NONPARAMETRIC SEQUENTIAL GOODNESS-OF-FIT TESTING UNDER HETEROGENEITY

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider sequential goodness-of-fit testing when the data generating distribution is an unknown mixture of several reference distributions, with heterogeneity arising through variation in the mixing proportions. In many contemporary applications, this reflects latent or structurally distinct subpopulations in the data that occur at different rates while their underlying distributions remain unchanged. Such changes in composition can substantially alter the overall data distribution and cause standard sequential procedures to spuriously reject. We develop a nonparametric sequential testing framework for detecting distributional departures in the presence of such heterogeneity. Our framework is based on a composite null that allows the data generating distribution to arise from the same collection of reference distributions but with possibly different mixing weights. Motivated by the principle of testing by betting, we propose Sequential Goodness-of-Fit Testing via Betting (SeqGoF), which uses kernel mean embeddings to quantify departures from this composite null and constructs a sequential betting strategy against it, without imposing a parametric model for the alternative. We show that SeqGoF has anytime-valid level-α control under arbitrary stopping and almost-sure power one against every fixed alternative outside the null, and we study its first-order stopping-time asymptotics. Numerical experiments and an application to hard-drive SMART telemetry demonstrate the strong finite-sample performance of SeqGoF.

1 INTRODUCTION

Sequential goodness-of-fit testing is designed to detect when an evolving data stream departs from a reference distribution. A central difficulty in practice is that the reference distribution itself may be heterogeneous: observations under the null hypothesis can arise from several distinct subpopulations, with the overall distribution determined by their mixing proportions. Changes in these proportions can substantially alter the marginal distribution even when the subpopulation-specific distributions remain unchanged. Standard sequential tests, such as Zhou & Liu (2024), are built for a single homogeneous reference, and may therefore signal a distributional change simply because the population composition has shifted.

This issue arises naturally in several applications. For instance, online behavioral data may contain several distinct modes of user activity, with changes in their relative prevalence altering the aggregate behavioral distribution even when the underlying modes themselves remain unchanged (Banerjee et al., 2025). Similarly, in single-cell studies, biological samples often contain multiple latent cell subpopulations whose expression distributions remain stable while their relative abundances vary, for example, when infection affects different subpopulations at different rates (Cavrois et al., 2017). In both settings, treating the reference distribution as homogeneous can confound compositional changes with genuine departures from the underlying reference family.

We address this problem by defining the null family $\mathcal{P}_0$ as all mixtures of a fixed collection of reference distributions with unknown mixing weights, yielding a composite null hypothesis (equation 1); formal definitions are given in Section 2. Motivated by the principle of testing by betting (Shafer, 2021), we propose SeqGoF (Section 3), which uses kernel mean embeddings to measure departures from this composite null and constructs a sequential betting strategy against it. In the RKHS, $\mathcal{P}_0$ is represented by the convex hull of the reference mean embeddings. At each time point, SeqGoF projects the empirical kernel mean onto this convex hull and uses the projection residual as a witness of departure

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

from the null family. This witness is then converted into a predictable betting score, producing an e-process that can be monitored continuously under arbitrary stopping without imposing a parametric model on the alternative.

These ingredients lead to both strong sequential guarantees and a practical implementation. Theoretically, we show that SeqGoF has anytime-valid level-α control under arbitrary stopping (Theorem 1), and almost-sure power one against every fixed alternative outside the null (Theorem 2), and we derive its first-order stopping-time asymptotics (Theorem 3). For characteristic kernels (Sriperumbudur et al., 2011), the kernel distance to the mixture family is positive if and only if the data-generating distribution lies outside the null, so the power guarantee applies to every fixed alternative beyond the composite family. Computationally, we develop a practical and efficient implementation based on a low-dimensional quadratic program and a one-dimensional betting optimization, which remains applicable when the reference distributions are available only through samples (Algorithm 1). Empirically, experiments with Gaussian and non-Gaussian reference distributions demonstrate accurate type-I error control and increasing power (Section 4), while an application to hard-drive SMART telemetry shows that SeqGoF accommodates heterogeneity among healthy drives while rapidly detecting near-failure behavior (Section 5).

Related work. Our work connects several strands of literature. Classical kernel two-sample and goodness-of-fit tests use RKHS discrepancies to test distributional equality or agreement with a specified reference distribution (Gretton et al., 2012; Chwialkowski et al., 2016; Liu et al., 2016). More recently, e-processes and testing-by-betting have provided a general framework for anytime-valid sequential inference (Ramdas et al., 2023; Ramdas & Wang, 2024; Grünwald et al., 2024), with applications to point nulls, moment constraints, exchangeability, two-sample testing, and independence testing (Shekhar & Ramdas, 2023; Podkopaev & Ramdas, 2023). Most closely related, Zhou & Liu (2024) develop a kernel-based sequential goodness-of-fit test using betting, but focus on a simple null defined by a single reference distribution. In contrast, we consider a composite null given by the convex hull of several reference distributions with unknown mixing weights, requiring both the witness construction and betting calibration to remain valid uniformly over the entire null family. A closely related framework is the Donsker-Varadhan (DV) e-process of Shekhar (2025), which develops anytime-valid tests for general convex nulls by optimizing over a class of test functions and calibrating against the null. Under suitable conditions, their procedure is first-order optimal. SeqGoF takes a different route tailored to mixture nulls: the witness is chosen geometrically by projecting the empirical kernel mean onto the convex hull of the reference mean embeddings, while null calibration reduces to component-wise expectations. This yields a substantially simpler implementation based on a low-dimensional quadratic program and a one-dimensional betting optimization. See Appendix B for details.

Composite nulls of the form considered here have been studied in batch two-sample testing under latent heterogeneity, including nearest-neighbor and graph-based methods (Banerjee et al., 2020; 2025; Karmakar et al., 2015), but not in a sequential setting. Related convex-hull ideas also arise in kernel mixture-proportion estimation (Blanchard et al., 2010; Ramaswamy et al., 2016), where the mixture model is typically assumed and the goal is to estimate the mixing weights. In contrast, we test whether the data-generating distribution belongs to the convex hull of the reference components while maintaining type-I error control under continuous monitoring. To our knowledge, this composite mixture null has not previously been studied in an anytime-valid sequential framework.

Main contributions. Our main contributions are fourfold. First, we formulate sequential goodness-of-fit testing under heterogeneity as testing a composite null defined by the convex hull of multiple reference distributions with unknown mixing weights. Second, we develop a new betting construction for this setting: SeqGoF projects the empirical kernel mean onto the convex hull of the reference embeddings, uses the projection residual as an RKHS witness, and calibrates a predictable bet uniformly over the null family. This extends kernel betting from a single reference distribution to composite mixture nulls. Third, we establish anytime-valid level-α control under arbitrary stopping, almost-sure power one against every fixed alternative outside the null, an explicit lower bound on the wealth growth rate, and first-order stopping-time asymptotics. Finally, we provide an efficient implementation of SeqGoF and show empirically that it controls type-I error across heterogeneous null compositions while rapidly detecting departures in both synthetic experiments and hard-drive SMART telemetry.

2 PROBLEM FORMULATION

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space and let $F_1, \ldots, F_K$ be K known probability distributions on $\mathcal{X}$ ("components"). Write $\Delta_K = \{w \in [0,1]^K : \sum_k w_k = 1\}$ and define the mixture null class

$$\mathcal{P}_0 = \Big\{P_w := \sum_{k=1}^{K} w_k F_k \;:\; w \in \Delta_K\Big\} = \mathrm{conv}\{F_1, \ldots, F_K\} \subset \mathcal{P}(\mathcal{X}). \tag{1}$$

We observe a stream $Y_1, Y_2, \ldots \overset{\text{i.i.d.}}{\sim} P$ for an unknown $P \in \mathcal{P}(\mathcal{X})$, and wish to test

$$H_0 : P \in \mathcal{P}_0 \text{ versus } H_1 : P \notin \mathcal{P}_0. \tag{2}$$

The null is composite and convex, while the alternative is nonparametric. Let $\mathcal{F}_t = \sigma(Y_1, \ldots, Y_t)$ denote the data filtration.

Our goal is to design a sequential test for equation 2. Unlike a fixed-sample test, a sequential test does not specify in advance how many observations will be collected. Instead, after each observation Y_t, the procedure may either continue sampling or stop and reject the null. The time at which rejection occurs is represented by a stopping time $\tau \in \mathbb{N} \cup \{\infty\}$ with respect to $(\mathcal{F}_t)_{t\geq 0}$; that is, the event $\{\tau \leq t\}$ is $\mathcal{F}_t$-measurable for every t. Thus, the decision whether to stop by time t can depend only on the observations available up to that time. We interpret $\tau < \infty$ as rejection of H_0, while $\tau = \infty$ means that the procedure never accumulates sufficient evidence to reject.

Given a significance level $\alpha \in (0,1)$, we seek a stopping time τ_α that controls the probability of ever rejecting under every distribution in the composite null: $\sup_{P_0 \in \mathcal{P}_0} \Pr_{P_0}(\tau_\alpha < \infty) \leq \alpha$. At the same time, under an alternative $P \notin \mathcal{P}_0$, we would like the procedure to eventually reject with probability one, $\Pr_P(\tau_\alpha < \infty) = 1$, and, beyond this qualitative power-one guarantee, to stop as quickly as possible.

3 SEQGOF: SEQUENTIAL GOODNESS-OF-FIT TESTING VIA BETTING

3.1 E-PROCESSES AND THE BETTING FORMULATION

A key challenge in sequential testing is to preserve type-I error control even when the data are inspected repeatedly and the stopping time is chosen adaptively. A general way to achieve this is through e-processes. In fact, Ramdas et al. (2020) show that e-processes, play a fundamental role in anytime-valid sequential inference.

Formally, a nonnegative adapted process $(M_t)_{t\geq 0}$ with $M_0 = 1$ is an e-process for $\mathcal{P}_0$ if, for every $P_0 \in \mathcal{P}_0$, there exists a nonnegative P_0-supermartingale $(\widetilde{M}_t^{P_0})_{t\geq 0}$, with $\widetilde{M}_0^{P_0} = 1$, such that

$$M_t \leq \widetilde{M}_t^{P_0} \qquad P_0\text{-almost surely for every } t \geq 0.$$

Ville's inequality (Ville, 1939) then gives

$$\sup_{P_0 \in \mathcal{P}_0} \mathbb{P}_{P_0}\left(\sup_{t\geq 0} M_t \geq \frac{1}{\alpha}\right) \leq \alpha.$$

Consequently, $\tau_\alpha := \inf\left\{t \geq 1 : M_t \geq \frac{1}{\alpha}\right\}$, $\inf \varnothing := \infty$, defines a level-$\alpha$ sequential test:

$$\sup_{P_0 \in \mathcal{P}_0} \mathbb{P}_{P_0}(\tau_\alpha < \infty) \leq \alpha.$$

Thus, constructing a valid sequential test reduces to constructing an e-process that grows when the observations provide evidence against the null. In this paper, we construct such an e-process through a sequence of multiplicative bets. Starting from $M_0 = 1$, write $M_t = M_{t-1} e_t$, where $e_t \geq 0$ is the one-step evidence factor at round t. If $\mathbb{E}_{P_0}[e_t \mid \mathcal{F}_{t-1}] \leq 1$ for every $P_0 \in \mathcal{P}_0$, then (M_t) is a nonnegative supermartingale under every null distribution, and hence an e-process for $\mathcal{P}_0$.

To enforce this condition, let $h_{t-1} \colon \mathcal{X} \to [0,1]$ be any $\mathcal{F}_{t-1}$-measurable score and define its largest possible null expectation by

$$m_{t-1} := \sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[h_{t-1}(X)],$$

where the expectation is taken with respect to the input X. Centering the score gives $g_{t-1}(x) := h_{t-1}(x) - m_{t-1}$, so that $\mathbb{E}_{P_0}[g_{t-1}(X)] \le 0$ for every $P_0 \in \mathcal{P}_0$. We may therefore take $e_t := 1 + \lambda_{t-1} g_{t-1}(Y_t)$, where $\lambda_{t-1} \ge 0$ is $\mathcal{F}_{t-1}$ measurable and chosen so that $e_t \ge 0$. Then

$$\mathbb{E}_{P_0}[e_t \mid \mathcal{F}_{t-1}] \le 1 \qquad \text{for every } P_0 \in \mathcal{P}_0.$$

Thus, validity is guaranteed once the score is centered at its worst-case null expectation. The remaining question is how to choose h_{t-1} so that, when the null is false, the resulting payoff tends to be positive. We answer this using kernel mean embeddings.

Predictable kernel witness. To construct h_{t-1}, we represent distributions as points in a reproducing kernel Hilbert space (RKHS). Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a measurable positive-definite kernel satisfying $\sup_{x \in \mathcal{X}} k(x,x) \le 1$, with RKHS $\mathcal{H}$. For any probability distribution R, define its kernel mean embedding by $\mu_R := \mathbb{E}_{X \sim R}[k(X, \cdot)]$. Because the embedding is linear in the distribution, the mixture null is represented by the convex polytope

$$\mathcal{M}_0 := \{\mu_{P_w} : w \in \Delta_K\} = \operatorname{conv}\{\mu_{F_1}, \dots, \mu_{F_K}\}.$$

At the beginning of round $t \ge 2$, using the observations available so far define the empirical embedding

$$\widehat{\mu}_{t-1} := \frac{1}{t-1} \sum_{i=1}^{t-1} k(Y_i, \cdot).$$

We project $\widehat{\mu}_{t-1}$ onto the null polytope and define

$$\widehat{\mu}^{\dagger}_{t-1} := \Pi_{\mathcal{M}_0}(\widehat{\mu}_{t-1}), \quad r_{t-1} := \widehat{\mu}_{t-1} - \widehat{\mu}^{\dagger}_{t-1}, \quad \widehat{\delta}_{t-1} := \|r_{t-1}\|_{\mathcal{H}}.$$

Geometrically, r_{t-1} points in the direction in which the empirical distribution departs from its closest mixture of the known components. We normalize this direction to obtain

$$h_{t-1}(x) := \begin{cases} \dfrac{1}{2}\left(1 + \dfrac{r_{t-1}(x)}{\widehat{\delta}_{t-1}}\right), & \widehat{\delta}_{t-1} > 0, \\ \dfrac{1}{2}, & \widehat{\delta}_{t-1} = 0. \end{cases} \tag{3}$$

By the reproducing property, $|r_{t-1}(x)| \le \|r_{t-1}\|_{\mathcal{H}} \sqrt{k(x,x)} \le \widehat{\delta}_{t-1}$, so $h_{t-1}(x) \in [0,1]$. Moreover, h_{t-1} depends only on $Y_1, \dots, Y_{t-1}$ and is therefore available before Y_t arrives.

Null calibration and betting size. The mixture representation of $\mathcal{M}_0$ makes the projection finite-dimensional. Since $\widehat{\mu}^{\dagger}_{t-1} = \Pi_{\mathcal{M}_0}(\widehat{\mu}_{t-1})$ belongs to $\mathcal{M}_0$, it can be written as a convex combination of the component embeddings. Accordingly, we choose $\widehat{w}_{t-1} \in \Delta_K$ such that

$$\begin{aligned} \widehat{w}_{t-1} &\in \operatorname*{arg\,min}_{w \in \Delta_K} \left\| \widehat{\mu}_{t-1} - \sum_{k=1}^{K} w_k \mu_{F_k} \right\|_{\mathcal{H}}^2, \\ \widehat{\mu}^{\dagger}_{t-1} &= \sum_{k=1}^{K} \widehat{w}_{t-1,k} \mu_{F_k}. \end{aligned} \tag{4}$$

The projected point $\widehat{\mu}^{\dagger}_{t-1}$ is unique, although its representation by $\widehat{w}_{t-1}$ need not be unique if the component embeddings are affinely dependent. Any minimizer in equation 4 yields the same projection residual.

To express this optimization using component-level quantities, define the Gram matrix

$$Q_{uv} := \langle \mu_{F_u}, \mu_{F_v} \rangle_{\mathcal{H}} = \mathbb{E}_{\substack{X \sim F_u \\ X' \sim F_v}}[k(X, X')],$$

and the component mean functions $\mu_{F_k}(y) := \mathbb{E}_{X \sim F_k}[k(y, X)]$. Also, define

$$(c_{t-1})_k := \langle \widehat{\mu}_{t-1}, \mu_{F_k} \rangle_{\mathcal{H}} = \frac{1}{t-1} \sum_{i=1}^{t-1} \mu_{F_k}(Y_i).$$

For any $w \in \Delta_K$, expanding the objective in equation 4 gives

$$\left\| \widehat{\mu}_{t-1} - \sum_{k=1}^{K} w_k \mu_{F_k} \right\|_{\mathcal{H}}^2 = \|\widehat{\mu}_{t-1}\|_{\mathcal{H}}^2 - 2c_{t-1}^\top w + w^\top Q w.$$

The first term does not depend on w. Hence the projection weights can be computed from the equivalent K-dimensional quadratic program

$$\widehat{w}_{t-1} \in \arg\min_{w \in \Delta_K} \left\{ w^\top Q w - 2c_{t-1}^\top w \right\}. \tag{5}$$

Because Q is a Gram matrix, it is positive semidefinite, so equation 5 is a convex quadratic program. Thus, although the witness is defined in the possibly infinite-dimensional space $\mathcal{H}$, computing its direction requires optimizing only over the K mixture weights.

Once the witness has been computed, it remains to center it at its largest possible expectation under the null. The same mixture structure again reduces this apparently infinite optimization to a finite calculation.

Lemma 1. *For every bounded measurable* $h\colon \mathcal{X} \to \mathbb{R}$, $\sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[h(X)] = \max_{1 \le k \le K} \mathbb{E}_{F_k}[h(X)]$.

Indeed, for $P_w = \sum_{k=1}^{K} w_k F_k$, the expectation $\mathbb{E}_{P_w}[h(X)]$ is affine in w, and hence its maximum over Δ_K is attained at a vertex. For the learned witness, define $a_{t,j} := \mathbb{E}_{F_j}[h_t(X)]$, $1 \le j \le K$. When $\widehat{\delta}_t > 0$, these expectations are

$$a_{t,j} = \frac{1}{2} + \frac{(c_t)_j - (Q\widehat{w}_t)_j}{2\widehat{\delta}_t}, \qquad 1 \le j \le K. \tag{6}$$

If $\widehat{\delta}_t = 0$, then $h_t \equiv 1/2$ and hence $a_{t,j} = 1/2$ for every j. Therefore,

$$m_t = \max_{1 \le j \le K} a_{t,j}, \qquad g_t := h_t - m_t. \tag{7}$$

By Lemma 1, $\mathbb{E}_{P_0}[g_t(X)] = \mathbb{E}_{P_0}[h_t(X)] - m_t \le 0$, $P_0 \in \mathcal{P}_0$. It remains to choose how aggressively to bet on the centered payoff. A fixed betting fraction would preserve validity, but may be inefficient because the magnitude and variability of the payoff are unknown. Since wealth accumulates multiplicatively, the natural criterion is its logarithmic growth. We therefore choose the betting fraction that would have maximized the cumulative log growth of the current payoff g_t on the observations seen so far.

Fix $\varepsilon \in (0, 1/2]$ and $\lambda_{\max} \ge 1/2$. Define

$$\begin{aligned} \overline{\lambda}_t &:= \min\left\{ \lambda_{\max}, \frac{1-\varepsilon}{m_t} \right\}, \\ \ell_t(\lambda) &:= \sum_{i=1}^{t} \log\{1 + \lambda g_t(Y_i)\}, \\ \lambda_t &\in \arg\max_{0 \le \lambda \le \overline{\lambda}_t} \ell_t(\lambda), \end{aligned} \tag{8}$$

where $(1-\varepsilon)/0 := +\infty$. The objective $\ell_t(\lambda)$ is the log wealth that would result from applying the same fraction λ to the current payoff g_t over the observed sample. Thus, λ_t adapts the size of the bet to the empirically observed strength of the separation. The optimization is one-dimensional and concave.

The restriction on the feasible range controls downside risk. Since $h_t \in [0,1]$ and $m_t \ge 0$, $g_t(x) = h_t(x) - m_t \ge -m_t$. Hence, whenever $m_t > 0$, $1 + \lambda_t g_t(x) \ge 1 - \lambda_t m_t \ge \varepsilon$. Thus, no single observation can reduce the wealth by more than a factor ε. If $m_t = 0$, then $g_t = h_t \ge 0$, so the factor is at least one.

The complete SeqGoF procedure is summarized in Algorithm 1. At every round, it learns a direction from the observations seen so far, calibrates that direction against the worst-case null component, and bets on the centered payoff when the next observation arrives. The most computationally expensive step is the quadratic program in equation 5. It is K-dimensional, independent of the ambient dimension of $\mathcal{X}$ and of t, and solvable in $O(K^3)$ worst case, typically in $O(K^2)$ with warm starting.

Algorithm 1 SeqGoF: Sequential Goodness-of-Fit Testing via Betting

Require: $F_1, \dots, F_K$, k, $\alpha, \varepsilon \in (0, 1/2]$, and $\lambda_{\max} \geq 1/2$.
Precompute Q and the functions $y \mapsto \mu_{F_k}(y)$.
Set $M_0 \leftarrow 1$, $h_0 \equiv 1/2$, $m_0 \leftarrow 1/2$, $g_0 \equiv 0$, and $\lambda_0 \leftarrow 0$.
for $t = 1, 2, \dots$ **do**
    Observe Y_t and set
$$M_t \leftarrow M_{t-1}\{1 + \lambda_{t-1} g_{t-1}(Y_t)\}.$$
    if $M_t \geq 1/\alpha$ **then**
        Reject H_0 and stop with $\tau_\alpha = t$.
    end if
    Compute $\widehat{\mu}_t$ and c_t from $Y_1, \dots, Y_t$.
    Solve equation 5 for $\widehat{w}_t$, and form r_t and $\widehat{\delta}_t = \|r_t\|_{\mathcal{H}}$.
    Construct h_t using equation 3.
    Compute $a_{t,j}$, m_t, and g_t using equation 6 and equation 7.
    Choose λ_t using equation 8.
end for

Remark 1. The preceding development assumes that the kernel quantities $Q_{u,v}$ and $\mu_{F_k}(x)$ are available in closed form. This holds, for example, for Gaussian reference distributions with Gaussian kernels and for several other analytically tractable kernel–distribution pairs. When closed-form expressions are unavailable but only samples from the F_k's are observed (e.g., the components are simulators or generative models), these quantities can instead be estimated offline from reference samples. Specifically, suppose that we observe an independent reference sample $X_1^{(k)}, \dots, X_m^{(k)} \sim F_k$ for each k. We replace the population kernel mean by $\widehat{\mu}_{F_k}(x) = m^{-1} \sum_{j=1}^{m} k(X_j^{(k)}, x)$, and estimate the Gram matrix of reference mean embeddings by

$$\widehat{Q}_{k\ell} = \frac{1}{m^2} \sum_{r=1}^{m} \sum_{s=1}^{m} k(X_r^{(k)}, X_s^{(\ell)}).$$

All subsequent quantities are computed exactly as in Algorithm 1 after replacing μ_{F_k} and Q by their empirical counterparts. In particular, the projection step remains a K-dimensional quadratic program, while the component-wise expectations used for null calibration are replaced by empirical averages over the corresponding reference samples. An explicit finite-sample validity correction is given in Theorem 4 of Appendix D.

3.2 Theoretical Guarantees

We first formalize the validity of the construction in Section 3.1.

Theorem 1. *Let $(M_t)_{t \geq 0}$ be the wealth process generated by Algorithm 1. Then $(M_t)_{t \geq 0}$ is a nonnegative supermartingale under every $P_0 \in \mathcal{P}_0$. Consequently, for every $\alpha \in (0, 1)$,*

$$\sup_{P_0 \in \mathcal{P}_0} \mathbb{P}_{P_0} \left(\sup_{t \geq 0} M_t \geq \frac{1}{\alpha} \right) \leq \alpha,$$

and hence $\sup_{P_0 \in \mathcal{P}_0} \mathbb{P}_{P_0}(\tau_\alpha < \infty) \leq \alpha$. Thus τ_α is a level-α sequential test.

We next study the behavior under an alternative. Define the kernel distance from P to the null by

$$\delta(P) := \operatorname{dist}_{\mathcal{H}}(\mu_P, \mathcal{M}_0) = \min_{w \in \Delta_K} \left\| \mu_P - \sum_{k=1}^{K} w_k \mu_{F_k} \right\|_{\mathcal{H}}.$$

Suppose $\delta(P) > 0$. Let $\mu^\dagger := \Pi_{\mathcal{M}_0}(\mu_P)$, $r^* := \mu_P - \mu^\dagger$, and define the population witness and centered payoff by

$$\begin{aligned} h^*(x) &:= \frac{1}{2} \left(1 + \frac{r^*(x)}{\delta(P)} \right), \\ m^* &:= \max_{1 \leq k \leq K} \mathbb{E}_{F_k}[h^*(X)], \qquad g^* := h^* - m^*. \end{aligned} \tag{9}$$

Lemma 2. *Suppose $\delta(P) > 0$, and let h^*, m^*, and g^* be defined in equation 9. Then*

$$\sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[g^*(X)] = 0, \text{ and } \mathbb{E}_P[g^*(Y)] = \frac{\delta(P)}{2} > 0.$$

Moreover, $\|g^*\|_\infty \le 1$.

To characterize the resulting growth rate, set $\overline{\lambda}_\varepsilon^* := \min\left\{\lambda_{\max}, \frac{1-\varepsilon}{m^*}\right\}$, with $(1-\varepsilon)/0 := +\infty$, and define

$$\Gamma_\varepsilon(P) := \max_{0 \le \lambda \le \overline{\lambda}_\varepsilon^*} \mathbb{E}_P\left[\log\{1 + \lambda g^*(Y)\}\right]. \tag{10}$$

Theorem 2. *Fix $\varepsilon \in (0, 1/2]$ and $\lambda_{\max} \ge 1/2$. If $Y_1, Y_2, \ldots \overset{\text{i.i.d.}}{\sim} P$ and $\delta(P) > 0$, then*

$$\frac{1}{t} \log M_t \longrightarrow \Gamma_\varepsilon(P) \qquad P\text{-almost surely.} \tag{11}$$

Moreover,

$$\Gamma_\varepsilon(P) \ge \frac{\delta(P)^2}{16} > 0. \tag{12}$$

Consequently, $\mathbb{P}_P(\tau_\alpha < \infty) = 1$ for every $\alpha \in (0, 1)$.

The proof uses the almost-sure convergence of the empirical embedding to μ_P and the nonexpansiveness of projection onto $\mathcal{M}_0$, which together imply that the learned payoff converges uniformly to g^*. The empirical criterion equation 8 then converges to the population objective in equation 10. Full details are given in Appendix A.3.

The positive growth rate also determines the first-order time needed to cross the rejection boundary.

Theorem 3 (Stopping-time asymptotics)**.** *Under the conditions of Theorem 2,*

$$\frac{\tau_\alpha}{\log(1/\alpha)/\Gamma_\varepsilon(P)} \longrightarrow 1 \qquad P\text{-almost surely as } \alpha \downarrow 0. \tag{13}$$

Combining Theorems 3 and 2 gives

$$\lim_{\alpha \downarrow 0} \frac{\tau_\alpha}{16 \log(1/\alpha)/\delta(P)^2} = \frac{\delta(P)^2}{16\Gamma_\varepsilon(P)} \le 1 \qquad P\text{-almost surely.}$$

Thus the proposed procedure is a level-α power-one sequential test against every alternative with $\delta(P) > 0$, with the upper bound $\tau_\alpha \le \left(\frac{16}{\delta(P)^2} + o(1)\right) \log(1/\alpha)$ P-almost surely as $\alpha \downarrow 0$. If the kernel mean embedding uniquely determines probability distributions, then $\delta(P) > 0$ for every $P \notin \mathcal{P}_0$.

4 Numerical Experiments

Throughout, we set $\alpha = 0.05$ and $d = 10$, and report empirical type-I error and power over $r = 500$ Monte Carlo (MC) replications. We write $\mathbb{I}(A)$ for the indicator of event A. Here, SeqGoF is implemented with a Gaussian radial basis function kernel, $k(x, y) = \exp\left\{-0.5\|x - y\|_2^2/h^2\right\}$, with bandwidth $h = \sqrt{d}$. All experiments use Algorithm 1; when reference quantities are estimated from samples, we use their empirical counterparts and retain the original rejection threshold $1/\alpha$, without the optional correction in Appendix D.

Experiment 1: Gaussian reference distributions - We set $K = 3$ and consider $P_w = \sum_{k=1}^K w_k N_d(\boldsymbol{\mu}_k, \Sigma_k)$, where $\boldsymbol{\mu}_1 = \mathbf{0}_d$, $\boldsymbol{\mu}_2 = -\mathbf{1}_d$, and $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2$. Each Σ_k is symmetric positive definite with eigenvalues randomly generated from $[1, 10]$. The left panel of Figure 1 shows the empirical type-I error of SeqGoF under three null configurations. In Scenario 1, $w = (0.9, 0.05, 0.05)$; in Scenario 2, $w = (1, 0, 0)$; and in Scenario 3, $w = (0, 0.5, 0.5)$. Thus, $P_w \in \mathcal{P}_0$ in all three scenarios. SeqGoF maintains the nominal 5% type-I error across these markedly different mixture configurations, including distributions lying on the boundary of the null family.

The right panel of Figure 1 reports the empirical power of SeqGoF. In Scenario 1, $P = 0.5N_d(\boldsymbol{\mu}_1, \Sigma_1) + 0.5N_d(\boldsymbol{\mu}_4, \Sigma_3)$, $\boldsymbol{\mu}_4 = 0.5\boldsymbol{\mu}_3$, while in Scenario 2 the same construction is

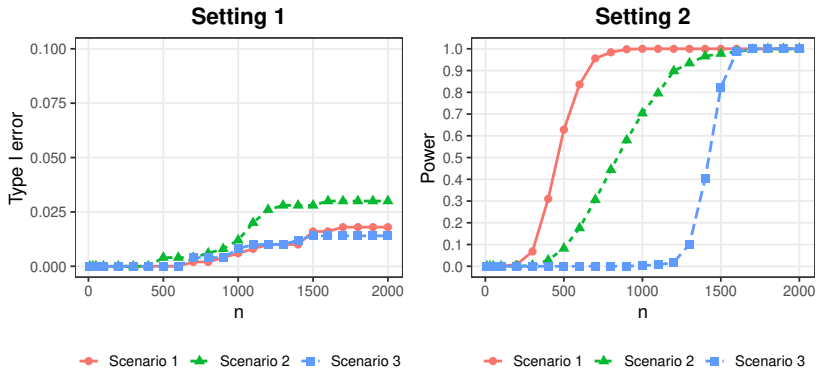

Figure 1: Empirical type-I error (left) and power (right) of SeqGoF under Gaussian reference distributions.

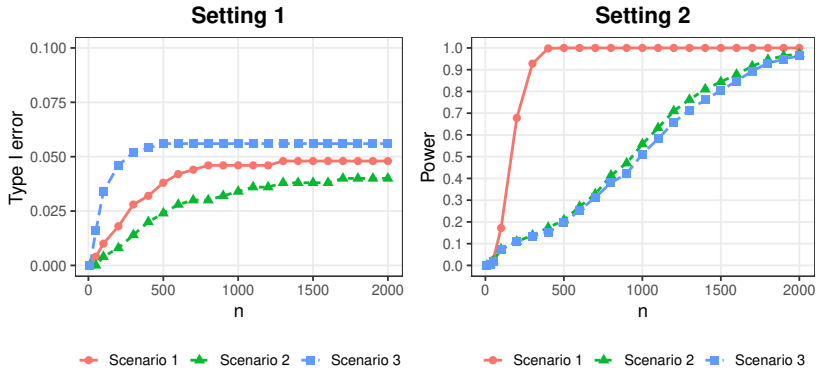

Figure 2: Empirical type-I error (left) and power (right) of SeqGoF under Gamma–Exponential reference distributions using $m = 50{,}000$ reference samples per component.

used with $\boldsymbol{\mu}_4 = 0.7\boldsymbol{\mu}_3$. Both distributions lie outside $\mathcal{P}_0$. Scenario 3 considers a change-point alternative: the observations are independent with $Y_t \sim \mathbb{I}(t \le t_0)P_1 + \mathbb{I}(t > t_0)P_2$ where $P_1 = 0.5N_d(\boldsymbol{\mu}_1, \Sigma_1) + 0.5N_d(\boldsymbol{\mu}_2, \Sigma_2)$, and $P_2 = 0.45N_d(\boldsymbol{\mu}_1, \Sigma_1) + 0.45N_d(\boldsymbol{\mu}_2, \Sigma_2) + 0.1N_d(\boldsymbol{\mu}_3, \Sigma_4)$. Here, Σ_4 is as above, and we set $t_0 = 1000$. Hence, the process follows the composite null up to t_0 and departs from it thereafter. Across these alternatives, the power curves illustrate that SeqGoF increasingly detects departures from the composite null as more observations become available. Moreover, Scenario 3 illustrates the adaptivity of SeqGoF to distributional changes occurring at an unknown time.

Experiment 2: Non-Gaussian reference distributions - The preceding Gaussian experiments admit closed-form expressions for the kernel quantities required by SeqGoF. We next consider Gamma and Exponential reference distributions and implement SeqGoF using empirical kernel estimates based on $m = 50{,}000$ reference samples from each distribution.

Specifically, we consider $P_w = w_1\,\mathrm{Gamma}_d(\mathrm{shape} = 5\mathbf{1}_d, \mathrm{rate} = \mathbf{1}_d, \Sigma_1) + w_2\,\mathrm{Exp}_d(\mathrm{rate} = \mathbf{1}_d, \Sigma_2)$, where $(\Sigma_1)_{ij} = 0.7^{|i-j|}$ and $(\Sigma_2)_{ij} = -0.9^{|i-j|}$. The left panel of Figure 2 reports the empirical type-I error under three null configurations: $w = (0.5, 0.5)$, $w = (1, 0)$, and $w = (0, 1)$. Since $P_w \in \mathcal{P}_0$ in all three cases, the results show that SeqGoF maintains the nominal 5% level across both interior and boundary points of the composite null. The right panel of Figure 2 reports empirical power. In Scenario 1 $P = 0.2\,\mathrm{Gamma}_d(\mathrm{shape} = 10\mathbf{1}_d, \mathrm{rate} = 0.5\mathbf{1}_d, \Sigma_1) + 0.8\,\mathrm{Exp}_d(\mathrm{rate} = \mathbf{1}_d, \Sigma_2)$, while Scenario 2 uses the same construction with mixture weight 0.1 on the Gamma component. In Scenario 3, $P = 0.1\,\mathrm{Gamma}_d(\mathrm{shape} = 7\mathbf{1}_d, \mathrm{rate} = 0.7\mathbf{1}_d, \Sigma_1) + 0.9\,\mathrm{Exp}_d(\mathrm{rate} = \mathbf{1}_d, \Sigma_2)$. Across these alternatives, SeqGoF exhibits increasing power with sample size. Together, the two panels show that the empirical-reference implementation retains accurate type-I error control while remaining sensitive to non-Gaussian departures from the composite null.

5 Real Data Illustration

We consider the Backblaze Drive Stats data, which contain daily Self-Monitoring, Analysis, and Reporting Technology (SMART) measurements for a large fleet of hard drives operating in production data centers. Each observation Y_t is a 7-dimensional SMART vector for one physical drive on one day, comprising SMART 5, 9, 187, 188, 194, 197, and 198 after transformation and standardization; consecutive observations may correspond to different drives. We restrict attention to a single drive model. Appendix C provides more details about this data and the features used.

To represent heterogeneity among healthy drives, we set $K = 3$ and partition healthy reference observations into three operating-age regimes based on tertiles of power-on hours (SMART 9). From each regime, we draw $m_{\text{ref}} = 20{,}000$ reference observations. We reserve $n_{\text{null}} = 5{,}000$ drive-day observations from disjoint never-failed drives to form the held-out null stream. For the evaluation stream, we prepend n_{pre} held-out healthy observations and then append up to $n_{\text{near_failure}} = 3{,}000$ near-failure drive-day observations, each recorded within the final 30 days before the corresponding drive failure. Thus, the near-failure segment begins at observation $n_{\text{pre}} + 1$.

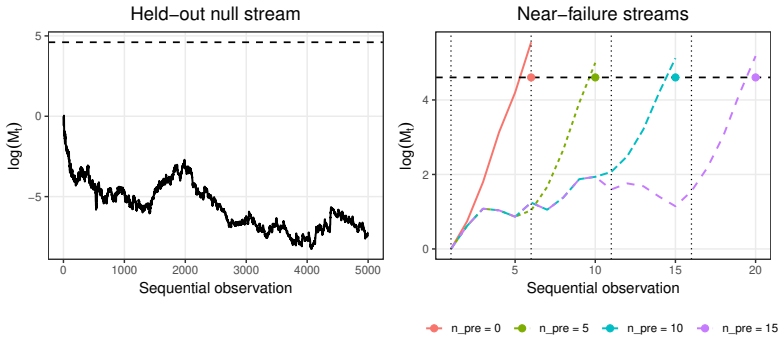

Figure 3: Log-wealth trajectories for the held-out healthy stream (left) and near-failure evaluation streams with varying healthy prefixes (right). The dashed horizontal line marks the rejection threshold, and dotted vertical lines indicate the onset of near-failure observations.

Figure 3 displays the log-wealth process $\log(M_t)$ for the held-out healthy stream (left) and for evaluation streams containing different numbers of healthy observations before the onset of near-failure telemetry (right). The horizontal dashed line denotes the rejection threshold $\log(1/\alpha)$, $\alpha = 0.01$, while the vertical dotted lines indicate the onset of near-failure observations. For the held-out healthy stream (left panel), the wealth remains well below the rejection threshold throughout, providing no evidence against the composite null. In contrast, after near-failure observations begin to arrive (right panel), the wealth increases rapidly and crosses the rejection threshold. In particular, for $n_{\text{pre}} = 0, 5$, and 10, detection occurs after approximately five near-failure observations. These results suggest that SeqGoF accommodates heterogeneity among healthy drives while rapidly detecting distributional changes associated with impending drive failure.

6 Discussion

We introduced SeqGoF, a nonparametric sequential goodness-of-fit procedure for heterogeneous composite nulls formed by mixtures of reference distributions with unknown weights. By combining kernel mean embeddings with testing by betting, SeqGoF achieves anytime-valid type-I error control under arbitrary stopping, consistency against fixed alternatives outside the null, and first-order stopping-time guarantees. Simulations show accurate type-I error control across diverse mixture configurations, including when the reference distributions are available only through samples. The Backblaze SMART application further demonstrates that SeqGoF can accommodate heterogeneity among healthy drives while rapidly detecting near-failure behavior. An important direction for future work is to extend these ideas to anytime-valid sequential two-sample testing under the same heterogeneous mixture setting, where both data streams may have unknown and potentially different mixing proportions.

AI USE STATEMENT

In this work, we used generative AI tools for editing the prose and polishing the writing. We have not used generative AI tools for idea generation, theoretical proofs, and coding, and the rest of the required disclosure tasks are not applicable to this work.

REPRODUCIBILITY STATEMENT

The analysis in sections 4 and 5 can be reproduced using the source code available at: https://www.dropbox.com/scl/fo/3l3u0wi2fmh5tsr5ekm1q/AI2IWeI7ToTPWOHinMCTxNI?rlkey=qlk27xrzeha1iw7dpxgrm8q1u&st=c2zjxeud&dl=0.

Appendix A provides the rigorous proof of all theoretical statements in the paper.

REFERENCES

Trambak Banerjee, Bhaswar B. Bhattacharya, and Gourab Mukherjee. A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data. *The Annals of Applied Statistics*, 14(4):1777 – 1805, 2020. doi: 10.1214/20-AOAS1362. URL https://doi.org/10.1214/20-AOAS1362.

Trambak Banerjee, Bhaswar B Bhattacharya, and Gourab Mukherjee. Bootstrapped edge count tests for nonparametric two-sample inference under heterogeneity. *Journal of Computational and Graphical Statistics*, 34(1):306–317, 2025.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. *Journal of Machine Learning Research*, 11:2973–3009, 2010.

Marielle Cavrois, Trambak Banerjee, Gourab Mukherjee, Nandhini Raman, Rajaa Hussien, Brandon Aguilar Rodriguez, Joshua Vasquez, Matthew H Spitzer, Nicole H Lazarus, Jennifer J Jones, et al. Mass cytometric analysis of hiv entry, replication, and remodeling in tissue cd4+ t cells. *Cell reports*, 20(4):984–998, 2017.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML*, 2016.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Peter Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society Series B*, 86(5):1091–1128, 2024.

Bikram Karmakar, Kumaresh Dhara, Kushal Kumar Dey, Analabha Basu, and Anil Kumar Ghosh. Tests for statistical significance of a treatment effect in the presence of hidden sub-populations. *Statistical Methods & Applications*, 24(1):97–119, 2015.

Qiang Liu, Jason Lee, and Michael I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *ICML*, 2016.

Aleksandr Podkopaev and Aaditya Ramdas. Sequential predictive two-sample and independence testing. In *NeurIPS*, 2023.

Sridhar Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *ICML*, 2016.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *arXiv:2410.23614*, 2024.

Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

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

S. Shekhar. Optimal anytime-valid tests for composite nulls. *arXiv:2512.20039*, 2025.

S. Shekhar and Aaditya Ramdas. Nonparametric two-sample testing by betting. *IEEE Transactions on Information Theory*, 70(2):1178–1203, 2023.

Bharath Sriperumbudur, Kenji Fukumizu, and Gert Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12:2389–2410, 2011.

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

Zhengyu Zhou and Weiwei Liu. Sequential kernel goodness-of-fit testing. In *Forty-first International Conference on Machine Learning*, 2024.

A Proofs

A.1 Proof of Lemma 1

Proof. Because h is bounded and measurable, all expectations below are finite. For any $P_0 = P_w \in \mathcal{P}_0$, with $w \in \Delta_K$, linearity of integration gives

$$\mathbb{E}_{P_w}[h(X)] = \sum_{k=1}^{K} w_k \mathbb{E}_{F_k}[h(X)] \leq \max_{1\leq k\leq K} \mathbb{E}_{F_k}[h(X)].$$

Taking the supremum over $w \in \Delta_K$ yields

$$\sup_{P_0\in\mathcal{P}_0} \mathbb{E}_{P_0}[h(X)] \leq \max_{1\leq k\leq K} \mathbb{E}_{F_k}[h(X)].$$

Conversely, each F_k belongs to $\mathcal{P}_0$, corresponding to the kth vertex of Δ_K. Hence

$$\sup_{P_0\in\mathcal{P}_0} \mathbb{E}_{P_0}[h(X)] \geq \max_{1\leq k\leq K} \mathbb{E}_{F_k}[h(X)].$$

Combining the two inequalities proves the result. □

A.2 Proof of Theorem 1

Proof. Fix $P_0 \in \mathcal{P}_0$. For every $t \geq 1$, the quantities h_{t-1}, m_{t-1}, and λ_{t-1} are $\mathcal{F}_{t-1}$-measurable. Moreover,

$$g_{t-1}(x) = h_{t-1}(x) - m_{t-1} \geq -m_{t-1}.$$

If $m_{t-1} > 0$, the constraint $\lambda_{t-1} \leq (1-\varepsilon)/m_{t-1}$ gives

$$e_t = 1 + \lambda_{t-1} g_{t-1}(Y_t) \geq \varepsilon.$$

If $m_{t-1} = 0$, then $g_{t-1} = h_{t-1} \geq 0$, so $e_t \geq 1$. Thus, $M_t \geq 0$.

Since Y_t is independent of $\mathcal{F}_{t-1}$ under the i.i.d. model, and since

$$m_{t-1} = \sup_{Q\in\mathcal{P}_0} \mathbb{E}_Q[h_{t-1}(X)],$$

we have

$$\begin{aligned}\mathbb{E}_{P_0}[e_t \mid \mathcal{F}_{t-1}] &= 1 + \lambda_{t-1}\left\{\mathbb{E}_{P_0}[h_{t-1}(X)] - m_{t-1}\right\} \\ &\leq 1.\end{aligned}$$

Therefore,

$$\mathbb{E}_{P_0}[M_t \mid \mathcal{F}_{t-1}] = M_{t-1}\mathbb{E}_{P_0}[e_t \mid \mathcal{F}_{t-1}] \leq M_{t-1}.$$

Hence $(M_t)_{t\geq 0}$ is a nonnegative supermartingale under P_0. By Ville's inequality the claim for τ_α follows immediately. □

A.3 Proof of Theorem 2

Proof. Write

$$\delta := \delta(P), \qquad \mu^\dagger := \Pi_{\mathcal{M}_0}(\mu_P), \qquad r^* := \mu_P - \mu^\dagger.$$

Step 1: Convergence of the learned payoff. First, the empirical mean embedding converges to its population counterpart almost surely:

$$\|\widehat{\mu}_t - \mu_P\|_{\mathcal{H}} \longrightarrow 0. \tag{14}$$

Indeed,

$$\mathbb{E}_P\|\widehat{\mu}_t - \mu_P\|_{\mathcal{H}}^2 = \frac{1}{t}\left\{\mathbb{E}_P[k(Y,Y)] - \|\mu_P\|_{\mathcal{H}}^2\right\} \le \frac{1}{t}.$$

Moreover, replacing one observation changes $\|\widehat{\mu}_t - \mu_P\|_{\mathcal{H}}$ by at most $2/t$. McDiarmid's inequality therefore gives, for $t \ge 2$,

$$\mathbb{P}_P\left(\|\widehat{\mu}_t - \mu_P\|_{\mathcal{H}} > \frac{1}{\sqrt{t}} + 2\sqrt{\frac{\log t}{t}}\right) \le \frac{1}{t^2}.$$

Equation equation 14 follows from the Borel–Cantelli lemma.

Because metric projection onto a closed convex set is nonexpansive,

$$\|\Pi_{\mathcal{M}_0}(\widehat{\mu}_t) - \Pi_{\mathcal{M}_0}(\mu_P)\|_{\mathcal{H}} \le \|\widehat{\mu}_t - \mu_P\|_{\mathcal{H}}.$$

Hence

$$\|r_t - r^*\|_{\mathcal{H}} \le 2\|\widehat{\mu}_t - \mu_P\|_{\mathcal{H}} \longrightarrow 0 \qquad \text{almost surely.}$$

It follows that

$$\widehat{\delta}_t = \|r_t\|_{\mathcal{H}} \longrightarrow \|r^*\|_{\mathcal{H}} = \delta > 0.$$

Therefore, almost surely, $\widehat{\delta}_t > 0$ for all sufficiently large t, and

$$\left\|\frac{r_t}{\widehat{\delta}_t} - \frac{r^*}{\delta}\right\|_{\mathcal{H}} \longrightarrow 0.$$

Using the reproducing property and $\sup_x k(x,x) \le 1$, we obtain

$$\|h_t - h^*\|_\infty \le \frac{1}{2}\left\|\frac{r_t}{\widehat{\delta}_t} - \frac{r^*}{\delta}\right\|_{\mathcal{H}} \longrightarrow 0.$$

Consequently,

$$\begin{aligned}|m_t - m^*| &\le \max_{1\le k\le K} |\mathbb{E}_{F_k}[h_t(X)] - \mathbb{E}_{F_k}[h^*(X)]| \\ &\le \|h_t - h^*\|_\infty \longrightarrow 0,\end{aligned}$$

and hence

$$\|g_t - g^*\|_\infty \le \|h_t - h^*\|_\infty + |m_t - m^*| \longrightarrow 0 \qquad \text{almost surely.} \tag{15}$$

Step 2: Asymptotic optimality of the empirical choice of λ_t. For brevity, write

$$b_t := \overline{\lambda}_t, \qquad b^* := \overline{\lambda}^*_\varepsilon.$$

Since the map

$$m \longmapsto \min\left\{\lambda_{\max}, \frac{1-\varepsilon}{m}\right\},$$

with value $\lambda_{\max}$ at $m = 0$, is continuous on $[0,1]$, the convergence $m_t \to m^*$ implies

$$b_t \longrightarrow b^* \qquad \text{almost surely.} \tag{16}$$

Because $b_t > 0$, write

$$u_t := \frac{\lambda_t}{b_t} \in [0,1].$$

By construction, u_t maximizes over $u \in [0,1]$ the criterion

$$\widehat{G}_t(u) := \frac{1}{t}\sum_{i=1}^{t} \log\left\{1 + u b_t g_t(Y_i)\right\}.$$

Define its population limit by

$$G(u) := \mathbb{E}_P\left[\log\left\{1 + u b^* g^*(Y)\right\}\right], \qquad u \in [0,1].$$

By a change of variables,

$$\max_{u\in[0,1]} G(u) = \Gamma_\varepsilon(P).$$

For every $u \in [0,1]$, the definitions of b_t and b^* give

$$1 + u b_t g_t(x) \geq \varepsilon, \qquad 1 + u b^* g^*(x) \geq \varepsilon.$$

Since $|g_t|, |g^*| \leq 1$, the logarithm is therefore uniformly Lipschitz on the relevant range, and

$$\begin{aligned}\sup_{u\in[0,1]} \left| \frac{1}{t}\sum_{i=1}^{t} \log\{1 + u b_t g_t(Y_i)\} - \frac{1}{t}\sum_{i=1}^{t} \log\{1 + u b^* g^*(Y_i)\} \right| \\ \leq \frac{|b_t - b^*| + \lambda_{\max}\|g_t - g^*\|_\infty}{\varepsilon} \longrightarrow 0. \end{aligned} \tag{17}$$

The class

$$\{y \mapsto \log\{1 + u b^* g^*(y)\} : u \in [0,1]\}$$

is uniformly bounded and uniformly Lipschitz in u. A finite-grid argument together with the strong law of large numbers therefore gives

$$\sup_{u\in[0,1]} \left| \frac{1}{t}\sum_{i=1}^{t} \log\{1 + u b^* g^*(Y_i)\} - G(u) \right| \longrightarrow 0$$

almost surely. Combining this with equation 17 yields

$$\sup_{u\in[0,1]} |\widehat{G}_t(u) - G(u)| \longrightarrow 0 \qquad \text{almost surely.} \tag{18}$$

Since u_t maximizes $\widehat{G}_t$, uniform convergence implies

$$G(u_t) \longrightarrow \max_{u\in[0,1]} G(u) = \Gamma_\varepsilon(P) \qquad \text{almost surely.} \tag{19}$$

Indeed, if u^* maximizes G, then

$$\Gamma_\varepsilon(P) - G(u_t) \leq 2 \sup_{u\in[0,1]} |\widehat{G}_t(u) - G(u)|.$$

Step 3: Convergence of the realized log wealth. Let

$$Z_t := \log\left\{1 + \lambda_{t-1} g_{t-1}(Y_t)\right\}.$$

Then

$$\log M_t = \sum_{s=1}^{t} Z_s.$$

Since λ_{t-1} and g_{t-1} are $\mathcal{F}_{t-1}$-measurable,

$$\begin{aligned} d_t &:= \mathbb{E}_P[Z_t \mid \mathcal{F}_{t-1}] \\ &= \mathbb{E}_P\left[\log\left\{1 + u_{t-1} b_{t-1} g_{t-1}(Y)\right\}\right]. \end{aligned}$$

The same Lipschitz bound used above, together with equation 15 and equation 16, shows that

$$|d_t - G(u_{t-1})| \longrightarrow 0.$$

By equation 19,

$$d_t \longrightarrow \Gamma_\varepsilon(P) \qquad \text{almost surely.} \tag{20}$$

The betting constraints imply

$$\varepsilon \le 1 + \lambda_{t-1} g_{t-1}(Y_t) \le 1 + \lambda_{\max}.$$

Thus the increments Z_t are uniformly bounded. Hence

$$D_t := Z_t - d_t$$

is a bounded martingale-difference sequence, and the martingale strong law gives

$$\frac{1}{t}\sum_{s=1}^{t} D_s \longrightarrow 0 \qquad \text{almost surely.}$$

Combining this with equation 20 and Cesàro's lemma yields

$$\begin{aligned}\frac{1}{t}\log M_t &= \frac{1}{t}\sum_{s=1}^{t} d_s + \frac{1}{t}\sum_{s=1}^{t} D_s \\ &\longrightarrow \Gamma_\varepsilon(P)\end{aligned}$$

almost surely, proving equation 11.

Step 4: Positivity of the growth rate. By Lemma 2,

$$\mathbb{E}_P[g^*(Y)] = \frac{\delta}{2}, \qquad \|g^*\|_\infty \le 1.$$

Moreover, $\|\mu_R\|_{\mathcal{H}} \le 1$ for every probability distribution R, so $\delta \le 2$. Therefore

$$\lambda^\circ := \frac{\delta}{4} \le \frac{1}{2}$$

is feasible: $\lambda^\circ \le \lambda_{\max}$, and, if $m^* > 0$,

$$\frac{1-\varepsilon}{m^*} \ge 1 - \varepsilon \ge \frac{1}{2}.$$

Using $\log(1+x) \ge x - x^2$ for $x \ge -1/2$, we obtain

$$\begin{aligned}\Gamma_\varepsilon(P) &\ge \mathbb{E}_P\left[\log\left\{1 + \frac{\delta}{4} g^*(Y)\right\}\right] \\ &\ge \frac{\delta}{4}\mathbb{E}_P[g^*(Y)] - \frac{\delta^2}{16}\mathbb{E}_P[g^*(Y)^2] \\ &\ge \frac{\delta^2}{8} - \frac{\delta^2}{16} = \frac{\delta^2}{16} > 0.\end{aligned}$$

This proves equation 12.

Finally,

$$\frac{1}{t}\log M_t \longrightarrow \Gamma_\varepsilon(P) > 0$$

implies $M_t \to \infty$ almost surely. Hence M_t eventually crosses every finite threshold, so

$$\mathbb{P}_P(\tau_\alpha < \infty) = 1 \qquad \text{for every } \alpha \in (0,1).$$

□

A.4 Proof of Lemma 2

Proof. Write $\delta = \delta(P)$. Since $\mu^\dagger = \Pi_{\mathcal{M}_0}(\mu_P)$ and $r^* = \mu_P - \mu^\dagger$, the variational characterization of the metric projection gives

$$\langle r^*, \mu - \mu^\dagger \rangle_{\mathcal{H}} \le 0 \qquad \text{for every } \mu \in \mathcal{M}_0.$$

Consequently,

$$\sup_{\mu \in \mathcal{M}_0} \langle r^*, \mu \rangle_{\mathcal{H}} = \langle r^*, \mu^\dagger \rangle_{\mathcal{H}}.$$

For every probability distribution R, the reproducing property gives

$$\mathbb{E}_R[h^*(X)] = \frac{1}{2} + \frac{\langle r^*, \mu_R \rangle_{\mathcal{H}}}{2\delta}.$$

Hence, by Lemma 1 and the identity $\{\mu_{P_0} : P_0 \in \mathcal{P}_0\} = \mathcal{M}_0$,

$$\begin{aligned} m^* &= \sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[h^*(X)] \\ &= \frac{1}{2} + \frac{\langle r^*, \mu^\dagger \rangle_{\mathcal{H}}}{2\delta}. \end{aligned}$$

It follows that

$$\begin{aligned} \mathbb{E}_P[g^*(Y)] &= \mathbb{E}_P[h^*(Y)] - m^* \\ &= \frac{\langle r^*, \mu_P - \mu^\dagger \rangle_{\mathcal{H}}}{2\delta} \\ &= \frac{\|r^*\|_{\mathcal{H}}^2}{2\delta} = \frac{\delta}{2}. \end{aligned}$$

Moreover,

$$\begin{aligned} \sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[g^*(X)] &= \sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[h^*(X)] - m^* \\ &= 0. \end{aligned}$$

Finally, for every $x \in \mathcal{X}$,

$$|r^*(x)| \le \|r^*\|_{\mathcal{H}} \sqrt{k(x,x)} \le \delta.$$

Thus $h^*(x) \in [0,1]$. Since also $m^* \in [0,1]$, we obtain $\|g^*\|_\infty \le 1$. □

A.5 Proof of Theorem 3

Proof. Write

$$\Gamma := \Gamma_\varepsilon(P) > 0, \qquad A_\alpha := \log(1/\alpha).$$

By Theorem 2,

$$\frac{1}{t} \log M_t \longrightarrow \Gamma \qquad P\text{-almost surely.}$$

Fix any $\eta \in (0, \Gamma)$. On an event of probability one, there exists a finite random time T_η such that

$$(\Gamma - \eta)t \le \log M_t \le (\Gamma + \eta)t$$

for every $t \ge T_\eta$.

For all sufficiently small α, the boundary A_α exceeds $\max_{t < T_\eta} \log M_t$. Hence the process cannot cross the boundary before T_η, and the preceding inequalities imply

$$\frac{A_\alpha}{\Gamma + \eta} \le \tau_\alpha \le \frac{A_\alpha}{\Gamma - \eta} + 1.$$

Dividing by A_α and taking lower and upper limits gives

$$\frac{1}{\Gamma + \eta} \le \liminf_{\alpha \downarrow 0} \frac{\tau_\alpha}{A_\alpha} \le \limsup_{\alpha \downarrow 0} \frac{\tau_\alpha}{A_\alpha} \le \frac{1}{\Gamma - \eta}.$$

Letting $\eta \downarrow 0$ proves equation 13. □

B Comparison with Shekhar (2025)

The Donsker-Varadhan (DV) e-process construction of Shekhar (2025) is closely related to SeqGoF. Given a weakly compact convex null $\mathcal{P}_0$ and a function class $\mathcal{F}$, their procedure selects

$$f_t \approx \arg\max_{f\in\mathcal{F}} \Big(\frac{1}{t-1}\sum_{i<t} f(Y_i) - \psi_0(f) \Big),$$
$$\psi_0(f) = \sup_{Q\in\mathcal{P}_0} \log \mathbb{E}_Q\big[e^{f(X)}\big],$$

and uses the e-factor $\exp(f_t(Y_t) - \psi_0(f_t))$. Under their Assumption 1, including the condition that the restricted DV objective over $\mathcal{F}$ recovers $\mathrm{KL}_{\inf}(P, \mathcal{P}_0)$, the resulting tests are first-order optimal. Their general formulation requires computing both the empirical maximizer over $\mathcal{F}$ and the null functional $\psi_0(f_t)$. Shekhar (2025, §3.2) discusses finite-dimensional approximations and heuristic procedures for these computations.

Our construction shares some features with this framework, but uses a different criterion for choosing the test function. For

$$f_{\lambda,h}(x) = \log\big(1 + \lambda(h(x) - m_h)\big),$$
$$m_h := \max_k \mathbb{E}_{F_k}[h(X)],$$

we have, for any $Q \in \mathcal{P}_0$,

$$\mathbb{E}_Q\big[e^{f_{\lambda,h}(X)}\big] = 1 + \lambda\big(\mathbb{E}_Q[h] - m_h\big) \le 1.$$

Moreover, because m_h is attained by at least one component,

$$\sup_{Q\in\mathcal{P}_0} \mathbb{E}_Q\big[e^{f_{\lambda,h}(X)}\big] = 1,$$
$$\psi_0(f_{\lambda,h}) = 0.$$

Thus the null functional for this betting class is determined by the component-wise expectations in Lemma 1. A related simplification appears in the bounded-mean example of Shekhar (2025, §B.2), although the null class and the construction of the witness are different.

The main difference lies in the selection of h. The DV procedure of Shekhar (2025) chooses the test function by empirical optimization of the DV objective. In contrast, we choose h_t from the geometry of the mixture null in the RKHS. Specifically, we project the empirical mean embedding onto the convex hull of the component mean embeddings and use the normalized projection residual as the witness. This leads to the K-dimensional quadratic program in equation 5. The compactness, boundedness, and learnability properties corresponding to (A1) and (A3)–(A5) in Shekhar (2025) are satisfied in our setting. We do not impose their condition (A2), since the present RKHS betting class is not assumed to recover $\mathrm{KL}_{\inf}(P, \mathcal{P}_0)$. Consequently, the asymptotic quantity studied here is $\gamma^*_\varepsilon(P)$. For the mixture null considered here, the relevant computations reduce to a finite collection of component expectations, a K-dimensional quadratic program, and a one-dimensional betting problem.

C Backblaze SMART telemetry data

We use the Q1 2025 Backblaze Drive Stats dataset, covering January 1–March 31, 2025, downloaded from the official Backblaze Hard Drive Test Data repository: https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data. The dataset contains daily snapshots of Self-Monitoring, Analysis, and Reporting Technology (SMART) attributes for hard drives operating in Backblaze data centers. Each observation corresponds to a particular drive on a particular day.

We use the raw values of the following seven SMART attributes:

- **SMART 5 (Reallocated Sector Count):** number of sectors that have been reallocated because of read/write errors;

- **SMART 9 (Power-On Hours):** cumulative number of hours the drive has been powered on;
- **SMART 187 (Reported Uncorrectable Errors):** number of errors that could not be corrected using hardware error-correction mechanisms;
- **SMART 188 (Command Timeout):** number of aborted operations caused by command timeouts;
- **SMART 194 (Temperature):** operating temperature of the drive;
- **SMART 197 (Current Pending Sector Count):** number of unstable sectors awaiting possible reallocation; and
- **SMART 198 (Offline Uncorrectable Sector Count):** number of sectors with uncorrectable read/write errors.

For each of the seven SMART attributes, we first apply the transformation

$$x \mapsto \operatorname{sign}(x)\log(1+|x|),$$

which reduces the strong right-skewness of the raw SMART measurements while accommodating possible signed values. Each transformed feature is then standardized using the mean and standard deviation computed from the reference sample only, and the same transformation and scaling are applied to the held-out healthy and near-failure observations.

SeqGoF is implemented with a Gaussian radial basis function kernel,

$$k(x,y) = \exp\left\{-\frac{\|x-y\|_2^2}{2h^2}\right\},$$

with bandwidth $h = \sqrt{d} = \sqrt{7}$. Observations with missing values in any of the seven selected SMART attributes are excluded.

D FINITE-SAMPLE VALIDITY WITH REFERENCE SAMPLES

This appendix gives an optional correction to the empirical-reference implementation of Algorithm 1 for applications requiring a finite-sample anytime-valid guarantee. The kernel projection and witness construction are unchanged; only the calibration, the associated betting choice, and the rejection threshold are adjusted.

Theorem 4 (Finite-sample validity with reference samples)**.** *Fix $0 < \eta < \alpha < 1$ and a kernel satisfying $\sup_x k(x,x) \le 1$. Let $X_1^{(k)}, \dots, X_m^{(k)}$ be i.i.d. from F_k for each k, independent of the online stream. The kernel and any preprocessing are fixed independently of these samples and the stream. Let h_t be the witness obtained from Algorithm 1 using the empirical reference embeddings and Gram matrix, and set*

$$\rho_m(\eta) := \frac{1+\sqrt{2\log(K/\eta)}}{2\sqrt{m}}, \qquad \widehat{m}_t := \max_{1\le k\le K} \frac{1}{m}\sum_{j=1}^{m} h_t(X_j^{(k)}),$$
$$m_t^+ := \min\{1, \widehat{m}_t + \rho_m(\eta)\}, \quad g_t^+ := h_t - m_t^+. \tag{21}$$

Choose λ_t^+ by equation 8 with (m_t, g_t) replaced by (m_t^+, g_t^+), and initialize $h_0 \equiv 1/2$, $\lambda_0^+ = 0$, and $M_0^+ = 1$. Define

$$M_t^+ := \prod_{s=1}^{t}\{1+\lambda_{s-1}^+ g_{s-1}^+(Y_s)\}, \qquad \tau_{\alpha,\eta}^+ := \inf\{t \ge 1 : M_t^+ \ge 1/(\alpha-\eta)\}.$$

Then, for an i.i.d. online stream from any $P_0 \in \mathcal{P}_0$,

$$\sup_{P_0\in\mathcal{P}_0} \mathbb{P}_{P_0}\big(\tau_{\alpha,\eta}^+ < \infty\big) \le \alpha,$$

where probability includes both the reference samples and the online stream.

The proof below bounds the calibration error uniformly over the RKHS witness class, so the same reference samples may be reused at every round. With $\eta = \alpha/2$, the rejection threshold is $2/\alpha$; the per-round correction equation 21 is also essential. The result establishes validity of the corrected sample-based test; the power and stopping-time guarantees in Section 3.2 concern the oracle procedure.

Proof. Write $\mathcal{R} = \sigma(X_j^{(k)} : 1 \le j \le m, 1 \le k \le K)$ and $\mathcal{G}_t = \mathcal{R} \vee \mathcal{F}_t$. For each component, let $D_k = \|\widehat{\mu}_{F_k} - \mu_{F_k}\|_{\mathcal{H}}$. Independence within each reference sample and the kernel bound give

$$\mathbb{E}[D_k^2] = \frac{\mathbb{E}_{F_k}[k(X,X)] - \|\mu_{F_k}\|_{\mathcal{H}}^2}{m} \le \frac{1}{m}, \qquad \mathbb{E}[D_k] \le m^{-1/2}.$$

Replacing one reference observation changes D_k by at most $2/m$. McDiarmid's inequality therefore yields

$$\mathbb{P}\{D_k > m^{-1/2} + u\} \le \exp(-mu^2/2), \qquad u > 0.$$

Taking $u = \sqrt{2\log(K/\eta)/m}$ and a union bound shows that the reference-measurable event

$$A := \left\{ \max_{1 \le k \le K} D_k \le 2\rho_m(\eta) \right\} \quad \text{satisfies} \quad \mathbb{P}(A) \ge 1 - \eta.$$

Every empirical witness has the form $h_t = (1 + f_t)/2$, where $\|f_t\|_{\mathcal{H}} \le 1$ (take $f_t = 0$ when the residual is zero). On A, the reproducing property and Cauchy–Schwarz give, simultaneously for all such witnesses and every component,

$$\left| \mathbb{E}_{F_k}[h_t(X)] - \frac{1}{m}\sum_{j=1}^{m} h_t(X_j^{(k)}) \right| = \frac{1}{2}\,|\langle f_t, \mu_{F_k} - \widehat{\mu}_{F_k}\rangle_{\mathcal{H}}| \le \rho_m(\eta).$$

This bound is uniform over the entire witness class, so adaptive selection using the same reference samples causes no additional error. Since $0 \le h_t \le 1$, on A we obtain

$$\sup_{P_0 \in \mathcal{P}_0} \mathbb{E}_{P_0}[h_t(X)] \le \min\{1, \widehat{m}_t + \rho_m(\eta)\} = m_t^+.$$

The betting constraint ensures $\varepsilon \le 1 + \lambda_{t-1}^+ g_{t-1}^+(Y_t) \le 1 + \lambda_{\max}$. Conditional on A, the reference samples are known at time zero, and under any i.i.d. null stream

$$\mathbb{E}_{P_0}\left[1 + \lambda_{t-1}^+ g_{t-1}^+(Y_t) \mid \mathcal{G}_{t-1}\right] \le 1 \quad \text{on } A.$$

Thus (M_t^+) is a nonnegative supermartingale under the conditional law given A. Ville's inequality gives $\mathbb{P}_{P_0}(\tau_{\alpha,\eta}^+ < \infty \mid A) \le \alpha - \eta$. Consequently,

$$\mathbb{P}_{P_0}(\tau_{\alpha,\eta}^+ < \infty) \le \mathbb{P}(A^c) + (\alpha - \eta)\mathbb{P}(A) \le \alpha.$$

The argument is uniform over $P_0 \in \mathcal{P}_0$. It establishes unconditional type-I error control, without asserting that M_t^+ is an unconditional supermartingale. □