ONLINE CHANGE-POINT DETECTION IN RKHS A NON-ASYMPTOTIC PERFORMANCE ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study online change-point detection based on the empirical regularized Pearson divergence in a reproducing kernel Hilbert space (RKHS). The proposed procedure compares a left reference window, representing the nominal regime, with a right test window, containing the most recent observations. Although the resulting sliding-window statistic admits a tractable closed-form expression, the main contribution of this paper is a non-asymptotic analysis of its sequential detection performance. Under a bounded-kernel assumption, we derive finite-sample concentration inequalities under the null and obtain explicit threshold calibrations guaranteeing either average-run-length control or uniform false-alarm control over an infinite horizon. Under the alternative, we establish non-asymptotic bounds on the per-time miss-detection probability and on the tail of the detection delay. The analysis highlights the roles of partial contamination of the test window, regularization, and the window lengths. Experiments support the theoretical findings.

1 INTRODUCTION

Detecting change-points in the statistical behavior of streaming time-series data is a central problem in signal processing, statistics, and machine learning (Tartakovsky et al., 2014; Xie et al., 2021). Such changes often correspond to critical events whose timely detection is essential for reliable monitoring and decision-making, with applications ranging from industrial systems (Chiang et al., 2012) and information systems (Tartakovsky et al., 2006) to communication networks (Lakhina et al., 2004), power networks (Chen et al., 2015), and social networks (Peel & Clauset, 2015).

Online change-point detection (CPD) aims to detect distributional changes as quickly as possible while controlling false alarms, a trade-off formalized by the theory of quickest CPD. Classical procedures such as the Shewhart chart (Shewhart, 1925; 2022), CUSUM (Page, 1954), and the Shiryaev–Roberts statistic (Shiryaev, 1963) typically rely on parametric models for the pre- and post-change distributions (Basseville et al., 1993; Tartakovsky et al., 2014). When these models are correctly specified, likelihood-ratio-based procedures enjoy well-understood optimality properties (Lorden, 1971; Moustakides, 1986; Pollak, 1985). Modern data streams, however, are often high-dimensional, multimodal, structured, or governed by distributions that are difficult to model parametrically, motivating nonparametric approaches based on generic measures of distributional discrepancy (Aminikhanghahi & Cook, 2017; Brodsky & Darkhovsky, 2013; Truong et al., 2020).

A prominent nonparametric approach casts CPD as a sequence of two-sample tests between consecutive segments of the stream. Kernel methods are particularly attractive in this setting because they embed distributions into a reproducing kernel Hilbert space (RKHS) (Harchaoui et al., 2013; 2008). The maximum mean discrepancy (MMD), for instance, compares kernel mean embeddings (Gretton et al., 2006; 2012; Kalinke & Gavioli-Akilagun, 2026), thereby avoiding restrictive distributional assumptions while handling multivariate data naturally.

Direct density-ratio estimation provides another major nonparametric paradigm for CPD, with applications to multivariate time series (Liu et al., 2013; Yamada et al., 2013; Wang et al., 2023; Ferrari et al., 2023) and graph-valued observations (Ferrari & Richard, 2020). Methods such as KLIEP (Sugiyama et al., 2008) and uLSIF (Kanamori et al., 2009; 2013) directly estimate density ratios and are closely related to the estimation of the Kullback–Leibler and Pearson χ^2 divergences, respectively (Nguyen et al., 2010; Sugiyama et al., 2012). The Pearson χ^2 divergence is particularly appealing for CPD. For small density-ratio perturbations, it agrees to second order

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.

with the Kullback–Leibler divergence, which governs the asymptotic detection delay of classical likelihood-ratio-based procedures (Lorden, 1971). Moreover, its quadratic variational formulation admits tractable least-squares estimation and yields a closed-form regularized RKHS statistic. The resulting inverse-operator weighting emphasizes changes along directions with small nominal second moments relative to MMD (Balasubramanian et al., 2021; Hagrass et al., 2024).

While recent work has established non-asymptotic performance guarantees for MMD-based online detectors (Kalinke & Gavioli-Akilagun, 2026), comparable results for Pearson χ^2-based CPD remain, to our knowledge, lacking. Related work has established asymptotic guarantees for kernel-based change-point analysis (Harchaoui et al., 2008) and non-asymptotic guarantees for regularized kernel two-sample tests, including principled threshold calibration and regularization (Hagrass et al., 2024). However, these results do not characterize the sequential performance of Pearson-based detectors. Existing online density-ratio methods have also been analyzed in the mean and mean-square sense (Ferrari et al., 2023), but without corresponding non-asymptotic sequential guarantees. In particular, explicit guarantees linking the Pearson detection statistic to average-run-length control, infinite-horizon false-alarm control, and detection delay, while accounting for window lengths and regularization, remain lacking. Consequently, these quantities have largely been calibrated and assessed empirically. This motivates our non-asymptotic analysis of the regularized Pearson statistic in an online sliding-window setting.

We address this gap through a regularized variational formulation of the Pearson χ^2 divergence in an RKHS. The resulting statistic admits a closed-form expression, connects directly to MMD and uLSIF, and enables a tractable non-asymptotic analysis of sequential detection performance. This yields the first such analysis for the Pearson χ^2 statistic, with the results extending directly to uLSIF through their equivalence. Our contributions are as follows:

- **RKHS Pearson statistic for online CPD.** We derive a closed-form expression for the regularized empirical Pearson χ^2 divergence and use it to construct an online sliding-window detector, while making explicit its connections to MMD and uLSIF.
- **Non-asymptotic false-alarm guarantees.** Under a bounded-kernel assumption, we derive finite-sample concentration bounds and explicit thresholds guaranteeing either a prescribed average run length or uniform false-alarm control over an infinite horizon.
- **Non-asymptotic detection guarantees and tuning guidelines.** Under the alternative, we bound the miss-detection probability and the tail of the detection delay. The analysis yields practical guidelines for choosing the window lengths, regularization parameter, and threshold, which are validated experimentally on synthetic and real high-dimensional data.

2 Online CPD with an RKHS Pearson Statistic

2.1 Problem statement

We consider a sequence of observations $Z_1, Z_2, \ldots$, arriving online, where each Z_t takes values in a measurable space $\mathcal{X} \subseteq \mathbb{R}^d$. We assume the existence of an unknown change time $t_0 \in \mathbb{N} \cup \{\infty\}$ such that $Z_t \sim P$ for $t \leq t_0$, and $Z_t \sim Q$ for $t > t_0$, where $P \neq Q$ are Borel probability measures on $\mathcal{X}$. We assume that the observations are mutually independent and identically distributed within each regime. The aim is to detect the occurrence of t_0 as soon as possible while avoiding premature false alarms. This yields the sequential hypothesis test:

$$\mathsf{H}_0 : Z_t \sim P \text{ for all } t \qquad \text{versus} \qquad \mathsf{H}_1 : \exists\, t_0 \text{ such that } Z_t \sim \begin{cases} P, & t \leq t_0, \\ Q, & t > t_0. \end{cases} \tag{1}$$

One principled approach to (1) is to cast CPD as a two-sample testing problem between consecutive segments of the stream. We therefore consider two adjacent windows of fixed lengths n and m. At each time step $t \geq n$, we define:

$$X_{1:n}^{(t)} := \{Z_{t-n+1}, \ldots, Z_t\} \qquad \text{and} \qquad Y_{1:m}^{(t)} := \{Z_{t+1}, \ldots, Z_{t+m}\}, \tag{2}$$

and evaluate their discrepancy using a divergence statistic $\mathcal{D}[X_{1:n}^{(t)} | Y_{1:m}^{(t)}]$. A change is declared at the detection index

$$T := \inf\left\{ t \geq n : \ \mathcal{D}[X_{1:n}^{(t)} | Y_{1:m}^{(t)}] > \nu_t \right\}, \tag{3}$$

for a suitable threshold $\nu_t > 0$, with the convention $\inf \varnothing = +\infty$. Since the statistic indexed by t uses observations up to Z_{t+m}, the detection index T corresponds to an alarm raised at time $T + m$.

Following standard sequential-detection criteria, we express the false-alarm conditions in terms of the detection index T:

- Average run length (ARL) control: $\mathbb{E}_\infty[T] \geq \gamma, \ \ \gamma \geq n$,
- Uniform false-alarm probability (UFAP) control: $\mathbb{P}_\infty(T < \infty) \leq \alpha, \ \ 0 < \alpha < 1$,

where $\mathbb{E}_\infty$ and $\mathbb{P}_\infty$ denote the expectation and the probability under the global null $t_0 = \infty$. Under either condition, for a change occurring at time t_0, the goal is to minimize the detection delay

$$D := (T + m - t_0 - 1)_+. \tag{4}$$

Thus, $D = 0$ when the alarm is raised upon observing the first post-change sample Z_{t_0+1}.

2.2 RKHS Pearson statistic

Assume that Q is absolutely continuous with respect to P, and let $r = dQ/dP$ denote the density ratio. As recalled in Appendix A.1.1, the Pearson χ^2 divergence admits the variational representation

$$\chi^2(P|Q) = \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}_Q[f(X)] - \mathbb{E}_P\left[f(X) + \frac{1}{2} f(X)^2 \right] \right\}, \tag{5}$$

where $\mathcal{F}$ is a suitable class of real-valued measurable functions on $\mathcal{X}$. The ordering $\chi^2(P|Q)$ is used to match the empirical notation in (6). The supremum in (5) is attained at $f^*(x) = r(x) - 1$, connecting this formulation to density-ratio estimation. The general ϕ-divergence derivation and the connections with uLSIF and MMD are given in Appendix A.1.2.

Since $r(x)$ is unknown, we restrict f to a function class that is rich enough to approximate $f^*(x)$, yet structured enough to allow tractable computation. To this end, we restrict the variational class in (5) to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ with kernel $k(x, x')$.

Assumption 2.1. *The kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite and continuous, and bounded in the sense there exists $K > 0$ such that $\sup_{x \in \mathcal{X}} k(x, x) \leq K$.*

Given two independent samples $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_m \sim Q$, each consisting of i.i.d. observations, we construct an empirical counterpart of (5) by replacing expectations with sample averages, restricting the supremum to $\mathcal{H}_k$, and adding Tikhonov regularization,

$$\hat{\chi}^2_\rho[X_{1:n}|Y_{1:m}] = \sup_{f \in \mathcal{H}_k} \left\{ \frac{1}{m} \sum_{j=1}^{m} f(Y_j) - \frac{1}{n} \sum_{i=1}^{n} \left[f(X_i) + \frac{1}{2} f(X_i)^2 \right] - \frac{\rho}{2} \|f\|^2_{\mathcal{H}_k} \right\}, \tag{6}$$

where $\rho > 0$ is a regularization parameter. To derive a closed-form expression for (6), we rewrite the estimator in terms of RKHS inner products. Define the empirical mean embeddings

$$\hat{\mu}_P := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i), \qquad \hat{\mu}_Q := \frac{1}{m} \sum_{j=1}^{m} k(\cdot, Y_j), \qquad \hat{\Delta} := \hat{\mu}_Q - \hat{\mu}_P, \tag{7}$$

and the empirical uncentered second-moment operator

$$\hat{T}_P := \frac{1}{n} \sum_{i=1}^{n} k(\cdot, X_i) \otimes k(\cdot, X_i). \tag{8}$$

For any positive-definite self-adjoint operator M, we use the notation $\|f\|^2_M := \langle f, Mf \rangle_{\mathcal{H}_k}$.

Lemma 2.1 (Closed-form RKHS Pearson statistic)**.** *Let $\hat{A} := \hat{T}_P + \rho I$ with $\rho > 0$. The optimization problem* (6) *admits the unique maximizer*

$$\hat{f}^*_\rho = \hat{A}^{-1} \hat{\Delta} = (\hat{T}_P + \rho I)^{-1} \hat{\Delta}, \qquad \hat{\chi}^2_\rho[X_{1:n}|Y_{1:m}] = \frac{1}{2} \|\hat{\Delta}\|^2_{\hat{A}^{-1}}. \tag{9}$$

The derivation is given in Appendix A.2. For comparison, consider the rescaled empirical squared MMD statistic

$$\widehat{\mathrm{MMD}}^2[X_{1:n}|Y_{1:m}] := \frac{1}{2}\|\hat{\Delta}\|^2_{\mathcal{H}_k}. \tag{10}$$

Thus, the two statistics differ through the data-dependent operator weighting $\hat{A}^{-1}$. Replacing this weighting in (9) by the identity recovers the MMD statistic in (10). This operator dependence is one of the main difficulties in the non-asymptotic analysis of $\hat{\chi}^2_\rho$ developed below.

3 False-alarm control

To calibrate the detection threshold, we analyze the empirical statistic $\hat{\chi}^2_\rho$ under the global null hypothesis $t_0 = \infty$, where the observations are i.i.d. according to P. The tail bound established in Theorem 3.1 provides the starting point and is then specialized to the two false-alarm criteria introduced in Section 2.1, namely ARL and UFAP control. All proofs are deferred to Appendix A.4.

Theorem 3.1. *Under Assumption 2.1, let $X_{1:n}$ and $Y_{1:m}$ be two independent samples of i.i.d. observations drawn from the same distribution P. Define the constants*

$$D := 4K\Big(\frac{1}{n} + \frac{1}{m}\Big), \qquad E := \frac{2}{3}\sqrt{K}\,\max\Big\{\frac{1}{n}, \frac{1}{m}\Big\}. \tag{11}$$

Then, for any regularization parameter $\rho > 0$ and threshold $\nu > 0$,

$$\mathbb{P}\big(\hat{\chi}^2_\rho[X_{1:n}|Y_{1:m}] > \nu\big) \le 2\exp\left(-\frac{\rho\nu}{D + E\sqrt{2\rho\nu}}\right). \tag{12}$$

Theorem 3.1 yields a constant threshold ensuring a prescribed ARL.

Corollary 3.1 (RKHS ARL control)**.** *Under the setting of Theorem 3.1, for any target ARL $\gamma \ge n$, it is sufficient to choose*

$$\nu \ge \frac{1}{2\rho}\left[E\log(4\gamma) + \sqrt{(E\log(4\gamma))^2 + 2D\log(4\gamma)}\right]^2, \tag{13}$$

in which case $\mathbb{E}_\infty[T] \ge \gamma$.

For UFAP control, consider the sliding-window statistic $\hat{\chi}^2_{\rho,t} := \hat{\chi}^2_\rho[X^{(t)}_{1:n}|Y^{(t)}_{1:m}]$. A constant threshold such as (13) provides ARL control on the detection-index scale but does not by itself guarantee uniform control of the false-alarm probability over an infinite horizon.

Corollary 3.2 (RKHS UFAP control)**.** *Assume the setting of Theorem 3.1. Fix $\alpha \in (0,1)$ and define the time-varying thresholds $\{\nu_t\}_{t\ge n}$ by*

$$\nu_t \ge \frac{1}{2\rho}\Big[EL_t + \sqrt{(EL_t)^2 + 2DL_t}\Big]^2, \qquad L_t := \log\Big(\frac{2t(t+1)}{\alpha}\Big). \tag{14}$$

Then the detection index $T := \inf\{t \ge n : \hat{\chi}^2_{\rho,t} > \nu_t\}$ satisfies $\mathbb{P}_\infty(T < \infty) \le \alpha$.

4 Miss-detection probability and detection delay

The preceding section established false-alarm guarantees under the global null. We now analyze the detector under the alternative hypothesis, where a change occurs at an unknown time t_0. Our goal is to derive bounds on the per-time miss-detection probability and the detection delay D defined in (4). All technical derivations and proofs are deferred to Appendix A.5.

After a change, the right test window may contain observations from both P and Q. At time t, let τ_t denote the number of pre-change observations remaining in the test window, and let ϵ_t denote the corresponding contamination level, defined as

$$\epsilon_t := \frac{\tau_t}{m}. \tag{15}$$

To describe the post-change signal, we introduce the population RKHS shift, the uncentered second-moment operator, and its regularized version,

$$\Delta := \mathbb{E}_Q[k(\cdot,Y)] - \mathbb{E}_P[k(\cdot,X)], \qquad T_P := \mathbb{E}_P[k(\cdot,X)\otimes k(\cdot,X)], \qquad A := T_P + \rho I. \tag{16}$$

As shown in Appendix A.3, the population regularized Pearson divergence can be written as

$$\chi^2_\rho(P|Q) = \frac{1}{2}\|\Delta\|^2_{A^{-1}}. \tag{17}$$

It is therefore convenient to express the post-change analysis in terms of $\|\Delta\|_{A^{-1}}$, which represents the population separation between P and Q associated with the statistic.

When the test window is contaminated, only a fraction $1-\epsilon_t$ of its observations is drawn from Q. The corresponding population mean-embedding shift is therefore reduced from Δ to $(1-\epsilon_t)\Delta$. As shown in Appendix A.5.1, this attenuation naturally appears through the effective signal

$$S_{\epsilon_t} := (1-\epsilon_t)\|\Delta\|_{A^{-1}}. \tag{18}$$

This quantity represents the post-change signal seen by the detector at time t. It is zero when the test window is entirely pre-change and increases toward $\|\Delta\|_{A^{-1}}$ as the contamination vanishes.

4.1 Miss-detection probability

Unless stated otherwise, we consider a constant threshold ν over the time scale relevant to detection. Define the miss-detection event and its per-time probability by

$$\mathcal{E}_{\mathrm{miss},t} := \big\{\|\hat{\Delta}_t\|_{\hat{A}_t^{-1}} \le \sqrt{2\nu}\big\}, \qquad p_{\mathrm{miss},t} := \mathbb{P}(\mathcal{E}_{\mathrm{miss},t}). \tag{19}$$

The following theorem provides a non-asymptotic upper bound on $p_{\mathrm{miss},t}$ while making explicit the distinct contributions of test-window fluctuations and perturbations affecting the empirical metric estimated from the reference window. The proof and the corresponding explicit constants are provided in Appendix A.5.2.

Theorem 4.1 (Per-time miss probability)**.** *Fix a time index t with $\epsilon_t \in (0,1)$. Assume the technical condition $0 < \rho \le K$. Define the effective signal margin a_t and deterministic floors c_0 and d_0 as*

$$a_t := \frac{S_{\epsilon_t} - \sqrt{2\nu}}{4}, \qquad c_0 := \sqrt{\frac{K}{\rho}}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right), \qquad d_0 := 2\sqrt{\frac{K\epsilon_t}{\rho m}}, \tag{20}$$

and the residual effective signal margins

$$r_t := (a_t - c_0)_+, \qquad r'_t := \frac{1}{\epsilon_t}(a_t - d_0)_+. \tag{21}$$

For $u \ge 0$, define

$$\Psi_\rho(u) := \min\left\{\frac{\rho^4\|\Delta\|^2_{\mathcal{H}_k}u^2}{K^4(K+\rho)}, \frac{\rho^3\|\Delta\|_{\mathcal{H}_k}u}{K^3\sqrt{K+\rho}}\right\}. \tag{22}$$

Under the sample-size condition (A.108) *in Appendix A.5.1, there exist strictly positive universal constants κ_1 and κ_2 such that*

$$\begin{aligned} p_{\mathrm{miss},t} \le \min\Big\{1,\ \kappa_1\Big[&\exp\left(-\kappa_2\frac{m\rho a_t^2}{K}\right) + \exp\left(-\kappa_2\frac{m\rho a_t^2}{K\epsilon_t}\right) \\ &+ \exp\left(-\kappa_2\, n\Psi_\rho(r_t)\right) + \exp\left(-\kappa_2\, n\Psi_\rho(r'_t)\right)\Big]\Big\}. \end{aligned} \tag{23}$$

The first two terms in (23) arise from fluctuations of the empirical mean embeddings in the test window. Their exponents contain an explicit linear factor in the test-window length m, with the second term additionally accounting for the contamination level ϵ_t. The last two terms account for empirical embedding errors and perturbations of the reference-dependent inverse metric $\hat{A}^{-1}$ from the reference window. Their exponents depend linearly on the reference-window length n and are governed by the Bernstein-type rate Ψ_ρ, which exhibits a quadratic regime for small residual margins and a linear regime for larger ones.

The two window lengths therefore play distinct roles. Increasing n stabilizes the reference metric and lowers the deterministic floor c_0, whereas increasing m improves concentration in the test window but slows the removal of pre-change samples through ϵ_t. This suggests using a relatively short test window together with a longer reference window.

The preceding bound also highlights the additional difficulty introduced by the data-dependent metric $\hat{A}^{-1}$. For MMD, this metric is replaced by the identity, so the corresponding metric-perturbation terms disappear.

Corollary 4.1 (MMD specialization). *For the MMD detector, the metric-perturbation terms in Theorem 4.1 vanish, while reference-embedding fluctuations remain. For $\epsilon_t \in (0,1)$, the per-time miss probability satisfies*

$$p_{\mathrm{miss},t} \le \min\left\{1,\ \exp\left(-\frac{\bar{r}_t^2}{2K(\frac{1}{n}+\frac{1}{m})}\right) + \exp\left(-\frac{m\epsilon_t(\bar{r}_t')^2}{4K}\right)\right\}, \tag{24}$$

where the MMD signal $\bar{S}_{\epsilon_t}$, margin $\bar{a}_t$, estimation floors $\bar{c}_0$ and $\bar{d}_0$, and residual margins $\bar{r}_t$ and $\bar{r}_t'$ are defined in Appendix A.5.3.

Unlike the Pearson bound, (24) contains no metric-perturbation terms because MMD uses the fixed identity metric rather than the data-dependent metric $\hat{A}^{-1}$. Nevertheless, the reference-window length n enters explicitly through the estimation of μ_P. For fixed m and ν, increasing n lowers the estimation floor $\bar{c}_0$, thereby increasing $\bar{r}_t$, and reduces the $1/n$ contribution in the first exponent's denominator. Both effects sharpen the bound, with diminishing gains as n tends to ∞, when the denominator approaches $2K/m$.

For a more direct interpretation of Theorem 4.1, and to obtain a tractable form for the detection-delay analysis in the next subsection, the following corollary isolates the two dominant contributions in the normalized-kernel regime.

Corollary 4.2 (Operational miss-probability bound). *Under the setting of Theorem 4.1, let $K = 1$ and assume that $r_t, r_t' > 0$ satisfy*

$$r_t \le \frac{\sqrt{1+\rho}}{2\rho\|\Delta\|_{\mathcal{H}_k}}, \qquad r_t' \le \frac{\sqrt{1+\rho}}{8\rho\|\Delta\|_{\mathcal{H}_k}}. \tag{25}$$

Then there exist strictly positive universal constants κ_3 and κ_4 such that

$$p_{\mathrm{miss},t} \le \min\left\{1,\ \kappa_3\Big[\exp\left(-\kappa_4\frac{m\rho a_t^2}{\max\{\epsilon_t,1/2\}}\right) + \exp\left(-\kappa_4\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n(\min\{r_t,r_t'\})^2}{1+\rho}\right)\Big]\right\}.$$

The proof and explicit constants are given in Appendix A.5.4. The first term is associated with fluctuations in the test window and contains an explicit linear factor in m, whereas the second arises from estimating the reference metric and contains an explicit linear factor in n.

The MMD specialization of this bound is also given in Appendix A.5.5.

4.2 Detection delay

We now turn from the per-time miss probability to the tail of the detection delay. Fix the change time t_0 and let $t = t_0 - m + 1$ denote the first split index whose test window contains a post-change observation. Assume $t \ge n$. For any integer $d \ge 0$,

$$\mathbb{P}(D > d) = \mathbb{P}(T > t + d) \le \mathbb{P}\left(\bigcap_{k=0}^{d} \mathcal{E}_{\mathrm{miss},t+k}\right) \le \min_{0\le k\le d} p_{\mathrm{miss},t+k} \le p_{\mathrm{miss},t+d}. \tag{26}$$

This bound follows from set inclusion alone and does not require independence across time. Combining (26) with Corollary 4.2 yields the following result.

Theorem 4.2 (Detection-delay tail). *Under the setting of Corollary 4.2, assume $t_0 - m + 1 \ge n$ and let $0 \le d \le m-2$ be an integer such that the range conditions* (25) *hold at time $t+d$. Then there exist strictly positive universal constants κ_3 and κ_4 such that*

$$\begin{aligned}\mathbb{P}(D > d) \le \min\Big\{1,\ \kappa_3\Big[&\exp\left(-\kappa_4\frac{m\rho a_{t+d}^2}{\max\{\epsilon_{t+d},1/2\}}\right)\\ &+ \exp\left(-\kappa_4\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n(\min\{r_{t+d},r_{t+d}'\})^2}{1+\rho}\right)\Big]\Big\}.\end{aligned} \tag{27}$$

The proof is given in Appendix A.5.6. During the warm-up phase, the contamination level ϵ_{t+d} decreases as pre-change observations leave the test window, while a_{t+d} and the residual margins increase. Thus, within the range where Corollary 4.2 applies, the bound in (27) tightens as the test window becomes increasingly post-change.

The corresponding detection-delay bound for the MMD detector is given in Appendix A.5.7.

5 DETECTOR-TUNING GUIDELINES

Our non-asymptotic bounds clarify how the threshold ν, window lengths n and m, and regularization parameter ρ affect false alarms, miss probability, and detection delay. Since the bounds are conservative, they primarily provide tuning guidelines rather than precise operating points. Calibration relies on nominal data from P, whereas performance under the unknown alternative Q is assessed for prescribed minimum signal strengths.

Threshold ν (false-alarm control). Corollary 3.1 provides an explicit, distribution-free threshold guaranteeing an ARL of at least γ. However, this conservative bound may yield a substantially larger realized ARL. A sharper operating point can be obtained through Monte Carlo calibration on nominal data, using the conditional false-alarm probability associated with each reference window. See Appendix A.6 for details. The resulting fixed-reference ARL curve is inverted to select a threshold targeting γ. Unlike the analytical threshold, this empirical calibration does not provide a distribution-free guarantee.

Reference window n (metric estimation). Increasing n improves estimation of the nominal embedding and second-moment operator. In Theorem 4.1, it strengthens the metric-perturbation bounds and reduces the detection floor c_0. However, gains diminish once the $1/\sqrt{m}$ contribution dominates. Choosing $n > m$ therefore provides a useful operating regime without requiring $n \gg m$.

Test window m (miss–delay trade-off). Increasing m improves concentration but also lengthens the transition period during which the test window contains pre-change observations. Indeed, its contamination level at lag ℓ is $\epsilon_{t+\ell} = (m - 1 - \ell)_+/m$. Thus, m should balance detection reliability against responsiveness, favoring the smallest window that achieves the desired sensitivity at the prescribed minimum signal strength.

Regularization ρ (metric conditioning). Too small a ρ amplifies estimation errors through the inverse operator, whereas excessive regularization drives the statistic toward an MMD-like limit, weakening its directional weighting. A practical strategy is to select ρ from the nominal spectrum of $\hat{T}_P$, for instance near its spectral knee, balancing numerical stability against preservation of informative directions.

6 NUMERICAL EXPERIMENTS

We evaluate the detector on four synthetic reference distributions to assess ARL calibration and the parameter trade-offs predicted by the theory. A complementary real-data evaluation on MNIST, including threshold calibration and sensitivity to n, m, and ρ, is reported in Appendix A.7. All experiments use the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|_2^2/(2\sigma^2))$, with bandwidth selected by the median heuristic (Schölkopf & Smola, 2002).

The experimental protocols follow (Kalinke & Gavioli-Akilagun, 2026; Wei & Xie, 2026).

6.1 DISTRIBUTIONAL ROBUSTNESS OF ARL CALIBRATION

We consider four nominal distributions in dimension $d = 20$, with different tail behaviors and supports: a **Gaussian mixture** (GM), $\frac{7}{8}\mathcal{N}(\frac{1}{4}\mathbf{1}_{20}, I_{20}) + \frac{1}{8}\mathcal{N}(\mathbf{0}_{20}, I_{20})$; and distributions with i.i.d. coordinates: **Laplace** $\mathrm{Lap}(\frac{1}{2}, \frac{1}{4})$ (location, scale), **Exponential** $\mathrm{Exp}(\frac{4}{5})$ (mean $\frac{4}{5}$), and **Uniform** $\mathcal{U}(-\frac{1}{2}, \frac{3}{2})$. Thresholds are calibrated using nominal data only, with $n = 80$, $m = 40$, and $\rho = 10^{-2}$.

Table 1 reports the realized ARL and relative error for targets $\gamma \in \{100, 200, 500, 1000\}$. Calibration is generally close to the target for GM, Laplace, and Uniform references, despite their different statistical properties. Variability increases at larger targets, as expected when false alarms become rare. For example, the GM reference yields 1032 ± 481 at $\gamma = 1000$.

The Exponential reference exhibits more conservative behavior, particularly for $\gamma \geq 200$, reaching 820 ± 70 at $\gamma = 500$. Strongly skewed references may therefore warrant a larger calibration budget, a finer threshold grid, or an adapted kernel bandwidth. Overall, empirical calibration provides

Table 1: Realized ARL ($\pm$ standard deviation over four independent evaluation sets) and relative error of the mean. Calibration uses nominal data only.

	Target ARL γ			
Reference P	100	200	500	1000
GM	82±19 (−18%)	178±41 (−11%)	558±226 (+12%)	1032±481 (+3%)
Laplace	91±24 (−9%)	224±115 (+12%)	503±285 (+1%)	1055±285 (+6%)
Exponential	108±12 (+8%)	267±34 (+34%)	820±70 (+64%)	1357±149 (+36%)
Uniform	84±10 (−16%)	187±51 (−6%)	416±99 (−17%)	975±179 (−2%)

Table 2: Empirical miss probability ($\pm$ standard deviation over five independent evaluation sets) versus reference length n, with $m = 30$.

Reference P (s)	$n = 15$	$n = 30$	$n = 60$	$n = 120$	$n = 240$
GM (1.4)	.533±.058	.251±.031	.074±.013	.047±.015	.025±.007
Laplace (0.65)	.296±.033	.042±.014	.001±.002	.000±.000	.000±.000
Exp (1.15)	.595±.043	.212±.039	.063±.018	.015±.010	.003±.002
Uniform (0.85)	.409±.032	.195±.046	.064±.022	.030±.017	.012±.003

useful operating points, although its accuracy may deteriorate in rare-event regimes. The analytical threshold remains available when distribution-free false-alarm guarantees are required.

6.2 Numerical verification of parameter trade-offs

We assess the predictions of Section 5 using the same four reference distributions. The post-change law Q differs from P by a mean shift of magnitude s along the first coordinate, leaving the remaining coordinates unchanged. The shift is selected separately for each experiment and distribution to obtain comparable detectability, and its value is reported in the tables. Moderate shifts are used for miss-probability experiments, whereas larger shifts ensure reliable in-window detection in the delay experiment. Thresholds are calibrated under P alone, and $\rho = 10^{-2}$ unless explicitly varied.

Reference window n. Table 2 reports the empirical miss probability with $m = 30$ fixed. Increasing n consistently improves detection across all four distributions. The largest gains occur at smaller reference sizes, while improvements progressively diminish as n increases. This agrees with the theoretical prediction that a longer reference window improves estimation of the nominal metric and reduces the detection floor $c_0 \propto 1/\sqrt{n} + 1/\sqrt{m}$. Once n becomes sufficiently large, the $1/\sqrt{m}$ contribution dominates, limiting the benefit of further increasing n.

Test window m. We evaluate the delay in a high-signal regime where in-window detection occurs with probability approximately one, with $n = 120$ and a per-window false-alarm probability of 0.01. Table 3 shows that the absolute delay increases with m for every reference distribution. However, the relative delay D/m decreases, indicating that longer windows require a smaller fraction of post-change samples to trigger detection. Since the miss probability is already negligible in this regime, increasing m mainly adds delay. These results support choosing the smallest test window that ensures reliable detection.

Regularization ρ. Table 4 reports the empirical miss probability for $n = 80$ and $m = 20$. Large values of ρ consistently degrade detection, in agreement with the predicted loss of directional weighting as the statistic approaches its MMD-like limit.

In contrast, the miss probability remains relatively stable for small ρ. Although the theoretical bounds indicate greater sensitivity to metric-estimation errors as ρ decreases, threshold calibration under P may partly compensate for this amplification. No pronounced loss of calibrated detection power is observed in this regime. Thus, for these synthetic distributions, small regularization is beneficial, provided numerical stability is maintained.

Table 3: Detection delay versus test length m, with $n = 120$ and in-window detection probability close to 1. Signal magnitude s in parentheses.

Reference P (s)	$m = 12$	$m = 16$	$m = 20$	$m = 30$	$m = 40$
GM (1.8)	7.3±.5	9.1±.2	11.2±.5	14.5±1.5	16.4±.9
Laplace (1.0)	5.4±.5	6.2±.4	7.2±.6	9.8±.6	11.0±.8
Exp (1.6)	7.1±.7	8.0±.5	9.9±.6	13.1±.7	15.6±1.5
Uniform (1.4)	6.0±.3	6.9±.6	8.7±.4	11.0±.3	12.0±1.1
D/m (Laplace)	0.45	0.39	0.36	0.33	0.28

Table 4: Empirical miss probability versus ρ, with $n = 80$ and $m = 20$.

Reference P (s)	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	3
GM (1.2)	.163±.023	.165±.019	.193±.037	.253±.024	.511±.036	.597±.038
Laplace (0.7)	.000±.000	.000±.000	.001±.001	.003±.004	.149±.030	.227±.039
Exp (1.0)	.101±.022	.110±.026	.137±.024	.408±.094	.795±.053	.811±.052
Uniform (0.55)	.418±.042	.423±.037	.461±.043	.474±.052	.518±.060	.541±.054

7 Conclusion and perspectives

We studied online CPD based on the empirical regularized Pearson divergence in an RKHS. We established a closed-form representation of the statistic, finite-sample concentration inequalities, and explicit thresholds ensuring false-alarm control. We also derived bounds on the miss-detection probability and the tail of the detection delay, providing a non-asymptotic performance characterization of the proposed sequential procedure.

Several extensions naturally follow from this work.

Finite-dimensional dictionary representation. An important computational direction is to restrict the RKHS optimization to a finite dictionary with controlled sparsity, reducing the problem to a finite-dimensional quadratic program. This may also enable sharper guarantees that exploit dictionary geometry, particularly the interplay between coherence, dictionary size, and the spectral properties of the Gram matrix.

Structured pre-change models. The distribution-free null thresholds derived here could be tightened when additional information about the pre-change distribution is available. For example, Gaussian pre-change models combined with Gaussian kernels allow kernel moments and variance proxies to be computed explicitly, potentially yielding less conservative thresholds.

Non-Euclidean input spaces. Although the present framework considers Euclidean observations, its kernel formulation naturally extends to Riemannian manifolds, Lie groups, graphs, and other structured domains through suitable positive-definite kernels. The concentration and false-alarm arguments extend to bounded kernels, while detection guarantees require appropriate conditions on the change signal. This opens the way to online CPD for structured data streams arising, for instance, in robotics and network monitoring.

Alternative divergences and function classes. The analytical approach is not restricted to Pearson χ^2 divergence or RKHSs. Potential extensions include other variational ϕ-divergences, integral probability metrics, and alternative function classes such as neural networks. A central challenge is to develop suitable concentration tools, while accounting for estimation and optimization errors, to retain explicit guarantees on false alarms and detection delays.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing the proposed methods and to review and improve the readability of the paper. Generative AI was not used for any other task requiring disclosure. In particular, all theoretical results and proofs are the authors' own work. We have reviewed all AI-assisted work: all AI-assisted code was tested by the authors, and all text edits were checked to preserve the technical content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical and advances our understanding of nonparametric online change-point detection. Although machine learning methods may have broad societal implications, we do not foresee any specific negative societal consequences arising directly from these results that warrant further discussion.

REPRODUCIBILITY STATEMENT

All assumptions underlying the theoretical results are stated in Section 2, and complete proofs are given in Appendix A. The evaluation protocol, data sets and all hyperparameters are described in Section 6 and Appendix A.7. The experiments use only synthetic data and the publicly available datasets. Source code to reproduce the results will be released upon acceptance.

REFERENCES

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51(2):339–367, 2017.

Krishnakumar Balasubramanian, Tong Li, and Ming Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 22(1):1–45, 2021.

Michele Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice hall Englewood Cliffs, 1993.

Emily Brodsky and Boris S Darkhovsky. *Nonparametric methods in change point problems*. Springer Science & Business Media, 2013.

Yu Christine Chen, Taposh Banerjee, Alejandro D Dominguez-Garcia, and Venugopal V Veeravalli. Quickest line outage detection and identification. *IEEE Transactions on Power Systems*, 31(1): 749–758, 2015.

Leo H Chiang, Evan L Russell, and Richard D Braatz. *Fault detection and diagnosis in industrial systems*. Springer Science & Business Media, 2012.

André Ferrari and Cédric Richard. Non-parametric community change-points detection in streaming graph signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5545–5549, 2020.

André Ferrari, Cédric Richard, Anthony Bourrier, and Ikram Bouchikhi. Online change-point detection with kernels. *Pattern Recognition*, 133:109022, 2023.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Omar Hagrass, Bharath K. Sriperumbudur, and Bing Li. Spectral regularized kernel two-sample tests. *Annals of Statistics*, 52(3):1076–1101, 2024.

Zaid Harchaoui, Eric Moulines, and Francis Bach. Kernel change-point analysis. *Advances in Neural Information Processing Systems*, 21, 2008.

Zaid Harchaoui, Francis Bach, Olivier Cappe, and Eric Moulines. Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Processing Magazine*, 30(4):87–97, 2013.

Florian Kalinke and Shakeel Gavioli-Akilagun. Optimal online change detection via random fourier features. In *Advances in Neural Information Processing Systems*, 2026.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10:1391–1445, 2009.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Computational complexity of kernel-based density-ratio estimation: A condition number analysis. *Machine Learning*, 90(3):431–460, 2013.

Anukool Lakhina, Mark Crovella, and Christophe Diot. Diagnosing network-wide traffic anomalies. *ACM SIGCOMM Computer Communication Review*, 34(4):219–230, 2004.

Song Liu, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

Gary Lorden. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics*, pp. 1897–1908, 1971.

George V Moustakides. Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, 14(4):1379–1387, 1986.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Leto Peel and Aaron Clauset. Detecting change points in the large-scale structure of evolving networks. In *AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Moshe Pollak. Optimal detection of a change in distribution. *Annals of Statistics*, pp. 206–227, 1985.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Walter A Shewhart. The application of statistics as an aid in maintaining quality of a manufactured product. *Journal of the American Statistical Association*, 20(152):546–548, 1925.

Walter A Shewhart. *Economic control of quality of manufactured product*. Barakaldo Books, 2022.

Albert N Shiryaev. On optimum methods in quickest detection problems. *Theory of Probability & Its Applications*, 8(1):22–46, 1963.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul Von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC press, 2014.

Alexander G Tartakovsky, Boris L Rozovskii, Rudolf B Blažek, and Hongjoong Kim. Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, 3(3):252–293, 2006.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

Xiuheng Wang, Ricardo Augusto Borsoi, Cédric Richard, and Jie Chen. Change point detection with neural online density-ratio estimator. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Song Wei and Yao Xie. Online kernel cusum for change-point detection, 2026. URL `https://arxiv.org/abs/2211.15070`.

Liyan Xie, Shaofeng Zou, Yao Xie, and Venugopal V Veeravalli. Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2):494–514, 2021.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural Computation*, 25(5): 1324–1370, 2013.

A Appendices

A.1 Pearson Divergence, uLSIF, and MMD

A.1.1 From ϕ-divergences to the Pearson variational form

Assume that Q is absolutely continuous with respect to P, and denote the corresponding density ratio by $r = dQ/dP$. The absolute continuity assumption guarantees that r exists as a genuine Radon–Nikodym derivative and is finite P-almost everywhere.

A broad family of discrepancies between P and Q is provided by ϕ-divergences. Given a proper convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that $\phi(1) = 0$, the ϕ-divergence between P and Q is defined as

$$\mathcal{D}_\phi(P|Q) = \int_{\mathcal{X}} \phi\left(\frac{dQ}{dP}(x)\right) dP(x) = \mathbb{E}_P[\phi(r(X))]\,. \tag{A.28}$$

The normalization $\phi(1) = 0$ ensures that $\mathcal{D}_\phi(P|P) = 0$, and Jensen's inequality applied to the convex ϕ yields $\mathcal{D}_\phi(P|Q) \geq 0$, so that $\mathcal{D}_\phi$ is a valid divergence though in general neither symmetric nor a metric.

By convexity, ϕ admits the Fenchel–Legendre representation

$$\phi(z) = \sup_{u\in\mathbb{R}} \big\{uz - \phi^*(u)\big\}, \tag{A.29}$$

where ϕ^* denotes the convex conjugate of ϕ. Applying this identity point-wise to $z = r(x)$, and interpreting the scalar variable u as a measurable function $f(x)$, yields the variational representation

$$\mathcal{D}_\phi(P|Q) = \sup_{f\in\mathcal{F}} \left\{\mathbb{E}_Q[f(X)] - \mathbb{E}_P[\phi^*(f(X))]\right\}, \tag{A.30}$$

where $\mathcal{F}$ is a suitable class of real-valued measurable functions on $\mathcal{X}$. This dual formulation is particularly useful because it replaces the explicit estimation of the density ratio r by the optimization of an empirical criterion involving only samples from P and Q.

In this work, we focus on the Pearson χ^2 divergence, which corresponds to $\phi(z) = \frac{1}{2}(z-1)^2$. This choice is particularly useful in practice because both ϕ and its convex conjugate admit simple quadratic forms. More precisely, we have $\phi^*(u) = u + \frac{1}{2}u^2$. Substituting this expression into the general variational representation gives

$$\chi^2(P|Q) = \sup_{f\in\mathcal{F}} \left\{\mathbb{E}_Q[f(X)] - \mathbb{E}_P\left[f(X) + \frac{1}{2}f(X)^2\right]\right\}. \tag{A.31}$$

The supremum in (A.31) is attained at $f^*(x) = r(x) - 1$. Therefore, estimating the optimizer f^* amounts to estimating the density ratio between Q and P. In the absolutely continuous case where P and Q admit densities p and q, respectively, this ratio can be written as

$$r(x) = \frac{q(x)}{p(x)}. \tag{A.32}$$

Since $r(x)$ is unknown, we restrict f to a function class that is rich enough to approximate $r(x) - 1$, yet structured enough to allow tractable computation. To this end, we embed (A.31) into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ with kernel $k(x, x')$.

A.1.2 Relation to uLSIF and MMD

The Pearson variational formulation is closely related to uLSIF. The latter estimates r by minimizing the square loss (Kanamori et al., 2009):

$$\mathrm{uLSIF}(P, Q) = \inf_{g \in \mathcal{H}_k} \tfrac{1}{2}\mathbb{E}_P\left[(g(X) - r(X))^2\right] \tag{A.33}$$

typically with an additional ℓ_2-regularization term. Setting $g = f + 1$ recovers the negated χ^2 dual in (A.31) up to the additive constant.

MMD, by contrast, is an integral probability metric (Gretton et al., 2012),

$$\mathrm{MMD}(P, Q) = \sup_{\|f\|_{\mathcal{H}_k} \le 1} \left\{\mathbb{E}_Q[f(X)] - \mathbb{E}_P[f(X)]\right\}. \tag{A.34}$$

Its statistic is therefore governed by the difference between kernel mean embeddings, whereas the Pearson χ^2 formulation targets the density ratio through a quadratic variational problem. At the empirical level, this difference appears through the operator weighting $\hat{A}^{-1}$ in (9), compared with the identity weighting in (10).

A.2 Empirical regularized Pearson divergence

Using the definitions of $\hat{\mu}_P$, $\hat{\mu}_Q$, $\hat{\Delta}$, and $\hat{T}_P$ in (7) and (8), the reproducing property yields

$$\langle f, \hat{T}_P f\rangle_{\mathcal{H}_k} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)^2, \qquad \langle f, \hat{\Delta}\rangle_{\mathcal{H}_k} = \frac{1}{m}\sum_{j=1}^{m} f(Y_j) - \frac{1}{n}\sum_{i=1}^{n} f(X_i). \tag{A.35}$$

Therefore, (6) can be rewritten as

$$\hat{\chi}^2_\rho[X_{1:n}|Y_{1:m}] = \sup_{f \in \mathcal{H}_k}\left\{\langle f, \hat{\Delta}\rangle_{\mathcal{H}_k} - \frac{1}{2}\langle f, \hat{A} f\rangle_{\mathcal{H}_k}\right\}. \tag{A.36}$$

Since $\hat{A}$ is self-adjoint and strictly positive for $\rho > 0$, the objective is strictly concave. Its unique maximizer $\hat{f}^*_\rho$ satisfies

$$\hat{A}\,\hat{f}^*_\rho = \hat{\Delta}, \qquad \text{i.e.,} \qquad \hat{f}^*_\rho = \hat{A}^{-1}\hat{\Delta} = (\hat{T}_P + \rho I)^{-1}\hat{\Delta}. \tag{A.37}$$

Substituting $\hat{f}^*_\rho$ into (A.36) yields

$$\hat{\chi}^2_\rho[X_{1:n}|Y_{1:m}] = \frac{1}{2}\langle \hat{\Delta}, \hat{A}^{-1}\hat{\Delta}\rangle_{\mathcal{H}_k}, \tag{A.38}$$

which proves Lemma 2.1.

A.3 Population regularized Pearson divergence

Consider the variational form of the RKHS-regularized Pearson χ^2 divergence:

$$\chi^2_\rho(P|Q) = \sup_{f \in \mathcal{H}_k}\left\{\mathbb{E}_Q[f(Y)] - \mathbb{E}_P\left[f(X) + \frac{1}{2}f(X)^2\right] - \frac{\rho}{2}\|f\|^2_{\mathcal{H}_k}\right\} \tag{A.39}$$

where $\mathcal{H}_k$ is the RKHS associated with kernel $k(\mathbf{x}, \mathbf{x}')$, assumed bounded as $\sup_{\mathbf{x}\in\mathcal{X}} k(\mathbf{x}, \mathbf{x}) \le K$, and $\rho > 0$ is a regularization parameter.

Define the mean embeddings

$$\mu_P := \mathbb{E}_P[k(\cdot, X)], \qquad \mu_Q := \mathbb{E}_Q[k(\cdot, Y)], \qquad \Delta := \mu_Q - \mu_P \in \mathcal{H}_k, \tag{A.40}$$

and the (uncentered) second-moment operator

$$T_P := \mathbb{E}_P\big[\, k(\cdot, X) \otimes k(\cdot, X)\,\big] \tag{A.41}$$

where $\otimes$ denotes the Hilbert-space outer product, meaning that, for any $f \in \mathcal{H}_k$,

$$(T_P f) = \mathbb{E}_P\Big[\big\langle f, k(\cdot, X)\big\rangle_{\mathcal{H}_k} k(\cdot, X)\Big] = \mathbb{E}_P[f(X)\, k(\cdot, X)]\,. \tag{A.42}$$

Let $A := T_P + \rho I$. Then the population regularized Pearson χ^2 divergence can be written as

$$\chi^2_\rho(P|Q) = \sup_{f \in \mathcal{H}_k} \left\{ \langle f, \Delta\rangle_{\mathcal{H}_k} - \frac{1}{2}\langle f, Af\rangle_{\mathcal{H}_k} \right\}. \tag{A.43}$$

Since A is self-adjoint and strictly positive because $\rho > 0$, the objective is strictly concave in f, and the unique maximizer f^*_ρ satisfies the first-order optimality condition

$$Af^*_\rho = \Delta, \qquad \text{i.e.,} \qquad f^*_\rho = A^{-1}\Delta = (T_P + \rho I)^{-1}\Delta. \tag{A.44}$$

Plugging f^*_ρ back into (A.43) yields

$$\chi^2_\rho(P|Q) = \frac{1}{2}\langle \Delta, A^{-1}\Delta\rangle_{\mathcal{H}_k}. \tag{A.45}$$

A.4 Proofs for False-Alarm Control

A.4.1 Proof of Theorem 3.1

Proof. Let $\|\cdot\|_{\mathrm{op}}$ denote the induced operator norm, defined for any bounded linear operator M as

$$\|M\|_{\mathrm{op}} := \sup_{\|f\|_{\mathcal{H}_k}=1} \|Mf\|_{\mathcal{H}_k}. \tag{A.46}$$

Since $\hat{A} = \hat{T}_P + \rho I \succeq \rho I$, it follows that $\hat{A}^{-1} \preceq \rho^{-1} I$ and therefore $\|\hat{A}^{-1}\|_{\mathrm{op}} \le 1/\rho$. Hence,

$$\hat{\chi}^2_\rho[X_{1:n}|Y_{1:m}] = \frac{1}{2}\langle \hat{\Delta}, \hat{A}^{-1}\hat{\Delta}\rangle_{\mathcal{H}_k} \le \frac{1}{2\rho}\|\hat{\Delta}\|^2_{\mathcal{H}_k}. \tag{A.47}$$

Consequently, the tail event satisfies the following inclusion

$$\left\{\hat{\chi}^2_\rho[X_{1:n}|Y_{1:m}] > \nu\right\} \subseteq \left\{\|\hat{\Delta}\|_{\mathcal{H}_k} > \sqrt{2\rho\nu}\right\}. \tag{A.48}$$

Under the null hypothesis with common distribution P, let $\mu_P := \mathbb{E}[k(\cdot, X)]$ denote the population mean embedding in $\mathcal{H}_k$ and define the independent zero-mean random elements in $\mathcal{H}_k$ by

$$\begin{aligned} Z_i &= -\frac{1}{n}\big[k(\cdot, X_i) - \mu_P\big], \quad i = 1, \dots, n, \\ Z_{n+j} &= \frac{1}{m}\big[k(\cdot, Y_j) - \mu_P\big], \qquad j = 1, \dots, m. \end{aligned} \tag{A.49}$$

Then $\hat{\Delta} = \sum_{i=1}^{n+m} Z_i$. By the reproducing property, since $\|k(\cdot, x)\|^2_{\mathcal{H}_k} \le K$,

$$\|\mu_P\|_{\mathcal{H}_k} \le \mathbb{E}\|k(\cdot, X)\|_{\mathcal{H}_k} \le \sqrt{K}. \tag{A.50}$$

Consequently, by the triangle inequality, we see that $\|Z_i\|_2$ is bounded for all i:

$$\|Z_i\|_{\mathcal{H}_k} \le \begin{cases} \dfrac{2\sqrt{K}}{n}, & i = 1, \dots, n, \\ \dfrac{2\sqrt{K}}{m}, & i = n+1, \dots, n+m. \end{cases} \tag{A.51}$$

Thus, $\|Z_i\|_{\mathcal{H}_k} \le 2\sqrt{K}\max\{\frac{1}{n}, \frac{1}{m}\} =: 3E$ (for notation convenience). Moreover,

$$\sum_{i=1}^{n+m} \mathbb{E}\|Z_i\|^2_{\mathcal{H}_k} \le n\frac{4K}{n^2} + m\frac{4K}{m^2} = 4K\left[\frac{1}{n} + \frac{1}{m}\right] =: D. \tag{A.52}$$

Applying a Hilbert-space Bernstein inequality to $\sum_{i=1}^{n+m} Z_i$ yields, for any $s > 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^{n+m} Z_i\right\|_{\mathcal{H}_k} \ge s\right) \le 2\exp\left(-\frac{s^2}{2(D + Es)}\right). \tag{A.53}$$

Setting $s = \sqrt{2\rho\nu}$ and using the event inclusion (A.48) yields (12). □

A.4.2 Proof of Corollary 3.1: RKHS ARL control

Proof. For a constant threshold $\nu > 0$, let

$$p(\nu) := \mathbb{P}_\infty \left(\hat{\chi}^2_\rho[X_{1:n}|Y_{1:m}] > \nu \right). \tag{A.54}$$

denote the per-test false-alarm probability. Since the first test occurs at $t = n$, define

$$N := T - n + 1.$$

Without assuming independence across successive tests, the union bound gives

$$\mathbb{P}_\infty(N \le k) \le k\, p(\nu), \qquad k \ge 1. \tag{A.55}$$

Therefore, $\mathbb{P}_\infty(N > k) \ge \big(1 - kp(\nu)\big)_+$. Using the tail-sum formula for the integer-valued random variable N,

$$\mathbb{E}_\infty[N] = \sum_{k=0}^{\infty} \mathbb{P}_\infty(N > k) \ge \sum_{k=0}^{\infty} \big(1 - kp(\nu)\big)_+ \ge \int_0^{1/p(\nu)} \big(1 - xp(\nu)\big)\, dx = \frac{1}{2p(\nu)}. \tag{A.56}$$

Since $T = N + n - 1$,

$$\mathbb{E}_\infty[T] \ge n - 1 + \frac{1}{2p(\nu)}. \tag{A.57}$$

In particular, the sufficient condition

$$p(\nu) \le \frac{1}{2\gamma} \tag{A.58}$$

implies $\mathbb{E}_\infty[T] \ge \gamma$.

By Theorem 3.1,

$$p(\nu) \le 2 \exp\left(-\frac{\rho\nu}{D + E\sqrt{2\rho\nu}} \right). \tag{A.59}$$

Thus, condition $p(\nu) \le \frac{1}{2\gamma}$ is ensured by

$$\frac{\rho\nu}{D + E\sqrt{2\rho\nu}} \ge \log(4\gamma). \tag{A.60}$$

Let $s := \sqrt{2\rho\nu}$. Since $\rho\nu = s^2/2$, the preceding inequality is equivalent to

$$s^2 - 2E\log(4\gamma)\, s - 2D\log(4\gamma) \ge 0. \tag{A.61}$$

It is therefore sufficient that

$$s \ge E\log(4\gamma) + \sqrt{\big(E\log(4\gamma)\big)^2 + 2D\log(4\gamma)}. \tag{A.62}$$

Squaring and dividing by 2ρ yields (13). □

A.4.3 Proof of Corollary 3.2: RKHS UFAP control

Proof. Let $\alpha_t := \frac{\alpha}{t(t+1)}$. Since $\sum_{t=1}^{\infty} \alpha_t = \alpha$, it is sufficient to ensure that the false-alarm probability at time t does not exceed α_t.

Under the global null, both windows at each $t \ge n$ contain i.i.d. observations drawn from P. Hence, by Theorem 3.1,

$$\mathbb{P}_\infty \left(\hat{\chi}^2_{\rho,t} > \nu_t \right) \le 2 \exp\left(-\frac{\rho\nu_t}{D + E\sqrt{2\rho\nu_t}} \right). \tag{A.63}$$

Requiring the right-hand side to be at most α_t gives

$$\frac{\rho\nu_t}{D + E\sqrt{2\rho\nu_t}} \ge \log\left(\frac{2t(t+1)}{\alpha} \right) = L_t. \tag{A.64}$$

Solving this inequality as in the proof of Corollary 3.1 yields (14).

Finally, the union bound gives

$$\mathbb{P}_\infty(T < \infty) \le \sum_{t=n}^{\infty} \mathbb{P}_\infty \left(\hat{\chi}^2_{\rho,t} > \nu_t \right) \le \sum_{t=1}^{\infty} \alpha_t = \alpha. \tag{A.65}$$

□

A.5 Proofs for Miss-Detection Probability and Detection Delay

A.5.1 Detailed post-change decomposition and auxiliary bounds

To formalize the contaminated-window model, we introduce separate symbols for the pre- and post-change observations within each window:

$$X^{(t)} = \{X_{t-n+1}, \dots, X_t\} \qquad \text{and} \qquad Y^{(t)} = \{\tilde{X}_{t+1}, \dots, \tilde{X}_{t+\tau}, Y_{t+\tau+1}, \dots, Y_{t+m}\}, \quad \text{(A.66)}$$

where $X_i, \tilde{X}_j \sim P$, and $Y_j \sim Q$, with $P \neq Q$. All variables are assumed mutually independent. Here $\tau \in \{1, \dots, m-1\}$ denotes the number of pre-change observations contaminating the right window, and the contamination level is $\epsilon := \tau/m$.

To isolate the contamination, we introduce an auxiliary *cleaned* version of $Y^{(t)}$, denoted $Y^{(t)}_{\text{clean}}$, by substituting the τ pre-change samples in (A.66) with τ independent samples drawn from Q:

$$Y^{(t)}_{\text{clean}} = \{\tilde{Y}_{t+1}, \dots, \tilde{Y}_{t+\tau}, Y_{t+\tau+1}, \dots, Y_{t+m}\}, \quad \text{(A.67)}$$

where $\tilde{Y}_j \sim Q$ are the replacement samples. We define the empirical mean embeddings of the left and right windows as

$$\begin{aligned}
\hat{\mu}_P^{(t)} &:= \frac{1}{n}\sum_{i=1}^{n} k(\cdot, X_{t-n+i}),\\
\tilde{\mu}_P^{(t)} &:= \frac{1}{\tau}\sum_{i=1}^{\tau} k(\cdot, \tilde{X}_{t+i}),\\
\tilde{\mu}_Q^{(t)} &:= \frac{1}{\tau}\sum_{j=1}^{\tau} k(\cdot, \tilde{Y}_{t+j}),\\
\hat{\mu}_Q^{(t)} &:= \frac{1}{m}\Big[\sum_{j=1}^{\tau} k(\cdot, \tilde{X}_{t+j}) + \sum_{j=\tau+1}^{m} k(\cdot, Y_{t+j})\Big].\\
\hat{\mu}^{(t)}_{Q,\text{clean}} &:= \frac{1}{m}\Big[\sum_{j=1}^{\tau} k(\cdot, \tilde{Y}_{t+j}) + \sum_{j=\tau+1}^{m} k(\cdot, Y_{t+j})\Big].
\end{aligned} \quad \text{(A.68)}$$

Accordingly, we introduce

$$\hat{\Delta}_t := \hat{\mu}_Q^{(t)} - \hat{\mu}_P^{(t)}, \qquad \hat{\Delta}_{t,\text{clean}} := \hat{\mu}^{(t)}_{Q,\text{clean}} - \hat{\mu}_P^{(t)}, \qquad \tilde{\Delta}_t := \tilde{\mu}_Q^{(t)} - \tilde{\mu}_P^{(t)}. \quad \text{(A.69)}$$

Finally, writing $\hat{A}_t := \hat{T}_P^{(t)} + \rho I$ (resp., $\tilde{A}_t := \tilde{T}_P^{(t)} + \rho I$) for the empirical regularized operator built from (8) using $X^{(t)}$ (resp., $\tilde{X}^{(t)}$), we have the representations

$$\begin{aligned}
\hat{\chi}^2_\rho[X^{(t)}|Y^{(t)}] &:= \frac{1}{2}\big\langle \hat{\Delta}_t,\ \hat{A}_t^{-1}\hat{\Delta}_t\big\rangle_{\mathcal{H}_k} = \frac{1}{2}\|\hat{\Delta}_t\|^2_{\hat{A}_t^{-1}},\\
\hat{\chi}^2_\rho[X^{(t)}|Y^{(t)}_{\text{clean}}] &:= \frac{1}{2}\big\langle \hat{\Delta}_{t,\text{clean}},\ \hat{A}_t^{-1}\hat{\Delta}_{t,\text{clean}}\big\rangle_{\mathcal{H}_k} = \frac{1}{2}\|\hat{\Delta}_{t,\text{clean}}\|^2_{\hat{A}_t^{-1}},\\
\hat{\chi}^2_\rho[\tilde{X}^{(t)}|\tilde{Y}^{(t)}] &:= \frac{1}{2}\big\langle \tilde{\Delta}_t,\ \tilde{A}_t^{-1}\tilde{\Delta}_t\big\rangle_{\mathcal{H}_k} = \frac{1}{2}\|\tilde{\Delta}_t\|^2_{\tilde{A}_t^{-1}}.
\end{aligned} \quad \text{(A.70)}$$

In addition, we recall the definition of the population regularized Pearson divergence; see Appendix A.3, equations (A.39)–(A.45):

$$\chi^2_\rho(P|Q) = \frac{1}{2}\langle \Delta, A^{-1}\Delta\rangle_{\mathcal{H}_k}, \quad \text{(A.71)}$$

where $\Delta := \mathbb{E}_Q[k(\cdot, Y)] - \mathbb{E}_P[k(\cdot, X)]$ and

$$A := T_P + \rho I, \qquad T_P := \mathbb{E}_P\big[\, k(\cdot, X) \otimes k(\cdot, X)\,\big]. \quad \text{(A.72)}$$

For readability, we work below with quantities of the form $\|\Delta\|_{A^{-1}}$ rather than with χ^2_ρ. The two notations are immediately related through (A.70)–(A.71).

Observe that $\hat{\Delta}_t = \hat{\Delta}_{t,\text{clean}} - \epsilon\,\tilde{\Delta}_t$ with $\epsilon = \tau/m$. Then, by the reverse triangle inequality,

$$\begin{aligned}\|\hat{\Delta}_t\|_{\hat{A}_t^{-1}} &\geq \|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}} - \epsilon\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}} \\ &\geq \mathbb{E}_t\|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}} - \Big|\|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}} - \mathbb{E}_t\|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}}\Big| \\ &\quad - \epsilon\Big(\mathbb{E}_t\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}} + \Big|\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}} - \mathbb{E}_t\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}}\Big|\Big),\end{aligned} \tag{A.73}$$

where $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|X^{(t)}]$, so that conditionally on $X^{(t)}$ the operator $\hat{A}_t$ is fixed and the randomness comes only from the samples in the right window. This conditioning is natural in the CPD setting: the left window contains exclusively pre-change observations from P and serves as the reference estimate, while the right window is where post-change evidence accumulates and where the detection-relevant randomness resides. Next, using triangle-inequality manipulations,

$$\mathbb{E}_t\|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}} \geq \|\Delta\|_{A^{-1}} - \Big|\mathbb{E}_t\|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}} - \|\Delta\|_{A^{-1}}\Big|, \tag{A.74}$$

$$\mathbb{E}_t\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}} \leq \|\Delta\|_{A^{-1}} + \Big|\mathbb{E}_t\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}} - \|\Delta\|_{A^{-1}}\Big|. \tag{A.75}$$

Combining these inequalities yields

$$\|\hat{\Delta}_t\|_{\hat{A}_t^{-1}} \geq (1-\epsilon)\|\Delta\|_{A^{-1}}, \tag{A.76a}$$

$$- \Big|\|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}} - \mathbb{E}_t\|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}}\Big| - \epsilon\Big|\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}} - \mathbb{E}_t\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}}\Big|, \tag{A.76b}$$

$$- \Big|\mathbb{E}_t\|\hat{\Delta}_{t,\text{clean}}\|_{\hat{A}_t^{-1}} - \|\Delta\|_{A^{-1}}\Big| - \epsilon\Big|\mathbb{E}_t\|\tilde{\Delta}_t\|_{\hat{A}_t^{-1}} - \|\Delta\|_{A^{-1}}\Big|. \tag{A.76c}$$

where the term in the right-hand-side of (A.76a) is the attenuated signal contribution, making explicit how contamination reduces the effective separation in feature space; the terms in (A.76b) are concentration contributions describing finite-sample fluctuations around the conditional means; and the terms in (A.76c) are approximation contributions capturing the mismatch between these conditional means in the random metric $\|\cdot\|_{\hat{A}_t^{-1}}$, and the population quantity $\|\Delta\|_{A^{-1}}$.

For readability, we rewrite empirical quantities such as $\hat{\square}_t$ and $\tilde{\square}_t$ simply as $\hat{\square}$ and $\tilde{\square}$, respectively, dropping the t-subscript whenever it creates no ambiguity. We also introduce a more compact notation for (A.76). Define the attenuated signal level

$$S_\epsilon := (1-\epsilon)\,\|\Delta\|_{A^{-1}}, \tag{A.77}$$

and the four deviation terms

$$Q_1 := \Big|\|\hat{\Delta}_{\text{clean}}\|_{\hat{A}^{-1}} - \mathbb{E}_t\|\hat{\Delta}_{\text{clean}}\|_{\hat{A}^{-1}}\Big|, \tag{A.78}$$

$$Q_2 := \Big|\|\tilde{\Delta}\|_{\hat{A}^{-1}} - \mathbb{E}_t\|\tilde{\Delta}\|_{\hat{A}^{-1}}\Big|, \tag{A.79}$$

$$Q_3 := \Big|\mathbb{E}_t\|\hat{\Delta}_{\text{clean}}\|_{\hat{A}^{-1}} - \|\Delta\|_{A^{-1}}\Big|, \tag{A.80}$$

$$Q_4 := \Big|\mathbb{E}_t\|\tilde{\Delta}\|_{\hat{A}^{-1}} - \|\Delta\|_{A^{-1}}\Big|. \tag{A.81}$$

With these definitions, (A.76) can be rewritten succinctly as

$$\|\hat{\Delta}\|_{\hat{A}^{-1}} \geq S_\epsilon - Q_1 - \epsilon Q_2 - Q_3 - \epsilon Q_4, \tag{A.82}$$

which we will use throughout the remainder of the derivation to keep expressions compact.

Bounds on Q_1 and Q_2. Consider the first term in (A.76b) denoted for convenience by Q_1 in (A.78). Since $\hat{A} = \hat{T}_P + \rho I \succeq \rho I$, it follows that $\|\hat{A}^{-1}\|_{\text{op}} \leq 1/\rho$. Let

$$f_{\text{clean}} := \|\hat{\Delta}_{\text{clean}}\|_{\hat{A}^{-1}}, \tag{A.83}$$

and let $f'_{\text{clean}} := \|\hat{\Delta}'_{\text{clean}}\|_{\hat{A}^{-1}}$ denote the same quantity computed after replacing a single sample in the right window with an independent copy. Then, by the reverse triangle inequality,

$$\left|f_{\text{clean}} - f'_{\text{clean}}\right| \le \|\hat{\Delta}_{\text{clean}} - \hat{\Delta}'_{\text{clean}}\|_{\hat{A}^{-1}} \le \sqrt{\|\hat{A}^{-1}\|_{\text{op}}}\,\frac{2\sqrt{K}}{m} \le \frac{2\sqrt{K}}{m\sqrt{\rho}}. \tag{A.84}$$

Therefore, by McDiarmid's inequality, for any $\delta \in (0,1)$, with $\mathbb{P}(\cdot|X^{(t)})$ probability at least $1-\delta$,

$$Q_1 \le \sqrt{\frac{2K}{m\rho}\log\Big(\frac{2}{\delta}\Big)}. \tag{A.85}$$

Since the right-hand side of (A.85) does not depend on $X^{(t)}$, the same bound holds with unconditional probability $\mathbb{P}(\cdot)$ at least $1-\delta$. The same bounded-differences argument applies to the second term in (A.76b) denoted for convenience by Q_2 in (A.79), with 2τ samples, leading to

$$Q_2 \le \sqrt{\frac{4K}{\tau\rho}\log\Big(\frac{2}{\delta}\Big)}. \tag{A.86}$$

Bound on Q_3. Consider the first term in (A.76c), denoted for conveniance by Q_3 in (A.80). By an add-and-subtract argument,

$$Q_3 \le \Big|\mathbb{E}_t\|\hat{\Delta}_{\text{clean}}\|_{\hat{A}^{-1}} - \|\Delta\|_{\hat{A}^{-1}}\Big| + \Big|\|\Delta\|_{\hat{A}^{-1}} - \|\Delta\|_{A^{-1}}\Big|. \tag{A.87}$$

We first control the first term on the right-hand side of (A.87). Applying Jensen's inequality, then the reverse triangle inequality, and finally the operator-Hilbert norm comparison,

$$\begin{aligned}\Big|\mathbb{E}_t\|\hat{\Delta}_{\text{clean}}\|_{\hat{A}^{-1}} - \|\Delta\|_{\hat{A}^{-1}}\Big| &\le \mathbb{E}_t\|\hat{\Delta}_{\text{clean}} - \Delta\|_{\hat{A}^{-1}} \\ &\le \mathbb{E}_t\Big[\sqrt{\|\hat{A}^{-1}\|_{\text{op}}}\,\|\hat{\Delta}_{\text{clean}} - \Delta\|_{\mathcal{H}_k}\Big] \\ &\le \frac{1}{\sqrt{\rho}}\,\mathbb{E}_t\|\hat{\Delta}_{\text{clean}} - \Delta\|_{\mathcal{H}_k}.\end{aligned} \tag{A.88}$$

Further, by the triangle inequality and since $\mathbb{E}_t$ conditions on the initial P-window so that $\hat{\mu}_P$ is fixed under $\mathbb{E}_t$, we have

$$\mathbb{E}_t\|\hat{\Delta}_{\text{clean}} - \Delta\|_{\mathcal{H}_k} \le \|\hat{\mu}_P - \mu_P\|_{\mathcal{H}_k} + \mathbb{E}_t\|\hat{\mu}_{Q,\text{clean}} - \mu_Q\|_{\mathcal{H}_k}. \tag{A.89}$$

Define $Z_i := k(\cdot, X_i)$, so that $\|Z_i\|_{\mathcal{H}_k} \le \sqrt{K}$. Let $\hat{\mu}_P = \frac{1}{n}\sum_{i=1}^n Z_i$ and $\mu_P = \mathbb{E}Z$. If, for some i, we replace X_i in $\hat{\mu}_P$ by an independent copy X'_i to form $\hat{\mu}'_P$, then

$$\|\hat{\mu}_P - \hat{\mu}'_P\|_{\mathcal{H}_k} \le \frac{2\sqrt{K}}{n}. \tag{A.90}$$

Hence, McDiarmid's inequality implies that for any $\delta \in (0,1)$, with probability at least $1-\delta/2$,

$$\|\hat{\mu}_P - \mu_P\|_{\mathcal{H}_k} \le \mathbb{E}\|\hat{\mu}_P - \mu_P\|_{\mathcal{H}_k} + \sqrt{\frac{2K}{n}\log\Big(\frac{2}{\delta}\Big)}. \tag{A.91}$$

To make this expression explicit, it remains to bound the mean term in (A.91). By Jensen's inequality,

$$\mathbb{E}\|\hat{\mu}_P - \mu_P\|_{\mathcal{H}_k} \le \sqrt{\mathbb{E}\|\hat{\mu}_P - \mu_P\|^2_{\mathcal{H}_k}}. \tag{A.92}$$

Moreover, by independence and zero cross terms,

$$\begin{aligned}\mathbb{E}\|\hat{\mu}_P - \mu_P\|^2_{\mathcal{H}_k} &= \mathbb{E}\Big\|\frac{1}{n}\sum_{i=1}^n (Z_i - \mathbb{E}Z_i)\Big\|^2_{\mathcal{H}_k} \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\|Z_i - \mathbb{E}Z_i\|^2_{\mathcal{H}_k} \\ &= \frac{1}{n}\,\mathbb{E}\|Z - \mathbb{E}Z\|^2_{\mathcal{H}_k} \le \frac{1}{n}\,\mathbb{E}\|Z\|^2_{\mathcal{H}_k} \le \frac{K}{n}.\end{aligned} \tag{A.93}$$

Thus $\mathbb{E}\|\hat{\mu}_P - \mu_P\|_{\mathcal{H}_k} \leq \sqrt{K/n}$. Combining with (A.91) yields that, with probability at least $1-\delta/2$,

$$\|\hat{\mu}_P - \mu_P\|_{\mathcal{H}_k} \leq \sqrt{\frac{K}{n}} + \sqrt{\frac{2K}{n}\log\Big(\frac{2}{\delta}\Big)}. \tag{A.94}$$

Consider now the second term on the right-hand-side of (A.89). Similarly to (A.93), by independence and zero cross terms,

$$\mathbb{E}_t\|\hat{\mu}_{Q,\mathrm{clean}} - \mu_Q\|_{\mathcal{H}_k} = \mathbb{E}\|\hat{\mu}_{Q,\mathrm{clean}} - \mu_Q\|_{\mathcal{H}_k} \leq \sqrt{\mathbb{E}\|\hat{\mu}_{Q,\mathrm{clean}} - \mu_Q\|^2_{\mathcal{H}_k}} \leq \sqrt{\frac{K}{m}}. \tag{A.95}$$

Finally, we obtain that, with probability at least $1-\delta/2$,

$$\mathbb{E}_t\|\hat{\Delta}_{\mathrm{clean}} - \Delta\|_{\mathcal{H}_k} \leq \sqrt{\frac{K}{n}} + \sqrt{\frac{2K}{n}\log\Big(\frac{2}{\delta}\Big)} + \sqrt{\frac{K}{m}}. \tag{A.96}$$

Consider now the second term on the right-hand side of (A.87). We start with a high-probability control of the operator perturbation $\hat{T}_P - T_P$.

Lemma A.1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\hat{T}_P - T_P\|_{\mathrm{op}} \leq \frac{K}{\sqrt{n}}\Big(2 + \sqrt{2\log\Big(\frac{1}{\delta}\Big)}\Big). \tag{A.97}$$

Proof. Let $U_i := k(\cdot, X_i) \otimes k(\cdot, X_i)$, so that $\hat{T}_P = \frac{1}{n}\sum_{i=1}^n U_i$ and $T_P = \mathbb{E}\,U$. Each U_i is a rank-one operator, hence Hilbert–Schmidt, and $\|U_i\|_{\mathrm{op}} = \|U_i\|_{\mathrm{HS}}$.

Therefore, since $\|u \otimes v\|_{\mathrm{HS}} = \|u\|_{\mathcal{H}_k}\,\|v\|_{\mathcal{H}_k}$,

$$\|U_i\|_{\mathrm{HS}} = \|k(\cdot, X_i)\|^2_{\mathcal{H}_k} \leq K, \qquad \|T_P\|_{\mathrm{HS}} \leq \mathbb{E}\|U\|_{\mathrm{HS}} \leq K. \tag{A.98}$$

If, for some i, we replace U_i in $\hat{T}_P$ by an independent copy U_i' to form $\hat{T}_P'$, then

$$\|\hat{T}_P - \hat{T}_P'\|_{\mathrm{HS}} = \frac{1}{n}\|U_i - U_i'\|_{\mathrm{HS}} \leq \frac{1}{n}\Big(\|U_i\|_{\mathrm{HS}} + \|U_i'\|_{\mathrm{HS}}\Big) \leq \frac{2K}{n}. \tag{A.99}$$

Consequently, McDiarmid's inequality implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|\hat{T}_P - T_P\|_{\mathrm{HS}} \leq \mathbb{E}\|\hat{T}_P - T_P\|_{\mathrm{HS}} + K\sqrt{\frac{2}{n}\log\Big(\frac{1}{\delta}\Big)}. \tag{A.100}$$

To make this expression explicit, it remains to bound the mean term in (A.100). By independence of U_i variables,

$$\mathbb{E}\|\hat{T}_P - T_P\|^2_{\mathrm{HS}} = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\|U_i - T_P\|^2_{\mathrm{HS}} = \frac{1}{n}\mathbb{E}\|U - T_P\|^2_{\mathrm{HS}}. \tag{A.101}$$

Moreover,

$$\|U - T_P\|_{\mathrm{HS}} \leq \|U\|_{\mathrm{HS}} + \|T_P\|_{\mathrm{HS}} \leq 2K. \tag{A.102}$$

Hence, by Jensen's inequality,

$$\mathbb{E}\|\hat{T}_P - T_P\|_{\mathrm{HS}} \leq \sqrt{\mathbb{E}\|\hat{T}_P - T_P\|^2_{\mathrm{HS}}} \leq \frac{2K}{\sqrt{n}}. \tag{A.103}$$

Combining with (A.100) yields that, with probability at least $1-\delta$,

$$\|\hat{T}_P - T_P\|_{\mathrm{HS}} \leq \frac{K}{\sqrt{n}}\Big(2 + \sqrt{2\log\Big(\frac{1}{\delta}\Big)}\Big). \tag{A.104}$$

Since $\|\cdot\|_{\mathrm{op}} \leq \|\cdot\|_{\mathrm{HS}}$, the same bound holds with $\|\cdot\|_{\mathrm{op}}$. □

To handle the second term on the right-hand side of (A.87), we exploit a first-order expansion of the inverse map $A \mapsto A^{-1}$ around any positive-definite operator, valid whenever the perturbation is sufficiently small in operator norm relative to the conditioning of A. Set

$$\hat{F} := \hat{A} - A = \hat{T}_P - T_P, \tag{A.105}$$

so that

$$\hat{A}^{-1} = (A + \hat{F})^{-1} = (I + A^{-1}\hat{F})^{-1}A^{-1}. \tag{A.106}$$

The convergence of the Neumann series for $(I + A^{-1}\hat{F})^{-1}$ requires $\|A^{-1}\hat{F}\|_{\mathrm{op}} < 1$. To work with a uniform contraction factor and to obtain explicit bounds on the higher-order remainder, we impose the slightly stronger condition defined by the event

$$\mathcal{E} := \big\{\|A^{-1}\hat{F}\|_{\mathrm{op}} \le 1/2\big\}. \tag{A.107}$$

Since $A \succeq \rho I$, we have $\|A^{-1}\|_{\mathrm{op}} \le 1/\rho$, and the inclusion $\{\|\hat{F}\|_{\mathrm{op}} \le \rho/2\} \subseteq \mathcal{E}$ holds. Lemma A.1 therefore implies that, under the sample-size condition

$$n \ge \frac{4K^2}{\rho^2}\left(2 + \sqrt{2\log\Big(\frac{4}{\delta}\Big)}\right)^2, \tag{A.108}$$

the event $\mathcal{E}$ holds with probability at least $1 - \delta/4$.

On $\mathcal{E}$, the Neumann series converges absolutely in operator norm,

$$(I + A^{-1}\hat{F})^{-1} = \sum_{k\ge 0}(-A^{-1}\hat{F})^k, \tag{A.109}$$

and yields the decomposition

$$\hat{A}^{-1} - A^{-1} = -A^{-1}\hat{F}A^{-1} + \hat{R}, \tag{A.110}$$

where the remainder is the tail of the series,

$$\hat{R} := \sum_{k\ge 2}(-A^{-1}\hat{F})^k A^{-1}, \tag{A.111}$$

and is of order $\|\hat{F}\|_{\mathrm{op}}^2$. Indeed, bounding (A.111) by a geometric-series argument on $\mathcal{E}$,

$$\|\hat{R}\|_{\mathrm{op}} \le \|A^{-1}\|_{\mathrm{op}} \sum_{k\ge 2}\|A^{-1}\hat{F}\|_{\mathrm{op}}^k \le \frac{1}{\rho}\cdot\frac{\|A^{-1}\hat{F}\|_{\mathrm{op}}^2}{1 - \|A^{-1}\hat{F}\|_{\mathrm{op}}} \le \frac{2\,\|\hat{F}\|_{\mathrm{op}}^2}{\rho^3}, \tag{A.112}$$

where we used $\sum_{k\ge 2} r^k = r^2/(1-r)$ for $|r| < 1$, and $\|A^{-1}\hat{F}\|_{\mathrm{op}} \le 1/2$.

Equation (A.110) is the key linearization for the analysis of the second term in (A.87). Combining it with the difference-of-squares identity $u - v = (u^2 - v^2)/(u + v)$ yields

$$\begin{aligned}\|\Delta\|_{\hat{A}^{-1}} - \|\Delta\|_{A^{-1}} &= \frac{\langle\Delta, \hat{A}^{-1}\Delta\rangle - \langle\Delta, A^{-1}\Delta\rangle}{\|\Delta\|_{\hat{A}^{-1}} + \|\Delta\|_{A^{-1}}} \\ &= \frac{\hat{L} + \langle\Delta, \hat{R}\Delta\rangle}{\|\Delta\|_{\hat{A}^{-1}} + \|\Delta\|_{A^{-1}}},\end{aligned} \tag{A.113}$$

with $\hat{L} := -\langle\Delta, A^{-1}\hat{F}A^{-1}\Delta\rangle$. Taking absolute values and applying the triangle inequality in the numerator gives

$$\big|\|\Delta\|_{\hat{A}^{-1}} - \|\Delta\|_{A^{-1}}\big| \le \frac{|\hat{L}| + |\langle\Delta, \hat{R}\Delta\rangle|}{\|\Delta\|_{\hat{A}^{-1}} + \|\Delta\|_{A^{-1}}}. \tag{A.114}$$

We bound the numerator terms $|\hat{L}|$ and $|\langle\Delta, \hat{R}\Delta\rangle|$ in high probability, and the denominator from below deterministically. Combining the three controls below yields the final bound.

Setting $h := A^{-1}\Delta$ and using the self-adjointness of A^{-1},

$$\hat{L} = -\langle h, \hat{F}\,h\rangle = -\frac{1}{n}\sum_{i=1}^{n}\big(\langle h, U_i h\rangle - \langle h, T_P h\rangle\big), \tag{A.115}$$

where $U_i := k(\cdot, X_i) \otimes k(\cdot, X_i)$. Thus $\hat{L}$ is the empirical average of independent centered random variables. Moreover

$$0 \le \langle h, U_i h \rangle = \langle h, k(\cdot, X_i) \rangle^2 \le K \|h\|_{\mathcal{H}_k}^2 \le \frac{K}{\rho^2} \|\Delta\|_{\mathcal{H}_k}^2. \tag{A.116}$$

By the triangle inequality and Jensen's inequality applied to the kernel mean embeddings,

$$\|\Delta\|_{\mathcal{H}_k} = \|\mu_P - \mu_Q\|_{\mathcal{H}_k} \le \|\mu_P\|_{\mathcal{H}_k} + \|\mu_Q\|_{\mathcal{H}_k} \le 2\sqrt{K}. \tag{A.117}$$

Combining these two bounds shows that each centered summand $\langle h, U_i h \rangle - \langle h, T_P h \rangle$ has range at most $4K^2/\rho^2$. Therefore, Hoeffding's inequality gives, with probability at least $1 - \delta/4$,

$$|\hat{L}| \le \frac{4K^2}{\rho^2} \sqrt{\frac{1}{2n} \log\Big(\frac{8}{\delta}\Big)}. \tag{A.118}$$

Using again $\|\Delta\|_{\mathcal{H}_k}^2 \le 4K$, we obtain

$$|\langle \Delta, \hat{R}\Delta \rangle| \le \|\Delta\|_{\mathcal{H}_k}^2 \|\hat{R}\|_{\mathrm{op}} \le \frac{8K}{\rho^3} \|\hat{F}\|_{\mathrm{op}}^2. \tag{A.119}$$

By Lemma A.1, with probability at least $1 - \delta/4$,

$$\|\hat{F}\|_{\mathrm{op}} \le \frac{K}{\sqrt{n}} \left(2 + \sqrt{2 \log\Big(\frac{4}{\delta}\Big)}\right). \tag{A.120}$$

Both (A.107) and (A.120) are consequences of the same application of Lemma A.1 at level $\delta/4$, hence they hold simultaneously on a single event of probability at least $1 - \delta/4$. On this event, substituting (A.120) into (A.119) yields

$$|\langle \Delta, \hat{R}\Delta \rangle| \le \frac{8K^3}{\rho^3 n} \left(2 + \sqrt{2 \log\Big(\frac{4}{\delta}\Big)}\right)^2. \tag{A.121}$$

Finally we consider the denominator in (A.114). Since

$$\|\Delta\|_{\hat{A}^{-1}} + \|\Delta\|_{A^{-1}} \ge \|\Delta\|_{A^{-1}}, \tag{A.122}$$

and

$$\|\Delta\|_{A^{-1}}^2 = \langle \Delta, A^{-1}\Delta \rangle \ge \frac{\|\Delta\|_{\mathcal{H}_k}^2}{\|A\|_{\mathrm{op}}} \ge \frac{\|\Delta\|_{\mathcal{H}_k}^2}{K + \rho}, \tag{A.123}$$

provided $\Delta \neq 0$, we obtain

$$\|\Delta\|_{\hat{A}^{-1}} + \|\Delta\|_{A^{-1}} \ge \frac{\|\Delta\|_{\mathcal{H}_k}}{\sqrt{K + \rho}}. \tag{A.124}$$

Combining (A.118), (A.121), (A.124), together with (A.114), we obtain, under the sample-size condition (A.108), with probability at least $1 - \delta/2$,

$$\Big| \|\Delta\|_{\hat{A}^{-1}} - \|\Delta\|_{A^{-1}} \Big| \le \frac{4K^2\sqrt{K+\rho}}{\rho^2 \|\Delta\|_{\mathcal{H}_k}} \sqrt{\frac{1}{2n} \log\Big(\frac{8}{\delta}\Big)} + \frac{8K^3\sqrt{K+\rho}}{\rho^3 \|\Delta\|_{\mathcal{H}_k} n} \left(2 + \sqrt{2 \log\Big(\frac{4}{\delta}\Big)}\right)^2. \tag{A.125}$$

Combining (A.96) and (A.125) via a union bound, we obtain the following high-probability bound for Q_3. Assume $\Delta \neq 0$ and that

$$n \ge \frac{4K^2}{\rho^2} \Big(2 + \sqrt{2 \log(4/\delta)}\Big)^2. \tag{A.126}$$

Then, with probability at least $1 - \delta$,

$$\begin{aligned} Q_3 \le{}& \frac{1}{\sqrt{\rho}} \left[\sqrt{\frac{K}{n}} + \sqrt{\frac{2K}{n} \log\Big(\frac{2}{\delta}\Big)} + \sqrt{\frac{K}{m}} \right] \\ &+ \frac{4K^2\sqrt{K+\rho}}{\rho^2 \|\Delta\|_{\mathcal{H}_k}} \sqrt{\frac{1}{2n} \log\Big(\frac{8}{\delta}\Big)} + \frac{8K^3\sqrt{K+\rho}}{\rho^3 \|\Delta\|_{\mathcal{H}_k} n} \left(2 + \sqrt{2 \log\Big(\frac{4}{\delta}\Big)}\right)^2. \end{aligned} \tag{A.127}$$

Bound on Q_4. Upon replacing $\hat{\Delta}_{\text{clean}}$ by $\tilde{\Delta}$, a similar analysis applies to the second term in (A.76c), denoted by Q_4 in (A.81) for convenience. By an add-and-subtract argument as in (A.87), Q_4 admits the decomposition

$$Q_4 \le \frac{1}{\sqrt{\rho}}\,\mathbb{E}_t\|\tilde{\Delta}-\Delta\|_{\mathcal{H}_k} + \big|\|\Delta\|_{\hat{A}^{-1}} - \|\Delta\|_{A^{-1}}\big|. \tag{A.128}$$

Since $\tilde{\Delta}$ depends only on the auxiliary τ-sample averages $(\tilde{X},\tilde{Y})$, which are independent of $X^{(t)}$, the first term is bounded in expectation by the same independence-and-zero-cross-terms argument as in (A.93) applied to each of the two auxiliary sample means and the triangle inequality:

$$\frac{1}{\sqrt{\rho}}\,\mathbb{E}_t\|\tilde{\Delta}-\Delta\|_{\mathcal{H}_k} \le \frac{2}{\sqrt{\rho}}\sqrt{\frac{K}{\tau}} = 2\sqrt{\frac{K}{\rho\tau}}. \tag{A.129}$$

The second term is identical to the second term of (A.87) and is therefore controlled by (A.125). Since the first term of (A.128) carries no stochastic event, the failure budget is entirely determined by (A.125). Therefore, under the sample-size condition (A.108), with probability at least $1-\delta/2$,

$$Q_4 \le 2\sqrt{\frac{K}{\rho\tau}} + \frac{4K^2\sqrt{K+\rho}}{\rho^2\,\|\Delta\|_{\mathcal{H}_k}}\sqrt{\frac{1}{2n}\log\Big(\frac{8}{\delta}\Big)} + \frac{8K^3\sqrt{K+\rho}}{\rho^3\,\|\Delta\|_{\mathcal{H}_k}\,n}\Big(2+\sqrt{2\log\Big(\frac{4}{\delta}\Big)}\Big)^2. \tag{A.130}$$

A.5.2 Proof of Theorem 4.1: Per-time miss probability

Having established high-probability bounds for the deviation terms $Q_1,\dots,Q_4$ in Appendix A.5.1, we now combine them to obtain an upper bound on the per-time miss probability. For completeness, we first restate Theorem 4.1 with the explicit constants omitted from its formulation in the main text.

Theorem A.1 (Per-time miss probability with explicit constants)**.** *Fix a time index t with $\epsilon_t \in (0,1)$. Assume the technical condition $0<\rho\le K$. Define the corresponding attenuated signal level*

$$S_{\epsilon_t} := (1-\epsilon_t)\|\Delta\|_{A^{-1}}, \tag{A.131}$$

and the effective signal margin

$$a_t := \frac{S_{\epsilon_t}-\sqrt{2\nu}}{4}, \tag{A.132}$$

where $\nu>0$ is a fixed detection threshold. Define the deterministic floors

$$c_0 := \sqrt{\frac{K}{\rho}}\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{m}}\right), \qquad d_0 := 2\sqrt{\frac{K\epsilon_t}{\rho m}}, \tag{A.133}$$

and the residual margins

$$r_t := (a_t-c_0)_+, \qquad r_t' := \frac{1}{\epsilon_t}(a_t-d_0)_+. \tag{A.134}$$

Assume the sample-size condition

$$n \ge \frac{4K^2}{\rho^2}\left(2+\sqrt{2\log\left(\frac{4}{\delta}\right)}\right)^2, \tag{A.135}$$

for some $\delta\in(0,1)$. Then

$$\begin{aligned} p_{\text{miss},t} \le \min\Big\{1,\ & 2\exp\left(-\frac{m\rho a_t^2}{2K}\right) + 2\exp\left(-\frac{m\rho a_t^2}{4K\epsilon_t}\right) \\ & + 8\exp\left(-\min\left\{\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t^2}{128K^4(K+\rho)}, \frac{\rho^3\|\Delta\|_{\mathcal{H}_k} n r_t}{256K^3\sqrt{K+\rho}}\right\}\right) \\ & + 8\exp\left(-\min\left\{\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t'^2}{32K^4(K+\rho)}, \frac{\rho^3\|\Delta\|_{\mathcal{H}_k} n r_t'}{256K^3\sqrt{K+\rho}}\right\}\right)\Big\}. \end{aligned} \tag{A.136}$$

Proof. The trivial bound $p_{\mathrm{miss},t} \le 1$ holds unconditionally and is absorbed into the outer $\min\{1,\cdot\}$. We now derive the non-trivial part under the assumption $S_{\epsilon_t} > \sqrt{2\nu}$, that is, $a_t > 0$. When $a_t \le 0$ the residual margins satisfy $r_t = r'_t = 0$, so the third and fourth terms in (A.136) each equal 8, and the outer $\min\{1,\cdot\}$ reduces the bound to the trivial value 1. The hypothesis $a_t > 0$ is therefore the regime in which the bound can be informative.

Recall the decomposition (A.82),

$$\|\hat{\Delta}\|_{\hat{A}^{-1}} \ge S_{\epsilon_t} - Q_1 - \epsilon_t Q_2 - Q_3 - \epsilon_t Q_4, \tag{A.137}$$

where $Q_1, \dots, Q_4$ are defined in (A.78)–(A.81). By inclusion of events,

$$\big\{\|\hat{\Delta}\|_{\hat{A}^{-1}} \le \sqrt{2\nu}\big\} \subseteq \big\{Q_1 + \epsilon_t Q_2 + Q_3 + \epsilon_t Q_4 \ge S_{\epsilon_t} - \sqrt{2\nu}\big\}. \tag{A.138}$$

A union bound with uniform splitting into four equal margins $a_t = (S_{\epsilon_t} - \sqrt{2\nu})/4$ yields

$$\mathbb{P}\Big(\|\hat{\Delta}\|_{\hat{A}^{-1}} \le \sqrt{2\nu}\Big) \le \mathbb{P}\big(Q_1 \ge a_t\big) + \mathbb{P}\big(\epsilon_t Q_2 \ge a_t\big) + \mathbb{P}\big(Q_3 \ge a_t\big) + \mathbb{P}\big(\epsilon_t Q_4 \ge a_t\big). \tag{A.139}$$

We now bound each term separately, retaining the trivial cap $\min\{1,\cdot\}$ throughout.

Bounds on Q_1 and Q_2. Using (A.85) and (A.86),

$$\begin{aligned} \mathbb{P}\big(Q_1 \ge a_t\big) &\le \min\Big\{1,\, 2\exp\Big(-\frac{m\rho\, a_t^2}{2K}\Big)\Big\}, \\ \mathbb{P}\big(\epsilon_t Q_2 \ge a_t\big) &\le \min\Big\{1,\, 2\exp\Big(-\frac{m\rho\, a_t^2}{4K\epsilon_t}\Big)\Big\}. \end{aligned} \tag{A.140}$$

The outer $\min\{1,\cdot\}$ is required because the factor 2 makes the exponential terms exceed 1 whenever $a_t^2 < 2K\log 2/(m\rho)$, respectively $a_t^2 < 4K\epsilon_t \log 2/(m\rho)$. These yield the first two terms of (A.136).

Bound on Q_3. By (A.127), with probability at least $1-\delta$ under the sample-size condition (A.108), the three logarithmic arguments $\log(2/\delta)$, $\log(4/\delta)$, $\log(8/\delta)$ appearing in (A.127) are all at most $L := \log(8/\delta)$. Moreover, for $\delta \le 1$ one has $L \ge \log 8$ and hence

$$\big(2 + \sqrt{2\log(4/\delta)}\big)^2 \le \big(2 + \sqrt{2L}\big)^2 \le 16\, L, \tag{A.141}$$

the last inequality being $4 + 4\sqrt{2L} \le 14L$, valid for $L \ge \log 8$. Substituting these into (A.127) gives

$$Q_3 \le c_0 + c_1\sqrt{L} + c_2\sqrt{L} + 16\, c_3\, L, \tag{A.142}$$

where

$$c_0 := \sqrt{\frac{K}{\rho}}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right), \quad c_1 := \sqrt{\frac{2K}{\rho n}}, \quad c_2 := \frac{4K^2\sqrt{K+\rho}}{\rho^2\,\|\Delta\|_{\mathcal{H}_k}\sqrt{2n}}, \quad c_3 := \frac{8K^3\sqrt{K+\rho}}{\rho^3\,\|\Delta\|_{\mathcal{H}_k}\, n}.$$

A direct comparison of c_1 and c_2 gives

$$\frac{c_1}{c_2} = \frac{\rho^{3/2}\,\|\Delta\|_{\mathcal{H}_k}}{2K^{3/2}\sqrt{K+\rho}}, \tag{A.143}$$

and using the universal bound $\|\Delta\|_{\mathcal{H}_k} \le 2\sqrt{K}$ from (A.117),

$$\frac{c_1}{c_2} \le \frac{\rho^{3/2}}{K\sqrt{K+\rho}}. \tag{A.144}$$

In particular, assuming the regime $0 < \rho \le K$, this ratio is at most one. Consequently $c_1 \le c_2$, and therefore $c_1\sqrt{L} + c_2\sqrt{L} \le 2c_2\sqrt{L}$. Equation (A.142) simplifies to

$$Q_3 \le c_0 + 2c_2\sqrt{L} + 16\, c_3\, L. \tag{A.145}$$

We now convert (A.145) into a tail bound on the event $\{Q_3 \ge a_t\}$. Recall $r_t = (a_t - c_0)_+$. If $r_t = 0$, then $a_t \le c_0$ and the trivial bound $\mathbb{P}(Q_3 \ge a_t) \le 1$ is retained. Assume henceforth $r_t > 0$,

and regard $L = \log(8/\delta)$ as a free parameter tuned through δ. By (A.145), on an event of probability at least $1 - \delta$,

$$Q_3 \le c_0 + 2c_2\sqrt{L} + 16\, c_3\, L. \tag{A.146}$$

Hence $Q_3 \le a_t$ on this event whenever

$$2c_2\sqrt{L} + 16\, c_3\, L \le r_t. \tag{A.147}$$

A sufficient condition for (A.147) is that each summand be at most $r_t/2$, that is $2c_2\sqrt{L} \le r_t/2$ and $16c_3L \le r_t/2$, equivalently

$$L \le L^\star := \min\left\{ \frac{r_t^2}{16c_2^2},\ \frac{r_t}{32c_3} \right\}. \tag{A.148}$$

If $L^\star \le \log 8$, the resulting exponential bound is at least one, and the trivial probability bound applies. Otherwise, choose $\log 8 < L < L^\star$. The sufficient inequality (A.147) is then strict, so

$$Q_3 < a_t$$

on the high-probability event. Consequently, $\mathbb{P}(Q_3 \ge a_t) \le 8e^{-L}$. Taking the limit as L approaches $L^\star$ from below yields

$$\mathbb{P}\big(Q_3 \ge a_t\big) \le \min\left\{1,\ 8\exp\Big(- \min\Big\{ \frac{r_t^2}{16c_2^2},\ \frac{r_t}{32c_3} \Big\}\Big)\right\}. \tag{A.149}$$

Substituting the explicit values of c_2 and c_3 yields the third term of (A.136).

Bound on Q_4. Since $\mathbb{P}(\epsilon_t Q_4 \ge a_t) = \mathbb{P}(Q_4 \ge a_t/\epsilon_t)$, the threshold for Q_4 is a_t/ϵ_t. By (A.130), with probability at least $1-\delta$ under the sample-size condition (A.108), and bounding the logarithmic arguments as in the Q_3 analysis by $L = \log(8/\delta)$,

$$Q_4 \le \frac{d_0}{\epsilon_t} + c_2\sqrt{L} + 16\, c_3\, L, \tag{A.150}$$

where the deterministic floor $d_0/\epsilon_t = 2\sqrt{K/(\rho\tau_t)}$ collects the contamination-window contribution, and the constants c_2, c_3 are those of Q_3.

Recall $r'_t = (a_t - d_0)_+/\epsilon_t$. If $r'_t = 0$, then $a_t \le d_0$ and $\mathbb{P}(\epsilon_t Q_4 \ge a_t) \le 1$ trivially. Assume henceforth $r'_t > 0$, so that $a_t > d_0$ and $(a_t - d_0)/\epsilon_t = r'_t$. By (A.150), on an event of probability at least $1 - \delta$,

$$Q_4 \le \frac{d_0}{\epsilon_t} + c_2\sqrt{L} + 16\, c_3\, L, \tag{A.151}$$

so $Q_4 \le a_t/\epsilon_t$ on this event whenever

$$c_2\sqrt{L} + 16\, c_3\, L \le \frac{a_t - d_0}{\epsilon_t} = r'_t. \tag{A.152}$$

A sufficient condition for (A.152) is that each summand be at most $r'_t/2$, equivalently

$$L \le L^\star := \min\left\{ \frac{r_t'^2}{4c_2^2},\ \frac{r'_t}{32c_3} \right\}. \tag{A.153}$$

If $L^\star \le \log 8$, the resulting exponential bound is at least one, and the trivial probability bound applies. Otherwise, choose $\log 8 < L < L^\star$. The sufficient inequality (A.152) is then strict, so

$$Q_4 < a_t/\epsilon_t$$

on the high-probability event. Consequently, $\mathbb{P}(\epsilon_t Q_4 \ge a_t) \le 8e^{-L}$. Taking the limit as L approaches $L^\star$ from below yields

$$\mathbb{P}\big(\epsilon_t Q_4 \ge a_t\big) = \mathbb{P}\big(Q_4 \ge a_t/\epsilon_t\big) \le \min\left\{1,\ 8\exp\Big(- \min\Big\{ \frac{r_t'^2}{4c_2^2},\ \frac{r'_t}{32c_3} \Big\}\Big)\right\}. \tag{A.154}$$

Substituting the explicit values of c_2 and c_3 yields the fourth term of (A.136).

Admissibility of the optimized confidence levels. The preceding substitutions $L = L^\star$, corresponding to $\delta^\star = 8e^{-L^\star}$, require verifying the sample-size condition (A.108) at the selected confidence level. Set $x := \rho\sqrt{n}/(2K)$. For $L > \log 8$, this condition is equivalent to

$$x \geq 2 + \sqrt{2(L - \log 2)}.$$

For Q_3, using $r_t \leq \|\Delta\|_{\mathcal{H}_k}/(4\sqrt{\rho})$ and $\|\Delta\|_{\mathcal{H}_k} \leq 2\sqrt{K}$, we obtain

$$L_3^\star \leq \frac{\rho^2 n}{128K^2} = \frac{x^2}{32}.$$

Thus, whenever $L_3^\star > \log 8$, the sample-size condition holds at $\delta^\star = 8e^{-L_3^\star}$.

For Q_4, note that A and $\hat{A}$ have spectra in $[\rho, K + \rho]$. Hence,

$$Q_4 \leq \frac{d_0}{\epsilon_t} + M_\rho, \qquad M_\rho := \|\Delta\|_{\mathcal{H}_k}\big(\rho^{-1/2} - (K + \rho)^{-1/2}\big).$$

If $r_t' > M_\rho$, the event $\{Q_4 \geq a_t/\epsilon_t\}$ is empty. Otherwise, using $r_t' \leq M_\rho$ gives

$$L_4^\star \leq \frac{\rho^2 n}{32K^2} = \frac{x^2}{8}.$$

Consequently, whenever $L_4^\star > \log 8$, the sample-size condition also holds at $\delta^\star = 8e^{-L_4^\star}$.

If either optimized value satisfies $L^\star \leq \log 8$, the corresponding exponential bound is at least one and the trivial probability bound applies. The optimized confidence levels are therefore admissible whenever their bounds are nontrivial. □

A.5.3 Proof of Corollary 4.1: MMD specialization of miss-detection bound

The analysis in Appendix A.5.2 can be specialized to the MMD detector by replacing the regularized Mahalanobis norm $\|\cdot\|_{\hat{A}^{-1}}$ with the RKHS norm $\|\cdot\|_{\mathcal{H}_k}$. The statistic becomes $\frac{1}{2}\|\tilde{\Delta}_t\|_{\mathcal{H}_k}^2$. Since the metric is fixed, the operator-perturbation terms vanish. However, fluctuations of both empirical mean embeddings remain.

The MMD-specific signal strength and margin are defined by

$$\bar{S}_{\epsilon_t} := (1 - \epsilon_t)\|\Delta\|_{\mathcal{H}_k}, \qquad \bar{a}_t := \frac{\bar{S}_{\epsilon_t} - \sqrt{2\nu}}{2}. \tag{A.155}$$

The corresponding estimation floors are

$$\bar{c}_0 := \sqrt{K}\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}}\right), \qquad \bar{d}_0 := 2\sqrt{\frac{K\epsilon_t}{m}}, \tag{A.156}$$

and the residual margins, for $\epsilon_t \in (0, 1)$, are

$$\bar{r}_t := (\bar{a}_t - \bar{c}_0)_+, \qquad \bar{r}_t' := \frac{1}{\epsilon_t}(\bar{a}_t - \bar{d}_0)_+. \tag{A.157}$$

Let $\tau_t = \epsilon_t m$. Using the construction of Appendix A.5.2, we have $\hat{\Delta}_t = \hat{\Delta}_{t,\mathrm{clean}} - \epsilon_t\tilde{\Delta}_t$, where the cleaned test window contains m independent observations from Q, and $\tilde{\Delta}_t$ is the difference between the auxiliary empirical embeddings formed from τ_t observations from Q and P.

By the triangle inequality,

$$\|\hat{\Delta}_t\|_{\mathcal{H}_k} \geq \|\hat{\Delta}_{t,\mathrm{clean}}\|_{\mathcal{H}_k} - \epsilon_t\|\Delta\|_{\mathcal{H}_k} - \epsilon_t\|\tilde{\Delta}_t - \Delta\|_{\mathcal{H}_k}. \tag{A.158}$$

Hence, on the miss event $\{\|\hat{\Delta}_t\|_{\mathcal{H}_k} \leq \sqrt{2\nu}\}$,

$$\big(\|\Delta\|_{\mathcal{H}_k} - \|\hat{\Delta}_{t,\mathrm{clean}}\|_{\mathcal{H}_k}\big) + \epsilon_t\|\tilde{\Delta}_t - \Delta\|_{\mathcal{H}_k} \geq \bar{S}_{\epsilon_t} - \sqrt{2\nu} =: 2\bar{a}_t. \tag{A.159}$$

At least one of the two terms must therefore be at least $\bar{a}_t$. Consequently,

$$p_{\mathrm{miss},t} \leq \mathbb{P}\big(\|\hat{\Delta}_{t,\mathrm{clean}}\|_{\mathcal{H}_k} \leq \|\Delta\|_{\mathcal{H}_k} - \bar{a}_t\big) + \mathbb{P}\big(\epsilon_t\|\tilde{\Delta}_t - \Delta\|_{\mathcal{H}_k} \geq \bar{a}_t\big). \tag{A.160}$$

For the first term, independence of the reference and cleaned test windows gives

$$\mathbb{E}\|\hat{\Delta}_{t,\text{clean}} - \Delta\|_{\mathcal{H}_k} \le \sqrt{K\left(\frac{1}{n}+\frac{1}{m}\right)} \le \bar{c}_0. \tag{A.161}$$

Thus, $\mathbb{E}\|\hat{\Delta}_{t,\text{clean}}\|_{\mathcal{H}_k} \ge \|\Delta\|_{\mathcal{H}_k} - \bar{c}_0$. The norm of the cleaned embedding difference has bounded differences $2\sqrt{K}/n$ for each reference observation and $2\sqrt{K}/m$ for each test observation. Applying the one-sided McDiarmid inequality yields

$$\mathbb{P}\big(\|\hat{\Delta}_{t,\text{clean}}\|_{\mathcal{H}_k} \le \|\Delta\|_{\mathcal{H}_k} - \bar{a}_t\big) \le \exp\left(-\frac{\bar{r}_t^2}{2K(\frac{1}{n}+\frac{1}{m})}\right). \tag{A.162}$$

For the second term, the auxiliary embedding difference satisfies

$$\mathbb{E}\|\tilde{\Delta}_t - \Delta\|_{\mathcal{H}_k} \le 2\sqrt{\frac{K}{\tau_t}}, \qquad \epsilon_t \mathbb{E}\|\tilde{\Delta}_t - \Delta\|_{\mathcal{H}_k} \le \bar{d}_0. \tag{A.163}$$

Its norm depends on $2\tau_t$ independent observations, each with bounded difference $2\sqrt{K}/\tau_t$. A second one-sided McDiarmid inequality therefore gives

$$\mathbb{P}\big(\epsilon_t\|\tilde{\Delta}_t - \Delta\|_{\mathcal{H}_k} \ge \bar{a}_t\big) \le \exp\left(-\frac{\tau_t(\bar{r}_t')^2}{4K}\right) = \exp\left(-\frac{m\epsilon_t(\bar{r}_t')^2}{4K}\right). \tag{A.164}$$

Combining these inequalities proves

$$p_{\text{miss},t} \le \min\left\{1,\ \exp\left(-\frac{\bar{r}_t^2}{2K(\frac{1}{n}+\frac{1}{m})}\right) + \exp\left(-\frac{m\epsilon_t(\bar{r}_t')^2}{4K}\right)\right\}, \tag{A.165}$$

which is precisely (24).

When $\epsilon_t = 0$, the contamination term disappears. The miss event is then controlled directly by the cleaned embedding difference, and the bound reduces to

$$p_{\text{miss},t} \le \exp\left(-\frac{(2\bar{a}_t - \bar{c}_0)_+^2}{2K(\frac{1}{n}+\frac{1}{m})}\right). \tag{A.166}$$

The first exponential controls the joint fluctuations of the reference and cleaned test embeddings, whereas the second accounts for contamination of the test window. Both residual margins must be positive for the two-term bound to be potentially informative.

Unlike the original MMD bound, (A.165) retains the reference-window fluctuation explicitly. For fixed m and ν, increasing n lowers $\bar{c}_0$ and enlarges $\bar{r}_t$, while reducing the $1/n$ contribution in the first exponent's denominator. These improvements diminish as n grows beyond m. Indeed, as $n \to \infty$, $\bar{c}_0 \to \sqrt{K/m}$, and the denominator of the first exponent approaches $2K/m$.

Compared with the Pearson detector, MMD avoids estimating the regularized inverse of the nominal second-moment operator, eliminating the associated metric-perturbation terms. In return, it uses the unweighted RKHS norm rather than adapting the discrepancy to the nominal second-moment structure. The reference-embedding estimation error, however, is common to both detectors.

A.5.4 Proof of Corollary 4.2: Operational miss-probability bound

Proof. From the explicit form of (A.136), with $K = 1$,

$$p_{\text{miss},t} \le \min\{1, T_1 + T_2 + T_3 + T_4\}, \tag{A.167}$$

where

$$T_1 := 2\exp\left(-\frac{m\rho a_t^2}{2}\right), \qquad T_2 := 2\exp\left(-\frac{m\rho a_t^2}{4\epsilon_t}\right), \tag{A.168}$$

and

$$T_3 := 8\exp\left(-\min\left\{\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t^2}{128(1+\rho)}, \frac{\rho^3\|\Delta\|_{\mathcal{H}_k} n r_t}{256\sqrt{1+\rho}}\right\}\right), \tag{A.169}$$

$$T_4 := 8\exp\left(-\min\left\{\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t'^2}{32(1+\rho)}, \frac{\rho^3\|\Delta\|_{\mathcal{H}_k} n r_t'}{256\sqrt{1+\rho}}\right\}\right). \tag{A.170}$$

For the first two terms,

$$\min\left\{\frac{m\rho a_t^2}{2}, \frac{m\rho a_t^2}{4\epsilon_t}\right\} = \frac{m\rho a_t^2}{4\max\{\epsilon_t, 1/2\}}, \tag{A.171}$$

and therefore

$$T_1 + T_2 \le 4\exp\left(-\frac{m\rho a_t^2}{4\max\{\epsilon_t, 1/2\}}\right). \tag{A.172}$$

For T_3, the ratio between the linear and quadratic exponents is

$$\frac{\rho^3\|\Delta\|_{\mathcal{H}_k} n r_t/\big(256\sqrt{1+\rho}\big)}{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t^2/\big(128(1+\rho)\big)} = \frac{\sqrt{1+\rho}}{2\rho\|\Delta\|_{\mathcal{H}_k} r_t}. \tag{A.173}$$

The first condition in (25) therefore ensures that the quadratic exponent is the smaller one, and hence

$$T_3 = 8\exp\left(-\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t^2}{128(1+\rho)}\right). \tag{A.174}$$

Similarly, for T_4 the corresponding ratio is

$$\frac{\sqrt{1+\rho}}{8\rho\|\Delta\|_{\mathcal{H}_k} r_t'}, \tag{A.175}$$

so that the second condition in (25) yields

$$T_4 = 8\exp\left(-\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t'^2}{32(1+\rho)}\right). \tag{A.176}$$

Since

$$\min\left\{\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t^2}{128(1+\rho)}, \frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n r_t'^2}{32(1+\rho)}\right\} \ge \frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n(\min\{r_t, r_t'\})^2}{128(1+\rho)}, \tag{A.177}$$

we obtain

$$T_3 + T_4 \le 16\exp\left(-\frac{\rho^4\|\Delta\|_{\mathcal{H}_k}^2 n(\min\{r_t, r_t'\})^2}{128(1+\rho)}\right). \tag{A.178}$$

Combining (A.172) and (A.178) and absorbing the numerical factors into strictly positive universal constants κ_3 and κ_4 yields the upper bound. □

A.5.5 MMD SPECIALIZATION OF THE OPERATIONAL MISS-PROBABILITY BOUND

For the MMD detector, let $K = 1$, $n \ge m$, and $\epsilon_t \in (0, 1)$, and define

$$\bar{r}_{t,\mathrm{op}} := \big(\bar{a}_t - \max\{\bar{c}_0, \bar{d}_0\}\big)_+.$$

From (24), we obtain

$$p_{\mathrm{miss},t} \le \min\left\{1,\ \exp\left(-\frac{m\bar{r}_{t,\mathrm{op}}^2}{2(1+m/n)}\right) + \exp\left(-\frac{m\bar{r}_{t,\mathrm{op}}^2}{4\epsilon_t}\right)\right\}. \tag{A.179}$$

Since

$$\min\left\{\frac{m\bar{r}_{t,\mathrm{op}}^2}{2(1+m/n)}, \frac{m\bar{r}_{t,\mathrm{op}}^2}{4\epsilon_t}\right\} = \frac{m\bar{r}_{t,\mathrm{op}}^2}{4\max\{\epsilon_t, (1+m/n)/2\}},$$

we obtain

$$p_{\mathrm{miss},t} \le \min\left\{1,\ 2\exp\left(-\frac{m\bar{r}_{t,\mathrm{op}}^2}{4\max\{\epsilon_t, (1+m/n)/2\}}\right)\right\}. \tag{A.180}$$

A.5.6 Proof of Theorem 4.2: Detection-delay tail

Proof. By the set-inclusion relation (26),

$$\mathbb{P}(D > d) \le p_{\mathrm{miss},t+d}. \tag{A.181}$$

Applying Corollary 4.2 at time $t+d$ gives

$$\begin{aligned}\mathbb{P}(D > d) \le \min\Big\{1,\ \kappa_3\Big[\exp\left(-\kappa_4 \frac{m\rho a_{t+d}^2}{\max\{\epsilon_{t+d}, 1/2\}}\right) \\ + \exp\left(-\kappa_4 \frac{\rho^4 \|\Delta\|_{\mathcal{H}_k}^2 n(\min\{r_{t+d}, r'_{t+d}\})^2}{1+\rho}\right)\Big]\Big\},\end{aligned} \tag{A.182}$$

provided that the range conditions (25) hold at time $t+d$. This is precisely (27). □

In the sliding-window regime,

$$\tau_{t+d} = (m-1-d)_+, \qquad \epsilon_{t+d} = \frac{(m-1-d)_+}{m}. \tag{A.183}$$

Hence ϵ_{t+d} is nonincreasing during the warm-up phase. Since

$$S_{\epsilon_{t+d}} = (1-\epsilon_{t+d})\|\Delta\|_{A^{-1}}, \tag{A.184}$$

the effective signal $S_{\epsilon_{t+d}}$, and therefore a_{t+d} and r_{t+d}, are nondecreasing as the contamination decreases. The same holds for r'_{t+d} as long as $\epsilon_{t+d} > 0$. Consequently, within the range where Corollary 4.2 applies, both exponents in (27) increase as the test window is progressively filled with post-change observations.

A.5.7 MMD specialization of the detection-delay tail

Let $t = t_0 - m + 1$ denote the first split index whose test window contains a post-change observation. For $0 \le d \le m-2$, the reference window at $t+d$ remains nominal, with contamination level

$$\epsilon_{t+d} = \frac{m-1-d}{m} \in (0,1).$$

The same set-inclusion argument gives $\mathbb{P}(D > d) \le p_{\mathrm{miss},t+d}$. Applying Corollary 4.1 yields

$$\mathbb{P}(D > d) \le \min\left\{1,\ \exp\left(-\frac{\bar{r}_{t+d}^2}{2K(\frac{1}{n}+\frac{1}{m})}\right) + \exp\left(-\frac{m\epsilon_{t+d}(\bar{r}'_{t+d})^2}{4K}\right)\right\}, \tag{A.185}$$

where all MMD-specific quantities are evaluated at ϵ_{t+d}.

For $K = 1$ and $n \ge m$, the operational bound (A.180) gives

$$\mathbb{P}(D > d) \le \min\left\{1,\ 2\exp\left(-\frac{m\bar{r}_{t+d,\mathrm{op}}^2}{4\max\{\epsilon_{t+d}, (1+m/n)/2\}}\right)\right\}. \tag{A.186}$$

At $d = m-1$, the test window is fully post-change, namely, $\epsilon_{t+d} = 0$, while the reference remains nominal. The contamination term disappears, yielding

$$\mathbb{P}(D > m-1) \le \exp\left(-\frac{(2\bar{a}_{t_0} - \bar{c}_0)_+^2}{2K(\frac{1}{n}+\frac{1}{m})}\right). \tag{A.187}$$

Unlike the Pearson bound, the MMD delay bound contains no metric-perturbation terms. Nevertheless, the reference-window length enters explicitly through the estimation of μ_P. For $d \ge m$, the sliding reference window may contain post-change observations, so the preceding bounds do not directly apply.

Table 5: MNIST $0 \to 8$: detection delay versus test length m, with $n = 120$ and in-window detection probability approximately one.

m	12	16	20	30	40	60
Mean D	6.7±.8	8.2±1.1	9.9±1.2	11.8±.7	13.6±.6	18.2±1.0
D/m	0.56	0.51	0.49	0.39	0.34	0.30

Table 6: MNIST $0 \to 8$: empirical miss probability versus regularization ρ, with $m = 4$.

n	$\rho = 10^{-3}$	10^{-2}	10^{-1}	1	10
40	.345±.110	.381±.158	.265±.156	.020±.017	.004±.004
150	.053±.027	.151±.038	.104±.039	.004±.004	.000±.000
400	.060±.039	.259±.196	.153±.114	.001±.002	.001±.002

A.6 Empirical threshold calibration

The analytical threshold in Corollary 3.1 guarantees an ARL of at least γ, but may be conservative because of the concentration inequalities and subsequent relaxations. We therefore consider empirical calibration using nominal data only.

For a fixed reference window $X_{1:n} \sim P$, define the conditional null exceedance probability

$$p(\nu; X_{1:n}) := \mathbb{P}_\infty\big(\hat{\chi}^2_\rho[X_{1:n} \mid Y_{1:m}] > \nu \mid X_{1:n}\big)\,, \tag{A.188}$$

where $Y_{1:m} \sim P$ is independent of the reference.

Under the fixed-reference approximation with independent successive analysis windows, the waiting time to the first exceedance is geometric, with conditional expectation $1/p(\nu; X_{1:n})$. Averaging over the reference distribution yields the calibration functional

$$\mathrm{ARL}_{\mathrm{fix}}(\nu) := \mathbb{E}_{X_{1:n}}\left[\frac{1}{p(\nu; X_{1:n})}\right]. \tag{A.189}$$

To estimate this functional, generate N_{ref} independent nominal reference windows $X^{(i)}_{1:n} \sim P$ and, for each reference, N_{ana} independent nominal analysis windows $Y^{(i,j)}_{1:m} \sim P$.

For each candidate threshold ν, estimate the conditional exceedance probability by

$$\hat{p}_i(\nu) := \max\left\{\frac{1}{N_{\mathrm{ana}}}\sum_{j=1}^{N_{\mathrm{ana}}} \mathbf{1}\left\{\hat{\chi}^2_\rho[X^{(i)}_{1:n} \mid Y^{(i,j)}_{1:m}] > \nu\right\}, p_{\min}\right\}. \tag{A.190}$$

The floor $p_{\min} > 0$ prevents infinite estimates when no exceedance is observed. In our experiments, we use $p_{\min} = \frac{1}{3N_{\mathrm{ana}}}$.

The empirical calibration curve is then

$$\hat{\mathrm{ARL}}_{\mathrm{fix}}(\nu) := \frac{1}{N_{\mathrm{ref}}}\sum_{i=1}^{N_{\mathrm{ref}}} \frac{1}{\hat{p}_i(\nu)}. \tag{A.191}$$

Evaluate this curve on a grid $\nu_1 < \cdots < \nu_R$ and select

$$\hat{\nu}_{\mathrm{ARL}}(\gamma) := \inf\left\{\nu_r : \hat{\mathrm{ARL}}_{\mathrm{fix}}(\nu_r) \geq \gamma\right\}. \tag{A.192}$$

Linear interpolation between adjacent grid points may be used for finer calibration. Since the same simulated statistics are reused across thresholds, the empirical curve is nondecreasing by construction. The probability floor also implies $\hat{\mathrm{ARL}}_{\mathrm{fix}}(\nu) \leq 3N_{\mathrm{ana}}$, so N_{ana} must be chosen sufficiently large relative to the target ARL.

Table 7: MNIST $0 \to 0$: realized ARL after empirical threshold calibration, with $n = 80$, $m = 40$, and $\rho = 10^{-2}$.

Target ARL γ	100	200	500	1000
Realized ARL	101 ± 22	239 ± 87	660 ± 167	1147 ± 83
Relative error	$+1\%$	$+20\%$	$+32\%$	$+15\%$

A.7 Additional experiments on MNIST

We evaluate the detector on MNIST images ($d = 784$), with normalized pixel intensities in $[0, 1]$. The nominal distribution P consists of digit zeros, while the post-change distribution Q consists of digit eights, following (Kalinke & Gavioli-Akilagun, 2026). The Gaussian bandwidth is selected using nominal zeros ($\sigma \approx 10$), and thresholds are calibrated on held-out zeros only. No post-change images are used for tuning.

Reference window n. With $m = 6$, the empirical miss probabilities are

$$0.301 \pm .033, \quad 0.161 \pm .023, \quad 0.044 \pm .015, \quad 0.018 \pm .009, \quad 0.002 \pm .004$$

for $n = 10$, 20, 40, 80, and 160, respectively. Standard deviations are computed over four independent evaluation sets. Increasing n substantially improves detection, consistently with improved estimation of the nominal embedding and metric.

Test window m. Table 5 reports the mean delay for $n = 120$ in the reliable in-window detection regime. The absolute delay increases with m, whereas D/m decreases from 0.56 to 0.30. Once reliable detection is achieved, a shorter test window therefore offers greater responsiveness.

Regularization ρ. Table 6 reports the empirical miss probability for $m = 4$ and three reference sizes. Unlike the synthetic experiments, stronger regularization substantially improves detection, particularly for short reference windows. The empirical reference operator has rank at most n, so small ρ can amplify poorly estimated directions. Larger regularization stabilizes the inverse metric, highlighting the importance of adapting ρ to the available reference data.

Threshold calibration to a target ARL. Finally, we calibrate the constant threshold on held-out zeros and evaluate the realized ARL on fresh nominal data. Table 7 reports the results for $n = 80$, $m = 40$, and $\rho = 10^{-2}$. The empirical threshold calibration remains effective in this high-dimensional setting, although deviations increase at larger targets.