On Continuous Monitoring of Risk Violations under Unknown Shift

Alexander Timans[*,1] **Rajeev Verma**[*,1] **Eric Nalisnick**[2] **Christian A. Naesseth**[1]

[1]UvA-Bosch Delta Lab, University of Amsterdam
[2]Department of Computer Science, Johns Hopkins University

Abstract

Machine learning systems deployed in the real world must operate under dynamic and often unpredictable distribution shifts. This challenges the validity of statistical safety assurances on the system's risk established beforehand. Common risk control frameworks rely on fixed assumptions and lack mechanisms to continuously monitor deployment reliability. In this work, we propose a general framework for the real-time monitoring of risk violations in evolving data streams. Leveraging the 'testing by betting' paradigm, we propose a sequential hypothesis testing procedure to detect violations of bounded risks associated with the model's decision-making mechanism, while ensuring control on the false alarm rate. Our method operates under minimal assumptions on the nature of encountered shifts, rendering it broadly applicable. We illustrate the effectiveness of our approach by monitoring risks in outlier detection and set prediction under a variety of shifts.

1 INTRODUCTION

The increasing demand for reliable predictions from machine learning systems has driven the development of statistical frameworks for *distribution-free risk control* [Angelopoulos et al., 2025, Bates et al., 2021a]. Such frameworks rely on data-driven inference to achieve their goal, leveraging representative held-out data to determine suitable parameters guiding an application-specific risk, *e.g.* selecting a threshold value for outlier flagging. The hope is that the user can then employ the determined settings indefinitely to aid in reliable decision-making. However, the common validation versus deployment mismatch in machine learning systems has the potential to thwart any 'quality assurance' stamp these methods derive from their static inference. Challenges like outliers, distribution shifts and feedback loops are commonplace [Koh et al., 2021]. In fact, van Amsterdam et al. [2025] argue that an effective machine learning model should *actively* affect the real-world—distribution shift is then not merely an artifact or deployment challenge, but rather a manifestation of a successfully operating system. Hence, any decision-making parameters necessitate *continuous monitoring* during deployment, and the user should be notified when statistical reliability is faltering.

We address this problem by proposing a general framework for the real-time continuous monitoring of bounded risks in evolving data streams, and raising a signal when desired risk levels are in danger of violation. Since alarm signals may trigger costly preventive measures, *e.g.* a production line stop in manufacturing or default loan denial in credit underwriting, it is crucial that false alarms are not raised too often, and our approach effectively controls this rate. We explicitly limit any assumptions on the deployment setting or nature of encountered data, rendering operability under arbitrary or *unknown shifts*. To achieve our goal we adopt the 'testing by betting' paradigm [Ramdas et al., 2023], and cast our monitoring task as a sequential hypothesis testing problem. Leveraging the framework's natural error control properties, our resulting monitoring procedure remains both efficient and statistically rigorous. To summarize, our contributions include:

- In § 3, we motivate sequential testing as a natural approach to continuous risk monitoring, place it in the context of 'testing by betting' and re-interpret the prior method of Feldman et al. [2023] under this lens.
- In § 4, we theoretically outline the statistical properties of our approach, including control over the false alarm rate, asymptotic consistency, and, under some conditions, bounds on the detection time of violations (Prop. 4.5).
- In § 6, we demonstrate the efficacy of our approach against baselines for risk monitoring in outlier detection

[*] Equal contributions. Corresponding authors: <a.r.timans@uva.nl, r.verma@uva.nl>

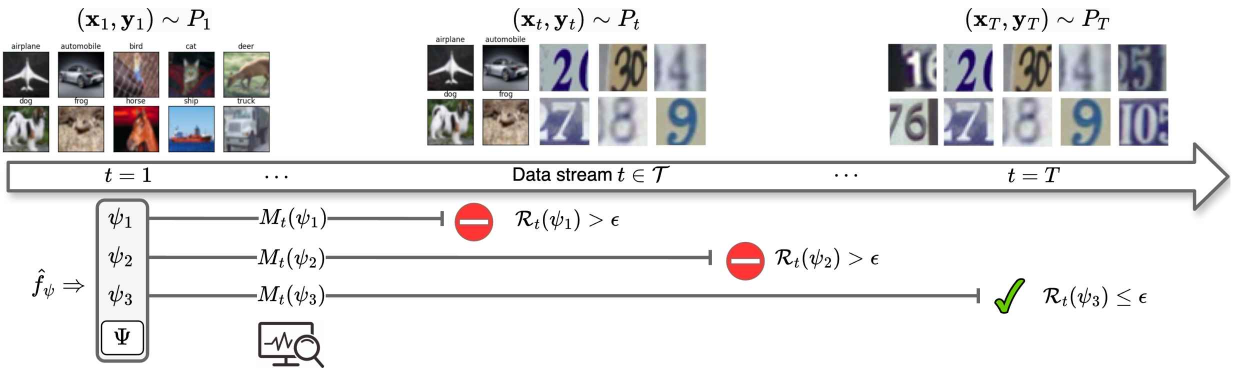
$(\mathbf{x}_1, \mathbf{y}_1) \sim P_1$
airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck
$(\mathbf{x}_t, \mathbf{y}_t) \sim P_t$
airplane
automobile
dog
frog
$(\mathbf{x}_T, \mathbf{y}_T) \sim P_T$
$t = 1$
···
Data stream $t \in \mathcal{T}$
···
$t = T$
$\hat{f}_\psi \Rightarrow$
ψ_1
ψ_2
ψ_3
Ψ
$M_t(\psi_1)$
$M_t(\psi_2)$
$M_t(\psi_3)$
$\mathcal{R}_t(\psi_1) > \epsilon$
$\mathcal{R}_t(\psi_2) > \epsilon$
$\mathcal{R}_t(\psi_3) \leq \epsilon$

Figure 1: We consider an evolving data stream $t = 1, \ldots, T$ susceptible to distribution shifts, *i.e.* observations are drawn from a time-dependent distribution P_t at each step. A predictor $\hat{f}_\psi$ is equipped with a decision-making mechanism governed by a threshold ψ (*e.g.* on outlier flagging). At deployment, we monitor each candidate $\psi \in \Psi$ using a sequential testing process $M_t(\psi)$ which collects evidence for or against risk violations. Risk-violating thresholds are then marked as unreliable.

(§ 6.1) and set prediction tasks (§ 6.2), employing real-world datasets and different shift scenarios including natural temporal shifts.

2 RISK AND PROBLEM FORMULATION

We next describe our notation, problem setting and task in detail, highlighting some key distinctions to existing work.

Notation and risk quantity. Let $\mathcal{X} \times \mathcal{Y}$ denote the sample space with a data-generating distribution P over it, and $\mathbf{x}, \mathbf{y}$ random variables with realizations $\boldsymbol{x}, \boldsymbol{y}$.[1] We consider access to the outputs of a base predictor $\hat{f} : \mathcal{X} \to \mathcal{S}$, where $\mathcal{S} \subseteq \mathbb{R}^{|\mathcal{Y}|}$ for classification or $\mathcal{S} \subseteq \mathbb{R}$ for regression. This model may have been explicitly trained by the user, but can in particular denote a pretrained model without internal access, *e.g.* accessible via an API. Next, similar to existing approaches for risk control [Angelopoulos et al., 2025, 2024a, Feldman et al., 2023], we equip the model with a general decision-making mechanism of the form

$$\hat{f}_\psi(\mathbf{x}) = g(\hat{f}(\mathbf{x}), \psi), \tag{1}$$

where $\psi \in \Psi$, $\Psi \subseteq [0, 1]$ denotes a particular threshold value and g a generic operator instantiated for each task-specific thresholding mechanism. For example, we can define g as a binary decision on outlier flagging given some outlier score computed using $\hat{f}$ (see § 6.1). Finally, a notion of error for $\hat{f}_\psi$ and any particular threshold ψ is captured by a problem-specific *supervised and bounded* loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \Psi \to \mathcal{L}$, $\mathcal{L} \subseteq [0, 1]$, and the resulting *true population risk* is given by the expected loss

$$\mathcal{R}(\psi) = \mathbb{E}_P[\ell(\hat{f}_\psi(\mathbf{x}), \mathbf{y})]. \tag{2}$$

Because $\ell \in \mathcal{L}$ is bounded it also follows that $\mathcal{R}(\psi) \in [0, 1]$. Boundedness of the loss constitutes our key restriction, but we place no conditions on the particular distribution of losses within those bounds [Waudby-Smith and Ramdas, 2024]. To simplify notation we additionally define $\mathrm{z} = \ell(\hat{f}_\psi(\mathbf{x}), \mathbf{y})$ as a random variable of the loss with realization z, and equivalently express the risk in Eq. 2 as $\mathcal{R}(\psi) = \mathbb{E}_P[\mathrm{z}]$. Crucially, $\mathcal{R}(\psi)$ denotes the quantity of interest for which safety assurances of some form are desired in order to robustify decisions made using $\hat{f}_\psi$ (and indirectly, $\hat{f}$).

Static risk control. Assume the deployment of $\hat{f}_\psi$ on new *i.i.d.* test data $\mathcal{D}_{test} \sim P_0$, and access to representative labelled *i.i.d.* calibration data $\mathcal{D}_{cal} \sim P_0$. Following existing frameworks of risk control such as *RCPS* [Bates et al., 2021a] or *Learn-then-Test* [Angelopoulos et al., 2025], $\mathcal{D}_{cal}$ can be leveraged to identify a subset $\hat{\Psi} \subseteq \Psi$ of *risk-controlling* thresholds which ensures a high-probability upper bound on the population risk. That is, for any $\hat{\psi} \in \hat{\Psi}$ we may state that $\mathbb{P}(\mathcal{R}(\hat{\psi}) \leq \epsilon) \geq 1 - \delta$ holds. The risk level $\epsilon \in (0, 1)$ and probability level $\delta \in (0, 1)$ are user-specified, and dictate how tightly the risk is to be controlled. For instance, selecting low values for both ϵ and δ will enforce strong guarantees but may result in overly conservative decision-making on the basis of a chosen $\hat{\psi}$. Crucially, these approaches operate in a *static* batch setting where the set $\hat{\Psi}$ is computed once and deployed indefinitely, and are limited by their assumption on a *static* distribution P_0 over time.

Data stream setting under shift. Instead, let us consider a more dynamic stream setting at deployment time. Given a time index set $\mathcal{T} = \{1, \ldots, T\}$, at every time step $t \in \mathcal{T}$ a covariate $\boldsymbol{x}_t$ is obtained, a decision is made using $\hat{f}_\psi(\boldsymbol{x}_t)$, and subsequently $\boldsymbol{y}_t$ is revealed and the loss z_t measured. Thus, the flow of information at each step follows as *covariate* → *decision* → *label* → *loss*, and at time t the observational history $\{(\boldsymbol{x}_i, \boldsymbol{y}_i, z_i)\}_{i=1}^{t-1}$ is available. If we as-

[1] Upright lettering denotes random variables and italic lettering their realizations. Boldening denotes multi-dimensional quantities.

sume that the test stream originates *i.i.d* $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{t\in\mathcal{T}} \sim P_0$, risk control frameworks as above could be directly applied after observing sufficient samples, and we elaborate further on this simpler setting in § 3.1. In this work, we address the challenging extension to the stream case under *time-dependent and unknown* distribution shifts. Specifically, we consider a data stream observed as $(\boldsymbol{x}_t, \boldsymbol{y}_t) \sim P_t$ for $t \in \mathcal{T}$, where samples at every time step originate from a time-dependent distribution P_t which may shift, and in particular tends to deviate away from any initial P_0. Our risk quantity of interest then becomes $\mathcal{R}_t(\psi) = \mathbb{E}_{P_t}[\mathrm{z}_t]$, the *time-dependent* true population risk at any given time t, and any obtained threshold set $\hat{\Psi}_t \subseteq \Psi$ is similarly time-dependent. We suppose minimal knowledge and place *no assumptions* on the nature of the shift, which may be caused by a single static jump, gradual, *etc.*, and originate in the covariates, labels, or both. Expectedly, the resulting high unpredictability on future risk development renders it substantially harder to provide safety assurances of any kind, but poses a commonly encountered problem setting in practice. Faced with such a challenge at deployment, we examine how to continuously monitor the true risk $\mathcal{R}_t(\psi)$ for candidates ψ and identify when violations of the form $\mathcal{R}_t(\psi) > \epsilon$ occur.

3 RISK MONITORING AS SEQUENTIAL TESTING BY BETTING

We next outline our approach to risk monitoring leveraging sequential hypothesis testing. We motivate how such a testing framework naturally arises by recasting our stream setting as a forecasting 'game' between two agents—the forecaster and nature—and formalizing the collected evidence as an error accumulation process. The procedure is then placed in the context of sequential 'testing by betting' [Ramdas et al., 2023], thereby enjoying the practicality as well as the rigour of the framework. Finally, we theoretically connect our approach to related methods in § 3.1.

A sequential forecasting game. Consider a game between two agents, the *forecaster* and *nature* (*i.e.* the environment). The forecaster provides a guess π_t for the true risk at time t given their knowledge of the observational history, formally encapsulated in the filtration $\mathcal{F}_{t-1} = \sigma(\{(z_1, \pi_1), (z_2, \pi_2), \ldots, (z_{t-1}, \pi_{t-1})\})$ (refer § A.1 for technical definitions). Should the forecaster desire to minimize the mean squared prediction error $\mathbb{E}_{P_t}[(\mathrm{z}_t - \pi_t)^2 \mid \mathcal{F}_{t-1}]$, their best guess is given by $\pi_t = \mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}]$. Nature then reveals the value of z_t, leading to an observable *discrepancy* $\delta_t = z_t - \pi_t$ representing the incurred forecasting error. As the game is repeated, a sequence of discrepancies $(\delta_t)_{t\in\mathcal{T}}$ is iteratively built. Crucially, if the forecaster continues to make their best guess at every step, the resulting discrepancy process forms a martingale difference sequence, *i.e.*, $\mathbb{E}_{P_t}[\delta_t \mid \mathcal{F}_{t-1}] = 0$ and hence asymptotically $\frac{1}{t}\sum_{i=1}^{t} \delta_i \to 0$ as $t \to \infty$. Thus, under the forecaster's best strategy asymptotic alignment between forecasts and actual outcomes is ensured, and systematic deviations in the discrepancies (*i.e.* error accumulation) can serve as evidence, or a testing signal, for such alignment.

Adopting the game to our problem setting, assume the forecaster's guess is upper-bounded as $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] \le \epsilon$. In general, nature has no obvious incentive to align its realizations of z_t with the forecaster. However, in our setting the outcomes are directly affected by the choice of threshold ψ since $\mathrm{z}_t = \ell(\hat{f}_\psi(\mathbf{x}_t), \mathbf{y}_t)$, rendering the associated discrepancy process useful for testing. For each candidate $\psi \in \Psi$, a formal hypothesis test on alignment at risk level ϵ can be formulated as

$$
\begin{aligned}
&H_0(\psi) : \mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] \le \epsilon \;\forall t \in \mathcal{T} \quad \text{(risk controlled)}\\
&H_1(\psi) : \exists t \in \mathcal{T} : \mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] > \epsilon, \quad \text{(risk violated)}
\end{aligned} \tag{3}
$$

and our game suggests that the threshold's discrepancy process $(\delta_t)_{t\in\mathcal{T}}$ can provide the necessary testing evidence.

Example test statistic. Given the sequential test in Eq. 3, how should the discrepancy sequence be leveraged to construct a test statistic? A straightforward choice is the cumulative process $M_t(\psi) = \sum_{i=1}^{t} \lambda_i \cdot \delta_i = \sum_{i=1}^{t} \lambda_i(\mathrm{z}_i - \epsilon)$, where λ_t denotes a non-negative weight associated with the 'trust' placed in the aggregated evidence at time t, dictating how $M_t(\psi)$ evolves. Intuitively, if $H_0(\psi)$ is true and the risk is indeed bounded by ϵ, then the forecaster's guesses should be well aligned and discrepancies exhibit little systematic effects. In that case, $M_t(\psi)$ forms a *supermartingale*, meaning that it is not expected to increase since $\mathbb{E}_{P_t}[\delta_t \mid \mathcal{F}_{t-1}] \le 0$. On the other hand, consistent evidence indicating the risk's growth beyond ϵ will accumulate and drive the growth of $M_t(\psi)$, signaling evidence for rejection in favour of $H_1(\psi)$. Thus the cumulative process provides an viable test statistic for Eq. 3, and we further expand on this approach in § A.2.

Test supermartingales and testing by betting. While the aforementioned summation statistic offers a valid testing procedure, it is not necessarily *efficient* in the sense of optimally accumulating evidence. That is, we want to accumulate the necessary evidence as fast as possible should a risk violation occur. To that end, a rich body of literature on sequential testing through the lens of 'testing by betting' can be leveraged [Ramdas et al., 2023]. Specifically, rather than via summation we may consider the multiplicative accumulation of discrepancies as

$$
M_t(\psi) = \prod_{i=1}^{t} (1 + \lambda_i \cdot \delta_i) = \prod_{i=1}^{t} (1 + \lambda_i (\mathrm{z}_i - \epsilon)), \tag{4}
$$

yielding a universal representation of a *test supermartingale* if we ensure $M_0 = 1$ and $(\lambda_t)_{t\in\mathcal{T}}$ to be a *predictable* process based only on past observations [Ramdas and Wang, 2024]. That is, λ_t may only depend on $\{z_i\}_{i=1}^{t-1}$ (and is thus measurable w.r.t. $\mathcal{F}_{t-1}$). A game-theoretic interpretation can be given to the sequential test and each component in Eq. 4[2]. The forecaster is actively betting against the null hypothesis starting from an initial wealth of $M_0 = 1$, and $M_t(\psi)$ describes the *wealth process* at every subsequent betting round t. The betting rate λ_t denotes the proportion of wealth gambled at each step, and $(z_t - \epsilon)$ the resulting pay-off once nature reveals z_t. Should $H_0(\psi)$ hold, then no betting strategy is expected to systematically increase wealth. On the other hand, a betting strategy resulting in meaningful wealth accumulation points towards evidence against the null. A rejection threshold can be employed to reach a final testing decision with *stopping time* $\tau(\psi) \in \mathcal{T}$, denoting the time step at which $H_0(\psi)$ has been ruled out by the wealth process.

Constructing threshold confidence sets. Since the risk associated with every threshold needs to be monitored simultaneously, we instantiate a number of wealth processes $M_t(\psi)$ in parallel, one for each candidate $\psi \in \varPsi$. Their joint behaviour can be encapsulated in a *confidence set* (ψ-CS) of valid thresholds at every time step t, constructed as

$$C_t^{\psi} = \{\psi \in \varPsi \ : \ M_t\left(\psi\right) < 1/\delta\}. \tag{5}$$

That is, using the predefined risk control parameters ϵ, δ, the confidence set C_t^{ψ} denotes the set of thresholds at time t for which $H_0(\psi)$ has not yet been rejected. It is, in effect, the equivalent of the threshold set $\hat{\varPsi}_t \subseteq \varPsi$ described in § 2 using the particular rejection threshold $1/\delta$. Crucially, by leveraging the stopping rule $1/\delta$ and test martingale properties of $M_t(\psi)$, any threshold that does *not* violate the risk level ϵ at time t is guaranteed to be included in C_t^{ψ} with high probability, *i.e.*, it holds that $\mathbb{P}_{H_0}(\forall t \in \mathcal{T} \ : \ M_t(\psi) < 1/\delta) > 1 - \delta$. We interpret this Type-I error control property as a *false alarm guarantee* on erroneous rejection, and elaborate upon it in § 4. In addition, the size of the ψ-CS can be interpreted as an indicator for the stream's shift intensity, and thus the underlying model's deployment reliability. A constant set size indicates temporally stable threshold choices are available, whereas a shrinkage of C_t^{ψ} towards zero implies that all thresholds eventually signal risk violation, necessitating a more substantial model update using the observational history. Since we are preoccupied with risk *monitoring* only, we leave the discussion on model updating, or *safe adaptation*, for future work.

Practical considerations. An important distinction to stress is that any false alarm guarantee holds *across time* for every threshold, and not *across thresholds* at every time step. Thus no guarantees can be given on an adaptive strategy to select a particular $\hat{\psi}_t \in C_t^{\psi}$ at every step, unless multiple testing corrections (which we do not consider here) are introduced to control for the multi-stream setting, *e.g.* drawing inspiration from Xu and Ramdas [2024], Dandapanthula and Ramdas [2025]. Empirically, one may adopt strategies such as selecting a stable threshold that persists over extended time horizons or a threshold to maximize significant results (*e.g.* the minimum value). These choices also relate to the *risk profile* of $\mathcal{R}_t(\psi)$, which dictates if a 'trivial' stable solution (that may be very conservative) is available, and facilitates the interpretability of the obtained ψ-CS. A preferable risk profile will behave both monotonically across time (*e.g.*, $\lim_{t\to 0} \mathcal{R}_t(\psi) = 0$ and $\lim_{t\to T} \mathcal{R}_t(\psi) = 1$) as well as across thresholds (*e.g.*, $\lim_{\psi\to 0} \mathcal{R}_t(\psi) = 0$ and $\lim_{\psi\to 1} \mathcal{R}_t(\psi) = 1$). However, we do not assume such conditions and our experiments in § 6 address non-monotonic behaviour in either argument.

3.1 RELATION TO OTHER APPROACHES

We next draw connections to other notions of risk control in the literature. Most notably, we leverage our formulation in terms of discrepancy processes to provide a novel interpretation of rolling risk control [Feldman et al., 2023] as an implicit, adaptive form of sequential testing with asymptotic guarantees (as opposed to finite-sample). We then contrast our shifting stream setting with the simpler *i.i.d.* case.

Sequential testing and rolling risk control. Proposed by Feldman et al. [2023] as an extension of Gibbs and Candès [2021] to bounded risks beyond the miscoverage rate, *rolling risk control* (RRC) aims to track the running estimate $\frac{1}{t}\sum_{i=1}^{t} z_i$ and ensure its asymptotic adherence to the risk level ϵ via the update rule $\psi_t = \psi_0 + \sum_{i=1}^{t-1} \gamma\,(z_i - \epsilon)$, where $\gamma > 0$ denotes a step size. The 'calibration parameter' ψ governs the behaviour of their set predictor, rendering it an instantiation of our threshold model $\hat{f}_{\psi}$ (see also § 6.2). Procedurally, the model is initialized with value ψ_0 and RRC incrementally updates the parameter at every time step following the rule. Leveraging our discrepancy process interpretation, we can directly observe that RRC accumulates evidence via discrepancies $\delta_t = z_t - \epsilon$ over time, and is mathematically analogous to the summation wealth process used as an example in § 3. More formally, we can denote the process $\psi_t = \psi_{t-1} + \gamma\ (\mathrm{z}_{t-1} - \epsilon)$, and under the null (Eq. 3) it follows that $\mathbb{E}\left[\psi_t \mid \mathcal{F}_{t-1}\right] \leq \psi_{t-1}$ indicates a risk-controlling parameter, whereas under the alternative $\mathbb{E}[\mathrm{z}_t|\mathcal{F}_{t-1}] > \epsilon$, and ψ_t thus grows as a martingale accumulating evidence against the null. The 'testing by betting' interpretation helps clarify the key distinction to our approach—how the designed wealth process is subsequently utilized. Whereas we take a testing decision on the basis of a rejection threshold, RRC does not enforce such a stopping rule but re-invests the wealth in an update step to dynamically adjust prediction set sizes. While offering a convenient step towards model adaptation, the rule is tied to the explicit monotonicity assumption underlying RRC,

[2] The evidence collection process $M_t(\psi)$ can be interchangeably referred to as a *test (super)martingale* by its mathematical properties, *wealth process* by its betting interpretation, or *E-process* in the context of the sequential testing literature.

wherein a larger ψ_t reduces the risk by enlargening the prediction set and vice versa. Such monotonic behaviour in the thresholds is desirable, but not always available.

Risk control under the i.i.d. data stream setting. In the simple case where the test stream originates *i.i.d* $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{t\in\mathcal{T}} \sim P_0$ rather than from time-dependent distributions P_t, we obtain that $\mathcal{R}_t(\psi) = \mathcal{R}_0(\psi)$ is a time-*independent* risk, and the independence between samples further simplifies the risk definition (we detail our argument in § A.3). We may then conveniently reverse the hypotheses pair from Eq. 3 to form the test

$$H_0(\psi) : \exists t \in \mathcal{T} : \mathcal{R}_0(\psi) > \epsilon,\ H_1(\psi) : \mathcal{R}_0(\psi) \leq \epsilon\ \forall t \in \mathcal{T}$$

and define a reverse wealth process as $M_t(\psi) = \prod_{i=1}^{t} (1 + \lambda_i (\epsilon - \mathrm{z}_i))$. The corresponding ψ-CS construction is given by $C_t^{\psi} = \{\psi \in \Psi : M_t(\psi) \geq 1/\delta\}$. Crucially, we can exploit the fact that P_0 is static and thus any drawn test conclusions on *non-violation* of the risk (now formalized in $H_1(\psi)$) hold indefinitely in the future. In other words, C_t^{ψ} will only grow as more thresholds are found to be safe, but never shrink. This permits leveraging the Type-I error control property of the wealth process to claim strong *time-uniform* risk control guarantees of the form $\mathbb{P}(\forall t \in \mathcal{T} : \mathcal{R}_0(\psi) \leq \epsilon) \geq 1 - \delta$, moving beyond static risk control assurances. This approach has been leveraged directly by Xu et al. [2024] to extend *RCPS* [Bates et al., 2021a] to the stream setting, and indirectly by Zecchin and Simeone [2025] to make *Learn-then-Test* [Angelopoulos et al., 2025] more adaptive. Naturally, such forward-looking assurances continue to only hold for the setting of a *static* distribution P_0, and are not applicable in our challenging shift setting.

4 THEORETICAL ANALYSIS

Next, we establish the theoretical guarantees our approach enjoys. To summarize, we suggest monitoring the risk using the wealth process $M_t(\psi)$ (Eq. 4) to detect risk violations and raise a signal when the stopping rule $1/\delta$ is met. Running this procedure simultaneously for all candidates $\psi \in \Psi$, a set of thresholds C_t^{ψ} (Eq. 5) deemed non-violating is returned at every time step t. In turn, $\Psi \backslash C_t^{\psi}$ denotes the thresholds for which a test decision on risk violation at time t has been made. In the following series of statements we provide insights on the *statistical validity* and *efficiency* of our approach, deferring all proofs to § A.4. We begin by stating the essential martingale properties of the process $M_t(\psi)$ which render it practical for risk monitoring.

Lemma 4.1 (Valid wealth process)**.** *For any $\psi \in \Psi$ such that $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] \leq \epsilon\ \forall t \in \mathcal{T}$ satisfies the null, the process $M_t(\psi)$ in Eq. 4 is a valid test supermartingale for the predictable betting rate $\lambda_t \in [0, 1/\epsilon)$.*

This follows directly from satisfying the conditions of a valid test supermartingale (§ A.1), and $\lambda_t \in [0, 1/\epsilon)$ ensures that $M_t(\psi)$ remains non-negative with its expected value upper-bounded by $M_0 = 1$ under the considered null hypothesis $H_0(\psi)$ (Eq. 3). We next characterize the *false alarm guarantee* that our procedure natively derives from Lemma 4.1 via its Type-I error control property, given as

Lemma 4.2 (False alarm guarantee)**.** *For any $\psi \in \Psi$ such that $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] \leq \epsilon\ \forall t \in \mathcal{T}$ satisfies the null, it holds that $\mathbb{P}(\exists t \in \mathcal{T} : M_t(\psi) \geq 1/\delta) \leq \delta$.*

The result ensures that a false positive or false alarm—when $M_t(\psi)$ crosses $1/\delta$ and *incorrectly* signals risk violation as the threshold ψ is in fact risk-controlling—occurs with at most probability δ. Consequently, any threshold that does *not* violate the risk level ϵ at time t is guaranteed to be included in C_t^{ψ} with high probability, *i.e.*, it holds that $\mathbb{P}_{H_0}(\forall t \in \mathcal{T} : M_t(\psi) < 1/\delta) > 1 - \delta$. Such high-probability safety assurances can be given under the null hypothesis $H_0(\psi)$ but do not translate to the alternative (see also § 3.1 for the reverse hypotheses). Leveraging the properties of $M_t(\psi)$, what statements can be made with respect to $H_1(\psi)$, indicating true risk violation? First, we establish the method's asymptotic consistency under regular conditions.

Lemma 4.3 (Asymptotic consistency)**.** *For any $\psi \in \Psi$ such that $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] \leq \epsilon$ for finitely many steps $t \in \mathcal{T}$ and $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] > \epsilon$ otherwise, it holds that $\mathbb{P}(\tau(\psi) < \infty) = 1$, where $\tau(\psi)$ denotes the stopping time.*

In words, our hypothesis test aligns with the classical *sequential test of power one* [Darling and Robbins, 1968] and assures that a risk-violating threshold will be inevitably detected (with finite stopping time). While this ensures asymptotic correctness, a practical application demands more than *eventual* detection—it requires efficiency in the evidence accumulation. We borrow such a notion of statistical efficiency by directly leveraging the property of *growth rate optimality* to guide our betting rate λ_t. Informally, an adaptive strategy for the bets $(\lambda_t)_{t\in\mathcal{T}}$ can be designed to directly maximize the growth of the wealth process under the alternative, ensuring efficient evidence accumulation [Waudby-Smith and Ramdas, 2024, Koolen and Grünwald, 2022]. We formally define the condition as

Definition 4.4 (Growth rate optimality (GRO))**.** *The betting rate λ_t is growth rate optimal if it satisfies the condition $\lambda_t = \arg\max_{\lambda\in[0,1/\epsilon)} \mathbb{E}_{H_1}[\log M_t(\psi)]$.*

We follow the GRO principle in our experiments (see § 6), and hence ensure our procedure retains maximal efficiency (or statistical power) among all possible wealth processes. Finally, under some additional conditions we propose characterizing the method's stopping time behaviour, and by extension its *detection delay* denoting the practical utility. Defining $\tau_*(\psi) = \inf\{t \in \mathcal{T} : \mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] > \epsilon\}$ as the (unknown) *true* stopping time of ψ (*i.e.*, when the true risk is violated), and $\tau(\psi) = \inf\{t \in \mathcal{T} \ : \ M_t(\psi) \geq 1/\delta\}$ as our stopping signal, the method's detection delay is given by $\tau(\psi) - \tau_*(\psi)$. We propose the following characterization:

Proposition 4.5 (Detection delay bound)**.** *A worst-case detection delay for the hypothesis pair in Eq. 3 and wealth process $M_t(\psi)$ in Eq. 4 is characterized by $(\tau(\psi) - \tau_*(\psi)) \approx \mathcal{O}((\log(1/\delta) + T)/(\lambda \cdot \mu))$, where μ denotes the risk violation intensity and T a shift changepoint.*

We elaborate on Prop. 4.5 in § A.4, but can intuitively observe *(i)* inverse proportionality to both the betting rate λ and the violation strength μ, indicating a faster detection when evidence grows adaptively and is strong; and *(ii)* direct proportionality to T, indicating a slower detection when shifts occur later in the stream as the initial evidence favouring $H_0(\psi)$ needs to be overcome (slowing the wealth's growth).

5 RELATED WORK

Waudby-Smith and Ramdas [2024] offer an in-depth study on detecting deviations in the means of bounded quantities for stream settings, providing useful tools (*e.g.* in terms of betting rate design) for the testing of risks following Eq. 2. Very recently, Fan et al. [2025] explore some settings for additional bounds on the variance. Whereas they consider data to originate from a fixed source distribution P with unknown mean μ to be estimated, our observations originate from variable, time-dependent distributions P_t, and we simultaneously monitor multiple means (corresponding to the threshold-dependent risks $\mathcal{R}_t(\psi)$). Thus our underlying hypothesis dictating the test design is subtly, but distinctly different. Yet, strong results on the universal representation of test martingales (*e.g.* stated by Waudby-Smith and Ramdas [2024], Prop. 3) render the process structure in Eq. 4 useful for a wide range of testing problems. We attempt to intuitively motivate this via a 'forecasting game' in § 3.

Sequential testing and risk monitoring. Xu et al. [2024], Zecchin and Simeone [2025] leverage above results on deviations in means to provide strong time-uniform risk control for *i.i.d* streams, discussed further in § 3.1. Closely related to our work, Podkopaev and Ramdas [2022] monitor a *running risk* of the form $\mathcal{R}_r(\psi) = \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}_{P_i}[\mathrm{z}_i]$ under the sequential testing framework, with similar false alarm guarantees. However, we consider the more challenging instantaneous true risk $\mathcal{R}_t(\psi)$ at any given time step, which can recover $\mathcal{R}_r(\psi)$ but not vice versa. Furthermore, their experimental design tends to distinguish between benign and harmful shifts caused by a dominant shift initiated at P_0 (akin to changepoint detection), whereas we incorporate a broader variety of shifts. Finally, we do not impose sample independence. Their approach was reformulated by Amoukou et al. [2024] for unlabelled streams, and relatedly Bar et al. [2024] suggest an unsupervised covariate shift detector on the basis of entropy-matching. Particular to out-of-distribution detection, Vishwakarma et al. [2024], Sun et al. [2024] also leverage martingale-based constructions.

Other sequential testing under shift. Stream data denotes a particular test setting under the 'testing-by-betting' framework, lending itself naturally to the use of test martingales[3] [Ramdas et al., 2023, Ramdas and Wang, 2024]. Within that framework, previous work has considered a variety of testing problems, such as on exchangeability [Vovk, 2021, Saha and Ramdas, 2024], independence [Podkopaev and Ramdas, 2023], two-sample testing [Shekhar and Ramdas, 2021, Pandeva et al., 2024a,b, Luo et al., 2024], or changepoint detection [Shekhar and Ramdas, 2023b, 2024, Vovk et al., 2021, Volkhonskiy et al., 2017, Shin et al., 2023]. Theses works differ from ours in terms of the hypotheses they address, their data settings (*e.g.* by simultaneously observing two separate data streams), or experimental designs (*e.g.* detecting a single changepoint or shift). Additional related works on static risk control and extensions to stream settings not using sequential testing can be found in Appendix B.

6 EMPIRICAL RESULTS

We empirically validate our risk monitoring approach on two tasks, outlier detection (§ 6.1) and set prediction (§ 6.2). The first experiment induces shifts by mixture sampling in order to demonstrate monitoring behaviour for explicit scenarios, while the latter is based on naturally occuring temporal shifts. Evaluating diverse, real-world datasets, we find that the method ensures both timely detection of risk violations and a controlled false alarm rate. We next outline our baselines and practical design choices, followed by each experiment in more detail (see also Appendix C). Our code is publicly available at `https://github.com/alextimans/risk-monitor`.

Baselines. We compare our primary monitoring approach, the wealth process $M_t(\psi)$ described by Eq. 4, to the following (empirical) risk tracking mechanisms:

(i) An empirical estimate of the *unobservable oracle* or true population risk $\mathcal{R}_t(\psi)$, computed as $\hat{\mathcal{R}}_t(\psi) = \frac{1}{B_*}\sum_{b=1}^{B_*} z_{t,b},\ z_{t,b} \sim P_t$ for a batch draw B_* of large size (*e.g.* $B_* = 1000$). We desire for $M_t(\psi)$ to emulate the monitoring behaviour of $\hat{\mathcal{R}}_t(\psi)$ as closely as possible while controlling false alarms by Lemma 4.2.

(ii) An empirical estimate of the running risk $\mathcal{R}_r(\psi)$, accumulated over the data stream for a given time step t as $\hat{\mathcal{R}}_r(\psi) = \frac{1}{t}\sum_{i=1}^{t} z_i$. This is the risk quantity evaluated both by Podkopaev and Ramdas [2022] as a tractable estimate of the running risk, and Feldman et al. [2023]

[3] Also referred to therein as *sequential anytime-valid inference*.

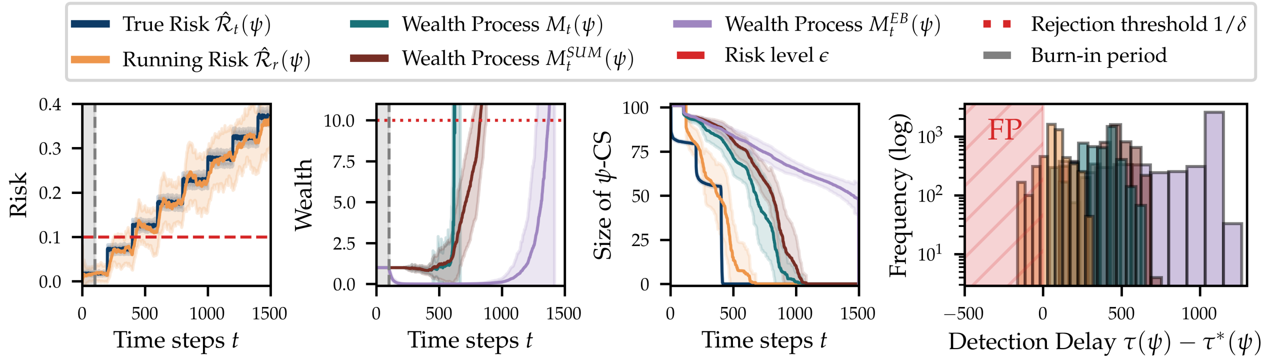

Figure 2: Results for **outlier detection with a stepwise shift** (§ 6.1). *From left to right:* Visuals of the growing risk and wealth process behaviour with respective rejection thresholds ϵ and $1/\delta$, for a single threshold candidate (here $\psi = 0.50$); the behaviour of the valid threshold set ψ-CS (Eq. 5), which eventually shrinks to zero signalling a model update; and the empirical distributions of detection delays $\tau(\psi) - \tau_*(\psi)$ across all $\psi \in \Psi$, including the false alarm region (FP). We also have $B = 1, S = 50$ and $t_{out} = 200$, with results evaluated over $R = 50$ trials (mean and std. deviation).

directly as a rolling risk target (§ 3.1). Note that since we are monitoring the more challenging instantaneous risk $\mathcal{R}_t(\psi)$, the estimator $\hat{\mathcal{R}}_r(\psi)$ is nominally void of any false alarm guarantees.

(iii) The summation wealth process illustrated in § 3, given by $M_t^{SUM}(\psi) = \sum_{i=1}^{t} \lambda_i (z_i - \epsilon)$. This process retains the same false alarm guarantees as $M_t(\psi)$, but tends to be less adaptive as evidence is accumulated additively.

(iv) The *predictably-mixed Empirical-Bernstein* wealth process from Waudby-Smith and Ramdas [2024], given by $M_t^{EB}(\psi) = \prod_{i=1}^{t} \exp\{\lambda_i (z_i - \epsilon) - v_i\, \rho(\lambda_i)\}$ where $v_i = 4\,(z_i - \hat{\mu}_{i-1})^2$, $\rho(\lambda_i) = 1/4\,(-\log(1-\lambda_i) - \lambda_i)$, and we use the *predictable plug-in* betting rate $\lambda_i^{EB} = \min\left\{\sqrt{\frac{2\,\log(2/\delta)}{\hat{\sigma}_{i-1}^2\; i\; \log(1+i)}}, \frac{1}{2}\right\}$. A similar method is also derived by Podkopaev and Ramdas [2022] to estimate confidence bounds on $\mathcal{R}_r(\psi)$ in their problem setting, but we employ its direct form as a sequential test.

Choice of betting rate. We follow the growth rate optimality (GRO) condition outlined in Definition 4.4 to guide our choice of betting rate. Whereas selecting λ_t based on direct wealth maximization is possible, the approach can be computationally expensive to re-evaluate for every candidate ψ and time step t. Instead, we leverage a suggested approximation by Waudby-Smith and Ramdas [2024], yielding the closed-form expression

$$\lambda_t^{AGR} = \max\left\{0, \min\left\{\frac{\hat{\mu}_{t-1} - \epsilon}{\hat{\sigma}_{t-1}^2 + (\hat{\mu}_{t-1} - \epsilon)^2}, \frac{1/2}{\epsilon}\right\}\right\},$$

where $\hat{\mu}_{t-1}$ and $\hat{\sigma}_{t-1}^2$ denote the estimated running mean and variance over $\{z_i\}_{i=1}^{t-1}$. Intuitively, the betting rate increases when the running mean is far from ϵ, and is further amplified by a small variance. λ_t^{AGR} is *approximately* GRO [Shekhar and Ramdas, 2023a] and performs empirically similar to direct maximization. A range of other suitable bets is discussed in Waudby-Smith and Ramdas [2024], and we briefly touch upon this in Appendix C.

Batching, sliding window and burn-in. Instead of a data stream where samples arrive individually at every time step, we may also consider the arrival of small batches of size $B \ll B_*$, *i.e.* we sample $\{(\boldsymbol{x}_{t,b}, \boldsymbol{y}_{t,b})\}_{b=1}^{B} \sim P_t$. The batch-wise evidence at every time step can be easily aggregated by, for instance, averaging, which tends to both reduce the variance of the tracking process and improve the detection delay $\tau(\psi) - \tau_*(\psi)$ with respect to the true risk even for small batches ($B = 10$). Similarly, delays can be reduced by enhancing the adaptivity of any tracker via a sliding window of size S, wherein only the most recent observations for time steps $i \in [t - S, t]$ are considered. Intuitively, the observational history is truncated by discarding past information deemed irrelevant for the current shift environment. This renders the tracking process more reactive (*e.g.*, via the betting rate parameters $\hat{\mu}_{t-1}, \hat{\sigma}_{t-1}^2$) but also increases sensitivity to the retained samples, heightening the chance of false alarms if the resulting evidence is misleading. The choice of B and S can sometimes be delicate, and results for different combinations are provided in Appendix D. Finally, we introduce an initial number of *burn-in* time steps $t_{burn} = \lfloor 100/B \rfloor$ during which any risk tracker (aside of the true risk) does not test for risk violation but merely accumulates samples, in order to stabilize any running quantities such as $\hat{\mathcal{R}}_r(\psi)$.

6.1 MONITORING THE TOTAL ERROR RATE FOR OUTLIER DETECTION

We first consider the task of outlier detection, and instantiate the threshold predictor from Eq. 1 as $\hat{f}_\psi(\mathbf{x}) = \mathbb{1}[\texttt{out}(\mathbf{x}) \geq \psi]$, where $\texttt{out} : \mathcal{S} \to [0, 1]$ maps the predictor's output to a bounded outlier score and $\mathbb{1}[\cdot]$ is the indicator function. When $\texttt{out}(\mathbf{x}) \geq \psi$ evaluates

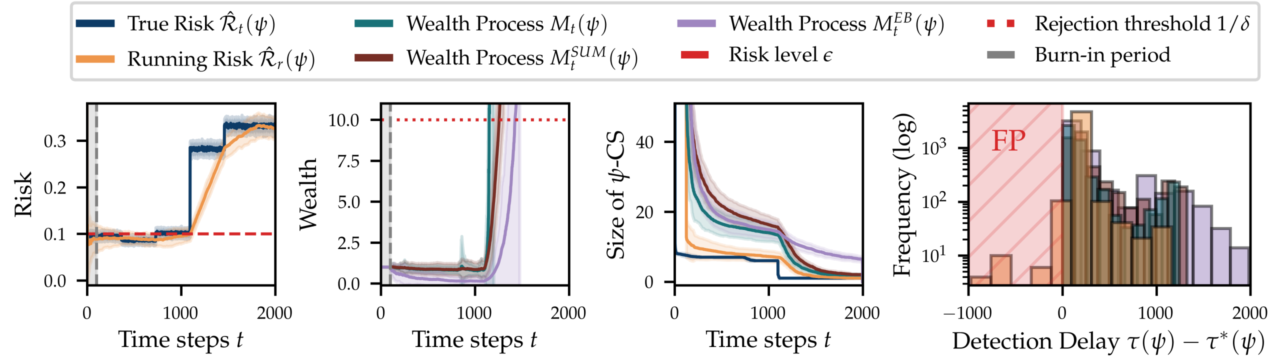

Figure 3: Results for **set prediction with a temporal shift on FMoW** (§ 6.2). *From left to right:* Visuals of the growing risk and wealth process behaviour with respective rejection thresholds ϵ and $1/\delta$, for a single threshold candidate (here $\psi = 0.08$); the behaviour of the valid threshold set ψ-CS (Eq. 5), which eventually tends to zero signalling a model update; and the empirical distributions of detection delays $\tau(\psi) - \tau_*(\psi)$ across all $\psi \in \Psi$, including the false alarm region (FP). We also have $B = 1$ and $S = 365$ (one year), with results evaluated over $R = 50$ trials (mean and std. deviation).

true we declare the sample an outlier. Given a classification setting, we define out as the normalized entropy of the base model's predictive distribution $\hat{p}(\mathbf{y} \mid \mathbf{x})$[4]. The target risk to monitor is given by the *total error rate*, accounting for both cases of inlier (false positives, FP) and outlier misclassification (false negatives, FN) via the loss variable

$$\mathrm{z}_t = \begin{cases} 1, & \text{if } \mathtt{out}(\mathbf{x}_t) \geq \psi \text{ and } (\mathbf{x}_t, \mathbf{y}_t) \sim P_{in}, & \text{(FP)} \\ 1, & \text{if } \mathtt{out}(\mathbf{x}_t) < \psi \text{ and } (\mathbf{x}_t, \mathbf{y}_t) \sim P_{out}, & \text{(FN)} \\ 0 & \text{else.} \end{cases}$$

P_{in} and P_{out} denote inlier and outlier distributions, and the shifting stream is characterized by a time-dependent outlier probability π_t^{out} such that $(\boldsymbol{x}_t, \boldsymbol{y}_t) \sim (1 - \pi_t^{out})\, P_{in} + \pi_t^{out}\, P_{out}$ for $t \in \mathcal{T}$ is generated by mixture sampling. We consider three distinct shift settings: *(i)* an *i.i.d* stream where trivially $\pi_t^{out} = 0$ across all time steps; *(ii)* an immediate stark outlier shift where $\pi_t^{out} = 1$ early on; and *(iii)* a stepwise shift with $\pi_t^{out} \in \{0, 0.05, 0.1, \ldots, 1\}$ increasing every t_{out} time steps. Risk parameters are set to common values $\epsilon = 0.1, \delta = 0.1$, and we simulate for $T = 1500$ steps. P_{in} and P_{out} are given by CIFAR-10 [Krizhevsky et al., 2009] and SVHN [Netzer et al., 2011] respectively, with a base classifier (ResNet-50) trained on CIFAR-10.

Our results in Fig. 2 for the stepwise shift assert that as the shift intensity increases, so does the number of risk-violating thresholds, leading to a gradual shrinkage of the ψ-CS towards zero. Among risk trackers the running risk $\hat{\mathcal{R}}_r(\psi)$ emulates the true risk well but tends to misinterpret evidence, resulting in an undesirable number of false alarms. In contrast, all martingale-based trackers uphold the guarantee, at the cost of increased detection delays. Among them, the monitoring behaviour of the wealth process $M_t(\psi)$ most closely aligns with the true risk, striking a good trade-off. Results for other shift settings can be found in Appendix D, where as anticipated *(i)* for the *i.i.d* case most thresholds remain valid and the ψ-CS stabilizes over the full data stream; and *(ii)* for the immediate shift all thresholds are rejected as soon as possible, correctly identifying $\hat{f}_\psi$ as highly unreliable.

6.2 MONITORING THE MISCOVERAGE RATE FOR SET PREDICTION

Next we consider set prediction tasks on data subject to *natural temporal shifts*, both for the classification and regression setting.

Functional Map of the World. For classification, we instantiate $\hat{f}_\psi$ as a set predictor of the form

$$\hat{f}_\psi(\mathbf{x}) = \{\boldsymbol{y} \in \mathcal{Y} : \hat{p}(\mathbf{y} = \boldsymbol{y} \mid \mathbf{x}) \geq \psi\},$$

and the base classifer once more returns a predictive distribution $\hat{p}(\mathbf{y} \mid \mathbf{x})$ used to determine class inclusion in the set. A natural risk to monitor here is the *miscoverage rate* with loss variable $\mathrm{z}_t = \mathbb{1}[\boldsymbol{y}_t \notin \hat{f}_\psi(\mathbf{x}_t)]$, where $\boldsymbol{y}_t$ denotes the true label. We consider the *Functional Map of the World* dataset (FMoW) [Christie et al., 2018], a large-scale satellite image dataset on building and land use over 16 years, and employ a time-dependent partitioning proposed by Yao et al. [2022]. Therein a natural shift occurs as the *same* satellite image locations capture land use changes over time. The classifier (DenseNet-121) is trained on the first 11 years, and we increase the test stream frequency by sampling chronologically from the final five years every 365 time steps (simulating daily observations). We again set $\epsilon = 0.1, \delta = 0.1$ and run for $T = 2000$ steps.

Our results in Fig. 3 draw similar conclusions as in § 6.1, that is, the proposed wealth process $M_t(\psi)$ produces

[4] This is merely one possible choice, and can be easily swapped for other scoring mechanisms satisfying the required bounds.

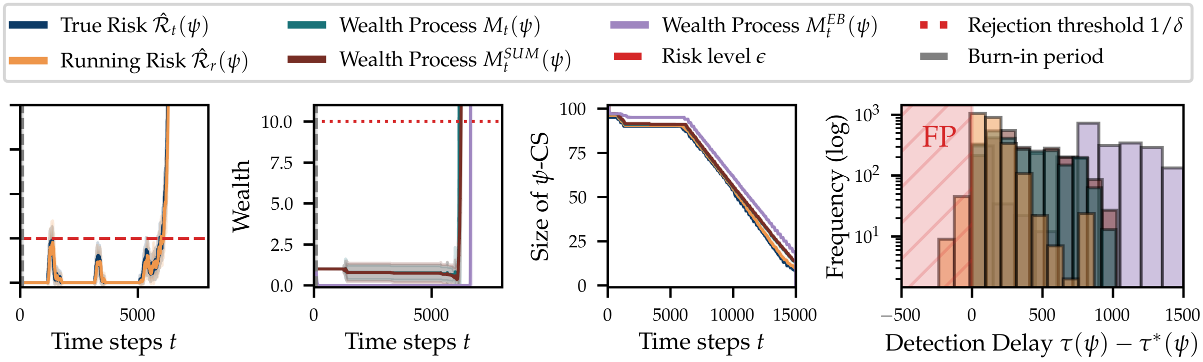

Figure 4: Results for **set prediction with a temporal shift on Naval propulsion** (§ 6.2). *From left to right:* Visuals of the growing risk and wealth process behaviour with respective rejection thresholds ϵ and $1/\delta$, for a single threshold candidate (here $\psi = 0.005$); the behaviour of the valid threshold set ψ-CS (Eq. 5), which eventually tends to zero signalling a model update; and the empirical distributions of detection delays $\tau(\psi) - \tau_*(\psi)$ across all $\psi \in \mathcal{H}$, including the false alarm region (FP). We also have $B = 1$ and $S = 50$, with results evaluated over $R = 50$ trials (mean and std. deviation).

the lowest detection delays among all risk monitoring processes with false alarm control, while the running risk prematurely rejects some threshold candidates. Interestingly, the natural temporal shift induces a non-monotonic risk profile, wherein miscoverage for the second year slightly drops, but thereafter starkly increases. We elaborate on the connection between risk profiles and threshold behaviour in Appendix C, and provide complete results in Tab. 4.

Naval propulsion system. For regression, the set predictor takes the interval form $\hat{f}_\psi(\mathbf{x}) = [\hat{f}(\mathbf{x}) - \psi,\ \hat{f}(\mathbf{x}) + \psi]$, with $\hat{f}(\mathbf{x})$ returning point estimates. The target risk remains the miscoverage rate, and we consider predictive maintenance data on naval gas turbine behaviour [Cipollini et al., 2018]. This tabular time series consists of $\sim 12\,000$ recordings for various turbine system parameters, and an associated turbine compressor degradation coefficient denoting the compressor's health. Over time this degradation coefficient steadily increases, denoting a gradual equipment decay. We train a Random Forest regressor on the initial 'healthy' compressor state (enriched via jittered resampling) which expectedly fails to extrapolate as the degradation worsens, resulting in decreased performance in line with a temporal shift. Once more we have $\epsilon = 0.1, \delta = 0.1$ and run our monitoring process for the full time series.

Our results in Fig. 4 and summarized in Tab. 5 draw consistent conclusions with other experiments. Specifically, we observe *(i)* incurred false positives by the running risk, in particular for more adaptive tracking windows (smaller S); and *(ii)* lowest detection delays for the wealth process $M_t(\psi)$ among all trackers with false alarm guarantees. Furthermore, the gap between running risk and wealth process remains fairy narrow under most realistic settings (*i.e.*, small B and large S). Visually, the ψ-CS (Eq. 5) stabilizes during the initial healthy state, but consistently shrinks as turbine compressor degradation and thus distributional shift worsens. The visualized threshold, being very small, displays sensitive risk behaviour even during early time steps.

7 DISCUSSION

We investigate sequential testing-based approaches to monitor an unobservable, time-dependent bounded risk $\mathcal{R}_t(\psi)$ in a dynamic data stream setting, challenged by unknown and repeated distribution shifts. Motivated by the 'testing by betting' framework [Waudby-Smith and Ramdas, 2024], our martingale-based monitoring process ensures timely detection of risk violations whilst providing finite-sample control over the false alarm rate. This renders a statistically rigorous and yet practical procedure for risk monitoring.

However, we are inherently limited in our safety assurances by the unpredictability of any occuring shift, and the minimal assumptions we impose on it. More informative forward-looking assurances can potentially be obtained if additional restrictions are considered, such as constraints on the shift origin or its intensity and growth rate. Similarly, rephrasing our problem statement in terms of a different hypothesis might simplify the task and offer more efficient or *unsupervised* monitoring, *e.g.* by drawing inspiration from Bar et al. [2024]'s entropy-matching idea or Amoukou et al. [2024]'s label-free quantile test. Possible unsupervised extensions may include recasting the task as two-sample testing [Pandeva et al., 2024a,b], using generalization estimation [Baek et al., 2022, Rosenfeld and Garg, 2023], or leveraging calibration properties [Gupta et al., 2020]. Similarly, the model update step can be integrated into the framework, *e.g.* via test-time adapation [Schirmer et al., 2024] or online and continual learning principles [Wang et al., 2024]. Ultimately, the provision of practical safety assurances for robust model behaviour at deployment *under arbitrary shift* is a challenging problem [Fang et al., 2022].

Author Contributions

AT and RV contributed together to ideation and methodology. AT developed and conducted all experiments and led paper writing, while RV initiated the problem statement and the preliminary approach, developed the theoretical motivation and proofs and co-authored sections of the paper. EN provided project guidance and general feedback. CN contributed to ideation, theoretical development, project guidance and general feedback.

Acknowledgements

We thank Teodora Pandeva for initial discussions that helped shape the project. Additionally, we thank Wouter Koolen and Yaniv Romano for insightful discussions. This project was generously supported by the Bosch Center for Artificial Intelligence. Eric Nalisnick did not use any resources from Johns Hopkins for this work.

References

Salim I Amoukou, Tom Bewley, Saumitra Mishra, Freddy Lecue, Daniele Magazzeni, and Manuela Veloso. Sequential Harmful Shift Detection Without Labels. *Neural Information Processing Systems*, 2024.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 2023a.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal Prediction: A Gentle Introduction. *Foundations and Trends® in Machine Learning*, 2023b.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. *International Conference on Learning Representations*, 2024a.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn Then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *The Annals of Applied Statistics*, 2025.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. *International Conference on Machine Learning*, 2024b.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 2022.

Yarin Bar, Shalev Shaer, and Yaniv Romano. Protected Test-Time Adaptation via Online Entropy Matching: A Betting Approach. *Neural Information Processing Systems*, 2024.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 2023.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, Risk-controlling Prediction Sets. *Journal of the ACM*, 2021a.

Stephen Bates, Emmanuel J. Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 2021b.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World. *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2018.

Francesca Cipollini, Luca Oneto, Andrea Coraddu, Alan John Murphy, and Davide Anguita. Condition-based maintenance of naval propulsion systems with supervised data analysis. *Ocean Engineering*, 2018.

Sanjit Dandapanthula and Aaditya Ramdas. Multiple testing in multi-stream sequential change detection. *arXiv Preprint (arXiv:2501.04130)*, 2025.

Donald A Darling and Herbert Robbins. Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences*, 1968.

Yixuan Fan, Zhanyi Jiao, and Ruodu Wang. Testing the mean and variance by e-processes. *Biometrika*, 2025.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? *Advances in Neural Information Processing Systems*, 2022.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving Risk Control in Online Learning Settings. *Transactions on Machine Learning Research*, 2023.

Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: A unified review of theory and new challenges. *Bernoulli*, 2023.

Isaac Gibbs and Emmanuel J. Candès. Adaptive Conformal Inference Under Distribution Shift. *Neural Information Processing Systems*, 2021.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 2023.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. *International Conference on Machine Learning*, 2021.

Wouter M Koolen and Peter Grünwald. Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

Rikard Laxhammar and Göran Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 2015.

Rachel Luo, Rohan Sinha, Yixiao Sun, Ali Hindy, Shengjia Zhao, Silvio Savarese, Edward Schmerling, and Marco Pavone. Online distribution shift detection via recency prediction. *International Conference on Robotics and Automation*, 2024.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Teodora Pandeva, Tim Bakker, Christian A. Naesseth, and Patrick Forré. E-Valuating Classifier Two-Sample Tests. *Transactions on Machine Learning Research*, 2024a.

Teodora Pandeva, Patrick Forré, Aaditya Ramdas, and Shubhanshu Shekhar. Deep anytime-valid hypothesis testing. *International Conference on Artificial Intelligence and Statistics*, 2024b.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. *Uncertainty in Artificial Intelligence*, 2021.

Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. *International Conference on Learning Representations*, 2022.

Aleksandr Podkopaev and Aaditya Ramdas. Sequential Predictive Two-Sample and Independence Testing. *Advances in Neural Information Processing Systems*, 2023.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal Validity Guarantees Exist for Any Data Distribution (and How to Find Them). *International Conference on Machine Learning*, 2024.

Aaditya Ramdas and Ruodu Wang. Hypothesis Testing with E-values. *arXiv Preprint (arXiv:2410.23614)*, 2024.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 2023.

Elan Rosenfeld and Saurabh Garg. (Almost) Provable Error Bounds Under Distribution Shift via Disagreement Discrepancy. *Advances in Neural Information Processing Systems*, 2023.

Aytijhya Saha and Aaditya Ramdas. Testing exchangeability by pairwise betting. *International Conference on Artificial Intelligence and Statistics*, 2024.

Mona Schirmer, Dan Zhang, and Eric Nalisnick. Test-time Adaptation with State-Space Models. *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024.

Shubhanshu Shekhar and Aaditya Ramdas. Nonparametric Two-Sample Testing by Betting. *IEEE Transactions on Information Theory*, 2021.

Shubhanshu Shekhar and Aaditya Ramdas. On the near-optimality of betting confidence sets for bounded means. *arXiv Preprint (arXiv:2310.01547)*, 2023a.

Shubhanshu Shekhar and Aaditya Ramdas. Sequential Changepoint Detection via Backward Confidence Sequences. *International Conference on Machine Learning*, 2023b.

Shubhanshu Shekhar and Aaditya Ramdas. Reducing sequential change detection to sequential estimation. *International Conference on Machine Learning*, 2024.

Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. E-detectors: A Nonparametric Framework for Sequential Change Detection. *The New England Journal of Statistics in Data Science*, 2023.

Sophia Huiwen Sun, Abishek Sankararaman, and Balakrishnan Murali Narayanaswamy. Online Adaptive Anomaly Thresholding with Confidence Sequences. *International Conference on Machine Learning*, 2024.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift. *Advances in Neural Information Processing Systems*, 2019.

Wouter AC van Amsterdam, Nan van Geloven, Jesse H Krijthe, Rajesh Ranganath, and Giovanni Ciná. When accurate prediction models yield harmful self-fulfilling prophecies. *Patterns*, 2025.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars Paris, 1939.

Harit Vishwakarma, Heguang Lin, and Ramya Korlakai Vinayak. Taming False Positives in Out-of-Distribution Detection with Human Feedback. *International Conference on Artificial Intelligence and Statistics*, 2024.

Denis Volkhonskiy, Evgeny Burnaev, Ilia Nouretdinov, Alexander Gammerman, and Vladimir Vovk. Inductive conformal martingales for change-point detection. *Conformal and Probabilistic Prediction and Applications*, 2017.

Vladimir Vovk. Testing Randomness Online. *Statistical Science*, 2021.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: Conformal test martingales for change-point detection. *Conformal and Probabilistic Prediction and Applications*, 2021.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. *International Conference on Artificial Intelligence and Statistics*, 2024.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. *Neural Information Processing Systems*, 2024.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-Time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 2022.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. *International Conference on Machine Learning*, 2022.

Matteo Zecchin and Osvaldo Simeone. Adaptive Learn-then-Test: Statistically Valid and Efficient Hyperparameter Selection. *International Conference on Machine Learning*, 2025.

On Continuous Monitoring of Risk Violations under Unknown Shift
— Supplementary Material —

CONTENTS

1 Introduction **1**

2 Risk and Problem Formulation **2**

3 Risk Monitoring as Sequential Testing By Betting **3**

3.1 Relation to Other Approaches 4

4 Theoretical Analysis **5**

5 Related Work **6**

6 Empirical Results **6**

6.1 Monitoring the Total Error Rate for Outlier Detection 7

6.2 Monitoring the Miscoverage Rate for Set Prediction 8

7 Discussion **9**

A Mathematical Details **14**

A.1 Definitions and Terminology 14

A.2 Details on the Sum-Process 14

A.3 Risk Control under the *i.i.d.* Data Stream Setting 14

A.4 Proofs 15

B Additional Related Work **18**

C Additional Experimental Design **19**

D Additional Experimental Results **21**

A MATHEMATICAL DETAILS

We provide relevant mathematical details to complement the main text, including *(i)* on the terminology of (super)martingales and the associated measure-theoretic objects, *(ii)* a more detailed description of the summation wealth process, *(iii)* insights on risk control under the *i.i.d* data stream setting, and finally *(iv)* formal proofs for our main theoretical statements.

A.1 DEFINITIONS AND TERMINOLOGY

Given a sequence of random variables $\mathbf{u}^t = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_t)$, we denote the smallest σ-field generated by $\mathbf{u}^t$ as $\mathcal{F}_t = \sigma(\mathbf{u}^t)$. The sequence of random variables then lead to the filtration $\mathcal{F} = (\mathcal{F}_t)_{t=0}^{\infty}$ defined as the increasing sequence of generated σ-fields $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$, where $\mathcal{F}_0$ is the trivial σ-field. A sequence of random variables $(M_t)_{t=0}^{\infty}$ is called a *martingale* if it is adapted to the filtration $\mathcal{F}$, *i.e.* each M_t is $\mathcal{F}_t$ measurable, each M_t is integrable, and satisfies $\mathbb{E}\left[M_t \mid \mathcal{F}_{t-1}\right] = M_{t-1}$. If this equality is replaced with $\leq$, then we call $(M_t)_{t=0}^{\infty}$ a *supermartingale*. Furthermore, we define a sequence $(\lambda_t)_{t=0}^{\infty}$ as a *predictable* sequence if λ_t is $\mathcal{F}_{t-1}$ measurable, meaning λ_t can only depend on the past information up to the time step $t-1$. Finally, we define a random variable $\tau : \Omega \to \mathbb{N} \cup \{\infty\}$ to be a *stopping time* with respect to the filtration $\mathcal{F}$ if, for every $t \geq 0$, the event $\{\tau \leq t\}$ belongs to the sigma-algebra $\mathcal{F}_t$, *i.e.*, $\{\tau \leq t\} \in \mathcal{F}_t$. This condition ensures that the decision to stop at time t can be made based only on the information available up to time t, meaning τ does not 'see into the future'. We also make use of the following martingale concentration inequality in our results:

Lemma A.1 (Azuma-Hoeffding Inequality)**.** *Let* $(\mathrm{v}_i)_{i=1}^{t}$ *be a martingale difference sequence adapted to a filtration* $(\mathcal{F}_i)_{i=0}^{t}$*, meaning:* $\mathbb{E}[\mathrm{v}_i|\mathcal{F}_{i-1}] = 0,\ \forall i$*. Suppose there exist constants* c_i *such that for all* i*,* $|\mathrm{v}_i| \leq c_i$ *almost surely. Then for any* $\eta > 0$ *we have that*

$$\mathbb{P}\left(\left|\sum_{i=1}^{t} \mathrm{v}_i\right| \geq \eta\right) \leq 2\exp\left(-\frac{\eta^2}{2\sum_{i=1}^{t} c_i^2}\right).$$

There also exists one-sided version of the above inequality as follows:

$$\mathbb{P}\left(\sum_{i=1}^{t} \mathrm{v}_i \geq \eta\right) \leq \exp\left(-\frac{\eta^2}{2\sum_{i=1}^{t} c_i^2}\right) \quad \text{and similarly} \quad \mathbb{P}\left(\sum_{i=1}^{t} \mathrm{v}_i \leq -\eta\right) \leq \exp\left(-\frac{\eta^2}{2\sum_{i=1}^{t} c_i^2}\right).$$

A.2 DETAILS ON THE SUM-PROCESS

As stated in § 3 the summation wealth process is given as $M_t(\psi) = \sum_{i=1}^{t} \lambda_i (\mathrm{z}_i - \epsilon)$, with $(\lambda_t)_{t\in\mathcal{T}}$ being the predictable betting rate. It is clear under the null $H_0(\psi)$ (Eq. 3) this forms a supermartingale, and hence does not grow. Furthermore, it is easy to see that $M_t(\psi)$ is a supermartingale sequence if $\mathbb{E}_{P_t}\left[\mathrm{z}_t(\psi) \mid \mathcal{F}_{t-1}\right] < \epsilon,\ \forall t \in \mathcal{T}$. Thus, we obtain an *if and only if* characterization of the risk control condition, and hence if one can deduce that $M_t(\psi)$ is not a supermartingale, then this gives the evidence that the desired risk control assurance is violated. If $M_t(\psi)$ was a martingale sequence, *i.e.* when $\mathbb{E}_{P_t}\left[\mathrm{z}_t(\psi) \mid \mathcal{F}_{t-1}\right] = \epsilon$, we may apply the one-sided Azuma-Hoeffding inequality to argue that the martingale sequence does not grow beyond a certain limit with high-probability. However, considering a sequence $(\tilde{\mathrm{z}}_t)_{t\in\mathcal{T}}$ such that $\mathbb{E}_{P_t}\left[\tilde{\mathrm{z}}_t \mid \mathcal{F}_{t-1}\right] = \epsilon$, it can be argued that $\tilde{M}_t(\psi) = \sum_{i=1}^{t} \lambda_i (\tilde{\mathrm{z}}_t - \epsilon) \geq \sum_{i=1}^{t} \lambda_i (\mathrm{z}_t - \epsilon)$, hence the one-sided Azuma-Hoeffding bound applied to $\tilde{M}_t(\psi)$ also extends to $M_t(\psi)$. Thus we can argue that

$$\mathbb{P}\left(\sum_{i=1}^{t} \lambda_i (\mathrm{z}_i - \epsilon) \geq \eta\right) \leq \exp\left(-\frac{\eta^2}{2t}\right),$$

where the bounded assumption $\mathrm{z} \in [0,1]$ and $\epsilon \in [0,1)$ result in boundedness of the difference sequence. Next, we can choose the threshold $\eta = \sqrt{2t \log \frac{1}{\delta}}$ and argue that $\mathbb{P}\left(M_t(\psi) \geq \eta\right) < \delta$. If $M_t(\psi)$ does indeed grow above η, then one can raise the alarm while retaining a false alarm control guarantee, as under the null $M_t(\psi)$ will not grow beyond η with probability of at least $1-\delta$. We further note that with high-probability, $M_t(\psi)$ will remain bounded as $\mathcal{O}(\sqrt{t})$.

A.3 RISK CONTROL UNDER THE *I.I.D.* DATA STREAM SETTING

Consider the simpler setting where the test stream originates *i.i.d* from a non-shifting test distribution as $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{t\in\mathcal{T}} \sim P_0$. The risk quantity to monitor from Eq. 3 directly simplifies due to independence as $\mathbb{E}_{P_t}\left[\mathrm{z}_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{P_t}\left[\mathrm{z}_t\right] = \mathcal{R}_t(\psi)$, and since $P_0 = P_t \; \forall t \in \mathcal{T}$ is static we have $\mathcal{R}_t(\psi) = \mathcal{R}_0(\psi)$ as a time-independent risk. We can now conveniently reverse our hypotheses under the sequential testing framework to exploit the fact that P_0 is static, and significant discoveries will thus hold even under future observations. That is, we can test for risk control directly by the hypothesis pair

$$H_0(\psi) : \exists t \in \mathcal{T} : \mathcal{R}_0(\psi) > \epsilon, \qquad H_1(\psi) : \mathcal{R}_0(\psi) \leq \epsilon \; \forall t \in \mathcal{T}, \tag{6}$$

and use the following (reversed) wealth process and ψ-CS construction:

$$M_t(\psi) = \prod_{i=1}^{t} \left(1 + \lambda_i \left(\epsilon - \mathrm{z}_i\right)\right) \quad \text{and} \quad C_t^{\psi} = \{\psi \in \Psi : M_t(\psi) \geq 1/\delta\}. \tag{7}$$

Observe how once sufficient evidence is collected to support that the risk associated with a particular candidate ψ does not exceed the tolerated risk level ϵ, that candidate ψ can be added to C_t^{ψ} safely and indefinitely since the evidence collected remains meaningful under a static P_0. We can then directly leverage the Type-I error control property under the sequential testing framework [Ramdas et al., 2023] (via Ville's Inequality) to state strong *time-uniform* or *anytime-valid* risk control guarantees. Specifically, it follows that for every $\psi \in \Psi$ we have

$$\mathbb{P}_{H_0}(\exists t \in \mathcal{T} : M_t(\psi) \geq 1/\delta) \leq \delta \;\Rightarrow\; \mathbb{P}_{H_0}(\exists t \in \mathcal{T} : \mathcal{R}_0(\psi) \leq \epsilon) \leq \delta \;\Rightarrow\; \mathbb{P}(\forall t \in \mathcal{T} : \mathcal{R}_0(\psi) \leq \epsilon) \geq 1 - \delta. \tag{8}$$

In words, the probability of claiming risk control (H_1) under risk violation (H_0) is upper bounded by δ, whereas under risk control we may perhaps mistakingly claim violation (and thus be overly conservative by excluding the associated ψ) but will not invalidate the risk level ϵ. Thus, the overall probability of risk violation is controlled at level $1 - \delta$, rendering a strong safety assurance. Since $\mathcal{R}_t(\psi)$ is not truly time-dependent neither is C_t^{ψ}, which will initially grow as evidence for each ψ is collected and a decision on inclusion is made, and eventually stabilize. It is then straightforward to also recommend a particular threshold choice if the risk profile is monotonic or in some sense predictable, such as $\hat{\psi}_t := \min C_t^{\psi}$ as the least conservative threshold in case of a monotonically increasing risk. In other words, $\hat{\psi}_t$ will quickly tend to an optimal fixed choice $\hat{\psi}$ after a sufficient number of observations are processed.

A.4 PROOFS

We provide proofs for our main theoretical statements from § 4 below. We first restate each result for self-containment.

PROOF OF LEMMA 4.1.

Lemma 4.1. (Valid wealth process). *For any $\psi \in \Psi$ such that $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] \leq \epsilon \; \forall t \in \mathcal{T}$ satisfies the null, the process $M_t(\psi)$ in Eq. 4 is a valid test supermartingale for the predictable betting rate $\lambda_t \in [0, 1/\epsilon)$.*

Proof. For any $\psi \in \Psi$, we consider the test-statistic of the form $M_t\left(\psi\right) = \prod_{i=1}^{t}\left(1 + \lambda_i\left(\mathrm{z}_i - \epsilon\right)\right)$ where $(\lambda_t)_{t \in \mathcal{T}}$ is a predictable process. We also have that $\mathrm{z}_t \in [0, 1]$ (*i.e.* we consider a bounded loss function; from § 2), and the restriction on $\lambda_t \in [0, 1/\epsilon)$ renders the term $1 + \lambda_t\left(\mathrm{z}_t - \epsilon\right)$ to be non-negative. Furthermore, $M_t\left(\psi\right)$ being adapated to the filtration follows from λ_t being predictable. Integrability of $M_t\left(\psi\right)$ follows from the boundedness assumption. Next we verify the supermartingale condition $\mathbb{E}_{P_t}\left[M_t\left(\psi\right) \mid \mathcal{F}_{t-1}\right] \leq M_{t-1}$. Since conditional on $\mathcal{F}_{t-1}$ randomness only originates from z_t, we have that $\mathbb{E}_{P_t}\left[M_t\left(\psi\right) | \mathcal{F}_{t-1}\right] = M_{t-1}\left(\psi\right) + \lambda_t \cdot M_{t-1}\left(\psi\right) \cdot \mathbb{E}_{P_t}\left[\mathrm{z}_t - \epsilon \mid \mathcal{F}_{t-1}\right] \leq M_{t-1}\left(\psi\right)$, where the last inequality follows from the fact that for a valid ψ we have $\mathbb{E}_{P_t}\left[\mathrm{z}_t - \epsilon \mid \mathcal{F}_{t-1}\right] \leq 0$. Hence, we have shown that $M_t\left(\psi\right)$ is a valid test supermartingale (or wealth process) for ψ. □

It is also easy to verify the converse direction for the following corollary:

Corollary. *$M_t(\psi)$ is a valid test supermartingale if and only if $\mathbb{E}_{P_t}[\mathrm{z}_t(\psi) \mid \mathcal{F}_{t-1}] \leq \epsilon, \quad \forall t \in \mathcal{T}$.*

Remark. So far in our approach, we have not put any restriction on the nature of stream, *i.e.* we can have arbitrary dependence between distributions P_t and $P_{t'}$, $t \neq t'$. Thus, the highlighted approach encompasses general and realistic shift settings in the stream. However, in the scenario where the data stream under shift satisfies an independence assumption, *i.e.* when samples from P_t and $P_{t'}$ are independent, the considered hypothesis pair further simplifies to $H_0\left(\psi\right) = \mathcal{R}_t\left(\psi\right) \leq \epsilon, \; \forall t \in \mathcal{T}, \; H_1\left(\psi\right) : \exists t \in \mathcal{T} : \mathcal{R}_t\left(\psi\right) > \epsilon$ as then $\mathbb{E}_{P_t}\left[\mathrm{z}_t \mid \mathcal{F}_{t-1}\right] = \mathbb{E}_{P_t}\left[\mathrm{z}_t\right]$. Hence, the hypothesis formulation in Eq. 3 is more general.

PROOF OF LEMMA 4.2.

Lemma 4.2. (False alarm guarantee). *For any $\psi \in \Psi$ such that $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] \leq \epsilon\ \forall t \in \mathcal{T}$ satisfies the null, it holds that $\mathbb{P}\left(\exists t \in \mathcal{T} : M_t(\psi) \geq 1/\delta\right) \leq \delta$.*

Proof. The above statement is a direct consequence of Ville's inequality which we state below for completion:

Ville's inequality [Ville, 1939]. Given a non-negative supermartingale sequence $(M_t)_{t\in\mathcal{T}}$ such that $M_0 = 1$, it holds that

$$\mathbb{P}(\exists t \in \mathcal{T}\ :\ M_t \geq 1/\delta) \leq \frac{\mathbb{E}[M_0]}{1/\delta} = \delta.$$

Similar to the Azuma-Hoeffding inequality (§ A.2), Ville's inequality gives probabilistic control on the growth of the supermartingale process. The false alarm guarantee then trivially follows from an interpretation of the obtained Type-I error control, as Lemma 4.1 asserts that $M_t\left(\psi\right)$ is a valid supermartingale for $\psi \in \Psi$ such that $\mathbb{E}_{P_t}[\mathrm{z}_t(\psi) \mid \mathcal{F}_{t-1}] \leq \epsilon, \forall t \in \mathcal{T}$. □

PROOF OF LEMMA 4.3.

Lemma 4.3. (Asymptotic consistency). *For any $\psi \in \Psi$ such that $\mathbb{E}_{P_t}\left[\mathrm{z}_t \mid \mathcal{F}_{t-1}\right] \leq \epsilon$ for finitely many steps $t \in \mathcal{T}$ and $\mathbb{E}_{P_t}\left[\mathrm{z}_t \mid \mathcal{F}_{t-1}\right] > \epsilon$ otherwise, it holds that $\mathbb{P}(\tau(\psi) < \infty) = 1$, where $\tau(\psi)$ denotes the stopping time.*

Proof. This is a simple consequence of the property of *sequential test of power one* [Darling and Robbins, 1968] as stated in the main text. However, we provide a simple proof below based on the growth rate of $M_t\left(\psi\right)$ under the alternative $(H_1\left(\psi\right))$. The proof closely follows the ideas outlined in the consistency results from Pandeva et al. [2024b] (Prop. 4.2 in the paper). For notational clarity, we suppress the dependence on ψ below.

We first note that $\mathbb{P}\{\tau = \infty\} = \mathbb{P}\{\cap_{t\geq 1}\{\tau > t\}\} \leq \mathbb{P}\{\tau > t\}$. Taking the limit, $\mathbb{P}\{\tau = \infty\} \leq \limsup_{t\to\infty} \mathbb{P}\{\tau > t\}$. Next, we will argue thet $\limsup_{t\to\infty} \mathbb{P}\{\tau > t\}$ goes to zero under the alternative *almost surely*. We have the wealth process $M_t = \prod_{i=1}^{t}\left(1 + \lambda_i \cdot \delta_i\right)$, $\delta_i = z_i - \epsilon$. Denoting $v_i = \log(1 + \lambda_i \cdot \delta_i)$, we define $S_t = \log M_t = \sum_{i=1}^{t} v_i$. Furthermore, let us denote $A_i = \mathbb{E}[\mathrm{v}_i \mid \mathcal{F}_{t-1}]$. With this notation in place, we consider the event $\mathbb{P}\{\tau > t\}$, *i.e.* the stopping condition as follows.

The stopping condition. The event $\mathbb{P}\{\tau > t\}$ is the probability that the stopping time is greater than t which from our stopping condition and general monotonicity arguments is $\mathbb{P}\{S_t < \log(1/\delta)\}$. Using our notation from above, we have $S_t = \sum_{i=1}^{t} v_i - A_i + \sum_{i=1}^{t} A_i$ which gives

$$\mathbb{P}\{\tau > t\} = \mathbb{P}\left\{\frac{1}{t}\sum_{i=1}^{t} v_i - A_i + \frac{1}{t}\sum_{i=1}^{t} A_i < \frac{\log(\frac{1}{\delta})}{t}\right\}.$$

Martingale difference sequence. It is clear that the sequence $(v_i - A_i)_{i\in\mathcal{T}}$ is a martingale difference sequence, and following the boundedness assumptions $|v_i - A_i| \leq \lambda_i$. And hence, we can apply the Azuma-Hoeffding's inequality to argue that $\mathbb{P}\left\{|\frac{1}{t}\sum_{i=1}^{t} v_i - A_i| > \frac{\eta}{t}\right\} \leq 2\exp\left(\frac{-\eta^2}{2\sum_{i=1}^{t}\lambda_i^2}\right)$ for some η. Given $\lambda_i \leq \lambda_{\text{max}}$ (leveraging bounded betting rates), then $\sum_{i=1}^{t}\lambda_i^2 = t \cdot \lambda_{\text{max}}^2$. Choosing $\eta = \sqrt{t\log t}$, and defining the event $G_t^c = \left\{|\frac{1}{t}\sum_{i=1}^{t} v_i - A_i| > \frac{\eta}{t}\right\}$ to be an undesirable event where the martingale is overly fluctuating, we may state that $\mathbb{P}(G_t^c) \leq 2\exp\{\frac{-\log t}{2\lambda_{\text{max}}^2}\}$ (a decaying rate in t). Defining G_t as a favourable event where the martingale difference remains controlled, we say taht $G_t = \left\{|\frac{1}{t}\sum_{i=1}^{t} v_i - A_i| \leq \frac{\eta}{t}\right\}$. Then, we can write $\mathbb{P}\{\tau > t\}$ using the law of total probability as below:

$$\begin{aligned}
\mathbb{P}\{\tau > t\} &\leq \mathbb{P}\left(\left\{\frac{1}{t}\sum_{i=1}^{t} A_i < \frac{\log 1/\delta}{t} + |\frac{1}{t}\sum_{i=1}^{t} v_i - A_i|\right\} \cap G_t\right) + \mathbb{P}\{G_t^c\} \\
&\leq \mathbb{P}\left(\left\{\frac{1}{t}\sum_{i=1}^{t} A_i < \frac{\log 1/\delta}{t} + \sqrt{\frac{\log t}{t}}\right\} \cap G_t\right) + \mathbb{P}\{G_t^c\} \\
&\leq \mathbb{P}\left(\left\{\frac{1}{t}\sum_{i=1}^{t} A_i < \frac{\log 1/\delta}{t} + \sqrt{\frac{\log t}{t}}\right\}\right) + \mathbb{P}\{G_t^c\}.
\end{aligned}$$

Next, taking the limit $\mathbb{P}\{\tau = \infty\} \leq \limsup_{t\to\infty} \mathbb{P}\{\tau > t\}$, we can bound the first term in the above expression, *i.e.*

$$\mathbb{P}\{\tau = \infty\} \leq \limsup_{t\to\infty} \mathbb{P}\{\tau > t\} \leq \mathbb{P}\left(\left\{\frac{1}{t}\sum_{i=1}^{t} A_i < \frac{\log 1/\delta}{t} + \sqrt{\frac{\log t}{t}}\right\}\right).$$

The alternative. We are given that $\mathbb{E}[\mathrm{z}_t - \epsilon \mid \mathcal{F}_{t-1}] > 0$ for infintely many steps t, and $\mathbb{E}[\mathrm{z}_t - \epsilon \mid \mathcal{F}_{t-1}] \leq 0$ for finitely many steps t. Denote $\mu = \inf_{t\in\mathcal{T}} \mathbb{E}_{H_1}[\mathrm{z}_t - \epsilon \mid \mathcal{F}_{t-1}] > 0$. We assume that the betting rate λ_i is small, and using the approximation $\log(1+x) \approx x$ we write $A_i = \mathbb{E}[\log(1 + \lambda_i \cdot \delta_i \mid \mathcal{F}_{i-1}] \approx \lambda_i \cdot \mathbb{E}[\mathrm{z}_i - \epsilon \mid \mathcal{F}_{i-1}]$. We further make the approximation that λ_i is not exactly zero. By *Cesàro means*, we have $\liminf_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} A_i \geq \liminf_{i\to\infty} A_i = \lambda\mu > 0$, where we use the definition of μ. Now, for sufficiently large t we have

$$\frac{1}{t}\sum_{i=1}^{t} A_i \gg \frac{\log 1/\delta}{t} + \sqrt{\frac{\log t}{t}},$$

and hence $\frac{1}{t}\sum_{i=1}^{t} A_i$ grows faster than the other two terms shrink, leading to the probability $\limsup \mathbb{P}\{\tau > t\} \to 0$. Therefore, we obtain $\mathbb{P}_{H_1}\{\tau = \infty\} = 0$ and thus $\mathbb{P}_{H_1}\{\tau < \infty\} = 1$. □

DETAILS OF DEFINITION 4.4.

We refer to the relevant works such as Waudby-Smith and Ramdas [2024], Koolen and Grünwald [2022], Shekhar and Ramdas [2023a] on the notion of *growth rate optimality* and related betting rates. Waudby-Smith and Ramdas [2024] also refer to the condition as *growth rate adaptive to the particular alternative* (GRAPA), whereas Koolen and Grünwald [2022] label it the *GROW* criterion.

PROOF OF PROPOSITION 4.5.

Proposition 4.5. (Detection delay bound). *A worst-case detection delay for the hypothesis pair in Eq. 3 and wealth process $M_t(\psi)$ in Eq. 4 is characterized by $(\tau(\psi) - \tau_*(\psi)) \approx \mathcal{O}((\log(1/\delta) + T)/(\lambda \cdot \mu))$, where μ denotes the risk violation intensity and T a shift changepoint.*

Proof. We first provide more clarification on the statement of this result. We consider a simplistic setting of risk violations, *i.e.* for some $\psi \in \Psi$, $\exists T \in \mathcal{T}$ such that $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] \leq \epsilon$ for $t \leq T$ (*i.e.* the risk is within control until time step T), and $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] > \epsilon$ for $t > T$ (*i.e.* the risk gets violated after time step T). Furthemore, we assume $\mathbb{E}_{P_t}[\mathrm{z}_t \mid \mathcal{F}_{t-1}] = \mu + \epsilon, \mu > 0,\ t > T$, *i.e.* we assume that the mean deviates above ϵ with some fixed positive quantity μ. Our setting now closely resembles a changepoint detection scenario. The goal of our detection delay argument is to study when a risk violation alarm will be raised, and ideally we do not want significant detection delays after reaching time period T. Our result will help characterize the flexibility of the methodology we employ to control these delays.

We first consider the simple summation test statistic described in § 3), and once more suppress dependency on ψ for notational clarity from now on. We then have the wealth process

$$M_t = \sum_{i=1}^{t} \lambda_i(\mathrm{z}_i - \epsilon) = \sum_{i=1}^{T} \lambda_i(\mathrm{z}_i - \epsilon) + \sum_{i=T+1}^{t} \lambda_i\left(\mathrm{z}_i - \epsilon\right).$$

Let us denote $\sum_{i=1}^{T} \lambda_i(\mathrm{z}_i - \epsilon)$ to be M_T, *i.e.* $M_t = M_T + \sum_{i=T+1}^{t} \lambda_i\left(\mathrm{z}_i - \epsilon\right)$. For simplicity, we consider a fixed betting rate λ moving forward, and we express the pay-off term $(\mathrm{z}_i - \epsilon)$ as $(\mathrm{z}_i - \mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}]) + (\mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}] - \epsilon)$. Thus,

$$M_t = M_T + \sum_{i=T+1}^{t} \lambda\left(\mathrm{z}_i - \mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}]\right) + \sum_{i=T+1}^{t} \lambda\left(\mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}] - \epsilon\right).$$

By the assumption in our setting, the last term resolves to $\lambda \cdot (t - T) \cdot \mu$, and hence the whole expression becomes

$$M_t = M_T + \lambda\mu(t - T) + \sum_{i=T+1}^{t} \lambda(\mathrm{z}_i - \mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}]).$$

We employ the same decomposition for M_T and obtain the following expression as

$$M_t = \underbrace{\sum_{i=1}^{T} \lambda \left(\mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}] - \epsilon \right)}_{E_T} + \lambda\mu \left(t - T \right) + \underbrace{\sum_{i=1}^{t} \lambda \left(\mathrm{z}_i - \mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}] \right)}_{S_t} .$$

Now, it can be seen that $(\mathrm{z}_i - \mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1})_{i \in \mathcal{T}}$ is a martingale difference sequence, and hence by use of Azuma-Hoeffding's inequality (§ A.1) will be contained. For some pre-specified threshold b, we define the stopping time $\tau = \inf\{t \ : \ M_t \geq b\}$. To argue for the detection delay, we consider the term $(\tau - T)$ and raise an alarm when

$$\begin{aligned} & E_T + \lambda\mu(\tau - T) + S_\tau = b, \\ \Leftrightarrow\ & \lambda\mu(\tau - T) = b - E_T - S_\tau, \\ \Leftrightarrow\ & \tau - T = \frac{b - E_T - S_\tau}{\lambda\mu}. \end{aligned}$$

Following this we can analyse the expectation of the detection delay $(\tau - T)$ as $\mathbb{E}\{\tau - T\} = \frac{b - E_T}{\lambda\mu}$ where $\mathbb{E}[S_\tau] = 0$. Since $E_T < 0$ surely, we have $E_T = 0$ when $\mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}] = \epsilon, \ \forall t \leq T$ (the best case setting). In the worst case, when $\mathbb{E}[\mathrm{z}_i \mid \mathcal{F}_{i-1}] = 0, \ \forall t \leq T$ we obtain $E_T = -\lambda T \epsilon$, leading to a worst-case expected detection delay of $(\tau - T) \approx \mathcal{O}(\frac{b+T}{\lambda\mu})$.

We can also give a high-probability argument using a one-sided Azuma-Hoeffding bound. We then have

$$\mathbb{P}\left(S_t \geq \lambda \cdot \eta\right) \leq \exp\left(\frac{-\eta^2}{2t}\right),$$

and taking $\eta = \sqrt{t}/\lambda$ we get $\mathbb{P}\left(S_t \geq \sqrt{t}\right) \leq \exp\left(-\frac{1}{2\lambda^2}\right)$. Considering the worst case setting, we have $(\tau - T) = \frac{b + \lambda T \epsilon - S_\tau}{\lambda\mu} \leq \frac{b + \lambda T \epsilon - \sqrt{\tau}}{\lambda\mu}$ with high-probability, which further results in $(\tau - T) \approx \mathcal{O}\left(\frac{b + \lambda T \epsilon}{\lambda\mu}\right)$.

We can adopt the exact same arguments for our primary wealth process $M_t = \prod_{i=1}^{t} \left(1 + \lambda_i \left(\mathrm{z}_i - \epsilon\right)\right)$ (Eq. 4) by considering $\log M_t$ and using the approximation $\log(1+x) \approx x$, that is valid for small enough λ. Considering $\log M_t \approx \sum_{i=1}^{t} \lambda \left(\mathrm{z}_i - \epsilon\right)$ this then equates the sum-process considered in our argument above. In this case, b will be replaced with $\log 1/\delta$, giving $(\tau - T) \approx \mathcal{O}\left(\frac{\log(1/\delta) + T}{\lambda\mu}\right)$. Some intituitive insights from this expression are that *(i)* the detection delay is directly proportional to T, the time step at which risk violation occurs—if the risk violations begin later, then a delay arises from overcoming the decay from the initial behavior; and *(ii)* the detection delay is inversely proportional to both the betting rate λ and the intensity of the violations μ. However, we note that our methodology comes with the flexibility to control these delays to some extent by leveraging a smart betting rate design. □

B ADDITIONAL RELATED WORK

Static risk control. The framework of *conformal prediction* constructs set predictors with upper bounds specifically on the miscoverage risk under *i.i.d.* or exchangeable data, with a substantial recent body of literature (see, *e.g.*, Angelopoulos et al. [2023b], Fontana et al. [2023]). The approach has been extended to more general bounded risks by Angelopoulos et al. [2024a] leveraging similar exchangeability arguments. Bates et al. [2021a] use concentration inequalities to provide high-probability assurances for monotonic expectation risks, and Angelopoulos et al. [2025] extend the idea to non-monotonic risks by reframing the task as a non-sequential multiple testing problem. Related in spirit, Angelopoulos et al. [2023a] propose the use of a hold-out unlabelled dataset to provide probability guarantees for confidence intervals on population-level parameters.

Risk control for stream data and under shift. Recent work on conformal prediction includes addressing non-exchangeable data sequences such as time series, *e.g.* by tracking and updating the tolerated miscoverage rate [Gibbs and Candès, 2021, Angelopoulos et al., 2024b, Zaffran et al., 2022] or different weighting schemes [Barber et al., 2023, Guan, 2023]. Particular applications also include outlier detection [Bates et al., 2021b, Laxhammar and Falkman, 2015]. Recent work on data shifts has included covariate shift [Tibshirani et al., 2019], label shift [Podkopaev and Ramdas, 2021] and their abstraction to more general shifts [Prinster et al., 2024]. All of the above work predominantly focuses on the miscoverage risk, with Feldman et al. [2023] being an interesting extension of Gibbs and Candès [2021] to more general bounded risks, discussed in § 3.1. Furthermore, obtainable guarantees are generally asymptotic or finite-sample only under relaxation (*e.g.*, with respect to a permitted coverage deviation from the targeted guarantee).

C ADDITIONAL EXPERIMENTAL DESIGN

Empirical-Bernstein wealth process. We directly adopt the formulation as a test supermartingale or wealth process described in Waudby-Smith and Ramdas [2024] (see Sec. 3.2 and Thm. 2 in their paper), given by

$$M_t^{EB}(\psi) = \prod_{i=1}^{t} \exp\{\lambda_i\,(z_i - \epsilon) - v_i\,\rho(\lambda_i)\} \tag{9}$$

with $v_i = 4\,(z_i - \hat{\mu}_{i-1})^2$ and $\rho(\lambda_i) = 1/4\,(-\log(1-\lambda_i) - \lambda_i)$ for $\lambda_i \in [0, 1)$, and using the suggested *predictable plug-in* betting rate $\lambda_i^{EB} = \min\{\sqrt{\frac{2\,\log(2/\delta)}{\hat{\sigma}_{i-1}^2\; i\,\log(1+i)}}, c\}$. The estimates $\hat{\mu}_{i-1}$ and $\hat{\sigma}_{i-1}^2$ denote the empirical running mean and variance over the observed loss sequence $\{z_1, \ldots, z_{i-1}\}$ up to time $i-1$, thus rendering λ_i^{EB} predictable at every time step. We select $c = 1/2$ as a recommended constant $c \in (0, 1)$, and omit the bias terms of $1/4$ and $1/2$ for $\hat{\mu}_{i-1}$ and $\hat{\sigma}_{i-1}^2$ respectively, which are negligable for sufficiently large streams. Podkopaev and Ramdas [2022] leverage the process in its confidence sequence-equivalent form to estimate bounds on the running risk $R_r(\psi)$ in their problem setting, motivating its inclusion as a baseline.

Choice of betting rate. We refer to Waudby-Smith and Ramdas [2024] on the particular technical details of various betting rate designs, in particular their App. B.2 for GRO and App. B.3 for *approximately GRO*. In essence, a direct optimization of the GRO condition (Definition 4.4) can be achieved by exhaustive root-finding over a fine grid of possible values for $\lambda_t \in [0, 1/\epsilon)$. The approximation to GRO takes an additional Taylor approximation and truncation step to derive a closed-form solution for the root, thus being substantially more efficient. Our final betting rate expression takes this approach for a particular truncation aligned with our problem setting (*i.e.* the permitted range of λ_t). Other approximations to the GRO objective are also possible, such as solving for a lower-bound to the wealth (see their App. B.4 and following).

Batched data stream. Assume we observe more than a single observation at every time t, *i.e.* the data stream with samples $\{(\boldsymbol{x}_{t,b}, \boldsymbol{y}_{t,b})\}_{b=1}^{B} \sim P_t$ for some batch size $B \ll B_*$. Then we can average over the evidence in each batch to obtain a more robust measure of evidence for risk violations as

$$M_t(\psi) = \prod_{i=1}^{t} \frac{1}{B} \sum_{b=1}^{B} (1 + \lambda_i(z_{i,b} - \epsilon)), \tag{10}$$

leading to a reduced variance of the wealth process as well as reducing the detection delay $\tau(\psi) - \tau_*(\psi)$ with respect to the true risk. However this does not necessarily equate lower sampling costs, since the total number of observations is $B \cdot t$.

False alarm rate. For a given experiment run or trial, a threshold candidate ψ raises a false alarm (or is labelled a false positive) if $\mathcal{R}_t(\psi) \leq \epsilon$ but $M_t(\psi) \geq 1/\delta$, thus erroneously claiming violation. Equivalently, we may state for the detection delay that $(\tau(\psi) - \tau_*(\psi)) < 0$ stops prematurely. The false alarm rate is then computed as the fraction of false alarms across R trials, *i.e.*

$$\%FP = \frac{1}{R} \sum_{r=1}^{R} \mathbb{1}[(\tau(\psi) - \tau_*(\psi)) < 0], \tag{11}$$

and compared to the tolerated false alarm rate $\delta \in (0, 1)$. If $\%FP > \delta$ then the Type-I error under $H_0(\psi)$ is uncontrolled, resp. any error control property violated.

Total error rate (TER). In § 6.1 the target risk to monitor is the *total error rate* (TER), accounting for both cases of inlier (false positives, FP) and outlier misclassification (false negatives, FN). That is, we define the true risk $\mathcal{R}_t(\psi) = \mathbb{E}_{P_t}[\mathrm{z}_t]$ with loss variable

$$\mathrm{z}_t = \begin{cases} 1, & \text{if } \mathtt{out}(\mathbf{x}_t) \geq \psi \;\text{ and }\; (\mathbf{x}_t, \mathbf{y}_t) \sim P_{in}, \quad \text{(FP)} \\ 1, & \text{if } \mathtt{out}(\mathbf{x}_t) < \psi \;\text{ and }\; (\mathbf{x}_t, \mathbf{y}_t) \sim P_{out}, \quad \text{(FN)} \\ 0 & \text{else.} \end{cases} \tag{12}$$

The TER is a complex risk quantity that is both non-monotonic across time *and* thresholds, since the FP and FN terms introduce competing objectives in terms of what constitutes a ‘safe’ threshold ψ. Under a stepwise shift with increasing outlier fraction, the FP term initially weighs stronger, motivating a higher threshold choice (since the chance of an inlier mislabelling is reduced). However, as the outlier fraction increases the FN term becomes more relevant, motivating a lower threshold choice (*i.e.* increasing the chance of an outlier label). Therefore, no clear ‘trivial’ safe threshold selection is available except for $\hat{\psi} = 1$ if $\pi_t^{out} = 0$ and $\hat{\psi} = 0$ if $\pi_t^{out} = 1$.

Miscoverage rate (MCR). In § 6.2 the target risk to monitor is the *miscoverage rate*, accounting for set exclusion of the correct label $\boldsymbol{y}_t$. That is, we define the true risk $\mathcal{R}_t(\psi) = \mathbb{E}_{P_t}[\mathrm{z}_t]$ with loss variable $\mathrm{z}_t = \mathbb{1}[\boldsymbol{y}_t \notin \hat{f}_\psi(\mathbf{x}_t)]$. In this case, the MCR is closer to monotonically increasing over time as the natural temporal shift caused by both FMoW and Naval propulsion degrades model performance, and clearly monotonic in the thresholds. For FMoW, an indefinitely valid 'safe' threshold with zero risk is given by $\hat{\psi} = 0$, resulting in a prediction set matching the full label space $\mathcal{Y}$, and thus $\mathrm{z}_t = 0$ at every time step. For Naval propulsion, since $\mathcal{Y} \subseteq [0, 1]$ and in practice $\boldsymbol{y}_t \in [0.95, 1.0]$ we instantiate a fine grid of threshold candidates in the range $\Psi := [0, 0.05]$. Thus for a sufficiently well-trained regressor, $\hat{\psi} = 0.05$ trivially ensures coverage (again returning the full response space) while thresholds towards zero place higher reliance on the prediction's accuracy at the risk of miscoverage. Clearly, such 'trival' threshold solutions are generally impractical, but the MCR's monotonic behaviour renders it a more interpretable quantity and easier to track than the TER in § 6.1.

Functional Map of the World dataset. We consider the *Functional Map of the World* dataset (FMoW) [Christie et al., 2018], a large-scale satellite image dataset with 62 categories of building and land use, collected across various geographic regions and over 16 years (2002 – 2017). We consider the time-dependent partitioning of FMoW proposed by Yao et al. [2022], wherein a natural shift occurs as land use for the *same* satellite image locations, surveyed repeatedly over several years, changes over time. The predictor (a DenseNet-121) is trained on earlier years (2002 – 2012), and we increase the test stream frequency to simulate daily observations by sampling chronologically from data in 2013 – 2017 every 365 time steps (equating each passing year). This induces a (slow) step-wise shift, and we observe that the classifier's predictive accuracy worsens as time progresses, in line with results reported by Yao et al. [2022].

Naval propulsion system dataset. We consider predictive maintenance (or equipment monitoring) data on naval gas turbine behaviour [Cipollini et al., 2018]. This tabular time series consists of $\sim 12\,000$ recordings for various turbine system parameters, and an associated turbine compressor degradation coefficient denoting the compressor's health. Over time this degradation coefficient steadily increases from 0.95 to 1.0, denoting a gradual equipment decay. We supplement the data via jittered resampling of early observations (the first 2000 samples) to enrich the initial 'healthy' compressor state, and train a Random Forest regressor on that data. Expectedly, as the compressor gradually degrades beyond its initial 'healthy' range the predictor fails to extrapolate, resulting in decreased performance in line with a temporal distribution shift.

D ADDITIONAL EXPERIMENTAL RESULTS

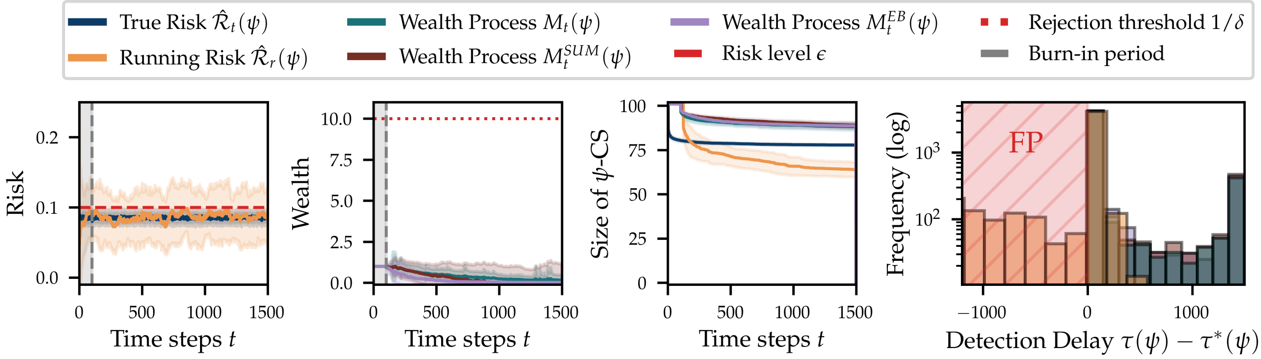

Figure 5: Results for **outlier detection with no shift** (§ 6.1). *From left to right:* Visuals of the steady risk and wealth process behaviour with respective rejection thresholds ϵ and $1/\delta$, for a single threshold candidate (here $\psi = 0.20$); the behaviour of the valid threshold set ψ-CS (Eq. 5), which slightly shrinks by eliminating clearly violating thresholds and then stabilizes, signalling robust detection performance for the *i.i.d.* stream; and the empirical distributions of detection delays $\tau(\psi) - \tau_*(\psi)$ across all $\psi \in \Psi$, including the false alarm region (FP). We also have $B = 1$ and $S = 50$, with results evaluated over $R = 50$ trials (mean and std. deviation).

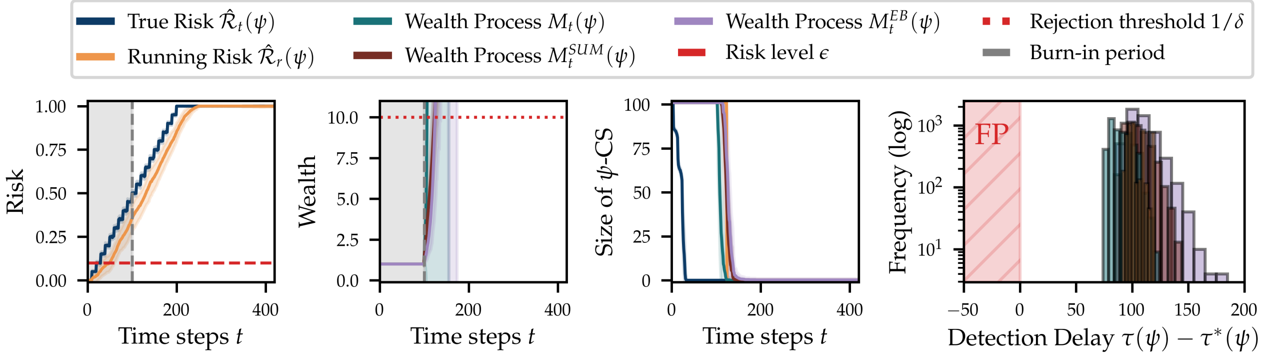

Figure 6: Results for **outlier detection with an immediate shift early on** (§ 6.1). *From left to right:* Visuals of the strongly growing risk and wealth process behaviour with respective rejection thresholds ϵ and $1/\delta$, for a single threshold candidate (here $\psi = 0.90$); the behaviour of the valid threshold set ψ-CS (Eq. 5), which almost immediately collapses to zero signalling a highly unreliable model in need of updating; and the empirical distributions of detection delays $\tau(\psi) - \tau_*(\psi)$ across all $\psi \in \Psi$, including the false alarm region (FP). Since the evidence for risk violation is very clear, no false alarms are incurred by any tracker. We also have $B = 1$ and $S = 50$, with results evaluated over $R = 50$ trials (mean and std. deviation).

Table 1: Results for **outlier detection with a stepwise shift** (§ 6.1). We provide monitoring results of the total error rate for $\epsilon = 0.1, \delta = 0.1$ and different combinations of the two key tracking parameters, sliding window size S and batch size B. The delay quantity $\bar{\tau}(\psi)$ denotes the mean and std. deviation of detection delays $\tau(\psi) - \tau_*(\psi)$ across repeated trials ($R = 50$), whereas $\%FP > 0$ and $\%FP > \delta$ denote the number of thresholds $\psi \in \Psi$ (in %) whose false alarm rate (see Eq. 11) is non-zero or exceeds the desired rate δ, respectively.

Params		**Running Risk** $\hat{\mathcal{R}}_r(\psi)$			**Wealth Process** $M_t(\psi)$			**Wealth Process** $M_t^{SUM}(\psi)$			**Wealth Process** $M_t^{EB}(\psi)$		
S	B	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$
None	1	408 ±201	13.86%	3.96%	563 ±207	0.00%	0.00%	694 ±230	0.00%	0.00%	807 ±334	0.00%	0.00%
None	10	393 ±214	6.93%	2.97%	441 ±210	0.00%	0.00%	495 ±219	0.00%	0.00%	708 ±307	0.00%	0.00%
None	50	393 ±213	0.99%	0.00%	414 ±215	0.00%	0.00%	451 ±220	0.00%	0.00%	655 ±273	0.00%	0.00%
200	1	180 ±84	25.74%	3.96%	355 ±110	0.00%	0.00%	469 ±128	0.00%	0.00%	831 ±337	0.00%	0.00%
200	10	159 ±74	6.93%	2.97%	210 ±80	0.00%	0.00%	257 ±92	0.00%	0.00%	698 ±301	0.00%	0.00%
200	50	156 ±70	0.99%	0.00%	171 ±73	0.00%	0.00%	200 ±75	0.00%	0.00%	604 ±240	0.00%	0.00%
50	1	77 ±83	79.21%	71.29%	349 ±123	0.00%	0.00%	444 ±127	0.00%	0.00%	833 ±333	0.00%	0.00%
50	10	68 ±46	15.84%	3.96%	176 ±83	0.00%	0.00%	214 ±79	0.00%	0.00%	695 ±300	0.00%	0.00%
50	50	74 ±50	0.99%	0.00%	131 ±81	0.00%	0.00%	167 ±77	0.00%	0.00%	598 ±238	0.00%	0.00%
10	1	37 ±79	80.20%	73.27%	394 ±151	0.00%	0.00%	465 ±139	0.00%	0.00%	865 ±326	0.00%	0.00%
10	10	18 ±42	75.25%	27.72%	184 ±93	0.00%	0.00%	204 ±78	0.00%	0.00%	685 ±292	0.00%	0.00%
10	50	41 ±29	5.94%	0.00%	120 ±86	0.00%	0.00%	160 ±78	0.00%	0.00%	598 ±239	0.00%	0.00%

Table 2: Results for **outlier detection with no shift** (§ 6.1). We provide monitoring results of the total error rate for $\epsilon = 0.1, \delta = 0.1$ and different combinations of the two key tracking parameters, sliding window size S and batch size B. The delay quantity $\bar{\tau}(\psi)$ denotes the mean and std. deviation of detection delays $\tau(\psi) - \tau_*(\psi)$ across repeated trials ($R = 50$), whereas $\%FP > 0$ and $\%FP > \delta$ denote the number of thresholds $\psi \in \Psi$ (in %) whose false alarm rate (see Eq. 11) is non-zero or exceeds the desired rate δ, respectively.

Params		**Running Risk** $\hat{\mathcal{R}}_r(\psi)$			**Wealth Process** $M_t(\psi)$			**Wealth Process** $M_t^{SUM}(\psi)$			**Wealth Process** $M_t^{EB}(\psi)$		
S	B	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$
None	1	53 ±277	17.82%	4.95%	176 ±437	0.00%	0.00%	197 ±463	0.00%	0.00%	185 ±446	0.00%	0.00%
None	10	70 ±289	6.93%	0.00%	127 ±389	0.00%	0.00%	137 ±398	0.00%	0.00%	176 ±451	0.00%	0.00%
None	50	84 ±319	2.97%	0.00%	112 ±373	0.00%	0.00%	118 ±376	0.00%	0.00%	180 ±463	0.00%	0.00%
200	1	11 ±233	17.82%	11.88%	175 ±434	0.00%	0.00%	197 ±462	0.00%	0.00%	185 ±445	0.00%	0.00%
200	10	56 ±243	6.93%	0.00%	127 ±389	0.00%	0.00%	137 ±398	0.00%	0.00%	176 ±451	0.00%	0.00%
200	50	80 ±304	2.97%	0.00%	111 ±372	0.00%	0.00%	118 ±376	0.00%	0.00%	180 ±463	0.00%	0.00%
50	1	−102 ±362	27.72%	21.78%	177 ±437	0.00%	0.00%	197 ±461	0.00%	0.00%	185 ±446	0.00%	0.00%
50	10	13 ±107	8.91%	4.95%	127 ±390	0.00%	0.00%	135 ±395	0.00%	0.00%	177 ±452	0.00%	0.00%
50	50	59 ±245	2.97%	0.00%	112 ±371	0.00%	0.00%	118 ±375	0.00%	0.00%	180 ±463	0.00%	0.00%
10	1	−226 ±478	37.62%	33.66%	183 ±447	0.00%	0.00%	198 ±461	0.00%	0.00%	199 ±470	0.00%	0.00%
10	10	−126 ±363	19.80%	17.82%	133 ±402	0.00%	0.00%	129 ±382	0.00%	0.00%	178 ±455	0.00%	0.00%
10	50	12 ±73	7.92%	2.97%	111 ±370	0.00%	0.00%	118 ±373	0.00%	0.00%	180 ±463	0.00%	0.00%

Table 3: Results for **outlier detection with an immediate shift early on** (§ 6.1). We provide monitoring results of the total error rate for $\epsilon = 0.1, \delta = 0.1$ and different combinations of the two key tracking parameters, sliding window size S and batch size B. The delay quantity $\bar{\tau}(\psi)$ denotes the mean and std. deviation of detection delays $\tau(\psi) - \tau_*(\psi)$ across repeated trials ($R = 50$), whereas $\%FP > 0$ and $\%FP > \delta$ denote the number of thresholds $\psi \in \Psi$ (in %) whose false alarm rate (see Eq. 11) is non-zero or exceeds the desired rate δ, respectively.

Params		**Running Risk** $\hat{\mathcal{R}}_r(\psi)$			**Wealth Process** $M_t(\psi)$			**Wealth Process** $M_t^{SUM}(\psi)$			**Wealth Process** $M_t^{EB}(\psi)$		
S	B	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$
None	1	104 ±9	0.00%	0.00%	91 ±9	0.00%	0.00%	108 ±9	0.00%	0.00%	110 ±79	0.00%	0.00%
None	10	38 ±9	0.00%	0.00%	28 ±10	0.00%	0.00%	40 ±12	0.00%	0.00%	49 ±11	0.00%	0.00%
None	50	36 ±9	0.00%	0.00%	23 ±11	0.00%	0.00%	32 ±11	0.00%	0.00%	50 ±13	0.00%	0.00%
200	1	104 ±9	0.00%	0.00%	91 ±9	0.00%	0.00%	108 ±9	0.00%	0.00%	110 ±79	0.00%	0.00%
200	10	38 ±9	0.00%	0.00%	28 ±10	0.00%	0.00%	40 ±12	0.00%	0.00%	49 ±11	0.00%	0.00%
200	50	36 ±9	0.00%	0.00%	23 ±11	0.00%	0.00%	32 ±11	0.00%	0.00%	50 ±13	0.00%	0.00%
50	1	104 ±9	0.00%	0.00%	90 ±9	0.00%	0.00%	105 ±40	0.00%	0.00%	110 ±57	0.00%	0.00%
50	10	38 ±9	0.00%	0.00%	27 ±9	0.00%	0.00%	37 ±9	0.00%	0.00%	48 ±11	0.00%	0.00%
50	50	36 ±9	0.00%	0.00%	22 ±10	0.00%	0.00%	30 ±9	0.00%	0.00%	50 ±13	0.00%	0.00%
10	1	104 ±9	0.00%	0.00%	90 ±10	0.00%	0.00%	110 ±88	0.00%	0.00%	110 ±57	0.00%	0.00%
10	10	26 ±5	0.00%	0.00%	18 ±6	0.00%	0.00%	27 ±6	0.00%	0.00%	46 ±11	0.00%	0.00%
10	50	25 ±3	0.00%	0.00%	11 ±4	0.00%	0.00%	22 ±5	0.00%	0.00%	48 ±12	0.00%	0.00%

Table 4: Results for **set prediction with a temporal shift on FMoW** (§ 6.2). We provide monitoring results of the miscoverage rate for $\epsilon = 0.1, \delta = 0.1$ and different combinations of the two key tracking parameters, sliding window size S and batch size B. The delay quantity $\bar{\tau}(\psi)$ denotes the mean and std. deviation of detection delays $\tau(\psi) - \tau_*(\psi)$ across repeated trials ($R = 50$), whereas $\%FP > 0$ and $\%FP > \delta$ denote the number of thresholds $\psi \in \Psi$ (in %) whose false alarm rate (see Eq. 11) is non-zero or exceeds the desired rate δ, respectively.

Params		**Running Risk** $\hat{\mathcal{R}}_r(\psi)$			**Wealth Process** $M_t(\psi)$			**Wealth Process** $M_t^{SUM}(\psi)$			**Wealth Process** $M_t^{EB}(\psi)$		
S	B	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$
None	1	166 ±192	3.96%	1.98%	262 ±313	0.00%	0.00%	340 ±359	0.00%	0.00%	324 ±351	0.00%	0.00%
None	10	82 ±196	1.98%	0.00%	103 ±263	0.00%	0.00%	126 ±281	0.00%	0.00%	213 ±380	0.00%	0.00%
None	50	75 ±200	0.00%	0.00%	75 ±239	0.00%	0.00%	91 ±249	0.00%	0.00%	201 ±381	0.00%	0.00%
365	1	137 ±121	3.96%	2.97%	239 ±280	0.00%	0.00%	314 ±322	0.00%	0.00%	329 ±364	0.00%	0.00%
365	10	58 ±132	1.98%	0.00%	81 ±225	0.00%	0.00%	104 ±243	0.00%	0.00%	216 ±388	0.00%	0.00%
365	50	52 ±140	0.00%	0.00%	52 ±190	0.00%	0.00%	68 ±203	0.00%	0.00%	200 ±381	0.00%	0.00%
50	1	101 ±133	6.93%	5.94%	239 ±284	0.00%	0.00%	309 ±312	0.00%	0.00%	334 ±380	0.00%	0.00%
50	10	35 ±54	3.96%	1.98%	75 ±223	0.00%	0.00%	94 ±230	0.00%	0.00%	217 ±391	0.00%	0.00%
50	50	37 ±91	0.00%	0.00%	45 ±182	0.00%	0.00%	60 ±191	0.00%	0.00%	196 ±370	0.00%	0.00%
10	1	89 ±180	6.93%	6.93%	258 ±296	0.00%	0.00%	327 ±314	0.00%	0.00%	384 ±469	0.00%	0.00%
10	10	9 ±126	4.95%	4.95%	83 ±244	0.00%	0.00%	90 ±222	0.00%	0.00%	218 ±391	0.00%	0.00%
10	50	25 ±28	1.98%	1.98%	47 ±191	0.00%	0.00%	61 ±195	0.00%	0.00%	195 ±368	0.00%	0.00%

Table 5: Results for **set prediction with a temporal shift on Naval propulsion** (§ 6.2). We provide monitoring results of the miscoverage rate for $\epsilon = 0.1, \delta = 0.1$ and different combinations of the two key tracking parameters, sliding window size S and batch size B. The delay quantity $\bar{\tau}(\psi)$ denotes the mean and std. deviation of detection delays $\tau(\psi) - \tau_*(\psi)$ across repeated trials ($R = 50$), whereas $\%FP > 0$ and $\%FP > \delta$ denote the number of thresholds $\psi \in \Psi$ (in %) whose false alarm rate is non-zero or exceeds the desired rate δ, respectively.

Params		**Running Risk** $\hat{\mathcal{R}}_r(\psi)$			**Wealth Process** $M_t(\psi)$			**Wealth Process** $M_t^{SUM}(\psi)$			**Wealth Process** $M_t^{EB}(\psi)$		
S	B	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$	Delay $\bar{\tau}(\psi)$	%FP > 0	%FP $> \delta$
None	1	1373 ±952	0.00%	0.00%	1416 ±957	0.00%	0.00%	1469 ±971	0.00%	0.00%	3527 ±2193	0.00%	0.00%
None	10	1364 ±963	0.00%	0.00%	1401 ±969	0.00%	0.00%	1450 ±983	0.00%	0.00%	3216 ±1969	0.00%	0.00%
None	50	1361 ±965	0.00%	0.00%	1400 ±970	0.00%	0.00%	1449 ±984	0.00%	0.00%	3169 ±1946	0.00%	0.00%
200	1	293 ±453	0.00%	0.00%	419 ±540	0.00%	0.00%	492 ±660	0.00%	0.00%	1178 ±1076	0.00%	0.00%
200	10	305 ±491	0.00%	0.00%	349 ±503	0.00%	0.00%	375 ±507	0.00%	0.00%	1104 ±992	0.00%	0.00%
200	50	308 ±492	0.00%	0.00%	325 ±498	0.00%	0.00%	346 ±501	0.00%	0.00%	1074 ±990	0.00%	0.00%
50	1	146 ±297	54.46%	8.91%	394 ±514	0.00%	0.00%	438 ±537	0.00%	0.00%	1077 ±1059	0.00%	0.00%
50	10	204 ±430	0.00%	0.00%	318 ±505	0.00%	0.00%	342 ±507	0.00%	0.00%	1011 ±973	0.00%	0.00%
50	50	232 ±482	0.00%	0.00%	284 ±498	0.00%	0.00%	312 ±501	0.00%	0.00%	915 ±866	0.00%	0.00%
10	1	58 ±147	81.19%	60.40%	416 ±544	0.00%	0.00%	451 ±566	0.00%	0.00%	1024 ±1055	0.00%	0.00%
10	10	99 ±186	1.98%	0.00%	330 ±511	0.00%	0.00%	336 ±507	0.00%	0.00%	985 ±968	0.00%	0.00%
10	50	191 ±463	0.00%	0.00%	284 ±500	0.00%	0.00%	311 ±502	0.00%	0.00%	888 ±861	0.00%	0.00%