\ln(6.3B/\alpha)}{\widehat{s}_b} \\
&\leq \frac{7\sqrt{\widehat{V}_b^+ \cdot \ln\ln\left(e\left(\widehat{V}_b^+ \vee 1\right)\right)} + 5.3\ln(6.3B/\alpha)}{\widehat{s}_b}.
\end{aligned}
$$

using that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ and $\ln\ln(ex) \leq \sqrt{x \ln\ln ex}$ for $x \geq 1$. Finally, we apply a union bound to get a guarantee that holds simultaneously for all regions of the sample-space partition. □

D.4 Calibration under Covariate Shift (including results in Section 5.4.4)

The results from Section 5.4.4 are proved in Appendix D.4.1 (Theorem 5.7) and D.4.3 (Proposition 6). To show Theorem 5.7, we first propose and analyze a slightly different estimator than (D.20) that is unbiased for $\pi_b^{(w)}$, but needs additional oracle access to the parameters $\{m_b\}_{b\in[B]}$ defined as

$$
m_b = \mathbb{P}_{P_X}(X \in \mathcal{X}_b) \,/\, \mathbb{P}_{\widetilde{P}_X}(X \in \mathcal{X}_b).
$$

m_b denotes the 'relative mass' of region $\mathcal{X}_b$. (For simplicity, we assume that $\mathbb{P}_{\widetilde{P}}(X \in \mathcal{X}_b) > 0$ for every b since otherwise the test-point almost surely does not belong to $\mathcal{X}_b$ and estimation in that bin is not relevant for a calibration guarantee.) We then show that m_b can be estimated using w, which would lead to the proposed estimator $\breve{\pi}_b^{(w)}$. First, we establish the following relationship between $\mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]$ and $\mathbb{E}_P[Y \mid X \in \mathcal{X}_b]$.

Proposition 9. *Under the covariate shift assumption, for any $b \in [B]$*

$$
\mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b] = m_b \cdot \mathbb{E}_P[w(X)Y \mid X \in \mathcal{X}_b].
$$

Proof. Observe that

$$
\frac{d\widetilde{P}(X \mid X \in \mathcal{X}_b)}{dP(X \mid X \in \mathcal{X}_b)} = \frac{d\widetilde{P}(X)}{dP(X)} \cdot \frac{\mathbb{P}_P(X \in \mathcal{X}_b)}{\mathbb{P}_{\widetilde{P}}(X \in \mathcal{X}_b)} = w(X) \cdot m_b.
$$

Thus we have,

$$\begin{aligned}
\mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right] &\overset{(1)}{=} \mathbb{E}_{\widetilde{P}}\left[\mathbb{E}_{\widetilde{P}}\left[Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(2)}{=} \mathbb{E}_{\widetilde{P}}\left[\mathbb{E}_{P}\left[Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(3)}{=} \mathbb{E}_{P}\left[\frac{d\widetilde{P}(X \mid X \in \mathcal{X}_b)}{dP(X \mid X \in \mathcal{X}_b)} \cdot \mathbb{E}_{P}\left[Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(4)}{=} m_b \cdot \mathbb{E}_{P}\left[w(X)\mathbb{E}_{P}\left[Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(5)}{=} m_b \cdot \mathbb{E}_{P}\left[\mathbb{E}_{P}\left[w(X)Y \mid X\right] \mid X \in \mathcal{X}_b\right] \\
&\overset{(6)}{=} m_b \cdot \mathbb{E}_{P}\left[w(X)Y \mid X \in \mathcal{X}_b\right],
\end{aligned}$$

where in (1) we use the tower rule, in (2) we use the covariate shift assumption, (3) can be seen by using the integral form of the expectation, (4) uses the observation at the beginning of the proof, (5) uses that $w(X)$ is a function of X and finally, (6) uses the tower rule. □

Let $\widehat{s}_b$ denote the number of calibration points from the source domain that belong to bin b. Given Proposition 9, a natural estimator for $\mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]$ is given by:

$$\widehat{\pi}_b^{(w)} := \frac{1}{\widehat{s}_b} \sum_{i:\mathcal{B}(X_i)=b} m_b w(X_i) Y_i. \tag{D.14}$$

Estimation properties of $\widehat{\pi}_b^{(w)}$ are given by the following theorem.

Theorem D.1. *Assume that* $\sup_x w(x) = U < \infty$*. For any* $\alpha \in (0,1)$*, with probability at least* $1-\alpha$*,*

$$\left|\widehat{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| \leq \sqrt{\frac{2\widehat{V}_b^{(w)} \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3 m_b U \ln(3B/\alpha)}{\widehat{s}_b}, \quad \textit{simultaneously for all } b \in [B],$$

where $\widehat{V}_b^{(w)} = \frac{1}{\widehat{s}_b}\sum_{i:\mathcal{B}(X_i)=b}(m_b w(X_i) Y_i - \widehat{\pi}_b^{(w)})^2$.

The proof is given in Appendix D.4.2. Next, we discuss a way of estimating m_b using likelihood ratio w instead of relying on oracle access. Observe that

$$\frac{d\widetilde{P}(X \mid X \in \mathcal{X}_b)}{dP(X \mid X \in \mathcal{X}_b)} = \frac{d\widetilde{P}(X)}{dP(X)} \cdot \frac{\mathbb{P}_P\left(X \in \mathcal{X}_b\right)}{\mathbb{P}_{\widetilde{P}}(X \in \mathcal{X}_b)} = w(X) \cdot m_b.$$

Thus we have,

$$\mathbb{E}_P\left[w(X) \mid X \in \mathcal{X}_b\right] = m_b^{-1}\mathbb{E}_P\left[\frac{d\widetilde{P}(X \mid X \in \mathcal{X}_b)}{dP(X \mid X \in \mathcal{X}_b)} \mid X \in \mathcal{X}_b\right] = m_b^{-1}, \tag{D.15}$$

which suggests a possible estimator for m_b given by

$$\widehat{m}_b = \left(\frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}{\widehat{s}_b} \right)^{-1}, \quad b \in [B]. \tag{D.16}$$

On substituting this estimate for m_b in (D.14), we get a new estimator

$$\frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)},$$

which is exactly $\breve{\pi}_b^{(w)}$. With this observation, we now prove Theorem 5.7.

D.4.1 Proof of Theorem 5.7

Let us define $r_b := 1/m_b$ and

$$\widehat{r}_b = \frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}{\widehat{s}_b}. \tag{D.17}$$

Step 1 (Uniform lower bound for $\widehat{s}_b$). Since the regions of the sample-space partition were constructed using uniform-mass binning, the guarantee of Theorem 5.5 holds. Precisely, we have that with probability at least $1 - \alpha/3$, simultaneously for every $b \in [B]$,

$$\widehat{s}_b \geq \frac{n}{2B} - \sqrt{\frac{n \ln(6B/\alpha)}{2}}.$$

Step 2 (Approximating r_b). Observe that the estimator (D.17) is an average of $\widehat{s}_b$ random variables bounded by the interval $[0, U]$. Let $E_{\mathcal{B}(x)}$ be the event that $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. On the event $E_{\mathcal{B}(x)}$, within each region $\mathcal{X}_b$, the number of point from the calibration set is known and the Y_i's in each bin represent independent Bernoulli random variables that share the same mean $\mathbb{E}[w(X) \mid X \in \mathcal{X}_b]$. Consider any fixed region $\mathcal{X}_b$, $b \in [B]$. By Hoeffding's inequality, it holds that

$$\mathbb{P}\left(|r_b - \widehat{r}_b| > \sqrt{\frac{U^2 \ln(6B/\alpha)}{2\widehat{s}_b}} \;\middle|\; E_{\mathcal{B}(x)} \right) \leq \alpha/(3B).$$

Applying union bound across all regions of the sample-space partition, we get that:

$$\mathbb{P}\left(\exists b \in [B] : \; |r_b - \widehat{r}_b| > \sqrt{\frac{U^2 \ln(6B/\alpha)}{2\widehat{s}_b}} \;\middle|\; E_{\mathcal{B}(x)} \right) \leq \alpha/3.$$

Because this is true for any $\mathcal{B}(x)$, we can marginalize to obtain that with probability at least $1-\alpha/3$,

$$\forall b \in [B],\ |r_b - \widehat{r}_b| \leq \sqrt{\frac{U^2 \ln(6B/\alpha)}{2\widehat{s}_b}}. \tag{D.18}$$

Step 3 (Going from r_b to m_b). Define $r^\star = \min_{b\in[B]} \mathbb{E}\left[w(X) \mid X \in \mathcal{X}_b\right]$. Suppose $\forall b \in [B]$, $|r_b - \widehat{r}_b| \leq \varepsilon$ and $\varepsilon < r^\star/2$. Then, we have with probability at least $1-\alpha/3$:

$$|m_b - \widehat{m}_b| = \left|\frac{1}{r_b} - \frac{1}{\widehat{r}_b}\right| = \left|\frac{r_b - \widehat{r}_b}{r_b \cdot \widehat{r}_b}\right| \leq \frac{\varepsilon}{r_b^2|1-\varepsilon/r_b|} \leq \frac{2\varepsilon}{r_b^2} = 2m_b^2\varepsilon, \quad \forall b \in [B]. \tag{D.19}$$

We now set $\varepsilon = \sqrt{\frac{U^2\ln(6B/\alpha)}{2\widehat{s}_b}}$ as specified in equation (D.18) and verify that $\varepsilon < r^\star/2$.

- First, from step 1, with probability at least $1-\alpha/3$, $\widehat{s}_{b^\star} = \Omega(n/B)$ and thus $\widehat{s}_b = \Omega(n/B)$ for every $b \in [B]$.
- By the condition in the theorem statement, for every $b \in [B]$,

$$\varepsilon = \sqrt{\frac{U^2\ln(6B/\alpha)}{2\widehat{s}_b}} = O\left(\sqrt{\frac{U^2 B\ln(6B/\alpha)}{n}}\right) = O\left(\sqrt{\frac{U^2 B\ln(6B/\alpha)}{\left(\frac{U^2 B\ln(6B/\alpha)}{L^2}\right)}}\right) = O\left(L\right).$$

  Finally recall that $L \leq r^\star$. Thus we can pick c in the theorem statement to be large enough such that $\varepsilon < L/2 \leq r^\star/2$.

Thus for $\varepsilon = \sqrt{\frac{U^2\ln(6B/\alpha)}{2\widehat{s}_b}}$, by a union bound over the event in (D.18) and step 1, the conditions for (D.19) are satisfied with probability at least $1-2\alpha/3$. Hence we have for some large enough constant $c > 0$,

$$|m_b - \widehat{m}_b| \leq c m_b^2 \cdot \sqrt{\frac{U^2 B\ln(6B/\alpha)}{2n}} \leq c \cdot \frac{U}{L^2}\sqrt{\frac{B\ln(6B/\alpha)}{2n}}.$$

The final inequality holds by observing that $m_b \leq 1/L$ which follows from relationship (D.15) and the assumption that $\inf_x w(x) \geq L$.

Step 4 (Computing the final deviation inequality for $\breve{\pi}_b^{(w)}$). Recall the definitions of the two estimators:

$$\widehat{\pi}_b^{(w)} := \frac{1}{\widehat{s}_b}\sum_{i:\mathcal{B}(X_i)=b} m_b w(X_i) Y_i,$$

and

$$\breve{\pi}_b^{(w)} := \frac{1}{\widehat{s}_b}\sum_{i:\mathcal{B}(X_i)=b} \widehat{m}_b w(X_i) Y_i,$$

which differ by replacing m_b by its estimator $\widehat{m}_b$ defined in (D.16). By triangle inequality,

$$|\breve{\pi}_b - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]| \leq \left|\breve{\pi}_b^{(w)} - \widehat{\pi}_b^{(w)}\right| + \left|\widehat{\pi}_b^{(w)} - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]\right|.$$

Theorem D.1 bounds the term $\left|\widehat{\pi}_b^{(w)} - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]\right|$ with high probability. In the proof of Theorem D.1, we can replace the empirical Bernstein's inequality by Hoeffding's inequality to obtain with probability at least $1 - \alpha/3$,

$$\left|\widehat{\pi}_b^{(w)} - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]\right| \leq \sqrt{\frac{U^2 \ln(6B/\alpha)}{2\widehat{s}_b}} \leq \left(\frac{U}{L}\right)^2 \sqrt{\frac{\ln(6B/\alpha)}{2\widehat{s}_b}},$$

simultaneously for all $b \in [B]$ (the last inequality follows since $L \leq 1 \leq U$). To bound $\left|\widehat{\pi}_b^{(w)} - \breve{\pi}_b^{(w)}\right|$, first note that:

$$\begin{aligned}
\left|\widehat{\pi}_b^{(w)} - \breve{\pi}_b^{(w)}\right| &= \left|\frac{1}{\widehat{s}_b} \sum_{i:\mathcal{B}(X_i)=b} (\widehat{m}_b - m_b)\, w(X_i) Y_i\right| \\
&\leq U \cdot \left|\frac{1}{\widehat{s}_b} \sum_{i:\mathcal{B}(X_i)=b} (\widehat{m}_b - m_b)\right| \\
&= U \cdot |\widehat{m}_b - m_b|\,.
\end{aligned}$$

Then we use the results from steps 1 and 3 to conclude that with probability at least $1 - 2\alpha/3$,

$$\left|\breve{\pi}_b^{(w)} - \widehat{\pi}_b^{(w)}\right| \leq c \cdot \left(\frac{U}{L}\right)^2 \sqrt{\frac{B \ln(6B/\alpha)}{2n}}, \text{ and } \widehat{s}_b \geq n/B - \sqrt{\frac{n \ln(6B/\alpha)}{2}}.$$

simultaneously for all $b \in [B]$. Thus by union bound, we get that it holds with probability at least $1 - \alpha$,

$$|\breve{\pi}_b - \mathbb{E}\left[Y \mid X \in \mathcal{X}_b\right]| \leq c \cdot \left(\frac{U}{L}\right)^2 \sqrt{\frac{B \ln(6B/\alpha)}{2n}},$$

simultaneously for all $b \in [B]$ and large enough absolute constant $c > 0$. This concludes the proof. □

D.4.2 Proof of Theorem D.1

Consider the event $E_{\mathcal{B}(x)}$ defined as $(\mathcal{B}(X_1), \ldots, \mathcal{B}(X_n)) = (\mathcal{B}(x_1), \ldots, \mathcal{B}(x_n))$. Conditioned on $E_{\mathcal{B}(x)}$, since $\sup_x w(x) \leq U$, we get that $\widehat{\pi}_b^{(w)}$ is an average of independent nonnegative random variables $m_b w(X_i) Y_i$ that are bounded by $m_b U$ and share the same mean $m_b\, \mathbb{E}_P\left[w(X)Y \mid X \in \mathcal{X}_b\right] = \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]$ (by Proposition 9).Using Theorem D.3 for a fixed $b \in [B]$, we obtain:

$$\mathbb{P}\left(\left|\widehat{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| > \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3 m_b U \ln(3B/\alpha)}{\widehat{s}_b} \;\Big|\; E_{\mathcal{B}(x)}\right) \leq \alpha/B.$$

Applying a union bound over all $b \in [B]$, we get:

$$\mathbb{P}\left(\forall b \in [B] : \left|\widehat{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| \leq \sqrt{\frac{2\widehat{V}_b \ln(3B/\alpha)}{\widehat{s}_b}} + \frac{3m_b U \ln(3B/\alpha)}{\widehat{s}_b} \;\middle|\; E_{\mathcal{B}(x)}\right) \geq 1 - \alpha.$$

Because this is true for any $\mathcal{B}(x)$, we can marginalize to obtain the assertion of the theorem in unconditional form. □

D.4.3 Proof of Proposition 6

Fix any $\alpha \in (0, 1)$. For any $k \in \mathbb{N}$ observe that by triangle inequality,

$$\left|\breve{\pi}_b^{(\widehat{w}_k)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| \leq \left|\breve{\pi}_b^{(w)} - \mathbb{E}_{\widetilde{P}}\left[Y \mid X \in \mathcal{X}_b\right]\right| + \left|\breve{\pi}_b^{(w)} - \breve{\pi}_b^{(\widehat{w}_k)}\right|.$$

Consider any $\varepsilon > 0$. Note that by Theorem 5.7, there exists sufficiently large n such that the first term is larger than $\varepsilon/2$ with probability at most $\alpha/2$ simultaneously for all $b \in [B]$. Hence, it suffices to show that there exists a large enough k such that the probability of the second term exceeding $\varepsilon/2$ is at most $\alpha/2$ simultaneously for all $b \in [B]$. While analyzing the second term, we treat n as a constant while leveraging the consistency of $\widehat{w}_k$ as $k \to \infty$. For simplicity, denote $\Delta_k = \sup_x |w(x) - \widehat{w}_k(x)|$. Then for any $b \in [B]$:

$$\begin{aligned}
\left|\breve{\pi}_b^{(w)} - \breve{\pi}_b^{(\widehat{w}_k)}\right| &= \left|\frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)} - \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}\right| \\
&\overset{(1)}{\leq} \left|\frac{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)} - \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}\right| \\
&\quad + \left|\frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)} - \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}\right| \\
&\overset{(2)}{\leq} n \cdot \Delta_k \cdot \left|\frac{1}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)}\right| \\
&\quad + \left|\frac{1}{\sum_{i:\mathcal{B}(X_i)=b} w(X_i)} - \frac{1}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)}\right| \left|\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}_k(X_i)Y_i\right| \\
&\overset{(3)}{\leq} \frac{n}{L} \cdot \Delta_k + \left(\frac{n \cdot \Delta_k}{(L - \Delta_k)L}\right) \cdot \left((U + \Delta_k) \cdot n\right),
\end{aligned}$$

where (1) is due to the triangle inequality, (2) is due to the facts that the number of points in any bin is at most n and that absolute difference between $\widehat{w}$ and w is at most Δ_k, (3) combines the aforementioned reasons in (2) and the assumptions: $L \leq \inf_x w(x) \leq \sup_x w(x) \leq U$. Since $\Delta_k \xrightarrow{P} 0$, clearly there exists a large enough k such that:

$$\mathbb{P}\left(\left|\breve{\pi}_b^{(w)} - \breve{\pi}_b^{(\widehat{w}_k)}\right| \geq \varepsilon/2\right) \leq \alpha/2.$$

Thus we conclude that $\breve{\pi}_b^{(\widehat{w}_k)}$ is asymptotically calibrated at level α. □

D.4.4 Preliminary Simulations

This section is structured as follows. We first describe the overall procedure for calibration under covariate shift. The finite-sample calibration guarantee of Theorem 5.7 holds for oracle w whereas in our experiments we will estimate w; to assess the loss in calibration due to this approximation, we introduce some standard techniques used in literature. The preliminary experiments are performed with simulated data which are described after this. Finally, we propose a modified estimator $\widetilde{\pi}_b^{(\widehat{w})}$ of $\mathbb{E}_{\widetilde{P}}[Y \mid X \in \mathcal{X}_b]$ which appears natural but has poor performance in practice.

Procedure. We describe how to construct approximately calibrated predictions practically. This involves approximating the importance weights w and the relatives mass terms $\{m_b\}_{b\in[B]}$. The summarized calibration procedure consists of the following steps:

1. Split the calibration set into two parts and use the first to perform *uniform mass* binning
2. Given unlabeled examples from both source and target domain, estimate $\widehat{w}$. The unconstrained Least-Squares Importance Fitting (uLSIF) procedure Kanamori et al. (2009) is used for this.
3. Compute for every $b \in [B]$, the estimator as per (5.16), replacing w with $\widehat{w}$:

$$\breve{\pi}_b^{(\widehat{w})} := \frac{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}(X_i) Y_i}{\sum_{i:\mathcal{B}(X_i)=b} \widehat{w}(X_i)}. \tag{D.20}$$

4. On a new test point from the target distribution, output the calibrated estimate $\breve{\pi}^{(\widehat{w})}_{\mathcal{B}(X_{n+1})}$.

Assessment through reliability diagrams and ECE. Given a test set (from the target distribution) of size m: $\{(X_i', Y_i')\}_{i\in[m]}$ and a function $g : \mathcal{X} \to [0,1]$ that outputs approximately calibrated probabilities, we consider the reliability diagram to estimate its calibration properties. A reliability diagram is constructed using splitting the unit interval $[0,1]$ into non-overlapping intervals $\{I_b\}_{b\in[B']}$ for some B' as

$$I_i = \left[\frac{i-1}{B'}, \frac{i}{B'}\right), \; i = 1, \ldots, B'-1 \;\text{ and }\; I_{B'} = \left[\frac{B'-1}{B'}, 1\right].$$

Let $\mathcal{B}' : [0,1] \to [B']$ denote the binning function that corresponds to this binning. We then compute the following quantities for each bin $b \in [B']$:

$$\begin{aligned} \mathrm{FP}(I_b) &= \frac{\sum_{i:\mathcal{B}'(X_i')=b} Y_i'}{|\{i : \mathcal{B}'(X_i') = b\}|} && \text{(fraction of positives in a bin),} \\ \mathrm{MP}(I_b) &= \frac{\sum_{i:\mathcal{B}'(X_i')=b} g(X_i')}{|\{i : \mathcal{B}'(X_i') = b\}|} && \text{(mean predicted probability in a bin).} \end{aligned}$$

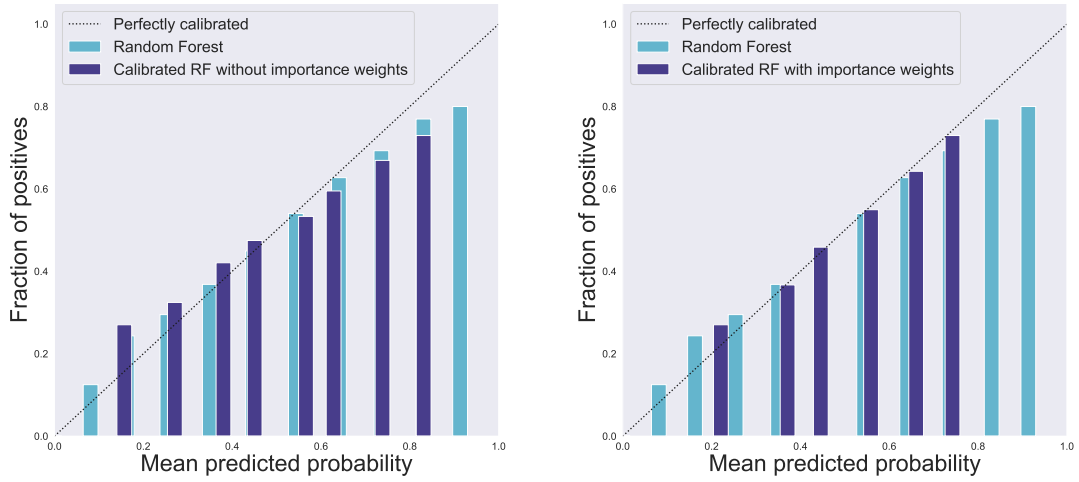
If g is perfectly calibrated, the reliability diagram is diagonal. Define the proportion of points that fall into various bins as:

$$\widehat{p}_b = \frac{|\{i : \mathcal{B}'(X_i') = b\}|}{m}, \quad b \in [B'].$$

Then ECE (or ℓ_1-ECE) is defined as:

$$\mathrm{ECE}(g) = \sum_{b \in [B']} \widehat{p}_b \cdot |\mathrm{MP}(I_b) - \mathrm{FP}(I_b)| \,.$$

ECE can also be defined in the ℓ_p sense and for multiclass problems but we limit our attention to the ℓ_1-ECE for binary problems.



(a) (b)

Figure D.1: In Figure D.1a uncalibrated Random Forest (ECE ≈ 0.023) is compared with calibration that does not take the covariate shift into account (ECE ≈ 0.047). In Figure D.1b uncalibrated Random Forest is compared with calibration that takes the covariate shift into account (ECE ≈ 0.015).

Simulations with synthetic data. We illustrate the performance of our proposed estimator (5.16) using the following simulated example, for which we can explicitly control the covariate shift. Consider the following data generation pipeline: for the source domain each component of the feature vector is drawn from Beta(α, β) where $\alpha = \beta = 1$, which corresponds to uniform draws from the unit cube. For the target distribution each component can be drawn independently from Beta(α', β'). If the dimension is d, the true likelihood ratio is given as

$$w(x) = \frac{d\widetilde{P}_X(x)}{dP_X(x)} = \frac{B^d(\alpha;\beta)}{B^d(\alpha';\beta')} \prod_{i=1}^{d} \frac{(x_{(i)})^{\alpha'-1}(1-x_{(i)})^{\beta'-1}}{(x_{(i)})^{\alpha-1}(1-x_{(i)})^{\beta-1}},$$

where $x_{(i)}$ are the coordinates of feature vector x. We set $d = 3$ and $\alpha' = 2, \beta' = 1$ so that $w(x) = 8 \cdot x_{(1)}x_{(2)}x_{(3)}$. The labels for both source and target distributions are assigned according to:

$$\mathbb{P}(Y = 1 \mid X = x) = \frac{1}{2}\left(1 + \sin\left(\omega\left(x_{(1)}^2 + x_{(2)}^2 + x_{(3)}^2\right)\right)\right),$$

for $\omega = 20$. As the underlying classifier we use a Random Forest with 100 trees (from `sklearn`). 14700 data points were used to train the underlying Random Forest classifier, 2000 data points from both source and target were used for the estimation of importance weights. The parameters σ and λ for uLSIF were tuned by leave-one-out cross-validation: we considered 25 equally spaced values on a log-scale in range $(10^{-2}, 10^2)$ for σ and 100 equally spaced values on a log-scale in range $(10^{-3}, 10^3)$ for λ. Uniform mass binning was performed with 10 bins and 1940 data points from the source domain were used to estimate the quantiles. 7840 source data points were used for the calibration and finally, 28000 data points from the target domain were used for evaluation purposes. We note that this simulation is a 'proof-of-concept'; the sample sizes we used are not necessarily optimal can presumably be improved.

We compare the unweighted estimator (5.12) which corresponds to weighing points in each bin equally as we would do if there was no covariate shift, and the estimator (5.16) that uses an estimate of w to account for covariate shift. The reliability diagrams are presented in Figure D.1, with the ECE reported in the caption. For the ECE estimation and reliability diagrams, we used $B' = 10$.

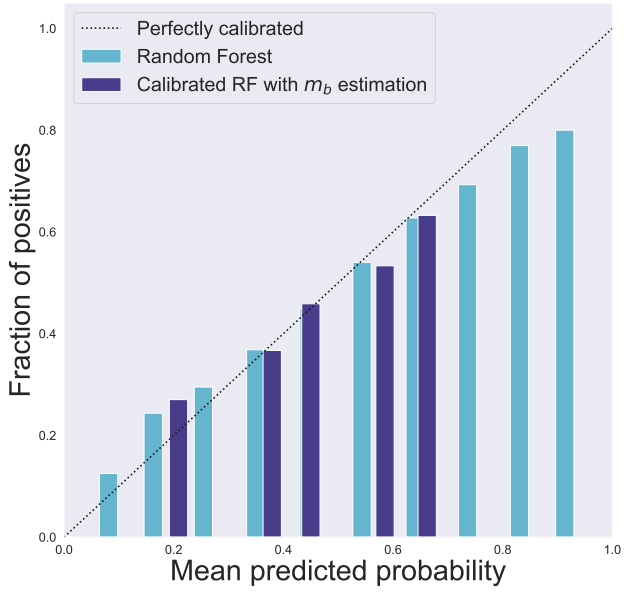


Figure D.2: Calibration of Random Forest with m_b estimated as per equation (D.16) (ECE ≈ 0.035).

Alternative estimator for m_b. Estimator (D.16) is one way of estimating m_b using the w values, that leads to (5.16). However, there exists another natural estimator which we propose and show some preliminary empirical results for. Suppose we have access to additional unlabeled data from the source and target domains ($\{X_i^s\}_{i\in[n_s]}$, and $\{X_i^t\}_{i\in[n_t]}$

respectively). From the definition of $m_b = \mathbb{P}_{P_X}(X \in \mathcal{X}_b)/\mathbb{P}_{\widetilde{P}_X}(X \in \mathcal{X}_b)$, a natural estimator is,

$$\widehat{m}_b = \frac{\frac{1}{n_s}\left|\{i \in [n_s] : \mathcal{B}(X_i^s) = b\}\right|}{\frac{1}{n_t}\left|\{i \in [n_t] : \mathcal{B}(X_i^t) = b\}\right|}, \quad b \in [B]. \tag{D.21}$$

In this case, the estimator (D.14) reduces to:

$$\widetilde{\pi}_b^{(\widehat{w})} = \frac{\widehat{m}_b}{\widehat{s}_b} \sum_{i:\mathcal{B}(X_i)=b} \widehat{w}(X_i) Y_i.$$

We show experimental results with this estimation procedure. We used 8500 data points from the source domain and 8000 points from the target domain to compute (D.21). The reliability diagram and ECE with this estimator is reported in Figure D.2. On our simulated dataset, we observe that the estimators $\widetilde{\pi}_b^{(\widehat{w})}$ perform significantly worse than the estimators $\breve{\pi}_b^{(\widehat{w})}$. While this is only a single experimental setup, we outline some drawbacks of this estimation method that may lead to poor performance in general.

1. $\widetilde{\pi}_b^{(\widehat{w})}$ requires access to additional unlabeled data from the source and target domains without leading to increase in performance.

2. The denominator of $\widehat{m}_b$ could be badly behaved if the number of points from the target domain in bin b are small. We could perform uniform-mass binning on the target domain to avoid this, but in this case $\widehat{s}_b$ may be small which would lead to the estimator $\widetilde{\pi}_b^{(\widehat{w})}$ performing poorly.

Our overall recommendation through these preliminary experiments is to use the estimator $\widehat{\pi}_b^{(\widehat{w})}$ as proposed in Section 5.4.4 instead of $\widetilde{\pi}_b^{(\widehat{w})}$.

D.5 Auxiliary results

D.5.1 Concentration Inequalities

Theorem D.2 (Howard et al. (2021), Theorem 4)**.** *Suppose $Z_t \in [a, b]$ a.s. for all t. Let $(\widehat{Z}_t)$ be any $[a, b]$-valued predictable sequence, and let u be any sub-exponential uniform boundary with crossing probability α for scale $c = b - a$. Then:*

$$\mathbb{P}\left(\forall t \geq 1 : \left|\overline{Z}_t - \mu_t\right| < \frac{u\left(\sum_{i=1}^t \left(Z_i - \widehat{Z}_i\right)^2\right)}{t}\right) \geq 1 - 2\alpha.$$

Theorem D.3 (Partial statement of Audibert et al. (2007), Theorem 1)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking their values in $[0, b]$. Let $\mu = \mathbb{E}[X_1]$ be their common expected value. Consider the empirical expectation $\overline{X}_n$*

and variance V_n defined respectively by

$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}, \quad \textit{and} \quad V_t = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)^2}{n}.$$

Then for any and $x > 0$, with probability at least $1 - 3e^{-x}$,

$$\left|\overline{X}_n - \mu\right| \leq \sqrt{\frac{2V_n x}{n}} + \frac{3bx}{n}.$$

D.5.2 Uniform-mass Binning

Kumar et al. (2019) defined well-balanced binning and showed that uniform mass-binning is well-balanced.

Definition 13 (Well-balanced binning)**.** *A binning scheme $\mathcal{B}$ of size B is β-well-balanced ($\beta \geq 1$) for some classifier g if*

$$\frac{1}{\beta B} \leq \mathbb{P}\left(g(X) \in I_b\right) \leq \frac{\beta}{B},$$

simultaneously for all $b \in [B]$.

To perform uniform-mass binning labeled examples are required at the stage of training the base classifier $g(\cdot)$. We denote this data as $\mathcal{D}^1_{\text{cal}}$. Procedures based on uniform-mass binning are well-balanced if $\left|\mathcal{D}^1_{\text{cal}}\right|$ is sufficiently large.

Lemma D.3.1 (Kumar et al. (2019), Lemma 4.3)**.** *For a universal constant $c > 0$, if $\left|\mathcal{D}^1_{cal}\right| \geq cB\ln(B/\alpha)$, then with probability at least $1 - \alpha$, the uniform mass binning scheme $\mathcal{B}$ is 2-well-balanced.*

The calibration guarantees in Section 5.4 depend on the minimum number of training points $\widehat{s}_{b^\star}$ in any bin. Uniform mass-binning guarantees that $\widehat{s}_{b^\star} = \Omega(n/B)$. This is used in the proof of Theorem 5.5.

Appendix E

Additional Results for Chapter 6

E.1 Importance Weights Estimation under Label Shift

Below we provide details about importance weights estimation procedures which are relevant mainly to Sections 6.2.2 and 6.3.2 of the paper. Estimation of the importance weights is performed using a held-out labeled set from the source distribution and an unlabeled set from the target distribution. Procedures, such as BBSE (Lipton et al., 2018) or RLLS (Azizzadenesheli et al., 2019), are based on estimation of the confusion matrix and yield consistent importance weights estimators with quantifiable estimation error under relatively mild assumptions. First, given a black-box predictor $f : \mathcal{X} \to \Delta_K$, define the corresponding expected confusion matrix $C_P(f) \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$:

$$[C_P(f)]_{ij} := \mathbb{E}_P \left[\mathbb{1}\left\{ \arg\max_k f_k(X) = i \right\} \cdot \mathbb{1}\left\{ Y = j \right\} \right].$$

We assume that

(A1) for every label $y \in \mathcal{Y}$, it holds that $q(y) > 0 \implies p(y) > 0$,

(A2) expected confusion matrix $C_P(f)$ is full-rank.

Assumption (A1) states that target label distribution is absolutely continuous with respect to the source. Indeed, reasoning properly about a class in the target domain which is not represented in the source domain is not possible. Assumption (A2) simply represents an identifiability condition. Lipton et al. (2018) show that under label shift assumption: $\mathbb{P}_Q\left(f(X) = i\right) = \sum_{j\in\mathcal{Y}} \left[C_P(f)\right]_{ij} w(j)$, or in matrix-vector notation:

$$\mu = C_P(f) w.$$

where $\mu \in \mathbb{R}^{|\mathcal{Y}|} : \mu_i = \mathbb{P}_Q\left(f(X) = i\right)$. BBSE is a simple plug-in procedure, which yields the following estimator of the importance weights:

$$\begin{aligned} \widehat{w} = & \ \widehat{C}^{-1}\, \widehat{\mu}, \\ \text{where} \quad \widehat{C}_{ij} = & \ \frac{1}{m} \sum_{p=1}^{m} \mathbb{1}\left\{f(X_p^s) = i \text{ and } Y_p^s = j\right\}, \\ \widehat{\mu}_i = & \ \frac{1}{l} \sum_{p=1}^{l} \mathbb{1}\left\{f(X_p^t) = i\right\}, \end{aligned}$$

where $\{(X_i^s, Y_i^s)\}_{i=1}^m$ is a labeled dataset from the source distribution and $\{(X_i^t)\}_{i=1}^l$ is unlabeled data from the target distribution. BBSE-hard described above can be trivially modified to the whole probability distribution output of f which is referred to as BBSE-soft procedure. Under aforementioned assumptions, Lipton et al. (2018) establish results with respect to consistency of BBSE and corresponding convergence rates.

A well-known alternative approach to directly estimate the importance weights which performs well in practice is MLLS (Saerens et al., 2002) and its recent variations that combine it with preceding calibration on the source domain (Alexandari et al., 2020). We refer the reader to Garg et al. (2020) for the theoretical analysis of MLLS and a detailed overview of the results for the importance weights estimation under label shift. For all simulations in this work we use BBSE-soft procedure motivated simply by its satisfactory empirical performance throughout all of the simulations we performed. Our modular approach to UQ allows to replace BBSE with any alternative choice.

E.2 Conformal Classification

Below, Section E.2.1 includes details about the tie-breaking rules for the oracle prediction sets, Section E.2.2 includes a discussion regarding the role of randomization for conformal classification, Section E.2.3 includes all necessary proofs for Sections 6.2.1 and 6.2.2 and Section E.2.4 includes details about the simulation on a real dataset mentioned in Section 6.2.2.

E.2.1 Tie-breaking RRules for the Oracle Prediction Set

In practice, when an estimator $\widehat{\pi}_y(x)$ is used in place of $\pi_y(x)$, one does not expect ties to be present but for completeness it is important to consider such scenario in the oracle setting. First, note that for any $\alpha \in (0, 1)$, the oracle prediction set clearly never include labels $y \in \mathcal{Y} : \pi_y(x) = 0$. Now, presence of ties can lead to a conservative prediction set for some $x \in \mathcal{X}$ if there is a subset of class labels $S(x) \subseteq \mathcal{Y}$ of size $L = |S(x)| > 1$, such that $\forall y, y' \in S(x) : \pi_y(x) = \pi_{y'}(x) > 0$ and

$$\begin{cases} \mathbb{P}\left(Y \in C_\alpha^{\text{oracle}}(X) \backslash S(X) \mid X = x\right) < 1 - \alpha, \\ \mathbb{P}\left(Y \in C_\alpha^{\text{oracle}}(X) \mid X = x\right) \geq (1 - \alpha). \end{cases}$$

In the oracle case ties can be broken arbitrarily in order to preserve the conditional coverage. One option is to break ties randomly, i.e. one can fix a random permutation of labels in $S(x)$: $\widetilde{y}_{i_1}, \ldots, \widetilde{y}_{i_l}$, and output a smaller oracle prediction set:

$$C_\alpha^{\text{oracle,new}}(X) := \left(C_\alpha^{\text{oracle}}(X) \backslash S(X)\right) \cup \{\widetilde{y}_{i_1}, \ldots, \widetilde{y}_{i_{l^\star}}\},$$

where $l^\star$ is the smallest index in $\{1, \ldots, l\}$ such that

$$\mathbb{P}\left(Y \in C_\alpha^{\text{oracle}}(X) \backslash S(X) \mid X = x\right) + \sum_{k=1}^{l^\star} \pi_{i_k}(x) \geq 1 - \alpha.$$

E.2.2 Note on Randomization and Conditional Coverage

As the number of works on conformal classification has seen a recent spurt, it is important to understand what exactly might be the benefits of using one nested sequence over another. For example, Angelopoulos et al. (2021) state in their Appendix B that "randomization is of little practical importance, since... output by the randomized procedure will differ from that of the non-randomized procedure by at most one element". However, we do not quite agree with their sentiment about it being of little practical importance for the following reason. While their observation is indeed accurate in the oracle setting, there is a noticeable difference in the empirical conditional coverage when the nested sequences are conformalized in practice (non-oracle setting). Roughly speaking, randomized scores better handle the heterogeneity of the conditional distribution of the response variable across the sample space. Note that this type of randomization has a different role from that of a randomized conformal p-value Vovk et al. (2005) which aims to improve possibly conservative marginal coverage. We believe that the reasoning below complements the one given in Romano et al. (2020) and, in particular, might help an unfamiliar reader to gain some useful insights (as well as arguably having simpler notation). For completeness, we start with an example of randomization in action. Consider a binary classification problem: $\mathcal{Y} = \{0, 1\}$, and fix target miscoverage level $\alpha = 0.05$. Now, assume that for some $x \in \mathcal{X}$:

- $\pi_0(x) = 0.99$, $\pi_1(x) = 0.01$. Then with probability $95/99$, we have $\tilde{C}_\alpha^{\text{oracle}}(x, u) = \{0\}$ and $\tilde{C}_\alpha^{\text{oracle}}(x, u) = \{\emptyset\}$ otherwise.
- $\pi_0(x) = 0.9$, $\pi_1(x) = 0.1$. Then with probability $1/2$, $\tilde{C}^{\text{oracle}}(x, u) = \{0, 1\}$ and $\tilde{C}^{\text{oracle}}(x, u) = \{0\}$ otherwise.

First, consider the marginal coverage of conformal prediction sets in the "null" case when $\widehat{\pi} \equiv \pi$. The marginal coverage guarantee of conformal prediction sets is due to Lemma E.1.1 which states a classic result for quantiles of exchangeable random variables and is tight when these variables are almost surely distinct. In the non-randomized setting for any point (X, Y), the corresponding non-conformity score are given by $\rho_Y(X; \pi)$. Such form might suggest that the marginal coverage could be conservative due to possible ties as whenever the predicted most likely

label appears to be the correct one, it holds that $\rho_Y(X;\pi) = 0$. However, if ties among non-conformity scores are present, they would typically occur only between zero-valued scores, and thus in a reasonable classification setup one should expect the marginal coverage to be tight even for non-randomized nested sequence as the calibrated threshold would typically be nonzero.

Next, before reasoning about conditional coverage of conformal sets, recall that the conditional distribution of the response is discrete in classification setting, and thus even in the null case it is hard to reason meaningfully about the distribution of non-conformity scores $\rho_Y(X;\pi)$. However, Romano et al. (2020) noticed that if randomization (6.4) is used, then it becomes possible to do at least in the null case. If $\widehat{\pi} \equiv \pi$, it is trivial to see the distribution of corresponding non-conformity scores $\rho_Y(X;\pi)+U\cdot\pi(X)$ is uniform conditional on X. Then, as the authors conjecture, it is intuitive that conformal prediction sets would recover the oracle ones under some consistency assumptions for $\widehat{\pi}$.

However, randomization is also performed when the prediction set is a singleton containing the most likely label only, and thus might yield non-interpretable and non-actionable empty prediction sets being purely the consequence of deploying randomization. Thus one might consider abstaining from dropping a label from the prediction set whenever it forms a singleton and perform randomization if and only if the oracle prediction set contains more than one label. While that decision can be embedded into either prediction step only or calibration step as well, we state explicitly that it should be done at the prediction step only for the aforementioned reasons.

Consider the binary toy example from Section 6.3.2 with focus on the source distribution only. As the true class-posterior probability $\pi_1^P(x)$ is known, we construct the non-randomized oracle prediction set C^{oracle} and compare it visually with the randomized version $\widetilde{C}^{\text{oracle}}$ on Figures E.1a and E.1b where randomization demonstrates desired behavior.

Consequently, we consider conformal prediction sets based on non-randomized sequence:

$$\begin{aligned} \mathcal{F}_{\tau^\star}(x,u;\widehat{\pi}) &= \{y \in \mathcal{Y} : r'(x,y) \le \tau^\star\}, \\ \tau^\star &= Q_{1-\alpha}\left(\{r_i'\}_{i\in\mathcal{I}_2} \cup \{1\}\right), \\ r'(x,y) &= \rho_y(x;\widehat{\pi}), \end{aligned} \tag{E.1}$$

and two randomized sequences where Scheme 1 performs randomization for all labels and was introduced before for conformal prediction sets (6.7) and Scheme 2 (added for completeness of comparison) performs randomization for all labels except the most likely one:

$$\begin{aligned} \mathcal{F}_{\tau^\star}(x,u;\widehat{\pi}) &= \{y \in \mathcal{Y} : r''(x,y) \le \tau^\star\}, \\ \tau^\star &= Q_{1-\alpha}\left(\{r_i''\}_{i\in\mathcal{I}_2} \cup \{1\}\right), \end{aligned} \tag{E.2}$$

where

$$r''(x,y) = \mathbb{1}\{\rho_y(x;\widehat{\pi}) > 0\} \cdot (\rho_y(x;\widehat{\pi}) + u \cdot \widehat{\pi}_y(x)).$$

We again use the Bayes-optimal classifier $\pi_y(x)$, and thus ignore the results that are due to estimation and focus purely on effects that are due to conformalization. For a single data draw we illustrate the resulting conformal prediction sets on Figures E.1c, E.1d and E.1e. While at first sight it might seem that non-randomized nested sequences is superior in terms of yielding prediction sets with smaller cardinality, it should be taken with a grain of salt. We repeatedly draw calibration and test data and track marginal characteristics for those sets. As expected, all three resulting prediction sets inherit $1-\alpha$ (marginal) coverage guarantee as confirmed on Figure E.2a. Moreover, Figure E.2b indeed confirms that randomization could yield larger prediction sets for not perfectly separable data. But Figure E.2c confirms that randomization proposed by Romano et al. (2020) (Scheme 1) demonstrates superior conditional coverage since for this example the true $\pi_y(x)$ is used, and thus the oracle prediction sets are recovered if $\tau^\star = 1-\alpha$. Figure E.2d confirms that oracle prediction sets are not recovered even when the size of the calibration set is increased.

E.2.3 Proofs

Proof of Theorem 6.1. By the definition of the conformal prediction set, $Y_{n+1} \in \mathcal{F}_{\tau^\star}(X_{n+1}, U_{n+1}; \widehat{\pi})$ if and only if:

$$r(X_{n+1}, Y_{n+1}, U_{n+1}; \widehat{\pi}) \leq Q_{1-\alpha}\left(\{r_i\}_{i\in\mathcal{I}_2} \cup \{1\}\right).$$

As the non-conformity scores $\{r_i\}_{i=1}^{n+1}$ are exchangeable random variables for any fixed $\widehat{\pi}$, Lemma E.1.1 implies the desired result conditional on $\{(X_i, Y_i)\}_{i\in\mathcal{I}_1}$. Finally, when randomization is performed, the scores are uniformly distributed, and thus Lemma E.1.1 implies that the marginal coverage is nearly tight. □

Proof of Theorem 6.2. First, recall the definition of weighted exchangeability (Tibshirani et al., 2019).

Definition 14 (Weighted exchangeability)**.** *Random variables $Z_1, \ldots, Z_n$ are said to be* weighted exchangeable*, with weight functions $\omega_1, \ldots, \omega_n$, if the density f of their joint distribution can be factorized as:*

$$f(z_1, \ldots, z_n) = \prod_{i=1}^{n} \omega_i(z_i) \cdot g(z_1, \ldots, z_n),$$

where g is any function that that invariant to permutations of its arguments, i.e., $g(z_{\sigma(1)}, \ldots, z_{\sigma(n)})$ for any permutation σ of $1, \ldots, n$.

Independent draws are always weighted exchangeable and it is easy to see that under label shift setting $Z_i = (X_i, Y_i, U_i)$, $i = 1, \ldots, n+1$ are weighted exchangeable with $\omega_i \equiv 1$, $i = 1, \ldots, n$ and $\omega_{n+1}((x, y)) = q(y)/p(y)$, for any pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Let $r_{n+1} := r\left(X_{n+1}, Y_{n+1}, U_{n+1}; \widehat{\pi}\right)$. By construction $Y_{n+1} \in \mathcal{F}_{\tau^\star}^{(w)}\left(X_{n+1}, U_{n+1}; \widehat{\pi}\right)$ if and only if:

$$r_{n+1} \leq Q_{1-\alpha}\left(\sum_{i=1}^{n} \tilde{p}_i^w(Y_{n+1})\delta_{r_i} + \tilde{p}_{n+1}^w(Y_{n+1})\delta_1\right).$$

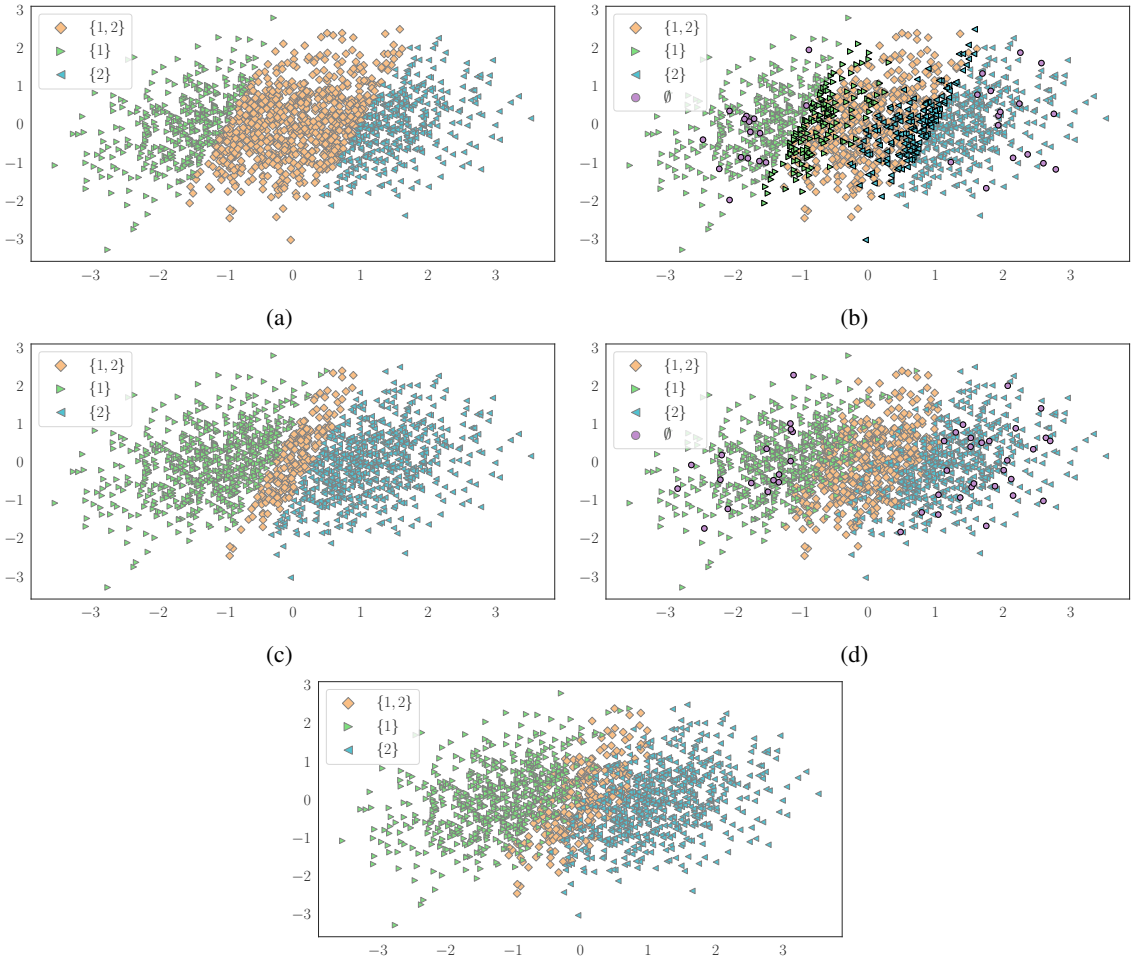


Figure E.1: Prediction sets corresponding to (a) the non-randomized oracle from (6.3); (b) the randomized oracle from (6.4); (c) the non-randomized conformal method (E.1); (d) the randomized conformal (scheme 1) method (6.7); (e) the randomized conformal (scheme 2) method (E.2). Notice that randomization acts differently in the oracle and conformal settings. While for the oracle setting randomization as per scheme 2 corresponds to recoloring the purple points to either green (leftmost color, class 0) or blue (rightmost color, class 1) depending on the most likely label, for the conformal setting two schemes yield conceptually different prediction sets. Presented visualizations might be misleading regarding the role of randomization for conformal classification as they suggest the non-randomized conformal method is the optimal one. See Figure E.2 and Section E.2.2 for more details.

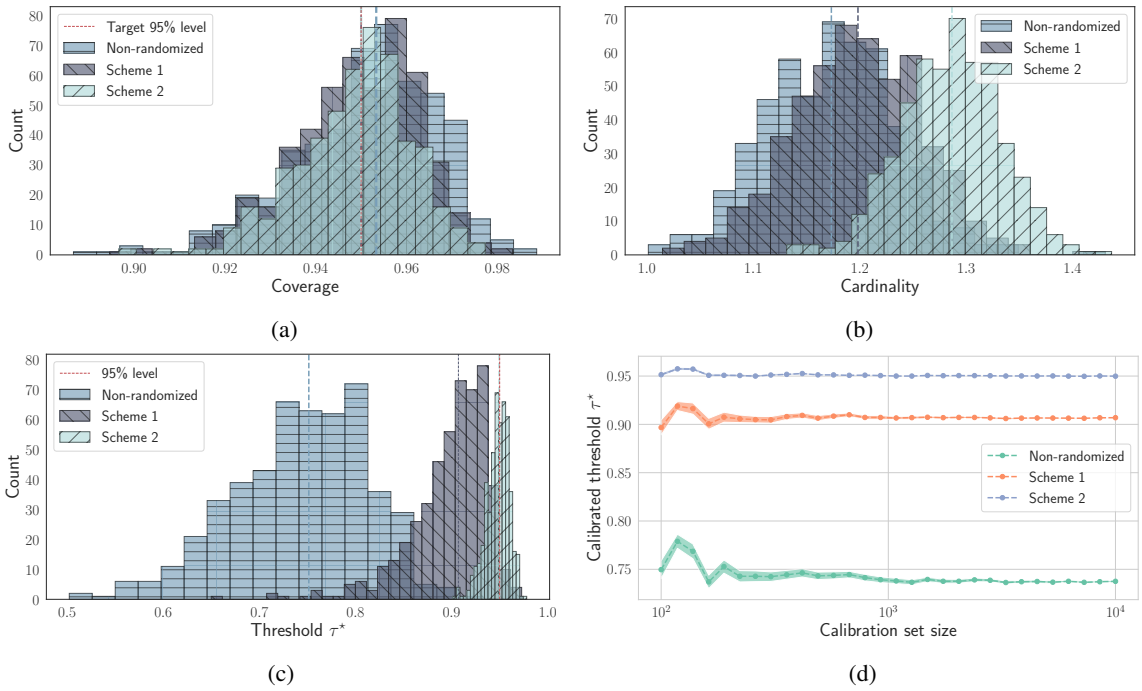


Figure E.2: Characteristics of conformal prediction sets for the simulation in Section E.2.2: (a) average marginal coverage, (b) average cardinality, (c) learned cut-off thresholds in each setting (appending empty prediction sets with the most-likely label does not impact the threshold), (d) learned cut-off thresholds in each setting when increasing the size of the calibration set. Key takeaways include: (i) marginal coverage requirement is met irrespective of whether conformal method performs randomization or not, (ii) the fact that randomization yields larger prediction sets, and thus is inferior is misleading, (iii) as in considered the example the conformal method recovers the oracle if learned threshold $\tau^\star = 0.95$, only randomized (scheme 1) one does it, (iv) the cut-off thresholds do not depend much on the size of the calibration dataset.

Under label shift assumption, weights (E.8) do simplify as

$$\begin{aligned} p_i^w(Z_1,\ldots,Z_{n+1}) &= \frac{\sum_{\sigma:\sigma(n+1)=i} w_{n+1}(Z_i)}{\sum_\sigma w_{n+1}(Z_{\sigma(n+1)})} \\ &= \frac{w(Y_i)}{\sum_{j=1}^n w(Y_j) + w(Y_{n+1})} \\ &= \tilde{p}_i^w(Y_{n+1}), \end{aligned}$$

for $i = 1,\ldots,n+1$ matching the ones stated in (6.9). The result follows by invoking Lemma E.1.2. As $\widehat{\pi}$ is fixed at the calibration step being pre-computed on a separate part of the dataset split, the result is conditional on $\{(X_i,Y_i)\}_{i\in\mathcal{I}_1}$. □

Proof of Corollary 6.2.1. As for the other results, here it is also conditional on the training data, and thus we omit writing $\{(X_i,Y_i)\}_{i\in\mathcal{I}_1}$ for succinctness and we use $r_{n+1} = r\left(X_{n+1},Y_{n+1},U_{n+1};\widehat{\pi}\right)$ to denote the radius for the test point. Choose an arbitrary $\varepsilon > 0$. We have:

$$\begin{aligned} & \mathbb{P}\left(Y_{n+1} \notin \mathcal{F}_{\tau^\star}^{(\widehat{w}_k)}\left(X_{n+1},U_{n+1};\widehat{\pi}\right)\right) \\ = \;& \mathbb{P}\left(r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1})\right) \\ = \;& \mathbb{P}\left(\left\{r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1})\right\} \cap \left\{r_{n+1}+\varepsilon > \tau^\star_w(Y_{n+1})\right\}\right) \\ + \;& \mathbb{P}\left(\left\{r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1}\right\}) \cap \left\{r_{n+1}+\varepsilon \leq \tau^\star_w(Y_{n+1})\right\}\right). \end{aligned} \tag{E.3}$$

We have that:

$$\mathbb{P}\left(r_{n+1} \geq \tau^\star_w(Y_{n+1})\right) = \mathbb{P}\left(r_{n+1} > \tau^\star_w(Y_{n+1})\right) < \alpha,$$

where equality is due to the fact that r_{n+1} in the randomized scheme has a continuous distribution and inequality is due to Theorem 6.2. For the first term in (E.3) we have:

$$\begin{aligned} & \mathbb{P}\left(\left\{r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1}\right\}) \cap \left\{r_{n+1}+\varepsilon > \tau^\star_w(Y_{n+1})\right\}\right) \\ = \;& \mathbb{P}\left(\left\{r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1}\right\}) \cap \left\{r_{n+1} > \tau^\star_w(Y_{n+1})-\varepsilon\right\}\right) \\ \leq \;& \mathbb{P}\left(r_{n+1} > \tau^\star_w(Y_{n+1})-\varepsilon\right), \end{aligned}$$

and for the second term we have that:

$$\mathbb{P}\left(\left\{r_{n+1} > \tau^\star_{\widehat{w}_k}(Y_{n+1})\right\} \cap \left\{r_{n+1} \leq \tau^\star_w(Y_{n+1})-\varepsilon\right\}\right) \leq \mathbb{P}\left(\left|\tau^\star_{\widehat{w}_k}(Y_{n+1}) - \tau^\star_w(Y_{n+1})\right| \geq \varepsilon\right).$$

Note that ε was chosen arbitrarily, so we can let $\varepsilon \to 0$. By the continuous mapping theorem, consistency of $\widehat{w}_k$ implies that of $\tau^{\star}_{\widehat{w}_k}(y)$, $y \in \mathcal{Y}$. Thus,

$$\lim_{k\to\infty} \mathbb{P}\left(Y_{n+1} \in \mathcal{F}^{(\widehat{w}_k)}_{\tau^\star}\left(X_{n+1}, U_{n+1}; \widehat{\pi}\right)\right) \geq 1-\alpha,$$

which concludes the proof of the Corollary. □

E.2.4 Simulation on Real Data

For the simulation in Section 6.2.2 we use `wine quality` dataset (Cortez et al., 2009) to illustrate the performance of the conformal prediction sets when label shift is (not) taken into account. We focus on white wine dataset only, which has 4898 instances with 11 features and construct a 3-class classification problem by keeping classes 5,6,7 only to avoid complications arising due to high imbalance in the dataset (less than 10% of the data points were removed). Other important aspects include

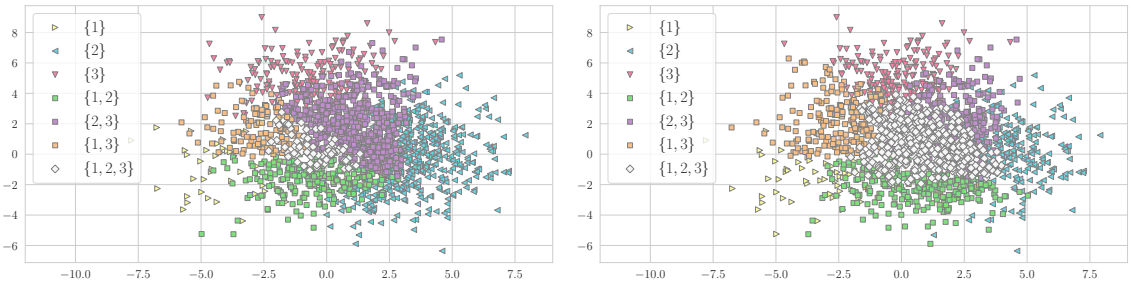
1. **Data Split:** First, the original dataset $\mathcal{D}$ is split into two disjoint and approximately equal sets $\mathcal{D}_1$ and $\mathcal{D}_2$. Then label shift is simulated via resampling according to considered class proportions yielding $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$ of the same size. Finally, the former dataset is split at random into sets for training (≈ 1000 instances), calibration (≈ 100 instances) and importance weights estimation (≈ 700 instances) and the latter is split is split at random into importance weight estimation (≈ 1000 instances; recall that only labels from the target are used) and test (≈ 1600 instances) sets.

2. **Model:** We use a standard Feed Forward Neural Network with 3 hidden layers with (128,64,32) neurons and ℓ_2-regularization in each as an underlying model. We use Adam optimizer with default parameters, set the maximum number of training epochs to 500 and deploy Early Stopping with patience for 25 epochs.

3. **Estimating label shift:** We use BBSE-soft (Lipton et al., 2018) for estimating importance weights.

E.2.5 Marginal Conformal versus Label-conditional Conformal

Various procedures of performing label-conditional conformal prediction have been proposed in a series of works (Vovk et al., 2005, 2016; Sadinle et al., 2019; Guan and Tibshirani, 2022). Those are based on a slight modification of the standard conformal p-value used to determine whether there is enough evidence to exclude given label from the prediction set. Roughly speaking, for each candidate label y instead of looking whether a pair (X_{n+1}, y) conforms well to the whole collection of points $\widetilde{\mathcal{D}} = \{(X_i, Y_i)\}_{i\in\mathcal{I}}$, one considers only the subcollection that shares the same label y. Since the standard exchangeability argument immediately implies validity, the difference then lies in a particular choice for the underlying (non-)conformity score. For example, one could design a score that aims to minimize expected size of the prediction set Sadinle et al. (2019); Guan and Tibshirani (2022).



(a) (b)

Figure E.3: (a) Conformal prediction sets with marginal coverage guarantee, (b) Conformal prediction sets with class-specific coverage guarantee. Stronger coverage comes at the price of larger the prediction sets in certain areas.

We now apply label-conditional split-conformal framework to the setting discussed in this work and focus on the case of not well-separated data. Consider, for example, the data simulation pipeline from Section 6.2.2. First, we fix $\alpha_y = \alpha = 0.1$ for all $y \in \mathcal{Y}$ and illustrate the difference between label-conditional conformal (6.10) and standard conformal (6.7) prediction sets with the same randomized non-conformity scores (6.6) for a fair comparison on Figure E.3. In both cases a shallow MLP (two layers with 100 hidden units in each) is used as an underlying predictor. In this particular example a stronger requirement of conditional validity forces many prediction sets to be larger and to contain the least populated class 1.

Then we perform 1000 simulations and compare label-conditional conformal against marginal conformal in two settings (in all cases prediction sets are forced to contain the most likely label for a fair comparison). First, we set the calibration set size to be ≈ 350 data points and compare two procedures depending on whether class proportions change, and in the former case we perform reweighting of the non-conformity scores as described in Section 6.2.2. On Figure E.4b we observe that when class proportions do not change label-conditional conformal yields larger prediction sets as opposed to standard marginal conformal due to a stronger coverage requirement. However, when class proportions change, after performing the reweighting with the true label likelihood ratios, both procedures output prediction sets of similar size on average as illustrated on Figure E.4d. Motivated by reasons related to the practical limits of data resources when keeping a sufficiently large held-out set per label could become prohibitive, we also consider a setting when the calibration set contains ≈ 100 data points (total). Smaller calibration set size results in losses of statistical power when testing whether a given label should be included into the prediction set, and thus, might yield larger prediction sets as observed on Figure E.4f.

To summarize, label-conditional conformal is a complementary (and a powerful) technique to label-shifted conformal that is inherently robust to changes in class proportions. It does not require importance weights, and thus can yield exact finite-sample guarantees. Still, it has certain limitations: (a) it might be potentially a bit conservative in certain areas of the sample space where classes overlap, (b) it requires further splitting of the calibration set that could have negative impact, especially when the number of classes K is large, a common setting for the modern datasets.

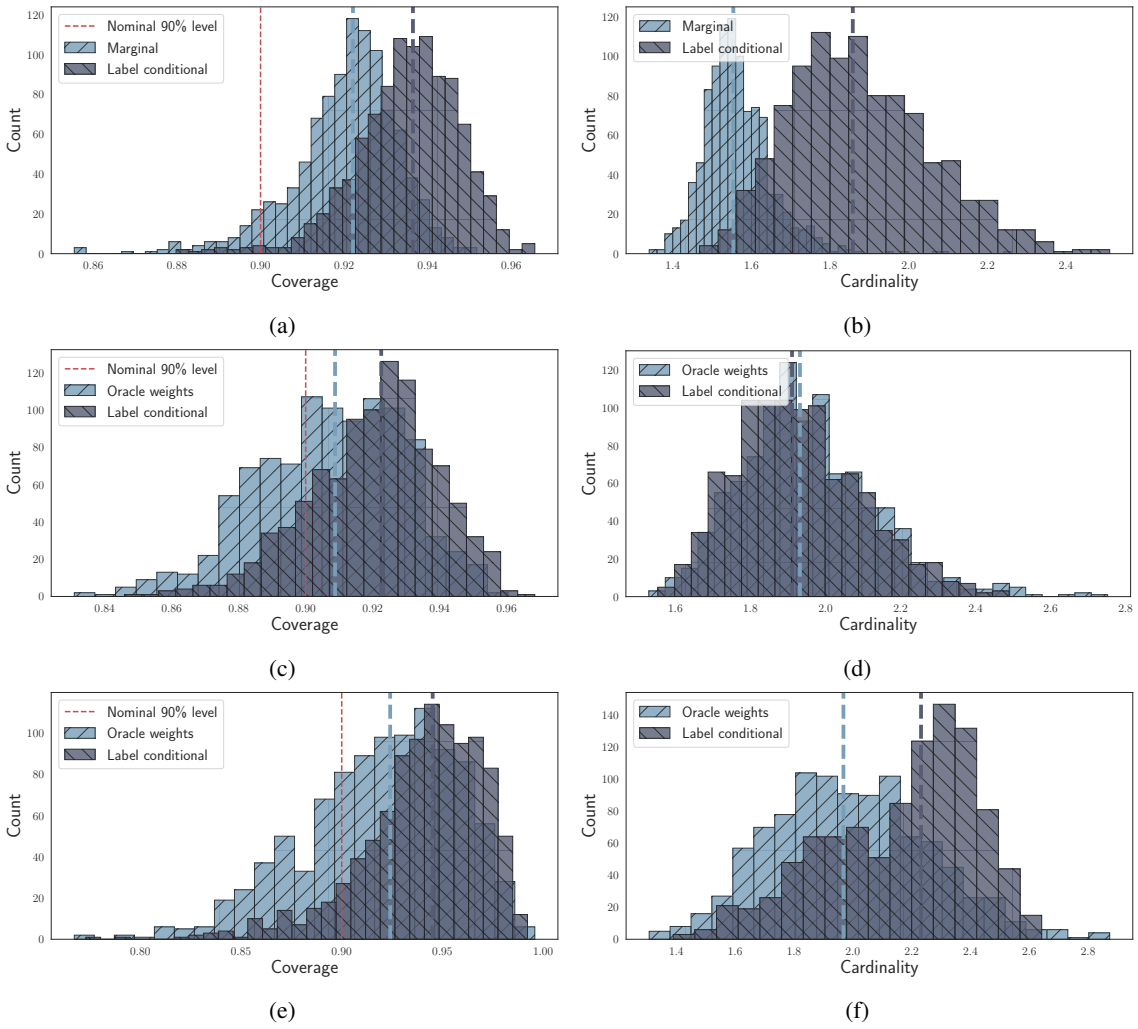


Figure E.4: Empirical coverage and average cardinality of conformal prediction sets: (a-b) source distribution and $\approx$ 350 calibration data points total, (c-d) target distribution and $\approx$ 350 calibration data points total, (e-f) target distribution and $\approx$ 100 calibration data points total. Complete comparison of the results is given in Section E.2.5.

E.3 Calibration

Section E.3.1 includes all proofs for Sections 6.3.1 and 6.3.2 and Section E.3.2 includes details about the simulation on a real dataset mentioned in Section 6.3.2.

E.3.1 Proofs

Proof of Theorem 6.3. Recall that $g : \mathcal{X} \to \mathcal{M}$ denotes the bin-mapping function. Let E be the event that $(g(X_1), \ldots, g(X_n)) = (g(x_1), \ldots, g(x_n))$. On this event, the number of calibration points N_m within each bin B_m is known and for each bin labels are i.i.d. with corresponding class probabilities given by $\pi^P_{y,m} = \mathbb{P}(Y = y \mid f(X) \in B_m)$ for all $y \in \mathcal{Y}$. Thus, a vector corresponding of label frequencies has multinomial distribution with parameters N_m and $\left\{\pi^P_{y,m}\right\}_{y\in\mathcal{Y}}$. Theorem E.1 yields that conditional on E

$$\sum_{y=1}^{K} \left|\widehat{\pi}^P_{y,m} - \pi^P_{y,m}\right| \geq \frac{2}{\sqrt{N_m}} \sqrt{\frac{1}{2} \ln\left(\frac{M2^K}{\alpha}\right)},$$

with probability at most α/M. Invoking union bound, we get that, conditional on E, with probability at least $1 - \alpha$,

$$\sum_{y=1}^{K} \left|\widehat{\pi}^P_{y,m} - \pi^P_{y,m}\right| \leq \frac{2}{\sqrt{N_m}} \sqrt{\frac{1}{2} \ln\left(\frac{M2^K}{\alpha}\right)},$$

simultaneously for all $m \in \mathcal{M}$. Since it is true for any E, we can marginalize to obtain the first assertion of the Proposition. The second assertion simply represents a consideration of the case when multiple bins happen to have the same calibrated output which is needed to state the desired calibration guarantee. Let

$$\varepsilon^\star = \sup_{m\in\mathcal{M}} \varepsilon_m$$

denote the worst-case bound. Note that $\varepsilon^\star$ is in fact random and to be fully rigorous we, first, perform next steps conditional on E and then marginalize to obtain the assertion. Now, for any $y \in \mathcal{Y}$:

$$\begin{aligned}
& \left|\mathbb{P}(Y = y \mid h(X)) - h_y(X)\right| \\
= \; & \left|\mathbb{E}\left[\mathbb{1}\{Y = y\} \mid h(X)\right] - h_y(X)\right| \\
\overset{(a)}{=} \; & \left|\mathbb{E}\left[\mathbb{1}\{Y = y\} \mid h(X)\right] - \mathbb{E}\left[h_y(X) \mid h(X)\right]\right| \\
\overset{(b)}{=} \; & \left|\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\{Y = y\} \mid g(X)\right] \mid h(X)\right] - \mathbb{E}\left[h_y(X) \mid h(X)\right]\right| \\
\overset{(c)}{=} \; & \left|\mathbb{E}\left[\left[\pi^P_{y,g(X)} - h_y(X)\right] \mid h(X)\right]\right| \\
\overset{(d)}{\leq} \; & \mathbb{E}\left[\left|\pi^P_{y,g(X)} - \widehat{\pi}_{y,g(X)}\right| \mid h(X)\right],
\end{aligned} \tag{E.4}$$

where $(a), (b)$ are due to the tower rule (h is a function of g), (c) is due to linearity of conditional expectation and due to definition of $\pi^P_{y,m}$ and, finally, (d) is due to Jensen's inequality. Consider the event:

$$E_1: \quad \left\|\widehat{\pi}^P_m - \pi^P_m\right\|_1 \leq \varepsilon_m,$$

simultaneously for all $m \in \mathcal{M}$. Note that the first assertion of the Proposition states event E_1 happens with probability at least $1-\alpha$ for chosen ε_m: $\mathbb{P}(E_1) \geq 1-\alpha$. Let E_2 be the following event:

$$E_2: \quad \sum_{y=1}^{K} \left|\mathbb{P}\left(Y = y \mid h(X)\right) - h_y(X)\right| \leq \varepsilon^\star.$$

Summing up over labels $y \in \mathcal{Y}$, (E.4) yields that on E_1 it holds with probability 1:

$$\sum_{y=1}^{K} \left|\mathbb{P}\left(Y = y \mid h(X)\right) - h_y(X)\right| \leq \mathbb{E}\left[\left\|\pi^P_{g(X)} - \widehat{\pi}_{g(X)}\right\|_1 \mid h(X)\right] \leq \mathbb{E}\left[\varepsilon^\star \mid h(X)\right] = \varepsilon^\star,$$

since $\varepsilon^\star$ is a constant. We get that $E_1 \subseteq E_2$, and thus $\mathbb{P}(E_2) \geq \mathbb{P}(E_1)$, and the assertion of the Proposition follows. $\square$

Proof of Proposition 7. The Proposition is a straightforward combination of the Bayes rule and label shift assumption. Given a predictor f, for any class label $y \in \mathcal{Y}$ and any bin B_m, $m \in \mathcal{M} = \{1, \ldots, M\}$ one can equivalently represent conditional probabilities with respect to the target distribution as:

$$\begin{aligned}
& \mathbb{P}_Q\left(Y = y \mid f(X) \in B_m\right) \\
\overset{(a)}{=} \; & \mathbb{P}_Q\left(f(X) \in B_m \mid Y = y\right) \cdot \frac{\mathbb{P}_Q\left(Y = y\right)}{\mathbb{P}_Q\left(f(X) \in B_m\right)} \\
\overset{(b)}{=} \; & \mathbb{P}_P\left(f(X) \in B_m \mid Y = y\right) \cdot \frac{\mathbb{P}_Q\left(Y = y\right)}{\mathbb{P}_Q\left(f(X) \in B_m\right)} \\
\overset{(c)}{=} \; & \mathbb{P}_P\left(Y = y \mid f(X) \in B_m\right) \cdot \frac{\mathbb{P}_Q\left(Y = y\right)}{\mathbb{P}_P\left(Y = y\right)} \cdot \frac{\mathbb{P}_P\left(f(X) \in B_m\right)}{\mathbb{P}_Q\left(f(X) \in B_m\right)} \\
= \; & \mathbb{P}_P\left(Y = y \mid X \in B_m\right) \cdot w(y) \cdot V_m,
\end{aligned}$$

where $w(y)$ is the importance weight of label y and V_m is the 'relative volume' of bin B_m. Steps (a), (c) are due to the Bayes rule, (b) is due to label shift assumption. Normalization: $\sum_{k=1}^{K} \mathbb{P}_Q\left(Y = k \mid f(X) \in B_m\right) = 1$, implies that:

$$V_m = \frac{1}{\sum_{k=1}^{K} \pi^P_{k,m} \cdot w(k)}.$$

Thus for all bins $m \in \mathcal{M}$ and labels $y \in \mathcal{Y}$ it holds:

$$\pi^Q_{y,m} = \frac{\pi^P_{y,m} \cdot w(y)}{\sum_{k=1}^{K} \pi^P_{k,m} \cdot w(k)},$$

which concludes the proof of the Proposition. □

Proof of Theorem 6.4. By triangle inequality, one obtains that for any bin $m \in \mathcal{M}$:

$$\sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{(\widehat{w})} - \pi_{y,m}^{Q} \right| \leq \quad \sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{(w)} - \pi_{y,m}^{Q} \right| + \sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{(\widehat{w})} - \widehat{\pi}_{y,m}^{(w)} \right|. \tag{E.5}$$

Consider the first term in (E.5). For any $y \in \mathcal{Y}$:

$$\begin{aligned}
& \left| \widehat{\pi}_{y,m}^{(w)} - \pi_{y,m}^{Q} \right| \\
= \quad & \left| \frac{w(y) \cdot \widehat{\pi}_{y,m}^{P}}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{w(y) \cdot \pi_{y,m}^{P}}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \right| \\
= \quad & \left| \frac{\widehat{\pi}_{y,m}^{P}}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{\pi_{y,m}^{P}}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \right| \cdot w(y) \\
= \quad & \left| \frac{\widehat{\pi}_{y,m}^{P}}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{\pi_{y,m}^{P} - \widehat{\pi}_{y,m}^{P} + \widehat{\pi}_{y,m}^{P}}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \right| \cdot w(y) \\
\overset{(a)}{\leq} \quad & \left| \frac{1}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{1}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \right| \cdot \widehat{\pi}_{y,m}^{P} \cdot w(y) + \; w(y) \cdot \left| \frac{\pi_{y,m}^{P} - \widehat{\pi}_{y,m}^{P}}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \right|,
\end{aligned}$$

where (a) is due to triangle inequality. We infer that:

$$\begin{aligned}
& \sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{(w)} - \pi_{y,m}^{Q} \right| \\
\leq \quad & \left| 1 - \frac{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}_{k,m}^{P}}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \right| + \frac{\sum_{y=1}^{K} w(y) \left| \pi_{y,m}^{P} - \widehat{\pi}_{y,m}^{P} \right|}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \\
= \quad & \frac{\left| \sum_{k=1}^{K} w(k) \cdot \left(\widehat{\pi}_{k,m}^{P} - \pi_{l,m}^{P} \right) \right|}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} + \frac{\sum_{y=1}^{K} w(y) \left| \pi_{y,m}^{P} - \widehat{\pi}_{y,m}^{P} \right|}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \\
\overset{(a)}{\leq} \quad & 2 \cdot \frac{\sum_{y=1}^{K} w(y) \left| \pi_{y,m}^{P} - \widehat{\pi}_{y,m}^{P} \right|}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}} \\
\overset{(b)}{\leq} \quad & 2 \cdot \frac{(\sup_k w(k)) \cdot \sum_{y=1}^{K} \left| \pi_{y,m}^{P} - \widehat{\pi}_{y,m}^{P} \right|}{\sum_{l=1}^{K} w(l) \cdot \pi_{l,m}^{P}},
\end{aligned}$$

where (a) is due to triangle inequality and (b) is due to Hölder's inequality. Observe that for any $m \in \mathcal{M}$:

$$\frac{1}{\sum_{k=1}^{K} w(k) \cdot \pi_{k,m}^{P}} \leq \quad \frac{1}{\left(\inf\limits_{k : w(k) \neq 0} w(k) \right) \cdot \sum_{l=1}^{K} \pi_{l,m}^{P}} = \quad \frac{1}{\inf\limits_{k : w(k) \neq 0} w(k)},$$

as $\sum_{l=1}^{K} \pi_{l,m}^{P} = 1, \forall m \in \mathcal{M}$. Hence, for any $m \in \mathcal{M}$,

$$\sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{(w)} - \pi_{y,m}^{Q} \right| \leq \quad 2 \cdot \frac{\sup_k w(k)}{\inf\limits_{k:w(k)\neq 0} w(k)} \cdot \sum_{y=1}^{K} \left| \pi_{y,m}^{P} - \widehat{\pi}_{y,m}^{P} \right|. \tag{E.6}$$

Now, consider the second term in (E.5). Observe that:

$$\begin{aligned}
\left| \widehat{\pi}_{y,m}^{(\widehat{w})} - \widehat{\pi}_{y,m}^{(w)} \right| &= \left| \frac{\widehat{w}(y) \cdot \widehat{\pi}_{y,m}^{P}}{\sum_{k=1}^{K} \widehat{w}(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{w(y) \cdot \widehat{\pi}_{y,m}^{P}}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \right| \\
&= \left| \frac{\widehat{w}(y)}{\sum_{k=1}^{K} \widehat{w}(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{w(y)}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \right| \cdot \widehat{\pi}_{y,m}^{P} \\
&= \left| \frac{\widehat{w}(y)}{\sum_{k=1}^{K} \widehat{w}(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{w(y) - \widehat{w}(y) + \widehat{w}(y)}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \right| \cdot \widehat{\pi}_{y,m}^{P} \\
&\overset{(a)}{\leq} \left| \frac{1}{\sum_{k=1}^{K} \widehat{w}(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{1}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \right| \cdot \widehat{\pi}_{y,m}^{P} \cdot \widehat{w}(y) + \frac{\widehat{\pi}_{y,m}^{P} \cdot |w(y) - \widehat{w}(y)|}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}},
\end{aligned}$$

where (a) is due to triangle inequality. Thus,

$$\begin{aligned}
\sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{(\widehat{w})} - \widehat{\pi}_{y,m}^{(w)} \right| &\leq \left| \frac{1}{\sum_{k=1}^{K} \widehat{w}(k) \cdot \widehat{\pi}_{k,m}^{P}} - \frac{1}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \right| \cdot \sum_{y=1}^{K} \widehat{\pi}_{y,m}^{P} \cdot \widehat{w}(y) + \frac{\sum_{y=1}^{K} \widehat{\pi}_{y,m}^{P} \cdot |w(y) - \widehat{w}(y)|}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \\
&= \left| 1 - \frac{\sum_{y=1}^{K} \widehat{w}(y) \cdot \widehat{\pi}_{y,m}^{P}}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \right| + \frac{\sum_{y=1}^{K} \widehat{\pi}_{y,m}^{P} \cdot |w(y) - \widehat{w}(y)|}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \\
&= \frac{\left| \sum_{y=1}^{K} (w(y) - \widehat{w}(y)) \cdot \widehat{\pi}_{y,m}^{P} \right|}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} + \frac{\sum_{y=1}^{K} \widehat{\pi}_{y,m}^{P} \cdot |w(y) - \widehat{w}(y)|}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}} \\
&\leq \frac{2 \left\| \widehat{w} - w \right\|_{\infty}}{\sum_{l=1}^{K} w(l) \cdot \widehat{\pi}_{l,m}^{P}},
\end{aligned}$$

since $\sum_{k=1}^{K} \widehat{\pi}_{k,m}^{P} = 1, \forall m \in \mathcal{M}$. Similarly, for any $m \in \mathcal{M}$:

$$\frac{1}{\sum_{k=1}^{K} w(k) \cdot \widehat{\pi}_{k,m}^{P}} \leq \frac{1}{\left(\inf_{l:w(l)\neq 0} w(l)\right) \cdot \sum_{k=1}^{K} \widehat{\pi}_{k,m}^{P}} = \frac{1}{\inf_{l:w(l)\neq 0} w(l)}.$$

Thus, we get that for any $m \in \mathcal{M}$:

$$\sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{(\widehat{w})} - \widehat{\pi}_{y,m}^{(w)} \right| \leq \frac{2 \left\| \widehat{w} - w \right\|_{\infty}}{\inf_{l:w(l)\neq 0} w(l)}. \tag{E.7}$$

Combining bounds (E.6) and (E.7) with the bound (E.5), we obtain that for any $m \in \mathcal{M}$:

$$\sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{(\widehat{w})} - \pi_{y,m}^{Q} \right| \leq \quad 2\kappa \cdot \sum_{y=1}^{K} \left| \widehat{\pi}_{y,m}^{P} - \pi_{y,m}^{P} \right| + \frac{2 \left\| \widehat{w} - w \right\|_{\infty}}{\inf_{l:w(l)\neq 0} w(l)},$$

which concludes the proof of the Theorem. □

E.3.2 Simulation on Real Data

For the simulation mentioned in Section 6.3.2 we use `wine quality` dataset (Cortez et al., 2009). The original dataset contains ratings for white wines and we reduce it to a binary classification problem by treating wine as good if the corresponding rating is at least 7 on a 10-point scale. Logistic regression is used as an underlying predictor and for each pass the original dataset $\mathcal{D}$ is, first, split into two disjoint and approximately equal sets $\mathcal{D}_1$ and $\mathcal{D}_2$. Label shift is simulated via resampling of $\widetilde{\mathcal{D}}_1$ with class proportions $p = (0.8, 0.2)$ and $\widetilde{\mathcal{D}}_2$ with class proportions $(0.5, 0.5)$. Final splitting resulted in ≈ 1350 instances used for both training and calibration, ≈ 700 and ≈ 400 instances used for importance weights estimation on the source and the target respectively and ≈ 1100 instances used for the test. Uniform-mass binning with 10 bins was used for calibration purposes. For 4 random data splits the resulting reliability curves are presented on Figure E.5 illustrating that calibration with proper reweighting leads to approximate calibration on the target domain and uncorrected fails to do so.

E.4 Auxiliary Results

Note Lemma E.1.1 and Lemma E.1.2 were originally formulated for possibly unbounded non-conformity scores. It is easy to see that we can safely replace point masses δ_∞ by δ_1 in the conformal classification setting considered in this work.

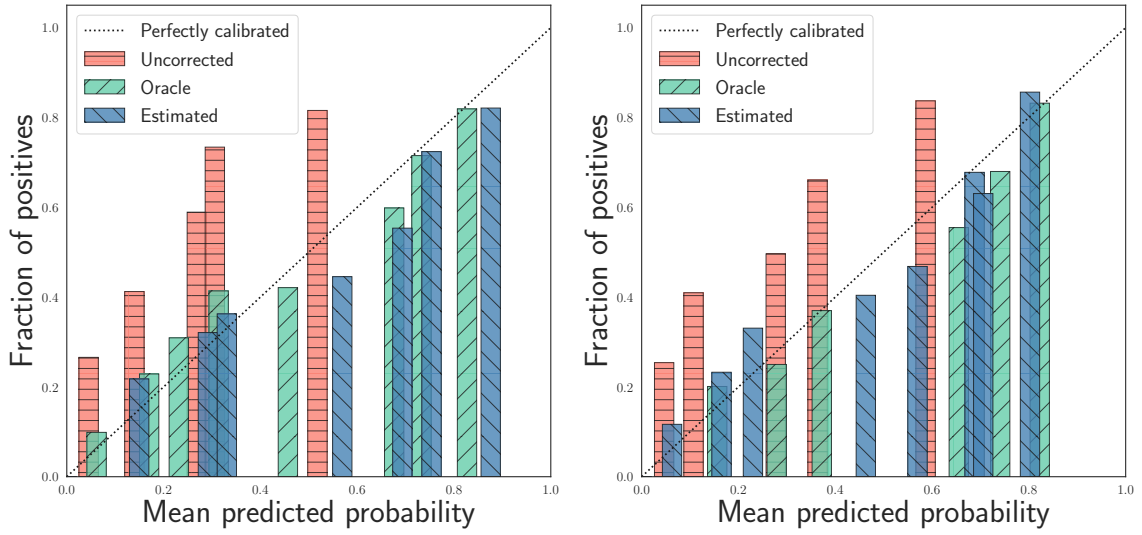
Theorem E.1 (Bretagnolle-Huber-Carol inequality (van der Vaart and Wellner, 1996))**.** *If the random vector* $(N_1, \ldots, N_k)$ *is multinomially distributed with parameters* n *and* $(p_1, \ldots, p_k)$*, then*

$$\mathbb{P}\left(\sum_{i=1}^{k} |N_i - np_i| \geq 2\sqrt{n}\lambda \right) \leq 2^k e^{-2\lambda^2}, \quad \lambda > 0.$$

Lemma E.1.1 (Lemma 1 (Tibshirani et al., 2019))**.** *Assume* $Z_1, \ldots, Z_{m+1}$ *are exchangeable random variables supported on* $[0, 1]$*. Then for any* $\beta \in (0, 1)$*,*

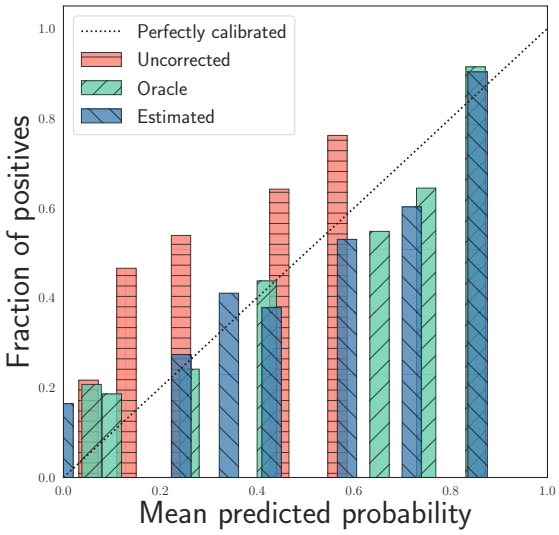
$$\mathbb{P}\left(Z_{m+1} \leq Q_\beta\left(Z_{1:m} \cup \{1\} \right) \right) \geq \beta.^*$$

Moreover, if Z_i*,* $i = 1, \ldots, m+1$ *are almost surely distinct, then the above probability is upper bounded by* $\beta + \frac{1}{m+1}$*.*

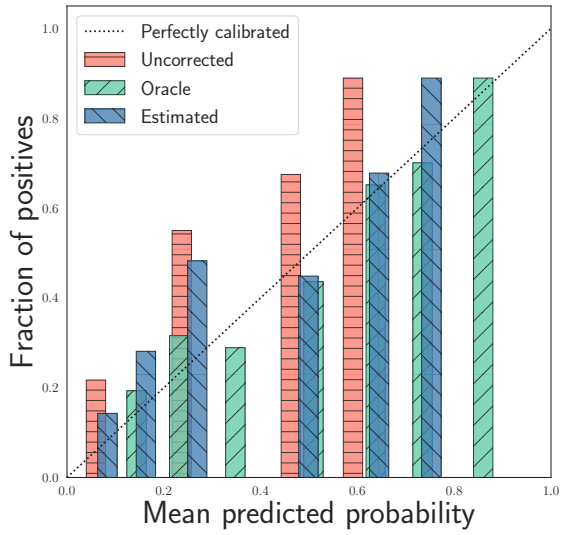


(a)

(b)



(c)



(d)

Figure E.5: Reliability curves for the simulation on the `wine quality` dataset obtained for several data splits. Notice that the bars indicating calibration using oracle and estimated importance weights are quite similar to each other, but most importantly that both are very close to the ideal diagonal line (perfect calibration). In contrast, the uncorrected bars are poorly calibrated, demonstrating both the need for handling label shift and the relative success of our procedures in doing so. See Section E.3.2 for details.

Lemma E.1.2 (Lemma 3 (Tibshirani et al., 2019)). *Let Z_i, $i = 1, \ldots, n+1$ be weighted exchangeable random variables with weight functions $w_1, \ldots, w_{n+1}$ and supported on $[0,1]$. Let $V_i = S(Z_i, Z_{-i})$, where $Z_{-i} = Z_{1:(n+1)} \backslash \{Z_i\}$, $i = 1, \ldots, n+1$ and S is an arbitrary score function. Define*

$$p_i^w(z_1, \ldots, z_{n+1}) = \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)})}, \tag{E.8}$$

for $i = 1, \ldots, n+1$, where summations are taken over permutations σ of $1, \ldots, n+1$. Then for any $\beta \in (0,1)$,

$$\mathbb{P}(V_{n+1} \leq Q_\beta(G_n)) \geq 1 - \beta,$$

where the distribution G_n is defined as

$$G_n := \sum_{i=1}^{n} p_i^w(Z_1, \ldots, Z_{n+1}) \delta_{V_i} + p_{n+1}^w(Z_1, \ldots, Z_{n+1}) \delta_1.$$