Model-X Sequential Testing for Conditional Independence via Testing by Betting

Shalev Shaer[*,1] **Gal Maman**[*,1] **Yaniv Romano**[1,2]
[1]Department of Electrical and Computer Engineering, Technion–Israel Institute of Technology
[2]Department of Computer Science, Technion–Israel Institute of Technology

Abstract

This paper develops a model-free sequential test for conditional independence. The proposed test allows researchers to analyze an incoming i.i.d. data stream with any arbitrary dependency structure, and safely conclude whether a feature is conditionally associated with the response under study. We allow the processing of data points online, as soon as they arrive, and stop data acquisition once significant results are detected, rigorously controlling the type-I error rate. Our test can work with any sophisticated machine learning algorithm to enhance data efficiency to the extent possible. The developed method is inspired by two statistical frameworks. The first is the model-X conditional randomization test, a test for conditional independence that is valid in offline settings where the sample size is fixed in advance. The second is testing by betting, a "game-theoretic" approach for sequential hypothesis testing. We conduct synthetic experiments to demonstrate the advantage of our test over out-of-the-box sequential tests that account for the multiplicity of tests in the time horizon, and demonstrate the practicality of our proposal by applying it to real-world tasks.

1 INTRODUCTION

A central problem in data analysis is to rigorously find conditional associations in complex data sets with nonlinear dependencies. This problem lies at the heart of causal discovery (Pearl et al., 2000; Peters et al., 2017), variable selection (Barber and Candès, 2015; Candes et al., 2018), machine learning interpretability (Burns et al., 2020; Lu et al., 2018), economics (Angrist and Kuersteiner, 2011; Wang and Hong, 2018), and genetics studies (Sesia et al., 2019; Bates et al., 2020), to name a few. In such applications, the data are often collected online, and, naturally, researchers are interested in analyzing the data points immediately after they are observed so that further data acquisition can be terminated as soon as significant results are detected. This experimental setting, for example, is typical in decision-making (Nikolakopoulou et al., 2018; Bhui, 2019) and clinical trials (Park et al., 2018), where the need for additional samples to obtain accurate statistical inference must frequently be balanced with experimental costs.

To formalize the problem, suppose we are given a stream of data points (X_t, Y_t, Z_t) for $t \in \mathbb{N} = 1, 2, \ldots$, where each triplet contains a response $Y_t \in \mathbb{R}$, a feature $X_t \in \mathbb{R}$, and a vector of covariates $Z_t \in \mathbb{R}^d$. We assume the observations are sampled i.i.d. from $P_{YXZ} = P_{Y|XZ}P_{XZ}$, where $P_{Y|XZ}$ is unknown. Given such an online data stream, our goal is to test for *conditional independence* (CI), where the null hypothesis is given by

$$H_0 : X_t \perp\!\!\!\perp Y_t \mid Z_t \text{ for all } t \in \mathbb{N}.$$

In words, we say that H_0 is true if X_t is independent of the response Y_t after accounting for the effect of the covariates Z_t, *simultaneously for all time steps* t. We refer to X_t that satisfies H_0 as an 'unimportant' feature. Analogously, the alternative hypothesis implies that X_t carries new information on the response Y_t beyond what is already contained in Z_t, i.e., $X_t \not\perp\!\!\!\perp Y_t \mid Z_t$. Therefore, we say that such a feature X_t is 'important'.

The goal of sequential hypothesis testing is to formulate a concrete decision rule on whether we can confidently reject the null at each time step t, by monitoring and accumulating the evidence collected at each step against the null using past data $\{(X_s, Y_s, Z_s)\}_{s=1}^t$ (Wald, 1945). This allows the analyst great flexibility, as she can decide, at each step, whether new data should be collected to support the question under study. Key to this setting is the need to provide the analyst with a tool that rigorously controls the type-I error rate—defined as the probability of rejecting the null when it is in fact true—at any given desired level α, simul-

*Equal contribution. Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

taneously for all time steps t. This requirement should not be confused with the premise of classic offline tests for CI that attain type-I error rate control *only when the sample size is fixed in advance*. We refer to these as offline tests, emphasizing that one cannot naively monitor the outcome of a classic test—a p-value—and reject the null at an optional time step t without accounting for the multiplicity of the tests across the time horizon; this strategy would result in inflation of the type-I error rate. Beyond online type-I error rate control, ideally, we wish to have a powerful test that would reject the null when it is false, and we want it to do so as early as possible.

Our contribution

In this paper, we develop a novel sequential test for CI. Our proposal takes inspiration from two powerful and attractive statistical tools that are gaining increasing attention in recent years. The first is the model-X conditional randomization test (CRT) by Candes et al. (2018), an *offline* test for CI. The second is testing by betting (Shafer and Vovk, 2019; Grünwald et al., 2020), a "game-theoretic" approach for sequential hypothesis testing, where our proposal is very much inspired by the line of work reported in (Ramdas and Wenbe, 2020; Ramdas et al., 2022). The method we introduce in this paper, presented in Section 4, generalizes the offline CRT to the challenging online setting, resulting in a new test with the following features.

Safe testing with early stopping: building on recent advances in sequential testing using e-values and martingales, detailed in Section 3, the proposed CI test is guaranteed to control the type-I error rate at any time step. In particular, the analyst is allowed to track the outcome of the test over time, and safely reject the null if it exceeds a user-defined significance level, preventing a wasteful collection of unnecessary new data points.

Model-X setting: similar to the offline CRT method, described in Section 2, the online test we propose does not make any assumptions on the conditional distribution of $Y \mid X, Z$. For instance, we do not make unrealistic assumptions that the relationship between Y and (X, Z) is linear, or that $Y \mid X, Z$ is Gaussian. However, this advantage comes at the cost of assuming that the distribution of $X \mid Z$ is known. This assumption is common to all tests belonging to the family of model-X knockoffs, including the CRT, and it is manageable when (i) large unlabeled data are available in contrast to labeled data, or (ii) when we have good prior knowledge about the distribution of $X \mid Z$ (Candes et al., 2018; Sesia et al., 2019; Romano et al., 2020). We discuss this in more detail in Section 4.2.

Online learning from past experience: the proposed test can leverage any machine learning algorithm to powerfully discover violations of the CI null. In particular, when a new triplet (Y_t, X_t, Z_t) is observed, we use online learning techniques to efficiently update the running predictive model, instead of fitting a new model from scratch. This way, the whole data stream is used for training in a computationally efficient manner. The proposed framework also falls under the umbrella of interactive tests (Lei and Fithian, 2018; Lei et al., 2021; Duan et al., 2022), providing the analyst the liberty to look at past data and decide how to modify the learning algorithm at any time step—e.g., to switch to a model that is more robust to outliers—to better discriminate the null and alternative hypotheses when applied to future test points.

Optimized software package: we provide a python code that implements our testing framework, is available at `https://github.com/shaersh/ecrt`. The package includes important design principles: an automatic hyper-parameter tuning that does not require fitting the machine learning model from scratch (Supplementary Section E); an ensemble procedure for improving the power of the test by averaging multiple martingales (Section 4.2); and a de-randomization procedure that also improves power by reducing inherent algorithmic randomness due to a sampling mechanism that is necessary to formulate the test (Section 4.2).

2 MODEL-X CI TESTING

The CRT, developed by Candes et al. (2018), is an *offline* test for CI that we build upon in this work. A key advantage of the CRT is that it assumes nothing on the conditional distributions of $Y \mid X, Z$ and $Y \mid Z$. This test, however, assumes that the conditional distribution of $X \mid Z$ is known. The CRT procedure, described in Algorithm 3 in Supplementary Section B, resembles classic permutation tests and has two key components: a test statistic function $T(\cdot)$ and a function that samples dummy features $\tilde{X}$ from $P_{X|Z}$. Since $\tilde{X}$ is sampled without looking at Y, the dummy triplets $(\tilde{X}, Y, Z)$ satisfy $\tilde{X} \perp\!\!\!\perp Y \mid Z$ by construction. Hence, by comparing the test statistic evaluated on the original $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ and dummy $\{(\tilde{X}_i, Y_i, Z_i)\}\}_{i=1}^n$ triplets, the CRT generates a valid p-value p_n, controlling the CI null at level α when the sample size n is fixed in advance (Candes et al., 2018), i.e.,

$$\mathbb{P}[p_n \leq \alpha \mid \text{the null is true}] \leq \alpha \text{ for a fixed } n. \quad (1)$$

Put differently, when all n observations are available before testing, one can use p_n to rigorously control the type-I error. However, future observations cannot be utilized to generate a new p-value (e.g., in cases where the null is not rejected) without a proper correction that ensures the validity of the sequential test. To see this, suppose for simplicity that under the null $p_n \sim \text{Uniform}(0, 1)$ is distributed uniformly over the $[0, 1]$ interval for any fixed n, satisfying (1). Next, let τ be a data-dependent stopping time, given by

$$\tau = \{\min n : p_n \leq \alpha, n \in \mathbb{N}\}.$$

Now, observe that with this choice of stopping time, $\mathbb{P}[p_\tau \leq \alpha \mid \text{the null is true}]$ cannot be bounded by α anymore: there exists τ such that a rejection rule $p_\tau \leq \alpha$ would result in an invalid α-level test.

In many applications, however, one is interested in applying the test online to obtain reliable data-driven conclusions as soon as possible. This motivates us to adopt a fresh statistical approach for hypothesis testing, called testing by betting, briefly described in the next section.

3 TESTING BY BETTING

Before diving into the mathematical principles of the testing by betting approach, we follow Shafer and Vovk (2019) and Shafer (2021) and present an intuitive interpretation of this framework. Imagine we are playing a game, in which we start with initial toy money. At each time step, we place a bet against the null hypothesis, and then reality reveals the truth. If this bet turns out to be correct, our wealth is increased by the money we risk in the bet; otherwise, we lose and the wealth is decreased accordingly. If our wealth at time t is at least $1/\alpha$ times as large as the initial toy money we started with (e.g., we have managed to multiply our initial money by a factor of $1/0.05 = 20$ for $\alpha = 0.05$) we can confidently reject the null, knowing that the type-I error is guaranteed to be controlled at level α. A property important to the formulation of the above game is this: if the null is true, the game must be fair in the sense that it is unlikely we will be able to significantly increase our initial toy money, no matter how sophisticated our betting strategy is.

A mathematical object that is crucial to formalize a fair game is a *test martingale*, defined below.

Definition 1. *A random process $\{S_t : t \in \mathbb{N}_0\}$ is a test martingale for a given null hypothesis H_0 if it satisfies the following conditions: (i) $S_0 = 1$, (ii) $S_t \geq 0$, $\forall t \in \mathbb{N}_0$, and (iii) $\{S_t : t \in \mathbb{N}_0\}$ is a supermartingale under H_0.*

In the view of testing by betting, the initial value of the test martingale S_0 represents the initial toy money in the game, and S_t corresponds to our wealth at time t. Now, suppose we are handed a valid test martingale $\{S_t : t \in \mathbb{N}_0\}$, and let $\tau \geq 1$ be a data-dependent optional stopping time. By invoking the *optional stopping theorem* we get

$$\mathbb{E}_{H_0}[S_\tau] \leq \mathbb{E}_{H_0}[S_0] = 1, \tag{2}$$

meaning that S_τ is a non-negative random variable whose expected value is bounded by one for any stopping time $\tau \geq 1$. In the literature on testing by betting, S_τ is often referred to as an *e-value* (Vovk and Wang, 2021; Wang and Ramdas, 2022; Grünwald et al., 2020). Importantly, the consequence of (2) is that, under the null, the game is fair since the expected value of our wealth S_t at any time step t is bounded by the initial toy money S_0. Moreover, since $\{S_t : t \in \mathbb{N}_0\}$ is a non-negative supermartingale under H_0, we can apply Ville's inequality (Ville, 1939) and get

$$\mathbb{P}_{H_0}(\exists t \geq 1 : S_t \geq 1/\alpha) \leq \alpha \mathbb{E}_{H_0}[S_0] = \alpha, \tag{3}$$

for any $\alpha \in (0, 1)$. Therefore, the ability to form a valid test martingale allows us to rigorously test for H_0 and reject the null if $S_t \geq 1/\alpha$ at any time step, with the premise that the type-I error would not exceed the level α. Crucially, when the null is false, S_t can largely grow depending on how successful our betting strategy is. In Section 4 we formulate a valid test martingale and design a powerful betting strategy.

Related work. Sequential testing has a long standing history (Wald, 1945; Lai, 1984; Naghshvar and Javidi, 2013; Lhéritier and Cazals, 2018), where the sequential probability ratio test of Wald (1945) is perhaps one of the first sequential hypothesis tests. More recently, the *testing by betting* methodology (Shafer and Vovk, 2019; Shafer, 2021) has led to the design of new powerful nonparametric approaches for constructing confidence sequences, e.g., (Jun and Orabona, 2019; Waudby-Smith and Ramdas, 2023), for testing a single hypothesis, as well as for testing multiple hypotheses; see (Waudby-Smith and Ramdas, 2023, Section 6) for a detailed summary.

Related methods to our proposal are offline and online two-sample tests that are based on martingales (Balsubramani and Ramdas, 2016; Turner et al., 2021; Shekhar and Ramdas, 2021; Duan et al., 2022). Specifically, Shekhar and Ramdas (2021) studied the problem of designing martingale-based sequential nonparametric one- and two-sample tests that are consistent, i.e., these sequential tests can attain *power one* under certain conditions. In our work, we build on the foundations of Shekhar and Ramdas (2021), and extend this framework to CI testing. Recently, Ren and Barber (2022) suggested using e-values to de-randomize the outcome of the knockoff filter—a sister method to the CRT that focuses on false discovery rate control (FDR) in an offline setting. In our work, we aggregate e-values to de-randomize our test, where the e-values we define take a different form than those proposed by Ren and Barber (2022), as we focus on sequential testing of a single feature. Lastly, independent work by Grünwald et al. (2022), which has been developed and posted in parallel to ours, also offers a martingale-based sequential test under the model-X setting, although suggesting a different test martingale. In Supplementary Section D we provide a more detailed discussion about the relation of our proposal to that of Grünwald et al. (2022), along with empirical comparisons.

4 THE PROPOSED e-CRT

In this section, we introduce e-CRT: a sequential test for CI based on martingales and e-values. Suppose we are given a

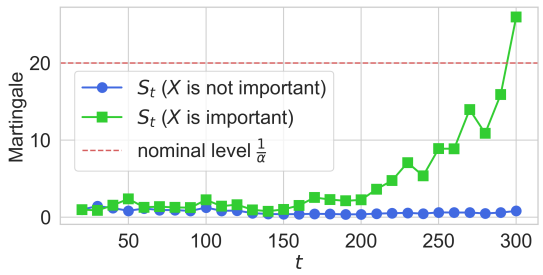


Figure 1: **Illustration of the test martingale (wealth) S_t as a function of t.** The blue (resp. green) curve represents the test martingale evaluated on simulated `null data` (resp. `non-null data`).

machine learning model $\hat{f}_t$, fitted on an initial batch of labeled data $\{(X_s, Y_s, Z_s) : s \leq t-1\}$ to provide an estimate of Y given (X, Z). At a high level, the test is initialized with toy money $S_0 = 1$ and proceeds as follows.

1. **Collect** a fresh test triplet (X_t, Y_t, Z_t).
2. **Generate a dummy feature** $\tilde{X}_t \sim P_{X|Z}(X_t \mid Z_t)$, and form the dummy triplet $(\tilde{X}_t, Y_t, Z_t)$.
3. **Compute a betting score** W_t. Use $\hat{f}_t$ to bet against the null, where the bet is that the prediction error of $\hat{f}_t$ (or any other test statistic), evaluated on the dummy triplet $(\tilde{X}_t, Y_t, Z_t)$, would be higher than that of the original triplet (X_t, Y_t, Z_t). A positive (resp. negative) score indicates that our bet is successful (resp. unsuccessful).
4. **Update the current wealth (test martingale)** S_t: if the betting score is positive, the previous S_{t-1} is increased by the money we risked on placing the bet; otherwise, the previous wealth S_{t-1} is decreased analogously.
5. **Update the predictive model** $\hat{f}_t$ and get $\hat{f}_{t+1}$, e.g., by taking one (or more) gradient steps to minimize a loss evaluated on $\{(X_s, Y_s, Z_s) : s \leq t\}$.
6. If $S_t \geq 1/\alpha$ reject H_0 and stop. Otherwise, increase t and return to step (1).

In what follows, we describe each of the above components in depth, define the proposed test martingale, and prove its validity. Later, in Section 4.2, we provide additional design principles that improve the power of the test.

Before doing so, we pause to provide a small synthetic experiment that showcases how the wealth process S_t behaves under the null and the alternative. To this end, we generate two different data sets. The first satisfies H_0, which we refer to as `null data` in which X is *un*important. The second satisfies the alternative, which we call `non-null data` in which X is important. The data generation process for each case and the implementation details are described in Section 5.1. Next, we apply e-CRT on each data set, and present in Figure 1 the wealth process S_t as a function of t. When the test is applied to the `null data`, the value of S_t remains close to the initial wealth $S_0 = 1$ for all presented time steps t. In particular, S_t does not exceed the value $1/\alpha = 20$, and thus H_0 cannot be rejected. By contrast, when the test is applied to the `non-null data`, the wealth process grows as the testing procedure proceeds, until reaching a target value of $1/\alpha = 20$. In this case, we reject the null and report that X is indeed important. This experiment illustrates the advantage of monitoring the value of S_t over time: we can safely terminate the test after collecting 300 samples and avoid a wasteful collection of new data.

4.1 Formulating the Test Martingale

Our procedure exploits the dummy feature $\tilde{X}_t$, sampled from the conditional distribution of $X \mid Z$ to form a fair game. In the sequel, we state key properties of the dummies, which we will use to define our test. The proofs of all statements given in this section are provided in Supplementary Section A. We start by emphasizing that we sample $\tilde{X}_t \sim P_{X|Z}(X_t \mid Z_t)$ without looking at Y_t, and so $\tilde{X}_t \perp\!\!\!\perp X_t \mid Z_t$ for all $t \in \mathbb{N}$ by construction. Therefore, X_t and its dummy $\tilde{X}_t$ are exchangeable conditional on Z_t; that is, $(X_t, \tilde{X}_t) \mid Z_t \overset{d}{=} (\tilde{X}_t, X_t) \mid Z_t$, where $\overset{d}{=}$ reads as 'equal in distribution'. This implies that it is impossible to distinguish between X_t and its dummy $\tilde{X}_t$ when viewing Z_t, for any time step t. Furthermore, under the null, this exchangeability property holds not only conditionally on Z_t but also on Y_t.

Lemma 1. *Take $(X_t, Y_t, Z_t) \sim P_{XYZ}$, and let $\tilde{X}_t$ be drawn independently from $P_{X|Z}$ without looking at Y_t. If $Y_t \perp\!\!\!\perp X_t \mid Z_t$, then $(X_t, \tilde{X}_t, Y_t, Z_t) \overset{d}{=} (\tilde{X}_t, X_t, Y_t, Z_t)$.*

The above result lies at the heart of the knockoff and CRT frameworks, and its proof follows (Candes et al., 2018, Lemma 3.2), (Barber et al., 2020, Lemma 1). Lemma 1 implies that, if the null is true, it is impossible to tell which is the original feature and which is the dummy when viewing the full observation, at any time step t. This result is essential for proving the validity of the CRT p-value, as well as for formulating our test martingale, as we do next.

Denote by $\mathcal{F}_t = \sigma(\{X_s, Y_s, Z_s\}_{s=1}^t)$ the sigma-algebra generated by observations collected up to time t, where $\mathcal{F}_0$ is the trivial sigma-algebra. Let $q_t = T(X_t, Y_t, Z_t; \hat{f}_t) \in \mathbb{R}$ and $\tilde{q}_t = T(\tilde{X}_t, Y_t, Z_t; \hat{f}_t) \in \mathbb{R}$ be the test statistics evaluated on the original and dummy triplets, respectively. Importantly, $T(\cdot; \hat{f}_t)$ can be any function, and its choice may affect the power of the test. For instance, one can define $T(\cdot; \hat{f}_t)$ as the squared prediction error evaluated on the current sample $T(x, y, z; \hat{f}_t) = (\hat{f}_t(x, z) - y)^2$ using a model $\hat{f}_t$ trained on past data $\{X_s, Y_s, Z_s\}_{s=1}^{t-1}$. Observe that $\hat{f}_t$ is not fitted on the new triplet (X_t, Y_t, Z_t), thus it is considered as a fixed function once conditioning on $\mathcal{F}_{t-1}$.

We then proceed by evaluating a **betting score**

$$W_t = g(q_t, \tilde{q}_t), \tag{4}$$

where the function $g(a,b) \in [-m,m]$ is antisymmetric $g(a,b) = -g(b,a)$, satisfying $g(a,b) > 0$ if $b > a$ and $g(a,b_1) \geq g(a,b_2)$ for $b_1 \geq b_2$. For example, $g(a,b) = m \cdot \text{sign}(b-a)$. The hyper-parameter $0 < m \leq 1$ controls the magnitude of the score. As in the knockoff filter, our design of g ensures it follows the *flip sign property*, requiring that a swap of the original feature X_t and its dummy $\tilde{X}_t$ will flip the sign of W_t (Candes et al., 2018).

Under the alternative, one should interpret a strictly positive betting score $W_t > 0$ as a successful bet, which will increase our wealth. This means that we gain some evidence that X_t carries extra predictive power about Y_t beyond what is already known in Z_t. Analogously, a strictly negative $W_t < 0$ indicates an erroneous bet, which will reduce our wealth even though the null is false. Crucially, under the null, W_t will be zero on average, no matter how accurate $\hat{f}_t$ is. In other words, it is impossible to have a systematically positive W_t when H_0 is true.

Lemma 2. *Under the same conditions as in Lemma 1, if H_0 is true then $\mathbb{E}_{H_0}[W_t \mid \mathcal{F}_{t-1}] = 0$ for all $t \in \mathbb{N}$.*

The core idea behind the proof of the above lemma is that, under the null, W_t has a symmetric distribution about zero conditional on $\mathcal{F}_{t-1}$, and thereby its expected value is zero; see (Ramdas et al., 2020) for a related property of symmetric distributions. In particular, W_t is equally likely to have positive and negative values, which is a well-known result in the knockoff literature with the important difference that in our case we show it holds conditionally on $\mathcal{F}_{t-1}$.

Armed with the betting score W_t at time t, we turn to define a test martingale $\{S_t : t \in \mathbb{N}_0\}$ for H_0. The martingale can be thought of as the wealth process, initialized by toy money $S_0 = 1$, and our ultimate goal is to maximize it. We begin with defining the *base martingale* as follows:

$$S_t^v := \prod_{j=1}^{t} (1 + v \cdot W_j), \tag{5}$$

where $v \in [0,1]$ is a fixed amount of toy money that we are willing to risk at step t.[1] Proposition 2 in Supplementary Section C.1 shows that $\{S_t^v : t \in \mathbb{N}_0\}$ in (5) is a valid test martingale. As a result, following Ville's inequality in (3), one can monitor S_t^v and control the type-I error for any choice of $v \in [0,1]$. Importantly, the amount of toy money v that we risk when placing the bet affects the power.

The above immediately raises the question of how should we choose v? Ideally, we want to set the best constant v^* so that $S_t^{v^*}$ is maximized under the alternative. The problem is that we are not allowed to look at the current betting score W_t, so it is impossible to find such an ideal data-dependent v^* in foresight. As a thought experiment, consider the simplest choice for g as the sign function for which $W_t \in \{+1,-1\}$, and suppose we adopt an aggressive betting strategy with $v = 1$. With this choice, when we win a bet we will increase S_t^v by the maximal amount possible at step t. However, if we lose a bet even once, we will have $S_t^v = 0$, resulting in a powerless test; to see this, assign $W_t = -1$ in (5). We give a concrete example that visualizes this discussion in Supplementary Section C.2.

As a way out, we formulate a powerful betting strategy using the mixture-method of Shekhar and Ramdas (2021), which is intimately connected to universal portfolio optimization (Cover, 2011). The mixture-method is defined as the average over S_t^v for all $v \in [0,1]$:

$$S_t = \int_0^1 S_t^v \cdot h(v) dv, \tag{6}$$

where $h(v)$ is a probability density function (pdf) whose support is on the $[0,1]$ interval, e.g., a uniform distribution. We adopt the mixture method betting strategy to formulate our test martingale since it has appealing power properties, which we discuss soon. Before doing so, however, we shall first prove that the test martingale in (6) is valid. The theorem presented below states that by monitoring S_t one can safely reject the null the first time S_t exceeds $1/\alpha$, while rigorously controlling the type-I error simultaneously for all optional stopping times. This result holds in finite samples, without making any modeling assumptions on the conditional distribution of $Y \mid X$, and for any machine learning model $\hat{f}_t$, which we use to bet against the null. The proof follows (Shekhar and Ramdas, 2021, Section 2.2).

Theorem 1. *Under the same conditions as in Lemma 1, if the null hypothesis H_0 is true then for any $\alpha \in (0,1)$,*

$$\mathbb{P}_{H_0}(\exists t : S_t \geq 1/\alpha) \leq \alpha.$$

Having established the validity of the test, we turn to discuss the key advantage of the mixture method betting strategy. The idea behind this approach is that one of the base martingales S_t^v in (6) must hit the best constant v^*, which, in turn, drives the average martingale S_t upwards. We demonstrate this visually in Supplementary Section C.2. In fact, Shekhar and Ramdas (2021) proved that S_t is not only dominated by $S_t^{v^*}$, but can also provably form a consistent test that achieves *power one* in the limit of infinite data.

Proposition 1 (Shekhar and Ramdas (2021))**.** *If $\liminf_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} W_s > 0$ under the alternative H_1. Then, $\mathbb{P}_{H_1}(\exists t : S_t \geq 1/\alpha) = 1$ for any $\alpha \in (0,1)$.*

The condition of $\liminf_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} W_s > 0$ implies that it suffices that only on average the predictive model will be able to tell apart the original and dummy triplets, so at the limit of infinite data we will attain a consistent test.

[1] We can set a different v_t for each time step, yet v_t must be chosen without looking at the current (X_t, Y_t, Z_t) as otherwise the test will cease to be valid. Intuitively, in such a case one can always set $v_t = 0$ when W_t is negative and $v_t = 1$ otherwise, and increase the wealth regardless on whether the null is true or false.

Algorithm 1 Betting score evaluation

Input: Data batch $\{(X_s, Y_s, Z_s)\}_{s=1}^{b}$; conditional distribution $P_{X|Z}$; test statistic $T(\cdot)$; fixed predictive model $\hat{f}$; de-randomization parameter K; betting score function $g(\cdot)$.

Compute $q \leftarrow T(\{(X_s, Y_s, Z_s)\}_{s=1}^{b}; \hat{f})$
for $k = 1, \ldots, K$ **do**
  Sample $\tilde{X}_s \sim P_{X|Z}(X_s \mid Z_s)$ for $s = 1, \ldots, b$
  Compute $\tilde{q} \leftarrow T(\{(\tilde{X}_s, Y_s, Z_s)\}_{s=1}^{b}; \hat{f})$
  Compute a betting score $W^{(k)} \leftarrow g(q, \tilde{q})$

Output An average betting score $W \leftarrow \frac{1}{K}\sum_{k=1}^{K} W^{(k)}$

Algorithm 2 e-CRT: sequential test for CI

Input: Data stream $(X_t, Y_t, Z_t), t = 1, \ldots$; test level $\alpha \in (0, 1)$; set of batch sizes $\mathcal{B}$.

Train a predictive model $\hat{f}_1$ on an initial batch of data $\{(X_{t'}, Y_{t'}, Z_{t'})\}_{t'=1}^{n_{\text{init}}}$
Set $S_{0,b} \leftarrow 1$ for all $b \in \mathcal{B}$
for $t = 1, 2, \ldots$ **do**
  for b in $\mathcal{B}$ **do**
    Set $t' \leftarrow \lfloor t/b \rfloor$
    Set $\mathcal{D}_{t,b} = \{(X_j, Y_j, Z_j)\}_{j=(t'-1)\cdot b+1}^{t'\cdot b}$
    Compute the average betting score $W_{t',b}$ by applying Algorithm 1 to $\mathcal{D}_{t,b}$ with $\hat{f}_{(t'-1)\cdot b+1}$
    Update $S_{t,b} \leftarrow \int_0^1 \prod_{s=1}^{t'}(1 + v \cdot W_{s,b}) \cdot h(v)dv$
  Compute the ensemble-over-batches martingale $S_t \leftarrow \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} S_{t,b}$
  if $S_t \geq 1/\alpha$ **then**
    Reject the null hypothesis H_0 and stop
  else
    Obtain $\hat{f}_{t+1}$ by online updating $\hat{f}_t$ after adding the most recent point (X_t, Y_t, Z_t) to the train set

4.2 Practical Considerations

In this section, we provide design principles that improve the power of the test while maintaining its validity. For ease of reference, we provide a pseudo code that implements the following ideas in Algorithm 2.

De-randomization. To reduce the algorithmic randomness induced by the generated dummy feature $\tilde{X}_t$, we (i) sample $K > 1$ independent dummy copies of X_t; (ii) compute the corresponding betting scores $W_t^{(k)}$, $k = 1, \ldots, K$; and (iii) evaluate the average betting score $W_t = \frac{1}{K}\sum_{k=1}^{K} W_t^{(k)}$. We refer to K as the de-randomization hyper-parameter. Importantly, this strategy preserves the validity of the test martingale S_t in (6), since the expected value of the average betting score is also equal to zero under the null, i.e., $\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{H_0}[W_t^{(k)} \mid \mathcal{F}_{t-1}] = 0$. The ablation study presented in Section 5.3 demonstrates that the above de-randomization procedure improves the power of the test, even for a moderate choice of K.

Ensemble over batches. The betting score W_t presented in (4) is evaluated on a *single* data point. This choice, however, might be inferior to evaluating W_t on a *batch* of several data points, as working with a batch is often less sensitive to the randomness in the data. On the other hand, a larger batch size reduces the total number of updates of the wealth process S_t that we can make, and this may result in slower growth of the total wealth for a given number of data points. To mitigate the above trade-off, we suggest an ensemble approach by formulating the test martingale

$$S_t = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} S_{t,b}, \quad t \in \mathbb{N} \tag{7}$$

as an average of $|\mathcal{B}|$ test martingales $S_{t,b}$, where $\mathcal{B}$ is a set containing the batch-sizes and $|\cdot|$ returns the set size. Each of the batch martingales $S_{t,b}$ is evaluated analogously to (6), however on a batch of size b instead of a single data point. We refer the reader to Supplementary Section F for more details on the ensemble approach, where we rigorously explain why (7) is a valid test martingale. This procedure is summarized in Algorithm 2. The ablation study in Section 5.3 demonstrates the above trade-off and the advantage of our ensemble procedure.

Unknown conditional. As in the CRT, to generate the dummy $\tilde{X}$ we assume that we have access to $P_{X|Z}$, however, in many real applications this distribution may not be known precisely. Here, we briefly discuss several use-cases where it is sensible to assume we have reasonable knowledge about $X \mid Z$, and we refer the reader to Candes et al. (2018) for a more detailed discussion. One such use-case is controlled experiments, e.g., genetic crossing experiments (Haldane and Waddington, 1931), sensitivity analysis of numerical models (Saltelli et al., 2008), and gene knockout experiments (Cong et al., 2013). Another important use-case is when we have a large number of unlabeled observations of (X, Z), so we can estimate $P_{X|Z}$ before applying the test. This is a reasonable assumption in various economic or genetic applications as we can collect covariates from different populations, or leverage previous studies that have acquired the same (X, Z) however with different response variables. In such situations, we can utilize powerful machine learning techniques to estimate $P_{X|Z}$ using the available data, as suggested in previous work on model-X tests (Tansey et al., 2022; Gimenez et al., 2019; Bellot and van der Schaar, 2019; Romano et al., 2020; Sesia et al., 2019). We will rely on such ideas in our experiments with real data. Importantly, the above line of research demonstrates the robustness of the CRT and knockoffs to errors in the estimation of $P_{X|Z}$. We believe it will be striking to provide a rigorous robustness theory for our e-CRT, possibly by following the approach presented in (Barber et al., 2020; Berrett et al., 2020).

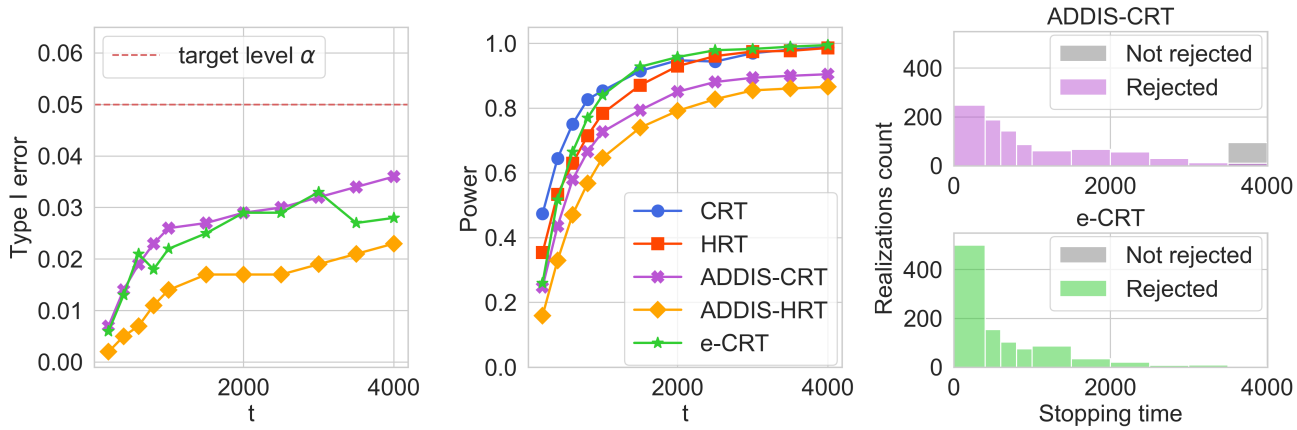


Figure 2: **Performance evaluation with simulated data**, evaluated over 1000 independent trials. Left: type-I error. Middle: empirical power. Right: histogram of the stopping times evaluated on the `non-null data`. We set $\alpha = 0.05$.

5 SYNTHETIC EXPERIMENTS

5.1 Experimental Setup

In this section, we evaluate the performance of e-CRT in a controlled synthetic setting, where $P_{X|Z}$ and $P_{Y|XZ}$ are known. We generate a sequence of i.i.d. data points $\{(X_t, Y_t, Z_t)\}_{t=1}^{n}$, where n is the maximal number of samples that can be collected. The features X_t, Z_t are sampled as follows: $Z_t \sim \mathcal{N}(0, I_d)$ and $X_t \mid Z_t \sim \mathcal{N}(u^\top Z_t, 1)$ where $u \sim \mathcal{N}(0, I_d)$, so as X_t and Z_t are dependent by design (Javanmard and Mehrabi, 2021). Unless specified otherwise, we fix the number of covariates to $d = 19$, so in total we have 20 features. Next, we consider two different conditional models for $P_{Y|XZ}$. The first model is used to examine the validity of our test by evaluating the type-I error rate, which we refer to as `null data` model. Specifically, we sample Y_t such that $Y_t \perp\!\!\!\perp X_t \mid Z_t$ as follows: $Y_t \mid X_t, Z_t \sim \mathcal{N}((w^\top Z_t)^2, 1)$, where $w \sim \mathcal{N}(0, 1)$. To evaluate the empirical power—i.e., the rate we reject H_0 when applying a test with significance level α—we define the following `non-null data` model in which $Y_t \not\perp\!\!\!\perp X_t \mid Z_t$ by construction: $Y_t \mid X_t, Z_t \sim \mathcal{N}((w^\top Z_t)^2 + 3X_t, 1)$, where $w \sim \mathcal{N}(0, 1)$. We compare the performance of our e-CRT to offline CRT (Candes et al., 2018), detailed in Section 2, as well as to the holdout randomization test (HRT) (Tansey et al., 2022), which is a computationally efficient variant of the offline CRT that often comes at the price of reduced power due to data splitting. All the methods use lasso regression model to compute the test statistics, whereas in e-CRT we fit the model online as described in Supplementary Section E. Additional implementation details on all methods are provided in Supplementary Section G.1.

We also compare e-CRT to out-of-the-box sequential versions of the offline CRT/HRT that allows monitoring the p-value p_t over time. Towards that end, we apply the state-of-the-art ADDIS-spending approach (Tian and Ramdas, 2021), which rigorously adjusts the p-value at time t by accounting for multiple comparisons in the time horizon, controlling H_0. We refer to the ADDIS-spending version of the CRT and HRT as ADDIS-CRT and ADDIS-HRT; see Supplementary Section G.1 for implementation details. Unfortunately, we find it infeasible to generate a p-value p_t for each time step t due to the high computational complexity of the CRT: it requires fitting M predictive models from scratch for each t, where we set $M = 1000$ in our experiments to have a reasonable resolution for the p-value corrected by ADDIS-spending. Therefore, we apply ADDIS on p-values evaluated using CRT/HRT over a grid of 11 times steps in total.

5.2 Type-I Error, Power, and Early Stopping

Type-I error. Recall Figure 1 from Section 4, illustrating that the test martingale S_t does not grow significantly over time for a single realization of the `null data`. Here, we expand this experiment by reporting the type-I error rates of all the sequential tests as a function of t, evaluated on 1000 independent realizations of the `null data` model. Following Figure 2 (left), we can see how the type-I error of our e-CRT is controlled and falls below the level $\alpha = 0.05$ for all time steps t, as expected. The same conclusion holds for the sequential tests ADDIS-CRT and ADDIS-HRT. We also present in Supplementary Figure 6 the type-I error of the offline CRT and offline HRT, showing these also control the type-I error rate, however for a fixed sample size t.

Robustness. In practice, we do not have access to the sampling distribution of $X \mid Z$, and thus it is important to study the robustness of the test to approximation error in the sampling of $\tilde{X}$. In Supplementary Section G.5.1 we conduct such an experiment, showing that inflation in the type-I error occurs only when the estimation of $P_{X|Z}$ is far from the true distribution. Further, in Supplementary

Section G.5.2 we consider a more challenging $X \mid Z$ that follows a student-t distribution and show how a recent non-parametric method (Rosenberg et al., 2022) can be used to effectively estimate the conditional distribution. There, we demonstrate how the type-I error is controlled, and also that the power grows with t. A related experiment is given in Section 6, where we apply our method to real data for which $P_{X|Z}$ is unknown and thus must be estimated from data.

Power. The middle panel in Figure 2 presents the empirical power as a function of t, evaluated on 1000 independent realizations of the `non-null data` model. Observe how the offline tests outperform the sequential ones when the sample size t is relatively small. Yet, for a larger number of samples with $t > 500$, the e-CRT tends to outperform the offline HRT; this may be due to the sample inefficiency of the HRT as it involves data splitting. Interestingly, the e-CRT has comparable performance to CRT for $t > 1000$, and the three tests nearly achieve power one when the sample size is large. Importantly, the e-CRT outperforms both ADDIS-CRT and ADDIS-HRT for all t, highlighting the advantages of our specialized test for CI compared to more out-of-the-box sequential solutions.

Early stopping. The strength of any sequential test—including our e-CRT—is the ability to monitor the outcome of the test and reject the null as soon as it exceeds a predefined threshold. Of course, the earlier the rejection happens the better the sample efficiency of the test. The right panel in Figure 2 presents the histogram of the stopping times of ADDIS-CRT and e-CRT over the 1000 realizations of the `non-null data` used in the power experiments. As can be seen, e-CRT tends to reject the null earlier than ADDIS-CRT, indicating superior sample efficiency.

Additional experiments. Supplementary Section G.3 studies the effect of the number of covariates d on the performance of e-CRT, where the overall trend is that the power is decreased as d is increased, and the type-I error is controlled. We also examine the impact of the dependency strength between X and Z on the performance of our e-CRT in Supplementary Section G.4. There, we observe a decrease in power as the correlation increases, while controlling the type-I error. This trend is in line with previous studies that focus on the offline setting, see, e.g., Candes et al. (2018); Liu et al. (2022); Shaer and Romano (2023).

5.3 Ablation Study

The effect of the de-randomization parameter K. In Section 4.2 we suggest using an average betting score over K realizations of W_t to reduce the randomization induced by the generation of the dummy features. To study the effect of the de-randomization parameter K on the power of the e-CRT, we generate `non-null data` stream of length $n = 2000$ and present in Figure 3 (top) the power of e-CRT as a function of t for several values of K. It is evident that the power of the test is improved as K increases, with a maximal absolute improvement of about 20%.

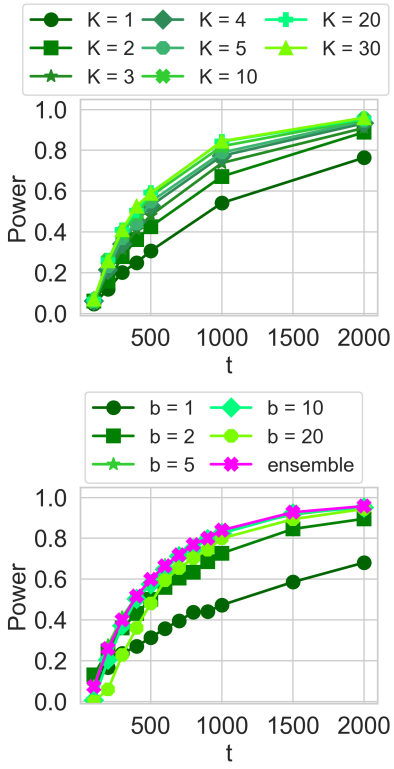


Figure 3: **Ablation study.** Empirical power of e-CRT ($\alpha = 0.05$), evaluated on 1000 realizations of the `non-null data`. Top: the effect of the de-randomization parameter K on power. Bottom: the effect of the batch size on power, in comparison to the batch-ensemble approach in (7).

The effect of the batch size. In Section 4.2 we discuss how different batch sizes might affect the performance of the e-CRT, and describe the trade-off between using small and large batch sizes. To visualize this trade-off, we apply the e-CRT on `non-null data` with and without the ensemble-over-batches approach, for several choices of batch sizes. Following Figure 3 (bottom), we can see that, for smaller sample sizes, e-CRT with small batches performs better than with large batches in terms of power. However, the opposite is true for larger sample sizes, in which larger batches are favored. Our proposed ensemble approach achieves a remarkable performance: it tends to follow the leading choice for all sample sizes tested.

6 REAL DATA EXPERIMENTS

6.1 Fund Replication

We begin with the task of identifying which stocks contribute to the performance of known index funds. Our study

follows Challet et al. (2021); Spector and Fithian (2022) who deployed the knockoff filter in this context. Although the samples are not i.i.d. and thus not satisfying the model-X assumptions, we present this application to illustrate how the e-CRT can be applied to real data. In our experiments, we focus on a technology sector index fund named XLK, and follow the data collection procedure of Spector and Fithian (2022). The data consists of the daily log returns of each stock in the S&P 500 since 2013 and the corresponding daily log return of XLK. We exclude samples and features with missing data, resulting in $n = 2421$ and $d = 457$. More details are in Supplementary Section H.1.1.

Table 1 in Supplementary Section H.1.2 summarizes the results obtained by applying e-CRT, CRT and HRT on each of the $d = 457$ stocks. We classify a stock as important if it currently belongs to the technology sector XLK. We report the p-values obtained by CRT and HRT for each stock, and the test martingale $S_{t_{\text{stop}}}$ for e-CRT, where t_{stop} is the stopping time for a test level of $\alpha = 0.05$. Following Supplementary Table 1, we can see that the e-CRT tends to reject the null for stocks that are currently in XLK, and avoids rejecting stocks that are currently not in XLK. Observe the advantage of early stopping: the e-CRT rejects some of the important stocks with a relatively small sample size t_{stop}.

6.2 HIV Drug Resistance

Herein, we consider the task of detecting genetic mutations in human immunodeficiency virus (HIV) of type-I that are associated with drug resistance (Rhee et al., 2006). We follow Romano et al. (2020) and study the resistance to the Lopinavir protease inhibitor drug, applying the same preprocessing steps to the raw data. Consequently, this data set consists of $n = 1555$ samples of $d = 150$ features. We consider the data points as if they arrive sequentially and apply the test on each feature. See Supplementary Section H.2.1 for details on the data and the tests we apply.

Table 2 in Supplementary Section H.2.2 summarizes the results obtained by running e-CRT, CRT, and HRT on each of the $d = 150$ mutations, analogously to the fund replication experiment. We classify each mutation by its effect on drug resistance, as reported in previous studies. Following Supplementary Table 2, we can see that our e-CRT tends to reject the null for mutations that have been previously reported to have a 'major' or 'accessory' effect and avoid the 'unknown' ones. The e-CRT rejects some of the 'major'/'accessory' mutations with a relatively small sample size t_{stop}, demonstrating the advantage of early stopping. Figure 11 in Supplementary Section H.2.2 portrays three test martingales for representative mutations. Figure 11a corresponds to a mutation that has not been reported in previous studies to have an effect on drug resistance. Indeed, the test martingale does not grow significantly. By contrast, Figure 11b corresponds to a mutation that has been reported to have a major effect on drug resistance, for which S_t grows fast and reaches $1/\alpha = 20$ using only 240 samples. Lastly, Figure 11c corresponds to a mutation that has been also reported to have a major effect. Here, S_t grows at a slower rate and does not reach the nominal level of $1/\alpha = 20$. Yet, it achieves a final value of 6.5, which provides substantial evidence against the null (Vovk and Wang, 2021).

7 CONCLUSION

In this paper, we develop e-CRT—a sequential CI test that allows processing each data point as soon as it arrives, supporting early stopping while controlling the type-I error for all t simultaneously. Our proposed test is inspired by the model-X randomization test Candes et al. (2018), and testing by betting Shafer and Vovk (2019); Shafer (2021). We prove the validity of e-CRT and propose several design choices to improve its power, including de-randomization and batch ensemble. Numerical experiments demonstrate the validity of our e-CRT, its superiority over existing out-of-the-box sequential tests for CI, and the impact of the design choices we make on the power of the test.

One important future direction is to theoretically analyze the robustness of the e-CRT to errors in the estimation of $P_{X|Z}$, rigorously characterizing the potential inflation in type-I error rate Barber et al. (2020); Grünwald et al. (2022). From a practical perspective, it would be illuminating to develop more robust betting scores, for example, by taking into account the error in the estimation of $P_{X|Z}$ or by borrowing ideas from the doubly robust literature Shah and Peters (2020); Shi et al. (2021); Niu et al. (2022). Another direction is to explore new ways to fit predictive models and form more powerful betting scores. In our experiments, we train the predictive model on the original data to minimize the MSE while ignoring the dummy features. Recently, in the context of HRT, Shaer and Romano (2023) developed a new loss function that takes into account the dummy features during training to improve the power of the test. We believe such an approach can be used in our context as well.

Acknowledgements

This research was supported by the ISRAEL SCIENCE FOUNDATION (grant No. 729/21). Y.R. also thanks the Career Advancement Fellowship, Technion, for providing additional research support.

References

J. D. Angrist and G. M. Kuersteiner. Causal effects of monetary shocks: Semiparametric conditional independence tests with a multinomial propensity score. *Review of Economics and Statistics*, 93(3):725–747, 2011.

A. Balsubramani and A. Ramdas. Sequential nonparametric testing with the law of the iterated logarithm. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI)*, 2016.

R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5): 2055–2085, 2015.

R. F. Barber, E. J. Candès, and R. J. Samworth. Robust inference with knockoffs. *The Annals of Statistics*, 48(3): 1409–1431, 2020.

S. Bates, M. Sesia, C. Sabatti, and E. Candès. Causal inference in genetic trio studies. *Proceedings of the National Academy of Sciences*, 117(39):24117–24126, 2020.

A. Bellot and M. van der Schaar. Conditional independence testing using generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.

T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 175–197, 2020.

R. Bhui. Testing optimal timing in value-linked decision making. *Computational Brain & Behavior*, 2(2):85–94, 2019.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

C. Burns, J. Thomason, and W. Tansey. Interpreting black box models via hypothesis testing. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pages 47–57, 2020.

E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

D. Challet, C. Bongiorno, and G. Pelletier. Financial factors selection with knockoffs: fund replication, explanatory and prediction networks. *Physica A: Statistical Mechanics and its Applications*, 580:126105, 2021.

L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, et al. Multiplex genome engineering using crispr/cas systems. *Science*, 339(6121):819–823, 2013.

T. M. Cover. Universal portfolios. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pages 181–209. World Scientific, 2011.

B. Duan, A. Ramdas, and L. Wasserman. Interactive rank testing by betting. In *Conference on Causal Learning and Reasoning*, pages 201–235. PMLR, 2022.

J. R. Gimenez, A. Ghorbani, and J. Zou. Knockoffs for the mass: new feature importance statistics with false discovery guarantees. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2125–2133. PMLR, 2019.

P. Grünwald, R. de Heide, and W. M. Koolen. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–54. IEEE, 2020.

P. Grünwald, A. Henzi, and T. Lardy. Anytime valid tests of conditional independence under model-x. *arXiv preprint arXiv:2209.12637*, 2022.

J. Haldane and C. Waddington. Inbreeding and linkage. *Genetics*, 16(4):357, 1931.

A. Javanmard and M. Mehrabi. Pearson chi-squared conditional randomization test. *arXiv preprint arXiv:2111.00027*, 2021.

K.-S. Jun and F. Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Conference on Learning Theory*, pages 1802–1823. PMLR, 2019.

T. Lai. Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. *Communications in Statistics-Theory and Methods*, 13(19):2355–2368, 1984.

L. Lei and W. Fithian. Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679, 2018.

L. Lei, A. Ramdas, and W. Fithian. A general interactive framework for false discovery rate control under structural constraints. *Biometrika*, 108(2):253–267, 2021.

A. Lhéritier and F. Cazals. A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370, 2018.

M. Liu, E. Katsevich, L. Janson, and A. Ramdas. Fast and powerful conditional randomization testing via distillation. *Biometrika*, 109(2):277–293, 2022.

Y. Lu, Y. Fan, J. Lv, and W. Stafford Noble. DeepPINK: reproducible feature selection in deep neural networks. *Advances in neural information processing systems*, 31, 2018.

M. Naghshvar and T. Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013.

A. Nikolakopoulou, D. Mavridis, M. Egger, and G. Salanti. Continuously updated network meta-analysis and statistical monitoring for timely decision-making. *Statistical methods in medical research*, 27(5):1312–1330, 2018.

Z. Niu, A. Chakraborty, O. Dukes, and E. Katsevich. Reconciling model-x and doubly robust approaches to conditional independence testing. *arXiv preprint arXiv:2211.14698*, 2022.

J. J. Park, K. Thorlund, and E. J. Mills. Critical concepts in adaptive clinical trials. *Clinical epidemiology*, 10:343, 2018.

J. Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

A. Ramdas and L. Wenbe. The lady keeps tasting coffee: randomization inference by betting. *Unpublished manuscript*, 2020.

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

A. Ramdas, J. Ruf, M. Larsson, and W. M. Koolen. Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109, 2022.

Z. Ren and R. F. Barber. Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*, 2022.

S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D. L. Brutlag, and R. W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46): 17355–17360, 2006.

Y. Romano, M. Sesia, and E. Candès. Deep knockoffs. *Journal of the American Statistical Association*, 115 (532):1861–1872, 2020.

A. A. Rosenberg, S. Vedula, Y. Romano, and A. M. Bronstein. Fast nonlinear vector quantile regression. *arXiv preprint arXiv:2205.14977*, 2022.

A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

M. Sesia, C. Sabatti, and E. J. Candès. Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1):1–18, 2019.

S. Shaer and Y. Romano. Learning to increase the power of conditional randomization tests. *Machine Learning*, pages 1–41, 2023.

G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021.

G. Shafer and V. Vovk. *Game-Theoretic Foundations for Probability and Finance*, volume 455. John Wiley & Sons, 2019.

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. 2020.

S. Shekhar and A. Ramdas. Game-theoretic formulations of sequential nonparametric one-and two-sample tests. *arXiv preprint arXiv:2112.09162*, 2021.

C. Shi, T. Xu, W. Bergsma, and L. Li. Double generative adversarial networks for conditional independence testing. *The Journal of Machine Learning Research*, 22 (1):13029–13060, 2021.

A. Spector and W. Fithian. Asymptotically optimal knockoff statistics via the masked likelihood ratio. 2022.

W. Tansey, V. Veitch, H. Zhang, R. Rabadan, and D. M. Blei. The holdout randomization test for feature selection in black box models. *Journal of Computational and Graphical Statistics*, 31(1):151–162, 2022.

J. Tian and A. Ramdas. Online control of the familywise error rate. *Statistical Methods in Medical Research*, 30 (4):976–993, 2021.

L. Tonelli. Sull'integrazione per parti. *Rend. Acc. Naz. Lincei*, 5(18):246–253, 1909.

R. Turner, A. Ly, and P. Grünwald. Two-sample tests that are safe under optional stopping, with an application to contingency tables. *arXiv preprint arXiv:2106.02693*, 2021.

J. Ville. *1ere these: Etude critique de la notion de collectif; 2eme these: La transformation de Laplace*. PhD thesis, Gauthier-Villars & Cie, 1939.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

A. Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 1945.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal statistical society: series B (Methodological)*, 2022.

X. Wang and Y. Hong. Characteristic function based testing for conditional independence: A nonparametric regression approach. *Econometric Theory*, 34(4):815–849, 2018.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Methodology), with discussion*, 2023.

S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

Supplementary Material: Model-X Sequential Testing for Conditional Independence via Testing by Betting

A MATHEMATICAL PROOFS

Proof of Lemma 1. Observe that it is equivalent to showing that

$$(X_t, \tilde{X}_t, Y_t) \mid Z_t \overset{d}{=} (\tilde{X}_t, X_t, Y_t) \mid Z_t, \tag{8}$$

since the marginal distribution P_Z is identical on both sides of (8). Below, we use discrete random variables for simplicity, as the continuous case can be proved analogously. From the law of total probability, we can write relation (8) as follows:

$$\mathbb{P}_{Y|X\tilde{X}Z}(y \mid a, b, z) \cdot \mathbb{P}_{X\tilde{X}|Z}(a, b \mid z) = \mathbb{P}_{Y|\tilde{X}XZ}(y \mid a, b, z) \cdot \mathbb{P}_{\tilde{X}X|Z}(a, b \mid z). \tag{9}$$

Now, recall that $(X_t, \tilde{X}_t) \mid Z_t \overset{d}{=} (\tilde{X}_t, X_t) \mid Z_t$ by construction, therefore the conditional distributions $\mathbb{P}_{\tilde{X}X|Z}$ and $\mathbb{P}_{X\tilde{X}|Z}$ on both sides of (9) are the same. As a result, it suffices to show that

$$Y_t \mid (X_t, \tilde{X}_t, Z_t) \overset{d}{=} Y_t \mid (\tilde{X}_t, X_t, Z_t).$$

The above relation holds once observing that

$$\begin{aligned}\mathbb{P}_{Y|X\tilde{X}Z}(y \mid a, b, z) &= \mathbb{P}_{Y|Z}(y \mid z) \\ &= \mathbb{P}_{Y|\tilde{X}XZ}(y \mid a, b, z),\end{aligned}$$

where the first and second equality hold since $Y_t \perp\!\!\!\perp X_t \mid Z_t$, $Y_t \perp\!\!\!\perp \tilde{X}_t \mid Z_t$, and $X_t \perp\!\!\!\perp \tilde{X}_t \mid Z_t$, implying that $Y_t \perp\!\!\!\perp (X_t, \tilde{X}_t) \mid Z_t$. This completes the proof. □

Proof of Lemma 2. Observe that the predictive model $\hat{f}_t$ is a fixed function given $\mathcal{F}_{t-1}$, as it is fitted to $\{(X_s, Y_s, Z_s)\}_{s=1}^{t-1}$. Thus, we can invoke Lemma 1, implying that under the null

$$g(q_t, \tilde{q}_t) \mid \mathcal{F}_{t-1} \overset{d}{=} g(\tilde{q}_t, q_t) \mid \mathcal{F}_{t-1}, \tag{10}$$

as $\hat{f}_t$ is a fixed function. Now, recall that $g(\cdot)$ is an antisymmetric function, i.e.,

$$g(q_t, \tilde{q}_t) = -g(\tilde{q}_t, q_t), \tag{11}$$

and observe that by combining (10) and (11) we get the following

$$g(q_t, \tilde{q}_t) \mid \mathcal{F}_{t-1} \overset{d}{=} -g(q_t, \tilde{q}_t) \mid \mathcal{F}_{t-1}.$$

This implies that, under the null, the density function of $g(q_t, \tilde{q}_t) \mid \mathcal{F}_{t-1}$ is symmetric about 0, and therefore

$$\mathbb{E}_{H_0}[g(q_t, \tilde{q}_t) \mid \mathcal{F}_{t-1}] = \mathbb{E}_{H_0}[W_t \mid \mathcal{F}_{t-1}] = 0.$$

□

Proof of Theorem 1. Note that $S_0 = 1$ and S_t in (6) is non-negative for all $t = \mathbb{N}$ by construction. According to Ville's inequality (3), it is enough to show that $\{S_t : t \in \mathbb{N}_0\}$ is a supermartingale under H_0 with respect to the filtration $\{\mathcal{F}_{t-1} : t \in \mathbb{N}\}$. This statement holds true since

$$\begin{aligned}
\mathbb{E}_{H_0}[S_t \mid \mathcal{F}_{t-1}] &= \mathbb{E}_{H_0}\left[\int_0^1 \prod_{s=1}^{t}(1 + v \cdot W_j) \cdot h(v)dv \mid \mathcal{F}_{t-1}\right] \\
&= \int_0^1 \prod_{j=1}^{t-1}(1 + v \cdot W_j) \cdot \mathbb{E}_{H_0}[1 + v \cdot W_t \mid \mathcal{F}_{t-1}] \cdot h(v)dv \\
&= \int_0^1 \prod_{j=1}^{t-1}(1 + v \cdot W_j) \cdot (1 + v \cdot \mathbb{E}_{H_0}[W_t \mid \mathcal{F}_{t-1}]) \cdot h(v)dv \\
&= \int_0^1 \prod_{j=1}^{t-1}(1 + v \cdot W_j) \cdot h(v)dv = S_{t-1}.
\end{aligned}$$

The second equality is due Tonelli's theorem Tonelli (1909), as $\prod_{j=1}^{t}(1 + v \cdot W_j) \cdot h(v)$ is non-negative for all $t \in \mathbb{N}$, and $h(v)$ is a probability density function. The last equality holds by invoking Lemma 2. □

B THE OFFLINE CONDITIONAL RANDOMIZATION TEST

Algorithm 3 Offline Conditional Randomization Test

Input: Data $\{(X_i, Y_i, Z_i)\}_{i=1}^n$; test statistic $T(\cdot)$; number of iterations M.

1: Set $t \leftarrow T(\{(X_i, Y_i, Z_i)\}_{i=1}^n)$
2: **for** $m = 1, \ldots, M$ **do**
3: Sample dummy variables $\tilde{X}_i \sim P_{X|Z}(X_i \mid Z_i)$ for $i = 1, \ldots, n$
4: Set $\tilde{t}^{(m)} \leftarrow T(\{\tilde{X}_i, Y_i, Z_i)\}_{i=1}^n)$

Output: A p-value $p_n = \frac{1}{1+M}\big(1 + \sum_{m=1}^{M} \mathbb{1}\{\tilde{t}^{(m)} \leq t\}\big)$.

C SUPPLEMENTARY DETAILS ON THE PROPOSED METHOD

C.1 Validity of the Base Martingale

In Section 4.1 we formulate the base martingale S_t^v in (5). Here, we prove in Proposition 2 that $\{S_t^v : t \in \mathbb{N}_0\}$ is a valid test martingale, according to Definition 1, for any $v \in [0, 1]$.

Proposition 2. *The base martingale* $\{S_t^v : t \in \mathbb{N}_0\}$ (5) *is a valid test martingale w.r.t the filtration* $\{\mathcal{F}_{t-1} : t \in \mathbb{N}\}$*, i.e., satisfying Definition 1, for any constant* $v \in [0, 1]$.

Proof. Let $v \in [0, 1]$. Note that $S_0^v = 1$ and $v \cdot W_t \geq 0$ for any $t \geq 1$, hence $S_t^v \geq 0, \forall t \in \mathbb{N}_0$ by construction. To conclude the prof, we show that $\{S_t^v : t \in \mathbb{N}_0\}$ is a supermartingale under the null with respect to $\{\mathcal{F}_{t-1} : t \in \mathbb{N}\}$:

$$\mathbb{E}_{H_0}[S_t^v \mid \mathcal{F}_{t-1}] = \prod_{j=1}^{t-1}(1 + v \cdot W_j) \cdot \mathbb{E}_{H_0}[1 + v \cdot W_t \mid \mathcal{F}_{t-1}] = \prod_{j=1}^{t-1}(1 + v \cdot W_j) \cdot (1 + v \cdot \mathbb{E}_{H_0}[W_t \mid \mathcal{F}_{t-1}]) = S_{t-1}^v.$$

□

C.2 Further Details on the Synthetic Experiment from Section 4.1

Here, we conduct an experiment that visualizes the advantage of the mixture-method S_t (6) over a constant v in S_t^v (5). To this end, we present in Figure 4 the wealth process S_t^v obtained for several values of v and the mixture-method test martingale S_t (6), for two different realizations of the `non-null data` generating model from Figure 1. We construct S_t by applying Algorithm 2 to the generated data, with the choice of $h(v)$ as the pdf of the uniform distribution on $[0, 1]$.

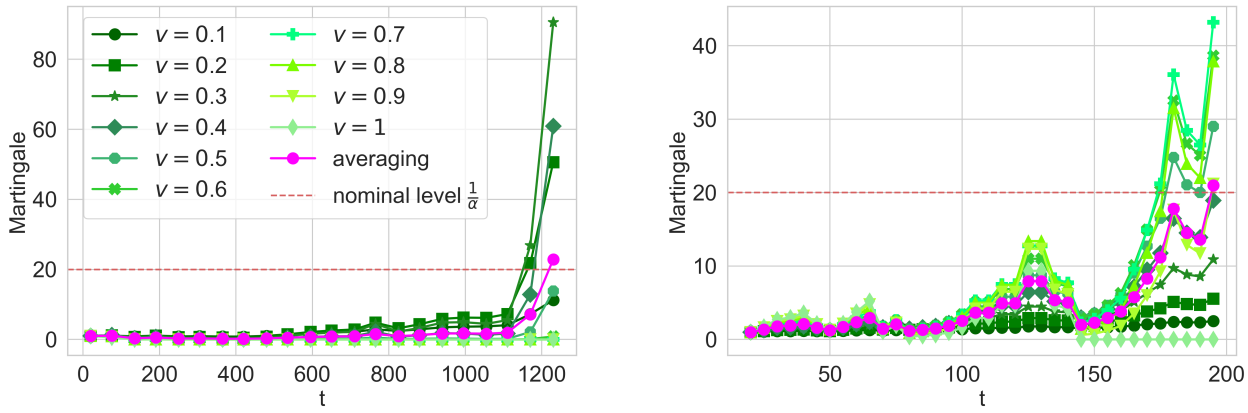


Figure 4: **The effect of the amount of toy money v we risk on the wealth, in comparison with the mixture-method martingale.** Each of the two panels corresponds to a different realization of the `non-null data`. The green curves represent S_t^v in (5) with different choices of v. The magenta curve represents the average test martingale S_t in (6).

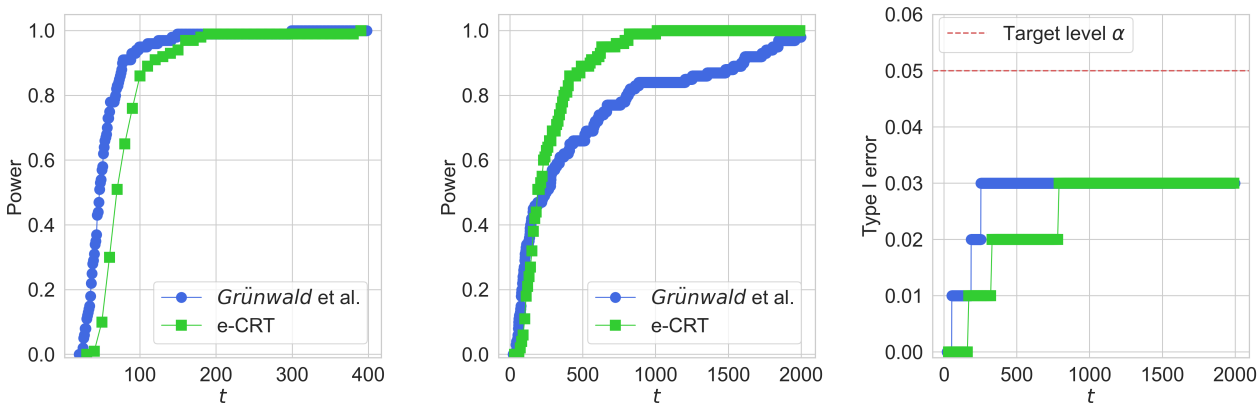
The base test martingales S_t^v are constructed in the same fashion but with (5) in line 8 of Algorithm 2 instead of the mixture approach. Below, we provide the implementation details of Algorithm 2 for this experiment.

- We set the betting score function in (4) to be $g(a, b) = \text{sign}(b - a)$.
- The online learning model for $\hat{f}_t$ takes the form of lasso regression using the hyper-parameter tuning approach described in Section 4.2; we trained $L = 20$ models, each corresponds to a different η, where the number of samples for initial training is set to be $n_{\text{init}} = 20$.
- The test statistic function is the mean squared error of a given batch $T(\{(X_s, Y_s, Z_s)\}_{s=1}^b; \hat{f}) = \frac{1}{b}\sum_{s=1}^b (\hat{f}(X_s) - Y_s)^2$, where we use a batch size of $b = 5$.
- We set the de-randomization parameter K, described in Section 4.2, to be equal to 20.

Although the data distribution is identical in both cases, the wealth processes presented in Figure 4 behave very differently: the left panel portrays that the best-performing constant is $v^* = 0.3$, whereas the right panel indicates that $v^* = 0.7$. Importantly, observe how the martingale $S_t^{v^*}$ grows exponentially with t, and thus has a strong traction force on the average martingale S_t. Observe also how the effect of the growing base martingales on S_t is stronger than that of the ones that do not grow with t. This demonstrates the advantage of the mixture-method S_t (6) over the base martingale with a constant v in S_t^v (5).

D SUPPLEMENTARY DISCUSSION ON RELATED WORK

The contemporary work by Grünwald et al. (2022) also offers an approach to test for CI sequentially based on test martingales. Although Grünwald et al. (2022) test martingale shares similarities with the method proposed in this work, there are several key differences. Grünwald et al. (2022) martingale can be conceptualized as a likelihood ratio process and it involves integration over $P_{X|Z}$. By contrast, ours resembles the knockoffs approach which measures differences of a test statistic evaluated on the original and dummy triplets, and is valid due to the anti-symmetry of the betting score (4). In terms of power, under the assumption of a fast converging estimator for $P_{Y|X,Z}$, Grünwald et al. (2022) prove their test martingale has a growth-rate optimality (Grünwald et al., 2020). On the other hand, following Proposition 1, we merely assume a weaker assumption to achieve power one asymptotically (should not be confused with growth-rate optimality): the model should distinguish the original and dummy triplets *on average*. Given the aforementioned variations, one may opt for the technique put forth by Grünwald et al. (2022) if there is a reasonable estimate for the conditional distribution of $Y \mid X, Z$. On the other hand, our method may be a more suitable choice if greater flexibility is desired for designing the machine learning model.



(a) (b) (c)

Figure 5: **Empirical power and type-I error rate of e-CRT and the method proposed by Grünwald et al. (2022)**, evaluated on 100 realizations of the data. (a) empirical power in a linear case, where $\varphi(X_t, Z_t) = \beta^\top Z_t + 3 \cdot X_t$. (b) empirical power in a non-linear case, where $\varphi(X_t, Z_t) = \beta^\top Z_t + 6 \cdot X_t \cdot |Z_t^{(1)}| \cdot |Z_t^{(2)}|$. (c) type-I error rate with $\varphi(X_t, Z_t) = \beta^\top Z_t$.

To support the above discussion, we compare the two methods based on simulated data. We focus on binary classification since Grünwald et al. (2022) present a concrete test martingale in this context. To this end, we generate X_t, Z_t akin to Section 5.1. The binary response Y_t is generated from a Bernoulli distribution with a probability obtained by applying the sigmoid function to $c \cdot \varphi(X_t, Z_t)$, where we conduct two experiments. In the first, we choose φ to be a linear function and $c = 1$, whereas in the second we choose φ to be a non-linear function and set $c = 0.8$. We deploy the method of Grünwald et al. (2022) as described in (Grünwald et al., 2022)[Section 3.3], and our e-CRT with $K = 20$, $\mathcal{B} = \{2, 5, 10\}$, $g(a, b) = \tanh(20 \cdot (b - a)/\max\{a, b\})$, and $T(\cdot)$ be the binary cross entropy loss. We use a neural network classifier for both testing methods.

We begin with a linear case, demonstrating a scenario where a reasonable estimation of $P_{Y|X,Z}$ can be achieved. In this experiment, we set $\varphi(X_t, Z_t) = \beta^\top Z_t + 3 \cdot X_t$ with $\beta \sim \mathcal{N}(0, I_d)$ and $d = 5$. Figure 5a presents the empirical power of both methods evaluated on 100 realizations of the described data. As can be seen, the method of Grünwald et al. (2022) tends to be more powerful than ours, indicating the advantage of Grünwald et al. (2022) in cases where a good estimation of $P_{Y|X,Z}$ is attainable.

Next, we consider a more complex non-linear interaction model, where $\varphi(X_t, Z_t) = \beta^\top Z_t + 6 \cdot X_t \cdot |Z_t^{(1)}| \cdot |Z_t^{(2)}|$. Here, $\beta \sim \mathcal{N}(0, I_d)$ and $d = 10$, where $Z_t^{(j)}$ represents the jth covariate of Z_t. The empirical power of the methods, evaluated on 100 realizations of the data, is depicted in Figure 5b. As portrayed, our e-CRT performs better in this case, indicating the superiority of our e-CRT when it is harder to attain a fast converging estimator for $P_{Y|X,Z}$.

Lastly, Figure 5c presents the type-I error of both e-CRT and the method proposed by Grünwald et al. (2022), evaluated on 100 realizations of the data with $\varphi(X_t, Z_t) = \beta^\top Z_t$ and $c = 0.8$ where $\beta \sim \mathcal{N}(0, I_d)$ and $d = 10$. There, the type-I error is controlled for both methods.

E ONLINE LEARNING WITH AUTOMATIC HYPER-PARAMETER TUNING

Recall that after we place a bet for a new test point (X_t, Y_t, Z_t), we update the predictive model $\hat{f}_t$ using (X_t, Y_t, Z_t) by leveraging online learning techniques and get $\hat{f}_{t+1}$. The online update is done sequentially and thus computationally efficient. For example, in our experiments we use lasso regression model, minimizing

$$\hat{\beta}_t := \arg\min_{\beta} \frac{1}{t} \sum_{s=1}^{t} (X_s^\top \beta - Y_s)^2 + \eta \|\beta\|_1, \tag{12}$$

where η is a hyper-parameter that controls the regularization strength. The above optimization problem is convex and it is minimized using an iterative solver Wright (2015); Boyd et al. (2011). To form a computationally efficient learning

algorithm, at each step t we initialize the iterative solver with the previous $\hat{\beta}_{t-1}$ and update the regression coefficients by applying a few additional steps with the squared error term in (12) that includes the new observed point. To obtain a powerful predictive model we should tune the hyper-parameter η, but tuning this parameter via standard cross-validation may break the sequential update of $\hat{\beta}_t$. As a way out, we apply the following procedure for tuning η. We train *online* a series of L models $\hat{f}^l_{t_{\text{tr}}}$ on $\{(X_s, Y_s, Z_s)\}_{s<t_{\text{tr}}}$ over a grid of possible values of η_l, $l = 1 \ldots, L$, where $t_{\text{tr}} < t$, and evaluate the models on the $t - t_{\text{tr}}$ recent holdout points. Next, we update the running model $\hat{f}_t$ by minimizing (12) with η_{l^*}, where l^* is the index of the model $\hat{f}^l_{t_{\text{tr}}}$ that achieves the smallest prediction error; this is done by applying a few steps of any iterative solver, initialized with the previous $\hat{\beta}_{t-1}$.

F SUPPLEMENTARY DETAILS ON ENSEMBLE OVER BATCHES

In Section 4.2 we present our approach of ensemble of batch martingales $S_{t,b}$. Each of $S_{t,b}$ is evaluated on a batch of size b instead of a single data point. In more detail,

$$S_{t,b} = \int_0^1 S^v_{t,b} \cdot h(v)dv, \text{ with } S^v_{t,b} := \prod_{s=1}^{\lfloor t/b \rfloor} (1 + v \cdot W_{s,b}), \tag{13}$$

and $\lfloor \cdot \rfloor$ is the floor function. Above, we use the convention that $\prod_{s=1}^{0}(1 + v \cdot W_{s,b}) = 1$. The betting score $W_{s,b}$ in (13) is evaluated similarly to (4) but on a batch via the following test statistic function

$$q_{s,b} = T(\{(X_j, Y_j, Z_j)\}_{j=(s-1)\cdot b+1}^{s\cdot b}; \hat{f}_{(s-1)\cdot b+1}) \in \mathbb{R},$$

that operated on the original batch of triplets; analogously, $\tilde{q}_{s,b}$ is obtained by invoking the same T function, however on the dummy triplets, resulting in $W_{s,b} = g(q_{s,b}, \tilde{q}_{s,b})$. For example, T can be a function returning the mean squared error of $\hat{f}_{(s-1)\cdot b+1}$ evaluated on the observed data. Importantly, the test martingale in (7) is valid since

$$\mathbb{E}_{H_0}\left[S_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{H_0}\left[\frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} S_{t,b} \mid \mathcal{F}_{t-1}\right] = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \mathbb{E}_{H_0}[S_{t,b} \mid \mathcal{F}_{t-1}] = \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} S_{t-1,b} = S_{t-1},$$

where the third equation is because each of the batch martingales $S_{t,b}$ is valid.

G SUPPLEMENTARY DETAILS ON SYNTHETIC EXPERIMENTS

G.1 Implementation Details

In Section 5.1 we describe the experimental setup of our synthetic experiments. Here we provide the implementation details of e-CRT and the baseline methods—CRT and HRT.

We implement the e-CRT procedure as described in Algorithm 2, with the following choices. We set the betting score function (4) to be $g(a, b) = \text{sign}(b - a)$. The martingale in (6) is evaluated by choosing $h(v)$ as the pdf of the uniform distribution on $[0, 1]$. The online learning model for $\hat{f}_t$ takes the form of lasso regression using the hyper-parameter tuning approach described in Section 4.2; we trained $L = 20$ models, each corresponds to a different η, where the number of samples for initial training is set to be $n_{\text{init}} = 20$. The test statistic function is the mean squared error of a given batch, defined as $T(\{(X_s, Y_s, Z_s)\}_{s=1}^b; \hat{f}) = \frac{1}{b}\sum_{s=1}^b (\hat{f}(X_s) - Y_s)^2$, where we apply the batch-ensemble approach with $\mathcal{B} = \{2, 5, 10\}$ in (7). Lastly, we set the de-randomization parameter K, described in Section 4.2, to be equal to 20.

The machine learning method we use in both CRT and HRT is a 5-fold cross-validated lasso regression algorithm. As for the ADDIS-Spending approach Tian and Ramdas (2021) for adjusting CRT and HRT, we use the software package available at `https://github.com/jinjint/onlineFWER`, with the default parameters.

G.2 Type-I Error of the Offline CRT and HRT

In Section 5.2 we present the empirical power of e-CRT compared to CRT, HRT and out-of-the-box sequential versions of them ADDIS-CRT and ADDIS-HRT, evaluated on simulated data. There, we present the type-I error only for the sequential tests: e-CRT, ADDIS-CRT and ADDIS-HRT. Here, we present in Figure 6 the type-I error of CRT and HRT evaluated on

the same data as in Section 5.2. Importantly, the presented type-I error is evaluated by treating the data at each presented time step as a fixed size dataset.

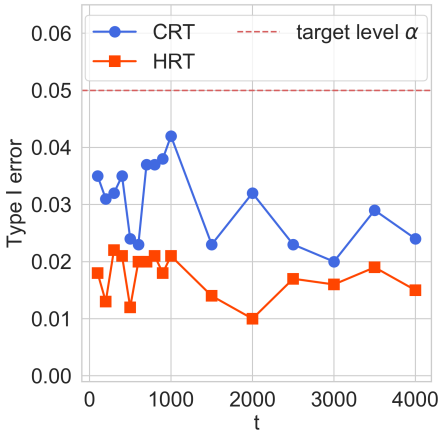
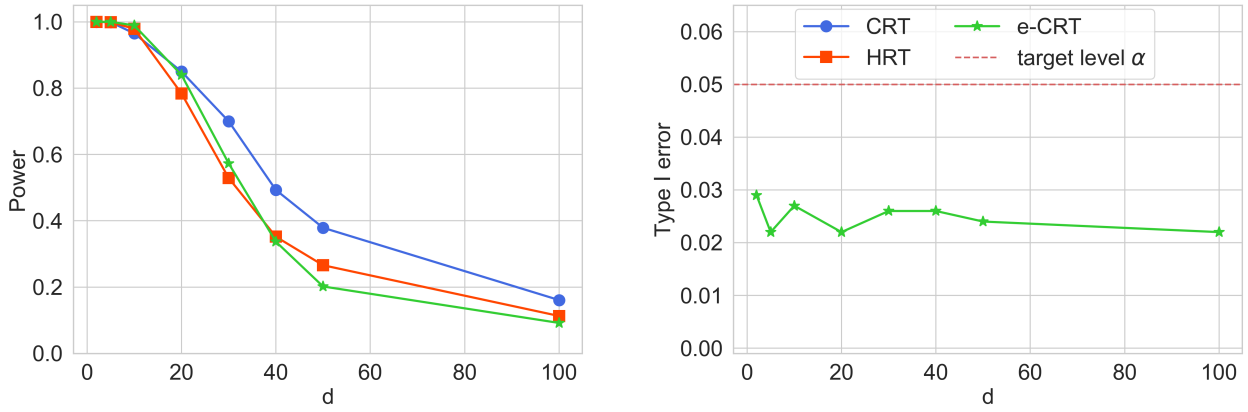


Figure 6: Type-I error of CRT and HRT evaluated over 1000 realizations of the `null data` model. Other details are as in Figure 2.

G.3 Additional Synthetic Experiment with Varying Number of Covariates

In this section we evaluate the performance of e-CRT as a function of the number of covariates d. To do so, we follow the data generation process described in Section 5.1 and sample $n = 1000$ data points of different dimensions d. Then, we apply the e-CRT to each data set and we also apply CRT and HRT on the whole generated data (i.e., only once) to serve as baseline for reference. Figure 7 presents the empirical power and the type-I error as a function of the number of covariates d. It can be seen that the type-I error is controlled for all d, and the empirical power is decreased as we increase the dimension d.



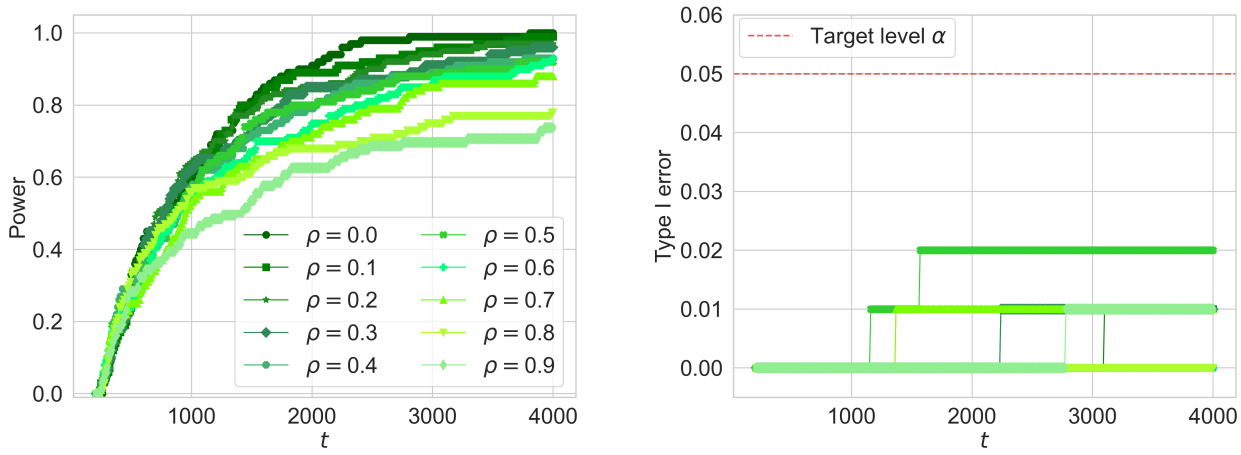
(a) Power

(b) Type-I error

Figure 7: **Empirical power and type-I error rate of e-CRT of level $\alpha = 0.05$ as a function of number of covariates d.** Left: empirical power evaluated on 1000 realizations of the `non-null data` model. Right: type-I error rate evaluated on 1000 realizations of the `null data` model.

G.4 Additional Synthetic Experiment with Increasing Correlation Between the Features

In this section, we study the effect of the dependency structure between X and Z on the performance of the proposed method. To this end, we sample $(X_t, Z_t) \in \mathbb{R}^d$ jointly from $\mathcal{N}(0, \Sigma)$, where X_t is the first covariate of the generated d-dimension vector, and Z_t are the rest $d-1$ covariates. We set the (i, j) entry in the covariance matrix to be $\Sigma_{i,j} = \rho^{|i-j|}$,



(a) Empirical power

(b) Type-I error

Figure 8: **Empirical power and type-I error rate of e-CRT as a function of the auto-correlation parameter ρ.** The empirical power and type-I error evaluated over 100 realizations of the data.

where $\rho \in [0, 1]$ is the auto-correlation parameter. The response Y_t is generated the same way as in Section 5.1. To examine the impact of the dependency strength, we vary the auto-correlation parameter ρ and apply the e-CRT to the generated data as described in Supplementary Section G.1. Figure 8 presents the empirical power and type-I error evaluated over 100 realizations of the data. Following that figure, one can see that the type-I error is controlled, as expected. One can also see that the power decreases as the correlation increases. This result aligns with previous analysis in the field of CI testing; see, for example, Candes et al. (2018); Liu et al. (2022); Shaer and Romano (2023).

G.5 Additional Synthetic Experiment on the Robustness of e-CRT

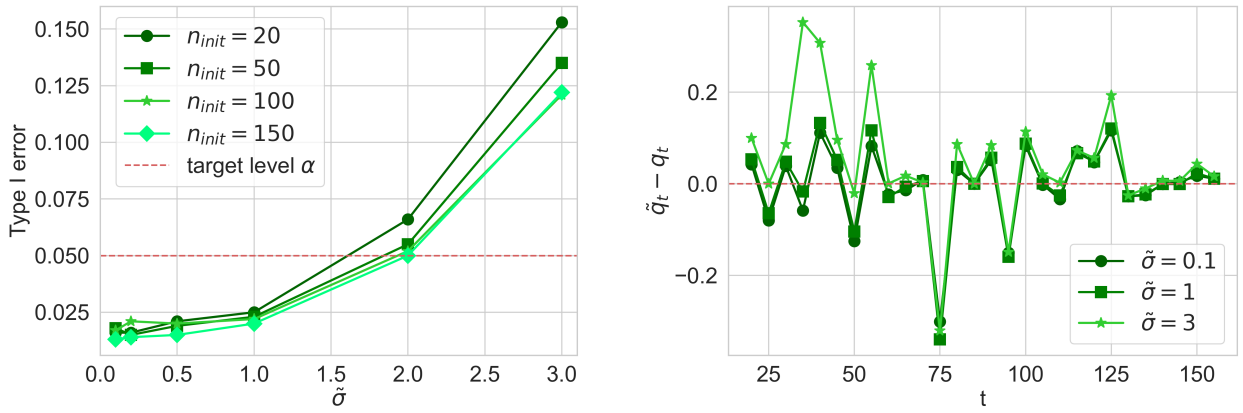
To implement e-CRT, we must generate dummy features $\tilde{X}_t$ from $P_{X|Z}$. In practice, when this conditional distribution is unknown it should be estimated from data. In general, there is no formal type-I error control in this case. Therefore, it is important to study the robustness of e-CRT to errors in the estimation of $P_{X|Z}$.

G.5.1 Parametric Estimation of $P_{X|Z}$

Here, we apply e-CRT to a sequence of data points generated from the `null data` model from Section 5.1, but instead of sampling $\tilde{X}_t$ from the true $P_{X|Z}$, we consider the following data distribution:

$$X_t \mid Z_t \sim \mathcal{N}(u^\top Z_t, \tilde{\sigma}), \quad \text{where} \quad u \sim \mathcal{N}(0, I_d).$$

When setting $\tilde{\sigma} = \sigma = 1$ we recover the true $P_{X|Z}$, and by increasing (resp. decreasing) $\tilde{\sigma}$ we move further away from the true conditional distribution. Figure 9a presents the type-I error rate as a function of $\tilde{\sigma}$, where each curve corresponds to a different number of samples used for initial training n_{init}. The test is applied to $n = 1000$ fresh samples, in addition to the n_{init} ones. Interestingly, observe how the type-I error is conservatively controlled for small values of $\tilde{\sigma} < \sigma = 1$, whereas inflation in the type-I error is reported for larger values of $\tilde{\sigma}$. Observe also how this type-I error inflation is mitigated when using more samples for initial training. To illustrate this behavior from a different angle, we present in Figure 9b the difference $\tilde{q}_t - q_t$ as a function of t for different values of $\tilde{\sigma}$, with the choice of $n_{\text{init}} = 20$. As displayed, the difference $\tilde{q}_t - q_t$ that corresponds to $\tilde{\sigma} = 0.1$ tends to be smaller than the one corresponding to the true $\tilde{\sigma} = \sigma = 1$, which is in line with the tendency of our method to construct conservative martingale when $\tilde{\sigma}$ is small. On the other hand, when $\tilde{\sigma} = 3$ the difference $\tilde{q}_t - q_t$ tends to be larger than that of $\tilde{\sigma} = 1$, and this gap decreases as t increases, where the difference is closer to zero for $t = 150$. This shows that the predictive model we use (lasso) tends to ignore the null feature with the increase of the sample size, and thus the type-I error is moderated for larger n_{init}.



(a) (b)

Figure 9: **Robustness experiments with simulated data.** The dummies $\tilde{X}_t$ are generated from a misspecify model of $P_{X|Z}$, where the farther $\hat{\sigma}$ from $\sigma = 1$ the larger the estimation error of $P_{X|Z}$. (a) Type-I error of e-CRT evaluated over 1000 realizations of the `null data` model, where each curve represents a different number of samples used for initial training of the predictive model. (b) $\tilde{q}_t - q_t$ of a single realization of the `null data` model as a function of t, with $n_{\text{init}} = 20$.

G.5.2 Non-Parametric Estimation of $P_{X|Z}$

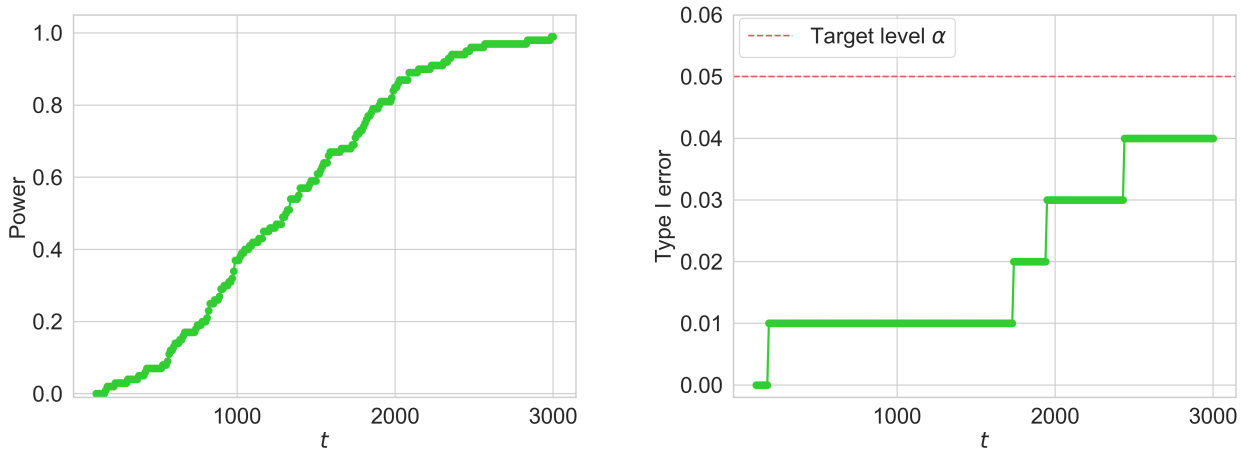
In this section, we consider a more challenging scenario in which $Z_t \sim \mathcal{N}(0, I_{19})$ and $X_t \mid Z_t$ follows a Student-t distribution with 5 degrees of freedom and mean equals to $Z_t^{(1)} Z_t^{(2)}$. Here, $Z_t^{(i)}$ refers to the i'th covariate of Z_t. We generate the response Y_t as described in Section 5.1. To estimate $P_{X|Z}$, we first generate 3000 unlabeled samples (X, Z) and use the non-linear density estimation method proposed by Rosenberg et al. (2022), called NL-VQR. For hyperparameter tuning, we train NL-VQR on 2000 of the unlabeled samples and subsequently evaluated its goodness-of-fit on the remaining 1000 unlabeled samples using the KDE-L1 metric as described by Rosenberg et al. (2022). We then train the NL-VQR with all the 3000 unlabeled samples using the chosen hyperparameters. We deploy the e-CRT on 3000 new, labeled data points, sampled from the same distribution, with a kernel ridge regression model $\hat{f}_t$ with a polynomial kernel of degree 2, whose parameters are tuned by 5-fold cross-validation. The model is fitted on $\{(X_s, Y_s, Z_s)\}_{s=1}^{t-1}$, at each time step t. Figure 10 presents the empirical power and type-I error obtained by the e-CRT. Observe how the type-I error rate grows slowly, but it is controlled even for a relatively large number of samples. Observe also how the power reaches 1 when the sample size is relatively large.

H SUPPLEMENTARY DETAILS ON REAL DATA EXPERIMENTS

H.1 Fund Replication Experiment

H.1.1 Supplementary Implementation Details

Here we provide supplementary details on the implementation of the fund replication experiment, described in Section 6.1. We denote by $X_t^j \in \mathbb{R}$ the tth log return of the jth stock, and by $Z_t^j \in \mathbb{R}^{d-1}$ the vector of the tth log returns of all the stocks except X_t^j. We deploy the e-CRT on the above data stream as in the synthetic experiments from Section 5.1, but with the following adaptations. Since real data sets tend to have outliers, we choose to work with larger batches and a more moderate betting function that takes into account the magnitude of the error, not only its direction as happen in the sign function used in the synthetic experiments. Specifically, we form the betting function as $g(a, b) = \tanh(20 \cdot (b - a)/\max\{a, b\})$, and implement our method with an ensemble over batches of size $\{5, 10, 20\}$. As a strong baseline for reference, we apply the offline CRT and offline HRT on the whole data set, and use lasso regression model with 5-fold cross-validation to tune its hyper-parameter. In contrast to the controlled synthetic experiments from Section 5, here $P_{X^j|Z^j}$ is unknown and thus we must estimate it from the data to generate $\tilde{X}^j$, both for e-CRT and for CRT and HRT. For the offline tests, we approximate it by fitting a multivariate Gaussian on all the samples. For e-CRT, we set $n_{\text{init}} = 500$, fit a multivariate Gaussian on the



(a) Empirical power

(b) Type-I error

Figure 10: **Robustness experiments with simulated data where $X \mid Z$ follows a Student-t distribution.** The dummies $\tilde{X}_t$ are generated from an estimated $P_{X|Z}$, using the density estimation method proposed by Rosenberg et al. (2022). The empirical power and type-I error are evaluated over 100 realizations of the data.

first 500 samples, and use the rest samples for testing.

H.1.2 Supplementary Table

Table 1: Summary of results obtained by CRT, HRT and e-CRT applied to each of the stocks in the data. The stocks belong to the Information Technology sector are highlighted in light green. For CRT and HRT, green (resp. red) value represents a p-value below (resp. above) $\alpha = 0.05$. For e-CRT, we color each value according to the categorization in Vovk and Wang (2021): red for insignificant, orange for 'worth a bare mention', blue and green indicate that the evidence against the null is substantial or strong, respectively.

Symbol	Sector	CRT p-value	HRT p-value	e-CRT $S_{t_{\text{stop}}}$	e-CRT t_{stop}
FTNT	Information Technology	0.035	1	52.8	2220
AAPL	Information Technology	0.005	0.001	52.6	580
CDNS	Information Technology	0.095	1	45.6	700
SNPS	Information Technology	0.005	0.011	40.9	1940
MSFT	Information Technology	0.005	0.001	37.7	560
ORCL	Information Technology	0.005	0.157	33.2	560
V	Information Technology	0.005	0.032	28.1	640
NVDA	Information Technology	0.005	0.001	27.7	840
AMAT	Information Technology	0.03	0.206	26.7	2260
QCOM	Information Technology	0.005	0.001	25.5	1160
HPQ	Information Technology	0.005	0.381	24.1	900
MA	Information Technology	0.005	0.008	23.5	2200
IBM	Information Technology	0.005	0.236	23.1	620
INTC	Information Technology	0.005	0.002	22.8	600
ADBE	Information Technology	0.005	0.006	22.7	1020
TER	Information Technology	0.005	0.036	22.5	2180
ACN	Information Technology	0.005	0.015	22.5	580
INTU	Information Technology	0.005	1	21.7	2180
CTXS	Information Technology	0.507	0.76	21.7	900
TXN	Information Technology	0.005	0.085	20.9	1660
MU	Information Technology	0.005	0.078	20.7	2200
CSCO	Information Technology	0.005	0.093	20.7	580

AVGO	Information Technology	0.005	0.281	20.6	1200
CTSH	Information Technology	0.891	0.495	20.2	680
CRM	Information Technology	0.005	0.008	20.1	900
KLAC	Information Technology	0.085	0.223	8.6	2421
WU	Information Technology	0.96	1	7.1	2421
FIS	Information Technology	0.005	0.011	4.7	2421
AMD	Information Technology	0.015	0.278	4.6	2421
TYL	Information Technology	0.02	1	4.5	2421
IT	Information Technology	0.02	0.154	3	2421
NXPI	Information Technology	0.572	1	2.3	2421
PAYX	Information Technology	0.731	1	1.6	2421
LRCX	Information Technology	0.005	0.165	1.2	2421
ADI	Information Technology	0.602	0.765	1	2421
TRMB	Information Technology	0.716	1	1	2421
APH	Information Technology	0.612	1	1	2421
BR	Information Technology	0.522	1	1	2421
ADP	Information Technology	0.428	1	0.9	2421
ADSK	Information Technology	0.06	1	0.9	2421
ENPH	Information Technology	0.537	1	0.9	2421
FISV	Information Technology	0.204	0.286	0.8	2421
ZBRA	Information Technology	0.199	1	0.8	2421
MPWR	Information Technology	0.164	1	0.8	2421
MSI	Information Technology	0.134	1	0.8	2421
GPN	Information Technology	0.378	0.723	0.6	2421
PTC	Information Technology	0.577	1	0.6	2421
ANSS	Information Technology	0.085	1	0.6	2421
TEL	Information Technology	0.005	0.067	0.6	2421
AKAM	Information Technology	0.522	1	0.5	2421
NTAP	Information Technology	0.383	1	0.5	2421
VRSN	Information Technology	0.204	0.674	0.5	2421
GLW	Information Technology	0.766	1	0.5	2421
FLT	Information Technology	0.507	0.731	0.4	2421
DXC	Information Technology	0.836	1	0.4	2421
STX	Information Technology	0.677	0.593	0.4	2421
MCHP	Information Technology	0.065	0.21	0.4	2421
IPGP	Information Technology	0.517	1	0.3	2421
NOW	Information Technology	0.403	1	0.3	2421
SWKS	Information Technology	0.776	1	0.3	2421
JNPR	Information Technology	0.428	1	0.3	2421
NLOK	Information Technology	0.085	0.395	0.2	2421
WDC	Information Technology	0.209	0.514	0.2	2421
JKHY	Information Technology	0.831	0.368	0.2	2421
FFIV	Information Technology	0.965	0.308	0.1	2421
CRL	Health Care	0.308	0.587	37.7	1780
PKI	Health Care	0.662	0.596	30.7	1660
GOOGL	Communication Services	0.045	0.392	27.8	1120
EL	Consumer Staples	0.005	1	27.4	2160
ALB	Materials	0.015	1	25.6	2300
VZ	Communication Services	0.005	0.386	25.4	740
DVN	Energy	0.03	1	24.7	2220
CMG	Consumer Discretionary	0.005	1	24.5	2140
LYB	Materials	0.194	1	23.3	880
T	Communication Services	0.01	0.797	22.7	700
LUMN	Communication Services	0.005	0.301	22.2	1760
GOOG	Communication Services	0.055	0.56	22.1	960

ROP	Industrials	0.005	1	22	2220
NDAQ	Financials	0.005	1	16.7	2421
NFLX	Communication Services	0.06	0.366	6	2421
LEN	Consumer Discretionary	0.96	0.165	5.8	2421
HUM	Health Care	0.169	1	4.6	2421
IPG	Communication Services	0.025	1	4.1	2421
GNRC	Industrials	0.104	1	3.9	2421
XYL	Industrials	0.498	0.563	3.9	2421
AOS	Industrials	0.124	0.956	3.5	2421
ATVI	Communication Services	0.677	0.624	3.4	2421
VLO	Energy	0.602	0.631	3.2	2421
EA	Communication Services	0.02	0.354	3.1	2421
PWR	Industrials	0.09	1	2.7	2421
FCX	Materials	0.015	0.189	2.7	2421
EBAY	Consumer Discretionary	0.01	0.432	2.6	2421
WAT	Health Care	0.781	1	2.2	2421
DVA	Health Care	0.005	1	2.2	2421
TROW	Financials	0.771	0.326	2.1	2421
WHR	Consumer Discretionary	0.761	1	1.9	2421
DISH	Communication Services	0.149	1	1.9	2421
CE	Materials	0.045	1	1.9	2421
PEAK	Real Estate	0.522	1	1.8	2421
WMB	Energy	0.04	1	1.8	2421
EXPD	Industrials	0.1	1	1.8	2421
REGN	Health Care	0.806	1	1.7	2421
MTCH	Communication Services	0.02	0.221	1.7	2421
EFX	Industrials	0.065	1	1.6	2421
ABBV	Health Care	0.328	1	1.5	2421
KEY	Financials	0.468	1	1.5	2421
INCY	Health Care	0.119	1	1.5	2421
FAST	Industrials	0.632	1	1.5	2421
DHI	Consumer Discretionary	0.886	1	1.5	2421
SWK	Industrials	0.164	1	1.4	2421
UAL	Industrials	0.587	0.473	1.4	2421
ZION	Financials	0.811	1	1.4	2421
BKR	Energy	0.303	1	1.4	2421
TJX	Consumer Discretionary	0.055	1	1.3	2421
RHI	Industrials	0.438	1	1.3	2421
DIS	Communication Services	0.134	1	1.3	2421
CMCSA	Communication Services	0.458	1	1.2	2421
NEM	Materials	0.194	1	1.2	2421
DRE	Real Estate	0.731	1	1.2	2421
BKNG	Consumer Discretionary	0.602	1	1.1	2421
LVS	Consumer Discretionary	0.741	1	1.1	2421
PHM	Consumer Discretionary	0.403	1	1.1	2421
PFG	Financials	0.443	1	1.1	2421
CME	Financials	0.975	1	1.1	2421
COP	Energy	0.786	1	1.1	2421
EIX	Utilities	0.313	1	1.1	2421
BAX	Health Care	0.323	1	1.1	2421
APTV	Consumer Discretionary	0.01	1	1.1	2421
TECH	Health Care	0.05	1	1.1	2421
MNST	Consumer Staples	0.313	1	1.1	2421
KMX	Consumer Discretionary	0.244	0.418	1.1	2421
PPG	Materials	0.617	1	1.1	2421

LOW	Consumer Discretionary	0.905	1	1.1	2421
NWL	Consumer Discretionary	0.468	1	1.1	2421
AMZN	Consumer Discretionary	0.015	0.18	1.1	2421
VTRS	Health Care	0.682	1	1.1	2421
AMP	Financials	0.572	1	1.1	2421
JNJ	Health Care	0.945	1	1.1	2421
LUV	Industrials	0.736	1	1	2421
LNT	Utilities	0.458	1	1	2421
LMT	Industrials	0.667	1	1	2421
COST	Consumer Staples	0.786	1	1	2421
LLY	Health Care	0.811	1	1	2421
EXR	Real Estate	0.229	1	1	2421
COO	Health Care	0.478	1	1	2421
CMS	Utilities	0.378	1	1	2421
MDT	Health Care	0.557	1	1	2421
EMR	Industrials	0.627	1	1	2421
MMC	Financials	0.468	1	1	2421
MMM	Industrials	0.711	1	1	2421
MO	Consumer Staples	0.677	1	1	2421
CLX	Consumer Staples	0.512	1	1	2421
MRK	Health Care	0.507	1	1	2421
CL	Consumer Staples	0.537	1	1	2421
MSCI	Financials	0.294	1	1	2421
MTB	Financials	0.02	1	1	2421
MTD	Health Care	0.279	1	1	2421
CI	Health Care	0.408	1	1	2421
ECL	Materials	0.453	1	1	2421
CMA	Financials	0.368	1	1	2421
KMB	Consumer Staples	0.771	1	1	2421
LIN	Materials	0.015	1	1	2421
DHR	Health Care	0.846	1	1	2421
HBI	Consumer Discretionary	0.766	1	1	2421
ETR	Utilities	0.726	1	1	2421
HBAN	Financials	0.179	1	1	2421
HAS	Consumer Discretionary	0.692	1	1	2421
DG	Consumer Discretionary	0.184	1	1	2421
GWW	Industrials	0.557	1	1	2421
DGX	Health Care	0.612	1	1	2421
EVRG	Utilities	0.388	1	1	2421
GPC	Consumer Discretionary	0.348	1	1	2421
EOG	Energy	0.866	1	1	2421
EW	Health Care	0.746	1	1	2421
GIS	Consumer Staples	0.428	1	1	2421
DLR	Real Estate	0.572	1	1	2421
GD	Industrials	0.687	1	1	2421
FRC	Financials	0.706	1	1	2421
FMC	Materials	0.761	1	1	2421
FITB	Financials	0.289	1	1	2421
EXC	Utilities	0.493	1	1	2421
HD	Consumer Discretionary	0.756	1	1	2421
HRL	Consumer Staples	0.552	1	1	2421
HST	Real Estate	0.746	1	1	2421
ESS	Real Estate	0.612	1	1	2421
LH	Health Care	0.209	1	1	2421
DUK	Utilities	0.846	1	1	2421

LEG	Consumer Discretionary	0.06	1	1	2421
L	Financials	0.587	1	1	2421
CVS	Health Care	0.144	1	1	2421
KMI	Energy	0.164	1	1	2421
F	Consumer Discretionary	0.129	1	1	2421
KIM	Real Estate	0.582	1	1	2421
K	Consumer Staples	0.756	1	1	2421
JCI	Industrials	0.687	1	1	2421
ITW	Industrials	0.602	1	1	2421
D	Utilities	0.308	1	1	2421
DD	Materials	0.473	1	1	2421
EQR	Real Estate	0.896	1	1	2421
ILMN	Health Care	0.279	1	1	2421
IDXX	Health Care	0.597	1	1	2421
ES	Utilities	0.736	1	1	2421
IEX	Industrials	0.771	1	1	2421
JPM	Financials	0.736	0.745	1	2421
SNA	Industrials	0.781	1	1	2421
AMT	Real Estate	0.791	1	1	2421
AON	Financials	0.512	1	1	2421
TGT	Consumer Discretionary	0.637	1	1	2421
TFX	Health Care	0.438	1	1	2421
TFC	Financials	0.836	1	1	2421
ARE	Real Estate	0.07	1	1	2421
SYK	Health Care	0.567	1	1	2421
STT	Financials	0.562	1	1	2421
AVB	Real Estate	0.617	1	1	2421
SO	Utilities	0.338	1	1	2421
SLB	Energy	0.761	1	1	2421
SJM	Consumer Staples	0.826	1	1	2421
SIVB	Financials	0.453	1	1	2421
SCHW	Financials	0.9	1	1	2421
SBUX	Consumer Discretionary	0.95	1	1	2421
RTX	Industrials	0.393	1	1	2421
RSG	Industrials	0.294	1	1	2421
AWK	Utilities	0.771	1	1	2421
ROL	Industrials	0.542	1	1	2421
AXP	Financials	0.542	1	1	2421
TSCO	Consumer Discretionary	0.333	1	1	2421
ALK	Industrials	0.677	1	1	2421
RF	Financials	0.572	1	1	2421
UNH	Health Care	0.801	1	1	2421
AAP	Consumer Discretionary	0.721	1	1	2421
YUM	Consumer Discretionary	0.905	1	1	2421
XRAY	Health Care	0.557	1	1	2421
XEL	Utilities	0.249	1	1	2421
ABC	Health Care	0.627	1	1	2421
ABT	Health Care	0.522	1	1	2421
ADM	Consumer Staples	0.473	1	1	2421
AEE	Utilities	0.557	1	1	2421
AEP	Utilities	0.662	1	1	2421
WFC	Financials	0.572	1	1	2421
WELL	Real Estate	0.582	1	1	2421
AFL	Financials	0.388	0.358	1	2421
AJG	Financials	0.731	1	1	2421

VRSK	Industrials	0.9	1	1	2421
VNO	Real Estate	0.637	1	1	2421
VMC	Materials	0.269	1	1	2421
VFC	Consumer Discretionary	0.169	1	1	2421
USB	Financials	0.164	1	1	2421
UNP	Industrials	0.308	1	1	2421
BAC	Financials	0.532	1	1	2421
NI	Utilities	0.607	1	1	2421
PSA	Real Estate	0.06	1	1	2421
O	Real Estate	0.612	1	1	2421
PNW	Utilities	0.045	1	1	2421
PNR	Industrials	0.876	1	1	2421
CAH	Health Care	0.04	1	1	2421
BK	Financials	0.348	1	1	2421
PPL	Utilities	0.428	1	1	2421
OMC	Communication Services	0.786	1	1	2421
PKG	Materials	0.667	1	1	2421
BRO	Financials	0.627	1	1	2421
PH	Industrials	0.532	1	1	2421
PM	Consumer Staples	0.552	1	1	2421
PLD	Real Estate	0.866	1	1	2421
BLK	Financials	0.458	1	1	2421
PEG	Utilities	0.353	1	1	2421
PVH	Consumer Discretionary	0.239	1	1	2421
PEP	Consumer Staples	0.109	1	1	2421
PG	Consumer Staples	0.94	1	1	2421
PFE	Health Care	0.687	1	1	2421
NTRS	Financials	0.403	1	1	2421
NSC	Industrials	0.557	1	1	2421
PNC	Financials	0.224	1	1	2421
HCA	Health Care	0.741	1	0.9	2421
TDG	Industrials	0.821	1	0.9	2421
MCK	Health Care	0.323	1	0.9	2421
BWA	Consumer Discretionary	0.846	1	0.9	2421
COF	Financials	0.04	0.959	0.9	2421
AIG	Financials	0.537	0.291	0.9	2421
CNC	Health Care	0.463	1	0.9	2421
DE	Industrials	0.134	1	0.9	2421
TTWO	Communication Services	0.443	1	0.9	2421
UDR	Real Estate	0.02	1	0.9	2421
DLTR	Consumer Discretionary	0.756	1	0.9	2421
UPS	Industrials	0.393	1	0.9	2421
HIG	Financials	0.731	1	0.9	2421
STZ	Consumer Staples	0.841	0.719	0.9	2421
HOLX	Health Care	0.9	1	0.9	2421
BIIB	Health Care	0.468	1	0.9	2421
WYNN	Consumer Discretionary	0.572	1	0.9	2421
BDX	Health Care	0.935	1	0.9	2421
LKQ	Consumer Discretionary	0.756	1	0.9	2421
PSX	Energy	0.234	1	0.9	2421
AVY	Materials	0.388	1	0.9	2421
JBHT	Industrials	0.652	1	0.9	2421
CVX	Energy	0.547	1	0.9	2421
SHW	Materials	0.617	1	0.9	2421
FDX	Industrials	0.697	1	0.9	2421

HSY	Consumer Staples	0.547	0.859	0.9	2421
HSIC	Health Care	0.04	1	0.9	2421
NEE	Utilities	0.025	1	0.9	2421
IFF	Materials	0.592	1	0.9	2421
HON	Industrials	0.463	1	0.9	2421
IP	Materials	0.502	1	0.9	2421
FBHS	Industrials	0.383	1	0.8	2421
CAT	Industrials	0.726	1	0.8	2421
MKTX	Financials	0.433	1	0.8	2421
FANG	Energy	0.642	1	0.8	2421
VTR	Real Estate	0.94	1	0.8	2421
BXP	Real Estate	0.463	1	0.8	2421
MPC	Energy	0.498	1	0.8	2421
NUE	Materials	0.189	1	0.8	2421
NVR	Consumer Discretionary	0.338	1	0.8	2421
WEC	Utilities	0.93	1	0.8	2421
MS	Financials	0.816	1	0.8	2421
REG	Real Estate	0.582	1	0.8	2421
FE	Utilities	0.065	0.287	0.8	2421
RE	Financials	0.856	1	0.8	2421
CPRT	Industrials	0.194	1	0.8	2421
SEE	Materials	0.025	1	0.8	2421
J	Industrials	0.736	1	0.8	2421
SPG	Real Estate	0.493	1	0.8	2421
ICE	Financials	0.935	0.348	0.8	2421
STE	Health Care	0.025	1	0.8	2421
MHK	Consumer Discretionary	0.09	1	0.8	2421
CTAS	Industrials	0.01	0.054	0.8	2421
HES	Energy	0.672	1	0.8	2421
LDOS	Industrials	0.607	1	0.8	2421
TMUS	Communication Services	0.612	1	0.8	2421
TRV	Financials	0.03	1	0.8	2421
TSLA	Consumer Discretionary	0.005	1	0.8	2421
AMGN	Health Care	0.537	1	0.8	2421
DOV	Industrials	0.522	1	0.8	2421
ROK	Industrials	0.647	1	0.8	2421
BA	Industrials	0.632	1	0.7	2421
CBRE	Real Estate	0.095	1	0.7	2421
CPB	Consumer Staples	0.234	1	0.7	2421
DTE	Utilities	0.915	1	0.7	2421
APD	Materials	0.682	1	0.7	2421
ALL	Financials	0.328	1	0.7	2421
CCL	Consumer Discretionary	0.493	1	0.7	2421
ZTS	Health Care	0.448	1	0.7	2421
A	Health Care	0.652	1	0.7	2421
HII	Industrials	0.881	0.397	0.7	2421
ED	Utilities	0.363	1	0.7	2421
OKE	Energy	0.592	1	0.7	2421
ORLY	Consumer Discretionary	0.562	1	0.7	2421
PGR	Financials	0.562	1	0.7	2421
MLM	Materials	0.249	1	0.7	2421
MET	Financials	0.214	1	0.7	2421
RCL	Consumer Discretionary	0.617	1	0.7	2421
SBAC	Real Estate	0.378	1	0.7	2421
SRE	Utilities	0.284	1	0.7	2421

LNC	Financials	0.667	1	0.7	2421
NRG	Utilities	0.199	0.448	0.7	2421
TMO	Health Care	0.045	0.28	0.7	2421
LHX	Industrials	0.756	1	0.7	2421
NOC	Industrials	0.348	1	0.7	2421
GPS	Consumer Discretionary	0.358	1	0.7	2421
WBA	Consumer Staples	0.234	1	0.7	2421
WM	Industrials	0.94	1	0.7	2421
WMT	Consumer Staples	0.915	0.349	0.7	2421
GE	Industrials	0.244	1	0.7	2421
WST	Health Care	0.493	1	0.7	2421
CHRW	Industrials	0.169	1	0.6	2421
BMY	Health Care	0.886	1	0.6	2421
HAL	Energy	0.254	1	0.6	2421
MRO	Energy	0.781	1	0.6	2421
EQIX	Real Estate	0.597	0.687	0.6	2421
POOL	Consumer Discretionary	0.761	1	0.6	2421
C	Financials	0.781	0.058	0.6	2421
CCI	Real Estate	0.9	1	0.6	2421
MDLZ	Consumer Staples	0.353	1	0.6	2421
PXD	Energy	0.542	1	0.6	2421
NCLH	Consumer Discretionary	0.662	1	0.6	2421
TXT	Industrials	0.697	1	0.6	2421
ROST	Consumer Discretionary	0.537	1	0.6	2421
IVZ	Financials	0.821	1	0.6	2421
TT	Industrials	0.139	1	0.6	2421
IRM	Real Estate	0.413	1	0.6	2421
ISRG	Health Care	0.139	0.394	0.6	2421
AES	Utilities	0.706	1	0.6	2421
WAB	Industrials	0.249	1	0.6	2421
MAA	Real Estate	0.592	1	0.6	2421
VRTX	Health Care	0.801	0.63	0.5	2421
ATO	Utilities	0.891	1	0.5	2421
WRB	Financials	0.776	1	0.5	2421
SYY	Consumer Staples	0.104	0.483	0.5	2421
SPGI	Financials	0.279	1	0.5	2421
TAP	Consumer Staples	0.035	1	0.5	2421
PRU	Financials	0.353	1	0.5	2421
TPR	Consumer Discretionary	0.826	0.538	0.5	2421
BEN	Financials	0.02	1	0.5	2421
AME	Industrials	0.408	1	0.5	2421
WY	Real Estate	0.174	1	0.5	2421
PCAR	Industrials	0.597	1	0.5	2421
CAG	Consumer Staples	0.383	1	0.5	2421
CNP	Utilities	0.065	1	0.5	2421
CINF	Financials	0.403	1	0.5	2421
GS	Financials	0.806	1	0.5	2421
DAL	Industrials	0.542	1	0.5	2421
GM	Consumer Discretionary	0.284	1	0.5	2421
CSX	Industrials	0.771	1	0.5	2421
GL	Financials	0.284	1	0.5	2421
MAR	Consumer Discretionary	0.174	1	0.5	2421
EXPE	Consumer Discretionary	0.711	0.369	0.5	2421
DRI	Consumer Discretionary	0.438	1	0.5	2421
GRMN	Consumer Discretionary	0.483	1	0.5	2421

CHTR	Communication Services	0.418	1	0.5	2421
ABMD	Health Care	0.637	1	0.4	2421
GILD	Health Care	0.493	1	0.4	2421
UHS	Health Care	0.01	0.272	0.4	2421
ODFL	Industrials	0.07	1	0.4	2421
DFS	Financials	0.005	1	0.4	2421
NLSN	Industrials	0.02	1	0.4	2421
TDY	Industrials	0.682	1	0.4	2421
CF	Materials	0.119	1	0.4	2421
CHD	Consumer Staples	0.597	1	0.4	2421
MCD	Consumer Discretionary	0.836	0.671	0.4	2421
RMD	Health Care	0.657	0.687	0.4	2421
MAS	Industrials	0.905	1	0.4	2421
RL	Consumer Discretionary	0.244	1	0.4	2421
EMN	Materials	0.045	1	0.4	2421
MKC	Consumer Staples	0.353	0.687	0.4	2421
BIO	Health Care	0.905	1	0.4	2421
MGM	Consumer Discretionary	0.637	1	0.4	2421
BBY	Consumer Discretionary	0.383	1	0.4	2421
AMCR	Materials	0.363	1	0.4	2421
CB	Financials	0.771	1	0.4	2421
AIZ	Financials	0.736	1	0.3	2421
AAL	Industrials	0.771	1	0.3	2421
FRT	Real Estate	0.01	0.198	0.3	2421
URI	Industrials	0.716	0.805	0.3	2421
ALGN	Health Care	0.592	1	0.3	2421
DPZ	Consumer Discretionary	0.856	1	0.3	2421
ULTA	Consumer Discretionary	0.259	0.353	0.3	2421
ZBH	Health Care	0.388	1	0.3	2421
ETN	Industrials	0.537	1	0.3	2421
DXCM	Health Care	0.214	0.327	0.3	2421
UAA	Consumer Discretionary	0.791	1	0.3	2421
KO	Consumer Staples	0.02	1	0.3	2421
NKE	Consumer Discretionary	0.204	0.586	0.3	2421
BSX	Health Care	0.517	0.834	0.3	2421
CBOE	Financials	0.015	1	0.3	2421
LYV	Communication Services	0.234	0.495	0.3	2421
AZO	Consumer Discretionary	0.02	0.245	0.3	2421
PENN	Consumer Discretionary	0.02	1	0.3	2421
CMI	Industrials	0.015	0.217	0.3	2421
MOS	Materials	0.537	1	0.2	2421
XOM	Energy	0.189	1	0.2	2421
OXY	Energy	0.632	1	0.2	2421
BBWI	Consumer Discretionary	0.617	0.605	0.2	2421
TSN	Consumer Staples	0.423	1	0.2	2421
RJF	Financials	0.592	1	0.2	2421
CTRA	Energy	0.383	1	0.2	2421
APA	Energy	0.672	1	0.2	2421
KR	Consumer Staples	0.562	0.719	0.2	2421
MCO	Financials	0.905	0.343	0.1	2421

H.2 Supplementary Details on the HIV Drug Resistance Experiment

H.2.1 Data and Implementation Details

In Section 6.2 we present an experiment of detection mutations in HIV that are associated with drug resistance. The data set[2] we consider there has not been collected sequentially, and thus it is not ideal to present the strength of our sequential testing procedure. Yet, we choose this task because of its importance, and since it has been studied in depth in the knockoff literature Barber and Candès (2015); Lu et al. (2018); Romano et al. (2020); Shaer and Romano (2023). In particular, this data set is convenient to analyze as the effect of each mutation on drug resistance—reported by previous scientific works—is summarized in `https://hivdb.stanford.edu/dr-summary/comments/PI/`.

We denote by (X_t^j, Z_t^j, Y_t) the tth sample, where $X_t^j \in \{0,1\}$ indicates the presence or absence of the jth mutation, and $Z_t^j \in \mathbb{R}^{d-1}$ is a vector that contains all the remaining measured mutations in locations $1, 2, \ldots, j-1, j+1, \ldots, d$. The response Y_t represents the log-fold increase in drug resistance. We deploy the e-CRT the same way as described in Section 6.1, with an additional adaptation; we fit a 5-fold cross-validated lasso model $\hat{f}_t$ on $\{(X_s, Y_s, Z_s)\}_{s=1}^{t-1}$ at each time step t, in contrast to the 1-fold cross-validation approach we used in the previous experiments; we use the latter to reduce computational cost, illustrating how to combine e-CRT with online learning algorithms. Naturally, the 5-fold cross validation approach leads to a better choice of lasso's hyper-parameter, and thus obtaining more accurate predictive models. Next, we approximate $P_{X^j|Z^j}$ as follows. Since X^j is binary, we sample $\tilde{X}_t^j$ from a Bernoulli distribution with probability of success $\hat{\pi}^j(Z_t^j)$, where $\hat{\pi}^j(Z_t^j)$ estimates $P_{X^j|Z^j}(X^j = 1 \mid Z^j)$. We formulate this estimator by fitting a logistic regression model on the unlabeled data $\{(X_t^j, Z_t^j)\}_{t=1}^n$, with an l_2 regularization whose penalty parameter is tuned via 10-fold cross-validation.

H.2.2 Supplementary Results

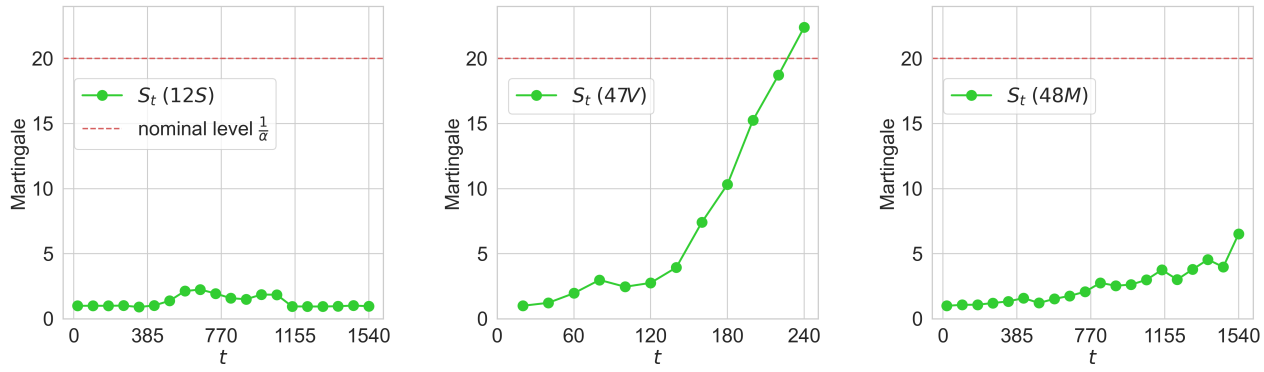
Table 2: Summary of the output of CRT, HRT and e-CRT applied to each of the HIV mutations in the data. The type of each mutation (Major, Minor, Accessory, Other, Unknown) represents the effect of the feature on drug resistance as reported by previous studies. The other details are as in Table 1.

Feature Name	Mutation Type	CRT p-value	HRT p-value	e-CRT $S_{t_{\text{stop}}}$	e-CRT t_{stop}
10F	Accessory	0.005	0.001	25.5	680
10I	Other	0.005	0.001	31.8	520
10V	Other	0.01	0.001	25.5	1280
11I	Other	0.005	0.009	5.5	1555
11L	Other	0.98	1	1	1555
12A	Unknown	0.94	1	1	1555
12I	Unknown	0.716	1	1	1555
12K	Unknown	0.806	1	0.9	1555
12N	Unknown	0.711	1	1	1555
12P	Unknown	0.458	1	1	1555
12S	Unknown	0.204	0.556	1	1555
13V	Unknown	0.005	0.001	23.8	1540
14R	Unknown	0.03	0.042	1.2	1555
15V	Unknown	0.925	0.929	0.2	1555
16A	Unknown	0.005	0.001	20.7	1020
16E	Unknown	0.647	1	0.8	1555
18H	Unknown	0.925	1	0.5	1555
19I	Unknown	0.438	1	0.4	1555
19P	Unknown	0.637	1	1	1555
19Q	Unknown	0.826	1	1	1555
19T	Unknown	0.816	1	1	1555
19V	Unknown	0.776	1	0.7	1555
20I	Other	0.01	0.033	1	1555

[2]Data is available online at `https://hivdb.stanford.edu/pages/published_analysis/genophenoPNAS2006`

20M	Other	1	1	1	1555
20R	Other	0.005	0.004	26	1240
20T	Accessory	0.005	0.002	40.9	1160
20V	Other	0.254	1	0.8	1555
22V	Unknown	0.025	0.116	5.7	1555
23I	Accessory	1	0.357	0.5	1555
24F	Accessory	0.005	0.001	20.9	1000
24I	Accessory	0.005	0.002	24.9	940
30N	Major	0.03	0.148	3.2	1555
32I	Major	0.01	0.007	29	880
33F	Accessory	0.005	0.001	27.3	600
33I	Other	0.045	0.096	1.9	1555
33V	Other	0.622	1	1	1555
34D	Other	0.09	1	1	1555
34Q	Other	0.03	0.002	27.3	1540
35D	Other	0.388	0.282	0.4	1555
35G	Other	0.826	0.887	1.2	1555
35N	Other	0.945	1	0.9	1555
35Q	Other	1	1	1	1555
36I	Other	0.925	0.777	0.4	1555
36L	Other	0.08	0.125	1.2	1555
36V	Other	0.667	0.423	2.7	1555
37C	Other	0.662	1	1	1555
37D	Other	0.015	0.1	3.2	1555
37E	Other	0.398	1	0.8	1555
37H	Other	0.672	1	1	1555
37S	Other	0.672	0.481	0.4	1555
37T	Other	0.776	1	0.5	1555
37X	Other	0.975	0.248	1	1555
39Q	Other	0.985	1	0.8	1555
39S	Other	0.965	0.392	1	1555
41K	Other	0.368	1	0.4	1555
43T	Accessory	0.005	0.001	21	880
45R	Unknown	0.209	0.223	0.8	1555
46I	Major	0.005	0.001	23.3	520
46L	Major	0.005	0.001	21	980
46V	Accessory	0.075	1	0.8	1555
47A	Major	0.005	0.028	8	1555
47V	Major	0.005	0.001	22.4	240
48M	Major	0.005	0.067	6.5	1555
48V	Major	0.005	0.003	23.6	660
50L	Major	0.005	0.001	20	700
50V	Major	0.005	0.001	21.3	440
53L	Accessory	0.025	1	37	1320
54A	Major	0.005	0.001	29.5	760
54L	Major	0.02	0.071	20.6	1160
54M	Major	0.005	0.003	21.5	960
54S	Major	0.095	0.363	1.8	1555
54T	Major	1	0.998	1	1555
54V	Major	0.005	0.001	33.1	180
55R	Unknown	1	1	0.6	1555
57G	Unknown	0.478	1	0.3	1555
57K	Unknown	0.1	0.07	0.9	1555
58E	Accessory	0.02	0.002	1.9	1555
60E	Unknown	0.577	1	0.5	1555

61E	Unknown	0.025	0.352	0.9	1555
61H	Unknown	0.756	0.881	0.5	1555
61N	Unknown	0.637	0.642	1.1	1555
62V	Unknown	0.129	0.057	1.1	1555
63A	Unknown	0.174	1	0.8	1555
63C	Unknown	0.925	1	1	1555
63H	Unknown	0.915	1	1	1555
63P	Unknown	0.005	0.072	28.4	160
63Q	Unknown	0.642	1	0.8	1555
63S	Unknown	0.701	1	1.3	1555
63T	Unknown	0.328	1	0.4	1555
63V	Unknown	0.826	1	1	1555
63X	Unknown	0.766	1	1	1555
64L	Unknown	0.005	0.011	28.6	1300
64M	Unknown	0.214	0.706	0.4	1555
64V	Unknown	0.95	0.442	0.5	1555
65D	Unknown	0.726	0.642	0.6	1555
66F	Unknown	0.01	0.144	3.1	1555
66V	Unknown	0.652	1	0.9	1555
67E	Unknown	0.91	1	0.9	1555
67F	Unknown	0.159	0.557	1.1	1555
69Q	Unknown	0.811	1	0.7	1555
69R	Unknown	0.279	1	0.8	1555
69Y	Unknown	1	1	0.8	1555
70E	Unknown	0.736	1	1	1555
70R	Unknown	0.085	0.061	1.4	1555
70T	Unknown	0.846	1	1	1555
71I	Other	0.488	0.546	1.6	1555
71L	Other	0.114	1	1	1555
71T	Other	0.03	0.097	0.8	1555
71V	Other	0.03	0.356	7.3	1555
72E	Unknown	0.716	1	0.6	1555
72M	Unknown	0.055	1	5.2	1555
72R	Unknown	0.756	1	0.8	1555
72T	Unknown	0.07	0.034	1	1555
72V	Unknown	0.612	0.699	0.6	1555
73A	Accessory	0.721	0.721	0.6	1555
73C	Accessory	0.005	0.004	20.3	1080
73S	Accessory	0.005	0.001	22.4	1260
73T	Accessory	0.01	1	3.4	1555
74A	Unknown	0.756	1	1	1555
74S	Other	0.284	0.404	2.5	1555
76V	Major	0.005	0.001	33.1	400
77I	Unknown	0.005	0.003	6.3	1555
79A	Unknown	0.806	0.341	1	1555
79S	Unknown	0.726	1	0.9	1555
82A	Major	0.005	0.001	21.1	480
82C	Major	0.005	0.001	9.2	1555
82F	Major	0.005	0.012	21.4	800
82I	Other	1	0.941	0.7	1555
82L	Major	0.836	1	0.6	1555
82M	Major	0.94	1	0.5	1555
82S	Major	0.229	0.544	1.3	1555
82T	Major	0.005	0.003	25.4	880
83D	Accessory	0.075	1	0.7	1555



(a) (b) (c)

Figure 11: **Real HIV data experiment.** The test martingale S_t as a function of t, evaluated on three representative mutations of HIV. (a) Mutation '12S', which has not been reported in previous studies to have an effect on drug resistance. (b) Mutation '47V', which has been reported to have a major effect. (c) Mutation '48M', which has been reported to have a major effect.

84A	Major	0.169	0.328	1.9	1555
84C	Major	0.746	0.683	0.9	1555
84V	Major	0.005	0.001	35.4	220
85V	Other	0.866	0.874	0.4	1555
88D	Accessory	0.005	0.059	20.1	1300
88G	Major	0.657	1	1	1555
88S	Major	1	0.172	0.7	1555
88T	Major	0.975	1	1	1555
89I	Unknown	0.005	0.044	1.4	1555
89M	Unknown	0.328	0.669	0.3	1555
89V	Accessory	0.01	0.026	20.6	600
90M	Major	0.005	0.001	35.7	780
91S	Unknown	0.189	1	1.1	1555
92K	Unknown	0.289	0.555	0.6	1555
92R	Unknown	0.736	0.62	0.7	1555
93L	Unknown	0.005	0.086	3.2	1555
95F	Unknown	0.408	0.514	0.9	1555