WHAT TO LEARN FOR SEQUENTIAL LOG-LIKELIHOOD RATIO TESTS?

Anonymous authors
Paper under double-blind review

ABSTRACT

In a learned log-likelihood ratio (LLR) test, one checkpoint, or a model state saved during training, must be selected to generate the LLR estimates used for threshold-based decisions. However, we find that conventional validation criteria can favor checkpoints whose estimated LLR trajectories poorly match the shape of the true LLR. Moreover, multiplying both the LLR and thresholds by the same positive constant leaves stopping times and decisions unchanged, thus test results alone cannot determine exact-scale LLR estimation error. To establish a proper validation metric for checkpoint selection, we therefore shift the estimand to the positive-projectivized LLR (P_+LLR), the true LLR up to positive rescaling. With correspondingly rescaled decision thresholds, P_+LLR gives exactly the same stopping decisions as the true LLR to preserve optimality guarantees of statistical tests. P_+LLR lets us introduce a validation metric Normalized Integrated Tradeoff Area (NITA), which is provably ordered with an oracle P_+LLR estimation error in a high-threshold detection task under an i.i.d. data stream with independent noise. We also establish a finite-sample checkpoint-selection guarantee for empirical NITA and provide a unified extension of the theory to broader non-i.i.d., multiclass data and classification tasks, accompanied by experiments on 12 diverse datasets.

1 INTRODUCTION

Two foundational sequential tests are the Sequential Probability Ratio Test (SPRT) for hypothesis testing and the Cumulative Sum (CUSUM) algorithm for Quickest Changepoint Detection (QCD) (Wald, 1945; Page, 1954). SPRT supports early classification (Xing et al., 2012; Mori et al., 2018), trading off latency against terminal accuracy, while CUSUM supports monitoring, trading off detection delay against false alarms. Both tests use the log-likelihood ratio (LLR)-based decision score: with the true LLR, SPRT minimizes expected sample size at prescribed error levels, while CUSUM minimizes detection delay subject to a false-alarm constraint (Wald & Wolfowitz, 1948; Moustakides, 1986).

In practice, the true LLR is unknown and must be estimated from data, for example through direct density-ratio estimation (DRE), class-posterior estimation, or surrogate scores (Kawahara & Sugiyama, 2012; Moustakides & Basioti, 2019; Banerjee & Tarokh, 2024). During training of the estimator, *checkpoints*, or model states with trained parameters, are saved at particular epochs. From a practitioner's point of view, selecting one best checkpoint for the deployment is an important problem. However, the checkpoint is commonly selected based on cross-entropy (CE) or classification accuracy (Acc), and are only evaluated afterwards via a speed-accuracy tradeoff (SAT) curve (Moustakides & Basioti, 2019; Ebihara et al., 2021; Gong et al., 2024), which create a model-selection problem.

We find that selecting checkpoints by CE or Acc can be highly sensitive to the validation protocol under finite data and limited model capacity, potentially favoring checkpoints whose estimated LLR trajectories poorly match the shape of the true trajectory (Fig. 2e). Although CE is a strictly proper *training* loss for class-posterior estimation (Gneiting & Raftery, 2007), and Acc is a natural per-time-step classification metric, they are not necessarily reliable *validation metrics for sequential LLR tests*: sequential decisions depend on how evidence accumulates until a threshold is crossed, whereas CE and Acc score individual predictions at each time step, even when averaged over time.

A natural alternative is to validate directly on the SAT curve, obtained by varying the thresholds. However, these operating curves are *positive-scale invariant*: doubling the LLR and all thresholds

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.

preserves every stopping time and decision. Such rescaling can be accommodated by *threshold calibration*, which chooses thresholds for a desired error or false-alarm level. Moreover, a distorted score may produce the same stopping time and decision on a particular sequence at one threshold (Fig. 1). Such agreement alone does not guarantee equivalent behavior on unseen data. We prove that as long as we insist on the *exact-scale* LLR as the estimand, no metric derived from operating curves can faithfully represent the LLR estimation error. We thus ask the central question of this paper:

What learning target permits principled checkpoint selection for calibrated sequential tests?

Our answer is that the target is not the exact-scale LLR trajectory, but its *positive ray*, or the set of all positive rescalings of the trajectory, which we call the *Positive-projectivized Log-likelihood Ratio (*P_+LLR*)*. P_+LLR preserves the shape of the true LLR trajectory while discarding a global scale that can be absorbed by threshold calibration, maintaining the stopping times and decisions unchanged.

The oracle error is *Scale-Invariant Normalized Error (SINE)*, which vanishes if and only if the learned trajectory lies on the same positive ray. The corresponding practical validation metric is *Normalized Integrated Tradeoff Area (NITA)*. Note that computing SINE requires the true LLR, whereas computing NITA does not, making it a practical error surrogate. Under binary i.i.d. QCD task with affine independent noise, we prove an *asymptotic antitone relation*: lower SINE implies larger NITA on common high-threshold regions. At large thresholds, false-alarm rate improves exponentially as more detection delay is allowed. Our corollary quantifies how small perturbations reduce this growth rate in proportion to squared SINE. Separately, we prove that the estimation error of empirical NITA is bounded uniformly across candidate checkpoints. This theorem thus provides guarantee for empirical checkpoint selection with finite samples, given a fixed validation protocol.

As a first step toward principled checkpoint selection through P_+LLR learning, the main theory in this paper is built on the cleanest but somewhat restrictive settings. However, we show the same SINE-NITA antitone can be established under broader settings, i.e., multiclass, non-i.i.d., and SPRT settings with setting-specific additional assumptions. Moreover, we present these assumptions as sufficient conditions to formulate the theory, instead of claiming that all are necessary in practice.

Indeed, extensive experiments support the SINE–NITA ordering beyond the binary i.i.d. QCD regime. Across 12 datasets spanning binary and multiclass tasks, i.i.d. and non-i.i.d. dynamics, synthetic and real data, both SPRT/CUSUM, and high-threshold/full-threshold settings, the predicted SINE–NITA antitone is statistically significant. On unseen test data, NITA-based checkpoint selection yields statistically significantly faster sequential decisions than CE/Acc-based selection and alternative baselines at both the 99% and 50% points of each setting's common attainable performance range.

2 Calibrated Sequential Tests with a Learned LLR

Data stream and LLR evidence. Observations arrive sequentially as $X_1, X_2, \ldots$, with $X_t \in \mathcal{X}$ for $t \geq 1$, where $\mathcal{X}$ is the observation space. We write $X^{(1,t)} := (X_1, \ldots, X_t)$ for the observations available by time t, and call a finite sequence $X^{(1,T)}$ an observation *record* of length T. We write $[T] := \{1, \ldots, T\}$. The natural filtration $(\mathcal{F}_t)_{t\geq 1}$, with $\mathcal{F}_t := \sigma(X^{(1,t)})$, represents the information available up to time t. Whether a sequential rule should stop by time t is determined by $\mathcal{F}_t$. The terminal decision is made only with $\mathcal{F}_\tau$, the information available at stopping.

The same LLR evidence is used both in SPRT and CUSUM. For SPRT, a fixed label or hypothesis $Y \in \{0, \ldots K\}$ governs the entire stream where $K \in \mathbb{N}$ is class size. For QCD, the label is 0 before an unknown changepoint ν and $k \in [K]$ from time ν onward. $\nu = \infty$ indicates no change. In the main theory we focus on $K = 1$ (binary). The one-step and cumulative LLR are

$$\lambda^t := \log \frac{p_1(X_t \mid X^{(1,t-1)})}{p_0(X_t \mid X^{(1,t-1)})}, \qquad \Lambda^{1:t} := \sum_{s=1}^{t} \lambda^s. \tag{1}$$

Wald's Sequential Probability Ratio Test (SPRT) selects one of the two fixed hypotheses as soon as the cumulative LLR crosses the decision threshold. Given positive numbers a and b, called the upper and lower *decision thresholds*, the SPRT continues while $\Lambda^{1:t} \in (-b, a)$ and stops at

$$\tau_{\mathrm{SPRT}}(a, b) := \inf\{t \geq 1 : \Lambda^{1:t} \notin (-b, a)\}. \tag{2}$$

At the stopping time it declares $d_{\tau_{\mathrm{SPRT}}} = 1$ or $d_{\tau_{\mathrm{SPRT}}} = 0$ after crossing the upper or lower threshold, respectively. Its speed is measured by the mean hitting time $\mathrm{MHT} := \mathbb{E}[\tau_{\mathrm{SPRT}}]$, and its decision quality by the *terminal* accuracy $\mathrm{Acc}_\tau := \mathbb{P}(d_{\tau_{\mathrm{SPRT}}} = Y)$. Here Acc_τ evaluates the decision at stopping time, whereas local Acc appeared in Introduction evaluates forced decisions at each time t using the accumulated LLR's sign (i.e., forced decisions by collapsing the threshold to zero).

Under the classical simple i.i.d. hypotheses with the true LLR, the SPRT minimizes expected sample size satisfying prescribed error constraints (Wald & Wolfowitz, 1948; Tartakovsky et al., 2014).

Page's Cumulative Sum (CUSUM) algorithm raises an alarm when there is enough accumulated evidence that a change has occurred. Given an alarm threshold $h > 0$, define

$$W^0 := 0, \qquad W^t := \max\{0, W^{t-1} + \lambda^t\}, \qquad \tau_{\mathrm{CUSUM}}(h) := \inf\{t \geq 1 : W^t \geq h\}. \tag{3}$$

The reset to zero discards accumulated evidence favoring the baseline and lets the procedure reconsider where an unknown change may have started. Let $\mathbb{E}_\infty$ denote expectation when no change occurs and $\mathbb{E}_{\nu=1}$ expectation when the change occurs immediately. We use

$$\mathrm{ARL}(h) := \mathbb{E}_\infty[\tau_{\mathrm{CUSUM}}(h)], \qquad \mathrm{ADD}_1(h) := \mathbb{E}_{\nu=1}[\tau_{\mathrm{CUSUM}}(h) - 1] \tag{4}$$

as the average run length to a false alarm and the average detection delay, respectively. Detection delay is measured from the changepoint, $\tau - \nu$ for $\tau \geq \nu$. Our theory studies the immediate-change case $\nu = 1$. See Appendix D.1 for how delay and false alarms are evaluated on finite observation.

Under the classical i.i.d. hypotheses with the true LLR, CUSUM is minimax optimal for worst-case detection delay subject to a false-alarm constraint (Lorden, 1971; Moustakides, 1986).

Threshold calibration is to choose the threshold to satisfy a desired false-alarm level or terminal accuracy, either analytically or by using held-out data. A **threshold sweep** is to evaluate the same trained model over a collection of threshold values. Each threshold gives an *operating point*: $(\mathrm{ADD}_1, \mathrm{ARL})$ for CUSUM or $(\mathrm{MHT}, \mathrm{Acc}_\tau)$ for SPRT. Collection of them delineate *threshold-swept operating curve*.

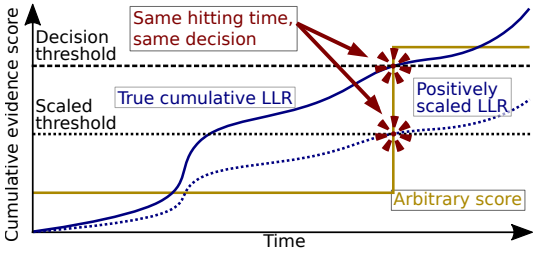


Figure 1: **Stopping outcomes does not identify the LLR.** The true LLR (blue) and its positive rescaling (dotted blue) cross their correspondingly scaled thresholds at the same time and select the same hypothesis. This rescaling holds for every sequence (Proposition 2.1). The arbitrary score (yellow) matches the stopping outcomes, although such agreement generally does not have guarantee to generalize on other unknown sequences.

The following elementary results are the starting point of our paper. Let $g = (g^t)_{t\geq 1}$ be a real-valued evidence process, where each one-step score g^t uses only observations available by time t. For a fixed constant $\alpha > 0$, shared across times and sequences, write $\alpha g := (\alpha g^t)_{t\geq 1}$. Define the SPRT and CUSUM rules by replacing λ^t with g^t in Eq. (2)–Eq. (3). Write $\delta_{\mathrm{SPRT}}(a, b; g) := (\tau_{\mathrm{SPRT}}(a, b; g), d_{\tau_{\mathrm{SPRT}}}(a, b; g))$ for the SPRT stopping time and terminal decision based on g.

Proposition 2.1 (Positive score rescaling is absorbed by threshold calibration)**.** *For every realized evidence sequence g and every $\alpha > 0$, $\delta_{\mathrm{SPRT}}(a, b; g) = \delta_{\mathrm{SPRT}}(\alpha a, \alpha b; \alpha g)$, $\tau_{\mathrm{CUSUM}}(h; g) = \tau_{\mathrm{CUSUM}}(\alpha h; \alpha g)$. If the threshold family is closed under positive rescaling, g and αg have identical threshold-swept operating curves. Let $\mathcal{C}(g)$ denote the operating points. Then*

$$\mathcal{C}(\alpha g) = \mathcal{C}(g) \qquad \textit{for every } \alpha > 0. \tag{5}$$

Proof. For SPRT, the accumulated score under αg is $\alpha \sum_{s=1}^t g^s$. Because $\alpha > 0$, crossing a or $-b$ is equivalent to crossing αa or $-\alpha b$, including which boundary is crossed. For CUSUM, let W_g^t and $W_{\alpha g}^t$ denote the two recursions. Induction gives $W_{\alpha g}^t = \alpha W_g^t$ for every t, therefore $W_g^t \geq h$ if and only if $W_{\alpha g}^t \geq \alpha h$, and sweeping the rescaled thresholds gives the same operating curve. □

Proposition 2.1 is exact without asymptotic approximations. It says that positive rescaling leaves attainable sequential performance unchanged (Fig. 1), even though an exact-scale estimation error may penalize it. Appendix B.1 gives the corresponding general statements for the multiclass procedures.

Corollary 2.2 (Exact scale is not identifiable from a calibrated operating curve)**.** *Use the operating-curve setup of Proposition 2.1, including a threshold family closed under positive rescaling. Write* $\mathbf{\Lambda}^\star := (\sum_{s=1}^t \lambda^s)_{t\geq 1},\ \widehat{\mathbf{\Lambda}} := (\sum_{s=1}^t g^s)_{t\geq 1}$ *for the true and learned cumulative score processes, respectively, with* $\mathbf{\Lambda}^\star$ *not identically zero.*
Let $S(\widehat{\mathbf{\Lambda}})$ *be a real-valued validation score depending only on the induced operating curve. Suppose an exact-scale error* $E_{\mathrm{exact}}(\widehat{\mathbf{\Lambda}}, \mathbf{\Lambda}^\star)$ *distinguishes the true LLR from at least one positive rescaling:*

$$E_{\mathrm{exact}}(\alpha\mathbf{\Lambda}^\star, \mathbf{\Lambda}^\star) \neq E_{\mathrm{exact}}(\mathbf{\Lambda}^\star, \mathbf{\Lambda}^\star) \quad \textit{for some } \alpha > 0,\ \alpha \neq 1. \tag{6}$$

Then no function ψ *mapping validation scores to errors can satisfy* $E_{\mathrm{exact}}(\widehat{\mathbf{\Lambda}}, \mathbf{\Lambda}^\star) = \psi\Big(S(\widehat{\mathbf{\Lambda}})\Big)$ *throughout any candidate family containing* $\{\alpha\mathbf{\Lambda}^\star : \alpha > 0\}$*. In particular,* S *cannot reproduce every strict ordering induced by the exact-scale error.*

The proof by contradiction is shown in Appendix C.1: Proposition 2.1 gives $S(\alpha\mathbf{\Lambda}^\star) = S(\mathbf{\Lambda}^\star)$, whereas Eq. (6) requires the corresponding exact-scale errors to differ.

3 Theory

3.1 Projective estimand, oracle error, and observable surrogate

Proposition 2.1 shows that positive rescaling of the cumulative LLR can be absorbed by corresponding threshold rescaling. We therefore reframe all such rescalings as equivalent learning targets. For a fixed horizon T, we represent each realized cumulative trajectory as one vector $\mathbf{\Lambda}_T := (\Lambda^{1:1}, \ldots, \Lambda^{1:T}) \in \mathbb{R}^T$, with one coordinate per time step. For a nonzero trajectory, its positive multiples form a *positive ray*: a half-line in this vector space extending from the origin through $\mathbf{\Lambda}_T$, excluding the origin.

Definition 3.1 ($\mathrm{P}_+\mathrm{LLR}$)**.** For a nonzero cumulative LLR trajectory $\mathbf{\Lambda}_T$, its *positive-projectivized LLR* is the positive ray

$$[\mathbf{\Lambda}_T]_+ := \{\alpha\mathbf{\Lambda}_T : \alpha > 0\} \subset \mathbb{R}^T. \tag{7}$$

A learned trajectory $\widehat{\mathbf{\Lambda}}_T$ is *projectively correct* if $\widehat{\mathbf{\Lambda}}_T \in [\mathbf{\Lambda}_T]_+$, meaning that it equals to the true cumulative LLR trajectory up to one global positive scale.

$\mathrm{P}_+\mathrm{LLR}$ preserves the direction and relative shape of accumulated evidence, while discarding only its recalibratable scale. The multiplier α is one global constant shared across all records and times.

Oracle projective error: SINE. Because the target is a ray rather than one point, the relevant error is the distance from the learned trajectory to the closest positive multiple of the true trajectory.

Definition 3.2 (Scale-Invariant Normalized Error (SINE))**.**

$$\mathrm{SINE}(\hat{v}, v) := \frac{\min_{\beta\geq 0} \|\hat{v} - \beta v\|}{\|\hat{v}\|} = \sqrt{1 - \rho(\hat{v}, v)^2}, \tag{8}$$

where v and $\hat{v}$ are nonzero elements of a real Hilbert space. $\rho(\hat{v}, v) := \max\left\{0, \frac{\langle \hat{v}, v\rangle}{\|\hat{v}\|\,\|v\|}\right\}$ is the nonnegative part of their cosine similarity.

SINE lies in $[0, 1]$. For cumulative trajectories, take $v = \mathbf{\Lambda}_T$ and $\hat{v} = \widehat{\mathbf{\Lambda}}_T$. The resulting error is zero exactly when $\widehat{\mathbf{\Lambda}}_T = \beta\mathbf{\Lambda}_T$ for some $\beta > 0$. When their inner product is positive, SINE is literally the sine function of the angle between the learned trajectory and the true positive ray. SINE requires the ground-truth LLR and is therefore an *oracle diagnostic*: it is only available in synthetic settings.

Observable operating-curve surrogate: NITA. The practical checkpoint selector must use only validation outcomes from the threshold sweep described in Section 2. Let Θ denote the prespecified collection of thresholds used in the sweep (a finite grid in implementation). For $\theta \in \Theta$, where $\theta = h$ for CUSUM or $\theta = (a, b)$ for SPRT, let $(x(\theta), y(\theta))$ be the resulting operating point, with smaller x and larger y preferred. Fix a *common evaluation rectangle* $[x_0, X^*] \times [0, Y^*]$, where $0 \leq x_0 < X^*$ and $Y^* > 0$. The rectangle is part of the validation protocol and is held fixed across checkpoints. Raw operating values outside it are clipped to its boundary before the area is computed. Define the *upper envelope* $y^\star(x) := \sup\{y(\theta) : x(\theta) \leq x,\ \theta \in \Theta\},\ x \in [x_0, X^*]$, or the largest y attainable with cost no greater than x. If no threshold satisfies the constraint, the supremum is defined as 0.

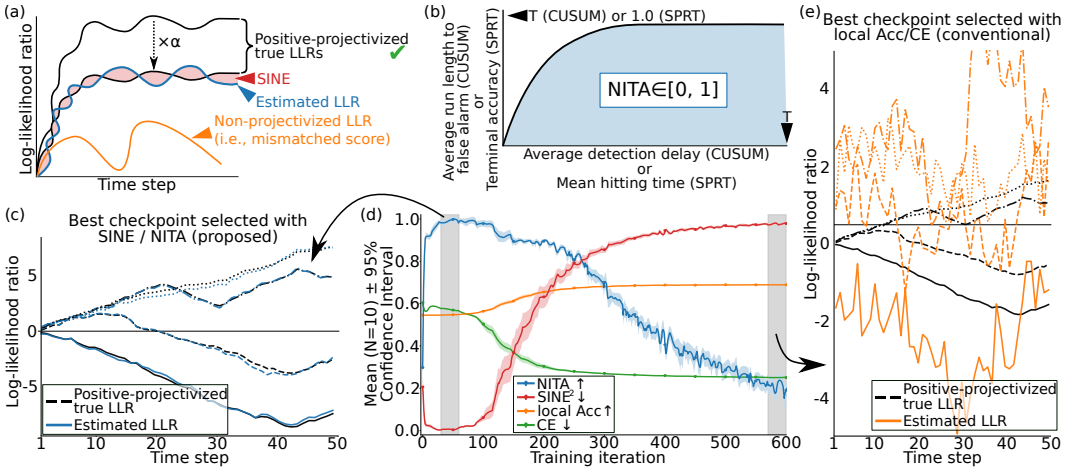


Figure 2: **NITA selects checkpoints that better preserve the true LLR trajectory shape.** (a) SINE measures the normalized distance from an estimated cumulative LLR trajectory to the true LLR's positive ray. (b) NITA is the normalized area under the operating curve's upper envelope, which is computable without true LLR. (d) During the training of LLR estimator, CE and local Acc improve while squared SINE rises and NITA falls (mean traces with 95% confidence intervals over 10 runs are shown). (c,e) In the illustrated runs, checkpoints selected by NITA or oracle SINE yield trajectories close to positively rescaled true LLRs (c), whereas CE or local Acc selects distorted trajectories (e).

Definition 3.3 (Normalized Integrated Tradeoff Area (NITA))**.**

$$\mathrm{NITA}_{Y^*}^{[x_0,X^*]} := \frac{1}{(X^*-x_0)Y^*}\int_{x_0}^{X^*} y^\star(x)\,dx. \tag{9}$$

NITA is the fraction of the common rectangle lying below the best attainable operating frontier, thus it belongs to $[0,1]$ and larger is better. For CUSUM we use $(x,y) = (\mathrm{ADD}_1, \mathrm{ARL})$. For SPRT we use $(x,y) = (\mathrm{MHT}, \mathrm{Acc}_\tau)$. When $x_0 = 0$, we abbreviate Eq. (9) as NITA and call it *full-box NITA*. When $x_0 > 0$, we call it *high-threshold NITA*. In QCD this focuses comparison on the high-false-alarm-protection region analyzed by our main theorem. All checkpoints must use the same rectangle, because checkpoint-dependent normalizers can reverse an otherwise valid frontier ordering (Appendix B.4). Unlike SINE, NITA requires no true LLR, serving as a practical validation metric.

The central theoretical question is now precise: under what conditions does a checkpoint with smaller oracle projective error (SINE) also have a larger observable operating area (NITA)? We answer this first in the cleanest setting—binary i.i.d. QCD with one-sided CUSUM—and use the later extensions and experiments to examine how far the same mechanism persists beyond this regime.

3.2 Main theorem: binary i.i.d. QCD

We now connect cumulative-trajectory SINE to high-threshold NITA under a stochastic model of learned scores. Let P_0 and P_1, with densities p_0 and p_1 relative to a common reference measure, denote the distributions of a single observation before and after the change, respectively. Observations are independent, with the same distribution within each regime. Assume $P_0 \neq P_1$ and mutual absolute continuity, written $P_0 \sim P_1$, meaning that the two assign zero probability to the same events.

To compare cumulative trajectories across validation records, let Q_{val} be a fixed probability distribution over length-T records, shared across the compared scores, and let $V \sim Q_{\mathrm{val}}$ denote one such random record. Its true cumulative LLR vector is $\mathbf{\Lambda}_T(V)$. We use the population Hilbert space

$$\mathcal{H}_\Sigma := L^2(Q_{\mathrm{val}};\mathbb{R}^T), \qquad \langle f,g\rangle_{\mathcal{H}_\Sigma} := \mathbb{E}_{V\sim Q_{\mathrm{val}}}\big[\langle f(V),g(V)\rangle_{\mathbb{R}^T}\big], \tag{10}$$

whose elements are trajectory-valued functions with finite expected squared Euclidean norm. Here $f(V)$ and $g(V)$ are the trajectory vectors assigned to record V by $f,g \in \mathcal{H}_\Sigma$. We use the same notation SINE_Σ for Definition 3.2: the minimizing coefficient β is shared across all records and time steps. Empirical SINE_Σ replaces the expectation in Eq. (10) by the sample average (Lemma B.10).

Assumption 3.4 (Affine independent-noise family)**.** There exist a fixed signal scale $\alpha > 0$ and a nuisance random variable ξ such that the one-step scores are $g_\sigma = \alpha\lambda + \sigma\xi,\ 0 \le \sigma \le \bar{\sigma} < \infty$.

Here σ controls the nuisance magnitude, while λ and ξ remain unchanged across the family.

(1) Noise distribution. Under P_0, ξ is independent of the true LLR λ. Under both P_0 and P_1, it has the same distribution Q, with $\mathbb{E}_Q[\xi] = 0$, and is not almost surely constant.

(2) Exponential moments. The moment-generating function (MGF) $M_\xi(z) := \mathbb{E}_Q[e^{z\xi}]$, for real z, is finite on an open interval containing $[0, \bar{\sigma}/\alpha]$.

(3) Cumulative geometry. Under each fixed regime, $(\lambda^t, \xi^t)_{t\ge1}$ are i.i.d. copies of (λ, ξ). Define the cumulative nuisance by $\boldsymbol{\eta}_{\Sigma,T} := (\sum_{s=1}^t \xi^s)_{t\in[T]}$. Summing $g_\sigma^t = \alpha\lambda^t + \sigma\xi^t$ gives the first identity below, while we assume the second: $\widehat{\boldsymbol{\Lambda}}_{\sigma,T} = \alpha\boldsymbol{\Lambda}_T + \sigma\boldsymbol{\eta}_{\Sigma,T},\ \langle\boldsymbol{\eta}_{\Sigma,T}, \boldsymbol{\Lambda}_T\rangle_{\mathcal{H}_\Sigma} = 0$. We require $\boldsymbol{\Lambda}_T, \boldsymbol{\eta}_{\Sigma,T} \in \mathcal{H}_\Sigma$ and $\boldsymbol{\Lambda}_T \neq 0$. The orthogonality is in population, not separately on each record.

This is a sufficient perturbation model, not a literal description of every neural estimation error: independence simplifies the analysis, while population orthogonality makes the SINE calculation explicit. The observed SINE-NITA ordering persists under temporally correlated and heteroskedastic nuisance (Appendix D.13), indicating that temporal independence and homoskedasticity are not strict requirements in these experiments. Appendix F.1 further discusses the implications.

Assumption 3.5 (Common evaluation box and CUSUM asymptotics)**.** Every NITA comparison uses one evaluation box shared by the models being compared. For the CUSUM rule in binary i.i.d., assume the standard first-order large-threshold expansions (Lai, 1998; Fuh, 2004; Tartakovsky, 2021):

$$\log \mathrm{ARL}(h;\sigma) = \kappa(\sigma)\, h + O(1), \qquad \mathrm{ADD}_1(h;\sigma) = \frac{h}{\mu_1(\sigma)} + O(1), \qquad h \to \infty. \tag{11}$$

Here $\mu_1(\sigma) := \mathbb{E}_1[g_\sigma], \mathbb{E}_0[e^{\kappa(\sigma)g_\sigma}] = 1, \mathcal{E}(\sigma) := \kappa(\sigma)\mu_1(\sigma)$, where $\mu_1(\sigma)$ is the post-change mean score increment and $\kappa(\sigma) > 0$ is the unique positive solution of the exponential-moment equation.

The two expansions say that increasing the threshold by one unit multiplies false-alarm protection at exponential rate $\kappa(\sigma)$, while detection delay grows at reciprocal drift $1/\mu_1(\sigma)$. Their product $\mathcal{E}(\sigma)$ is therefore the first-order slope of the log-ARL versus delay frontier that we call as the *effective slope*. Remark B.13 gives a change-of-measure interpretation of $\kappa(\sigma)$ and background on these asymptotics.

Let x denote a delay cost, and y_σ^* be the upper envelope. Eliminating h from Eq. (11) yields

$$\log y_\sigma^\star(x) = \mathcal{E}(\sigma)\, x + b_\sigma(x), \qquad b_\sigma(x) = O(1) \quad \text{as } x \to \infty, \tag{12}$$

where b_σ is the bounded remainder. We use the term *asymptotically antitone* to mean exactly this pairwise high-threshold ordering: once the operating point is taken sufficiently deep into the asymptotic regime, the better-slope frontier dominates, and the corresponding NITA computed on any such common high-threshold comparison interval is ordered the same way. The theorem below is central to our theory, connecting this eventual NITA ordering to the opposite SINE ordering.

Theorem 3.6 (SINE–NITA asymptotic antitone)**.** *Under Assumptions 3.4–3.5, the following hold.*

(1) ***Projective error is monotone in noise.*** *Define* $\mathrm{SINE}_\Sigma(\sigma) := \mathrm{SINE}(\widehat{\boldsymbol{\Lambda}}_{\sigma,T}, \boldsymbol{\Lambda}_T)$ *in* $\mathcal{H}_\Sigma$*. Then*

$$\mathrm{SINE}_\Sigma(\sigma)^2 = \frac{\sigma^2\|\boldsymbol{\eta}_{\Sigma,T}\|_{\mathcal{H}_\Sigma}^2}{\alpha^2\|\boldsymbol{\Lambda}_T\|_{\mathcal{H}_\Sigma}^2 + \sigma^2\|\boldsymbol{\eta}_{\Sigma,T}\|_{\mathcal{H}_\Sigma}^2}. \tag{13}$$

In particular, $\mathrm{SINE}_\Sigma(\sigma)$ *is nondecreasing in* σ*, and it is strictly increasing whenever* $\boldsymbol{\eta}_{\Sigma,T} \neq 0$*.*

(2) ***The effective slope decreases with noise.*** *The post-change drift is constant in* σ*,* $\mu_1(\sigma) = \alpha\,\mathbb{E}_1[\lambda]$*, whereas the root* $\kappa(\sigma)$ *is strictly decreasing in* σ*. Therefore* $\mathcal{E}(\sigma)$ *is strictly decreasing in* σ*.*

(3) ***NITA is asymptotically antitone to SINE on non-saturated high-threshold boxes.*** *For every pair* $0 \le \sigma_1 < \sigma_2 \le \bar{\sigma}$*, there exists* $x_0 < \infty$ *such that*

$$y_{\sigma_1}^\star(x) > y_{\sigma_2}^\star(x) \qquad \forall x \ge x_0. \tag{14}$$

Thus, for any $X^* > x_0$ *and common non-saturating* $Y^* > \max_{i\in\{1,2\}} y_{\sigma_i}^\star(x)$*,*

$$\mathrm{NITA}_{Y^*}^{[x_0,X^*]}(\sigma_1) > \mathrm{NITA}_{Y^*}^{[x_0,X^*]}(\sigma_2). \tag{15}$$

If $\boldsymbol{\eta}_{\Sigma,T} \neq 0$*, then on the same box*

$$\mathrm{SINE}_\Sigma(\sigma_1) < \mathrm{SINE}_\Sigma(\sigma_2) \iff \mathrm{NITA}_{Y^*}^{[x_0,X^*]}(\sigma_1) > \mathrm{NITA}_{Y^*}^{[x_0,X^*]}(\sigma_2). \tag{16}$$

Proof sketch. The first claim is geometric. By Assumption 3.4, the learned LLR decomposes into a signal on the P_+LLR ray and an orthogonal nuisance trajectory, thus SINE increases monotonically with nuisance magnitude. The second claim is sequential. Since $\mathbb{E}_Q[\xi] = 0$, the post-change drift is unchanged: $\mu_1(\sigma) = \mathbb{E}_1[g_\sigma] = \alpha\mathbb{E}_1[\lambda]$. Under Assumption 3.4, writing moment generating functions (MGFs) $M_\lambda(u) := \mathbb{E}_0[e^{u\lambda}]$ and $M_\xi(v) := \mathbb{E}_Q[e^{v\xi}]$, the pre-change moment factorizes as $\mathbb{E}_0[e^{\kappa g_\sigma}] = \mathbb{E}_0[e^{\alpha\kappa\lambda}]\,\mathbb{E}[e^{\sigma\kappa\xi}] = M_\lambda(\alpha\kappa)\,M_\xi(\sigma\kappa)$. Because ξ is centered and nondegenerate, $M_\xi(v) > 1$ for every $v > 0$, and the positive root $\kappa(\sigma)$ and $\mathcal{E}(\sigma)$ decrease with σ. For the third claim, Eq. (11) implies Eq. (12), thus for any $\sigma_1 < \sigma_2$, $\log y^\star_{\sigma_1}(x) - \log y^\star_{\sigma_2}(x) = \big(\mathcal{E}(\sigma_1) - \mathcal{E}(\sigma_2)\big)x + O(1)$. By part (2) of this theorem, $\mathcal{E}(\sigma_1) > \mathcal{E}(\sigma_2)$, the slope gap asymptotically dominates the remainder, yielding Eq. (14). Integrating over common high-threshold evaluation box then gives Eq. (15), and part (1) transfers the pairwise ordering to SINE when $\boldsymbol{\eta}_{\Sigma,T} \neq 0$. See Appendix C.1 for details.

Corollary 3.7 (Local quadratic loss of the log-ARL–delay slope)**.** *Under the assumptions of Theorem 3.6, assume $\|\boldsymbol{\eta}_{\Sigma,T}\|_{\mathcal{H}_\Sigma} > 0$ and that M_λ and M_ξ are twice continuously differentiable near 1 and 0, respectively. Hold T, Q_{val}, α, and the signal–nuisance law fixed. Then, as $\sigma \downarrow 0$,*

$$\mathbb{E}_1[\lambda] - \mathcal{E}(\sigma) = c_{\mathcal{E}}\,\mathrm{SINE}_\Sigma(\sigma)^2 + o(\mathrm{SINE}_\Sigma(\sigma)^2), \tag{17}$$

where $c_{\mathcal{E}} := \|\boldsymbol{\Lambda}_T\|^2_{\mathcal{H}_\Sigma}\,\mathrm{Var}_Q(\xi)/(2\|\boldsymbol{\eta}_{\Sigma,T}\|^2_{\mathcal{H}_\Sigma}) > 0$, and $\mathrm{Var}_Q(\xi)$ is the nuisance variance under Q.

Corollary 3.7 follows from Theorem B.18 and tells that halving a small SINE approximately quarters the gap from the noise-free slope ($\mathcal{E}(0) = \mathbb{E}_1[\lambda]$, or the true average LLR increment), quantifying convergence to the true-LLR asymptotic detection tradeoff. Appendix C.2 provide the proof, and Appendix B.5 distinguishes it from the idealized NITA calculation.

3.3 Practical checkpoint-selection: finite-sample reliability of empirical NITA

In practice, how accurately can a finite validation set estimate NITA for checkpoint selection? The following guarantee requires neither the affine-noise model nor high-threshold asymptotics.

For finite-validation selection, fix checkpoints $\mathcal{M} = \{1, \ldots, N_{\mathrm{ckpt}}\}$, a common normalized grid $\Theta = \{\theta_1, \ldots, \theta_{N_\theta}\}$, and box $[0, X^*] \times [0, Y^*]$, with $X^*, Y^* > 0$. Write NITA_m, $\widehat{\mathrm{NITA}}_m$ for population and empirical grid-NITA, and $m^\star, \widehat{m}$ for their maximizers. Operating points are estimated by sample means or Kaplan–Meier (KM) estimates, which account for censored sequences (Appendix A.3).

Theorem 3.8 (Finite-sample guarantee for grid-based NITA)**.** *Use the sample-mean or KM construction in Appendix A.3, with n_x, n_y independent validation units. In the KM case, event times (e.g., detection delay or false-alarm run length), and follow-up limits (i.e., the last time step that the event is observable) are independent, and follow-up reaches each time before the common restriction horizon H_ι with probability at least $\gamma_\iota > 0$, for coordinate $\iota \in \{x, y\}$. Set $B_x = X^* > 0, B_y = Y^* > 0$. For any $\delta_{\mathrm{conc}} \in (0, 1)$, with probability at least $1 - \delta_{\mathrm{conc}}$,*

$$\sup_{m \in \mathcal{M}} \left| \widehat{\mathrm{NITA}}_m - \mathrm{NITA}_m \right| \leq \varepsilon_{\mathrm{conc}}, \tag{18}$$

where $\varepsilon_{\mathrm{conc}} := \omega_x + \omega_y$ and

$$\omega_\iota := \begin{cases} \sqrt{\dfrac{\log(4N_{\mathrm{ckpt}}N_\theta/\delta_{\mathrm{conc}})}{2n_\iota}}, & \textit{sample mean}, \\ \dfrac{3H_\iota}{B_\iota\gamma_\iota}\sqrt{\dfrac{\log(8N_{\mathrm{ckpt}}N_\theta/\delta_{\mathrm{conc}})}{2n_\iota}}, & \textit{KM}. \end{cases} \tag{19}$$

On the same event,

$$\mathrm{NITA}_{m^\star} - \mathrm{NITA}_{\widehat{m}} \leq 2\varepsilon_{\mathrm{conc}}. \tag{20}$$

Therefore, the selected checkpoint is within $2\varepsilon_{\mathrm{conc}}$ of the best population grid-NITA. Proof and empirical check are provided in Appendices C.3 and D.14, respectively.

3.4 Beyond the i.i.d. binary QCD settings

Theorem 3.6 deliberately studies a simple setup for a first, clean formalization of the P_+LLR learning and SINE–NITA antitone. However, these are sufficient conditions, not strict requirements.

Table 1: Monotonic association with SINE, measured by Spearman's ρ.

Task	**QCD**							**SPRT**				
Dynamics	i.i.d.				non-i.i.d.			i.i.d.		non-i.i.d.		
Classes	2	2	2	3	2	2	3	2	3	2	3	3
Settings	Fixed	Rand.	Min5	Min5	AR(4)	Kit.	AR(4)	Gauss.	Gauss.	AR(4)	AR(4)	Meat
NITA↑	-.992**	-.990**	-.991**	-.994**	-.558**	-.542**	-.540**	-.716**	-.800**	-.854**	-.912**	-.702**
CE↓	.993**	.994**	.995**	.996**	.904**	.234**	.544*	.492**	.814**	.891**	-0.296	.616**
Acc↑	.997**	0.061	.865**	.911**	.327*	0.100	-.953**	.491*	0.325	-.735**	-.943**	-.751**

Reading the results. Colors show statistically significant correlations for which optimizing the metric either minimizes SINE (blue) or instead increases it (red). Note that smaller CE is better, whereas larger NITA and Acc are better. Black entries are not significant. The experimental results of dotted columns are visualized in Fig. 3. One and two asterisks denote $p < 0.05$ and $p < 0.01$. *Settings.* Fixed, Rand., and Min5 specify segment lengths, where a segment is a consecutive block with one class label: Fixed uses 10-step segments. Rand. allows random lengths. Min5 requires at least five steps per segment. AR(4): fourth-order autoregressive. Gauss: Gaussian SPRT. Gray columns are real data, Kit: Kitsune. Meat: UCR Meat. Full protocol are in Appendix D.

Proposition 3.9 (Unified transfer principle (informal))**.** *Consider a family of learned LLR scores ordered by a nuisance level. Suppose that greater nuisance strictly increases SINE and that the high-threshold or high-accuracy operating frontier is governed at leading order by a positive* effective slope*, with larger values indicating better performance. Suppose this slope strictly decreases with nuisance and the pairwise remainder differences are lower order than the leading term. Then, for each pair of scores, sufficiently far into the asymptotic regime, the lower-SINE score has the better operating frontier and therefore the larger NITA on a common, non-saturated comparison region.*

Appendix B.7 states the formal transfer principle and summarizes its instantiations for multiclass QCD, Markov QCD, binary and multiclass SPRT, and Markov SPRT in Table 2. Appendices B.8–B.11 derive the effective slopes and establish their decrease with nuisance under setting-specific assumptions. The proposition is the shared final step to prove SINE–NITA antitone, as in Theorem 3.6.

4 Experiments and Results

The purpose of our experiment is to evaluate NITA-based checkpoint selection for LLR estimators on practical streaming data. Reproducing classical optimality of the sequential tests on controlled data is not the main experimental scope, yet in Appendix F.5 we evaluate the selected checkpoints under classical first-alarm CUSUM. See Appendices D and E for protocols and statistical test details, respectively.

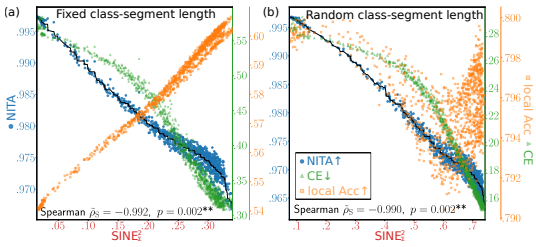


Figure 3: **Conventional metrics can be misaligned with the learning target of sequential tests.** In two validation datasets with slightly different changepoint statistics, Acc and CE can be unstable or misaligned, despite identical training data and setup. Black NITA curves are isotonic fits, with run-level Spearman ρ and p-values.

Does NITA track the projective target? (Table 1 and Fig. 3) We use 12 settings spanning synthetic i.i.d. Gaussian and non-i.i.d. autoregressive streams, binary and multiclass tasks, QCD and SPRT, and two real-world benchmarks: Kitsune network attack and UCR Meat (Mirsky et al., 2018; Dau et al., 2018). While synthetic data provide exact LLRs, on real data we use kernel density estimation (Botev et al., 2010)-based LLRs to evaluate SINE. NITA and all test-time decisions use estimated LLRs. The estimators share the same SPRT-TANDEM

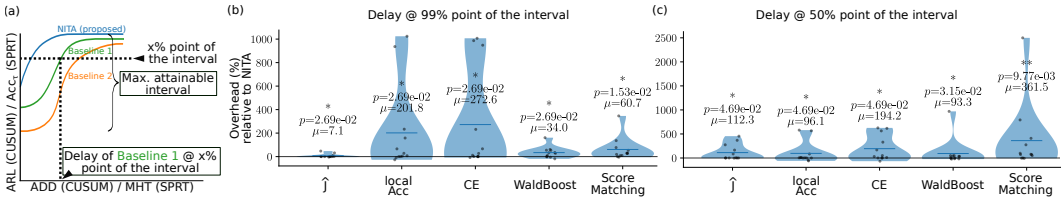


Figure 4: **Performance gain of NITA-selected checkpoints.** (a) On held-out test data, we measure time (abscissa) to reach $x\%$ of a *common* attainable operating range (ordinate), where the range is computed per dataset from pooled results of all baselines. Results are aggregated over the 12 datasets in Table 1. (b,c) Baseline overheads relative to NITA at the 99% (b) and 50% (c) points. Positive values indicate slower baselines. Violin plots show the density, mean (μ), and p-values showing significantly positive median overhead for all baselines. See Appendices D and E for details.

DRE backbone (Ebihara et al., 2021; 2023), so that NITA, CE, local Acc, and $\hat{J}$ (Bayesian-risk surrogate baseline (Cooper & Meyn, 2024)) differ only as checkpoint-selection rules.

With 10 runs per setting, Table 1 shows significant and consistently negative SINE–NITA associations in all 12 settings, whereas CE and local Acc have associations whose signs and significance vary across settings. Figure 3 illustrates the Fixed and Rand conditions in Table 1. While NITA preserves a strong correlation, local Acc and CE are unstable and consistently favoring worse SINE, respectively. The remaining checkpoint-level scatter plots appear in Appendix D.9 (Figs. 7 and 8).

Does NITA improve performance on held-out datasets? (Fig. 4) We sweep thresholds on unseen test data and measure the delay needed to reach the 99% and 50% points of each dataset's pooled attainable performance range (ARL for QCD, Acc_τ for SPRT. See Fig. 4a).

We compare NITA with CE-, Acc-, and $\hat{J}$-selected checkpoints on the same posterior-DRE backbone and with WaldBoost (Sochman & Matas, 2005) and Hyvärinen score matching (Hyvärinen, 2005). Details and an additional result for the baseline M-KLIEP (Givens et al., 2023) are in Appendix D.

Relative to NITA-selected checkpoints, all baselines show statistically significant positive delay overheads, even with low-threshold operating point (Fig. 4b,c). Thus NITA-based checkpoint selection improves performance at both stringent and moderate operating points, indicating robustness beyond the strict high-threshold regime of Theorem 3.6. See Appendix F.2 for detailed discussion.

Ablation tests. First, we repeat the experiments in Table 1 and Fig. 4 with the *full-box* NITA. The result is strikingly similar to that of high-threshold NITA (Fig. 13). The two NITA variants are almost perfectly rank-aligned with each other (Fig. 12), supporting the use of full-box NITA as the practical validation criterion beyond the theory, even without finding the common high-threshold region.

Second, we violate the independence and constant noise variance assumptions of Assumption 3.4 by using temporally correlated and heteroskedastic noise. The observed SINE-NITA ordering is still preserved under these noises (Appendix D.13). Together, these ablations support the observed SINE-NITA relationship beyond the high-threshold and independent noise conditions of Theorem 3.6.

Third, we ablate the normalization choices of SINE and NITA: removing the scale-normalizing denominator from SINE breaks the antitone relation almost completely (Fig. 10a), while checkpoint-dependent NITA normalization can become unstable at early training stages and only partially recovers later once attainable operating ranges stabilize (Fig. 10b), justifying the design choice.

Finally, we replace the TANDEMformer by the LSTM-based B2Bsqrt-LSTM to recover the same SINE-NITA antitone, showing that it is not an idiosyncrasy of one architecture (Appendix D.12.2).

5 Conclusion

For calibrated SPRT/CUSUM, the decision-relevant target is $\mathrm{P}_+\mathrm{LLR}$. SINE and NITA provably connect $\mathrm{P}_+\mathrm{LLR}$ to oracle geometry, asymptotic sequential performance, and finite-sample checkpoint selection, supported by experiments in broader settings even outside the binary i.i.d. QCD regime.

AI USE STATEMENT

Motivated by a practical problem, the authors developed the core research idea, formulated the theorem statements, and constructed the mathematical arguments and the proofs underlying the paper, independently of generative AI. We used large language models (LLMs) to refine wording and phrasing, prepare preliminary section drafts, search for related work, obtain feedback on readability, and discuss how to maintain mathematical integrity. We reviewed all the AI-assisted materials and kept LLM-suggested wordings where we judged them to accurately convey the intended content. The authors take full responsibility for the entire manuscript, including all AI-assisted contents.

REPRODUCIBILITY STATEMENT

Code is included in the supplementary material to ensure reproducibility. We repeat experiments with different random seeds, 10 runs in most settings, and report aggregated results and statistical-test p-values rather than relying on single runs. Appendix D documents data preparation, model configurations, and training and evaluation protocols, while Appendix E details the statistical analyses. Explicit assumptions and complete proofs are provided in the main text and Appendices B and C.

REFERENCES

Y. Achenchabe, A. Bondu, A. Cornuéjols, and A. Dachraoui. Early classification of time series: Cost-based optimization criterion and algorithms. *Mach. Learn.*, 110(6):1481–1504, June 2021. ISSN 0885-6125.

P. Armitage. Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12(1):137–144, 1950.

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26(4):641–647, 1955. ISSN 00034851.

T. Banerjee and V. Tarokh. Bayesian quickest change detection for unnormalized and score-based models. *Sequential Analysis*, 43:1–20, 08 2024.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. John Wiley & Sons, London, New York, 1972. ISBN 978-0471049708.

Z. Botev, J. Grotowski, and D. Kroese. Kernel density estimation via diffusion. *The Annals of Statistics*, 38, 11 2010.

H. P. Chan and T. L. Lai. Asymptotic approximations for error probabilities of sequential or fixed sample size tests in exponential families. *The Annals of Statistics*, 28(6):1638 – 1669, 2000.

A. Cooper and S. Meyn. Reinforcement learning for optimal stopping in pomdps with application to quickest change detection, 2025.

A. Cooper and S. P. Meyn. Quickest change detection using mismatched cusum extended abstract. *2024 60th Annual Allerton Conference on Communication, Control, and Computing*, pp. 1–7, 2024.

H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The UCR time series classification archive, October 2018.

V. P. Dragalin, A. G. Tartakovsky, and V. V. Veeravalli. Multihypothesis sequential probability ratio tests. i. asymptotic optimality. *IEEE Transactions on Information Theory*, 45(7):2448–2461, November 1999.

A. F. Ebihara, T. Miyagawa, K. Sakurai, and H. Imaoka. Sequential density ratio estimation for simultaneous optimization of speed and accuracy. In *The International Conference on Learning Representations (ICLR)*, 2021.

A. F. Ebihara, T. Miyagawa, K. Sakurai, and H. Imaoka. Toward asymptotic optimality: Sequential unsupervised regression of density ratio for early classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

R. S. Ellis. Large deviations for a general class of random vectors. *The Annals of Probability*, 12(1):1–12, 1984. ISSN 00911798, 2168894X.

C.-D. Fuh. SPRT and CUSUM in hidden Markov models. *The Annals of Statistics*, 31(3):942 – 977, 2003.

C.-D. Fuh. Asymptotic operating characteristics of an optimal change point detection in hidden markov models. *The Annals of Statistics*, 32(5):2305–2339, 2004. ISSN 00905364.

J. Gärtner. On large deviations from the invariant measure. *Theory of Probability & Its Applications*, 22(1): 24–39, 1977.

I. V. Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability & Its Applications*, 5(3):285–301, 1960.

J. Givens, S. Liu, and H. W. J. Reeve. Density ratio estimation and neyman–pearson classification with missing data. In F. Ruiz, J. Dy, and J.-W. van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8645–8681. PMLR, 2023.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

T. Gong, J. Lee, X. Cheng, and Y. Xie. Neural network-based cusum for online change-point detection, 2024.

A. Guntuboyina and B. Sen. Nonparametric shape-restricted regression. *Statistical Science*, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016a.

K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pp. 630–645, 2016b.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459, 1537274X.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 01 2005.

K. Ito. 109. Stochastic integral. *Proceedings of the Imperial Academy*, 20(8):519–524, 1944.

Y. Kawahara and M. Sugiyama. Sequential change-point detection based on direct density-ratio estimation. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(2):114–127, 2012.

I. Kontoyiannis and S. P. Meyn. Large deviations asymptotics and the spectral theory of multiplicatively regular markov processes. *Electronic Journal of Probability*, 10:61–123, 2005.

T. L. Lai and D. Siegmund. A Nonlinear Renewal Theory with Applications to Sequential Analysis I. *The Annals of Statistics*, 5(5):946 – 954, 1977.

T. L. Lai and D. Siegmund. A Nonlinear Renewal Theory with Applications to Sequential Analysis II. *The Annals of Statistics*, 7(1):60 – 76, 1979.

T. L. Lai. Sequential Changepoint Detection in Quality Control and Dynamical Systems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(4):613–644, November 1995. ISSN 1369-7412, 1467-9868.

T. L. Lai. Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Inf. Theory*, 44:2917–2929, 1998.

Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, June 2018.

G. Lorden. Procedures for Reacting to a Change in Distribution. *The Annals of Mathematical Statistics*, 42(6): 1897 – 1908, 1971.

G. Lorden. Nearly-optimal sequential tests for finitely many parameter values. *Annals of Statistics*, 5:1–21, 01 1977.

P. Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *Annals of Probability*, 18:1269–1283, 1990.

V. F. Melfi. Nonlinear Markov Renewal Theory with Statistical Applications. *The Annals of Probability*, 20(2): 753 – 771, 1992.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2 edition, 2009. ISBN 9780521731829.

Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai. Kitsune: An ensemble of autoencoders for online network intrusion detection. *CoRR*, abs/1802.09089, 2018.

T. Miyagawa and A. F. Ebihara. The power of log-sum-exp: Sequential density ratio matrix estimation for speed-accuracy optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 7792–7804, 2021.

T. Miyagawa and A. F. Ebihara. Accurate evaluation of quickest changepoint detectors via non-parametric survival analysis. In *Forty-third International Conference on Machine Learning*, 2026.

U. Mori, A. Mendiburu, S. Dasgupta, and J. A. Lozano. Early classification of time series by simultaneously optimizing the accuracy and earliness. *IEEE Transactions on Neural Networks and Learning Systems*, 29(10): 4569–4578, Oct 2018. ISSN 2162-237X.

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

G. V. Moustakides and K. Basioti. Training neural networks for likelihood/density ratio estimation. *ArXiv*, abs/1911.00405, 2019.

B. Øksendal. *Stochastic Differential Equations: An Introduction with Applications (Universitext)*. Springer, 6th edition, January 2014. ISBN 3540047581.

T. Oskiper and H. Poor. Online activity detection in a multiuser environment using the matrix cusum algorithm. *IEEE Transactions on Information Theory*, 48(2):477–493, 2002.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–115, 1954.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33 (3):1065–1076, 1962. ISSN 00034851.

M. Rosenblatt. Conditional probability density and regression estimators. In P. R. Krishnaiah (ed.), *Multivariate analysis, II*, pp. 25–31. Academic Press, New York, 1969. (Dayton, OH, 17–22 June 1968). MR:254987.

M. Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837, 1956.

J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. Sdr – half-baked or well done? *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2018.

D. Siegmund. *Sequential Analysis: Tests and Confidence Intervals*. Springer Series in Statistics. Springer, New York, NY, 1985. ISBN 978-0-387-96134-7.

J. Sochman and J. Matas. Waldboost - learning for time constrained sequential detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pp. 150–156 vol. 2, June 2005.

C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

A. Tartakovsky, I. Nikiforov, and M. Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman & Hall/CRC, 1st edition, 2014.

A. G. Tartakovsky. An asymptotic theory of joint sequential changepoint detection and identification for general stochastic models. *IEEE Transactions on Information Theory*, 67:4768–4783, 2021.

A. G. Tartakovsky and V. V. Veeravalli. General asymptotic bayesian theory of quickest change detection. *Theory of Probability and Its Applications*, 49:458–497, 2005.

A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 06 1945.

A. Wald and J. Wolfowitz. Optimum character of the sequential probability ratio test. *Ann. Math. Statist.*, 19(3): 326–339, 09 1948.

A. Wald. *Sequential Analysis*. John Wiley and Sons, 1st edition, 1947.

W. Wang, C. Chen, W. Wang, P. Rai, and L. Carin. Earliness-aware deep convolutional networks for early time series classification. *arXiv preprint arXiv:1611.04578*, 2016.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

M. Woodroofe. *Nonlinear Renewal Theory in Sequential Analysis*, volume 39 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982. ISBN 978-0-89871-180-6.

Z. Xing, J. Pei, and P. S. Yu. Early classification on time series. *Knowledge and Information Systems*, 31(1): 105–127, April 2012.

E. Zitzler, D. Brockhoff, and L. Thiele. The Hypervolume Indicator Revisited: On the Design of Pareto-compliant Indicators Via Weighted Integration. In S. Obayashi et al. (eds.), *Conference on Evolutionary Multi-Criterion Optimization (EMO 2007)*, volume 4403 of *LNCS*, pp. 862–876, Berlin, 2007. Springer.

APPENDIX

CONTENTS

1 Introduction **1**

2 Calibrated Sequential Tests with a Learned LLR **2**

3 Theory **4**

3.1 Projective estimand, oracle error, and observable surrogate 4

3.2 Main theorem: binary i.i.d. QCD 5

3.3 Practical checkpoint-selection: finite-sample reliability of empirical NITA 7

3.4 Beyond the i.i.d. binary QCD settings 7

4 Experiments and Results **8**

5 Conclusion **9**

A Mathematical Foundations **17**

A.1 Pairwise evidence: one-vs-baseline and all-rivals rules 17

A.2 Additional Conventions Shared with the Main Text 18

A.3 Finite-grid empirical NITA for checkpoint-selection 20

B Supplementary Theoretical Materials **22**

B.1 Decision-theoretic sufficiency of $\mathrm{P}_+\mathrm{LLR}$ 22

B.2 Tensor-valued positive-projective invariance for multiclass SPRT 23

B.3 Auxiliary geometric and asymptotic lemmas 24

B.4 Scope and failure modes of the SINE–NITA alignment 27

B.5 A local effective-slope theorem and an idealized NITA illustration 27

B.5.1 Numerical illustration of the idealized local area calculation 29

B.6 Auxiliary lemmas for the finite-sample guarantee 29

B.7 Formal transfer principle for the structured extensions 30

B.8 All-rivals matrix QCD: correct-declaration effective slopes 31

B.8.1 Decision rule and performance quantities 31

B.8.2 Learned-score assumptions and directional roots 33

B.8.3 Effective slope and SINE–NITA ordering 34

B.9 Markov-order-N QCD: effective slopes and SINE–NITA ordering 36

B.9.1 Conditional evidence and stationary score model 36

B.9.2 SCGF roots and effective slope 37

B.9.3 Increment-level SINE and NITA ordering 39

B.9.4 Initial-block correction and regularity conditions 40

B.10 Binary SPRT: effective slopes and SINE–NITA ordering 41

B.10.1 Learned-score model and directional roots 41
B.10.2 Equal-error calibration and effective slope 42
B.10.3 High-accuracy SINE–NITA ordering 43
B.10.4 Aggregate-accuracy and symmetric-threshold calibration 44
B.11 Multiclass and Markov SPRT: effective slopes and SINE–NITA ordering 45
B.11.1 Multiclass SPRT: the hardest-rival slope 45
B.11.2 Markov SPRT: directional SCGF slopes 48

C Proofs **51**
C.1 Proofs related to Section 3 51
C.2 Proofs for the local effective-slope (Corollary 3.7) 53
C.3 Proofs for the finite-sample guarantee (Theorem 3.8) 54
C.4 Proofs for the all-rivals matrix-QCD extension 57
C.5 Proofs for the Markov-order-N extension 61
C.6 Proofs for the SPRT extension 62
C.7 Proofs for the multiclass and Markov-order-N SPRT extensions 67

D Experimental Details **69**
D.1 Shared QCD evaluation protocol 69
D.1.1 KM operating coordinates and NITA 70
D.2 Synthetic data: i.i.d. Gaussian regimes 71
D.3 Synthetic data: vector autoregression (VAR) 71
D.4 Isotonic regression for visualization 72
D.5 Training-fitted KDE pseudo-LLR references 72
D.6 Additional baseline: KLIEP / M-KLIEP-style density-ratio estimation 73
D.7 Architecture and hyperparameter settings 73
D.8 Dataset details 74
D.9 Checkpoint-level scatter plots underlying Table 1 76
D.10 Additional real-world datasets 76
D.11 SPRT evaluation and NITA 77
D.11.1 Class weighting and sampling 78
D.12 Ablation details 78
D.12.1 Main high-performance-tail protocol and full-box ablation 79
D.12.2 Backbone ablation. 82
D.13 Stress tests of Assumption 3.4 82
D.13.1 Temporally correlated nuisance 82
D.13.2 Heteroskedastic and inverse-heteroskedastic nuisance 83
D.14 Empirical finite-sample stability of NITA 85

E Statistical tests **85**

E.1 Two-sided tests on run-level Spearman coefficients 85

E.2 One-sided median test of delay overheads 86

F Supplementary Discussion **87**

F.1 The affine-noise model: theoretical role and empirical evidence 87

F.2 High-threshold and full-box NITA: theoretical motivation and practical use 89

F.3 Why LSEL does not resolve the CE-validation issue? 90

F.4 Scope of the positive-projective learning target 91

F.5 Streaming checkpoint selection and transfer to first-alarm CUSUM 91

F.6 Novelty, impact, and the future of P_+LLR learning 92

A MATHEMATICAL FOUNDATIONS

We use the same definition for observation, filtration, and stopping-rule notation from Section 2 and specify the additional multiclass, representation, and finite-validation conventions needed below. For collections of records, $m \in [N_{\text{seq}}]$ indexes $X_m^{(1,T)} = (X_{m,1}, \ldots, X_{m,T})$, with $X_{m,t} \in \mathcal{X} \subseteq \mathbb{R}^d$. Background on the sequential rules is available in Page (1954); Tartakovsky et al. (2014); Wald & Wolfowitz (1948).

A.1 PAIRWISE EVIDENCE: ONE-VS-BASELINE AND ALL-RIVALS RULES

Pairwise evidence. Extend the binary notation of Section 2 to distinct classes k, ℓ in the class set $\mathcal{Y}$:

$$\lambda_{k\ell}^t := \log \frac{p_k(X_t \mid X^{(1,t-1)})}{p_\ell(X_t \mid X^{(1,t-1)})}, \qquad \Lambda_{k\ell}^{s:t} := \sum_{i=s}^{t} \lambda_{k\ell}^i, \quad 1 \le s \le t.$$

Positive evidence favors k over ℓ. We use $p_k(X_1 \mid X^{(1,0)}) := p_k(X_1)$ and $\Lambda_{k\ell}^{t+1:t} := 0$ for $t \ge 0$. In particular, $\Lambda_{k\ell}^{1:t} = \log\{p_k(X^{(1,t)})/p_\ell(X^{(1,t)})\}$. Segment sums keep the original conditioning history. Hats denote learned counterparts, and binary notation suppresses the pair $(k, \ell) = (1, 0)$.

One-vs-baseline: compare each alternative with the same reference. In the one-vs-baseline QCD model, $\mathcal{Y} = \{0\} \cup [K]$: class 0 is the designated pre-change regime, and $k \in [K]$ indexes possible post-change regimes. Each candidate is compared with class 0. This is the multichart CUSUM construction obtained by running Page's recursion for each post-change alternative against a common baseline (Page, 1954; Tartakovsky et al., 2014). The alarm asks whether *any* candidate has accumulated enough evidence against that baseline. It does not require the candidate to defeat the other post-change alternatives.

Let $\mathbb{P}_{\nu,k}$ and $\mathbb{E}_{\nu,k}$ denote the law and expectation when the conditional density switches from p_0 to p_k at time ν, and let $\mathbb{P}_\infty, \mathbb{E}_\infty$ denote the no-change law. For binary QCD, write $\mathbb{P}_\nu = \mathbb{P}_{\nu,1}$ and $\mathbb{E}_\nu = \mathbb{E}_{\nu,1}$. The class-$k$ reference chart is

$$W_0^{(k)} := 0, \qquad W_t^{(k)} = \max\{0, W_{t-1}^{(k)} + \lambda_{k0}^t\} = \max_{1 \le s \le t+1} \Lambda_{k0}^{s:t}.$$

For $h > 0$, its stopping time and terminal decision are

$$\tau_h := \inf\{t \ge 1 : \max_{k \in [K]} W_t^{(k)} \ge h\}, \qquad d_{\tau_h} \in \arg\max_{k \in [K]} W_{\tau_h}^{(k)},$$

with deterministic tie breaking on $\{\tau_h < \infty\}$. For $K = 1$, this is the binary rule in Section 2, with $W_t = W_t^{(1)}$.

All-rivals: a candidate must defeat every alternative. An all-rivals rule compares each candidate k with every $\ell \in \mathcal{Y} \setminus \{k\}$, rather than using one common baseline. For multiclass SPRT, $\mathcal{Y} = [K]$. For example, on $\mathcal{Y} = \{0, 1, 2\}$, a one-vs-baseline rule for class 1 compares it only with class 0. An all-rivals rule requires class 1 to defeat both class 0 and class 2.

For QCD, the recursive matrix-CUSUM construction of Oskiper & Poor (2002) uses a separate pairwise CUSUM chart $W_{k\ell,t}$ for each ordered pair, with $\lambda_{k\ell}^t$ in the recursion above. The *all-rivals reduction* combines these pairwise statistics into one score per candidate:

$$C_{k,t} := \min_{\ell \in \mathcal{Y} \setminus \{k\}} W_{k\ell,t}, \qquad C_{k,t} \ge h \iff W_{k\ell,t} \ge h \quad \text{for every } \ell \ne k.$$

The minimum is the candidate's weakest evidence, corresponding to its hardest rival. Strong evidence against one class cannot compensate for weak evidence against another. A subsequent maximum over candidates tests whether *at least one* class satisfies all its comparisons. Pairwise CUSUM reflection precedes the rival minimum, thus each pair has its own evidence history.

The all-rivals comparison on accumulated LLRs is the classical matrix-SPRT construction, tracing to Armitage (1950) and developed in Lorden (1977); Dragalin et al. (1999). It uses accumulated LLRs rather than reflected CUSUM scores: with a common threshold h, class k qualifies when

$\min_{\ell \neq k} \Lambda_{k\ell}^{1:t} \geq h$. The reduction specifies when a class qualifies, whereas the output and reset policies are separate choices. Appendix B.8 specifies the theoretical QCD declaration-and-restart procedure, while Appendix D.1 defines the continuous empirical confirmation protocol, including its initialization.

A.2 Additional Conventions Shared with the Main Text

The main theorem measures projective error in cumulative LLR trajectories, while the extensions also consider one-step increments and multiclass evidence. We introduce a common notation for applying SINE to these representations, then distinguish finite-record vectors from population geometry and specify the empirical operating-curve conventions.

Evidence representations for SINE. At a fixed horizon T, collect the chosen LLR values into a single vector, called an *evidence stack*. For example, a binary record gives the increment vector $(\lambda^1, \ldots, \lambda^T)$ or the cumulative trajectory $(\Lambda^{1:1}, \ldots, \Lambda^{1:T})$. Multiclass evidence additionally includes the class-pair indices. Let $\mathcal{Z}_T$ be the finite-dimensional real vector space of the chosen input stacks, and let $\boldsymbol{\ell}, \widehat{\boldsymbol{\ell}} \in \mathcal{Z}_T$ denote their true and learned values.

A fixed linear map specifies the representation used to compute SINE:

$$\mathsf{R} : \mathcal{Z}_T \to \mathcal{V}_{\mathsf{R}}, \qquad u_{\mathsf{R}} := \mathsf{R}\boldsymbol{\ell}, \qquad \hat{u}_{\mathsf{R}} := \mathsf{R}\widehat{\boldsymbol{\ell}}.$$

Thus, u_{R} and $\hat{u}_{\mathsf{R}}$ are the true and learned evidence in the chosen representation. For a fixed R, we abbreviate them as u and $\hat{u}$, which are not unit vectors. The same map is applied to both and held fixed across the comparison.

For an input stack of increments, R_I is the identity, whereas R_Σ takes prefix sums, giving $u_{\mathsf{R}_\Sigma} = \boldsymbol{\Lambda}_T$ in the binary setting of the main theorem. Conversely, R_Δ takes successive differences of a cumulative path with initial value zero. These choices let the same SINE definition compare either individual evidence increments or their accumulated trajectory.

The represented space $\mathcal{V}_{\mathsf{R}}$ uses the Euclidean inner product and norm, $\langle v, w \rangle_{\mathcal{V}_{\mathsf{R}}} = v^\top w$, $\|v\|_{\mathcal{V}_{\mathsf{R}}} = \|v\|_2$. For nonzero represented vectors, write $\mathrm{SINE}_{\mathsf{R}}(\widehat{\boldsymbol{\ell}}, \boldsymbol{\ell}) := \mathrm{SINE}(\hat{u}_{\mathsf{R}}, u_{\mathsf{R}})$. Projective correctness in this representation means $\hat{u}_{\mathsf{R}} \in [u_{\mathsf{R}}]_+$. Any component in $\ker(\mathsf{R})$ is mapped to zero and therefore does not affect this comparison. For a finite index set $\mathcal{I}$, $\mathrm{vec}((a_i)_{i \in \mathcal{I}})$ collects its entries in one fixed order, shared by the true and learned evidence. A common reordering leaves their Euclidean SINE unchanged.

Population SINE across validation records. The preceding convention describes evidence for one record. To measure projective error across records, let $V \sim Q_{\mathrm{val}}$, where Q_{val} is the validation distribution. The realized evidence $u_{\mathsf{R}}(V) = \mathsf{R}\boldsymbol{\ell}(V)$ is a vector in $\mathcal{V}_{\mathsf{R}}$, whereas u_{R} and $\hat{u}_{\mathsf{R}}$ are functions assigning true and learned represented evidence to each record. We compare these functions in

$$\mathcal{H}_{\mathsf{R}} := L^2(Q_{\mathrm{val}}; \mathcal{V}_{\mathsf{R}}), \qquad \langle f, g \rangle_{\mathcal{H}_{\mathsf{R}}} := \mathbb{E}_{V \sim Q_{\mathrm{val}}}[\langle f(V), g(V) \rangle_{\mathcal{V}_{\mathsf{R}}}].$$

This space contains measurable functions with finite expected squared norm, and $\|f\|^2_{\mathcal{H}_{\mathsf{R}}} = \mathbb{E}_{Q_{\mathrm{val}}} \|f(V)\|_2^2$. Two functions are identified in this space when they produce the same represented vector with probability one under Q_{val}.

Population SINE uses this inner product, with nonzero population norms for both the true and learned evidence. For each learned score, its projection uses one scalar $\beta \geq 0$ shared across all records and represented coordinates. Population orthogonality means that the expected record-wise inner product is zero. For binary cumulative trajectories, $\mathcal{H}_{\mathsf{R}_\Sigma} = \mathcal{H}_\Sigma$ from Eq. (10). We hold R, T, and Q_{val} fixed when comparing learned scores.

Empirical SINE replaces the expectation defining the inner product with its validation-sample average, then performs the projection and normalization. It is therefore not an average of separately computed record-wise SINE values (Lemma B.10).

QCD delay convention. Theoretical QCD delays use immediate-change laws: $\mathrm{ADD}_1(h) = \mathbb{E}_{\nu=1}[\tau_h - 1]$ in binary QCD and $\mathrm{ADD}_{k,1}(h) = \mathbb{E}_{\nu=1,k}[\tau_h - 1]$ for the one-vs-baseline reference

rule. We abbreviate $\mathbb{P}_k = \mathbb{P}_{\nu=1,k}$ and $\mathbb{E}_k = \mathbb{E}_{\nu=1,k}$. Appendix B.8 instead specifies $\mathrm{ADD}^+_{k,1} = \mathbb{E}_k[T_k^+] - 1$ for actually emitted correct declarations under its restart policy. For the main regime-switching benchmark, empirical $\mathrm{ADD}_{\mathcal{P}}$ is computed from KM survival for continuous correct-class confirmation lag, using the finite-stream convention in Appendix D.1.1. The separate binary first-alarm experiment in Appendix F.5 evaluates ADD_1 using the original checkpoint selections.

Posterior representation for multiclass sequential testing. When a time-invariant class label $Y \in [K]$ generates the entire sequence, define the posterior

$$\pi_k^{(t)} := \mathbb{P}(Y = k \mid X^{(1,t)}), \qquad k \in [K].$$

Under fixed class priors, posterior vectors and accumulated pairwise LLRs are equivalent:

$$\Lambda_{k\ell}^{1:t} = \log \frac{\pi_k^{(t)}}{\pi_\ell^{(t)}} - \log \frac{\mathbb{P}(Y = k)}{\mathbb{P}(Y = \ell)}, \qquad k \neq \ell. \tag{21}$$

Therefore, depending on the algorithm, one may work with either posteriors or accumulated pairwise LLRs as the sequential state. In this paper we use the LLR representation, since the $\mathrm{P}_+\mathrm{LLR}$ target is defined directly on it.

Finite-horizon convention for multiclass-SPRT. If a multiclass-SPRT procedure does not cross any decision boundary before a finite horizon T, a common convention is to force a terminal decision at T using the final class margin. With

$$m_k^{(t)} := \min_{\ell \in [K]\setminus\{k\}} \Lambda_{k\ell}^{1:t} - a_k^{(t)},$$

the forced terminal decision is

$$d_T \in \arg\max_{k\in[K]} m_k^{(T)} = \arg\max_{k\in[K]} \left(\min_{\ell \in [K]\setminus\{k\}} \Lambda_{k\ell}^{1:T} - a_k^{(T)} \right), \tag{22}$$

If several classes share the largest final margin, we select the first according to a fixed ordering of class labels (e.g., the smallest index). We refer to this as *tie-breaking rule*. This is a general margin-based convention. The implemented final-frame comparison sets all thresholds to zero, as described in Appendix D.11; numerical tie perturbations are documented separately there.

Truncated stopping times. While asymptotic results use untruncated stopping times, finite-record SPRT uses

$$\tau^T := \min\{\tau, T\}, \tag{23}$$

for mean hitting time, with a forced terminal decision. The main regime-switching QCD benchmark instead maintains an observed time and event flag, and sums Kaplan-Meier survival to common restriction horizons (see Appendices A.3 and D.1.1). Its empirical $\mathrm{ADD}_{\mathcal{P}}, \mathrm{ARL}_{\mathcal{P}}$ are KM-based operating coordinates under that finite-stream convention, not the uncapped expectations in Eq. (11).

Accuracy notation: local versus terminal. Unsubscripted Acc denotes the threshold-free local validation accuracy used as a checkpoint-selection baseline. Let $\mathcal{I}_{\mathrm{val}}$ be the set of valid local validation units, e.g. time steps or causal windows. For checkpoint m, let $\hat{y}_{m,i}^{\mathrm{loc}}$ be the local class/regime prediction for unit $i \in \mathcal{I}_{\mathrm{val}}$ and let y_i^{loc} be the corresponding ground-truth local target. We define

$$\widehat{\mathrm{Acc}}(m) := \frac{1}{|\mathcal{I}_{\mathrm{val}}|} \sum_{i \in \mathcal{I}_{\mathrm{val}}} \mathbf{1}\{\hat{y}_{m,i}^{\mathrm{loc}} = y_i^{\mathrm{loc}}\}.$$

This quantity is computed before running any SPRT/CUSUM thresholded sequential rule; it has no threshold parameter and no stopping time.

By contrast, terminal sequential accuracy is a property of the stopped SPRT decision rule. For threshold parameter θ, with stopping time τ_θ and terminal decision d_{τ_θ}, define

$$\mathrm{Acc}_\tau(\theta) := \mathbb{P}(d_{\tau_\theta} = Y).$$

In finite-horizon experiments this is evaluated using the truncated stopping time $\tau_\theta^T = \min\{\tau_\theta, T\}$ and the forced-decision convention described above. Thus local Acc is a checkpoint-selection baseline,

whereas Acc_τ is an operating-curve coordinate used in SPRT-NITA. Prior-weighted terminal variants are denoted Acc_π. The NITA implementation reports macro recall, namely Acc_π with $\pi_k = 1/K$: it averages classwise terminal recall rather than pooling all correct-record indicators. These coincide under balanced class counts. Empirical plot labels Acc_τ use this macro-recall convention; see Appendix D.11.1.

Multiple changepoints. The event and censoring definitions in Appendix D.1 use continuously kept evidence; labels select scoring intervals, not resets. These definitions apply throughout the main regime-switching benchmark; the separate binary first-alarm experiment is specified in Appendix F.5.

A.3 Finite-grid empirical NITA for checkpoint-selection

We define the operating coordinate estimators used in Theorem 3.8. The main QCD benchmark uses the same Kaplan–Meier (KM) computation to handle censored event times, with the finite-stream convention specified in Appendix D.1.1. We use a term *follow-up* to denote the period over which we observe whether the event of interest occurs. In QCD, the observation begins at the record's beginning for ARL, or at a changepoint for ADD. It ends when the relevant event occurs or valid observation ends at the next changepoint or record boundary. If no event has occurred by the last valid observation, the sequence is *right-censored*: we know only that it exceeds the available follow-up.

Training information and common threshold units. Let $\mathscr{T}$ contain the training data, training randomness, and training-only choices of checkpoints, positive score scales, threshold grids, validation horizons, and common comparison regions. These choices are frozen before validation. For checkpoint m, write g_m for its learned evidence process and $s_m > 0$ for its training-derived scale. Conditional on $\mathscr{T}$, the normalized checkpoint $\bar{g}_m := g_m/s_m$ is fixed. Applying a common normalized threshold θ_j to $\bar{g}_m$ is equivalent to applying $s_m\theta_j$ to g_m, with componentwise multiplication for threshold pairs or matrices (Proposition 2.1). The validation units are independent of $\mathscr{T}$. The scale-setting training records may have been used to fit g_m; independence between the fitted checkpoint and its scale is not required. Throughout this subsection and the finite-sample proof, probabilities and expectations are conditional on $\mathscr{T}$, with this conditioning is omitted in the notation for simplicity. Population coordinates refer to fresh validation units for the realized checkpoint–scale pairs.

Let $\mathcal{M} := \{1, \dots, N_{\mathrm{ckpt}}\}$ and $\Theta := \{\theta_1, \dots, \theta_{N_\theta}\}$. The common operating box is $[0, X^*] \times [0, Y^*]$, where $X^*, Y^* > 0$. Write $B_x := X^*$, $B_y := Y^*$, and let $\iota \in \{x, y\}$ select an operating coordinate. For each checkpoint–threshold pair (m, j), denote the clipped population coordinates by $x_{m,j}, y_{m,j}$ and their estimates by $\widehat{x}_{m,j}, \widehat{y}_{m,j}$. The estimator type for each coordinate is fixed across (m, j). There is no independence requirement across checkpoints, thresholds, or the two coordinates.

Sample-mean construction. When the measurements being averaged are fully observed, we use

$$\widehat{x}_{m,j} := \frac{1}{n_x}\sum_{r=1}^{n_x} \phi^x_{m,j}(U_r), \qquad \widehat{y}_{m,j} := \frac{1}{n_y}\sum_{s=1}^{n_y} \phi^y_{m,j}(V_s). \tag{24}$$

Here, $\phi^x_{m,j}$ and $\phi^y_{m,j}$ give the measurements for the two operating coordinates. For each fixed (m, j), the measurements are independent within each average and bounded in $[0, B_x]$ and $[0, B_y]$, respectively. The population coordinates are the averages of the corresponding measurement expectations, thus $\mathbb{E}[\widehat{x}_{m,j}] = x_{m,j}$ and $\mathbb{E}[\widehat{y}_{m,j}] = y_{m,j}$.

For SPRT, these measurements are the capped stopping time $\tau^T = \min\{\tau, T\}$ and terminal correctness, recorded as 1 for a correct prediction and 0 otherwise. Both are known even when no threshold is crossed before T, because the protocol forces a terminal decision at that horizon. Averaging the correctness indicators with equal record weights gives ordinary terminal accuracy. The reported macro recall instead averages classwise accuracies; its class weights and corresponding concentration bound are specified in Appendix D.11.1. It coincides with the equally weighted record average when class counts are balanced. No unbiasedness assumption is imposed on the KM estimator below.

KM construction: event times and incomplete follow-up. Following the QCD survival formulation of Miyagawa & Ebihara (2026), fix (m, j) and a coordinate ι. For one evaluation unit, let T

be its nonnegative integer *event time*, allowing $\mathsf{T} = \infty$, and let C be the last nonnegative integer time through which the event is observable. Here T and C are event and follow-up variables, not the record-length symbol T or the matrix margin $C_{m,k,t}$. From n_ι units we observe

$$\mathsf{U}_i := \min\{\mathsf{T}_i, \mathsf{C}_i\}, \qquad \mathsf{D}_i := \mathbf{1}\{\mathsf{T}_i \le \mathsf{C}_i\}, \qquad i \in [n_\iota].$$

A no-hit unit has $\mathsf{U}_i = \mathsf{C}_i$ and $\mathsf{D}_i = 0$: it is right-censored, not recorded as an event at C_i. An actual event observed at the last valid time has $\mathsf{D}_i = 1$. For each integer $t \ge 0$, define

$$N_\iota^{\mathrm{risk}}(t) := \sum_{i=1}^{n_\iota} \mathbf{1}\{\mathsf{U}_i \ge t\}, \qquad N_\iota^{\mathrm{event}}(t) := \sum_{i=1}^{n_\iota} \mathbf{1}\{\mathsf{U}_i = t, \mathsf{D}_i = 1\}.$$

The risk count is the number of records or intervals whose event has not occurred before t and remains observable at t; the event count is the number of those with an observed event exactly at t. Their ratio estimates the conditional event probability used in the KM survival update. With $0/0 := 0$ for an empty risk set, the KM estimate of $S^{\mathrm{event}}(t) := \Pr(\mathsf{T} > t)$ is

$$\widehat{S}^{\mathrm{KM}}(t) := \prod_{u=0}^{t} \left(1 - \frac{N_\iota^{\mathrm{event}}(u)}{N_\iota^{\mathrm{risk}}(u)}\right).$$

Events precede censoring at tied times. An empty risk set keeps the last survival value, which may be positive, so tha it does not force an event. For a common integer restriction horizon $H_\iota \ge 1$, define

$$\mu_{\iota,m,j}^{\mathrm{res}} := \mathbb{E}[\min\{\mathsf{T}, H_\iota\}] = \sum_{t=0}^{H_\iota - 1} S^{\mathrm{event}}(t), \qquad \widehat{\mu}_{\iota,m,j}^{\mathrm{KM}} := \sum_{t=0}^{H_\iota - 1} \widehat{S}^{\mathrm{KM}}(t).$$

The horizon H_ι restricts the event-time target through $\min\{\mathsf{T}, H_\iota\}$, whereas B_ι bounds the resulting operating coordinate. Specifically, clip each population or estimated mean v to $[0, B_\iota]$ by replacing it with $\min\{B_\iota, \max\{0, v\}\}$. This clipping is applied to the mean, not separately to each event time.

KM sampling and follow-up assumptions. The following conditions support the KM estimation bound in Theorem 3.8. For each coordinate $\iota \in \{x, y\}$ and checkpoint–threshold pair, assume that the event-time/follow-up pairs $(\mathsf{T}_i, \mathsf{C}_i)$ are independent and identically distributed across evaluation units, with $\mathsf{T} \perp \mathsf{C}$ within each pair. Fix common restriction horizons H_ι before validation and assume that there exists $\gamma_\iota > 0$ such that

$$\inf_{\substack{m \in \mathcal{M},\ j \in [N_\theta] \\ 0 \le t < H_\iota}} \Pr(\mathsf{C} \ge t) \ge \gamma_\iota.$$

Here H_ι is a protocol choice, whereas γ_ι is a population follow-up coverage bound used in the analysis, not a tuning parameter. Computing KM estimates and selecting checkpoints by empirical NITA do not require its value; a justified numerical lower bound is needed only to evaluate the theorem's error bound. For example, guaranteed follow-up through H_ι permits $\gamma_\iota = 1$.

The condition ensures positive follow-up coverage at every time used in the restricted mean. It concerns the probability of sufficient observation duration, not terminal accuracy or the observed-event fraction. H_ι is not necessarily equal to the operating cap B_ι. If the protocol includes only certain records or episodes, the assumptions and probabilities refer to that included population.

Follow-up support and flat-tail extension. Positive population coverage does not require a nonempty risk set at every time in every realized sample. In contrast, if the observation design gives $\Pr(\mathsf{C} \ge t) = 0$ at a time included in the restriction horizon, preserving a positive KM plateau there is a *flat-tail extension*. It does not identify event-time survival beyond the follow-up support, and Theorem 3.8 does not cover this extrapolated contribution. Appendix F.5 distinguishes controlled finite-validation reliability from practical streaming evaluation using this numerical convention.

Complete follow-up: connection to sample means. When the protocol provides follow-up through H_ι for every unit, the KM restricted mean equals the ordinary average of capped event times:

$$\widehat{\mu}_{\iota,m,j}^{\mathrm{KM}} = \frac{1}{n_\iota} \sum_{i=1}^{n_\iota} \min\{\mathsf{T}_i, H_\iota\}.$$

Thus the sample-mean bound applies when $H_\iota \le B_\iota$, even if a KM routine computes the estimate. The argument uses Hoeffding's inequality for averages of independent bounded measurements (Appendix C.3). If $H_\iota > B_\iota$, the corresponding normalized error bound is multiplied by H_ι/B_ι, since the summands are bounded by H_ι rather than B_ι; subsequent mean-level clipping cannot increase the error. With censoring before H_ι, use the KM bound in Eq. (19) under the preceding assumptions.

From operating coordinates to checkpoint selection. We now apply the same finite-grid NITA construction to the population coordinates and their validation estimates. For KM coordinates, the population values use $\mathbb{E}[\min\{\mathsf{T}, H_\iota\}]$ and the specified box clipping. The guarantee therefore handles this fixed-horizon evaluation, not unrestricted expected event times.

For the clipped population points, define

$$F_m(x) := \max\{y_{m,j} : x_{m,j} \le x,\ j \in [N_\theta]\}, \qquad x \ge 0, \tag{25}$$

$$\mathrm{NITA}_m := \frac{1}{X^* Y^*} \int_0^{X^*} F_m(x)\, dx. \tag{26}$$

Here $F_m(x)$ is the best performance attainable on the grid at cost no greater than x. The empirical quantities replace the population coordinates by their estimates:

$$\widehat{F}_m(x) := \max\{\widehat{y}_{m,j} : \widehat{x}_{m,j} \le x,\ j \in [N_\theta]\}, \qquad x \ge 0, \tag{27}$$

$$\widehat{\mathrm{NITA}}_m := \frac{1}{X^* Y^*} \int_0^{X^*} \widehat{F}_m(x)\, dx. \tag{28}$$

Both maxima use $\max \emptyset = 0$. The population-optimal candidate and empirical selector are

$$m^\star \in \arg\max_{m \in \mathcal{M}} \mathrm{NITA}_m, \qquad \widehat{m} \in \arg\max_{m \in \mathcal{M}} \widehat{\mathrm{NITA}}_m.$$

Exact finite-grid area. Sort the clipped empirical points so that $\widehat{x}_{m,(1)} \le \cdots \le \widehat{x}_{m,(N_\theta)}$, and define

$$\widetilde{y}_{m,(j)} := \max_{1 \le i \le j} \widehat{y}_{m,(i)}, \qquad j \in [N_\theta].$$

With $\widehat{x}_{m,(0)} = 0$, $\widetilde{y}_{m,(0)} = 0$, and $\widehat{x}_{m,(N_\theta+1)} = X^*$, the step envelope gives

$$\widehat{\mathrm{NITA}}_m = \frac{1}{X^* Y^*} \sum_{j=0}^{N_\theta} \widetilde{y}_{m,(j)} \big(\widehat{x}_{m,(j+1)} - \widehat{x}_{m,(j)}\big). \tag{29}$$

This is the exact integral of the finite-grid envelope, including the interval from the last operating point to X^*.

Scale invariance. A map is *positively homogeneous of degree one* if multiplying its input by any $\alpha > 0$ multiplies its output by the same factor. For a training-derived scale, this means that replacing g_m by αg_m on the same training reference records changes s_m to αs_m. The raw thresholds $s_m \theta_j$ then scale by α as well. For the stopping rules considered here, joint score–threshold rescaling preserves the event and censoring observations, and therefore the KM coordinates and grid-NITA. The restriction horizons and operating box remain fixed. Appendix C.1 gives the corresponding stopping-rule argument; numerical tie perturbations have the separate scope stated in Appendix D.11.

B Supplementary Theoretical Materials

This section shows the auxiliary results and structured extensions supporting the main theory. Their proofs are given in Appendix C.

B.1 Decision-theoretic sufficiency of $\mathrm{P}_+\mathrm{LLR}$

The following results extend the binary rescaling statement of Proposition 2.1 to the multiclass SPRT and CUSUM rules specified below.

Theorem B.1 (Decision-theoretic sufficiency of $\mathrm{P}_+\mathrm{LLR}$)**.** *Let $\{\delta_\theta\}_{\theta\in\Theta}$ be a threshold-calibrated LLR-based sequential test family, where θ denotes the threshold parameter(s). Assume that the rule is one of the following:*

1. *binary SPRT with boundary widths $a, b > 0$,*
2. *multiclass SPRT with threshold family $\{a_k^{(t)}\}_{t,k}$,*
3. *binary CUSUM with threshold h,*
4. *multiclass CUSUM with threshold h.*

Let $u = u_{\mathsf{R}}$ denote the corresponding represented accumulated-LLR path from Appendix A.2. The identity below holds for each realized path and therefore for the associated random path under any evaluation law. Then, for every $\alpha > 0$, there exists a rescaled threshold parameter $\mathcal{S}_\alpha(\theta)$ such that

$$\big(\tau_\theta(u),\, d_\theta(u)\big) = \big(\tau_{\mathcal{S}_\alpha(\theta)}(\alpha u),\, d_{\mathcal{S}_\alpha(\theta)}(\alpha u)\big) \qquad \textit{pathwise}. \tag{30}$$

More explicitly, $\mathcal{S}_\alpha(a,b) = (\alpha a, \alpha b)$ (binary SPRT), $\mathcal{S}_\alpha(a_k^{(t)}) = (\alpha a_k^{(t)})_{t,k}$ (multiclass SPRT), and $\mathcal{S}_\alpha(h) = \alpha h$ (binary/multiclass CUSUM).

Consequently, $[u]_+$ contains all represented information relevant to the stopping-time/decision family generated by $\{\delta_\theta\}_{\theta\in\Theta}$.

Proof. See Appendix C.1, proof of Theorem B.1.

Corollary B.2 (Operating-set invariance and calibration by threshold sweep)**.** *Let Φ_1, Φ_2 be performance functionals depending only on the law of (τ, d_τ) (e.g., $(\mathrm{MHT}, \mathrm{Acc}_\tau)$ for SPRT or $(\mathrm{ADD}_1, \mathrm{ARL})$ for CUSUM), and define*

$$\mathcal{C}(u) := \big\{(\Phi_1(\delta_\theta(u)),\, \Phi_2(\delta_\theta(u))) : \theta \in \Theta\big\}. \tag{31}$$

If Θ is closed under positive rescaling, then

$$\mathcal{C}(\alpha u) = \mathcal{C}(u) \qquad \forall \alpha > 0. \tag{32}$$

Thus unknown positive scale does not reduce attainable operating performance under threshold sweep.

Remark B.3 (Calibration versus sufficiency). Positive rescaling preserves attainable operating performance, but recovering the positive ray alone does not determine the numerical threshold for a prescribed error or false-alarm level. Thresholds must be expressed or calibrated in the recovered score units.

Remark B.4 (Sufficiency versus Fisher–Neyman sufficiency). Theorem B.1 is *decision-theoretic* sufficiency: $\mathrm{P}_+\mathrm{LLR}$ preserves threshold-calibrated SPRT/CUSUM decisions after threshold rescaling. It is distinct from Fisher–Neyman sufficiency.

Empirical all-rivals matrix rule. The pairwise reflections and hardest-rival reductions in Appendix D.1 are positively homogeneous. Scaling all scores and thresholds by one positive constant preserves the eligible class sets and the initial-prefix/post-change confirmation indicators. No unique emitted-class policy is required for these empirical events. This invariance does not substitute the streaming evaluation into the restarted asymptotic model. Numerical perturbations in the SPRT helper have the separate scope in Appendix D.11.

B.2 Tensor-valued positive-projective invariance for multiclass SPRT

The main text states the positive-projective invariance of threshold-calibrated sequential tests in a compact form. For multiclass sequential testing, it is sometimes convenient to write this invariance directly at the level of the accumulated pairwise LLR tensor. For multiclass SPRT, we write

$$\mathbf{\Lambda}_{\mathrm{MS}} := \big(\Lambda_{k\ell}^{1:t}\big)_{t\in[T],\, k,\ell\in[K],\, k\neq\ell}, \qquad u_{\mathrm{MS}} := \mathrm{vec}(\mathbf{\Lambda}_{\mathrm{MS}}), \tag{33}$$

For one record, $u_{\mathrm{MS}} \in \mathbb{R}^{TK(K-1)}$ is the represented vector of Appendix A.2, thus its $\mathrm{P}_+\mathrm{LLR}$ estimand is $[u_{\mathrm{MS}}]_+$. Its population version is the record-to-vector function used in Assumption B.49.

Proposition B.5 (Tensor-valued invariance of multiclass-SPRT under positive rescaling)**.** *Fix batch size $B \in \mathbb{N}$, horizon $T \in \mathbb{N}$, and number of classes $K \in \mathbb{N}$. Let*

$$\mathbf{\Lambda} \in \mathbb{R}^{B\times T\times K\times K}, \qquad \mathbf{\Lambda}_{b,t,k,\ell} = \Lambda^{1:t}_{k\ell}(X_b^{(1,t)}), \qquad b \in [B],\ t \in [T],\ k,\ell \in [K],\ k \neq \ell,$$

denote the accumulated pairwise LLR tensor. Let $\{a_k^{(t)}\}_{t\in[T],\,k\in[K]}$ be a family of time-dependent thresholds, and define the class margin

$$m_{b,k}^{(t)}(\mathbf{\Lambda}; a) := \min_{\ell\in[K]\setminus\{k\}} \mathbf{\Lambda}_{b,t,k,\ell} - a_k^{(t)}. \tag{34}$$

For each batch element b, define the stopping time

$$\tau_b(\mathbf{\Lambda}; a) := \inf\Big\{t \in [T] : \max_{k\in[K]} m_{b,k}^{(t)}(\mathbf{\Lambda}; a) \geq 0\Big\}, \tag{35}$$

and terminal decision

$$d_b(\mathbf{\Lambda}; a) \in \arg\max_{k\in[K]} m_{b,k}^{(\tau_b)}(\mathbf{\Lambda}; a), \tag{36}$$

with any deterministic tie-breaking rule. If no boundary is hit by time T, one may instead force the terminal decision

$$d_b^T(\mathbf{\Lambda}; a) \in \arg\max_{k\in[K]} m_{b,k}^{(T)}(\mathbf{\Lambda}; a).$$

Let $\alpha > 0$ and define the positively rescaled tensor and thresholds by

$$\widetilde{\mathbf{\Lambda}} := \alpha\,\mathbf{\Lambda}, \qquad \widetilde{a}_k^{(t)} := \alpha\, a_k^{(t)}.$$

Then, for every $b \in [B]$,

$$\tau_b(\widetilde{\mathbf{\Lambda}}; \widetilde{a}) = \tau_b(\mathbf{\Lambda}; a), \qquad d_b(\widetilde{\mathbf{\Lambda}}; \widetilde{a}) = d_b(\mathbf{\Lambda}; a), \tag{37}$$

and the same invariance holds for the forced finite-horizon decision d_b^T.

Proposition B.5 makes explicit that the multiclass-SPRT trajectory depends on the tensor $\mathbf{\Lambda}$ only through its positive projective class once the thresholds are rescaled accordingly.

Corollary B.6 (Threshold-swept operating curves are invariant on positive rays)**.** *Assume the setting of Proposition B.5. Let the threshold family be positively homogeneous in a scalar parameter $h > 0$, i.e.,*

$$a_k^{(t)}(h) = h\, a_{0,k}^{(t)} \qquad \text{for some fixed template } \{a_{0,k}^{(t)}\}_{t,k}. \tag{38}$$

Let $\delta_h(\mathbf{\Lambda})$ denote the multiclass-SPRT induced by $\mathbf{\Lambda}$ and thresholds $\{a_k^{(t)}(h)\}$, and let $\widehat{\delta}_h(\widehat{\mathbf{\Lambda}})$ denote the analogous rule driven by

$$\widehat{\mathbf{\Lambda}} = c\,\mathbf{\Lambda}, \qquad c > 0.$$

Then

$$\widehat{\delta}_{ch}(\widehat{\mathbf{\Lambda}}) = \delta_h(\mathbf{\Lambda}) \qquad \text{pathwise for every } h > 0. \tag{39}$$

Consequently, any operating characteristic that depends only on the joint law of the stopping time and terminal decision is invariant under unknown global positive rescaling after threshold sweeping.

Remark B.7 (Unknown scale and threshold calibration). Corollary B.6 implies that if a learned score recovers $c\,\mathbf{\Lambda}$ rather than $\mathbf{\Lambda}$ for some unknown $c > 0$, then post-hoc threshold calibration or sweeping still reaches the same attainable operating set. What is lost is only any *analytic* threshold formula that depends on the absolute LLR scale. This is precisely why the $\mathrm{P}_+\mathrm{LLR}$ viewpoint is natural for learned sequential testing.

B.3 Auxiliary geometric and asymptotic lemmas

Proposition B.8 (Exact minimizers of SINE)**.** *Fix $u = u_{\mathsf{R}} \in \mathcal{V}_{\mathsf{R}} \setminus \{0\}$ as in Appendix A.2. For $\hat{u} \in \mathcal{V}_{\mathsf{R}} \setminus \{0\}$, $\mathrm{SINE}(\hat{u}, u) \geq 0$, with the condition for equality to hold is $\hat{u} = \beta u$ for some $\beta > 0$. Equivalently,*

$$\arg\min_{\hat{u}\in\mathcal{V}_{\mathsf{R}}\setminus\{0\}} \mathrm{SINE}(\hat{u}, u) = [u]_+.$$

For raw stacks with $\mathsf{R}\boldsymbol{\ell} \neq 0$, the zero set is

$$\arg\min_{\substack{\widehat{\boldsymbol{\ell}} \in \mathcal{Z}_T \\ \mathsf{R}\widehat{\boldsymbol{\ell}} \neq 0}} \mathrm{SINE}_{\mathsf{R}}(\widehat{\boldsymbol{\ell}}, \boldsymbol{\ell}) = \mathsf{R}^{-1}([\mathsf{R}\boldsymbol{\ell}]_+).$$

Thus zero SINE means correctness in the represented space, and also raw projective correctness when R is injective on the evidence class. The same minimizer statement holds in $\mathcal{H}_{\mathsf{R}}$, with $\beta > 0$ a single constant and equality understood almost surely.

We next record auxiliary lemmas used in the proof of Theorem 3.6.

Lemma B.9 (Weighted-angle representation of represented SINE)**.** *Use the matrix of R in the finite-stack coordinates of Appendix A.2 and define the positive semidefinite matrix*

$$W := \mathsf{R}^\top \mathsf{R}.$$

For raw stacks $\widehat{\boldsymbol{\ell}}, \boldsymbol{\ell}$ with nonzero represented images, the cosine term in Definition 3.2 can be written as

$$\rho(\mathsf{R}\widehat{\boldsymbol{\ell}}, \mathsf{R}\boldsymbol{\ell}) = \max\left\{0,\ \frac{\widehat{\boldsymbol{\ell}}^\top W \boldsymbol{\ell}}{\sqrt{\widehat{\boldsymbol{\ell}}^\top W \widehat{\boldsymbol{\ell}}}\sqrt{\boldsymbol{\ell}^\top W \boldsymbol{\ell}}}\right\}. \tag{40}$$

For population SINE, the numerator and the two squared norms in Eq. (40) are each replaced by their expectations under Q_{val}, before forming the cosine. In particular, if C is the prefix-sum matrix, the finite-stack cosine underlying SINE_Σ uses the weighted inner product $v^\top C^\top C w$ on raw increments.

Lemma B.10 (Empirical finite-stack SINE is a plug-in estimate of population SINE)**.** *Let $V_1, V_2, \ldots$ be i.i.d. validation units with law Q_{val}, and let $f, g \in L^2(Q_{\mathrm{val}}; \mathcal{V}_{\mathsf{R}})$ with $\|f\|_{\mathcal{H}_{\mathsf{R}}} > 0$ and $\|g\|_{\mathcal{H}_{\mathsf{R}}} > 0$. Write $\mathrm{SINE}_{\mathcal{H}_{\mathsf{R}}}(f, g)$ for Definition 3.2 using the Hilbert-space inner product. Define the empirical inner product*

$$\langle f, g\rangle_n := \frac{1}{n}\sum_{i=1}^{n} \langle f(V_i), g(V_i)\rangle_{\mathcal{V}_{\mathsf{R}}},$$

Write $\|f\|_n^2 := \langle f, f\rangle_n$, and define $\mathrm{SINE}_n(f, g)$ by replacing the population inner products in Definition 3.2 with $\langle\cdot,\cdot\rangle_n$ whenever the empirical norms are nonzero. Then

$$\mathrm{SINE}_n(f, g) \xrightarrow{\text{a.s.}} \mathrm{SINE}_{\mathcal{H}_{\mathsf{R}}}(f, g).$$

For any finite collection of pairs (f_j, g_j), the convergence is simultaneous. The same conclusion holds for stationary ergodic validation units by replacing the i.i.d. law of large numbers with the ergodic theorem.

Lemma B.11 (Monotonicity of the capped-exponential box map)**.** *Fix $X^*, Y^* > 0$ and $A_0 \in (0, Y^*]$. Define*

$$\Phi_{A_0}(\mathcal{E}) := \frac{1}{X^* Y^*}\int_0^{X^*} \min\{A_0 e^{\mathcal{E}x}, Y^*\}\, dx, \qquad \mathcal{E} > 0. \tag{41}$$

Then $\Phi_{A_0}(\mathcal{E})$ is nondecreasing in $\mathcal{E}$. Moreover, it is strictly increasing whenever $A_0 < Y^*$; if $A_0 = Y^*$, then $\Phi_{A_0}(\mathcal{E}) \equiv 1$.

The next lemma makes explicit the factorized moment-generating-function argument behind the strict decrease of the Cramér root.

Lemma B.12 (Factorized moment-generating function and monotonicity of $\kappa(\sigma)$)**.** *Under Assumption 3.4, and write $\lambda := \log \frac{p_1(X)}{p_0(X)}$. Define*

$$M_\lambda(u) := \mathbb{E}_0[e^{u\lambda}], \qquad M_\xi(v) := \mathbb{E}_Q[e^{v\xi}].$$

Then the Cramér-root equation

$$\mathbb{E}_0[e^{\kappa(\sigma) g_\sigma}] = 1$$

factorizes as

$$M_\lambda(\alpha\kappa(\sigma))\, M_\xi(\sigma\kappa(\sigma)) = 1. \tag{42}$$

Moreover, if ξ is centered and nondegenerate, then $M_\xi(v) > 1$ for every $v > 0$, and the positive root $\kappa(\sigma)$ is strictly decreasing in σ. Consequently, the efficiency constant

$$\mathcal{E}(\sigma) := \kappa(\sigma)\mu_1(\sigma)$$

is strictly decreasing in σ because $\mu_1(\sigma)$ is constant.

Remark B.13 (Exponential tilting behind the false-alarm exponent). Fix σ, let $g_1, g_2, \dots$ be i.i.d. copies of g_σ, and define

$$S_n := \sum_{t=1}^{n} g_t, \qquad T_h := \inf\{n \geq 1 : S_n \geq h\}, \qquad \Lambda_{0,\sigma}(\theta) := \log \mathbb{E}_0[e^{\theta g_\sigma}].$$

For any θ with $\Lambda_{0,\sigma}(\theta) < \infty$, define the exponentially tilted law on $\mathcal{F}_n := \sigma(g_1, \dots, g_n)$ by

$$\frac{dP_0^{(\theta)}}{dP_0}\Big|_{\mathcal{F}_n} = \exp\big(\theta S_n - n\Lambda_{0,\sigma}(\theta)\big).$$

At the positive Cramér root $\kappa(\sigma)$ from Assumption 3.5, one has $\Lambda_{0,\sigma}(\kappa(\sigma)) = 0$, thus

$$\frac{dP_0^{(\kappa(\sigma))}}{dP_0}\Big|_{\mathcal{F}_n} = e^{\kappa(\sigma) S_n}.$$

Moreover,

$$\mathbb{E}_0^{(\kappa(\sigma))}[g_\sigma] = \Lambda'_{0,\sigma}(\kappa(\sigma)) > 0,$$

by convexity together with $\mathbb{E}_0[g_\sigma] < 0$, thus the upcrossing that is rare under P_0 becomes typical under the tilted law. Writing the overshoot as

$$\chi_h := S_{T_h} - h \geq 0,$$

a standard change-of-measure identity gives

$$P_0(T_h < \infty) = \mathbb{E}_0^{(\kappa(\sigma))}\left[e^{-\kappa(\sigma) S_{T_h}};\, T_h < \infty\right] = e^{-\kappa(\sigma)h}\mathbb{E}_0^{(\kappa(\sigma))}\left[e^{-\kappa(\sigma)\chi_h};\, T_h < \infty\right].$$

Standard overshoot / nonlinear-renewal theory then yields

$$P_0(T_h < \infty) = C_\sigma e^{-\kappa(\sigma)h}(1 + o(1))$$

for some $C_\sigma \in (0, \infty)$, which is the source of the false-alarm exponent in Eq. (11). For the reflected CUSUM process, the usual excursion decomposition then gives

$$\log \mathrm{ARL}(h; \sigma) = \kappa(\sigma)h + O(1).$$

The companion delay expansion

$$\mathrm{ADD}_1(h; \sigma) = \frac{h}{\mu_1(\sigma)} + O(1)$$

is the standard positive-drift Wald / overshoot estimate under P_1. In the positively scaled exact-LLR case $g = \alpha\lambda$, one has $\kappa = 1/\alpha$, and the tilted one-step law coincides with P_1. See, e.g., (Lai & Siegmund, 1977; Siegmund, 1985; Woodroofe, 1982; Lai, 1998; Tartakovsky et al., 2014).

Lemma B.14 (Relative nuisance parameterization)**.** *Use the population convention of Appendix A.2, with fixed $u = u_{\mathsf{R}} \in \mathcal{H}_{\mathsf{R}} \setminus \{0\}$ and $\eta_{\mathsf{R}} \in \mathcal{H}_{\mathsf{R}}$. Assume*

$$g_s = a(s)\lambda + b(s)\xi, \qquad \hat{u}_s = a(s)u + b(s)\eta_{\mathsf{R}},$$

where $a(s) > 0$, $b(s) \geq 0$, ξ satisfies the same centered independent-noise conditions as in Assumption 3.4, and $\langle \eta_{\mathsf{R}}, u \rangle_{\mathcal{H}_{\mathsf{R}}} = 0$. Let $\zeta(s) := b(s)/a(s)$. Then

$$\mathrm{SINE}(\hat{u}_s, u)^2 = \frac{\zeta(s)^2 \|\eta_{\mathsf{R}}\|^2_{\mathcal{H}_{\mathsf{R}}}}{\|u\|^2_{\mathcal{H}_{\mathsf{R}}} + \zeta(s)^2 \|\eta_{\mathsf{R}}\|^2_{\mathcal{H}_{\mathsf{R}}}}.$$

Moreover, the Cramér-root equation can be written as

$$M_\lambda(\tilde{\kappa}(\zeta)) M_\xi(\zeta\tilde{\kappa}(\zeta)) = 1,$$

and the first-order QCD slope is

$$\mathcal{E}(s) = \tilde{\kappa}(\zeta(s))\, \mathbb{E}_1[\lambda].$$

Thus SINE is nondecreasing and $\mathcal{E}$ is nonincreasing whenever $\zeta(s)$ is nondecreasing.

B.4 Scope and failure modes of the SINE–NITA alignment

Theorem 3.6 establishes pairwise ordering within its stated score family on sufficiently high, common non-saturated comparison regions. The following points explain the role of the score-family assumptions and common normalization. Broader empirical behavior is examined in Appendix D.13 and the full-box ablation in Appendix D.12.1.

Remark B.15 (Comparisons within a common score family.). The main theorem compares scores within a common affine independent noise. Separately decomposing arbitrary checkpoints into a signal component and an orthogonal residual does not establish common nuisance direction or distributional assumptions. Varying positive signal scales can be accommodated through the relative nuisance parameter $b(s)/a(s)$, as in Lemma B.14.

Section 4 evaluates actual learned checkpoints without imposing this common family construction, while Appendix D.13 reports preserved SINE-NITA associations under temporally correlated and heteroskedastic nuisance. These experiments assess empirical applicability beyond the sufficient conditions. Appendix F.1 discusses their implications for the scope of the framework.

Remark B.16 (Common normalizers are essential). If the NITA normalizers are allowed to depend on the model, order preservation can fail even for perfectly ordered frontiers. For example, fix $X^* > 0$ and consider the frontier family

$$y_{\mathcal{E}}(x) = e^{\mathcal{E}x}, \qquad \mathcal{E} > 0.$$

If one normalizes each model by its own maximum value

$$Y_{\max}(\mathcal{E}) := y_{\mathcal{E}}(X^*) = e^{\mathcal{E}X^*},$$

then the normalized area becomes

$$\frac{1}{X^* Y_{\max}(\mathcal{E})} \int_0^{X^*} y_{\mathcal{E}}(x)\, dx = \frac{1 - e^{-\mathcal{E}X^*}}{\mathcal{E}X^*}, \tag{43}$$

which is strictly *decreasing* in $\mathcal{E}$. Thus a better frontier can be assigned a smaller normalized area if the evaluation box is allowed to vary with the model. This is why NITA must be computed with common, protocol-fixed normalizers.

B.5 A local effective-slope theorem and an idealized NITA illustration

This subsection gives the supporting statement for Corollary 3.7, which quantifies the local effective-slope loss in terms of squared SINE. Under the fixed population geometry and local regularity conditions below, Theorem B.18 establishes that expansion and an additional ordered-limit relation for the actual uncapped CUSUM frontier. We then illustrate its consequence for the area under a separately defined capped-exponential curve.

In the binary i.i.d. QCD setting, let

$$g_\sigma = \alpha\,\lambda + \sigma\,\xi, \qquad \mathcal{E}(\sigma) := \kappa(\sigma)\mu_1(\sigma), \qquad \mu_1(\sigma) = \alpha I_{\mathrm{KL}}, \qquad I_{\mathrm{KL}} := \mathbb{E}_1[\lambda].$$

Recall that $\kappa(\sigma) > 0$ is the positive root of

$$\mathbb{E}_0[e^{\kappa(\sigma) g_\sigma}] = 1.$$

For the cumulative-trajectory geometry in $\mathcal{H}_\Sigma$, write

$$p := \|\mathbf{\Lambda}_T\|^2_{\mathcal{H}_\Sigma}, \qquad q := \|\boldsymbol{\eta}_{\Sigma,T}\|^2_{\mathcal{H}_\Sigma}. \tag{44}$$

The horizon T, validation law Q_{val}, and signal–nuisance law are held fixed as $\sigma \downarrow 0$, as specified in the assumption below.

Assumption B.17 (Local regularity for the effective slope)**.** In addition to the assumptions of Theorem 3.6, assume:

1. **Smooth MGFs.** The functions

$$M_\lambda(u) := \mathbb{E}_0[e^{u\lambda}], \qquad M_\xi(v) := \mathbb{E}_Q[e^{v\xi}]$$

are twice continuously differentiable in neighborhoods of $u = 1$ and $v = 0$, respectively.

2. **Fixed orthogonal cumulative-trajectory model.** Hold T, Q_{val}, α, and the joint signal–nuisance law fixed as $\sigma \downarrow 0$. In the same population Hilbert space $\mathcal{H}_\Sigma$ as Theorem 3.6,

$$\widehat{\mathbf{\Lambda}}_{\sigma,T} = \alpha \mathbf{\Lambda}_T + \sigma \boldsymbol{\eta}_{\Sigma,T},\ \langle \boldsymbol{\eta}_{\Sigma,T}, \mathbf{\Lambda}_T \rangle_{\mathcal{H}_\Sigma} = 0,\ p = \|\mathbf{\Lambda}_T\|^2_{\mathcal{H}_\Sigma} > 0,\ q = \|\boldsymbol{\eta}_{\Sigma,T}\|^2_{\mathcal{H}_\Sigma} > 0. \tag{45}$$

Theorem B.18 (Local effective-slope loss in terms of squared SINE)**.** *Assume the binary i.i.d. QCD setting and Assumption B.17. Let $s_\sigma := \mathrm{SINE}_\Sigma(\sigma)^2$. Then, as $\sigma \downarrow 0$,*

$$I_{\mathrm{KL}} - \mathcal{E}(\sigma) = c_{\mathcal{E}}\, s_\sigma + o(s_\sigma), \tag{46}$$

where

$$c_{\mathcal{E}} = \frac{p \operatorname{Var}_Q(\xi)}{2q} > 0. \tag{47}$$

In particular, for all sufficiently small positive σ,

$$\frac{c_{\mathcal{E}}}{2} s_\sigma \le I_{\mathrm{KL}} - \mathcal{E}(\sigma) \le \frac{3c_{\mathcal{E}}}{2} s_\sigma.$$

Let $y^\star_\sigma(d)$ denote the actual uncapped *CUSUM upper envelope in Eq. (12), with all positive thresholds available. Then*

$$\lim_{\sigma \downarrow 0} \left[\lim_{d \to \infty} \frac{\log y^\star_0(d) - \log y^\star_\sigma(d)}{d\, s_\sigma} \right] = c_{\mathcal{E}}.$$

The inner limit is taken with each positive σ fixed. No interchange of these limits, or fixed-box NITA expansion, is asserted.

Idealized area illustration. To illustrate how the local slope loss translates into an area loss, we apply Theorem B.18 to a capped-exponential curve. This curve is stipulated for the illustration, rather than identified with the actual CUSUM operating envelope. For protocol-fixed $D^* > 0$ and $0 < A_0 < F^*$, define

$$\widetilde{F}_\sigma(d) := \min\{A_0 e^{\mathcal{E}(\sigma) d}, F^*\}, \qquad d \in [0, D^*]. \tag{48}$$

Its box-normalized area is described by

$$\Phi(\mathcal{E}) := \frac{1}{D^* F^*} \int_0^{D^*} \min\{A_0 e^{\mathcal{E} x}, F^*\}\, dx, \qquad \mathcal{E} > 0, \tag{49}$$

and we write

$$\widetilde{\mathrm{NITA}}(\sigma) := \Phi(\mathcal{E}(\sigma)).$$

The strict inequality $A_0 < F^*$ excludes full saturation at the left endpoint and leaves a nonzero unsaturated interval.

Area derivative. The unsaturated length is

$$C := \log \frac{F^*}{A_0} > 0, \qquad m(\mathcal{E}) := \min\left\{ D^*, \frac{C}{\mathcal{E}} \right\}. \tag{50}$$

For every $\mathcal{E} > 0$,

$$\Phi'(\mathcal{E}) = \frac{A_0}{D^* F^*} \int_0^{m(\mathcal{E})} x e^{\mathcal{E} x}\, dx. \tag{51}$$

This formula covers an unreached cap, an interior cap, and a cap at the right endpoint. Its value at the noise-free slope is strictly positive:

Proposition B.19 (Positivity of the idealized area derivative)**.** *For $D^* > 0$, $0 < A_0 < F^*$, and $I_{\mathrm{KL}} > 0$, the map Φ in Eq. (49) is continuously differentiable at $\mathcal{E} = I_{\mathrm{KL}}$, and*

$$\Phi'(I_{\mathrm{KL}}) = \frac{A_0}{D^* F^*} \int_0^{m(I_{\mathrm{KL}})} x e^{I_{\mathrm{KL}} x}\, dx > 0, \tag{52}$$

where $m(\cdot)$ is defined in Eq. (50). Equivalently,

$$\Phi'(I_{\mathrm{KL}}) = \frac{A_0}{D^* F^* I^2_{\mathrm{KL}}} \left(I_{\mathrm{KL}} m(I_{\mathrm{KL}}) e^{I_{\mathrm{KL}} m(I_{\mathrm{KL}})} - e^{I_{\mathrm{KL}} m(I_{\mathrm{KL}})} + 1 \right) > 0. \tag{53}$$

Idealized local NITA calculation. Under the assumptions of Theorem B.18, the separately defined area of Eq. (48) satisfies

$$\widetilde{\mathrm{NITA}}(0) - \widetilde{\mathrm{NITA}}(\sigma) = c_N\, s_\sigma + o(s_\sigma), \tag{54}$$

where

$$c_N = \Phi'(I_{\mathrm{KL}})\, c_{\mathcal{E}} > 0. \tag{55}$$

Consequently, there exist $c_-, c_+ > 0$ and $\sigma_0 > 0$ such that

$$c_- s_\sigma \le \widetilde{\mathrm{NITA}}(0) - \widetilde{\mathrm{NITA}}(\sigma) \le c_+ s_\sigma, \qquad 0 < \sigma \le \sigma_0. \tag{56}$$

These area statements follow by the chain rule; their proof is separated from the effective-slope proof in Appendix C.2.

Relation to actual CUSUM NITA. Corollary 3.7 is about the effective slope of the actual CUSUM frontier. The area expansion Eq. (54), in contrast, is about the separately defined $\widetilde{\mathrm{NITA}}$. Obtaining the same coefficient for actual NITA on a fixed box would additionally require control of the noise-dependent difference between these areas, which the large-threshold assumptions do not provide.

B.5.1 Numerical illustration of the idealized local area calculation

We evaluate the capped-exponential curve Eq. (48) and compare its area gap with $c_N s_\sigma$, where $s_\sigma = \mathrm{SINE}_\Sigma(\sigma)^2$ and $c_N = \Phi'(I_{\mathrm{KL}})c_{\mathcal{E}}$. The three cap placements are defined relative to the noise-free curve $A_0 e^{I_{\mathrm{KL}} d}$:

$$\begin{aligned} &\text{interior cap:} && A_0 < F^* < A_0 e^{I_{\mathrm{KL}} D^*}, \\ &\text{endpoint cap:} && F^* = A_0 e^{I_{\mathrm{KL}} D^*}, \\ &\text{unreached cap:} && F^* > A_0 e^{I_{\mathrm{KL}} D^*}. \end{aligned}$$

The last case is labeled "cap above endpoint" in the figure. Each chosen cap is held fixed as σ varies.

Figure 5 shows that the analytic prediction tracks the area gap over several orders of squared SINE. The relative error tends to zero and the gap-to-s_σ ratio approaches c_N as σ decreases. At the largest displayed perturbations, deviations become visible, reaching approximately 10% in the interior-cap case. Thus the local coefficient remains informative over the plotted finite-noise range without implying exact proportionality there. The cap-ratio sweep also shows agreement across the endpoint boundary, with numerical sensitivity near $F^*/A_0 = 1$, where c_N vanishes.

B.6 Auxiliary lemmas for the finite-sample guarantee

The proof of Theorem 3.8 separates operating-coordinate estimation error from its effect on NITA. The first lemma bounds coordinate errors with high probability, using Hoeffding's inequality for sample means and a survival-function bound for KM. The second converts coordinate-error bounds into a NITA-error bound, regardless of how the points were estimated. All probability statements are conditional on the training information fixed in Appendix A.3. We use $B_x = X^*$ and $B_y = Y^*$ as in the main theorem.

Lemma B.20 (Concentration of sample-mean and KM operating points)**.** Sample means. *Under the sample-mean construction of Appendix A.3, for any $\varepsilon_x \in (0, X^*]$ and $\varepsilon_y \in (0, Y^*]$,*

$$\Pr\left(\max_{m\in\mathcal{M},\, j\in[N_\theta]} |\widehat{x}_{m,j} - x_{m,j}| > \varepsilon_x \right) \le 2N_{\mathrm{ckpt}} N_\theta \exp\left(-\frac{2n_x \varepsilon_x^2}{(X^*)^2} \right), \tag{57}$$

$$\Pr\left(\max_{m\in\mathcal{M},\, j\in[N_\theta]} |\widehat{y}_{m,j} - y_{m,j}| > \varepsilon_y \right) \le 2N_{\mathrm{ckpt}} N_\theta \exp\left(-\frac{2n_y \varepsilon_y^2}{(Y^*)^2} \right). \tag{58}$$

KM estimates. *Fix a coordinate $\iota \in \{x, y\}$ and a checkpoint–threshold pair satisfying the KM conditions in Appendix A.3. All survival functions and means below refer to this fixed choice. For any $\eta_{\mathrm{KM}} \in (0, 1)$, with probability at least $1 - \eta_{\mathrm{KM}}$,*

$$\sup_{0\le t<H_\iota} |\widehat{S}^{\mathrm{KM}}(t) - S^{\mathrm{event}}(t)| \le \frac{3}{\gamma_\iota}\sqrt{\frac{\log(4/\eta_{\mathrm{KM}})}{2n_\iota}},$$

$$|\widehat{\mu}^{\mathrm{KM}} - \mu^{\mathrm{res}}| \le \frac{3H_\iota}{\gamma_\iota}\sqrt{\frac{\log(4/\eta_{\mathrm{KM}})}{2n_\iota}}.$$

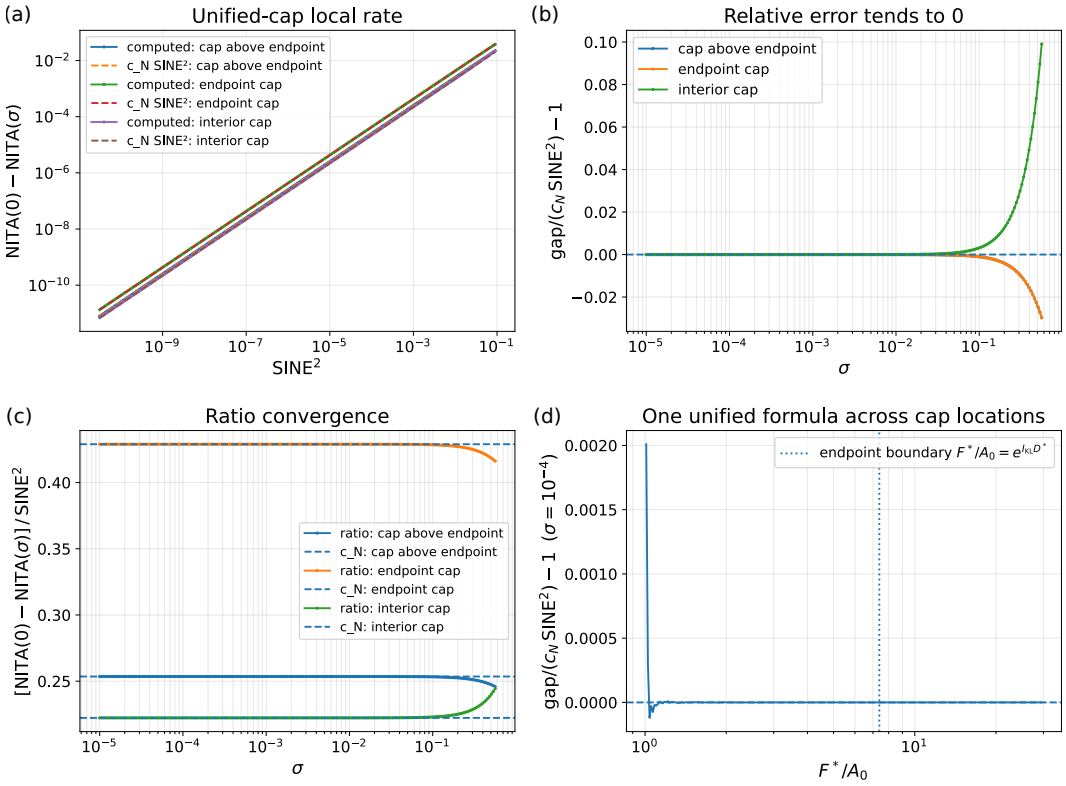


Figure 5: **Local area scaling for the idealized capped-exponential curve.** All areas refer to $\widetilde{\text{NITA}}$, not measured CUSUM NITA. (a) Area gap versus s_σ, compared with $c_N s_\sigma$ on log–log axes. (b) Relative error, or the NITA area gap divided by $c_N s_\sigma$, minus one. (c) Area gap divided by s_σ, compared with the analytic constants c_N. (d) Relative error versus F^*/A_0 at $\sigma = 10^{-4}$. The dotted line marks the noise-free endpoint-cap boundary $F^*/A_0 = e^{I_{\text{KL}} D^*}$.

The bounds may be capped at 1 *and* H_ι*, respectively. They allow tied event times, event time zero, and positive probability of no eventual event.*

Lemma B.21 (Deterministic stability of grid-NITA)**.** *Let* $\varepsilon_\iota \in [0, B_\iota]$ *for* $\iota \in \{x, y\}$*. Suppose that*

$$\max_{m\in\mathcal{M},\, j\in[N_\theta]} |\widehat{x}_{m,j} - x_{m,j}| \le \varepsilon_x, \qquad \max_{m\in\mathcal{M},\, j\in[N_\theta]} |\widehat{y}_{m,j} - y_{m,j}| \le \varepsilon_y.$$

Then

$$\sup_{m\in\mathcal{M}} |\widehat{\text{NITA}}_m - \text{NITA}_m| \le \sum_{\iota\in\{x,y\}} \frac{\varepsilon_\iota}{B_\iota}. \tag{59}$$

This implication holds for every collection of operating points satisfying the coordinate-error bounds; it requires no additional probability statement. Each coordinate's largest error contributes after division by its box bound. Combining this implication with Lemma B.20 gives the high-probability NITA bound in Theorem 3.8.

B.7 Formal transfer principle for the structured extensions

This subsection formalizes Proposition 3.9. The extensions have different effective slopes, but share the same final message: once the frontier expansion and slope ordering are established, smaller SINE gives larger NITA on a common comparison region. The subsequent subsections derive the setting-specific slopes and their ordering under setting-specific approximations..

Let Ξ label a sequential setting, including its data model, stopping rule, and operating coordinates, such as one of the settings in Table 2. The setting is fixed throughout each comparison, and σ varies the nuisance magnitude within its score family.

Choose the representation map R and validation law Q_{val} from Appendix A.2. Define the true and learned population evidence functions by

$$u_{\Xi}(V) = \mathsf{R}\boldsymbol{\ell}(V), \qquad \hat{u}_{\Xi,\sigma}(V) = \mathsf{R}\widehat{\boldsymbol{\ell}}_{\sigma}(V).$$

Their values on a record V lie in $\mathcal{V}_{\mathsf{R}}$, while the functions themselves lie in $\mathcal{H}_{\mathsf{R}}$. The nuisance η_{Ξ} is an element of the same population space. The horizon, representation map, validation law, true evidence, and nuisance are fixed as σ varies. All inner products and norms below are population quantities in $\mathcal{H}_{\mathsf{R}}$. For binary cumulative-trajectory SINE, this gives $u_{\Xi} = \mathbf{\Lambda}_T$ and $\mathcal{H}_{\mathsf{R}} = \mathcal{H}_{\Sigma}$.

Write $\mathrm{SINE}_{\Xi}(\sigma) := \mathrm{SINE}(\hat{u}_{\Xi,\sigma}, u_{\Xi})$. The notation $\mathfrak{N}_{\Xi}(\sigma; \mathcal{R})$ denotes the same NITA metric as before, now displaying the setting and the common operating-plane comparison region $\mathcal{R}$.

Proposition B.22 (Unified transfer principle)**.** *Assume*

$$\hat{u}_{\Xi,\sigma} = \alpha u_{\Xi} + \sigma\eta_{\Xi}, \qquad \alpha > 0,\ \sigma \geq 0, \qquad \langle \eta_{\Xi}, u_{\Xi} \rangle_{\mathcal{H}_{\mathsf{R}}} = 0, \qquad u_{\Xi} \neq 0. \tag{60}$$

Assume that $\mathcal{E}_{\Xi}(\sigma) > 0$ is strictly decreasing. Suppose the appropriate frontier expansion below holds, with the stated remainder condition for every fixed pair $\sigma_1 < \sigma_2$.

$$\textit{QCD:} \quad \log F^{\star}_{\sigma}(x) = \mathcal{E}_{\Xi}(\sigma)x + r_{\sigma}(x), \qquad r_{\sigma_1}(x) - r_{\sigma_2}(x) = o(x), \quad x \to \infty, \tag{61}$$

$$\textit{SPRT:} \quad T_{\sigma}(\gamma) = \frac{L_{\Xi}(\gamma)}{\mathcal{E}_{\Xi}(\sigma)} + r_{\sigma}(\gamma), \qquad r_{\sigma_1}(\gamma) - r_{\sigma_2}(\gamma) = o(L_{\Xi}(\gamma)), \quad \gamma \uparrow 1, \tag{62}$$

Here $F^{\star}_{\sigma}(x)$ gives the attainable false-alarm level at given delay budget x. γ is the target terminal accuracy, while $T_{\sigma}(\gamma)$ gives the mean decision time at γ and is monotone in that level. For each fixed setting Ξ, $L_{\Xi}(\gamma)$ is a function of γ and does not depend on σ. It grows without bound as γ approaches one. For example, binary equal-error SPRT uses $L_{\Xi}(\gamma) = \log\{\gamma/(1-\gamma)\}$.

$$\mathrm{SINE}_{\Xi}(\sigma)^2 = \frac{\sigma^2 \|\eta_{\Xi}\|^2_{\mathcal{H}_{\mathsf{R}}}}{\alpha^2 \|u_{\Xi}\|^2_{\mathcal{H}_{\mathsf{R}}} + \sigma^2 \|\eta_{\Xi}\|^2_{\mathcal{H}_{\mathsf{R}}}}, \tag{63}$$

thus SINE is nondecreasing, and strictly increasing if $\eta_{\Xi} \neq 0$. Moreover, for every $\sigma_1 < \sigma_2$ there is a common non-saturated asymptotic region $\mathcal{R}_{\sigma_1,\sigma_2}$ such that

$$\mathfrak{N}_{\Xi}(\sigma_1; \mathcal{R}_{\sigma_1,\sigma_2}) > \mathfrak{N}_{\Xi}(\sigma_2; \mathcal{R}_{\sigma_1,\sigma_2}). \tag{64}$$

If $\eta_{\Xi} \neq 0$, then on the same region

$$\mathrm{SINE}_{\Xi}(\sigma_1) < \mathrm{SINE}_{\Xi}(\sigma_2) \iff \mathfrak{N}_{\Xi}(\sigma_1; \mathcal{R}_{\sigma_1,\sigma_2}) > \mathfrak{N}_{\Xi}(\sigma_2; \mathcal{R}_{\sigma_1,\sigma_2}). \tag{65}$$

Proof is in Appendix C.1.

The same calculation applies to a fixed finite stack with $\mathcal{H}_{\mathsf{R}}$ replaced by $\mathcal{V}_{\mathsf{R}}$, provided its affine decomposition is orthogonal in that finite-stack inner product. This is a separate condition, not a consequence of population orthogonality; empirical SINE is related to population SINE by Lemma B.10.

B.8 All-rivals matrix QCD: correct-declaration effective slopes

We apply Proposition B.22 to the all-rivals matrix-CUSUM construction introduced in Appendix A.1 (Oskiper & Poor, 2002). A candidate must accumulate sufficient evidence against every rival. Our analysis concerns learned pairwise scores: we derive the false- and correct-declaration asymptotics, identify the effective slope, and show that its decrease with nuisance gives the SINE–NITA ordering.

B.8.1 Decision rule and performance quantities

We specify the decision cycles, the elapsed times used to evaluate them, and the represented evidence whose positive rescaling preserves these quantities.

All-rivals decision rule. Let $[K]$ be the set of $K \geq 2$ possible regimes or classes. For i.i.d. observations, let $\lambda_{k\ell,t} = \log(p_k(X_t)/p_\ell(X_t))$ and write $g^{\sigma}_{k\ell,t}$ for the learned, or estimated increments. A *decision cycle* starts from the zero state and ends when one class is declared. Within a cycle, define

$$W^{\sigma}_{k\ell,0} = 0, \qquad W^{\sigma}_{k\ell,t} = \max\{0, W^{\sigma}_{k\ell,t-1} + g^{\sigma}_{k\ell,t}\},$$
$$C^{\sigma}_{k,t} = \min_{\ell \neq k} W^{\sigma}_{k\ell,t}, \qquad \tau^{\sigma}_{h} = \inf\{t \geq 1 : \max_{k} C^{\sigma}_{k,t} \geq h\},$$

Table 2: Instantiation of the effective slopes for Proposition B.22.

Setting	Evidence path	NITA $\mathfrak{N}_\Xi$ axes	Effective slope $\mathcal{E}_\Xi$	Derivation
Binary QCD	$(\Lambda^{1:t})_{t\in[T]}$	$(\mathrm{ADD}_1, \mathrm{ARL})$	$\kappa(\sigma)\mu_1(\sigma)$	Thm. 3.6/ App. C.1
All-rivals matrix QCD	$\mathrm{vec}((\Lambda^{1:t}_{k\ell})_{t,(k,\ell)\in\mathcal{I}_K^{\neq}})$	$(D^+_\psi, \mathrm{ARL}_\Phi)$	$\kappa_\Phi(\sigma)/c_\psi$	App. B.8/ C.4
Markov QCD	$(\Lambda^{N+1:t}_{\mathrm{cond}})_t$	$(\mathrm{ADD}_1, \mathrm{ARL})$	$I_{\mathrm{rate}}\vartheta_\star(\zeta)$	App. B.9/ C.5
Binary SPRT	$(\Lambda^{1:t})_{t\in[T]}$	(T_π, γ)	$\mathrm{HM}_\pi\big(\kappa_0(\sigma)\mu_1, \kappa_1(\sigma)\mu_0\big)$	App. B.10/ C.6
Multiclass SPRT	$\mathrm{vec}((\Lambda^{1:t}_{k\ell})_{t,(k,\ell)\in\mathcal{I}_K^{\neq}})$	(T_π, γ_π)	$\Big(\sum_{k=1}^{K}\pi_k/\underline{\mathcal{E}}_k(\sigma)\Big)^{-1}$	App. B.11/ C.7
Markov SPRT	$(\Lambda^{1:t})_{t\in[T]}$	(T_π, γ)	$\mathrm{HM}_\pi\big(I_{10}\vartheta_0^\star(\zeta), I_{01}\vartheta_1^\star(\zeta)\big)$	App. B.11/ C.7

Notation. Entries show evidence path on one record and the operating coordinates (axes) used for NITA. Here $\mathcal{I}_K^{\neq} := \{(k,\ell) \in [K]^2 : k \neq \ell\}$. The Markov rows show detector paths, whereas SINE uses conditional increments and their stationary population geometry from Proposition B.35; for Markov QCD, $\mathsf{R}_\Delta(\Lambda^{N+1:t}_{\mathrm{cond}})_t = (\lambda^t)_{t=N+1}^T$. D_ψ is a homogeneous delay aggregate with slope c_ψ; $\mathrm{HM}_\pi(a,b) := (\pi/a + (1-\pi)/b)^{-1}$; T_π is prior π-weighted MHT, γ_π is prior-weighted multiclass accuracy, and γ is binary equal-error accuracy; $\kappa, \kappa_0, \kappa_1, \kappa_{\ell\to k}$ denote Cramér roots. In matrix QCD, $\kappa_b := \min_{k\neq b}\kappa_{b\to k}$ and κ_Φ is the exponent of the stated fixed-regime, worst-regime, or arithmetic-mean ARL aggregate; D^+_ψ uses correct-declaration delays with hardest-rival drift $d_k = \alpha\min_{\ell\neq k} I_{k\ell}$. The table's matrix-QCD law includes the joint accessibility and restart-after-declaration conditions of Appendix B.8; $\mu_1, \mu_0, \mu_{k\ell}(\sigma)$ are sign-corrected drifts; for multiclass SPRT, $\mathcal{E}_{k\ell}(\sigma) := \kappa_{\ell\to k}(\sigma)\mu_{k\ell}(\sigma)$, and $\underline{\mathcal{E}}_k(\sigma) := \min_{\ell\neq k}\mathcal{E}_{k\ell}(\sigma)$ is the class-k hardest-rival slope; $\vartheta_\star, \vartheta_0^\star, \vartheta_1^\star$ are scaled cumulant generating function (SCGF) roots replacing the i.i.d. Cramér roots at relative nuisance scale $\zeta = \sigma/\alpha$; $I_{\mathrm{rate}}, I_{10}, I_{01}$ are information rates.

and emit $J_h^\sigma \in \arg\max_k C^\sigma_{k,\tau_h^\sigma}$, using a fixed deterministic tie rule, independent of true labels. For simplicity, we assume that all rival comparisons use the same threshold $h > 0$. After *every* each declaration, reset all evidence and any score state to the initial state before starting the next cycle. The restart action is identical for correct and incorrect declarations. Under a constant regime and the i.i.d. score assumptions below, the successive cycle pairs are independent.

False and correct declarations. For a stream in regime b, write

$$q_b(h;\sigma) := \mathbb{P}_b(J_h^\sigma \neq b), \qquad m_b(h;\sigma) := \mathbb{E}_b[\tau_h^\sigma].$$

Let $(\tau^\sigma_{h,r}, J^\sigma_{h,r})_{r\geq 1}$ be the successive cycles under that fixed regime. Define

$$R_b(h;\sigma) := \sum_{r=1}^{N_b^-}\tau^\sigma_{h,r}, \quad N_b^- := \inf\{r \geq 1 : J^\sigma_{h,r} \neq b\},$$

$$T_b^+(h;\sigma) := \sum_{r=1}^{N_b^+}\tau^\sigma_{h,r}, \quad N_b^+ := \inf\{r \geq 1 : J^\sigma_{h,r} = b\}.$$

Thus $\mathrm{ARL}_b := \mathbb{E}_b[R_b]$ measures elapsed time to the first false declaration, including the durations of preceding correct declarations. The correct-declaration delay is $\mathrm{ADD}^+_{k,1} := \mathbb{E}_k[T_k^+ - 1]$. Importantly, a wrong declaration is never a successful detection. Its elapsed time is recorded, and contributes to T_k^+ only when correct class is declared later. The labels b, k specify the regimes under which performance is evaluated. Note that the decision and restart rules do not consult them.

Represented evidence and positive rescaling. The represented evidence is the full accumulated pairwise tensor:

$$\mathbf{\Lambda}_{\mathrm{mc}} := (\Lambda_{k\ell}^{1:t})_{t\in[T],\,(k,\ell)\in\mathcal{I}_K^{\neq}} \in \mathbb{R}^{T\times K(K-1)}, \qquad u_{\mathrm{mc}} := \mathrm{vec}(\mathbf{\Lambda}_{\mathrm{mc}}). \tag{66}$$

Here u_{mc} is a realized vector, whose population counterpart assigns this vector to each record as in Appendix A.2. Taking differences between consecutive times recovers the increments for each class pair. For the learned cumulative scores, these increments drive the pairwise CUSUM recursions defined above. Each pair is updated separately, and the minimum over rivals is taken after these updates.

As in Theorem B.1, positive rescaling gives $W_{k\ell,t}(cg) = cW_{k\ell,t}(g)$ for $c > 0$. With threshold ch, every cycle length and emitted class including fixed ties is unchanged. The next cycle starts at the same observation in both procedures, therefore induction over declarations also preserves R_b and T_k^+. Thus the relevant target is the positive ray of the full represented tensor.

B.8.2 Learned-score assumptions and directional roots

The mean pairwise increments describe typical evidence accumulation, while false declarations depend on rare upward fluctuations. We use moment-generating functions (MGFs) to analyze these fluctuations. For a real random variable Z, its MGF is the function $v \mapsto \mathbb{E}[e^{vZ}]$, wherever this expectation is finite. Exponential moments allow the proof to bound the probability that accumulated misleading evidence crosses a large threshold. The positive solutions of the MGF equations below determine the pairwise exponential rates used in this analysis. The following assumption specifies the score family and population SINE geometry. The root lemma and joint accessibility condition then provide the pairwise and all-rivals ingredients of the false-declaration analysis.

Assumption B.23 (i.i.d. affine all-pair scores and moments)**.** The K observation laws are distinct and mutually absolutely continuous, with $0 < I_{k\ell} < \infty$ for all $k \neq \ell$. Under each fixed regime, the full score vectors are i.i.d. across time and

$$g_{k\ell,t}^{\sigma} = \alpha\lambda_{k\ell,t} + \sigma\varepsilon_{k\ell,t}, \qquad \alpha > 0, \quad 0 \le \sigma \le \bar{\sigma} < \infty, \tag{67}$$

with the signal and nuisance vector laws fixed as σ varies. The nuisance is antisymmetric, $\varepsilon_{\ell k} = -\varepsilon_{k\ell}$, thus $g_{\ell k}^{\sigma} = -g_{k\ell}^{\sigma}$. For every ordered pair,

$$\mathbb{E}_k[\varepsilon_{k\ell}] = \mathbb{E}_\ell[\varepsilon_{k\ell}] = 0,$$

Under P_ℓ, the nuisance $\varepsilon_{k\ell}$ is nondegenerate and independent of $\lambda_{k\ell}$. Its MGF $v \mapsto \mathbb{E}_\ell[e^{v\varepsilon_{k\ell}}]$ is finite on an open neighborhood of $[-\bar{\sigma}/\alpha, \bar{\sigma}/\alpha]$.

For each fixed σ and ordered pair (k, ℓ), assume that the MGF of the learned increment satisfies

$$\mathbb{E}_\ell[e^{\kappa g_{k\ell}^{\sigma}}] = 1$$

at a unique positive value $\kappa = \kappa_{\ell\to k}(\sigma)$, called the *directional Cramér root*. This excludes the trivial solution $\kappa = 0$. The direction $\ell \to k$ indicates that regime ℓ is true while the score favors class k.

Under each regime P_b, the joint MGF

$$(v_{ij})_{i\neq j} \longmapsto \mathbb{E}_b\left[\exp\left\{\sum_{i\neq j} v_{ij} g_{ij}^{\sigma}\right\}\right]$$

is finite on an open neighborhood of the zero vector. When $b = \ell$, it is also finite on an open neighborhood of each vector $\kappa_{\ell\to k}(\sigma)e_{k\ell}$, where $e_{k\ell}$ has value one in coordinate (k, ℓ) and zero elsewhere. This vector corresponds to reweighting the original score distribution by $e^{\kappa_{\ell\to k}(\sigma) g_{k\ell}^{\sigma}}$. The root equation makes this weight have expectation one. The neighborhood conditions provide the exponential tail bounds and finite expected score magnitudes under this reweighting used in the proof.

Dependence between pairwise entries is allowed, including the dependence that arises when all pairs are computed from a shared set of class scores.

For a fixed finite horizon and fixed validation law, let $\mathcal{H}_{\mathrm{mc}} := L^2(Q_{\mathrm{val}}; \mathbb{R}^{TK(K-1)})$ and assume

$$\hat{u}_{\mathrm{mc},\sigma} = \alpha u_{\mathrm{mc}} + \sigma\eta_{\mathrm{mc}}, \qquad \langle \eta_{\mathrm{mc}}, u_{\mathrm{mc}}\rangle_{\mathcal{H}_{\mathrm{mc}}} = 0, \qquad 0 < \|u_{\mathrm{mc}}\|_{\mathcal{H}_{\mathrm{mc}}} < \infty, \tag{68}$$

with $\eta_{\mathrm{mc}} \in \mathcal{H}_{\mathrm{mc}}$ fixed across the family. This is the full-matrix counterpart of the geometry in the main theorem.

For later use, write

$$\kappa_b(\sigma) := \min_{j \neq b} \kappa_{b\to j}(\sigma), \qquad d_k := \min_{\ell \neq k} \mu_{k\ell}, \qquad \mu_{k\ell} := \alpha I_{k\ell}, \qquad I_{k\ell} := \mathrm{KL}(P_k \| P_\ell) > 0.$$

Directional roots. For each ordered pair, increasing the centered independent nuisance moves the positive MGF solution toward zero, by the binary argument in Lemma B.12. The next lemma records this ordering and shows that it also holds for the finite minima and maximum used to aggregate false-alarm performance.

Lemma B.24 (Directional roots and their extrema are monotone)**.** *Under Assumption B.23, each $\kappa_{b\to k}(\sigma)$, $b \neq k$, is strictly decreasing on $[0, \bar{\sigma}]$. Therefore each $\kappa_b(\sigma)$, their maximum, and*

$$\kappa_{\min}(\sigma) := \min_b \kappa_b(\sigma) = \min_{b \neq k} \kappa_{b\to k}(\sigma) \tag{69}$$

are strictly decreasing. The finiteness of the index sets is essential to this elementary strict-order argument; no common active pair is required.

From pairwise crossings to all-rivals false declarations. To analyze false declarations, fix the true regime b and work under P_b. Declaring any $k \neq b$ is then an error. Such a declaration requires $W^\sigma_{k\ell,t} \geq h$ for every $\ell \neq k$ at the declaration time. In particular, the comparison against the true regime must satisfy $W^\sigma_{kb,t} \geq h$, but this alone does not ensure that the remaining comparisons pass.

The directional root describes this pairwise crossing. To connect it to an all-rivals false declaration, the next assumption requires a joint condition for at least one candidate with the smallest directional root. Under the exponential reweighting defined below, every comparison in that candidate's row must have a mean increment at least as large as its positive mean increment against b. This condition concerns the joint score distribution and does not require independence between pairwise entries.

Assumption B.25 (Joint tilted-row accessibility)**.** Use the affine score family and directional roots of Assumption B.23. For each constant regime b and each $\sigma \in [0, \bar{\sigma}]$, at least one

$$k_\star \in \arg\min_{k \neq b} \kappa_{b\to k}(\sigma)$$

satisfies the following condition. Tilt the *whole one-step score-vector law* under $\mathbb{P}_b$ by

$$\frac{d\mathbb{Q}_{b,k_\star,\sigma}}{d\mathbb{P}_b} = \exp\{\kappa_{b\to k_\star}(\sigma) g^\sigma_{k_\star b}\}, \qquad v^{(b,\sigma)}_{k_\star \ell} := \mathbb{E}_{\mathbb{Q}_{b,k_\star,\sigma}} g^\sigma_{k_\star \ell}.$$

The root equation normalizes this density. Require

$$v^{(b,\sigma)}_{k_\star \ell} \geq v^{(b,\sigma)}_{k_\star b} > 0 \qquad \text{for every } \ell \neq k_\star.$$

Thus, under a least-cost pairwise false-declaration tilt, every rival in the candidate row accumulates evidence at least as quickly as the tilted b comparison. This is a joint restriction, not a consequence of the marginal root equations. Ties are allowed. It is required for at least one root-minimizing candidate for each b, σ, not for every false candidate. For envelope statements, use a large-threshold interval on which the population operating quantities are finite and locally bounded, and use common non-saturating comparison boxes.

Without this joint condition, the proof still gives $\liminf_{h\to\infty}[-\log q_b(h;\sigma)]/h \geq \kappa_b(\sigma)$, but not the matching equality needed to identify the effective slope. The Gaussian example below verifies the condition explicitly.

B.8.3 Effective slope and SINE–NITA ordering

We compare false-alarm protection with a specified aggregate of the classwise correct-declaration delays, as an evaluation of tradeoff. The following elementary step converts their first-order coefficients into the delay coefficient used by the transfer principle.

Lemma B.26 (Aggregation of correct-declaration delay)**.** *Let $\psi : \mathbb{R}^K_+ \to \mathbb{R}_+$ be coordinatewise nondecreasing, positively homogeneous of degree one, and locally Lipschitz. Suppose that, for fixed $d_k > 0$,*

$$\mathrm{ADD}^+_{k,1}(h) = \frac{h}{d_k} + o(h), \qquad k \in [K]. \tag{70}$$

Then

$$D^+_\psi(h) := \psi\big((\mathrm{ADD}^+_{k,1}(h))_{k=1}^K\big) = c_\psi h + o(h), \qquad c_\psi := \psi(d_1^{-1}, \dots, d_K^{-1}). \tag{71}$$

We require $c_\psi > 0$ when using delay as a frontier coordinate. Fixed positive weighted means and the maximum are examples.

Choice of false-alarm aggregation. Choose the false-alarm coordinate separately from the delay aggregate, and keep that choice fixed across compared scores. The theorem below gives the exponent in the right column:

$\mathrm{ARL}_\Phi(h;\sigma)$	$\kappa_\Phi(\sigma)$
$\mathrm{ARL}_{b_0}(h;\sigma)$	$\kappa_{b_0}(\sigma)$
$\min_b \mathrm{ARL}_b(h;\sigma)$	$\min_b \kappa_b(\sigma)$
$\sum_b \omega_b \mathrm{ARL}_b(h;\sigma)$	$\max_b \kappa_b(\sigma)$

In the last row, $\omega_b > 0$ and $\sum_b \omega_b = 1$ are fixed. The arithmetic mean is governed by the largest exponent, whereas the worst-regime ARL is governed by the smallest. These are the specified population functionals, not interchangeable summaries of pooled records.

Theorem B.27 (Effective slope and SINE–NITA ordering for all-rivals matrix QCD)**.** *Consider the declared-cycle and reset procedure above under Assumptions B.23 and B.25. For every fixed noise level, as $h \to \infty$:*

1. ***Full-matrix projective error.***

$$\mathrm{SINE}_{\mathrm{mc}}(\sigma)^2 = \frac{\sigma^2 \|\eta_{\mathrm{mc}}\|^2_{\mathcal{H}_{\mathrm{mc}}}}{\alpha^2 \|u_{\mathrm{mc}}\|^2_{\mathcal{H}_{\mathrm{mc}}} + \sigma^2 \|\eta_{\mathrm{mc}}\|^2_{\mathcal{H}_{\mathrm{mc}}}}. \tag{72}$$

   This is strictly increasing when $\eta_{\mathrm{mc}} \neq 0$.

2. ***Correct-declaration delay and false-declaration exponent.*** *With $d_k = \alpha \min_{\ell \neq k} I_{k\ell}$,*

$$m_k(h;\sigma) = h/d_k + o(h), \qquad -\log q_b(h;\sigma) = \kappa_b(\sigma)h + o(h).$$

   The regenerative identities $\mathrm{ARL}_b = m_b/q_b$ and $\mathrm{ADD}^+_{k,1} + 1 = m_k/(1-q_k)$ hold and yield

$$\log \mathrm{ARL}_b(h;\sigma) = \kappa_b(\sigma)h + o(h), \qquad \mathrm{ADD}^+_{k,1}(h;\sigma) = \frac{h}{d_k} + o(h), \tag{73}$$

   Incorrect declarations contribute elapsed time but do not count as correct detections.

3. ***Effective slope.*** *For D^+_ψ from Lemma B.26,*

$$c_\psi = \psi\left(\frac{1}{\alpha \min_{\ell \neq 1} I_{1\ell}}, \dots, \frac{1}{\alpha \min_{\ell \neq K} I_{K\ell}}\right) > 0 \tag{74}$$

   is independent of σ. For the chosen ARL aggregate,

$$\mathcal{E}_{\psi,\Phi}(\sigma) := \frac{\kappa_\Phi(\sigma)}{c_\psi} \tag{75}$$

   is strictly decreasing. The uncapped large-threshold upper envelope obeys

$$\log F^\star_{\psi,\Phi}(d;\sigma) = \mathcal{E}_{\psi,\Phi}(\sigma) d + o(d).$$

4. ***Common non-saturated tail ordering.*** *For each $\sigma_1 < \sigma_2$, there is $d_0 < \infty$ such that*

$$F^\star_{\psi,\Phi}(d;\sigma_1) > F^\star_{\psi,\Phi}(d;\sigma_2), \qquad d \geq d_0. \tag{76}$$

   For $D^* > d_0$, choose one finite cap satisfying

$$Y^* > \sup_{d \in [d_0, D^*]} \max_{i \in \{1,2\}} F^\star_{\psi,\Phi}(d;\sigma_i).$$

   Then

$$\mathrm{NITA}^{[d_0,D^*]}_{\psi,\Phi,Y^*}(\sigma_1) > \mathrm{NITA}^{[d_0,D^*]}_{\psi,\Phi,Y^*}(\sigma_2). \tag{77}$$

If $\eta_{\mathrm{mc}} \neq 0$, the same comparison region gives

$$\mathrm{SINE}_{\mathrm{mc}}(\sigma_1) < \mathrm{SINE}_{\mathrm{mc}}(\sigma_2) \iff \mathrm{NITA}^{[d_0,D^*]}_{\psi,\Phi,Y^*}(\sigma_1) > \mathrm{NITA}^{[d_0,D^*]}_{\psi,\Phi,Y^*}(\sigma_2). \tag{78}$$

No fixed-horizon or fully saturated area ordering is asserted.

A verified Gaussian instantiation. The joint accessibility condition and the effective slope can both be computed explicitly in the following family. Take $P_k = \mathcal{N}(m_0 e_k, I_d)$, $d \geq K$, $m_0 > 0$, thus $\lambda_{k\ell} = m_0(X_k - X_\ell)$, as in the signal construction of Appendix D.2. Add independent class-score noise $Z_1, \ldots, Z_K \overset{\text{iid}}{\sim} \mathcal{N}(0, v)$, $v > 0$, independent of X, and set $\varepsilon_{k\ell} = Z_k - Z_\ell$. Write $a = m_0^2$ and $V_\sigma = \alpha^2 a + \sigma^2 v$. For every $b \neq k$,

$$\kappa_{b\to k}(\sigma) = \frac{\alpha a}{V_\sigma}, \qquad d_k = \alpha a, \qquad v_{k\ell}^{(b,\sigma)} = \alpha a \quad (\ell \neq k).$$

The equality of the tilted row drifts verifies Assumption B.25, including at $\sigma = 0$. For a weighted-mean delay aggregate with weights summing to one, each of the listed ARL aggregates has

$$\mathcal{E}_{\psi,\Phi}(\sigma) = \frac{\alpha^2 a^2}{\alpha^2 a + \sigma^2 v}, \qquad \mathcal{E}_{\psi,\Phi}(0) = a.$$

Independent centered class-score noise also gives population orthogonality on fixed cumulative stacks under an independently chosen regime/record law. This construction verifies the assumptions for an explicit score family. See Appendix C.4 for details.

Remark B.28 (Declaration policy and empirical evaluation). The theorem concerns the specified declaration-and-restart rule on a fixed-regime i.i.d. stream. These assumptions make successive cycles independent. Resetting a statistic alone does not ensure independence under temporal dependence. The delay is measured from a fresh initial state, rather than over worst-case inherited histories.

QCD benchmark uses the same pairwise-reflection and all-rivals evidence construction with continuous confirmation scoring. Its event definitions and the separate binary first-alarm transfer experiment are in Appendices D.1 and F.5. The constant-regime renewal identities above are specific to the stated restart policy, even though the pathwise scaling identity also persists through regime switches.

The proof appendix also records the alternative all-rivals margins covered by the same argument and a first-cycle delay loss with a growing horizon (Appendix C.4).

B.9 Markov-order-N QCD: effective slopes and SINE–NITA ordering

We extend the binary QCD analysis to temporally dependent observations. The i.i.d. moment-generating function is replaced by the scaled cumulant generating function (SCGF), which describes exponential-moment growth per time step for accumulated evidence. The resulting effective slope is $I_{\text{rate}}\vartheta_\star(\zeta)$; the results below connect its noise ordering to SINE and NITA.

B.9.1 Conditional evidence and stationary score model

For an order-N model, the conditional increment is $\lambda^t = \log\{p_1(X_t \mid X_{t-N:t-1})/p_0(X_t \mid X_{t-N:t-1})\}$. The augmented state S_t records the observation history needed for these transitions. We use $\hat{\lambda}^t = \alpha\lambda^t + \sigma\varepsilon_t$, with fixed $\alpha > 0$ and relative nuisance magnitude $\zeta = \sigma/\alpha$. CUSUM uses the accumulated conditional evidence

$$\Lambda_{\text{cond}}^{N+1:t} := \sum_{s=N+1}^{t} \lambda^s, \qquad \widehat{\Lambda}_{\text{cond}}^{N+1:t} := \sum_{s=N+1}^{t} \hat{\lambda}^s, \qquad t \geq N+1.$$

SINE is evaluated on the one-step conditional increments, obtained by applying R_Δ to the same accumulated path. This distinguishes the detector input from the geometry used in Proposition B.35. For the stationary asymptotic sums below, the conditional increments are indexed from one after the initial block.

Assumption B.29 (Stationarity, moments, and strict SCGF-noise regularity)**.** Fix a finite relative-noise comparison range $[0, \bar{\zeta}]$. The following hold. Items 4–7 use the standard scaled cumulant generating function (SCGF) formalism for additive functionals of Markov chains and Markov-additive processes (Gärtner, 1977; Ellis, 1984; Kontoyiannis & Meyn, 2005; Meyn & Tweedie, 2009).

1. Under each regime $y \in \{0, 1\}$, the augmented chain (S_t) admits a stationary law π_y and is ergodic when started from π_y.

2. The noise process (ε_t) is stationary, nondegenerate, independent of (S_t), and satisfies

$$\mathbb{E}[\varepsilon_1] = 0, \qquad \mathbb{E}[\varepsilon_1^2] = 1.$$

3. For some fixed stationary reference law Q for the geometry calculation,
$$\mathbb{E}_Q[(\lambda^1)^2] < \infty.$$

4. The scaled cumulant generating functions
$$\Gamma(\vartheta) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_0 \exp\left(\vartheta \sum_{t=1}^{n} \lambda^t\right), \tag{79}$$
$$\Psi(v) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp\left(v \sum_{t=1}^{n} \varepsilon_t\right) \tag{80}$$
exist and are finite on neighborhoods of the relevant arguments, and are convex and continuously differentiable there.

5. For each $\zeta \in [0, \bar{\zeta}]$, the equation
$$\Gamma(\vartheta) + \Psi(\zeta\vartheta) = 0 \tag{81}$$
has a unique positive root $\vartheta_\star(\zeta)$.

6. The post-change information rate
$$I_{\mathrm{rate}} := \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_1 \left[\sum_{t=1}^{n} \lambda^t\right] \in (0, \infty) \tag{82}$$
exists and is finite.

7. **Strict long-run nuisance variability.** Let
$$F(\vartheta, \zeta) := \Gamma(\vartheta) + \Psi(\zeta\vartheta). \tag{83}$$
On the comparison range $\zeta \in [0, \bar{\zeta}]$, the positive root has the standard Cramér-root sign pattern: $F(\vartheta, \zeta) < 0$ for all sufficiently small $\vartheta > 0$, and $\vartheta_\star(\zeta)$ is the unique positive upward-crossing root. In particular, the root is regular in the sense that
$$\partial_\vartheta F(\vartheta_\star(\zeta), \zeta) > 0. \tag{84}$$
Moreover, the nuisance has nontrivial long-run SCGF contribution on the relevant positive tilted arguments: for every $v \in (0, \bar{\zeta}\vartheta_\star(0)]$,
$$\Psi(v) > 0, \qquad \Psi'(v) > 0. \tag{85}$$

Remark B.30 (One-step nondegeneracy is not enough under Markov dependence). Item 7 of Assumption B.29 is a long-run condition, not a one-step variance condition. This distinction matters only in the dependent Markov-additive setting. A stationary nuisance process can be one-step nondegenerate and centered while contributing zero exponential growth rate to its partial sums. For example, if $\varepsilon_t = U_t - U_{t-1}$, where (U_t) is bounded, i.i.d., centered, and independent of the Markov state process, then
$$\sum_{t=1}^{n} \varepsilon_t = U_n - U_0, \qquad \Psi(v) = \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} e^{v(U_n - U_0)} = 0.$$
The root equation $\Gamma(\vartheta) + \Psi(\zeta\vartheta) = 0$ is then independent of ζ, thus the CUSUM large-deviation exponent does not necessarily degrade even though the increment-level SINE increases with ζ. The strict SCGF-noise condition excludes precisely such asymptotically invisible coboundary nuisances.

B.9.2 SCGF ROOTS AND EFFECTIVE SLOPE

We first derive the SCGF root and its noise ordering, then relate them to ARL and delay through the stated first-passage assumptions.

SCGF decomposition. Independence between the full signal and nuisance processes factors the exponential moment of their partial sums. Taking the logarithm per time step gives the following decomposition.

Lemma B.31 (SCGF decomposition under independent additive noise)**.** *Under the Markov-order-N setup above, define*

$$\Gamma(\vartheta) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_0 \exp\left(\vartheta \sum_{t=1}^{n} \lambda^t\right), \tag{86}$$

$$\Psi(v) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp\left(v \sum_{t=1}^{n} \varepsilon_t\right), \tag{87}$$

and

$$\Gamma_{\hat{\lambda}}(\theta) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_0 \exp\left(\theta \sum_{t=1}^{n} \hat{\lambda}^t\right), \qquad \hat{\lambda}^t = \alpha\lambda^t + \sigma\varepsilon_t. \tag{88}$$

Then

$$\Gamma_{\hat{\lambda}}(\theta) = \Gamma(\alpha\theta) + \Psi(\sigma\theta). \tag{89}$$

Equivalently, after writing $\vartheta = \alpha\theta$ and $\zeta = \sigma/\alpha$, the positive root equation for the large-deviation exponent becomes

$$\Gamma(\vartheta) + \Psi(\zeta\vartheta) = 0. \tag{90}$$

Root monotonicity. The strict long-run nuisance condition determines how the root changes with ζ. The derivative below makes this ordering explicit.

Lemma B.32 (Monotonicity of the SCGF root)**.** *Let*

$$F(\vartheta, \zeta) := \Gamma(\vartheta) + \Psi(\zeta\vartheta), \qquad \vartheta > 0, \quad \zeta \in [0, \bar{\zeta}].$$

Assume that for each $\zeta \in [0, \bar{\zeta}]$ there exists a unique positive Cramér root $\vartheta_\star(\zeta)$ satisfying

$$F(\vartheta_\star(\zeta), \zeta) = 0,$$

with the standard root sign pattern: $F(\vartheta, \zeta) < 0$ for all sufficiently small $\vartheta > 0$, and the root crosses upward. Assume further that

$$\partial_\vartheta F(\vartheta_\star(\zeta), \zeta) > 0 \qquad \forall \zeta \in [0, \bar{\zeta}], \tag{91}$$

and that the nuisance SCGF is strictly positive and strictly increasing on the relevant positive tilted range:

$$\Psi(v) > 0, \qquad \Psi'(v) > 0, \qquad \forall v \in (0, \bar{\zeta}\vartheta_\star(0)]. \tag{92}$$

Then $\vartheta_\star(\zeta)$ is strictly decreasing on $[0, \bar{\zeta}]$. More precisely, for every $\zeta \in (0, \bar{\zeta}]$,

$$\vartheta'_\star(\zeta) = -\frac{\vartheta_\star(\zeta)\, \Psi'(\zeta\vartheta_\star(\zeta))}{\Gamma'(\vartheta_\star(\zeta)) + \zeta\, \Psi'(\zeta\vartheta_\star(\zeta))} < 0. \tag{93}$$

In addition, $\vartheta_\star(\zeta) < \vartheta_\star(0)$ for every $\zeta > 0$.

From the SCGF root to sequential performance. The SCGF root and the first-passage expansions have separate roles. Lemma B.31 identifies the learned-score root as $\vartheta_\star(\zeta)/\alpha$, while centering leaves the post-change drift αI_{rate} unchanged. The following assumption relates these quantities to ARL and delay. The relevant renewal theory includes the i.i.d. results of Lai & Siegmund (1977; 1979), Markov-additive results in Melfi (1992); Fuh (2003; 2004), and general quickest-detection analyses (Lai, 1998; Chan & Lai, 2000; Tartakovsky & Veeravalli, 2005).

Assumption B.33 (Regularity for standard Markov-additive CUSUM asymptotics)**.** Under Assumption B.29, and under the standard Markov-additive / nonlinear-renewal regularity conditions for first-passage asymptotics (Lai, 1995; 1998; Fuh, 2003; 2004; Tartakovsky & Veeravalli, 2005; Kontoyiannis & Meyn, 2005), for the CUSUM stopping time T_h driven by the accumulated conditional evidence path $\widehat{\Lambda}_{\mathrm{cond}}^{N+1:t}$ (equivalently, by its one-step increments $\hat{\lambda}^t$) in the Markov-order-N setup, as $h \to \infty$,

$$\log \mathrm{ARL}(h; \zeta) = \frac{\vartheta_\star(\zeta)}{\alpha}\, h + o(h), \qquad \mathrm{ADD}_1(h; \zeta) = \frac{h}{\alpha I_{\mathrm{rate}}} + o(h). \tag{94}$$

Theorem B.34 (Asymptotic log-ARL–ADD_1 slope under Markov dependence)**.** *Under Assumptions B.29 and B.33, including the strict long-run nuisance-variability condition in Assumption B.29, the root $\vartheta_\star(\zeta)$ is strictly decreasing on the comparison range $[0, \bar{\zeta}]$, and the large-threshold CUSUM tradeoff satisfies*

$$\log \mathrm{ARL}(h;\zeta) = \big(I_{\mathrm{rate}}\,\vartheta_\star(\zeta)\big)\,\mathrm{ADD}_1(h;\zeta) + o(h) \qquad \text{as } h \to \infty. \tag{95}$$

Therefore the first-order slope of the asymptotic $(\log \mathrm{ARL})$–(ADD_1) frontier is

$$\mathcal{E}_{\text{Markov-QCD}}(\zeta) := I_{\mathrm{rate}}\,\vartheta_\star(\zeta), \tag{96}$$

and $\mathcal{E}_{\text{Markov-QCD}}(\zeta)$ is strictly decreasing in ζ.

B.9.3 Increment-level SINE and NITA ordering

For the projective-geometry side, we evaluate SINE on the one-step conditional increments, equivalently after applying the linear difference representation R_Δ to the accumulated conditional-LLR path.

Proposition B.35 (Markov-family SINE monotonicity on the increment representation)**.** *Consider the affine score family*

$$\hat{\lambda}^t = \alpha\lambda^t + \sigma\varepsilon_t, \qquad \alpha > 0,\ \sigma \geq 0,$$

under the Markov-order-N setup above. Assume that under a reference law Q, the joint process

$$(\lambda^t, \varepsilon_t)_{t\geq 1}$$

is stationary and ergodic, with

$$\mathbb{E}[\varepsilon_t] = 0, \qquad \mathbb{E}[\varepsilon_t^2] = 1, \qquad 0 < \mathbb{E}_Q[(\lambda^1)^2] < \infty,$$

and that ε_t is independent of λ^t for each t. Define

$$\rho_T := \max\left\{0,\ \frac{\frac{1}{T}\sum_{t=1}^T \hat{\lambda}^t\lambda^t}{\sqrt{\frac{1}{T}\sum_{t=1}^T(\hat{\lambda}^t)^2}\,\sqrt{\frac{1}{T}\sum_{t=1}^T(\lambda^t)^2}}\right\}, \qquad \mathrm{SINE}_T^2 := 1 - \rho_T^2. \tag{97}$$

Then, with $\zeta := \sigma/\alpha$,

$$\mathrm{SINE}_T^2 \xrightarrow[T\to\infty]{\text{a.s.}} \mathrm{SINE}_\infty^2(\zeta) := \frac{\zeta^2}{\mathbb{E}_Q[(\lambda^1)^2] + \zeta^2}. \tag{98}$$

In particular, $\mathrm{SINE}_\infty(\zeta)$ is strictly increasing in ζ.

Equation Eq. (97) uses the Euclidean geometry of one realized increment stack. Its limit is population SINE between $\alpha\lambda^1 + \sigma\varepsilon_1$ and λ^1 in $L^2(Q)$. By stationarity, the population increment-stack SINE has this same value for every fixed T: each squared norm and the inner product acquire the same factor T. This does not assert exact empirical orthogonality on a finite stack or replace increment geometry by cumulative-trajectory geometry.

Asymptotic NITA for Markov-order-N QCD. The effective-slope ordering now enters the common comparison argument of Proposition B.22. Let $F_{\mathrm{M}}^\star(d;\zeta)$ denote the upper untruncated $(\mathrm{ADD}_1, \mathrm{ARL})$ frontier induced by the Markov-order-N CUSUM rule before truncation by any comparison box.

Corollary B.36 (Asymptotic NITA ordering for Markov-order-N QCD)**.** *Under Assumptions B.29 and B.33, for $\zeta \in [0, \bar{\zeta}]$,*

$$\log F_{\mathrm{M}}^\star(d;\zeta) = \mathcal{E}_{\text{Markov-QCD}}(\zeta)\,d + o(d), \qquad d \to \infty, \tag{99}$$

where

$$\mathcal{E}_{\text{Markov-QCD}}(\zeta) = I_{\mathrm{rate}}\vartheta_\star(\zeta)$$

is strictly decreasing in ζ. Consequently, for every pair $0 \leq \zeta_1 < \zeta_2 \leq \bar{\zeta}$, there exists $d_0 < \infty$ such that

$$F_{\mathrm{M}}^\star(d;\zeta_1) > F_{\mathrm{M}}^\star(d;\zeta_2) \qquad \forall d \geq d_0 \tag{100}$$

throughout the high-threshold region. Therefore, on any common non-saturated high-delay comparison box $[d_0, D^*] \times [0, Y^*]$ *with* $D^* > d_0$*, the induced asymptotic NITA*

$$\mathrm{NITA}^{[d_0,D^*]}_{\text{Markov-QCD},Y^*}(\zeta) := \frac{1}{(D^*-d_0)Y^*}\int_{d_0}^{D^*} F^{\star}_{\mathrm{M}}(x;\zeta)\,dx \tag{101}$$

satisfies

$$\mathrm{NITA}^{[d_0,D^*]}_{\text{Markov-QCD},Y^*}(\zeta_1) > \mathrm{NITA}^{[d_0,D^*]}_{\text{Markov-QCD},Y^*}(\zeta_2). \tag{102}$$

If the increment-level SINE assumptions of Proposition B.35 also hold, then this asymptotic NITA ordering is antitone to $\mathrm{SINE}_\infty(\zeta)$ *along the affine Markov-order-*N *family.*

Connection to Proposition B.22. In the Markov-QCD setting, the QCD-type frontier representation Eq. (61) in Proposition B.22 is supplied by Corollary B.36, with

$$\mathcal{E}_\Xi(\zeta) = \mathcal{E}_{\text{Markov-QCD}}(\zeta) = I_{\mathrm{rate}}\vartheta_\star(\zeta).$$

The detector/frontier side uses the accumulated conditional-evidence path $\widehat{\Lambda}^{N+1:t}_{\mathrm{cond}}$, while Proposition B.35 supplies the increment-level SINE side via R_Δ. Thus the Markov-QCD extension uses the same asymptotic NITA logic as the binary i.i.d. QCD theorem, with the i.i.d. Cramér root replaced by the Markov SCGF root.

B.9.4 Initial-block correction and regularity conditions

For finite-memory models, the exact pathwise Markov LLR differs from the accumulated conditional-LLR path only through the initial block. The following proposition records this explicitly.

Proposition B.37 (Initial-block correction for Nth-order Markov LLRs)**.** *Fix* $N \geq 1$ *and suppose that under regime* $y \in \{0,1\}$ *the law of the first* N *samples admits a density* $\mu_y(x_{1:N})$*, while for* $t \geq N+1$ *the transition density is* $p_y(x_t \mid x_{t-N:t-1})$*. Assume* $\mathbb{P}_1^{(n)} \ll \mathbb{P}_0^{(n)}$ *for every* $n \geq N$*. Then for every* $n \geq N$*,*

$$\log \frac{d\mathbb{P}_1^{(n)}}{d\mathbb{P}_0^{(n)}}(X_{1:n}) = B_N(X_{1:N}) + \sum_{t=N+1}^{n} \lambda^t, \tag{103}$$

where

$$B_N(X_{1:N}) := \log \frac{\mu_1(X_{1:N})}{\mu_0(X_{1:N})}, \qquad \lambda^t = \log \frac{p_1(X_t \mid X_{t-N:t-1})}{p_0(X_t \mid X_{t-N:t-1})}. \tag{104}$$

In particular, if the two regimes share the same initial law, i.e. $\mu_1 = \mu_0$*, then* $B_N \equiv 0$ *and the accumulated conditional LLR is exact. If instead* B_N *is integrable under both regimes, then it contributes only an* $O(1)$ *boundary term and does not affect asymptotic rates normalized by* n *or large-threshold asymptotics normalized by* h*.*

Thus, on the detector/frontier side, using accumulated conditional-LLR language is first-order asymptotically equivalent to using the exact pathwise Markov LLR.

Verifying the Markov regularity conditions. The model assumptions can be checked using Markov-additive spectral and renewal theory. One standard route uses the following conditions:

1. the augmented chain (S_t) under $\mathbb{P}_0$ is geometrically ergodic on X^N;
2. the additive functional λ^t has finite exponential moments on a neighborhood of $[0,1]$;
3. the tilted kernel
$$(Q_\vartheta f)(s) := \mathbb{E}_0\left[e^{\vartheta\lambda^{t+1}} f(S_{t+1}) \,\middle|\, S_t = s\right] \tag{105}$$
has finite spectral radius and satisfies the usual multiplicative drift / irreducibility conditions;
4. the noise process (ε_t) has a finite SCGF in a neighborhood of 0 and satisfies the strict long-run variability condition in Eq. (85);
5. the CUSUM score process has negative drift under $\mathbb{P}_0$ and positive drift under $\mathbb{P}_1$.

Under these conditions, $\Gamma(\vartheta) = \log r(Q_\vartheta)$ exists and is convex (and typically C^1 on the relevant domain), while large-threshold CUSUM asymptotics follow from standard Markov renewal / Markov-random-walk theory and non-i.i.d. quickest detection theory; see, e.g., Melfi (1992); Fuh (2003; 2004); Lai (1998); Chan & Lai (2000); Kontoyiannis & Meyn (2005); Meyn & Tweedie (2009).

B.10 BINARY SPRT: EFFECTIVE SLOPES AND SINE–NITA ORDERING

We connect the two directional error rates of binary SPRT to a single speed–accuracy slope. Classical stopped-sum and threshold-calibration arguments (Wald, 1947; Siegmund, 1985; Tartakovsky et al., 2014) supply the sequential framework, and the affine-noise analysis identifies how the learned-score slope decreases with nuisance. We first use equal-error calibration, then treat aggregate accuracy and a symmetric threshold sweep.

B.10.1 LEARNED-SCORE MODEL AND DIRECTIONAL ROOTS

Binary SPRT uses the hypothesis-label convention $\mathcal{Y} = \{0, 1\}$, and $\lambda_{10} = \log(p_1/p_0)$ denotes evidence for class/hypothesis 1 against class/hypothesis 0.

Directional error exponents. With positive boundary widths (a, b), the accumulated score continues inside $(-b, a)$ and declares class 1 or 0 at the upper or lower boundary, respectively. The score-based stopping time is specified in Eq. (107). For the affine score family

$$g_\sigma = \alpha\lambda + \sigma\xi, \qquad \lambda := \log \frac{p_1(X)}{p_0(X)}, \qquad \alpha > 0,$$

the two directional large-deviation roots entering the SPRT analysis are

$$\mathbb{E}_0[e^{\kappa_0(\sigma) g_\sigma}] = 1, \qquad \mathbb{E}_1[e^{-\kappa_1(\sigma) g_\sigma}] = 1. \tag{106}$$

They correspond, respectively, to the upper-boundary tail under H_0 and the lower-boundary tail under H_1. We also write

$$D_{10} := \mathbb{E}_1[\lambda] = \mathrm{KL}(P_1 \| P_0), \qquad D_{01} := -\mathbb{E}_0[\lambda] = \mathrm{KL}(P_0 \| P_1).$$

Assumption B.38 (Affine SPRT score family)**.** The SPRT score admits the affine representation

$$g_\sigma = \alpha\lambda + \sigma\xi, \qquad \lambda := \log \frac{p_1(X)}{p_0(X)}, \qquad \alpha > 0,\ \sigma \geq 0,$$

where ξ is centered, nondegenerate, and independent of λ under both $\mathbb{P}_0$ and $\mathbb{P}_1$. Moreover, g_σ has finite first moment under both $\mathbb{P}_0$ and $\mathbb{P}_1$, the directional root equations in Eq. (106) admit unique positive solutions, and the exponential moments needed in a neighborhood of these roots are finite.

Accumulated score and stopping time. For each fixed σ, let $g_{\sigma,1}, g_{\sigma,2}, \ldots$ be i.i.d. copies of g_σ, let

$$\mathcal{F}_t^{(\sigma)} := \sigma(g_{\sigma,1}, \ldots, g_{\sigma,t}),$$

and define the stopped score random walk

$$S_n^{(\sigma)} := \sum_{t=1}^{n} g_{\sigma,t}, \qquad \tau_{a,b}^{(\sigma)} := \inf\{n \geq 1 : S_n^{(\sigma)} \notin (-b, a)\}. \tag{107}$$

In the exact-LLR case $\sigma = 0$, one has $S_n^{(0)} = \alpha\Lambda^{1:n}$.

Lemma B.39 (Factorized root equations and monotonicity)**.** *Assume the setting of Assumption B.38. Define*

$$M_0(u) := \mathbb{E}_0[e^{u\lambda}], \qquad M_1(v) := \mathbb{E}_1[e^{-v\lambda}], \tag{108}$$

$$N_+(s) := \mathbb{E}[e^{s\xi}], \qquad N_-(s) := \mathbb{E}[e^{-s\xi}]. \tag{109}$$

Then the root equations Eq. (106) become

$$M_0(\alpha\kappa_0(\sigma))\, N_+(\sigma\kappa_0(\sigma)) = 1, \qquad M_1(\alpha\kappa_1(\sigma))\, N_-(\sigma\kappa_1(\sigma)) = 1. \tag{110}$$

Moreover, if ξ is centered and nondegenerate, then both $\kappa_0(\sigma)$ and $\kappa_1(\sigma)$ are strictly decreasing in σ.

B.10.2 EQUAL-ERROR CALIBRATION AND EFFECTIVE SLOPE

Fix a binary class prior $\pi \in (0,1)$ and define

$$\mathrm{Acc}_\pi := \pi\,\mathbb{P}_1(d_\tau = 1) + (1-\pi)\,\mathbb{P}_0(d_\tau = 0), \qquad T_\pi := \pi\,\mathbb{E}_1[\tau] + (1-\pi)\,\mathbb{E}_0[\tau].$$

Let

$$\gamma := 1 - \alpha_{\mathrm{I}} = 1 - \beta_{\mathrm{II}} \in (1/2, 1), \qquad L(\gamma) := \log \frac{\gamma}{1-\gamma}.$$

We calibrate the upper and lower boundary widths so that the false-positive and false-negative probabilities are both equal to $1-\gamma$. Under this calibration, both class-conditional accuracies equal γ, thus $\mathrm{Acc}_\pi = \gamma$ for every fixed class prior π. Equal-error calibration gives one accuracy parameter for the two-boundary family. Proposition B.45 and Corollary B.46 below show how the ordering extends to aggregate accuracy and symmetric thresholds.

Assumption B.40 (Large-threshold equal-error calibration for SPRT)**.** For each fixed σ, let $a_\sigma(\gamma) > 0$ and $b_\sigma(\gamma) > 0$ denote the upper and lower SPRT boundary widths, thus the continuation interval is $(-b_\sigma(\gamma), a_\sigma(\gamma))$, calibrated so that the type-I and type-II errors satisfy

$$\alpha_I = \beta_{II} = 1 - \gamma, \qquad \gamma \in (1/2, 1).$$

As $\gamma \uparrow 1$,

$$a_\sigma(\gamma) = \frac{L(\gamma)}{\kappa_0(\sigma)} + O(1), \qquad b_\sigma(\gamma) = \frac{L(\gamma)}{\kappa_1(\sigma)} + O(1), \qquad L(\gamma) := \log \frac{\gamma}{1-\gamma},$$

where $\kappa_0(\sigma)$ and $\kappa_1(\sigma)$ are the directional roots defined in the SPRT setup. Moreover, the expected overshoot contributions under P_1 and P_0 are absorbed into the $O(1)$ terms above, and the associated stopping time $\tau^{(\sigma)}_{a_\sigma(\gamma), b_\sigma(\gamma)}$ has finite mean under both P_1 and P_0. For finite-boundary SPRTs with nonzero drifts and the moment conditions assumed above, this is standard; see, e.g., (Wald, 1947, Ch. III), (Siegmund, 1985, Ch. III), and (Tartakovsky et al., 2014, Secs. 3.2–3.3).

Stopped-sum identity. Define

$$\mu_1(\sigma) := \mathbb{E}_1[g_\sigma], \qquad \mu_0(\sigma) := -\mathbb{E}_0[g_\sigma].$$

For any fixed σ, boundary pair (a, b), and regime $y \in \{0,1\}$, if $\mathbb{E}_y[\tau^{(\sigma)}_{a,b}] < \infty$, then the standard Wald identity for stopped i.i.d. sums gives

$$\mathbb{E}_y\left[S^{(\sigma)}_{\tau^{(\sigma)}_{a,b}}\right] = \mathbb{E}_y[g_\sigma]\,\mathbb{E}_y\left[\tau^{(\sigma)}_{a,b}\right], \qquad y \in \{0,1\}, \tag{111}$$

see, e.g., (Wald, 1947, Ch. III), (Siegmund, 1985, Ch. III), and (Tartakovsky et al., 2014, Sec. 3.2); for completeness, Appendix C.6 records a direct proof tailored to the present notation. In particular,

$$\mathbb{E}_1\left[S^{(\sigma)}_{\tau^{(\sigma)}_{a,b}}\right] = \mu_1(\sigma)\,\mathbb{E}_1\left[\tau^{(\sigma)}_{a,b}\right], \qquad \mathbb{E}_0\left[S^{(\sigma)}_{\tau^{(\sigma)}_{a,b}}\right] = -\mu_0(\sigma)\,\mathbb{E}_0\left[\tau^{(\sigma)}_{a,b}\right]. \tag{112}$$

For the calibrated family $\tau^{(\sigma)} = \tau^{(\sigma)}_{a_\sigma(\gamma), b_\sigma(\gamma)}$ used below, Assumption B.40 supplies the required finite-mean condition.

Proposition B.41 (High-accuracy class-conditional speed scaling)**.** *Under Assumptions B.38 and B.40, let*

$$\tau^{(\sigma)} := \tau^{(\sigma)}_{a_\sigma(\gamma), b_\sigma(\gamma)}.$$

Then the class-conditional mean hitting times satisfy, as $\gamma \uparrow 1$,

$$\mathbb{E}_1[\tau^{(\sigma)}] = \frac{L(\gamma)}{\mathcal{E}_{10}(\sigma)} + o(L(\gamma)), \qquad \mathbb{E}_0[\tau^{(\sigma)}] = \frac{L(\gamma)}{\mathcal{E}_{01}(\sigma)} + o(L(\gamma)), \tag{113}$$

where

$$\mathcal{E}_{10}(\sigma) := \kappa_0(\sigma)\mu_1(\sigma), \qquad \mathcal{E}_{01}(\sigma) := \kappa_1(\sigma)\mu_0(\sigma).$$

Consequently, the equal-weight average hitting time satisfies

$$T_{\mathrm{eq}}(\gamma;\sigma) = \frac{L(\gamma)}{\mathcal{E}_{\mathrm{SPRT}}(\sigma)} + o(L(\gamma)), \qquad \mathcal{E}_{\mathrm{SPRT}}(\sigma) := \frac{2}{\frac{1}{\mathcal{E}_{10}(\sigma)} + \frac{1}{\mathcal{E}_{01}(\sigma)}}. \tag{114}$$

Corollary B.42 (Exact-LLR special case). *If $\sigma = 0$ so that $g_0 = \alpha\lambda$ is a positively scaled exact LLR, then*

$$\kappa_0(0) = \kappa_1(0) = \frac{1}{\alpha}, \qquad \mathcal{E}_{10}(0) = D_{10}, \qquad \mathcal{E}_{01}(0) = D_{01},$$

and therefore

$$\mathcal{E}_{\mathrm{SPRT}}(0) = \frac{2}{\frac{1}{D_{10}} + \frac{1}{D_{01}}}. \tag{115}$$

If, in addition, the equal-error design uses symmetric boundary widths

$$a = b = A,$$

and overshoot contributes only lower-order terms, then

$$\mathbb{E}_1[\tau] = \frac{g(\gamma)}{D_{10}} + O(1), \qquad \mathbb{E}_0[\tau] = \frac{g(\gamma)}{D_{01}} + O(1), \qquad T_{\mathrm{eq}}(\gamma) = \frac{g(\gamma)}{\mathcal{E}_{\mathrm{SPRT}}(0)} + O(1), \tag{116}$$

where

$$g(\gamma) := (2\gamma - 1)\log\frac{\gamma}{1-\gamma}. \tag{117}$$

Since $g(\gamma) = L(\gamma) + o(L(\gamma))$ as $\gamma \uparrow 1$, Eq. (116) recovers the corresponding high-accuracy equal-error scaling stated in Proposition B.41.

Corollary B.43 (Mean hitting time under a fixed class prior). *For a fixed prior weight $\pi \in (0,1)$, define the prior-weighted mean hitting time*

$$T_\pi(\gamma;\sigma) := \pi\,\mathbb{E}_1[\tau^{(\sigma)}] + (1-\pi)\,\mathbb{E}_0[\tau^{(\sigma)}]. \tag{118}$$

Then, under the same assumptions as Proposition B.41,

$$T_\pi(\gamma;\sigma) = \frac{L(\gamma)}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma)} + o(L(\gamma)), \qquad \mathcal{E}_{\mathrm{SPRT},\pi}(\sigma) := \frac{1}{\frac{\pi}{\mathcal{E}_{10}(\sigma)} + \frac{1-\pi}{\mathcal{E}_{01}(\sigma)}}. \tag{119}$$

In the exact-LLR case $\sigma = 0$, this reduces to

$$\mathcal{E}_{\mathrm{SPRT},\pi}(0) = \frac{1}{\frac{\pi}{D_{10}} + \frac{1-\pi}{D_{01}}}. \tag{120}$$

B.10.3 High-accuracy SINE–NITA ordering

For an accuracy interval $\mathcal{U} \subset (1/2, 1)$, we evaluate Definition 3.3 after restricting the threshold sweep to operating points whose ordinary accuracy Acc_π lies in $\mathcal{U}$, with a shared normalization box across the compared models on the (T_π, γ) axes. By Lemma B.39, both directional roots are strictly decreasing in σ, while centered noise leaves the directional drifts unchanged. Thus,

$$\mathcal{E}_{10}(\sigma), \qquad \mathcal{E}_{01}(\sigma), \qquad \mathcal{E}_{\mathrm{SPRT},\pi}(\sigma)$$

are strictly decreasing in σ for every fixed class prior $\pi \in (0,1)$. The next proposition applies the comparison step in Proposition B.22. It gives eventual high-accuracy ordering and a fixed-interval criterion in which the slope gap must exceed the remainder difference.

Proposition B.44 (Asymptotic high-accuracy ordering for binary SPRT-NITA). *Fix a class prior $\pi \in (0,1)$ and let $T_\pi(\gamma;\sigma)$ denote the corresponding mean hitting time from Corollary B.43. Under Assumptions B.38 and B.40, the effective slopes*

$$\mathcal{E}_{10}(\sigma), \qquad \mathcal{E}_{01}(\sigma), \qquad \mathcal{E}_{\mathrm{SPRT},\pi}(\sigma)$$

are strictly decreasing in σ. Consequently, for every pair $\sigma_1 < \sigma_2$, there exists $\gamma_0 \in (1/2, 1)$ such that for every $\gamma_1 \in (\gamma_0, 1)$,

$$T_\pi(\gamma;\sigma_1) < T_\pi(\gamma;\sigma_2) \qquad \forall \gamma \in [\gamma_0, \gamma_1]. \tag{121}$$

Therefore

$$\mathrm{NITA}_{\mathrm{SPRT}}^{[\gamma_0,\gamma_1]}(\sigma_1) > \mathrm{NITA}_{\mathrm{SPRT}}^{[\gamma_0,\gamma_1]}(\sigma_2) \tag{122}$$

when both are evaluated on the common high-accuracy interval $[\gamma_0, \gamma_1]$ with a shared normalization box on the (T_π, γ) axes.

More generally, fix an accuracy interval $\mathcal{U} = [\gamma_0, \gamma_1] \subset (1/2, 1)$ contained in the regime where the high-accuracy approximation is used. Assume that for each model index σ,

$$T_\pi(\gamma; \sigma) = \frac{L(\gamma)}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma)} + r(\gamma; \sigma), \qquad \gamma \in \mathcal{U}, \tag{123}$$

where $L(\gamma) = \log \frac{\gamma}{1-\gamma}$ and $r(\cdot; \sigma)$ is a remainder term. Suppose that for two models σ_1, σ_2,

$$\sup_{\gamma \in \mathcal{U}} |r(\gamma; \sigma_1) - r(\gamma; \sigma_2)| \le \Delta \tag{124}$$

for some $\Delta \ge 0$. If

$$\inf_{\gamma \in \mathcal{U}} L(\gamma) \left(\frac{1}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma_2)} - \frac{1}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma_1)} \right) > \Delta, \tag{125}$$

then

$$T_\pi(\gamma; \sigma_1) < T_\pi(\gamma; \sigma_2) \qquad \forall \gamma \in \mathcal{U}. \tag{126}$$

Consequently, the interval-restricted NITA from Definition 3.3 is larger for σ_1 than for σ_2 on the common interval $\mathcal{U}$ with a shared normalization box on the (T_π, γ) axes.

Under the fixed orthogonal population score family in Proposition B.22, SINE increases with nuisance. Combining that geometry with the slope ordering above gives the binary SPRT SINE–NITA relation. Proofs of the sequential expansions and the comparison criteria are in Appendix C.6.

B.10.4 Aggregate-accuracy and symmetric-threshold calibration

The preceding statements use equal-error calibration to parameterize the binary-SPRT high-accuracy frontier. The same ordering mechanism also holds for the ordinary aggregate accuracy used in sequential classification. For a boundary pair (a, b), define the directional errors and prior-weighted performance coordinates

$$\alpha_\sigma(a, b) := \mathbb{P}_0(d_{\tau_{a,b}^{(\sigma)}} = 1), \qquad \beta_\sigma(a, b) := \mathbb{P}_1(d_{\tau_{a,b}^{(\sigma)}} = 0), \tag{127}$$

$$\gamma_{\pi,\sigma}(a, b) := 1 - (1-\pi)\alpha_\sigma(a, b) - \pi\beta_\sigma(a, b), \quad T_{\pi,\sigma}(a, b) := \pi\mathbb{E}_1[\tau_{a,b}^{(\sigma)}] + (1-\pi)\mathbb{E}_0[\tau_{a,b}^{(\sigma)}]. \tag{128}$$

Here $\gamma_{\pi,\sigma}$ is ordinary prior-weighted terminal accuracy. Let

$$T^\star_{\pi,\mathrm{agg}}(\gamma; \sigma) := \inf\{T_{\pi,\sigma}(a, b) : a, b > 0,\ \gamma_{\pi,\sigma}(a, b) \ge \gamma\} \tag{129}$$

be the two-boundary aggregate-accuracy frontier. This is the binary $K = 2$ analogue of the weighted-accuracy frontier used in multiclass SPRT.

Proposition B.45 (Two-boundary aggregate-accuracy binary SPRT)**.** *Assume Assumption B.38. Suppose that the standard large-boundary SPRT estimates hold uniformly on the high-accuracy feasible region: as $\min\{a, b\} \to \infty$,*

$$\alpha_\sigma(a, b) = \exp\{-\kappa_0(\sigma)a + O(1)\}, \qquad \beta_\sigma(a, b) = \exp\{-\kappa_1(\sigma)b + O(1)\}, \tag{130}$$

$$T_{\pi,\sigma}(a, b) = \frac{\pi a}{\mu_1(\sigma)} + \frac{(1-\pi)b}{\mu_0(\sigma)} + O(1). \tag{131}$$

Then, as $\gamma \uparrow 1$,

$$T^\star_{\pi,\mathrm{agg}}(\gamma; \sigma) = \frac{L(\gamma)}{\mathcal{E}_{\mathrm{agg},\pi}(\sigma)} + o(L(\gamma)), \tag{132}$$

where

$$\mathcal{E}_{\mathrm{agg},\pi}(\sigma) := \left[\frac{\pi}{\kappa_0(\sigma)\mu_1(\sigma)} + \frac{1-\pi}{\kappa_1(\sigma)\mu_0(\sigma)} \right]^{-1} = \mathcal{E}_{\mathrm{SPRT},\pi}(\sigma). \tag{133}$$

Consequently, under the affine centered-noise family, $\mathcal{E}_{\mathrm{agg},\pi}(\sigma)$ is strictly decreasing in σ. Thus, for every $\sigma_1 < \sigma_2$ there exists $\gamma_0 \in (1/2, 1)$ such that on every common high-accuracy interval $[\gamma_0, \gamma_1]$, $\gamma_1 \in (\gamma_0, 1)$, the aggregate-accuracy NITA induced by Eq. (129) is larger for σ_1 than for σ_2.

Corollary B.46 (Symmetric-threshold aggregate-accuracy binary SPRT)**.** *In the setting of Proposition B.45, restrict the boundary family to the scalar sweep $(a, b) = (c, c)$. Define*

$$\gamma^{\mathrm{sym}}_{\pi,\sigma}(c) := \gamma_{\pi,\sigma}(c, c), \qquad T^{\mathrm{sym}}_{\pi,\sigma}(c) := T_{\pi,\sigma}(c, c), \tag{134}$$

and let

$$T^{\star}_{\pi,\mathrm{sym}}(\gamma; \sigma) := \inf\{T^{\mathrm{sym}}_{\pi,\sigma}(c) : c > 0,\ \gamma^{\mathrm{sym}}_{\pi,\sigma}(c) \geq \gamma\} \tag{135}$$

be the corresponding high-accuracy frontier along this one-dimensional sweep. Then

$$T^{\star}_{\pi,\mathrm{sym}}(\gamma; \sigma) = \frac{L(\gamma)}{\mathcal{E}_{\mathrm{sym},\pi}(\sigma)} + o(L(\gamma)), \qquad \gamma \uparrow 1, \tag{136}$$

where

$$\mathcal{E}_{\mathrm{sym},\pi}(\sigma) := \frac{\min\{\kappa_0(\sigma), \kappa_1(\sigma)\}}{\frac{\pi}{\mu_1(\sigma)} + \frac{1-\pi}{\mu_0(\sigma)}}. \tag{137}$$

Under the affine centered-noise family, $\mathcal{E}_{\mathrm{sym},\pi}(\sigma)$ is strictly decreasing in σ. Therefore the same high-accuracy NITA ordering as in Proposition B.44 holds on sufficiently high common accuracy intervals for the symmetric scalar threshold sweep.

Remark B.47 (How the three binary-SPRT calibrations relate). Equal-error calibration is a convenient canonical parameterization of the two-boundary binary SPRT frontier, because it makes both directional errors equal to a single scalar $1 - \gamma$. Proposition B.45 shows that the equality of the two errors is not required for the leading high-accuracy slope: ordinary prior-weighted accuracy forces both directional errors to be $O(1 - \gamma)$ when $\pi \in (0, 1)$, and therefore gives the same first-order slope $\mathcal{E}_{\mathrm{agg},\pi} = \mathcal{E}_{\mathrm{SPRT},\pi}$. Corollary B.46 gives the corresponding statement for the scalar sweep $a = b = c$. This symmetric path is generally not the first-order optimal two-boundary frontier when $\kappa_0(\sigma) \neq \kappa_1(\sigma)$, but its effective slope is governed by the slower directional root, and this root also decreases with nuisance level. Thus the qualitative SINE–NITA antitone ordering persists for both two-boundary aggregate calibration and the symmetric threshold family used as a simple experimental protocol.

Remark B.48 (Exact LLRs and learned directional slopes). For a positively scaled exact LLR, both directional roots are exactly $1/\alpha$, and the symmetric equal-error design recovers the classical one-parameter scaling by the prior-weighted information number from Corollary B.43. For general learned scores, however, the two roots $\kappa_0(\sigma)$ and $\kappa_1(\sigma)$ do not necessarily agree, therefore the high-accuracy mean-hitting-time frontier is indexed by the two directional slopes through their prior-weighted harmonic combination $\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma)$. This is why the exact-LLR formula Eq. (120) is a useful special case, but not the general learned-score law.

B.11 Multiclass and Markov SPRT: effective slopes and SINE–NITA ordering

We extend the binary-SPRT argument in two directions. For multiclass SPRT, each class must defeat every rival, thus the slowest pairwise comparison determines its leading decision time. For order-N Markov SPRT, the two directional roots come from SCGFs rather than one-step MGFs. Both constructions lead to a prior-weighted effective slope and the common SINE–NITA comparison of Proposition B.22.

Throughout, the comparison uses a fixed high-accuracy interval $\mathcal{U} \subset (0, 1)$ and one normalization box shared by the compared scores. Multiclass accuracy is the prior-weighted quantity γ_π. binary Markov-SPRT uses the equal-error accuracy γ defined in Appendix B.10. These are the ordinate values of the operating points that can be found in Definition 3.3.

B.11.1 Multiclass SPRT: the hardest-rival slope

The all-rivals decision rule is the matrix-SPRT construction introduced in Appendix A.1 (Lorden, 1977; Dragalin et al., 1999). We specify the learned-score family and the pairwise high-accuracy conditions separately, then derive their consequence for SINE and NITA.

Assumption B.49 (Affine score family for multiclass SPRT)**.** For every ordered pair (k, ℓ) with $k \neq \ell$, the learned score has the form

$$\hat{\lambda}_{k\ell} = \alpha\, \lambda_{k\ell} + \sigma\, \varepsilon_{k\ell}, \qquad \alpha > 0,\ \sigma \geq 0.$$

Assume that

$$\mathbb{E}_\ell[\varepsilon_{k\ell}] = 0, \qquad \mathbb{E}_k[\varepsilon_{k\ell}] = 0,$$

that under $\mathbb{P}_\ell$ the variable $\varepsilon_{k\ell}$ is nondegenerate and independent of $\lambda_{k\ell}$, and that there exists $s_{0,k\ell} > 0$ such that

$$\mathbb{E}_\ell[e^{s\varepsilon_{k\ell}}] < \infty \qquad \forall |s| < s_{0,k\ell}.$$

Assume also that for each $\sigma \geq 0$ the directional root equation

$$\mathbb{E}_\ell[e^{\kappa\hat{\lambda}_{k\ell}}] = 1$$

admits a unique positive solution $\kappa_{\ell\to k}(\sigma)$. For the accumulated tensor in Eq. (33), use $u_{\mathrm{MS}}(V) := \mathrm{vec}(\mathbf{\Lambda}_{\mathrm{MS}}(V))$ and its learned counterpart in the fixed population space

$$\mathcal{H}_{\mathrm{MS}} := L^2(Q_{\mathrm{val}}; \mathbb{R}^{TK(K-1)}).$$

This is $\mathcal{H}_{\mathsf{R}}$ for the full cumulative-pairwise representation in Appendix A.2. Take $u_{\mathrm{MS}} \neq 0$ and a fixed nuisance $\eta_{\mathrm{MS}} \in \mathcal{H}_{\mathrm{MS}}$, with the horizon, validation law, and signal–nuisance law fixed across σ. Assume

$$\hat{u}_{\mathrm{MS},\sigma} = \alpha u_{\mathrm{MS}} + \sigma\eta_{\mathrm{MS}}, \qquad \langle \eta_{\mathrm{MS}}, u_{\mathrm{MS}} \rangle_{\mathcal{H}_{\mathrm{MS}}} = 0.$$

Directional roots and their ordering.

Lemma B.50 (Root equations and monotonicity for multiclass SPRT)**.** *Assume the affine score family of Assumption B.49:*

$$\hat{\lambda}_{k\ell} = \alpha\,\lambda_{k\ell} + \sigma\,\varepsilon_{k\ell}, \qquad k \neq \ell.$$

For each ordered pair (k, ℓ), define

$$M_{\ell\to k}(u) := \mathbb{E}_\ell[e^{u\lambda_{k\ell}}], \qquad N_{k\ell}(s) := \mathbb{E}_\ell[e^{s\varepsilon_{k\ell}}]. \tag{138}$$

Then the directional root

$$\kappa_{\ell\to k}(\sigma) > 0 \qquad \textit{defined by} \qquad \mathbb{E}_\ell[e^{\kappa_{\ell\to k}(\sigma)\hat{\lambda}_{k\ell}}] = 1$$

satisfies

$$M_{\ell\to k}(\alpha\kappa_{\ell\to k}(\sigma))\, N_{k\ell}(\sigma\kappa_{\ell\to k}(\sigma)) = 1. \tag{139}$$

If $\mathbb{E}_\ell[\varepsilon_{k\ell}] = 0$ and $\varepsilon_{k\ell}$ is nondegenerate under $\mathbb{P}_\ell$, then $\kappa_{\ell\to k}(\sigma)$ is strictly decreasing in σ. Consequently,

$$\mathcal{E}_{k\ell}(\sigma) := \kappa_{\ell\to k}(\sigma)\mu_{k\ell}(\sigma)$$

is strictly decreasing in σ because

$$\mu_{k\ell}(\sigma) = \mathbb{E}_k[\hat{\lambda}_{k\ell}] = \alpha I_{k\ell}, \qquad I_{k\ell} := \mathbb{E}_k[\lambda_{k\ell}].$$

Assumption B.51 (Pairwise high-accuracy multiclass-SPRT regime)**.** Let $\pi_1, \ldots, \pi_K > 0$ with $\sum_{k=1}^K \pi_k = 1$, and define

$$\gamma_\pi := \sum_{k=1}^K \pi_k\, \mathbb{P}_k(d_\tau = k), \qquad T_\pi := \sum_{k=1}^K \pi_k\, \mathbb{E}_k[\tau]. \tag{140}$$

Let $L_K : (0,1) \to (0,\infty)$ be a fixed function of accuracy, independent of σ, with $L_K(\gamma_\pi) \to \infty$ as $\gamma_\pi \uparrow 1$. For each true class k and rival $\ell \neq k$, let $\tau_{k\ell}(\gamma_\pi; \sigma)$ denote the pairwise certification time required to separate class k from rival ℓ under the same high-accuracy calibration. Assume:

1. **Pairwise high-accuracy expansions.** For every ordered pair (k, ℓ) with $k \neq \ell$,

$$\tau_{k\ell}(\gamma_\pi; \sigma) = \frac{L_K(\gamma_\pi)}{\mathcal{E}_{k\ell}(\sigma)} + R_{k\ell}(\gamma_\pi; \sigma), \qquad \mathbb{E}_k[|R_{k\ell}(\gamma_\pi; \sigma)|] = o(L_K(\gamma_\pi)), \tag{141}$$

uniformly over the finite set of ordered pairs and locally uniformly over the nuisance comparison range, where

$$\mathcal{E}_{k\ell}(\sigma) := \kappa_{\ell\to k}(\sigma)\, \mu_{k\ell}(\sigma), \qquad \mu_{k\ell}(\sigma) := \mathbb{E}_k[\hat{\lambda}_{k\ell}]. \tag{142}$$

2. **All-rivals reduction.** Under $\mathbb{P}_k$, the MSPRT stopping time satisfies

$$\mathbb{E}_k[\tau] = \mathbb{E}_k\left[\max_{\ell \neq k} \tau_{k\ell}(\gamma_\pi; \sigma)\right] + o(L_K(\gamma_\pi)). \tag{143}$$

This condition captures the high-accuracy fact that a correct class-k decision requires separating k from every rival, while wrong-branch contributions are lower order.

3. **Common comparison region.** The asymptotic MSPRT-NITA comparison is carried out on a common high-accuracy interval together with a shared normalization box across models.

Define the class-k hardest-rival slope and the prior-weighted multiclass-SPRT slope by

$$\mathcal{E}_k(\sigma) := \min_{\ell \in [K]\setminus\{k\}} \mathcal{E}_{k\ell}(\sigma), \qquad \mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma) := \left(\sum_{k=1}^{K} \frac{\pi_k}{\mathcal{E}_k(\sigma)}\right)^{-1}. \tag{144}$$

From pairwise hitting times to the multiclass slope. Each pairwise hitting time records when the accumulated evidence for class k against one rival first reaches its calibrated threshold. The all-rivals rule requires every comparison to meet its threshold at the same time. Since a score can fall below its threshold after an earlier crossing, simultaneous acceptance need not occur at the largest pairwise hitting time. Under the stated high-accuracy reduction and remainder conditions, the next proposition derives the leading classwise mean time and identifies its effective slope as the smallest pairwise effective slope.

Proposition B.52 (Hardest-rival reduction for multiclass SPRT)**.** *Fix a class $k \in [K]$. For each rival $\ell \neq k$, let $\tau_{k\ell}(\gamma;\sigma)$ denote the first time the accumulated learned evidence for class k against ℓ reaches the threshold prescribed by the high-accuracy calibration. Assume that*

$$\tau_{k\ell}(\gamma;\sigma) = \frac{L_K(\gamma)}{\mathcal{E}_{k\ell}(\sigma)} + R_{k\ell}(\gamma;\sigma), \qquad \mathbb{E}_k[|R_{k\ell}(\gamma;\sigma)|] = o(L_K(\gamma)) \tag{145}$$

for every rival $\ell \neq k$, uniformly over the finite rival set. If the class-k stopping time $\tau_k(\gamma;\sigma)$ satisfies the all-rivals reduction

$$\mathbb{E}_k[\tau_k(\gamma;\sigma)] = \mathbb{E}_k\left[\max_{\ell \neq k} \tau_{k\ell}(\gamma;\sigma)\right] + o(L_K(\gamma)), \tag{146}$$

then

$$\mathbb{E}_k[\tau_k(\gamma;\sigma)] = \frac{L_K(\gamma)}{\mathcal{E}_k(\sigma)} + o(L_K(\gamma)), \qquad \mathcal{E}_k(\sigma) := \min_{\ell \neq k} \mathcal{E}_{k\ell}(\sigma). \tag{147}$$

Corollary B.53 (Weighted multiclass-SPRT speed scaling)**.** *Under the assumptions of Proposition B.52, the weighted mean hitting time*

$$T_\pi(\gamma_\pi;\sigma) := \sum_{k=1}^{K} \pi_k\, \mathbb{E}_k[\tau_k(\gamma_\pi;\sigma)]$$

satisfies

$$T_\pi(\gamma_\pi;\sigma) = \frac{L_K(\gamma_\pi)}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma)} + o(L_K(\gamma_\pi)), \qquad \mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma) := \left(\sum_{k=1}^{K} \frac{\pi_k}{\mathcal{E}_k(\sigma)}\right)^{-1}. \tag{148}$$

Theorem B.54 (Asymptotic tail antitone relation for multiclass SPRT)**.** *Under Assumptions B.49 and B.51, the following hold.*

1. ***Projective error increases with noise.*** *For the population multiclass-SPRT object in $\mathcal{H}_{\mathrm{MS}}$,*

$$\mathrm{SINE}_{\mathrm{MS}}(\sigma)^2 = \frac{\sigma^2 \|\eta_{\mathrm{MS}}\|^2_{\mathcal{H}_{\mathrm{MS}}}}{\alpha^2 \|u_{\mathrm{MS}}\|^2_{\mathcal{H}_{\mathrm{MS}}} + \sigma^2 \|\eta_{\mathrm{MS}}\|^2_{\mathcal{H}_{\mathrm{MS}}}},$$

thus $\mathrm{SINE}_{\mathrm{MS}}(\sigma)$ is nondecreasing in σ, and strictly increasing whenever $\eta_{\mathrm{MS}} \neq 0$.

2. ***The effective multiclass-SPRT slope decreases with noise.*** *Every pairwise directional slope*

$$\mathcal{E}_{k\ell}(\sigma) := \kappa_{\ell\to k}(\sigma)\,\mu_{k\ell}(\sigma)$$

is strictly decreasing in σ. Therefore the hardest-rival slopes $\mathcal{E}_k(\sigma)$ and the weighted harmonic aggregate

$$\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma) := \left(\sum_{k=1}^{K} \frac{\pi_k}{\mathcal{E}_k(\sigma)}\right)^{-1}$$

are strictly decreasing in σ.

3. ***The weighted speed–accuracy frontier has first-order slope*** $\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma)$. *In particular,*

$$T_\pi(\gamma_\pi;\sigma) = \frac{L_K(\gamma_\pi)}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma)} + o(L_K(\gamma_\pi)), \qquad \gamma_\pi \uparrow 1.$$

4. ***Asymptotic antitone ordering on sufficiently high accuracies.*** *Fix $\sigma_1 < \sigma_2$. The strict slope ordering in Item (2) gives a $\gamma_0 \in (0,1)$ such that for every $\gamma_1 \in (\gamma_0, 1)$,*

$$T_\pi(\gamma_\pi;\sigma_1) < T_\pi(\gamma_\pi;\sigma_2) \qquad \forall \gamma_\pi \in [\gamma_0, \gamma_1].$$

Consequently, for every such γ_1,

$$\mathrm{NITA}^{[\gamma_0,\gamma_1]}_{\mathrm{MSPRT},\pi}(\sigma_1) > \mathrm{NITA}^{[\gamma_0,\gamma_1]}_{\mathrm{MSPRT},\pi}(\sigma_2)$$

when both are evaluated on the common high-accuracy interval $[\gamma_0, \gamma_1]$ with a shared normalization box.

If additionally $\eta_{\mathrm{MS}} \neq 0$, then, because $\mathrm{SINE}_{\mathrm{MS}}(\sigma)$ is strictly increasing in σ, the asymptotic ordering in Item (4) is antitone to SINE along the affine multiclass-SPRT family.

A criterion on a fixed accuracy interval. When the approximation remainders are bounded on a chosen interval, the following comparison makes the required slope gap explicit.

Proposition B.55 (Fixed-interval bounded-remainder criterion for multiclass-SPRT NITA ordering)**.** *Fix an accuracy interval $\mathcal{U} = [\gamma_0, \gamma_1] \subset (0,1)$ contained in the regime where the asymptotic scaling is used. Assume that for each model index σ,*

$$T_\pi(\gamma_\pi;\sigma) = \frac{L_K(\gamma_\pi)}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma)} + r_K(\gamma_\pi;\sigma), \qquad \gamma_\pi \in \mathcal{U}, \tag{149}$$

and that for two models σ_1, σ_2,

$$\sup_{\gamma_\pi \in \mathcal{U}} |r_K(\gamma_\pi;\sigma_1) - r_K(\gamma_\pi;\sigma_2)| \le \Delta_K \tag{150}$$

for some $\Delta_K \ge 0$. If

$$\inf_{\gamma_\pi \in \mathcal{U}} L_K(\gamma_\pi) \left(\frac{1}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma_2)} - \frac{1}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma_1)}\right) > \Delta_K, \tag{151}$$

then

$$T_\pi(\gamma_\pi;\sigma_1) < T_\pi(\gamma_\pi;\sigma_2) \qquad \forall \gamma_\pi \in \mathcal{U},$$

and therefore the interval-restricted NITA from Definition 3.3 is larger for σ_1 than for σ_2 when both are evaluated on the common interval $\mathcal{U}$ with a shared normalization box on the (T_π, γ_π) axes.

B.11.2 Markov SPRT: directional SCGF slopes

Use the conditional affine-score model of Appendix B.9, with $\zeta = \sigma/\alpha$ and the accumulated score driving the two-sided stopping rule. The positive information rates I_{10} and I_{01} are the long-run means of λ^t under P_1 and of $-\lambda^t$ under P_0, respectively. The directional SCGF roots replace the i.i.d. roots in the binary argument (Fuh, 2003; 2004). The next lemma gives sufficient root conditions, followed by the assumed high-accuracy calibration and the resulting SINE–NITA theorem.

Lemma B.56 (Directional SCGF roots for Markov-order-N SPRT)**.** *Use the Markov-order-N affine score model just specified. Define the two directional SCGFs*

$$\Gamma_0(\vartheta) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_0 \exp\left(\vartheta \sum_{t=1}^{n} \lambda^t\right), \tag{152}$$

$$\Gamma_1(v) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_1 \exp\left(-v \sum_{t=1}^{n} \lambda^t\right), \tag{153}$$

$$\Psi(s) := \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E} \exp\left(s \sum_{t=1}^{n} \varepsilon_t\right). \tag{154}$$

Then the two directional positive roots $\vartheta_0^\star(\zeta)$ and $\vartheta_1^\star(\zeta)$ are characterized by

$$\Gamma_0(\vartheta_0^\star(\zeta)) + \Psi(\zeta\vartheta_0^\star(\zeta)) = 0, \qquad \Gamma_1(\vartheta_1^\star(\zeta)) + \Psi(-\zeta\vartheta_1^\star(\zeta)) = 0. \tag{155}$$

Assume that these roots are unique and that

$$\partial_\vartheta\big(\Gamma_0(\vartheta) + \Psi(\zeta\vartheta)\big)\big|_{\vartheta=\vartheta_0^\star(\zeta)} > 0, \qquad \partial_v\big(\Gamma_1(v) + \Psi(-\zeta v)\big)\big|_{v=\vartheta_1^\star(\zeta)} > 0.$$

If $\Psi'(s) > 0$ for $s > 0$ and $\Psi'(s) < 0$ for $s < 0$, then both $\vartheta_0^\star(\zeta)$ and $\vartheta_1^\star(\zeta)$ are strictly decreasing in ζ.

Assumption B.57 (High-accuracy Markov-SPRT regime)**.** Fix a binary class prior $\pi \in (0,1)$ and define

$$\mathrm{Acc}_\pi := \pi\,\mathbb{P}_1(d_\tau = 1) + (1-\pi)\,\mathbb{P}_0(d_\tau = 0), \qquad T_\pi := \pi\,\mathbb{E}_1[\tau] + (1-\pi)\,\mathbb{E}_0[\tau].$$

Use $L(\gamma) = \log\{\gamma/(1-\gamma)\}$ from the binary SPRT calibration. Under the Markov-order-N affine model above, assume:

1. the directional SCGFs
$$\Gamma_0(\vartheta), \qquad \Gamma_1(v), \qquad \Psi(s)$$
exist on neighborhoods of the relevant roots;
2. the positive roots $\vartheta_0^\star(\zeta)$ and $\vartheta_1^\star(\zeta)$ solving
$$\Gamma_0(\vartheta_0^\star(\zeta)) + \Psi(\zeta\vartheta_0^\star(\zeta)) = 0, \qquad \Gamma_1(\vartheta_1^\star(\zeta)) + \Psi(-\zeta\vartheta_1^\star(\zeta)) = 0$$
are unique and strictly decreasing in ζ;
3. along the equal-error high-accuracy calibration, the class-conditional mean hitting times satisfy
$$\mathbb{E}_1[\tau^{(\zeta)}] = \frac{L(\gamma)}{\mathcal{E}_{10}(\zeta)} + o(L(\gamma)), \qquad \mathbb{E}_0[\tau^{(\zeta)}] = \frac{L(\gamma)}{\mathcal{E}_{01}(\zeta)} + o(L(\gamma)),$$
where
$$\mathcal{E}_{10}(\zeta) := I_{10}\vartheta_0^\star(\zeta), \qquad \mathcal{E}_{01}(\zeta) := I_{01}\vartheta_1^\star(\zeta),$$
and, under this equal-error calibration, the ordinary accuracy satisfies $\mathrm{Acc}_\pi = \gamma$;
4. the asymptotic Markov-SPRT NITA comparison is carried out on a common high-accuracy interval together with a shared normalization box across models.

Corollary B.58 (Mean hitting time under a fixed class prior for Markov-SPRT)**.** *Fix a binary class prior $\pi \in (0,1)$. Under Assumption B.57,*

$$T_\pi(\gamma;\zeta) = \frac{L(\gamma)}{\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta)} + o(L(\gamma)), \qquad \mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta) := \frac{1}{\frac{\pi}{\mathcal{E}_{10}(\zeta)} + \frac{1-\pi}{\mathcal{E}_{01}(\zeta)}}, \tag{156}$$

where

$$\mathcal{E}_{10}(\zeta) := I_{10}\vartheta_0^\star(\zeta), \qquad \mathcal{E}_{01}(\zeta) := I_{01}\vartheta_1^\star(\zeta).$$

If $\zeta = 0$ and the exact Markov conditional LLR is used up to positive scale, then

$$\vartheta_0^\star(0) = \vartheta_1^\star(0) = 1$$

and the effective slope reduces to

$$\mathcal{E}_{\text{Markov-SPRT},\pi}(0) = \frac{1}{\frac{\pi}{I_{10}} + \frac{1-\pi}{I_{01}}}. \tag{157}$$

Theorem B.59 (Asymptotic tail antitone relation for Markov-order-N SPRT)**.** *Under Assumption B.57 and the increment-level geometry assumptions of Proposition B.35, the following hold.*

1. ***Projective error increases with noise.*** *The corresponding Markov-order-N SINE limit is strictly increasing in ζ.*
2. ***The effective SPRT slope decreases with noise.*** *Since I_{10} and I_{01} are fixed and the SCGF roots $\vartheta_0^\star(\zeta)$ and $\vartheta_1^\star(\zeta)$ are strictly decreasing, the directional slopes*
$$\mathcal{E}_{10}(\zeta) = I_{10}\vartheta_0^\star(\zeta), \qquad \mathcal{E}_{01}(\zeta) = I_{01}\vartheta_1^\star(\zeta)$$
and their prior-weighted harmonic aggregate
$$\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta) := \frac{1}{\frac{\pi}{\mathcal{E}_{10}(\zeta)} + \frac{1-\pi}{\mathcal{E}_{01}(\zeta)}}$$
are strictly decreasing in ζ.
3. ***The mean hitting time / accuracy frontier has first-order slope*** $\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta)$. *In particular, with*
$$T_\pi(\gamma;\zeta) := \pi\,\mathbb{E}_1[\tau^{(\zeta)}] + (1-\pi)\,\mathbb{E}_0[\tau^{(\zeta)}],$$
we have
$$T_\pi(\gamma;\zeta) = \frac{L(\gamma)}{\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta)} + o(L(\gamma)), \qquad \gamma \uparrow 1.$$
4. ***Asymptotic antitone ordering on sufficiently high accuracies.*** *Fix $\zeta_1 < \zeta_2$. The strict slope ordering in Item (2) gives a $\gamma_0 \in (0,1)$ such that for every $\gamma_1 \in (\gamma_0, 1)$,*
$$T_\pi(\gamma;\zeta_1) < T_\pi(\gamma;\zeta_2) \qquad \forall \gamma \in [\gamma_0, \gamma_1].$$
Consequently, for every such γ_1,
$$\text{NITA}^{[\gamma_0,\gamma_1]}_{\text{Markov-SPRT}}(\zeta_1) > \text{NITA}^{[\gamma_0,\gamma_1]}_{\text{Markov-SPRT}}(\zeta_2)$$
when both are evaluated on the common high-accuracy interval $[\gamma_0, \gamma_1]$ with a shared normalization box on the (T_π, γ) axes.

Because $\text{SINE}_\infty(\zeta)$ is strictly increasing in ζ, the asymptotic ordering in Item (4) is antitone to SINE along the affine Markov-order-N family.

A criterion on a fixed accuracy interval. The same remainder comparison applies to the Markov-SPRT frontier.

Corollary B.60 (Fixed-interval bounded-remainder criterion for Markov-SPRT NITA ordering)**.** *Fix a binary class prior $\pi \in (0,1)$. Assume that for each ζ,*
$$T_\pi(\gamma;\zeta) = \frac{L(\gamma)}{\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta)} + r_{\text{M}}(\gamma;\zeta), \qquad \gamma \in \mathcal{U} \subset (1/2, 1), \tag{158}$$
and that for two models ζ_1, ζ_2,
$$\sup_{\gamma\in\mathcal{U}} |r_{\text{M}}(\gamma;\zeta_1) - r_{\text{M}}(\gamma;\zeta_2)| \le \Delta_{\text{M}}.$$
If
$$\inf_{\gamma\in\mathcal{U}} L(\gamma)\left(\frac{1}{\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta_2)} - \frac{1}{\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta_1)}\right) > \Delta_{\text{M}},$$
then the interval-restricted NITA from Definition 3.3 is larger for ζ_1 than for ζ_2 on the common interval $\mathcal{U}$ with a shared normalization box on the (T_π, γ) axes.

Remark B.61 (Compositional extension: multiclass Markov-order-N SPRT). The two extensions combine when the pairwise Markov models satisfy the corresponding root, high-accuracy, and population-geometry conditions. Indeed, one may replace each i.i.d. root $\kappa_{\ell\to k}(\sigma)$ in Eq. (139) by the corresponding SCGF root under the appropriate Markov conditional LLR process. The hardest-rival reduction and weighted aggregation arguments are unchanged. Combining Lemma B.56 with the pairwise comparison and aggregation in Proposition B.52 yields the corresponding setting-specific slope and SINE–NITA ordering. The composed statement uses the corresponding pairwise Markov regularity, high-accuracy, and population-geometry conditions.

Proofs for both groups are in Appendix C.7.

C PROOFS

C.1 PROOFS RELATED TO SECTION 3

Proof of Theorem B.1. Fix a realized score path and $\alpha > 0$. Multiplication by α preserves signs, ordering, and ties. Thus, if joint score–threshold rescaling multiplies every stopping margin and candidate decision score by α, the first stopping time and the selected action are unchanged. We verify this directly for the stated rules.

For binary SPRT, the margins $\Lambda^{1:t} - a$ and $-b - \Lambda^{1:t}$ become $\alpha(\Lambda^{1:t} - a)$ and $\alpha(-b - \Lambda^{1:t})$. For multiclass SPRT,

$$\min_{\ell \neq k} \alpha \Lambda_{k\ell}^{1:t} - \alpha a_k^{(t)} = \alpha \left(\min_{\ell \neq k} \Lambda_{k\ell}^{1:t} - a_k^{(t)} \right).$$

Thus, the boundary conditions and the maximizing class are preserved at every time, with the same deterministic tie-breaking convention.

For CUSUM, use the one-vs-baseline notation of Appendix A.1. Both recursions start at zero. If $\widetilde{W}_{t-1}^{(k)} = \alpha W_{t-1}^{(k)}$, then

$$\widetilde{W}_t^{(k)} = \max\{0, \alpha W_{t-1}^{(k)} + \alpha \lambda_{k0}^t\} = \alpha W_t^{(k)}.$$

Induction gives this identity for every branch and time. Scaling h to αh therefore preserves the alarm condition and the maximizing branch. In the binary case the terminal action is simply an alarm.

The stopping times consequently agree, including when no boundary is hit. Terminal decisions agree whenever the common stopping time is finite. The final-margin forced decision under the stated finite-horizon SPRT convention is preserved by the same ordering argument. This proves Eq. (30). □

Proof of Corollary B.2. Theorem B.1 preserves the stopping/decision law at corresponding thresholds, thus both Φ_1 and Φ_2. Closure under scaling gives $\mathcal{C}(u) \subseteq \mathcal{C}(\alpha u)$; applying the same argument to α^{-1} gives the reverse inclusion and Eq. (32). □

Proof of Corollary 2.2. If the stated map ψ existed, Proposition 2.1 would give $S(\alpha \mathbf{\Lambda}^\star) = S(\mathbf{\Lambda}^\star)$, and therefore $E_{\text{exact}}(\alpha \mathbf{\Lambda}^\star, \mathbf{\Lambda}^\star) = E_{\text{exact}}(\mathbf{\Lambda}^\star, \mathbf{\Lambda}^\star)$ for every $\alpha > 0$. This contradicts Eq. (6). □

Proof of Proposition B.5. For every batch element, class, and time, $m_{b,k}^{(t)}(\alpha \mathbf{\Lambda}; \alpha a) = \alpha m_{b,k}^{(t)}(\mathbf{\Lambda}; a)$, since $\alpha > 0$ commutes with the minimum. Thus every stopping condition and every maximizing class, including the forced decision at T, is unchanged under the stated deterministic tie-breaking convention. □

Proof of Corollary B.6. The threshold template obeys $ca_k^{(t)}(h) = a_k^{(t)}(ch)$. Proposition B.5 therefore gives the pathwise identity Eq. (39); taking laws preserves every stopping/decision-based operating characteristic. □

Proof of Lemma B.9. Use $W = \mathsf{R}^\top \mathsf{R}$ in $\langle \mathsf{R}\widehat{\boldsymbol{\ell}}, \mathsf{R}\boldsymbol{\ell} \rangle = \widehat{\boldsymbol{\ell}}^\top W \boldsymbol{\ell}$ and the corresponding quadratic forms for the two squared norms. Substitution in the clipped cosine definition gives the weighted-angle expression. □

Proof of Lemma B.10. By Cauchy–Schwarz, $\langle f(V_i), g(V_i) \rangle_{\mathcal{V}_\mathsf{R}}$, $\|f(V_i)\|_{\mathcal{V}_\mathsf{R}}^2$, and $\|g(V_i)\|_{\mathcal{V}_\mathsf{R}}^2$ are integrable. The strong law of large numbers gives

$$\langle f, g \rangle_n \to \langle f, g \rangle_{\mathcal{H}_\mathsf{R}}, \qquad \|f\|_n^2 \to \|f\|_{\mathcal{H}_\mathsf{R}}^2, \qquad \|g\|_n^2 \to \|g\|_{\mathcal{H}_\mathsf{R}}^2$$

almost surely. Since the population norms are nonzero, the empirical norms are nonzero for all sufficiently large n almost surely. The truncated cosine map

$$(a, b, c) \mapsto \max\{0, a/(\sqrt{b}\sqrt{c})\}$$

is continuous on $b, c > 0$, and $r \mapsto \sqrt{1 - r^2}$ is continuous on $[0, 1]$. Therefore the empirical SINE converges almost surely to the population Hilbert-space SINE. A finite collection follows by applying the same argument to each pair and intersecting the probability-one events. The stationary ergodic variant follows from Birkhoff's ergodic theorem. □

Proof of item (1) of Theorem 3.6. In $\mathcal{H}_\Sigma$, orthogonality gives

$$\|\widehat{\mathbf{\Lambda}}_{\sigma,T} - \beta\mathbf{\Lambda}_T\|^2 = (\alpha-\beta)^2\|\mathbf{\Lambda}_T\|^2 + \sigma^2\|\boldsymbol{\eta}_{\Sigma,T}\|^2, \quad \|\widehat{\mathbf{\Lambda}}_{\sigma,T}\|^2 = \alpha^2\|\mathbf{\Lambda}_T\|^2 + \sigma^2\|\boldsymbol{\eta}_{\Sigma,T}\|^2.$$

The unique minimizer over $\beta \geq 0$ is $\beta = \alpha > 0$. Dividing its residual by the second expression proves Eq. (13). The resulting ratio increases with σ strictly when $\boldsymbol{\eta}_{\Sigma,T} \neq 0$. □

Proof of Lemma B.12. Independence under P_0 gives $\mathbb{E}_0 e^{\kappa g_\sigma} = M_\lambda(\alpha\kappa)M_\xi(\sigma\kappa)$, proving Eq. (42). Strict Jensen and convexity, with $M_\xi(0) = 1$ and $M'_\xi(0) = 0$, make $M_\xi(v) > 1$ and strictly increasing on its positive finite domain. For $\sigma_2 > \sigma_1$, the product at $\kappa(\sigma_1)$ is therefore strictly greater than one under σ_2. Its derivative at zero is $\alpha\mathbb{E}_0[\lambda] = -\alpha D_{\mathrm{KL}}(P_0\|P_1) < 0$. Convexity and uniqueness of the positive root imply $\kappa(\sigma_2) < \kappa(\sigma_1)$. Finally, $\mathbb{E}_1[\xi] = 0$ gives the constant positive drift $\mu_1 = \alpha\mathbb{E}_1[\lambda]$, thus $\mathcal{E} = \kappa\mu_1$ is strictly decreasing. □

Proof of Lemma B.11. For $\mathcal{E}_2 > \mathcal{E}_1 > 0$, the capped integrand $\min\{A_0 e^{\mathcal{E}x}, Y^*\}$ is pointwise nondecreasing in $\mathcal{E}$. If $A_0 < Y^*$, it is strictly larger for $\mathcal{E}_2$ on the positive-length interval $0 < x < \min\{X^*, \log(Y^*/A_0)/\mathcal{E}_2\}$. Integration gives strict increase. If $A_0 = Y^*$, the integrand is identically Y^* and $\Phi_{A_0} \equiv 1$. □

Proof of item (3) of Theorem 3.6. Let $\sigma_1 < \sigma_2$. By item (2),

$$\Delta := \mathcal{E}(\sigma_1) - \mathcal{E}(\sigma_2) > 0.$$

By Eq. (12), there exist $B < \infty$ and $x_1 < \infty$ such that

$$\log y^\star_{\sigma_1}(x) - \log y^\star_{\sigma_2}(x) \geq \Delta x - B \qquad \forall x \geq x_1.$$

Choosing $x_0 > \max\{x_1, B/\Delta\}$ yields Eq. (14). Integrating over any common high-threshold comparison box $[x_0, X^*] \times [0, Y^*]$ with $X^* > x_0$ gives Eq. (15). Combining this with item (1) yields Eq. (16) whenever $\boldsymbol{\eta}_{\Sigma,T} \neq 0$. □

Representation-general geometric form. The geometric calculation is valid in the population space $\mathcal{H}_{\mathsf{R}}$ of Appendix A.2 whenever a fixed map gives $\hat{u}_{\mathsf{R},\sigma} = \alpha u_{\mathsf{R}} + \sigma\eta_{\mathsf{R}}$, $u_{\mathsf{R}} \neq 0$, and $\langle \eta_{\mathsf{R}}, u_{\mathsf{R}} \rangle_{\mathcal{H}_{\mathsf{R}}} = 0$. Taking $\mathsf{R} = \mathsf{R}_I$ on raw increments gives increment geometry; $\mathsf{R} = \mathsf{R}_\Sigma$ gives the cumulative-trajectory geometry above. The same algebra applies to finite vectors in $\mathcal{V}_{\mathsf{R}}$ when orthogonality holds in that space. Note that population orthogonality alone does not supply this finite-stack condition.

Proof of Proposition B.22. Work in the fixed population space $\mathcal{H}_{\mathsf{R}}$ specified in Appendix B.7. Orthogonality gives

$$\|\hat{u}_{\Xi,\sigma} - \beta u_\Xi\|^2_{\mathcal{H}_{\mathsf{R}}} = (\alpha-\beta)^2\|u_\Xi\|^2_{\mathcal{H}_{\mathsf{R}}} + \sigma^2\|\eta_\Xi\|^2_{\mathcal{H}_{\mathsf{R}}}.$$

Because $u_\Xi \neq 0$ and $\alpha > 0$, the unique minimizer over $\beta \geq 0$ is $\beta = \alpha$. Also $\|\hat{u}_{\Xi,\sigma}\|^2_{\mathcal{H}_{\mathsf{R}}} = \alpha^2\|u_\Xi\|^2_{\mathcal{H}_{\mathsf{R}}} + \sigma^2\|\eta_\Xi\|^2_{\mathcal{H}_{\mathsf{R}}}$. This proves Eq. (63) and its monotonicity. Replacing $\mathcal{H}_{\mathsf{R}}$ by $\mathcal{V}_{\mathsf{R}}$ throughout proves the finite-stack version under its stated orthogonality condition.

We next prove the deterministic frontier-to-area step. Fix $\sigma_1 < \sigma_2$. Since $\mathcal{E}_\Xi$ is strictly decreasing, $\Delta := \mathcal{E}_\Xi(\sigma_1) - \mathcal{E}_\Xi(\sigma_2) > 0$. In the QCD-type case, Eq. (61) gives

$$\log F^\star_{\sigma_1}(x) - \log F^\star_{\sigma_2}(x) = \Delta x + o(x),$$

which is positive for all sufficiently large x. Therefore there is a high-threshold interval $[x_0, X^*]$ on which the untruncated frontier for σ_1 pointwise dominates that for σ_2. With a common non-saturated vertical normalizer, integrating the two upper envelopes over this same interval yields $\mathfrak{N}_\Xi(\sigma_1; \mathcal{R}_{\sigma_1,\sigma_2}) > \mathfrak{N}_\Xi(\sigma_2; \mathcal{R}_{\sigma_1,\sigma_2})$.

In the SPRT-type case, set $\mathcal{E}_i := \mathcal{E}_\Xi(\sigma_i)$. Since $\mathcal{E}_1 > \mathcal{E}_2$,

$$T_{\sigma_2}(\gamma) - T_{\sigma_1}(\gamma) = L_\Xi(\gamma)\Big(\frac{1}{\mathcal{E}_2} - \frac{1}{\mathcal{E}_1}\Big) + o(L_\Xi(\gamma)),$$

which is positive for all sufficiently high accuracies γ. Thus the σ_1 frontier reaches every accuracy in a high-accuracy interval no later than the σ_2 frontier. Equivalently, after inversion to the upper accuracy envelope on the common mean-time axis, the σ_1 envelope pointwise dominates the σ_2 envelope on a common high-accuracy comparison region. Integrating over that shared, non-saturated NITA box again gives Eq. (64).

If $\eta_\Xi \neq 0$, the SINE formula is strictly increasing in σ. Combining this strict projective ordering with the NITA ordering just proved gives the equivalence in Eq. (65). □

C.2 Proofs for the local effective-slope (Corollary 3.7)

Proof of Proposition B.19. With $C = \log(F^*/A_0) > 0$ and $m(\mathcal{E}) = \min\{D^*, C/\mathcal{E}\}$, split at the cap:

$$\Phi(\mathcal{E}) = \frac{1}{D^* F^*} \left(\int_0^{m(\mathcal{E})} A_0 e^{\mathcal{E}x}\, dx + \int_{m(\mathcal{E})}^{D^*} F^*\, dx \right), \tag{159}$$

Differentiation on each side of $\mathcal{E}D^* = C$ gives

$$\Phi'(\mathcal{E}) = \frac{A_0}{D^* F^*} \int_0^{m(\mathcal{E})} x e^{\mathcal{E}x}\, dx.$$

For an interior cap the moving-boundary terms cancel because $A_0 e^{\mathcal{E} m(\mathcal{E})} = F^*$; the same formula gives matching one-sided derivatives at the endpoint cap. Thus Φ is C^1 on $(0, \infty)$. Evaluating the integral proves Eq. (53). Since $m(I_{\mathrm{KL}}) > 0$, the integrand is positive on a set of positive measure, giving Eq. (52) and its strict positivity. □

Proof of Theorem B.18. *Projective geometry.* Apply item (1) of Theorem 3.6 with $p = \|\mathbf{\Lambda}_T\|^2_{\mathcal{H}_\Sigma}$ and $q = \|\boldsymbol{\eta}_{\Sigma,T}\|^2_{\mathcal{H}_\Sigma}$:

$$s_\sigma = \mathrm{SINE}_\Sigma(\sigma)^2 = \frac{\sigma^2 q}{\alpha^2 p + \sigma^2 q}. \tag{160}$$

Solving this identity gives

$$\sigma^2 = \frac{\alpha^2 p}{q} \frac{s_\sigma}{1 - s_\sigma} = \frac{\alpha^2 p}{q} s_\sigma + o(s_\sigma) \qquad (\sigma \downarrow 0). \tag{161}$$

Root expansion. Set $f(\kappa, \sigma) = \log M_\lambda(\alpha\kappa) + \log M_\xi(\sigma\kappa)$, using the MGFs from the factorized root lemma. At $(1/\alpha, 0)$, the likelihood-ratio identity, change of measure, and centered nuisance give

$$f = 0, \qquad f_\kappa = \alpha \mathbb{E}_0[\lambda e^\lambda] = \alpha I_{\mathrm{KL}} > 0, \qquad f_\sigma = 0, \qquad f_{\sigma\sigma} = \mathrm{Var}_Q(\xi)/\alpha^2.$$

The assumed C^2 regularity permits the implicit function theorem. Its root branch remains positive near zero and agrees with the unique positive Cramér root for $\sigma \geq 0$. Differentiating $f(\kappa(\sigma), \sigma) = 0$ gives $\kappa'(0) = 0$ and $\kappa''(0) = -f_{\sigma\sigma}/f_\kappa = -\mathrm{Var}_Q(\xi)/(\alpha^3 I_{\mathrm{KL}})$. Thus

$$\kappa(\sigma) = \frac{1}{\alpha} - \frac{\mathrm{Var}_Q(\xi)}{2\alpha^3 I_{\mathrm{KL}}} \sigma^2 + o(\sigma^2). \tag{162}$$

Effective slope. Since $\mu_1(\sigma) = \alpha I_{\mathrm{KL}}$, multiplying the root expansion by this drift yields

$$I_{\mathrm{KL}} - \mathcal{E}(\sigma) = \frac{\mathrm{Var}_Q(\xi)}{2\alpha^2} \sigma^2 + o(\sigma^2). \tag{163}$$

Substitution of Eq. (161) gives

$$I_{\mathrm{KL}} - \mathcal{E}(\sigma) = \frac{p \mathrm{Var}_Q(\xi)}{2q} s_\sigma + o(s_\sigma).$$

The coefficient is strictly positive and finite by $p, q > 0$, the smooth MGF, and nondegeneracy of ξ. Thus the stated two-sided bounds follow by bounding the convergent ratio between one-half and three-halves of its positive limit, and Eq. (46) follows.

Ordered frontier limits. For each fixed small $\sigma > 0$, Eq. (12) gives

$$\frac{\log y_0^\star(d) - \log y_\sigma^\star(d)}{d s_\sigma} = \frac{I_{\mathrm{KL}} - \mathcal{E}(\sigma)}{s_\sigma} + \frac{b_0(d) - b_\sigma(d)}{d s_\sigma}.$$

The last term tends to zero as $d \to \infty$, with $s_\sigma > 0$ fixed. Then letting $\sigma \downarrow 0$ gives the asserted coefficient. No uniformity in σ or exchange of limits is used. □

Proof of the idealized local NITA calculation. By Proposition B.19, Φ is C^1 at I_{KL} with positive derivative. Combining its first-order Taylor expansion with Eq. (46) gives

$$\begin{aligned}\widetilde{\mathrm{NITA}}(0) - \widetilde{\mathrm{NITA}}(\sigma) &= \Phi'(I_{\mathrm{KL}})\{I_{\mathrm{KL}} - \mathcal{E}(\sigma)\} + o(I_{\mathrm{KL}} - \mathcal{E}(\sigma)) \\ &= c_N s_\sigma + o(s_\sigma), \qquad c_N = \Phi'(I_{\mathrm{KL}})c_{\mathcal{E}} > 0.\end{aligned}$$

This proves Eq. (54). For small enough $\sigma > 0$, the ratio to s_σ lies between $c_N/2$ and $3c_N/2$; choosing these as c_-, c_+ proves Eq. (56). □

For comparison with actual NITA on the same fixed box, write $R(\sigma) := \mathrm{NITA}(\sigma) - \widetilde{\mathrm{NITA}}(\sigma)$. Then

$$\mathrm{NITA}(0) - \mathrm{NITA}(\sigma) = c_N s_\sigma + o(s_\sigma) + R(0) - R(\sigma).$$

Thus, the same coefficient transfers if $R(\sigma) - R(0) = o(s_\sigma)$, a condition not supplied by the stated large-threshold asymptotics.

C.3 Proofs for the finite-sample guarantee (Theorem 3.8)

We condition on the training information $\mathscr{T}$ as in Appendix A.3.

Proof of the sample-mean part of Lemma B.20. For any fixed pair (m, j), the estimator

$$\widehat{x}_{m,j} = \frac{1}{n_x}\sum_{r=1}^{n_x} \phi^x_{m,j}(U_r)$$

is the average of independent random variables bounded in $[0, X^*]$. Therefore Hoeffding's inequality (Hoeffding, 1963) gives

$$\Pr(|\widehat{x}_{m,j} - x_{m,j}| > \varepsilon_x) \le 2\exp\left(-\frac{2n_x\varepsilon_x^2}{(X^*)^2}\right).$$

Taking a union bound over all $N_{\mathrm{ckpt}}N_\theta$ model-threshold pairs yields

$$\Pr\left(\max_{m\in\mathcal{M},\, j\in[N_\theta]} |\widehat{x}_{m,j} - x_{m,j}| > \varepsilon_x\right) \le 2N_{\mathrm{ckpt}}N_\theta \exp\left(-\frac{2n_x\varepsilon_x^2}{(X^*)^2}\right).$$

The proof of Eq. (58) is identical, replacing X^*, n_x, ϕ^x by Y^*, n_y, ϕ^y. □

Proof of the KM part of Lemma B.20. Fix a coordinate and checkpoint–threshold pair and use the notation of Appendix A.3, suppressing these indices. Set $S^{\mathrm{event}}(-1) = \widehat{S}^{\mathrm{KM}}(-1) := 1$. Define

$$\begin{aligned} G_t^{\mathrm{obs}} &:= \Pr(\mathsf{C} \ge t), \quad r_t^{\mathrm{obs}} := \Pr(\mathsf{U} \ge t), \qquad d_t^{\mathrm{obs}} := \Pr(\mathsf{U} = t, \mathsf{D} = 1), \\ \widehat{r}_t^{\mathrm{obs}} &:= N^{\mathrm{risk}}(t)/n, \quad \widehat{d}_t^{\mathrm{obs}} := N^{\mathrm{event}}(t)/n, \quad \widehat{p}_t^{\mathrm{KM}} := \widehat{d}_t^{\mathrm{obs}}/\widehat{r}_t^{\mathrm{obs}}, \end{aligned}$$

using $0/0 := 0$. Independent censoring gives

$$r_t^{\mathrm{obs}} = S^{\mathrm{event}}(t-1)G_t^{\mathrm{obs}}, \qquad d_t^{\mathrm{obs}} = \Pr(\mathsf{T} = t)G_t^{\mathrm{obs}}, \qquad G_t^{\mathrm{obs}} \ge \gamma \quad (0 \le t < H).$$

Let $F_1^{\mathrm{obs}}(t) := \Pr(\mathsf{U} \le t, \mathsf{D} = 1)$ and let $\widehat{F}_1^{\mathrm{obs}}$ be its empirical version. Set

$$a_n^{\mathrm{obs}} := \sup_{0\le t<H} |\widehat{F}_1^{\mathrm{obs}}(t) - F_1^{\mathrm{obs}}(t)|, \qquad b_n^{\mathrm{obs}} := \sup_{0\le t<H} |\widehat{r}_t^{\mathrm{obs}} - r_t^{\mathrm{obs}}|.$$

The Dvoretzky–Kiefer–Wolfowitz–Massart inequality gives probability at most $2e^{-2nz^2}$ that an empirical cumulative distribution function has maximum error exceeding $z > 0$ (Massart, 1990). Apply it first to U. For the observed-event function, send censored observations to the sentinel H; its cumulative distribution below H is F_1^{obs}. The two applications and a union bound give

$$\Pr(a_n^{\mathrm{obs}} > z \text{ or } b_n^{\mathrm{obs}} > z) \le 4e^{-2nz^2}.$$

This inequality holds for discrete distributions; no continuity or no-ties assumption is needed.

It remains to bound the survival error using these two deviations. Set $e_t^{\mathrm{KM}} := \widehat{S}^{\mathrm{KM}}(t) - S^{\mathrm{event}}(t)$. The KM recursion and $\widehat{r}_t^{\mathrm{obs}}\widehat{p}_t^{\mathrm{KM}} = \widehat{d}_t^{\mathrm{obs}}$ yield the exact identity

$$e_t^{\mathrm{KM}} = (1 - \widehat{p}_t^{\mathrm{KM}})e_{t-1}^{\mathrm{KM}} + \frac{d_t^{\mathrm{obs}} - \widehat{d}_t^{\mathrm{obs}} + (\widehat{r}_t^{\mathrm{obs}} - r_t^{\mathrm{obs}})\widehat{p}_t^{\mathrm{KM}}}{G_t^{\mathrm{obs}}}.$$

For fixed t, define

$$w_{u,t}^{\mathrm{KM}} := \frac{\prod_{v=u+1}^{t}(1 - \widehat{p}_v^{\mathrm{KM}})}{G_u^{\mathrm{obs}}}, \qquad 0 \le u \le t.$$

These weights are nonnegative and nondecreasing in u: the denominator is nonincreasing, while the product loses factors in $[0, 1]$. Also $w_{u,t}^{\mathrm{KM}} \le 1/G_t^{\mathrm{obs}}$. Iterating the recursion from $e_{-1}^{\mathrm{KM}} = 0$ gives

$$e_t^{\mathrm{KM}} = \sum_{u=0}^{t} w_{u,t}^{\mathrm{KM}}(d_u^{\mathrm{obs}} - \widehat{d}_u^{\mathrm{obs}}) + \sum_{u=0}^{t} w_{u,t}^{\mathrm{KM}}(\widehat{r}_u^{\mathrm{obs}} - r_u^{\mathrm{obs}})\widehat{p}_u^{\mathrm{KM}}.$$

The cumulative sums of $d_u^{\mathrm{obs}} - \widehat{d}_u^{\mathrm{obs}}$ have absolute value at most a_n^{obs}. Summation by parts and monotonicity of the weights therefore bound the first sum by $2a_n^{\mathrm{obs}}/G_t^{\mathrm{obs}}$. For the second, the identity

$$\sum_{u=0}^{t} \widehat{p}_u^{\mathrm{KM}} \prod_{v=u+1}^{t} (1 - \widehat{p}_v^{\mathrm{KM}}) = 1 - \widehat{S}^{\mathrm{KM}}(t) \le 1$$

gives the bound $b_n^{\mathrm{obs}}/G_t^{\mathrm{obs}}$. Therefore, deterministically,

$$\sup_{0 \le t < H} G_t^{\mathrm{obs}} |\widehat{S}^{\mathrm{KM}}(t) - S^{\mathrm{event}}(t)| \le 2a_n^{\mathrm{obs}} + b_n^{\mathrm{obs}}.$$

Take $z = \sqrt{\log(4/\eta_{\mathrm{KM}})/(2n)}$ in the distribution-function bound and use $G_t^{\mathrm{obs}} \ge \gamma$ to obtain the survival bound. Summing its errors over $t = 0, \dots, H-1$ proves the restricted-mean bound. □

Proof of Lemma B.21. Fix a model $m \in \mathcal{M}$ and a point $x \in [0, X^*]$. Suppose that $\widehat{x}_{m,j} \le x$. Then

$$x_{m,j} \le x + \varepsilon_x.$$

Also, by the assumed bound on the y-coordinates,

$$\widehat{y}_{m,j} \le y_{m,j} + \varepsilon_y.$$

Thus

$$\begin{aligned} \widehat{F}_m(x) &= \max\{\widehat{y}_{m,j} : \widehat{x}_{m,j} \le x,\ j \in [N_\theta]\} \\ &\le \max\{y_{m,j} + \varepsilon_y : x_{m,j} \le x + \varepsilon_x,\ j \in [N_\theta]\} \\ &\le F_m(x + \varepsilon_x) + \varepsilon_y. \end{aligned} \tag{164}$$

Conversely, if $x_{m,j} \le x$, then

$$\widehat{x}_{m,j} \le x + \varepsilon_x, \qquad y_{m,j} \le \widehat{y}_{m,j} + \varepsilon_y,$$

thus

$$F_m(x) \le \widehat{F}_m(x + \varepsilon_x) + \varepsilon_y. \tag{165}$$

Integrating Eq. (164) over $x \in [0, X^*]$ gives

$$\begin{aligned} \widehat{\mathrm{NITA}}_m &= \frac{1}{X^*Y^*} \int_0^{X^*} \widehat{F}_m(x)\, dx \\ &\le \frac{1}{X^*Y^*} \int_0^{X^*} \big(F_m(x + \varepsilon_x) + \varepsilon_y\big)\, dx \\ &= \frac{1}{X^*Y^*} \left(\int_{\varepsilon_x}^{X^* + \varepsilon_x} F_m(u)\, du + X^*\varepsilon_y \right). \end{aligned}$$

Because $0 \le F_m(u) \le Y^*$ for all u,

$$\int_{\varepsilon_x}^{X^*+\varepsilon_x} F_m(u)\,du \le \int_0^{X^*} F_m(u)\,du + \varepsilon_x Y^*.$$

Thus

$$\widehat{\mathrm{NITA}}_m \le \mathrm{NITA}_m + \frac{\varepsilon_x}{X^*} + \frac{\varepsilon_y}{Y^*}.$$

Similarly, integrating Eq. (165) yields

$$\mathrm{NITA}_m \le \widehat{\mathrm{NITA}}_m + \frac{\varepsilon_x}{X^*} + \frac{\varepsilon_y}{Y^*}.$$

Combining the two inequalities gives

$$|\widehat{\mathrm{NITA}}_m - \mathrm{NITA}_m| \le \frac{\varepsilon_x}{X^*} + \frac{\varepsilon_y}{Y^*}.$$

Taking the supremum over $m \in \mathcal{M}$ proves Eq. (59). □

Proof of Theorem 3.8. Condition on $\mathscr{T}$ as in Appendix A.3. For each coordinate and checkpoint–threshold pair allocate failure probability $\delta_{\mathrm{conc}}/(2N_{\mathrm{ckpt}}N_\theta)$. Hoeffding's inequality for a sample mean gives the first radius in Eq. (19) after dividing the coordinate error by B_ι. For KM, use the KM part of Lemma B.20 with $\eta_{\mathrm{KM}} = \delta_{\mathrm{conc}}/(2N_{\mathrm{ckpt}}N_\theta)$. Summing the survival error to H_ι and dividing by B_ι gives the second radius. Clipping both the estimate and its target to $[0, B_\iota]$ cannot increase their difference. A union bound therefore gives, with conditional probability at least $1 - \delta_{\mathrm{conc}}$,

$$\max_{m,j} |\widehat{x}_{m,j} - x_{m,j}| \le X^* \min\{1, \omega_x\}, \qquad \max_{m,j} |\widehat{y}_{m,j} - y_{m,j}| \le Y^* \min\{1, \omega_y\}.$$

Lemma B.21 then gives

$$\sup_m |\widehat{\mathrm{NITA}}_m - \mathrm{NITA}_m| \le \min\{1, \omega_x\} + \min\{1, \omega_y\} \le \varepsilon_{\mathrm{conc}},$$

which proves Eq. (18). No unbiasedness of KM, or independence across checkpoint–threshold pairs, has been used.

To prove the finite-sample guarantee, write

$$\mathrm{NITA}_{m^\star} - \mathrm{NITA}_{\widehat{m}} = \big(\mathrm{NITA}_{m^\star} - \widehat{\mathrm{NITA}}_{m^\star}\big) + \big(\widehat{\mathrm{NITA}}_{m^\star} - \widehat{\mathrm{NITA}}_{\widehat{m}}\big) + \big(\widehat{\mathrm{NITA}}_{\widehat{m}} - \mathrm{NITA}_{\widehat{m}}\big).$$

Since $\widehat{m}$ maximizes $\widehat{\mathrm{NITA}}_m$,

$$\widehat{\mathrm{NITA}}_{m^\star} - \widehat{\mathrm{NITA}}_{\widehat{m}} \le 0.$$

Thus

$$\mathrm{NITA}_{m^\star} - \mathrm{NITA}_{\widehat{m}} \le |\mathrm{NITA}_{m^\star} - \widehat{\mathrm{NITA}}_{m^\star}| + |\widehat{\mathrm{NITA}}_{\widehat{m}} - \mathrm{NITA}_{\widehat{m}}| \le 2 \sup_{m \in \mathcal{M}} |\widehat{\mathrm{NITA}}_m - \mathrm{NITA}_m|.$$

Combining this with Eq. (18) proves Eq. (20). □

Fixing the training output. The proof first fixes the training information $\mathscr{T}$ defined in Appendix A.3. This includes the training data and randomness, the resulting checkpoints and score scales, and the training-only choices of validation horizons, threshold grid, and common comparison region. Probabilities then refer to the independent validation data. For almost every training outcome, both conclusions of Theorem 3.8 hold simultaneously with conditional probability at least $1 - \delta_{\mathrm{conc}}$. Averaging over training outcomes gives the same probability bound when both training and validation are random. The population target is grid-NITA of the checkpoint–scale pairs actually produced by training. Because the scales are part of this target, no separate scale-estimation error term is needed.

Sample means as the complete-follow-up case. For an event-time coordinate $\iota \in \{x, y\}$, suppose the protocol guarantees follow-up through H_ι for every validation observation. For each fixed checkpoint–threshold pair, the KM risk-set ratios telescope to give

$$\widehat{\mu}^{\mathrm{KM}}_{\iota,m,j} = \frac{1}{n_\iota} \sum_{i=1}^{n_\iota} \min\{\mathsf{T}_i, H_\iota\}.$$

Thus the ordinary average of capped event times is exactly the complete-follow-up special case of the KM estimator. When $H_\iota \le B_\iota$, the summands satisfy the sample-mean range condition, thus Hoeffding's inequality gives the sharper sample-mean bound even if a KM routine computes the estimate. When both operating coordinates satisfy the sample-mean conditions, the error bound in Eq. (19) is

$$\varepsilon_{\mathrm{conc}} = \sqrt{\frac{\log(4N_{\mathrm{ckpt}}N_\theta/\delta_{\mathrm{conc}})}{2}}\left(\frac{1}{\sqrt{n_x}}+\frac{1}{\sqrt{n_y}}\right).$$

This sharper bound uses the sample-mean representation rather than direct substitution into the general KM bound.

Restricted and uncapped means. The KM construction estimates the restricted mean $\mathbb{E}[\min\{\mathsf{T}, H_\iota\}]$, followed by the specified operating-coordinate clipping. The theorem bounds estimation error for these coordinates. Extending the guarantee to the uncapped mean $\mathbb{E}[\mathsf{T}]$ would additionally require a bound on the contribution of waiting beyond H_ι, which the present result does not provide.

Fixed comparison regions. The theorem uses a fixed threshold grid and common operating box. Integrating the same envelope over a horizontal subinterval $[x_0, X^*]$ fixed before validation gives the same argument, with the horizontal error divided by $X^* - x_0$ instead of X^*. The same area argument applies to a threshold subset fixed conditional on $\mathscr{T}$. The high-performance region in Appendix D.12.1 is determined using training data only and frozen before validation. This differs from selecting or discarding points according to their validation-estimated performance, for which coordinate concentration alone does not establish stability.

C.4 Proofs for the all-rivals matrix-QCD extension

Proof of Lemma B.26. Set $u = (d_1^{-1}, \dots, d_K^{-1})$ and write the vector of classwise delays as $hu + r(h)$, where $r(h) = o(h)$ coordinatewise and therefore in norm, since $K < \infty$. Positive homogeneity and local Lipschitz continuity give

$$\psi(hu + r(h)) = h\psi(u + r(h)/h) = h\psi(u) + o(h).$$

This proves Eq. (71). If the component remainders happen to be $O(1)$, the same argument gives an $O(1)$ aggregate remainder, but that stronger condition is not needed below. □

Proof of Lemma B.24. Fix $b \neq k$. Under P_b, let

$$M_{b\to k}(u) := \mathbb{E}_b[e^{u\lambda_{kb}}], \qquad N_{kb}^{(b)}(v) := \mathbb{E}_b[e^{v\varepsilon_{kb}}].$$

Independence within this ordered pair gives

$$\mathbb{E}_b[e^{\kappa g_{kb}^\sigma}] = M_{b\to k}(\alpha\kappa)N_{kb}^{(b)}(\sigma\kappa).$$

The centered nondegenerate nuisance has a strictly increasing MGF on its positive domain, with $N_{kb}^{(b)}(v) > 1$ for $v > 0$. Also

$$\mathbb{E}_b[g_{kb}^\sigma] = -\alpha\mathrm{KL}(P_b\|P_k) < 0.$$

The exact-score root is $1/\alpha$. For positive σ the MGF at $1/\alpha$ exceeds one, thus $0 < \kappa_{b\to k}(\sigma) < 1/\alpha$. The stipulated MGF domain therefore permits comparison of the two noise levels at the smaller-noise root. If $\sigma_2 > \sigma_1$, the larger-noise MGF is strictly greater than one at $\kappa_{b\to k}(\sigma_1)$. Its derivative at zero is negative, and uniqueness of its positive root implies $\kappa_{b\to k}(\sigma_2) < \kappa_{b\to k}(\sigma_1)$. This argument is marginal and requires no independence between different matrix entries. A finite minimum is strictly decreasing by evaluating at a minimizer at σ_1. A finite maximum is strictly decreasing by evaluating at a maximizer at σ_2. This proves all stated extrema claims. □

Proof of Theorem B.27. We derive the first-cycle duration and error exponent, pass to repeated declarations using the restart policy, and then compare effective slopes. Fix σ and suppress it until the slope-ordering step.

Step 1: positive-drift row completion and first-cycle duration. Within a cycle let

$$S_{ij}(n) = \sum_{t=1}^{n} g_{ij,t}, \qquad W_{ij,n} = S_{ij}(n) - \min_{0 \le t \le n} S_{ij}(t).$$

For i.i.d. increments with finite first moment, the strong law gives $W_{ij,n}/n \to \max\{\mathbb{E}[g_{ij}], 0\}$ almost surely. For positive mean the running minimum is eventually constant. For negative mean the running minimum divided by n converges to the same negative mean as $S_{ij}(n)/n$. These statements also follow by applying the uniform form of the strong law to partial sums on $\{0, \dots, n\}$. Under P_k, every correct-row drift is

$$\mu_{k\ell} = \mathbb{E}_k g_{k\ell} = \alpha I_{k\ell} > 0, \quad \ell \neq k, \qquad C_{k,n}/n \longrightarrow d_k := \min_{\ell \neq k} \mu_{k\ell}.$$

For a wrong row $j \neq k$, $C_{j,n} \le W_{jk,n}$ and $\mathbb{E}_k g_{jk} = -\mu_{kj} < 0$ by antisymmetry. Thus, for every fixed $B < \infty$,

$$\frac{1}{h} \max_{1 \le n \le Bh} C_{j,n} \longrightarrow 0 \quad \text{almost surely.}$$

The true row reaches h at time $h/d_k + o(h)$ almost surely, while no wrong row can reach h by any smaller fixed multiple of h for all sufficiently large h on the same sample path. Consequently

$$\tau_h/h \longrightarrow d_k^{-1} \quad \text{almost surely under } P_k, \qquad \mathbb{P}_k(J_h = k) \longrightarrow 1.$$

This argument concerns simultaneous row completion. It does not replace it by the maximum of separate first pairwise hitting times, which does not necessarily coincide because of recrossings.

For convergence of expectations, choose $\theta_{k\ell} > 0$ small enough that $a_{k\ell} := -\log \mathbb{E}_k e^{-\theta_{k\ell} g_{k\ell}} > 0$. The moment assumptions and positive drifts permit this choice. Since $W_{k\ell,n} \ge S_{k\ell}(n)$,

$$\mathbb{P}_k(\tau_h > n) \le \mathbb{P}_k(C_{k,n} < h) \le \sum_{\ell \neq k} \exp\{\theta_{k\ell} h - a_{k\ell} n\}.$$

For $n \ge Bh$ with sufficiently large fixed B, this is an exponentially decaying tail in n. It gives uniform integrability of τ_h/h as $h \to \infty$, and thus

$$m_k(h) = \mathbb{E}_k \tau_h = h/d_k + o(h).$$

It also supplies all tail estimates used below.

Step 2: a universal upper bound on first-cycle error probability. Under P_b, a false declaration of $k \neq b$ by time n requires a k-versus-b reflected crossing by that time. For any starting time s,

$$\exp\left\{ \kappa_{b\to k} \sum_{t=s}^{s+r-1} g_{kb,t} \right\}, \qquad r = 0, 1, \dots,$$

is a nonnegative unit-mean martingale. The maximal inequality bounds its probability of ever reaching $e^{\kappa_{b\to k} h}$ by $e^{-\kappa_{b\to k} h}$. A union bound over the at most n starts in $\{1, \dots, n\}$ gives

$$\mathbb{P}_b\left(\max_{1 \le t \le n} W_{kb,t} \ge h \right) \le n e^{-\kappa_{b\to k} h}.$$

For $n = \lceil Bh \rceil$, Step 1 bounds $\mathbb{P}_b(\tau_h > n)$ by $O(e^{-Ah})$, with any prescribed $A > 0$ if B is chosen sufficiently large. It follows that

$$q_b(h) \le (K-1)\lceil Bh \rceil e^{-\kappa_b h} + O(e^{-Ah}), \qquad \kappa_b = \min_{k \neq b} \kappa_{b\to k}.$$

Taking $A > \kappa_b$ proves $\liminf_{h\to\infty} [-\log q_b(h)]/h \ge \kappa_b$. This one-sided conclusion does not require joint tilted-row accessibility.

Step 3: matching error exponent from joint accessibility. Choose the root-minimizing $k_\star$ supplied by Assumption B.25. Write $\kappa = \kappa_{b\to k_\star} = \kappa_b$, $\mathbb{Q} = \mathbb{Q}_{b,k_\star,\sigma}$, and $v = v_{k_\star b}^{(b,\sigma)} > 0$. Under the product tilted law the increments remain i.i.d. and all means in row $k_\star$ are at least v. Each competing row $j \neq k_\star$ contains $g_{jk_\star} = -g_{k_\star j}$, whose tilted mean is strictly negative. Applying the same strong-law reasoning as in Step 1, for any fixed $\epsilon > 0$ and $n_h = \lceil (1+\epsilon)h/v \rceil$,

$$\mathbb{Q}(\tau_h \le n_h,\ J_h = k_\star) \longrightarrow 1, \qquad S_{k_\star b}(n_h)/h \longrightarrow 1 + \epsilon.$$

In particular the event

$$A_h := \{\tau_h \le n_h,\ J_h = k_\star, S_{k_\star b}(n_h) \le (1+2\epsilon)h\}$$

has $\mathbb{Q}(A_h) \to 1$. The finite-time change of measure is

$$\left.\frac{d\mathbb{Q}}{d\mathbb{P}_b}\right|_{\mathcal{F}_{n_h}} = \exp\{\kappa S_{k_\star b}(n_h)\}.$$

Because $k_\star \neq b$, A_h is a genuine wrong *emitted* decision, not merely an eligible but unselected row. Therefore

$$q_b(h) \ge \mathbb{P}_b(A_h) = \mathbb{E}_{\mathbb{Q}}[e^{-\kappa S_{k_\star b}(n_h)}\mathbf{1}_{A_h}] \ge e^{-\kappa(1+2\epsilon)h}\mathbb{Q}(A_h).$$

Letting $h \to \infty$ and then $\epsilon \downarrow 0$ proves $\limsup[-\log q_b(h)]/h \le \kappa_b$. Together with Step 2 this yields $-\log q_b(h) = \kappa_b h + o(h)$.

Step 4: elapsed time to a false or correct declaration. For all sufficiently large h, $0 < q_b(h) < 1$. The prescribed resets and i.i.d. observations make the cycle pairs i.i.d. under each fixed regime. The cycle length may be dependent on its emitted class. The dependence does not affect the following renewal identities. If the first cycle is not the requested terminal type, the remaining mean waiting time is a fresh independent copy. Equivalently, summing over the independent previous cycles gives

$$\mathrm{ARL}_b(h) = \frac{m_b(h)}{q_b(h)}, \qquad \mathbb{E}_k T_k^+(h) = \frac{m_k(h)}{1-q_k(h)}.$$

Steps 1–3 and $\log m_b(h) = o(h)$ now prove

$$\log \mathrm{ARL}_b(h) = \kappa_b h + o(h), \qquad \mathrm{ADD}^+_{k,1}(h) = h/d_k + o(h).$$

A wrong declaration contributes its full cycle duration to T_k^+. Subtracting one converts elapsed observations to immediate-change lag and has no effect on the first-order coefficient.

The repeated-declaration policy is essential for these clocks. A one-shot procedure that stops forever after an incorrect declaration has infinite time to a correct declaration on that event. Stopping forever after a correct declaration likewise precludes a later false declaration. Finite restricted means are still defined, but they are not the renewal means.

Step 5: aggregation, elimination of the threshold, and ordering. For each of the three ARL aggregations in Appendix B.8, finite minima or the log-sum-exp identity give

$$\log \mathrm{ARL}_\Phi(h;\sigma) = \kappa_\Phi(\sigma)h + o(h).$$

Lemma B.26 gives $D^+_\psi(h;\sigma) = c_\psi h + o(h)$. Centered nuisance leaves every $\mu_{k\ell}$, thus d_k and c_ψ, unchanged as σ varies. Lemma B.24 shows that κ_Φ is strictly decreasing.

For completeness, passage to the upper envelope does not require a strict monotonicity assertion about a finite-threshold correct-declaration delay. Fix σ, write $c = c_\psi > 0$ and $\kappa = \kappa_\Phi(\sigma) > 0$, and let the envelope range over the large-threshold family. For sufficiently large h,

$$(c-\epsilon)h \le D^+_\psi(h) \le (c+\epsilon)h, \qquad (\kappa-\epsilon)h \le \log \mathrm{ARL}_\Phi(h) \le (\kappa+\epsilon)h.$$

A feasible threshold at delay budget d is bounded above by $d/(c-\epsilon)$ outside a fixed compact set. The ARL on that compact set is bounded by the assumed local boundedness. Conversely, taking $h = d/(c+2\epsilon)$ supplies a feasible large threshold for all large d. Letting $\epsilon \downarrow 0$ gives

$$\log F^\star_{\psi,\Phi}(d;\sigma) = (\kappa_\Phi(\sigma)/c_\psi)d + o(d).$$

For $\sigma_1 < \sigma_2$, the difference of the two log frontiers is a strictly positive slope gap times d, plus $o(d)$. It is positive for all sufficiently large d. Integrating over a common non-saturating box gives Eq. (77). The population orthogonal-projection calculation from the main theorem, now on $\mathcal{H}_{\mathrm{mc}}$, gives Eq. (72). Its strict increase when $\eta_{\mathrm{mc}} \neq 0$ yields Eq. (78). This completes the proof. □

Other all-rivals evidence updates covered by the same proof. Set $S_{k\ell}(t) := \sum_{s=1}^{t} g_{k\ell,s}$ within a fresh cycle. The proof above also applies to a positively homogeneous, observation-adapted row margin $\widetilde{C}_{k,t}$ satisfying the two-sided bound

$$\min_{\ell \neq k} S_{k\ell}(t) \leq \widetilde{C}_{k,t} \leq \min_{\ell \neq k} W_{k\ell,t}. \tag{166}$$

This includes the unreflected accumulated margin $\min_{\ell \neq k} S_{k\ell}(t)$ and the common-start reflected margin $\max_{1 \leq r \leq t+1} \min_{\ell \neq k} \sum_{s=r}^{t} g_{k\ell,s}$, as well as the pairwise-Page margin used above. These updates do not necessarily have the same finite-threshold outcomes. The theorem covers their first-order laws under the stated joint accessibility and the same first-declaration/restart policy. It does not allow arbitrary withholding of an eligible declaration or resets based on the true class.

Verification. For any margin satisfying Eq. (166), the true-row strong-law limit follows by dividing the lower and upper bounds by t and using their common limit. The upper bound implies that a wrong-row crossing requires its comparison against the true regime to cross, because $\widetilde{C}_{k,t}$ cannot exceed that scalar reflected entry. The lower bound by cumulative row sums gives the same late-cycle Chernoff bound and the same tilted witness event. Therefore all Steps 1–5 apply. Positive homogeneity additionally supplies pathwise projective invariance. The recurrence, fixed declaration rule, and restart policy remain part of the procedure. Eq. (166) does not justify a different emission policy.

Verification of the Gaussian instantiation. Use $a = m_0^2$ and $V_\sigma = \alpha^2 a + \sigma^2 v$ from Appendix B.8.3. Under P_b, for $k \neq b$,

$$g^{\sigma}_{kb} \sim \mathcal{N}(-\alpha a, 2V_\sigma), \qquad \log \mathbb{E}_b[e^{\theta g^{\sigma}_{kb}}] = -\alpha a \theta + V_\sigma \theta^2.$$

Thus $\kappa_{b \to k} = \alpha a / V_\sigma$. For $\ell \notin \{b, k\}$,

$$\mathbb{E}_b g^{\sigma}_{k\ell} = 0, \qquad \mathrm{Cov}_b(g^{\sigma}_{k\ell}, g^{\sigma}_{kb}) = V_\sigma.$$

Exponential tilting of the Gaussian vector by κg_{kb} shifts each mean by its covariance with g_{kb} times κ. Therefore the tilted b comparison has mean $-\alpha a + 2\kappa V_\sigma = \alpha a$, and every other entry in the row has mean $\kappa V_\sigma = \alpha a$. Joint accessibility follows, with equality permitted. Under P_k, every correct-row mean is αa. All required moments are finite. The noise differences are antisymmetric and mutually dependent, demonstrating that independence between pairwise comparisons was not assumed. Finally, conditional on a signal record, the cumulative noise tensor has mean zero, thus its population inner product with the true cumulative tensor is zero. Substitution yields the displayed strictly decreasing effective slope. □

Terminal episodes with a growing delay cap. A separate first-cycle correct-declaration loss can be related to the delay coefficient without crediting wrong decisions. Let $H = H(h)$ and set the loss to $\tau_h - 1$ on $\{\tau_h \leq H, J_h = k\}$ and to H otherwise. If $H/h \to \infty$ and $\log H = o(h)$, its expectation under P_k is $h/d_k + o(h)$ under the above assumptions. The proof bounds its difference from $\mathbb{E}_k(\tau_h - 1)$ by $(H+1)q_k + \mathbb{E}_k[(\tau_h - 1 - H)_+]$. This is a delay-side, increasing-horizon statement only. It does not turn a fixed finite-record false-alarm coordinate into the regenerative ARL, and does not supply a full fixed-horizon NITA theorem.

Proof. Let $L_{k,H}$ denote the loss just defined. Pointwise,

$$|L_{k,H} - (\tau_h - 1)| \leq (H+1)\mathbf{1}\{J_h \neq k\} + (\tau_h - 1 - H)_+.$$

The error bound in Step 2 implies $(H+1)q_k(h) = o(h)$ when $\log H = o(h)$. The late-cycle Chernoff bound from Step 1 gives

$$\mathbb{E}_k[(\tau_h - 1 - H)_+] \leq \sum_{\ell \neq k} \frac{\exp\{\theta_{k\ell} h - a_{k\ell} H\}}{1 - e^{-a_{k\ell}}} = o(h)$$

when $H/h \to \infty$. Thus, $\mathbb{E}_k L_{k,H} = \mathbb{E}_k \tau_h - 1 + o(h) = h/d_k + o(h)$. The result concerns the delay loss, not a capped false-alarm mean. □

C.5 Proofs for the Markov-order-N extension

Proof of Proposition B.37. For $n \geq N$, write the path density under regime $y \in \{0, 1\}$ as

$$d\mathbb{P}_y^{(n)}(x_{1:n}) = \mu_y(x_{1:N}) \prod_{t=N+1}^{n} p_y(x_t \mid x_{t-N:t-1})\, d\nu(x_{1:n}),$$

for a common dominating measure ν. Taking the log-ratio gives

$$\log \frac{d\mathbb{P}_1^{(n)}}{d\mathbb{P}_0^{(n)}}(X_{1:n}) = \log \frac{\mu_1(X_{1:N})}{\mu_0(X_{1:N})} + \sum_{t=N+1}^{n} \log \frac{p_1(X_t \mid X_{t-N:t-1})}{p_0(X_t \mid X_{t-N:t-1})},$$

which is exactly Eq. (103). If $\mu_1 = \mu_0$, the boundary term vanishes. If B_N is integrable, then it is $O_{\mathbb{P}}(1)$ and therefore disappears after normalization by n or h in the relevant asymptotic limits. □

Proof of Proposition B.35. Under the affine model,

$$\hat{\lambda}^t = \alpha \lambda^t + \sigma \varepsilon_t.$$

By ergodicity of the joint stationary process under the reference law Q,

$$\frac{1}{T} \sum_{t=1}^{T} (\lambda^t)^2 \xrightarrow[T\to\infty]{\text{a.s.}} \mathbb{E}_Q[(\lambda^1)^2], \tag{167}$$

$$\frac{1}{T} \sum_{t=1}^{T} \hat{\lambda}^t \lambda^t \xrightarrow[T\to\infty]{\text{a.s.}} \mathbb{E}_Q[\hat{\lambda}^1 \lambda^1], \tag{168}$$

$$\frac{1}{T} \sum_{t=1}^{T} (\hat{\lambda}^t)^2 \xrightarrow[T\to\infty]{\text{a.s.}} \mathbb{E}_Q[(\hat{\lambda}^1)^2]. \tag{169}$$

Independence of ε_1 and λ^1 together with $\mathbb{E}[\varepsilon_1] = 0$ gives

$$\mathbb{E}_Q[\hat{\lambda}^1 \lambda^1] = \alpha\, \mathbb{E}_Q[(\lambda^1)^2],$$

and

$$\mathbb{E}_Q[(\hat{\lambda}^1)^2] = \alpha^2 \mathbb{E}_Q[(\lambda^1)^2] + \sigma^2 \mathbb{E}[\varepsilon_1^2] = \alpha^2 \mathbb{E}_Q[(\lambda^1)^2] + \sigma^2.$$

Substituting these limits into Eq. (97) yields

$$\rho_T^2 \xrightarrow[T\to\infty]{\text{a.s.}} \frac{\alpha^2 \mathbb{E}_Q[(\lambda^1)^2]}{\alpha^2 \mathbb{E}_Q[(\lambda^1)^2] + \sigma^2},$$

and therefore

$$\mathrm{SINE}_T^2 \xrightarrow[T\to\infty]{\text{a.s.}} \frac{\sigma^2}{\alpha^2 \mathbb{E}_Q[(\lambda^1)^2] + \sigma^2} = \frac{\zeta^2}{\mathbb{E}_Q[(\lambda^1)^2] + \zeta^2}.$$

The right-hand side is strictly increasing in ζ. □

Proof of Lemma B.31. Independence of the entire nuisance process and Markov chain gives

$$\mathbb{E}_0 e^{\theta \sum_{t=1}^{n} \hat{\lambda}^t} = \mathbb{E}_0 e^{\alpha\theta \sum_{t=1}^{n} \lambda^t} \mathbb{E} e^{\sigma\theta \sum_{t=1}^{n} \varepsilon_t}.$$

Taking $n^{-1} \log$ and the assumed limits gives $\Gamma_{\hat{\lambda}}(\theta) = \Gamma(\alpha\theta) + \Psi(\sigma\theta)$. Substitute $\vartheta = \alpha\theta$ and $\zeta = \sigma/\alpha$ for the stated root equation. □

Proof of Lemma B.32. Write $F(\vartheta, \zeta) = \Gamma(\vartheta) + \Psi(\zeta\vartheta)$. At the noiseless positive root, for $\zeta > 0$,

$$F(\vartheta_\star(0), \zeta) = \Psi(\zeta \vartheta_\star(0)) > 0.$$

The assumed upward-crossing sign pattern therefore implies $\vartheta_\star(\zeta) < \vartheta_\star(0)$. In particular, $\zeta\vartheta_\star(\zeta)$ lies in the stipulated strict-variability range. The implicit function theorem now gives

$$\vartheta_\star'(\zeta) = -\frac{\vartheta_\star(\zeta)\Psi'(\zeta\vartheta_\star(\zeta))}{\Gamma'(\vartheta_\star(\zeta)) + \zeta\Psi'(\zeta\vartheta_\star(\zeta))} < 0 :$$

the numerator is positive by Eq. (92) and the denominator by Eq. (91). Together with the comparison to $\zeta = 0$, this proves strict decrease on the full interval. □

Proof of Theorem B.34. By Lemma B.32, $\vartheta_\star(\zeta)$ is strictly decreasing on $[0, \bar{\zeta}]$. Next, the post-change drift of the estimated score is

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathbb{E}_1\left[\sum_{t=1}^{n} \hat{\lambda}^t\right] &= \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_1\left[\sum_{t=1}^{n} (\alpha\lambda^t + \sigma\varepsilon_t)\right] \\ &= \alpha \lim_{n\to\infty} \frac{1}{n}\mathbb{E}_1\left[\sum_{t=1}^{n} \lambda^t\right] + \sigma \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}[\varepsilon_t] \\ &= \alpha I_{\mathrm{rate}}. \end{aligned} \tag{170}$$

Using Assumption B.33, in particular Eq. (94),

$$\log \mathrm{ARL}(h;\zeta) = \frac{\vartheta_\star(\zeta)}{\alpha} h + o(h), \qquad \mathrm{ADD}_1(h;\zeta) = \frac{h}{\alpha I_{\mathrm{rate}}} + o(h).$$

Therefore,

$$\begin{aligned} I_{\mathrm{rate}}\, \vartheta_\star(\zeta)\, \mathrm{ADD}_1(h;\zeta) &= I_{\mathrm{rate}}\, \vartheta_\star(\zeta) \left(\frac{h}{\alpha I_{\mathrm{rate}}} + o(h)\right) \\ &= \frac{\vartheta_\star(\zeta)}{\alpha} h + o(h) \\ &= \log \mathrm{ARL}(h;\zeta) + o(h), \end{aligned}$$

which proves Eq. (95). The monotonicity of $\mathcal{E}_{\text{Markov-QCD}}(\zeta)$ follows immediately. □

Proof of Corollary B.36. Theorem B.34 gives

$$\log \mathrm{ARL}(h;\zeta) = \mathcal{E}_{\text{Markov-QCD}}(\zeta)\, \mathrm{ADD}_1(h;\zeta) + o(h).$$

Since $\mathrm{ADD}_1(h;\zeta) = h/(\alpha I_{\mathrm{rate}}) + o(h)$, the parameter $d = \mathrm{ADD}_1(h;\zeta)$ diverges with h, and the same relation can be written on the upper untruncated frontier as

$$\log F^\star_{\mathrm{M}}(d;\zeta) = \mathcal{E}_{\text{Markov-QCD}}(\zeta) d + o(d), \qquad d \to \infty.$$

This proves Eq. (99). By Theorem B.34, $\mathcal{E}_{\text{Markov-QCD}}(\zeta)$ is strictly decreasing in ζ. Fix $\zeta_1 < \zeta_2$ and set

$$\Delta_{\mathrm{M}} := \mathcal{E}_{\text{Markov-QCD}}(\zeta_1) - \mathcal{E}_{\text{Markov-QCD}}(\zeta_2) > 0.$$

Then

$$\log F^\star_{\mathrm{M}}(d;\zeta_1) - \log F^\star_{\mathrm{M}}(d;\zeta_2) = \Delta_{\mathrm{M}} d + o(d),$$

which is positive for all sufficiently large d. This gives Eq. (100). Integrating the pointwise frontier dominance over any common non-saturated high-delay box $[d_0, D^*] \times [0, Y^*]$ yields Eq. (102). Finally, Proposition B.35 shows that $\mathrm{SINE}_\infty(\zeta)$ is strictly increasing in ζ, thus the asymptotic NITA ordering is antitone to SINE along the affine family. □

C.6 Proofs for the SPRT extension

Proof of Lemma B.39. Independence gives

$$\mathbb{E}_0 e^{\kappa g_\sigma} = M_0(\alpha\kappa) N_+(\sigma\kappa), \qquad \mathbb{E}_1 e^{-\kappa g_\sigma} = M_1(\alpha\kappa) N_-(\sigma\kappa).$$

For the upper-boundary root apply the comparison argument of Lemma B.12 to (P_0, λ, ξ). For the lower-boundary root apply it to $(P_1, -\lambda, -\xi)$. The two centered nondegenerate nuisance MGFs are strictly increasing on their positive finite domains. The initial derivatives are respectively $-\alpha D_{01} < 0$ and $-\alpha D_{10} < 0$. Uniqueness of each positive root therefore gives its strict decrease with σ. □

Proof of Eq. (111). The identity is the standard Wald lemma for stopped i.i.d. sums (Wald, 1947; Siegmund, 1985, Ch. III). We review the proof because the integrability conditions are used explicitly in Proposition B.41. Fix $y \in \{0, 1\}$ and write $\tau := \tau^{(\sigma)}_{a,b}$, assuming $\mathbb{E}_y[\tau] < \infty$. By Assumption B.38, $\mathbb{E}_y[|g_\sigma|] < \infty$. Therefore

$$\sum_{t\geq 1} \mathbb{E}_y[|g_{\sigma,t}|\, \mathbf{1}\{\tau \geq t\}] = \mathbb{E}_y[|g_\sigma|] \sum_{t\geq 1} \mathbb{P}_y(\tau \geq t) = \mathbb{E}_y[|g_\sigma|]\, \mathbb{E}_y[\tau] < \infty,$$

thus Tonelli's theorem gives

$$\mathbb{E}_y[S_\tau^{(\sigma)}] = \sum_{t\geq 1} \mathbb{E}_y[g_{\sigma,t}\,\mathbf{1}\{\tau \geq t\}]\,.$$

Since $\{\tau \geq t\} \in \mathcal{F}_{t-1}^{(\sigma)}$ and $g_{\sigma,t}$ is independent of $\mathcal{F}_{t-1}^{(\sigma)}$,

$$\mathbb{E}_y[g_{\sigma,t}\,\mathbf{1}\{\tau \geq t\}] = \mathbb{E}_y[g_{\sigma,t}]\,\mathbb{P}_y(\tau \geq t) = \mathbb{E}_y[g_\sigma]\,\mathbb{P}_y(\tau \geq t).$$

Summing over $t \geq 1$ and using the tail-sum identity $\sum_{t\geq 1} \mathbb{P}_y(\tau \geq t) = \mathbb{E}_y[\tau]$ yields Eq. (111), and Eq. (112) is the specialization to $y = 1$ and $y = 0$. □

Proof of Proposition B.41. Write

$$L(\gamma) := \log \frac{\gamma}{1-\gamma}, \qquad \tau := \tau^{(\sigma)} = \tau^{(\sigma)}_{a_\sigma(\gamma), b_\sigma(\gamma)}.$$

By Assumption B.40,

$$a_\sigma(\gamma) = \frac{L(\gamma)}{\kappa_0(\sigma)} + O(1), \qquad b_\sigma(\gamma) = \frac{L(\gamma)}{\kappa_1(\sigma)} + O(1).$$

Under H_1, the test stops at the upper boundary with probability γ and at the lower boundary with probability $1-\gamma$. With overshoot absorbed into the $O(1)$ term,

$$\mathbb{E}_1[S_\tau^{(\sigma)}] = \gamma\, a_\sigma(\gamma) - (1-\gamma)\, b_\sigma(\gamma) + O(1).$$

Since $\gamma \uparrow 1$,

$$\gamma\, a_\sigma(\gamma) = a_\sigma(\gamma) + o(L(\gamma)), \qquad (1-\gamma)\, b_\sigma(\gamma) = o(L(\gamma)),$$

therefore

$$\mathbb{E}_1[S_\tau^{(\sigma)}] = \frac{L(\gamma)}{\kappa_0(\sigma)} + o(L(\gamma)).$$

Similarly, under H_0,

$$\mathbb{E}_0[S_\tau^{(\sigma)}] = (1-\gamma)\, a_\sigma(\gamma) - \gamma\, b_\sigma(\gamma) + O(1) = -\frac{L(\gamma)}{\kappa_1(\sigma)} + o(L(\gamma)).$$

Now apply the stopped-sum identity Eq. (112) to the calibrated SPRT stopping time $\tau = \tau^{(\sigma)}$:

$$\mathbb{E}_1[S_\tau^{(\sigma)}] = \mu_1(\sigma)\,\mathbb{E}_1[\tau^{(\sigma)}], \qquad \mathbb{E}_0[S_\tau^{(\sigma)}] = -\mu_0(\sigma)\,\mathbb{E}_0[\tau^{(\sigma)}].$$

Therefore

$$\mathbb{E}_1[\tau^{(\sigma)}] = \frac{L(\gamma)}{\kappa_0(\sigma)\mu_1(\sigma)} + o(L(\gamma)) = \frac{L(\gamma)}{\mathcal{E}_{10}(\sigma)} + o(L(\gamma)),$$

and

$$\mathbb{E}_0[\tau^{(\sigma)}] = \frac{L(\gamma)}{\kappa_1(\sigma)\mu_0(\sigma)} + o(L(\gamma)) = \frac{L(\gamma)}{\mathcal{E}_{01}(\sigma)} + o(L(\gamma)).$$

Averaging the two expressions yields

$$T_{\mathrm{eq}}(\gamma;\sigma) = \frac{L(\gamma)}{2}\left(\frac{1}{\mathcal{E}_{10}(\sigma)} + \frac{1}{\mathcal{E}_{01}(\sigma)}\right) + o(L(\gamma)) = \frac{L(\gamma)}{\mathcal{E}_{\mathrm{SPRT}}(\sigma)} + o(L(\gamma)),$$

where

$$\mathcal{E}_{\mathrm{SPRT}}(\sigma) = \frac{2}{\frac{1}{\mathcal{E}_{10}(\sigma)} + \frac{1}{\mathcal{E}_{01}(\sigma)}}.$$

□

Proof of Corollary B.42. Let $\tau := \tau^{(0)}_{A,A}$ denote the symmetric-boundary stopping time. If $\sigma = 0$ and $g_0 = \alpha\lambda$, then under H_0,

$$\mathbb{E}_0[e^{\kappa g_0}] = \mathbb{E}_0[e^{\alpha\kappa\lambda}].$$

Since

$$\mathbb{E}_0[e^{\lambda}] = \int \frac{p_1(x)}{p_0(x)} p_0(x)\, dx = 1,$$

the unique positive root is $\alpha\kappa_0(0) = 1$, i.e.

$$\kappa_0(0) = \frac{1}{\alpha}.$$

Likewise,

$$\mathbb{E}_1[e^{-\lambda}] = \int \frac{p_0(x)}{p_1(x)} p_1(x)\, dx = 1,$$

thus

$$\kappa_1(0) = \frac{1}{\alpha}.$$

Therefore

$$\mathcal{E}_{10}(0) = \kappa_0(0)\mu_1(0) = \frac{1}{\alpha} \cdot \alpha D_{10} = D_{10}, \qquad \mathcal{E}_{01}(0) = \kappa_1(0)\mu_0(0) = \frac{1}{\alpha} \cdot \alpha D_{01} = D_{01},$$

and

$$\mathcal{E}_{\mathrm{SPRT}}(0) = \frac{2}{\frac{1}{D_{10}} + \frac{1}{D_{01}}}.$$

Under the symmetric boundary-width equal-error design $a = b = A$, with error probability $1 - \gamma$ and negligible overshoot, Assumption B.40 gives

$$A = \frac{L(\gamma)}{\kappa_0(0)} + O(1) = \alpha L(\gamma) + O(1),$$

since $\kappa_0(0) = 1/\alpha$. Under H_1,

$$\mathbb{E}_1[S_\tau] = \gamma A + (1-\gamma)(-A) + O(1) = (2\gamma - 1)A + O(1) = \alpha g(\gamma) + O(1),$$

where $g(\gamma) = (2\gamma - 1)L(\gamma)$. Under H_0,

$$\mathbb{E}_0[S_\tau] = (1-\gamma)A + \gamma(-A) + O(1) = -(2\gamma - 1)A + O(1) = -\alpha g(\gamma) + O(1).$$

Applying Eq. (112) at $\sigma = 0$ and the symmetric boundaries gives

$$\mathbb{E}_1[S_\tau] = \mu_1(0)\, \mathbb{E}_1[\tau], \qquad \mathbb{E}_0[S_\tau] = -\mu_0(0)\, \mathbb{E}_0[\tau],$$

Thus

$$\mu_1(0)\, \mathbb{E}_1[\tau] = \alpha g(\gamma) + O(1), \qquad \mu_0(0)\, \mathbb{E}_0[\tau] = \alpha g(\gamma) + O(1),$$

Since $\mu_1(0) = \alpha D_{10}$ and $\mu_0(0) = \alpha D_{01}$, and $\alpha > 0$ is fixed,

$$D_{10}\, \mathbb{E}_1[\tau] = g(\gamma) + O(1), \qquad D_{01}\, \mathbb{E}_0[\tau] = g(\gamma) + O(1),$$

thus

$$\mathbb{E}_1[\tau] = \frac{g(\gamma)}{D_{10}} + O(1), \qquad \mathbb{E}_0[\tau] = \frac{g(\gamma)}{D_{01}} + O(1).$$

Averaging gives

$$T_{\mathrm{eq}}(\gamma) = \frac{g(\gamma)}{\mathcal{E}_{\mathrm{SPRT}}(0)} + O(1).$$

Finally,

$$g(\gamma) = L(\gamma) - 2(1-\gamma)L(\gamma) = L(\gamma) + o(L(\gamma)),$$

since $(1-\gamma)L(\gamma) \to 0$. □

Proof of Corollary B.43. From Proposition B.41,

$$\mathbb{E}_1[\tau^{(\sigma)}] = \frac{L(\gamma)}{\mathcal{E}_{10}(\sigma)} + o(L(\gamma)), \qquad \mathbb{E}_0[\tau^{(\sigma)}] = \frac{L(\gamma)}{\mathcal{E}_{01}(\sigma)} + o(L(\gamma)).$$

Therefore

$$\begin{aligned} T_\pi(\gamma;\sigma) &= \pi\,\mathbb{E}_1[\tau^{(\sigma)}] + (1-\pi)\,\mathbb{E}_0[\tau^{(\sigma)}] \\ &= L(\gamma)\left(\frac{\pi}{\mathcal{E}_{10}(\sigma)} + \frac{1-\pi}{\mathcal{E}_{01}(\sigma)}\right) + o(L(\gamma)) \\ &= \frac{L(\gamma)}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma)} + o(L(\gamma)), \end{aligned}$$

where

$$\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma) = \frac{1}{\frac{\pi}{\mathcal{E}_{10}(\sigma)} + \frac{1-\pi}{\mathcal{E}_{01}(\sigma)}}.$$

If $\sigma = 0$, use $\mathcal{E}_{10}(0) = D_{10}$ and $\mathcal{E}_{01}(0) = D_{01}$. □

Proof of Proposition B.44. By Lemma B.39, both directional roots $\kappa_0(\sigma)$ and $\kappa_1(\sigma)$ are strictly decreasing in σ. By centered noise,

$$\mu_1(\sigma) = \alpha D_{10}, \qquad \mu_0(\sigma) = \alpha D_{01},$$

thus the directional slopes

$$\mathcal{E}_{10}(\sigma) = \kappa_0(\sigma)\mu_1(\sigma), \qquad \mathcal{E}_{01}(\sigma) = \kappa_1(\sigma)\mu_0(\sigma)$$

are strictly decreasing in σ. Therefore, for every fixed class prior $\pi \in (0,1)$, their prior-weighted harmonic aggregate

$$\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma) = \frac{1}{\frac{\pi}{\mathcal{E}_{10}(\sigma)} + \frac{1-\pi}{\mathcal{E}_{01}(\sigma)}}$$

is also strictly decreasing in σ.

Fix $\sigma_1 < \sigma_2$ and write

$$\mathcal{E}_j := \mathcal{E}_{\mathrm{SPRT},\pi}(\sigma_j), \qquad j \in \{1,2\}.$$

Then $\mathcal{E}_1 > \mathcal{E}_2$, thus

$$\delta := \frac{1}{\mathcal{E}_2} - \frac{1}{\mathcal{E}_1} > 0.$$

By Corollary B.43, for each $j \in \{1,2\}$ there exists a function $\varepsilon_j(\gamma) \to 0$ as $\gamma \uparrow 1$ such that

$$T_\pi(\gamma;\sigma_j) = L(\gamma)\left(\frac{1}{\mathcal{E}_j} + \varepsilon_j(\gamma)\right).$$

Choose $\gamma_0 \in (1/2, 1)$ close to 1 that

$$|\varepsilon_j(\gamma)| \le \frac{\delta}{4} \qquad \forall \gamma \in [\gamma_0, 1), \quad j \in \{1,2\}.$$

Then for every $\gamma \in [\gamma_0, 1)$,

$$\begin{aligned} T_\pi(\gamma;\sigma_2) - T_\pi(\gamma;\sigma_1) &= L(\gamma)\left(\frac{1}{\mathcal{E}_2} - \frac{1}{\mathcal{E}_1} + \varepsilon_2(\gamma) - \varepsilon_1(\gamma)\right) \\ &\ge L(\gamma)\left(\delta - \frac{\delta}{4} - \frac{\delta}{4}\right) = \frac{\delta}{2}L(\gamma) > 0. \end{aligned}$$

Thus

$$T_\pi(\gamma;\sigma_1) < T_\pi(\gamma;\sigma_2) \qquad \forall \gamma \in [\gamma_0, 1).$$

Fix any $\gamma_1 \in (\gamma_0, 1)$. Pointwise time dominance on $[\gamma_0, \gamma_1]$ implies frontier dominance on the common interval, and therefore

$$\mathrm{NITA}_{\mathrm{SPRT}}^{[\gamma_0,\gamma_1]}(\sigma_1) > \mathrm{NITA}_{\mathrm{SPRT}}^{[\gamma_0,\gamma_1]}(\sigma_2).$$

For the bounded-remainder extension on a fixed interval $\mathcal{U}$, observe that

$$T_\pi(\gamma;\sigma_2) - T_\pi(\gamma;\sigma_1) = L(\gamma)\left(\frac{1}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma_2)} - \frac{1}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma_1)}\right) + \big(r(\gamma;\sigma_2) - r(\gamma;\sigma_1)\big).$$

By Eq. (124),

$$r(\gamma;\sigma_2) - r(\gamma;\sigma_1) \geq -\Delta.$$

Therefore

$$T_\pi(\gamma;\sigma_2) - T_\pi(\gamma;\sigma_1) \geq L(\gamma)\left(\frac{1}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma_2)} - \frac{1}{\mathcal{E}_{\mathrm{SPRT},\pi}(\sigma_1)}\right) - \Delta.$$

If Eq. (125) holds, then the right-hand side is strictly positive for every $\gamma \in \mathcal{U}$, proving

$$T_\pi(\gamma;\sigma_1) < T_\pi(\gamma;\sigma_2) \qquad \forall \gamma \in \mathcal{U}.$$

This yields pointwise frontier dominance on the common interval and therefore a larger interval-restricted NITA on $\mathcal{U}$ for σ_1 than for σ_2. □

Proof of Proposition B.45. Fix σ and abbreviate $\kappa_i = \kappa_i(\sigma)$ and $\mu_i = \mu_i(\sigma)$. Let $\varepsilon := 1 - \gamma$. On the feasible set of Eq. (129),

$$(1-\pi)\alpha_\sigma(a,b) + \pi\beta_\sigma(a,b) \leq \varepsilon.$$

Since $\pi \in (0,1)$, both directional errors are individually $O(\varepsilon)$:

$$\alpha_\sigma(a,b) \leq \frac{\varepsilon}{1-\pi}, \qquad \beta_\sigma(a,b) \leq \frac{\varepsilon}{\pi}.$$

Using Eq. (130), this implies

$$a \geq \frac{\log(1/\varepsilon)}{\kappa_0} + O(1), \qquad b \geq \frac{\log(1/\varepsilon)}{\kappa_1} + O(1).$$

Because $L(\gamma) = \log(\gamma/(1-\gamma)) = \log(1/\varepsilon) + o(1)$, Eq. (131) gives the lower bound

$$T_{\pi,\sigma}(a,b) \geq L(\gamma)\left(\frac{\pi}{\kappa_0\mu_1} + \frac{1-\pi}{\kappa_1\mu_0}\right) + O(1)$$

uniformly over feasible (a,b). Therefore the same lower bound holds for $T^\star_{\pi,\mathrm{agg}}(\gamma;\sigma)$.

For the matching upper bound, choose

$$a_\gamma = \frac{L(\gamma)+C}{\kappa_0}, \qquad b_\gamma = \frac{L(\gamma)+C}{\kappa_1},$$

where C is a sufficiently large constant, independent of γ, so that Eq. (130) makes $(1-\pi)\alpha_\sigma(a_\gamma,b_\gamma) + \pi\beta_\sigma(a_\gamma,b_\gamma) \leq 1-\gamma$ for all γ close enough to one. Then (a_γ,b_γ) is feasible, and Eq. (131) gives

$$T^\star_{\pi,\mathrm{agg}}(\gamma;\sigma) \leq L(\gamma)\left(\frac{\pi}{\kappa_0\mu_1} + \frac{1-\pi}{\kappa_1\mu_0}\right) + O(1).$$

Combining the two bounds yields Eq. (132)–Eq. (133).

By Lemma B.39, $\kappa_0(\sigma)$ and $\kappa_1(\sigma)$ strictly decrease with σ, while centered noise leaves $\mu_1(\sigma) = \alpha D_{10}$ and $\mu_0(\sigma) = \alpha D_{01}$ fixed. Therefore $\mathcal{E}_{\mathrm{agg},\pi}(\sigma)$ is strictly decreasing. The high-accuracy NITA ordering follows by the same pointwise-frontier argument used in Proposition B.44: for $\sigma_1 < \sigma_2$, the gap between $1/\mathcal{E}_{\mathrm{agg},\pi}(\sigma_2)$ and $1/\mathcal{E}_{\mathrm{agg},\pi}(\sigma_1)$ dominates the lower-order terms for all sufficiently large $L(\gamma)$, giving smaller mean time for σ_1 throughout a common high-accuracy interval and thus larger interval-restricted NITA. □

Proof of Corollary B.46. Set $a = b = c$. By Eq. (130),

$$\begin{aligned}1 - \gamma^{\mathrm{sym}}_{\pi,\sigma}(c) &= (1-\pi)\alpha_\sigma(c,c) + \pi\beta_\sigma(c,c)\\ &= (1-\pi)\exp\{-\kappa_0(\sigma)c + O(1)\} + \pi\exp\{-\kappa_1(\sigma)c + O(1)\}\\ &= \exp\{-\kappa_{\min}(\sigma)c + O(1)\},\end{aligned}$$

where $\kappa_{\min}(\sigma) := \min\{\kappa_0(\sigma), \kappa_1(\sigma)\}$ and the last equality uses $\pi \in (0,1)$. Thus the threshold achieving aggregate accuracy γ along the symmetric path satisfies

$$c_\gamma(\sigma) = \frac{L(\gamma)}{\kappa_{\min}(\sigma)} + O(1).$$

The time expansion Eq. (131) gives

$$T^{\mathrm{sym}}_{\pi,\sigma}(c) = c\left(\frac{\pi}{\mu_1(\sigma)} + \frac{1-\pi}{\mu_0(\sigma)}\right) + O(1).$$

Substituting $c = c_\gamma(\sigma)$ yields Eq. (136)–Eq. (137). Since both $\kappa_0(\sigma)$ and $\kappa_1(\sigma)$ strictly decrease with σ, their minimum also strictly decreases. The denominator in Eq. (137) is constant in σ under the centered affine-noise family, thus $\mathcal{E}_{\mathrm{sym},\pi}(\sigma)$ strictly decreases. The resulting pointwise high-accuracy time ordering, and therefore the interval-restricted NITA ordering on a common high-accuracy box, follows exactly as in Proposition B.44. □

C.7 Proofs for the multiclass and Markov-order-N SPRT extensions

Proof of Lemma B.50. For each ordered pair, independence under $\mathbb{P}_\ell$ gives

$$\mathbb{E}_\ell e^{\kappa\hat{\lambda}_{k\ell}} = M_{\ell\to k}(\alpha\kappa)N_{k\ell}(\sigma\kappa).$$

Apply the root-comparison proof of Lemma B.12 with (P_0, λ, ξ) replaced by $(P_\ell, \lambda_{k\ell}, \varepsilon_{k\ell})$. The required initial derivative is $-\alpha D_{\mathrm{KL}}(P_\ell \| P_k) < 0$, and the assumed centered, nondegenerate nuisance and unique positive root give strict decrease of $\kappa_{\ell\to k}$. Centering under P_k gives $\mu_{k\ell} = \alpha I_{k\ell} > 0$, independent of σ. Thus, $\mathcal{E}_{k\ell} = \kappa_{\ell\to k}\mu_{k\ell}$ is strictly decreasing. □

Proof of Proposition B.52. For a finite set of rivals, the maximum is 1-Lipschitz in the maximum coordinate norm. Since $\max_{\ell\neq k} L_K(\gamma)/\mathcal{E}_{k\ell}(\sigma) = L_K(\gamma)/\mathcal{E}_k(\sigma)$, Eq. (145) implies

$$\left|\max_{\ell\neq k}\tau_{k\ell}(\gamma;\sigma) - \frac{L_K(\gamma)}{\mathcal{E}_k(\sigma)}\right| \le \max_{\ell\neq k}|R_{k\ell}(\gamma;\sigma)|.$$

The expectation of the right side is bounded by the finite sum of the assumed $o(L_K(\gamma))$ absolute-mean remainders. Taking expectations and applying Eq. (146) proves Eq. (147). □

Proof of Corollary B.53. Multiply the classwise expansion of Proposition B.52 by the fixed weights π_k and sum over the finite set of classes. The remainder remains $o(L_K(\gamma_\pi))$, and the leading coefficient is $\sum_k \pi_k/\mathcal{E}_k(\sigma) = 1/\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma)$, proving Eq. (148). □

Proof of Proposition B.55. For each $\gamma_\pi \in \mathcal{U}$,

$$\begin{aligned}T_\pi(\gamma_\pi;\sigma_2) - T_\pi(\gamma_\pi;\sigma_1) = L_K(\gamma_\pi)&\left(\frac{1}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma_2)} - \frac{1}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma_1)}\right)\\ &+ \big(r_K(\gamma_\pi;\sigma_2) - r_K(\gamma_\pi;\sigma_1)\big).\end{aligned}$$

By Eq. (150),

$$r_K(\gamma_\pi;\sigma_2) - r_K(\gamma_\pi;\sigma_1) \ge -\Delta_K.$$

Thus

$$T_\pi(\gamma_\pi;\sigma_2) - T_\pi(\gamma_\pi;\sigma_1) \ge L_K(\gamma_\pi)\left(\frac{1}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma_2)} - \frac{1}{\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma_1)}\right) - \Delta_K.$$

If Eq. (151) holds, the right-hand side is strictly positive for every $\gamma_\pi \in \mathcal{U}$, so

$$T_\pi(\gamma_\pi;\sigma_1) < T_\pi(\gamma_\pi;\sigma_2) \qquad \forall\gamma_\pi \in \mathcal{U}.$$

This yields pointwise dominance of the weighted-accuracy frontier and therefore larger NITA on the common interval with a shared normalization box. □

Proof of Lemma B.56. Define

$$F_0(\vartheta,\zeta) := \Gamma_0(\vartheta) + \Psi(\zeta\vartheta), \qquad F_1(v,\zeta) := \Gamma_1(v) + \Psi(-\zeta v).$$

By definition, the two roots satisfy

$$F_0(\vartheta_0^\star(\zeta),\zeta) = 0, \qquad F_1(\vartheta_1^\star(\zeta),\zeta) = 0.$$

Assuming the derivative conditions stated in the lemma, the implicit function theorem applies to both equations.

For the upper-boundary root,

$$\partial_\zeta F_0(\vartheta,\zeta) = \vartheta\,\Psi'(\zeta\vartheta), \qquad \partial_\vartheta F_0(\vartheta,\zeta) = \Gamma_0'(\vartheta) + \zeta\,\Psi'(\zeta\vartheta).$$

Therefore

$$(\vartheta_0^\star)'(\zeta) = -\frac{\vartheta_0^\star(\zeta)\Psi'(\zeta\vartheta_0^\star(\zeta))}{\Gamma_0'(\vartheta_0^\star(\zeta)) + \zeta\,\Psi'(\zeta\vartheta_0^\star(\zeta))}.$$

Because $\vartheta_0^\star(\zeta) > 0$, $\Psi'(\zeta\vartheta_0^\star(\zeta)) > 0$, and the denominator is positive by assumption, we have

$$(\vartheta_0^\star)'(\zeta) < 0.$$

For the lower-boundary root,

$$\partial_\zeta F_1(v,\zeta) = -v\,\Psi'(-\zeta v), \qquad \partial_v F_1(v,\zeta) = \Gamma_1'(v) - \zeta\,\Psi'(-\zeta v).$$

Thus

$$(\vartheta_1^\star)'(\zeta) = -\frac{-\vartheta_1^\star(\zeta)\Psi'(-\zeta\vartheta_1^\star(\zeta))}{\Gamma_1'(\vartheta_1^\star(\zeta)) - \zeta\,\Psi'(-\zeta\vartheta_1^\star(\zeta))}.$$

Since $\Psi'(s) < 0$ for $s < 0$, the numerator is positive and the denominator is positive by assumption, therefore

$$(\vartheta_1^\star)'(\zeta) < 0.$$

Thus both roots are strictly decreasing. □

Proof of Corollary B.58. The class-conditional speed expansions in Assumption B.57 are

$$\mathbb{E}_1[\tau^{(\zeta)}] = \frac{L(\gamma)}{\mathcal{E}_{10}(\zeta)} + o(L(\gamma)), \qquad \mathbb{E}_0[\tau^{(\zeta)}] = \frac{L(\gamma)}{\mathcal{E}_{01}(\zeta)} + o(L(\gamma)).$$

Therefore, for any fixed class prior $\pi \in (0,1)$,

$$T_\pi(\gamma;\zeta) = \pi\,\mathbb{E}_1[\tau^{(\zeta)}] + (1-\pi)\,\mathbb{E}_0[\tau^{(\zeta)}] = \frac{L(\gamma)}{\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta)} + o(L(\gamma)),$$

where

$$\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta) = \frac{1}{\frac{\pi}{\mathcal{E}_{10}(\zeta)} + \frac{1-\pi}{\mathcal{E}_{01}(\zeta)}}.$$

If $\zeta = 0$, the exact conditional LLR is used up to the positive scale α. Then the SCGF roots satisfy

$$\Gamma_0(1) = 0, \qquad \Gamma_1(1) = 0,$$

which implies

$$\vartheta_0^\star(0) = \vartheta_1^\star(0) = 1.$$

Therefore

$$\mathcal{E}_{10}(0) = I_{10}, \qquad \mathcal{E}_{01}(0) = I_{01},$$

and the prior-weighted harmonic expression Eq. (157) follows. □

Proof of Corollary B.60. The proof is identical to that of Proposition B.44, replacing $\mathcal{E}_{\text{SPRT},\pi}$ by $\mathcal{E}_{\text{Markov-SPRT},\pi}$ and the remainder term r by r_{M}. Indeed, for every $\gamma \in \mathcal{U}$,

$$T_\pi(\gamma;\zeta_2) - T_\pi(\gamma;\zeta_1) = L(\gamma)\left(\frac{1}{\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta_2)} - \frac{1}{\mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta_1)}\right) + \big(r_{\text{M}}(\gamma;\zeta_2) - r_{\text{M}}(\gamma;\zeta_1)\big),$$

and the stated gap condition implies pointwise time dominance, which in turn implies NITA dominance on the common interval with a shared normalization box. □

Proof of Theorem B.54. For item (1), apply the orthogonal-projection calculation in the proof of Theorem 3.6 item (1) in $\mathcal{H}_{\mathrm{MS}}$, using Assumption B.49. This gives the displayed SINE formula and its strict increase when $\eta_{\mathrm{MS}} \neq 0$.

For item (2), Lemma B.50 makes every positive $\mathcal{E}_{k\ell}(\sigma)$ strictly decreasing. Their finite minimum $\mathcal{E}_k(\sigma)$ is therefore strictly decreasing. The positive weighted sum $\sum_k \pi_k/\mathcal{E}_k(\sigma)$ strictly increases, thus its reciprocal $\mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma)$ strictly decreases. For item (3), apply Proposition B.52 under each $\mathbb{P}_k$ with $\tau_k = \tau$, then Corollary B.53. Their hypotheses are supplied by Assumption B.51.

For item (4), fix $\sigma_1 < \sigma_2$ and write $\mathcal{E}_j = \mathcal{E}_{\mathrm{MSPRT},\pi}(\sigma_j)$. Item (3) gives

$$T_\pi(\gamma_\pi;\sigma_2) - T_\pi(\gamma_\pi;\sigma_1) = L_K(\gamma_\pi)\big(\mathcal{E}_2^{-1} - \mathcal{E}_1^{-1} + o(1)\big).$$

The constant in parentheses is positive by item (2), thus the difference is positive for every $\gamma_\pi \geq \gamma_0$ for some $\gamma_0 < 1$. For each $\gamma_1 \in (\gamma_0, 1)$, inversion to the upper accuracy envelope and integration over the specified common comparison region give the NITA ordering, exactly as in the frontier part of Proposition B.22. Together with item (1), this is the asserted antitone ordering. □

Proof of Theorem B.59. Item (1) is Proposition B.35 under the stated stationary increment-geometry assumptions. For item (2), Lemma B.56 gives strictly decreasing directional roots. Multiplication by the fixed positive information rates preserves strict decrease. Taking the reciprocal of the positive weighted sum of their inverse slopes gives the decreasing harmonic aggregate. Item (3) follows from Corollary B.58.

For item (4), fix $\zeta_1 < \zeta_2$ and write $\mathcal{E}_j = \mathcal{E}_{\text{Markov-SPRT},\pi}(\zeta_j)$. Then

$$T_\pi(\gamma;\zeta_2) - T_\pi(\gamma;\zeta_1) = L(\gamma)\big(\mathcal{E}_2^{-1} - \mathcal{E}_1^{-1} + o(1)\big) > 0$$

for all sufficiently high γ, since $\mathcal{E}_1 > \mathcal{E}_2$. Choose $\gamma_0 < 1$ beyond this point. For any $\gamma_1 \in (\gamma_0, 1)$, the same common-region inversion and integration used in the preceding multiclass-SPRT proof yields the strict NITA ordering. Item (1) then transfers this ordering to the opposite SINE ordering. □

D Experimental Details

Experimental objective. Our experiments evaluate NITA as a checkpoint selector for LLR estimators using practical streaming data. The main QCD experiments therefore use finite streams with multiple changepoints, preserving network state and evidence history across changes. The streaming evaluation tests checkpoint selection directly on available labeled datasets without reconstructing separate long single-regime evaluation sequences. The controlled independent-unit study in Appendix D.14 separately examines finite-validation reliability. Appendix F.5 explains the complementary roles of these evaluations.

By Proposition 2.1, exact P_+LLR recovery preserves the stopping times and decisions of the corresponding true LLR-based test when the thresholds are calibrated. The empirical question is therefore whether NITA selects useful imperfect checkpoints, rather than whether classical SPRT/CUSUM optimality can be re-established experimentally. To complement the streaming evaluation, Appendix F.5 separately evaluates the frozen selected checkpoints under the classical CUSUM criteria: each run stops at its first alarm, and performance is measured by ARL and immediate-change ADD_1.

D.1 Shared QCD evaluation protocol

For record $m \in [N_{\mathrm{seq}}]$ of length T, let $Y_{m,t} \in [K]$, where $K = K_{\mathrm{reg}}$ counts all regimes. The network produces one continuous learned-LLR tensor per record. Neither network state nor LLR/CUSUM history is reset at true changepoints or threshold hits. A class is confirmed when its CUSUM evidence reaches the threshold against every rival, for both binary and multiclass QCD.

Learned evidence. One K-logit network supplies all ordered pairwise scores by logit subtraction. These are log softmax odds, treated as learned LLRs without explicit prior correction (Eq. (21)). Order $N = 0$ sums one-step differences. For $N > 0$, we follow Ebihara et al. (2021) to subtract the score of the existing length-N context from that of the same context plus the next observation; initial prefixes supply the first block. Write the resulting cumulative estimates as $\widehat{\Lambda}^{1:t}_{k\ell,m}$, with increments $\widehat{\lambda}^{t}_{k\ell,m} = \widehat{\Lambda}^{1:t}_{k\ell,m} - \widehat{\Lambda}^{1:t-1}_{k\ell,m}$.

Pairwise CUSUM, then the hardest rival. For $N = 0$, the pairwise reflected scores and class margins are

$$B_{m,k\ell,0} = 0, \qquad B_{m,k\ell,t} = \max\{0, B_{m,k\ell,t-1} + \widehat{\lambda}^{t}_{k\ell,m}\}, \qquad C_{m,k,t} = \min_{\ell \neq k} B_{m,k\ell,t}.$$

Equivalently, $B_{m,k\ell,t} = \widehat{\Lambda}^{1:t}_{k\ell,m} - \min_{0\le u\le t} \widehat{\Lambda}^{1:u}_{k\ell,m}$.

False-class and correct-class confirmation times. Let $\nu_{m,1} < \cdots < \nu_{m,J_m}$ be the label changes, with $\nu_{m,J_m+1} = T + 1$. Set $\nu^{\text{first}}_m = \nu_{m,1}$ when $J_m > 0$ and $T + 1$ otherwise, and $b_m = Y_{m,1}$. For change j, let $k_{m,j} = Y_{m,\nu_{m,j}}$, $e_{m,j} = \nu_{m,j+1}$, and $L^{+}_{m,j} = e_{m,j} - \nu_{m,j}$. All intervals read the same chart with its inherited history:

$$\tau^{-}_m(h) := \inf\{1 \le t < \nu^{\text{first}}_m : \max_{k \neq b_m} C_{m,k,t} \ge h\},$$
$$\tau^{+}_{m,j}(h) := \inf\{\nu_{m,j} \le t < e_{m,j} : C_{m,k_{m,j},t} \ge h\}.$$

Here $\inf \varnothing = \infty$. A wrong-class confirmation does not end correct-class follow-up, and simultaneous eligibility is allowed. A correct row already above threshold at $\nu_{m,j}$ has lag zero. A hit at or after $e_{m,j}$ cannot count for change j.

KM observations and pooling. The last valid follow-up times are $\mathsf{C}^{-}_m = \nu^{\text{first}}_m - 1$ for one-indexed run length and $\mathsf{C}^{+}_{m,j} = L^{+}_{m,j} - 1$ for zero-indexed lag. Record

$$\mathsf{U}^{-}_m(h) := \min\{\tau^{-}_m(h), \mathsf{C}^{-}_m\}, \qquad \mathsf{D}^{-}_m(h) := \mathbf{1}\{\tau^{-}_m(h) < \infty\},$$
$$\mathsf{U}^{+}_{m,j}(h) := \min\{\tau^{+}_{m,j}(h) - \nu_{m,j}, \mathsf{C}^{+}_{m,j}\}, \quad \mathsf{D}^{+}_{m,j}(h) := \mathbf{1}\{\tau^{+}_{m,j}(h) < \infty\}.$$

A no-hit interval has flag zero and is neither discarded nor treated as an event. For interpreting post-change censoring, the event time is the lag until the existing chart would confirm class $k_{m,j}$ if that class continued to generate observations. Network state and evidence are inherited at $\nu_{m,j}$, not reset. The next changepoint or record boundary ends available follow-up, not the underlying waiting time. Thus $\tau^{+}_{m,j}(h) = \infty$ above denotes no hit observed in that interval, not a claim that confirmation could never occur. No-change records contribute only initial-prefix follow-up, excluding padded entries.

False-event observations are pooled across pre-change prefixes, and delay observations across all post-change intervals, weighting transitions rather than first averaging within records. No interval is selected according to detection success.

D.1.1 KM OPERATING COORDINATES AND NITA

Apply the KM construction of Appendix A.3 to obtain $\widehat{S}^{\text{ADD}}_h$ and $\widehat{S}^{\text{ARL}}_h$, then use

$$\widehat{\text{ADD}}_{\mathcal{P}}(h) = \sum_{t=0}^{H_x - 1} \widehat{S}^{\text{ADD}}_h(t), \qquad \widehat{\text{ARL}}_{\mathcal{P}}(h) = \sum_{t=0}^{H_y - 1} \widehat{S}^{\text{ARL}}_h(t).$$

The protocol $\mathcal{P}$ fixes initialization, event definitions, and common restriction horizons H_x, H_y, independently of validation estimates. Replacing estimated survival by population survival defines the restricted-mean targets (Miyagawa & Ebihara, 2026). The guarantee for estimating them uses the conditions in Appendix A.3.

In the finite-stream computation, an empty risk set maitains the last KM survival value through the prescribed horizon, without inserting a detection or false alarm. Where the horizon exceeds the follow-up support, this is the flat-tail extension described in Appendix A.3. The same prescribed convention is used across checkpoints and for validation and test evaluation. These streaming comparisons assess checkpoint-selection utility under that convention, not a guarantee for the unobserved event-time tail. The supplementary code specifies the horizons and follow-up construction for each experiment.

Apply the common sorting, upper-envelope, and rectangle-sum construction in Eq. (29) to (d_j, f_j).

D.2 SYNTHETIC DATA: I.I.D. GAUSSIAN REGIMES

The generator returns observations $X_m^{(1:T)} \in \mathbb{R}^{T\times d}$, regime labels $Y_m^{(1:T)} \in [K]^T$, changepoints, and exact pairwise LLR references. Split sizes, sequence lengths, and validation variants are in Appendix D.8.

Regime generation. With minimum segment length $L_{\min} \geq 1$, draw

$$N_m^{\mathrm{CP}} \sim \mathrm{Unif}\{0, \ldots, (N_{\max}^{\mathrm{CP}})^{\mathrm{eff}}\}, \qquad (N_{\max}^{\mathrm{CP}})^{\mathrm{eff}} = \min\{N_{\max}^{\mathrm{CP}}, \lfloor T/L_{\min}\rfloor - 1\}.$$

Distribute the residual $R_m = T - (N_m^{\mathrm{CP}} + 1)L_{\min}$ by the specified random composition $\sum_j E_{m,j} = R_m$, $E_{m,j} \geq 0$, and set $L_{m,j} = L_{\min} + E_{m,j}$. Therefore, all segments meet the minimum length and sum to T. Their boundaries are $\nu_{m,j} = 1 + \sum_{u=1}^{j} L_{m,u}$. Draw $\ell_{m,1}$ uniformly from $[K]$, then $\ell_{m,j+1}$ uniformly from $[K] \setminus \{\ell_{m,j}\}$. Set $Y_{m,t} = \ell_{m,j}$ on $\sum_{u<j} L_{m,u} < t \leq \sum_{u\leq j} L_{m,u}$, thus each boundary is a genuine class change.

Gaussian observations and exact LLR references. Conditional on the regime labels, observations are independent across time, with

$$X_{m,t} \mid (Y_{m,t} = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, I_d), \qquad \boldsymbol{\mu}_k = m_{\mathrm{off}} e_k, \qquad d = 128, \quad m_{\mathrm{off}} = 2.0, \tag{171}$$

where e_k is the k-th standard basis vector. The exact pairwise LLR is

$$\lambda_{k\ell,m}^t = m_{\mathrm{off}}\big(X_{m,t}^{(k)} - X_{m,t}^{(\ell)}\big), \qquad k \neq \ell, \tag{172}$$

where $X_{m,t}^{(k)}$ denotes coordinate k. Prefix sums, as defined in Appendix A.1, give the cumulative references used for SINE; the operating curves use learned evidence.

D.3 SYNTHETIC DATA: VECTOR AUTOREGRESSION (VAR)

The generator combines piecewise-constant labels with a class-conditional truncated-normal VAR(N). Here d is feature dimension, T returned length, $B \geq N$ burn-in length, and $K = K_{\mathrm{reg}}$ the total number of classes. It simulates $T_{\mathrm{full}} = B + T$ observations and returns $X_m^{(1:T)}$, $Y_m^{(1:T)}$, changepoints, and optional exact conditional log-likelihoods and pairwise LLRs.

Regimes. Draw the changepoint count M_m uniformly from the configured integer range, whose lower endpoint is 0 or $M_{\min}$ and upper endpoint is $M_{\max}$, then clip it at $\lfloor T/L_{\min}\rfloor - 1$. Sample $M_m + 1$ segment lengths at least $L_{\min}$ summing to T. The boundaries and piecewise-constant labels follow the indexing in Appendix D.2. The initial label is fixed at y_{start} or uniform on $[K]$, as configured. Each subsequent label is uniform on the other $K - 1$ classes.

VAR parameters. For class k, parameters are $\Theta_k = \{\mathbf{A}_{k,1}, \ldots, \mathbf{A}_{k,N}, \mathbf{b}_k, \boldsymbol{\sigma}_k\}$, with lag matrices in $\mathbb{R}^{d\times d}$, drift in $\mathbb{R}^d$, and positive diagonal noise scales. Draw $(\mathbf{A}_{k,\ell})_{ij} \sim \mathrm{Unif}(-s/N, s/N)$. If the companion matrix $\mathbf{C}_k \in \mathbb{R}^{Nd\times Nd}$ has spectral radius above $r_{\mathrm{tgt}} < 1$, shrink all class-k lag matrices by a common factor and recompute until $\rho(\mathbf{C}_k) \leq r_{\mathrm{tgt}}$. This controls the linear dynamics before truncation and mean clipping.

For $K = 2$, independently draw $\mathbf{b}_1, \mathbf{g}_1, \mathbf{g}_2 \sim \mathcal{N}(0, \delta^2\mathbf{I})$ and set $\mathbf{b}_2 = \mathbf{g}_1 + \mathbf{g}_2$. For $K > 2$, draw $\mathbf{b}_k \sim \mathcal{N}(0, \sigma_b^2\mathbf{I})$ with $\sigma_b = \sqrt{3/2}\,\delta$. Both constructions give pairwise drift-difference variance $3\delta^2\mathbf{I}$; optional common rescaling matches mean squared pairwise distance to $3d\delta^2$. Draw $\varepsilon_i \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 0.05^2)$ and set $\boldsymbol{\sigma}_0 = \sigma_{\mathrm{scale}} \exp(\boldsymbol{\varepsilon})$ elementwise. This vector is shared across classes.

Conditional density. With shared coordinate bounds $l_i = x_{\mathrm{low}}$, $h_i = x_{\mathrm{high}}$, and clipping margin $m_{\mathrm{clip}} \geq 0$, let

$$\boldsymbol{\mu}_{m,u,k}^{\mathrm{lin}} = \mathbf{b}_k + \sum_{\ell=1}^{N} \mathbf{A}_{k,\ell}\mathbf{x}_{m,u-\ell}, \qquad \boldsymbol{\mu}_{m,u,k} = \mathrm{clip}(\boldsymbol{\mu}_{m,u,k}^{\mathrm{lin}}, \mathbf{l} + m_{\mathrm{clip}}, \mathbf{h} - m_{\mathrm{clip}}).$$

Given the active class $\widetilde{Y}_{m,u} = k$, sample the product density

$$p_k(\mathbf{x}_{m,u} \mid \mathbf{x}_{m,u-N:u-1}) = \prod_{i=1}^{d} \mathcal{N}_{[l_i,h_i]}(x_{m,u,i}; \mu_{m,u,k,i}, \sigma^2_{k,i}),$$

where

$$\mathcal{N}_{[l,h]}(x; \mu, \sigma^2) = \frac{\phi((x-\mu)/\sigma)}{\sigma\{\Phi((h-\mu)/\sigma) - \Phi((l-\mu)/\sigma)\}} \mathbf{1}\{l \le x \le h\}.$$

Here ϕ, Φ are the standard-normal density and distribution function. Clipping makes the autoregressive location parameter nonlinear in the lag history. Given the class and observed history, this parameter is known, thus the displayed truncated-normal density and the corresponding conditional LLRs can be evaluated directly.

Initialization and returned window. For $u \le N$, independently initialize the coordinates from $\mathcal{N}_{[l_i,h_i]}(\cdot; 0, \sigma^2_{\text{init}})$; for $u > N$, use the VAR dynamics. Set $\widetilde{Y}_{m,1:B} = Y_{m,1}$ and $\widetilde{Y}_{m,B+t} = Y_{m,t}$, and return $X_m^{(1:T)} = (\mathbf{x}_{m,B+1}, \ldots, \mathbf{x}_{m,B+T})$. The burn-in supplies a full lag context for the first returned observation.

Exact conditional-LLR reference. At returned time t, take the log ratio of the displayed conditional densities on the same actual lag history $\mathbf{h}_{m,t} = \mathbf{x}_{m,B+t-N:B+t-1}$ and accumulate it as in Appendix A.1. These references are exact for the generator's conditional law, including clipping and truncation. SINE compares the specified LLR representation, not CUSUM statistics. The initial-block distinction in Proposition B.37 and matched valid-time masks apply in both QCD and SPRT.

D.4 Isotonic regression for visualization

We fit NITA as a nonincreasing function of checkpoint-level SINE^2 by least-squares isotonic regression, computed with the pool-adjacent-violators algorithm (Barlow et al., 1972; Ayer et al., 1955; Guntuboyina & Sen, 2017). This fit is a visual aid, not a test of the relationship. Inference uses the original checkpoint measurements to compute within-run Spearman coefficients (i.e., run-level tests) and applies two-sided Wilcoxon tests across runs (Appendix E.1; Spearman 1904; Wilcoxon 1945), rather than treating pooled checkpoints as independent.

D.5 Training-fitted KDE pseudo-LLR references

Gaussian KDE provides approximate real-data references for SINE (Rosenblatt, 1956; 1969; Parzen, 1962). All standardization constants, template banks, class scales, and bandwidths are fitted on training data only. Write $(x_{b,t}, y_{b,t})$ for an observation, with $x_{b,t} \in \mathbb{R}^D$ and $y_{b,t} \in \mathcal{Y}$, $|\mathcal{Y}| = K$.

Standardization and template banks. Using training feature means $\bar{x}_j^{\text{tr}}$, standard deviations s_j^{tr}, and stabilizer $\epsilon_{\text{std}} > 0$, define

$$\widetilde{x}_{b,t,j} := \frac{x_{b,t,j} - \bar{x}_j^{\text{tr}}}{s_j^{\text{tr}} + \epsilon_{\text{std}}}, \qquad j = 1, \ldots, D. \tag{173}$$

All subsequent calculations use these standardized coordinates. Let $\mathcal{I}_k = \{(b,t) : y_{b,t} = k\}$ and $N_k = |\mathcal{I}_k|$. A global class bank $\mathcal{T}_k$ contains $M_k = \min\{256, N_k\}$ observations sampled uniformly without replacement. At time t, if $N_{k,t} := |\{b : y_{b,t} = k\}| \ge 16$, sample a bank $\mathcal{T}_{k,t}$ of $M_{k,t} = \min\{256, N_{k,t}\}$ observations uniformly without replacement from that class/time bucket. Otherwise use $\mathcal{T}_{k,t} = \mathcal{T}_k$ and $M_{k,t} = M_k$. Write its centers as $\mu_{k,t,i}$, $i \in [M_{k,t}]$.

Class scale and bandwidth. Let $\hat{m}_k = N_k^{-1} \sum_{(b,t)\in\mathcal{I}_k} \widetilde{x}_{b,t}$. The class scale and bandwidth are

$$\hat{v}_k := \frac{1}{N_k D} \sum_{(b,t)\in\mathcal{I}_k} \|\widetilde{x}_{b,t} - \hat{m}_k\|_2^2, \tag{174}$$

$$\widehat{b}^2_{k,t} := \max\left\{\epsilon^2_{\text{bw}}, c^2_{\text{bw}} n_{\text{eff},k,t}^{-2/(D+4)} \hat{v}_k\right\}, \qquad n_{\text{eff},k,t} = M_{k,t}. \tag{175}$$

Here $c_{\text{bw}} = 1$, $\epsilon_{\text{bw}} > 0$ is the bandwidth floor, and $n_{\text{eff},k,t} = M_k$ after global-bank fallback.

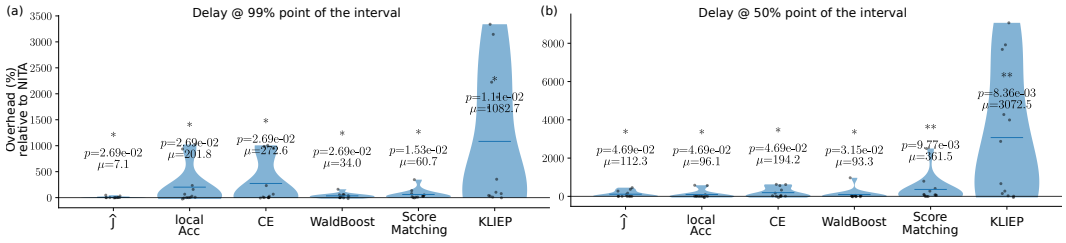


Figure 6: **Extended held-out test comparison including the KLIEP baseline.** Extended version of Fig. 4 with an additional KLIEP / M-KLIEP-style direct density-ratio baseline.

Density and pairwise references. The fitted density is the Gaussian mixture

$$\widehat{p}(\widetilde{x} \mid y = k, t) := \frac{1}{M_{k,t}} \sum_{i=1}^{M_{k,t}} \mathcal{N}(\widetilde{x}; \mu_{k,t,i}, \widehat{b}^2_{k,t} I_D), \tag{176}$$

where $\mathcal{N}(\cdot; \mu, \Sigma)$ denotes a Gaussian density. Its pairwise pseudo-LLRs are

$$\ell^{k,\ell}_{b,t} := \log \widehat{p}(\widetilde{x}_{b,t} \mid y = k, t) - \log \widehat{p}(\widetilde{x}_{b,t} \mid y = \ell, t), \qquad k \neq \ell. \tag{177}$$

Binary SINE uses one directed pair, whereas multiclass SINE stacks the off-diagonal pairs.

D.6 Additional baseline: KLIEP / M-KLIEP-style density-ratio estimation

We additionally evaluate a direct density-ratio estimator based on KLIEP (Sugiyama et al., 2008; Givens et al., 2023). Our no-missingness setting uses $\phi_0 \equiv \phi_1 \equiv 0$, reducing M-KLIEP to the unconstrained objective

$$\hat{r}_{\mathrm{KLIEP}} \in \arg\max_{r \in \bar{\mathcal{G}}} \left[\frac{1}{n_1} \sum_{i=1}^{n_1} \log r(X_i^{(1)}) - \log \left(\frac{1}{n_0} \sum_{j=1}^{n_0} r(X_j^{(0)}) \right) \right], \tag{178}$$

See Table 4 for the training setups. As shown in Fig. 6, It provides a ratio-level comparison to projective LLR validation. In the reported experiment, it has the largest variance in delay overhead and a positive shift relative to NITA at both the 99% and 50% operating points.

D.7 Architecture and hyperparameter settings

The directly comparable selectors (NITA, CE, local Acc, and $\hat{J}$) use the same saved checkpoints from the LLR estimator backbone (Ebihara et al., 2023). Table 3 gives the common architecture, optimizer, schedule, and checkpoint budget. One epoch exhausts an entire training dataset for parameter update. The best epoch is selected separately for each criterion.

Batch size is 1000 for the synthetic experiments. real-data exceptions are: Kitsune (2000), UCR Meat (10), UCR RefrigerationDevices (100), and SiW (1024). Selectors share seeds, splits, schedule, checkpoint budget, and the training-scaled grid and operating region of Appendix D.12.1.

Mixed-label training targets. For QCD windows containing several regime labels, the posterior-based estimator uses a sliding causal soft label: the proportions of classes within the window. Let $m_{i,t} = \mathbf{1}\{t \leq L_i\}$ mask the valid length L_i of record i, and define $\mathcal{C}_r(t) = \{\max(1, t-r+1), \ldots, t\}$ for $r \in [R]$, $R = N + 1$. With $\mathbf{e}_k$ the class-k one-hot vector,

$$\bar{\mathbf{y}}_i^{(t,r)} := \frac{\sum_{s \in \mathcal{C}_r(t)} m_{i,s} \mathbf{e}_{Y_{i,s}}}{\max\{\sum_{s \in \mathcal{C}_r(t)} m_{i,s}, 1\}}. \tag{179}$$

Cross-entropy to these targets averages all valid end-time/window-length pairs equally. Early windows use the available prefix, while later windows use only the last r positions, with no future labels. The population cross-entropy minimizer is the conditional expectation of this window-label vector, not necessarily the instantaneous posterior. This target choice does not change the subsequent LLR construction or checkpoint selector.

Table 3: Shared architecture and training setup for the directly comparable checkpoint-selection baselines (NITA, CE, Acc, and $\hat{J}$).

Component / hyperparameter	Value
Backbone	TANDEMformer + NSP
Transformer blocks	1
Attention heads	1
Hidden dimension	32
Feed-forward dimension	32
Dropout	0.1
Optimizer	AdamW
Learning rate	10^{-3}
LR schedule	CosineAnnealingWarmRestarts
Scheduler parameters	$T_0 = 3,\ T_{\text{mult}} = 1,\ \eta_{\min} = 0$
Weight decay	10^{-4}
Training epochs	50
Batch size	1000 by default
Gradient clipping	none
Warmup	none
Early stopping	none
Training loss	equal-weight CE + LSEL
Checkpoint evaluation cadence	every epoch
Threshold sweep for operating curves	1000 normalized thresholds training-derived scale per checkpoint (Appendix D.12.1)
Checkpointing	best epoch saved separately per criterion
Directly compared selectors	NITA, CE, Acc, $\hat{J}$

$\hat{J}$ baseline. The $\hat{J}$ baseline follows the practical prescription of Cooper & Meyn (2024; 2025). Its computation requires an exponential-tail parameter ρ, for which the authors use the heuristic value $\rho = 0.02$. Thus we adopt the same value throughout. Since the published $\hat{J}$ construction is derived for i.i.d. sequences, its use on non-i.i.d. datasets in our experiments should be interpreted as a heuristic extension rather than a matched theory-backed baseline.

Other estimator baselines. Table 4 gives the actual adapted settings for WaldBoost, class-conditional score matching, and the unconstrained M-KLIEP. Note that these compare estimator pipelines rather than selection criteria on an identical trained model.

D.8 Dataset details

Table 5 summarizes the datasets used in the main-text experiments. For all real-world datasets, the train/validation/test split is performed *before* sliding-window extraction, so that overlapping windows never cross split boundaries. For SPRT tasks, labels are constant over time ($N^{\text{CP}}_{\max} = 0$) by definition. for QCD tasks, sequences may contain up to $N^{\text{CP}}_{\max} = 3$ changepoints unless otherwise noted. Additional real-world datasets used only in the appendix are described in Appendix D.10.

Gaussian i.i.d. synthetic datasets. Use the parameters in Appendix D.2 and split sizes in Table 5. All i.i.d. QCD training splits use random changepoints with $L_{\min} = 1$. The binary experiments in Figs. 3b–c and 7a share those Gaussian parameters and the same training split, differing only in validation label durations:

- Equalized validation ('Fixed' in Table 1 / Fig. 3a). $T_{\text{val}} = 60$ comprises six ten-step segments. Binary labels alternate from a random initial class, giving exactly 30 observations per class in every record.
- Random validation ('Rand.' in Table 1 / Fig. 3b). $T_{\text{val}} = 60$, with unconstrained random changepoints and $L_{\min} = 1$, allowing one-step segments.

Table 4: Implementation details for the reference-level estimator baselines.

Baseline	Model / score parameterization	Optimization / implementation details
WaldBoost (Sochman & Matas, 2005)	Additive ensemble of weak learners, with each weak learner of the form $h_t(x) = \mathrm{sign}(w_t^\top x + b_t)$. We allow up to 200 weak learners in total; the fitted number can be smaller. For multiclass tasks, up to 200 weak learners are trained for each class combination.	We follow the original WaldBoost training rule and stopping mechanism as closely as possible, using linear-logit weak learners throughout.
Hyvärinen score matching (Hyvärinen, 2005; Banerjee & Tarokh, 2024)	One class-conditional energy network $E_k(x)$ per class. Each E_k is an MLP with hidden dimensions $(256, 256)$, LayerNorm, SiLU activations, and scalar output. Pseudo-log-likelihoods are defined by $\hat{\ell}_k(x) = -S_H(E_k; x)$, where S_H is the Hyvärinen score estimated with one Hutchinson sample. Instantaneous pairwise scores are $\hat{\lambda}^t_{k\ell} = \hat{\ell}_k(X_t) - \hat{\ell}_\ell(X_t)$, and are accumulated over time to form $\hat{\Lambda}^{1:t}_{k\ell}$.	Optimizer: AdamW, learning rate: 10^{-3}, batch size: 128, epochs: 5, weight decay: 0, gradient clipping: 1.0, Hutchinson samples = 1, feature standardization enabled, inference batch size: 256, $k_{\mathrm{scale}} = 1.0$.
Unconstrained KLIEP (Sugiyama et al., 2008; Givens et al., 2023)	Direct density ratio estimator M-KLIEP, using the same Transformer feature-extraction configuration as the directly comparable baselines.	Backbone: Transformer with the same configuration as the shared TANDEMformer encode. Optimizer: AdamW, learning rate: 3×10^{-4}, weight decay: 10^{-4}.

Table 5: Summary of datasets used in the main-text experiments

Dataset	Task	d	T	K	Splits $(N_{\mathrm{tr}}, N_{\mathrm{val}}, N_{\mathrm{te}})$
Gaussian i.i.d. synthetic	QCD / SPRT	128	50	2,3	$(20000, 5000, 20000)$
Gaussian AR(4) synthetic	QCD / SPRT	64	50	2,3	$(20000, 5000, 20000)$
Kitsune (Mirsky et al., 2018)	QCD	115	50	2	$(115493, 19249, 57747)$
UCR Meat (Dau et al., 2018)	SPRT	28	31	3	$(60, 30, 30)$

Here d is the per-step feature dimension, T is the maximum sequence length, and K is the number of classes. The split sizes $(N_{\mathrm{tr}}, N_{\mathrm{val}}, N_{\mathrm{te}})$ give the numbers of training, validation, and test records.

- Minimum-duration validation ('Min5' in Table 1 / Fig. 7a). $T_{\mathrm{val}} = 50$, with random changepoints and $L_{\min} = 5$.

All remaining i.i.d. settings use $T = 50$ unless noted otherwise. Three-class QCD (Fig. 7b) uses unconstrained random changepoints in train/test data and $L_{\min} = 5$ in validation.

Gaussian AR(4) synthetic datasets. Use the VAR generator in Appendix D.3 at order $N = 4$ with the dimensions and independent split sizes in Table 5. QCD allows up to three changes; SPRT

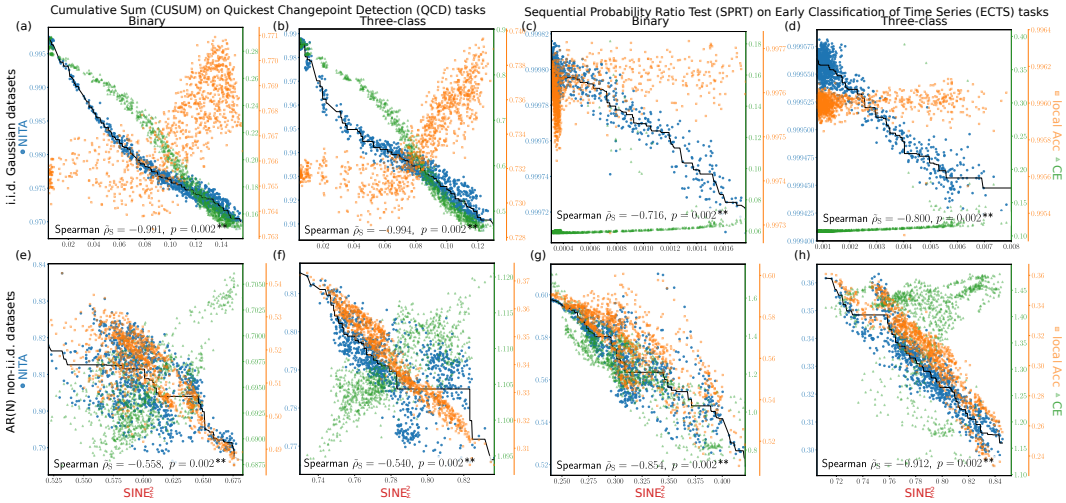


Figure 7: **Checkpoint-level associations in the synthetic settings.** Each point is one saved checkpoint, with cumulative-trajectory projective error SINE^2_Σ on the horizontal axis. Blue, green, and orange points show high-threshold NITA, CE, and local Acc, respectively. Black curves are decreasing isotonic fits to NITA only. The top row uses i.i.d. Gaussian data and the bottom row uses non-i.i.d. AR(N) data. Columns separate QCD and SPRT, with binary and three-class tasks. Table 1 summarizes the corresponding run-level Spearman coefficients.

has none. All AR(4) splits have minimum segment length five unless stated otherwise, with exact conditional-LLR references.

Kitsune cyberattack detection dataset (Mirsky et al., 2018). For binary QCD, use the four attack subsets with the smallest class imbalance: *Mirai*, *SSDP*, *ARP MitM*, and *Active Wiretap*. This controls class-imbalance confounding rather than testing robustness to it. Dimensions and split sizes are in Table 5. SINE uses the estimated LLR described in Appendix D.5.

UCR Meat (Dau et al., 2018). For three-class SPRT, split the raw examples before sliding-window embedding. Dimensions and split sizes are in Table 5.

D.9 Checkpoint-level scatter plots underlying Table 1

Figures 7 and 8 give the checkpoint-level plots underlying Table 1, beyond the motivating examples in Fig. 3. Black curves are decreasing isotonic fits to NITA only (Appendix D.4); panel annotations summarize the run-level association tests in Appendix E.1.

D.10 Additional real-world datasets

We additionally evaluate UCR RefrigerationDevices (Dau et al., 2018) and Spoofing in the Wild (SiW) (Liu et al., 2018) as SPRT benchmarks. KDE pseudo-LLRs are used for computing SINE, as in Appendix D.5. Run-level Spearman coefficients and two-sided Wilcoxon tests follow Appendix E.1.

UCR RefrigerationDevices (Dau et al., 2018). We use the RefrigerationDevices benchmark from the UCR archive (Dau et al., 2018) as an additional three-class SPRT task. Following our preprocessing pipeline, we form feature vectors of length 32 using a sliding window and then construct train, validation, and test splits of size $N_{\text{train}} = 600$, $N_{\text{val}} = 300$, $N_{\text{test}} = 600$. The resulting sequence has maximum horizon $T = 22$ and number of classes $K = 3$.

Spoofing in the Wild (SiW (Liu et al., 2018)). We use the SiW face anti-spoofing benchmark (Liu et al., 2018) as an additional binary SPRT task. Following the preprocessing of Ebihara et al. (2021), we extract a 512-dimensional feature vector from each video frame using a ResNet-152 backbone (He

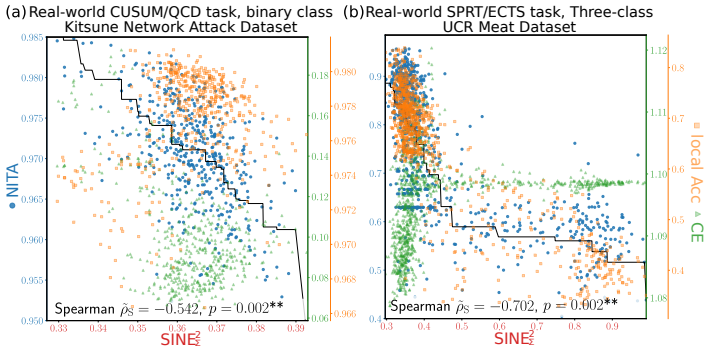


Figure 8: **Checkpoint-level associations on the primary real-world benchmarks.** Each point is one saved checkpoint. (a) Binary QCD on the Kitsune cyberattack dataset. (b) Three-class SPRT on the UCR Meat dataset. Cumulative-trajectory SINE^2_Σ is computed from training-fitted KDE pseudo-LLRs only for analysis. Colors and the NITA-only isotonic fits follow Fig. 7, and Table 1 reports the corresponding run-level Spearman coefficients.

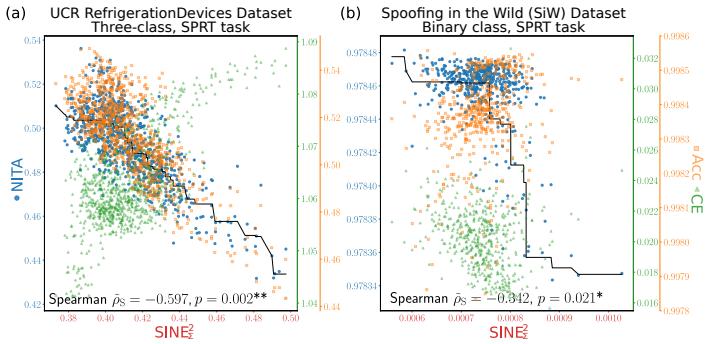


Figure 9: **Additional real-world SPRT benchmarks.** (a) UCR RefrigerationDevices (Dau et al., 2018), a three-class SPRT task, and (b) Spoofing in the Wild (SiW) (Liu et al., 2018), a binary SPRT task. As in Fig. 8, KDE-based pseudo-LLRs are used only to compute SINE^2 for analysis. The reported p-value is from a two-sided Wilcoxon signed-rank test for the alternative $H_1 : \mathrm{median}(\rho_r) \neq 0$ (Spearman, 1904; Wilcoxon, 1945). Both panels show a significant negative association between SINE^2 and NITA, consistent with the primary real-world results in Fig. 8.

et al., 2016a;b). We then construct train, validation, and test splits of size $N_{\text{train}} = 46729$, $N_{\text{val}} = 4968$, $N_{\text{test}} = 43878$. The maximum horizon is $T = 50$, and the number of classes is $K = 2$.

Figure 9 reports significant negative run-level SINE–NITA associations on both datasets, extending the primary real-data diagnostics.

D.11 SPRT EVALUATION AND NITA

Binary and multiclass early classification use the same matrix-SPRT implementation. The cumulative pairwise LLR tensor is constructed from shared logits using Appendix D.1. Each checkpoint's raw scalar thresholds are obtained based on scale parameter derived on a training data (Appendix D.12.1).

Acceptance and terminal decision. For checkpoint m, let $H^{\mathrm{SPRT}}_{k\ell,t}(h)$ have zero diagonal and common off-diagonal value h before T, and be zero at T. Define

$$s^{\mathrm{SPRT}}_{m,k,t}(h) = \min_{\ell \in [K]} \{\widehat{\Lambda}^{1:t}_{k\ell,m} - H^{\mathrm{SPRT}}_{k\ell,t}(h)\}, \qquad \tau^{\mathrm{SPRT}}_{m,k}(h) = \inf\{t \in [T] : s^{\mathrm{SPRT}}_{m,k,t}(h) = 0\}.$$

The zero diagonal caps a passing margin at zero without relaxing the rival conditions. Set $\tau^T_m(h) = \min_k \tau^{\mathrm{SPRT}}_{m,k}(h)$ and $d_{m,h} \in \arg\min_k \tau^{\mathrm{SPRT}}_{m,k}(h)$ with fixed class-order tie-breaking. For $h > 0$, antisymmetry prevents two classes from passing simultaneously before T. For $K = 2$, these are symmetric binary-SPRT boundaries.

Observed stopping times. All terminal outcomes are observed under the finite-horizon protocol. records reaching T are not censored for KM. The time coordinate is

$$\widehat{\mathrm{MHT}}(h) := \frac{1}{N_{\mathrm{seq}}} \sum_{m=1}^{N_{\mathrm{seq}}} \tau_m^T(h).$$

D.11.1 CLASS WEIGHTING AND SAMPLING

For each class k, let $n_k^{\mathrm{cls}} > 0$ be its record count and define

$$\widehat{p}_k^{\mathrm{correct}}(h) := \frac{1}{n_k^{\mathrm{cls}}} \sum_{m:Y_m=k} \mathbf{1}\{d_{m,h} = k\}, \qquad \widehat{t}_k^{\mathrm{stop}}(h) := \frac{1}{n_k^{\mathrm{cls}}} \sum_{m:Y_m=k} \tau_m^T(h).$$

The reported terminal accuracy is macro recall,

$$\widehat{\mathrm{Acc}}_\pi(h) := \sum_{k=1}^{K} \pi_k \widehat{p}_k^{\mathrm{correct}}(h), \qquad \pi_k = 1/K,$$

with population target $\sum_k \pi_k \mathbb{P}_k(d_{\tau^T} = k)$. SAT plot labels Acc_τ use this convention, not the local Acc checkpoint selector. The theoretical class-weighted time T_π has estimator $\sum_k \pi_k \widehat{t}_k^{\mathrm{stop}}$, whereas record-average MHT uses weights $n_k^{\mathrm{cls}}/N_{\mathrm{seq}}$.

Connection to the finite-sample guarantee. Macro recall weights classes equally, thus its record weights are unequal when class counts differ. For independent records with fixed positive class counts, the accuracy concentration step in Theorem 3.8 therefore uses

$$n_{\mathrm{eff}}^{\mathrm{acc}} = \left(\sum_{k=1}^{K} \frac{\pi_k^2}{n_k^{\mathrm{cls}}} \right)^{-1}, \qquad \pi_k = 1/K,$$

rather than automatically using N_{seq}. Balanced class counts give $n_{\mathrm{eff}}^{\mathrm{acc}} = N_{\mathrm{seq}}$. The weighted concentration argument is given alongside the sample-mean bound in Appendix C.3.

D.12 ABLATION DETAILS

SINE/NITA with improper normalization. We ablate the two normalization choices that make SINE and NITA comparable across checkpoints. First, we replace the bounded projective error by the *unnormalized* residual

$$\widetilde{\mathrm{SINE}}^2(\hat{\mathbf{L}}, \mathbf{L}) := \min_{\alpha \geq 0} \|\hat{\mathbf{L}} - \alpha \mathbf{L}\|_2^2, \tag{180}$$

which removes the denominator that makes SINE scale-invariant and bounded in $[0, 1]$. Second, we replace NITA by a *checkpoint-dependent* normalization

$$\widetilde{\mathrm{NITA}}(g) := \frac{1}{D_{\max}(g)\, F_{\max}(g)} \int_0^{D_{\max}(g)} F_g(x)\, dx, \tag{181}$$

where the caps are recomputed separately for each checkpoint instead of being fixed across models.

Both ablations impair the observed relation. Checkpoint-dependent NITA normalization can inflate early values when the attainable operating range is small. The unnormalized SINE residual loses its bounded, scale-comparable interpretation. Figure 10 illustrates both effects.

Increment-level versus cumulative-trajectory SINE. Using the binary i.i.d. QCD setup of Fig. 3b, repeat training ten times with the same data, splits, backbone, optimizer, schedule, and checkpoint cadence. For each checkpoint compute SINE_I^2 from one-step LLRs and SINE_Σ^2 from their cumulative paths. Both are compared with the same full-box NITA on a common fixed box, thus only the horizontal-axis representation changes, not the checkpoints or operating metric.

To quantify compatibility, for each run we compute Spearman's rank correlation between NITA and the chosen SINE variant across checkpoints, report the median run-level correlation $\tilde{\rho}_S$, and assess

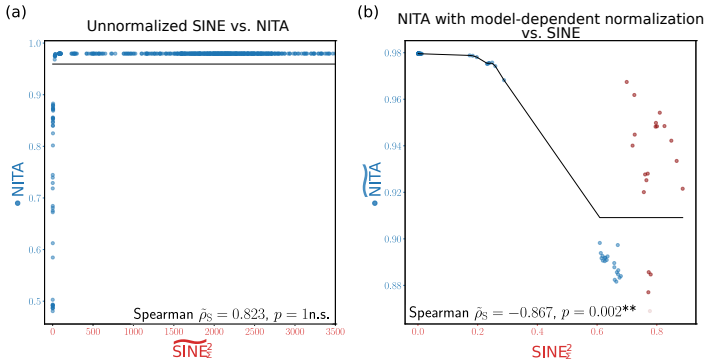


Figure 10: **Ablation of the normalization used in SINE and NITA.** (a) Unnormalized $\widetilde{\text{SINE}}^2$ versus correct NITA. Once the scale-normalizing denominator is removed, the anti-monotone trend is deteriorated. (b) Correct SINE^2 versus checkpoint-dependent $\widetilde{\text{NITA}}$. The first two training iterations are highlighted in red. Model-dependent normalization destabilizes the anti-monotone relation at the beginning of training, because weak early checkpoints have very small attainable operating ranges and can therefore be spuriously inflated by their own local range.

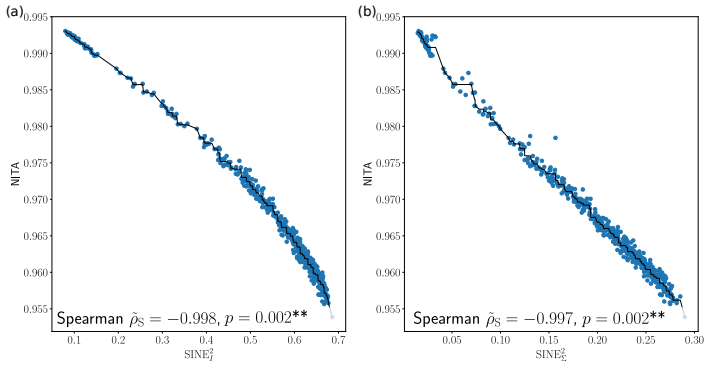


Figure 11: **(a) Increment-level versus (b) cumulative-trajectory SINE in the binary i.i.d. QCD experiment.** The training setup follows the binary i.i.d. QCD setting; this representation ablation uses full-box NITA. Each point corresponds to one checkpoint from one of the 10 independent runs; the black curves are isotonic regression fits. Both show an almost perfectly order-preserving anti-monotone relation with NITA ($\tilde{\rho}_S = -0.998$ for SINE^2_I and $\tilde{\rho}_S = -0.997$ for SINE^2_Σ).

significance with the same two-sided Wilcoxon signed-rank test used elsewhere for SINE–NITA anti-monotonicity, namely the alternative $H_1 : \text{median}(\rho_r) \neq 0$. As in Appendix D.4, the black curves are isotonic regression fits shown only for visualization. Figure 11 shows that both representations exhibit the same statistically significant anti-monotone trend: the increment-level metric yields $\tilde{\rho}_S = -0.998$ and the cumulative-trajectory metric yields $\tilde{\rho}_S = -0.997$, with $p = 0.002$ in both cases. This stress test indicates that, in the binary i.i.d. QCD setting, the observed SINE–NITA compatibility is not an artifact of reporting only SINE_Σ; the increment-level metric SINE_I preserves essentially the same checkpoint ordering.

D.12.1 Main high-performance-tail protocol and full-box ablation

Main-text NITA uses a common high-performance region corresponding to the high-ARL or high-accuracy frontier. The full-box NITA ablation includes the lower-performance portion without changing the other experimental settings.

Training-derived scales and a common normalized threshold grid. Fix a setting and training run, and let m index its saved checkpoints. Let $\mathcal{R}^{\text{tr}}_m$ be the latest training minibatch used for checkpoint m's scale estimate, separate from the validation records. Compute one scalar $s_m :=$

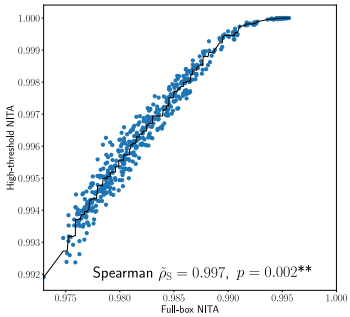


Figure 12: **Full-box NITA versus high-threshold NITA in the binary i.i.d. QCD experiment.** Each point corresponds to one checkpoint from one of the 10 independent runs. The black curve is an isotonic regression fit.

$\max_{X \in \mathcal{R}_m^{\mathrm{tr}},\ 1 \le t \le T_X,\ k \ne \ell} \left| \widehat{\Lambda}_{m,k\ell}^{1:t}(X^{(1,t)}) \right|$, $\theta_j := \frac{j-1}{999}$, $h_{m,j} := s_m \theta_j$, $j = 1, \ldots, 1000$, where T_X is valid record length. The maximum uses learned training scores over records, valid times, and off-diagonal pairs (one directed LLR in the binary case), not validation/test or reference-LLR values. One positive finite s_m is frozen per checkpoint before validation and kept for held-out evaluation. All boundary components use the same multiplier. Thus the normalized grid is common, while raw grids span $[0, s_m]$. We report ten independent runs per setting.

The operating caps X^*, Y^* are fixed before training. For setting d, a linear threshold sweep of a preliminary LLR estimator fitted and evaluated on training data only defines $\mathcal{Y}_d = [Y_d^{\mathrm{lo}}, Y_d^{\mathrm{hi}}]$, the common attainable ARL interval for QCD or terminal-accuracy interval for SPRT. Use $\mathcal{Y}_d^{\mathrm{tail}} = \left[Y_d^{\mathrm{lo}} + \frac{1}{2}(Y_d^{\mathrm{hi}} - Y_d^{\mathrm{lo}}), Y_d^{\mathrm{hi}}\right]$. The pilot operating point at the lower endpoint of $\mathcal{Y}_d^{\mathrm{tail}}$ determines the common lower cost x_0 and normalized threshold subset. These choices and the normalizer $(X^* - x_0)Y^*$ are frozen before validation and shared across checkpoints, repeated trials, and selectors.

Compute QCD restricted means with the common horizons and all eligible event/censoring observations in Appendix D.1.1. No interval is selected according to detection success. Apply Definition 3.3 to the envelope on the pilot-fixed threshold subset over $[x_0, X^*]$. Neither threshold membership nor the comparison region is recomputed from validation or test performance. The 50% cutoff is in performance units, not numerical-threshold units, throughout the high-threshold NITA experiments. Full-box NITA instead integrates the complete grid envelope over $[0, X^*]$ with normalizer X^*Y^*. This high-performance criterion is not a literal infinite-threshold limit. The supplementary code specifies the pilot configuration and fixed per-setting parameters.

For each checkpoint, validation supplies cumulative SINE_Σ^2, NITA, CE, and local Acc. Table 6 summaries use median run-level Spearman coefficients and the two-sided tests in Appendix E.1.

Full-box NITA versus high-threshold NITA. In the binary i.i.d. setting of Fig. 3, compute both NITA variants for every saved checkpoint using 1000 raw thresholds on $[0, s_m]$, with training-derived s_m and the same KM estimates and restriction horizons. The experiment is repeated 10 times. High-threshold NITA uses the training-fixed upper-performance region and common normalizer specified above, as in all other high-threshold experiments. Report the median run-level Spearman coefficient and its two-sided test (Appendix E.1). As in the main text, the isotonic curve is only a visual aid (Appendix D.4). Fig. 12 shows that the two quantities are almost perfectly rank-aligned ($\tilde{\rho}_S = 0.997$), and a two-sided Wilcoxon signed-rank test on the run-level Spearman coefficients strongly supports positive association ($p = 0.002$, $H_1 : \mathrm{median}(\rho_r) \ne 0$). The visible plateau near the upper-right corner shows a ceiling effect: once the high-threshold frontier nearly fills the tail box, the normalized tail area saturates at 1.0 and loses resolution among strong checkpoints. See also Appendix F.2 for discussion on using the full-box NITA.

Held-out checkpoint-selection performance. Full-box NITA evaluates the complete deployment range, whereas the tail-restricted experiment isolates the high-ARL or high-accuracy region empha-

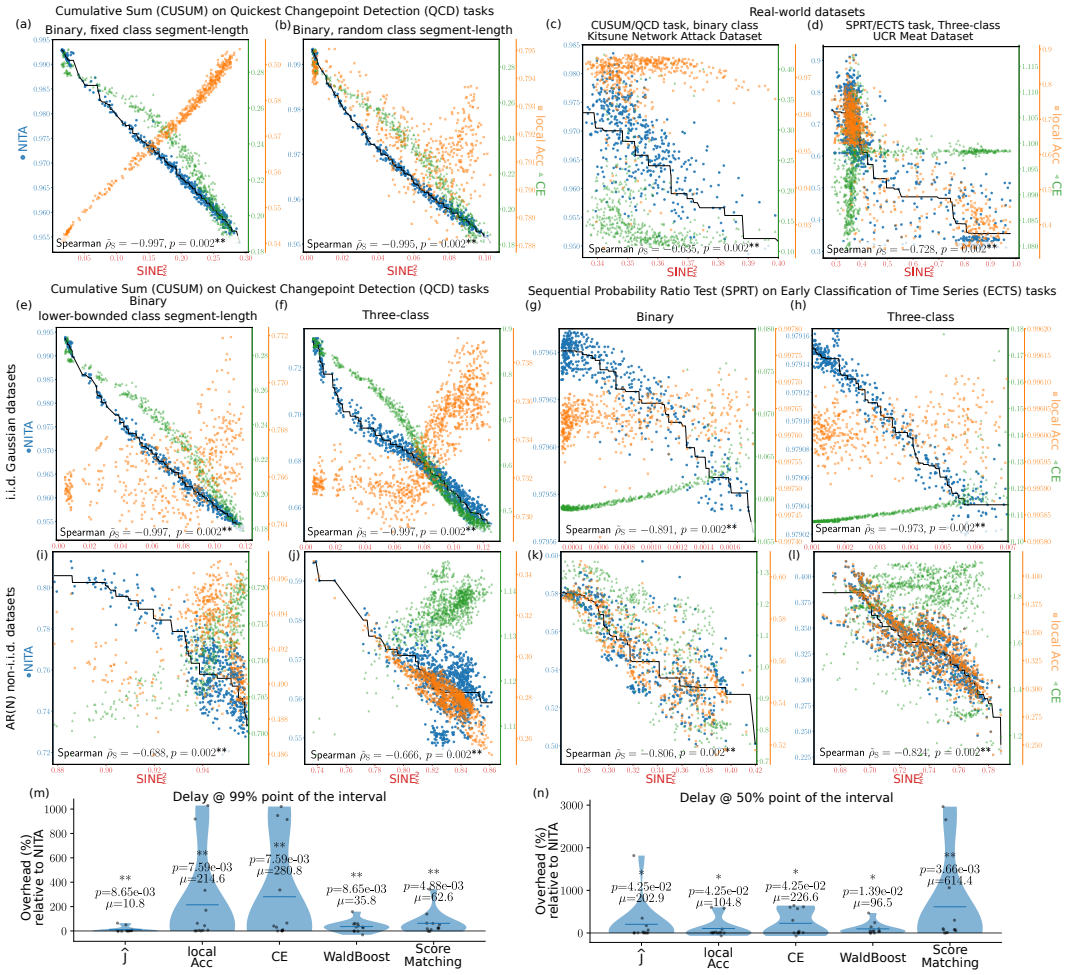


Figure 13: **Full-box NITA experiments.** Panels (a,b) repeat the settings of the main text experiments illustrated in Fig. 3(a,b). Panels (e–l) repeat Fig. 7(a–h), and panels (c,d) repeat Fig. 8(a,b). Experimental protocols are exactly the same except that full-box NITA, instead of high-threshold NITA, is used for checkpoint selection. Correlation tests are two-sided, while delay-overhead tests are one-sided with Holm–Bonferroni correction (Appendix E).

Table 6: **Full-box versus high-performance-tail SINE–NITA association.** Each entry $\widetilde{\rho}_S$ is the median of ten run-level Spearman correlations computed across checkpoints. The p-values are from two-sided Wilcoxon signed-rank tests of $H_1 : \mathrm{median}(\rho_S^{(r)}) \neq 0$. High-performance-tail coefficients agree with Table 1.

	Full-box NITA		high-threshold NITA	
Setting	$\widetilde{\rho}_S$	p	$\widetilde{\rho}_S$	p
i.i.d. QCD, binary, fixed segment length	-0.997	0.002	-0.992	0.002
i.i.d. QCD, binary, random segment length	-0.995	0.002	-0.990	0.002
i.i.d. QCD, binary, lower-bounded segment length	-0.997	0.002	-0.991	0.002
(d) i.i.d. QCD, three-class	-0.997	0.002	-0.994	0.002
i.i.d. SPRT, binary	-0.891	0.002	-0.716	0.002
i.i.d. SPRT, three-class	-0.973	0.002	-0.800	0.002
AR(N) QCD, binary	-0.688	0.002	-0.558	0.002
AR(N) QCD, three-class	-0.666	0.002	-0.540	0.002
AR(N) SPRT, binary	-0.806	0.002	-0.854	0.002
AR(N) SPRT, three-class	-0.824	0.002	-0.912	0.002
Kitsune, binary QCD	-0.635	0.002	-0.542	0.002
UCR Meat, three-class SPRT	-0.728	0.002	-0.702	0.002

sized by the asymptotic frontier analysis. Overall, the two altanertive evaluations give the same qualitative conclusions. In both evaluations, NITA is consistently antitone to SINE and selects checkpoints with favorable held-out sequential delay, as shown in Fig. 13.

D.12.2 Backbone ablation.

To test whether the SINE–NITA relation depends on the network architecture, we replace the TANDEMformer backbone (Ebihara et al., 2023) used in the main experiments with B2Bsqrt-LSTM (Ebihara et al., 2023). We recover the same qualitative anti-monotone relation between SINE and NITA, indicating that the surrogate behavior is not an architecture-specific artifact. This anti-monotonicity is statistically significant ($p = 0.002$; two-sided Wilcoxon signed-rank test on run-level Spearman coefficients, Appendix E.1).

D.13 Stress tests of Assumption 3.4

D.13.1 Temporally correlated nuisance

We replace the independent nuisance by a temporally correlated Gaussian perturbation in the binary i.i.d. QCD setting.

Use the same binary i.i.d. validation data and their exact one-step LLRs (not estimates). Perturb the increments, then accumulate them using the convention of Appendix A.1.

For each $\chi_{\mathrm{temp}} \in \{0, 0.25, 0.5, 0.75, 0.9, 1.0\}$, we generate a nuisance process $(\varepsilon^t)_{t=1}^T$ as follows. For $0 \leq \chi_{\mathrm{temp}} < 1$, we use a stationary AR(1)-type recursion

$$\varepsilon^1 \sim \mathcal{N}(0,1), \qquad \varepsilon^t = \chi_{\mathrm{temp}}\, \varepsilon^{t-1} + \sqrt{1-\chi_{\mathrm{temp}}^2}\, \eta^t, \qquad \eta^t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{182}$$

so that $\mathrm{Var}(\varepsilon^t) = 1$ for every $\chi_{\mathrm{temp}} < 1$. For the extreme case $\chi_{\mathrm{temp}} = 1$, which is not a stationary AR(1) process, we use a perfectly correlated-in-time stress test

$$\varepsilon^t \equiv \varepsilon^1 \qquad \text{for all } t. \tag{183}$$

Thus increasing χ_{temp} changes only the temporal dependence of the nuisance while keeping the marginal variance fixed. Using this nuisance, we define the perturbed one-step score

$$\hat{\lambda}^t_{\sigma,\chi_{\mathrm{temp}}} = \lambda^t + \sigma\, \varepsilon^t, \tag{184}$$

where $\sigma \geq 0$ is the global noise magnitude. We then reconstruct the perturbed cumulative LLRs by

$$\hat{\Lambda}^{1:t}_{\sigma,\chi_{\mathrm{temp}}} = \sum_{s=1}^{t} \hat{\lambda}^s_{\sigma,\chi_{\mathrm{temp}}}. \tag{185}$$

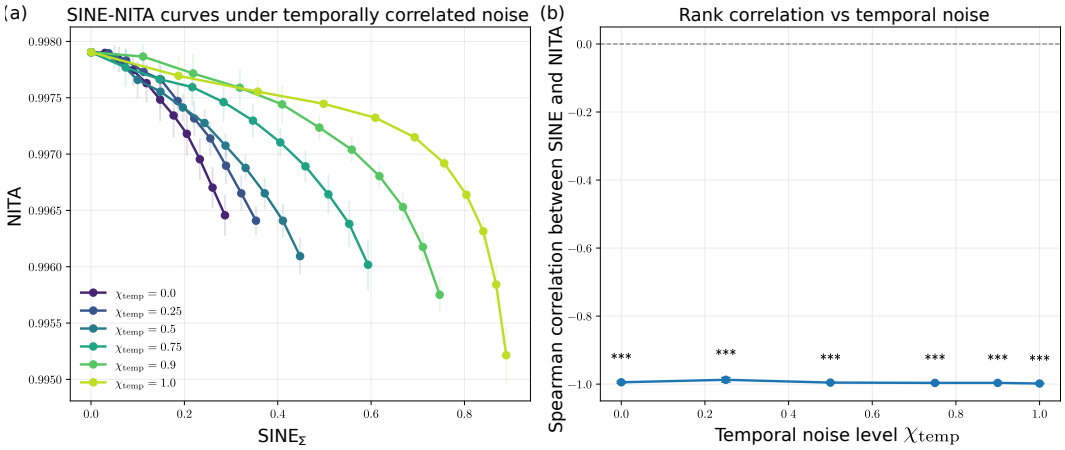


Figure 14: **Stress test under temporally correlated nuisance.** (a) SINE–NITA curves obtained by perturbing the true one-step LLR increments with Gaussian noise of varying global magnitude σ and temporal-correlation level χ_{temp}. Increasing χ_{temp} shifts the curve toward worse operating points and increases its curvature, but the antitone relation remains visually intact. (b) Spearman correlation between SINE and NITA across the σ-grid as a function of χ_{temp}, averaged over 10 repetitions. The correlation remains essentially -1 even under strong temporal dependence, suggesting that the temporal-independence part of Assumption 3.4 is stronger than necessary. The case $\chi_{\text{temp}} = 1$ is an extreme perfectly correlated stress test rather than a stationary AR(1) process.

Sweep σ over an eleven-point grid on $[0, 1.5]$, with ten repetitions per $(\chi_{\text{temp}}, \sigma)$. Compute cumulative-trajectory SINE against the true LLR and NITA with the same threshold-swept QCD protocol. Fig. 14 (a) connects the repetition means across σ at fixed χ_{temp}, while (b) reports the mean and standard deviation of within-repetition Spearman coefficients across the σ-grid. Figure 14 shows larger SINE, smaller NITA, and stronger curvature as correlation increases, while within-family correlations remain close to -1, including the perfectly correlated case.

D.13.2 Heteroskedastic and inverse-heteroskedastic nuisance

We vary nuisance magnitude with local LLR strength. Forward heteroskedasticity assigns more variance to stronger evidence, while inverse heteroskedasticity assigns more to weaker evidence.

Use the same validation data and exact increments as in Appendix D.13.1. From the true increment magnitude, form $q^t \in [0, 1]$, with larger values indicating stronger evidence; its normalization and clipping follow the accompanying notebook. Let $\chi_{\text{het}} \geq 0$ control the modulation:

$$m_t^{(+)}(\chi_{\text{het}}) := 1 + \chi_{\text{het}}\, q^t, \tag{186}$$

$$m_t^{(-)}(\chi_{\text{het}}) := 1 + \chi_{\text{het}}\, (1 - q^t), \tag{187}$$

corresponding respectively to forward and inverse heteroskedasticity. Given i.i.d. Gaussian base noise $\varepsilon^t \sim \mathcal{N}(0, 1)$, we define the variance-normalized perturbations

$$\widetilde{\varepsilon}_t^{(+)} := \frac{m_t^{(+)}(\chi_{\text{het}})\, \varepsilon^t}{\text{std}(m_t^{(+)}(\chi_{\text{het}})\, \varepsilon^t)}, \qquad \widetilde{\varepsilon}_t^{(-)} := \frac{m_t^{(-)}(\chi_{\text{het}})\, \varepsilon^t}{\text{std}(m_t^{(-)}(\chi_{\text{het}})\, \varepsilon^t)}, \tag{188}$$

so that changing χ_{het} redistributes noise across the sequence while keeping the global perturbation scale fixed. The perturbed one-step scores are then

$$\hat{\lambda}^t_{\sigma,\chi_{\text{het}},(\pm)} = \lambda^t + \sigma\, \widetilde{\varepsilon}_t^{(\pm)}, \tag{189}$$

where $\sigma \geq 0$ is the global noise magnitude, and the perturbed cumulative LLRs are reconstructed by

$$\hat{\Lambda}^{1:t}_{\sigma,\chi_{\text{het}},(\pm)} = \sum_{s=1}^{t} \hat{\lambda}^s_{\sigma,\chi_{\text{het}},(\pm)}. \tag{190}$$

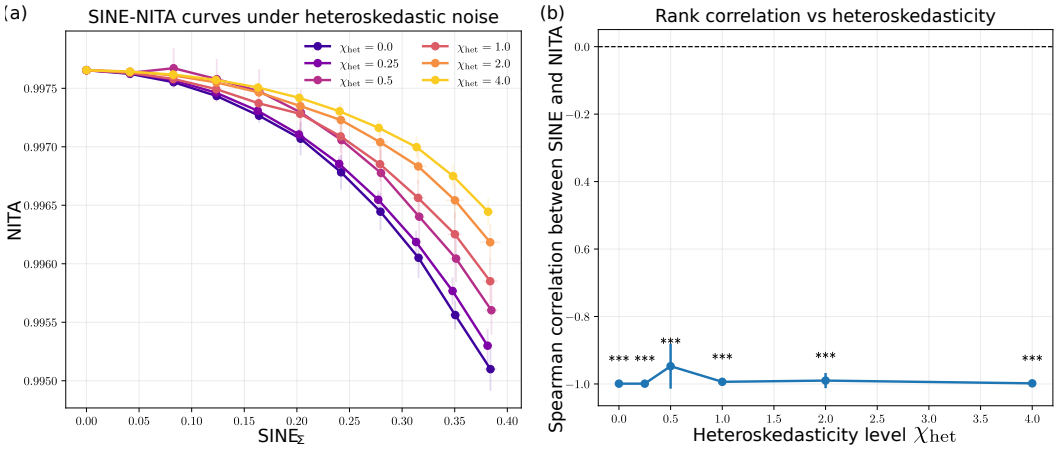


Figure 15: **Stress test under forward heteroskedastic nuisance.** (a) SINE–NITA curves obtained by perturbing the true one-step LLR increments with Gaussian noise of varying global magnitude σ and forward heteroskedasticity level χ_{het}, where larger local variance is assigned to stronger-evidence regions after global variance normalization. Increasing χ_{het} rotates the curve upward relative to the homoskedastic baseline $\chi_{\text{het}} = 0$, but the antitone relation remains visually intact. (b) Spearman correlation between SINE and NITA across the σ-grid as a function of χ_{het}, averaged over 10 repetitions. The correlation remains essentially -1, suggesting that the homoskedasticity part of Assumption 3.4 is stronger than necessary for preserving the SINE–NITA ordering in this structured perturbation family.

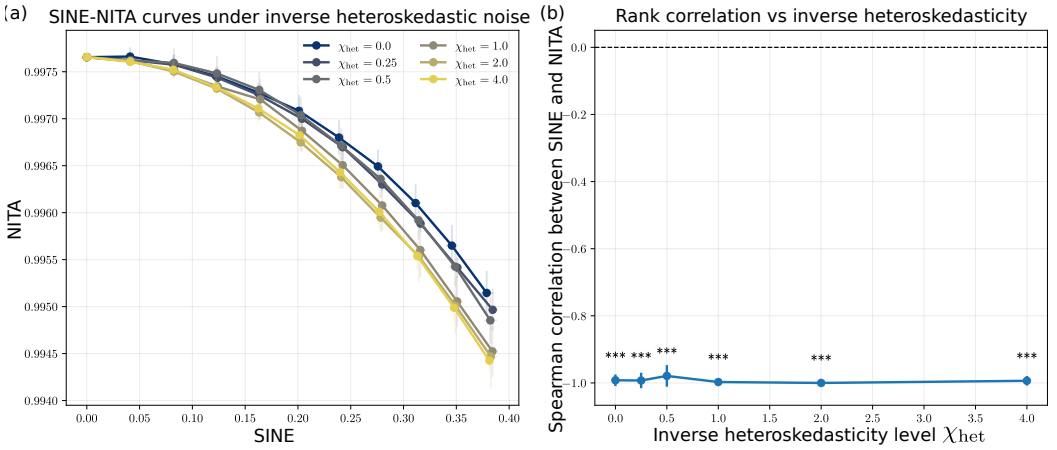


Figure 16: **Stress test under inverse heteroskedastic nuisance.** (a) SINE–NITA curves obtained by perturbing the true one-step LLR increments with Gaussian noise of varying global magnitude σ and inverse heteroskedasticity level χ_{het}, where larger local variance is assigned to weaker-evidence regions after global variance normalization. Increasing χ_{het} rotates the curve downward relative to the homoskedastic baseline $\chi_{\text{het}} = 0$, but again does not destroy the antitone relation. (b) Spearman correlation between SINE and NITA across the σ-grid as a function of χ_{het}, averaged over 10 repetitions. The correlation remains essentially -1, showing that even an adverse variance allocation changes the frontier geometry more than it changes the SINE–NITA ordering.

Use the same eleven-point σ-grid on $[0, 1.5]$ and ten repetitions per $(\chi_{\text{het}}, \sigma)$. Compute cumulative SINE and threshold-swept NITA as in the temporal test. In Figs. 15 and 16, (a) connect repetition means across σ, while (b) show the mean and standard deviation of within-run Spearman coefficients. Significance stars use two-sided run-level Wilcoxon tests as in Appendix E.1.

Figs. 15 and 16 show upward and downward changes relative to the homoskedastic curve, respectively, while correlations remain close to -1. Thus the variance allocation changes the curve under a common global noise budget without reversing its within-family ordering in the tested range.

Remark D.1 (Implications of the stress tests). Temporal correlation and signal-dependent variance allocation change the measured SINE-NITA curves, but strong negative rank associations across nuisance magnitude persist within all the three tested perturbation families. The observed relationship is therefore not confined to independent, homoskedastic nuisance. These controlled violations complement the checkpoint selection experiments in Section 4, which evaluate NITA without imposing the affine noise construction. The results support empirical applicability beyond Assumption 3.4 over the tested structures and ranges. A stronger theory under weaker assumptions remains a theoretical extension. Appendix F.1 discusses these complementary roles.

D.14 Empirical finite-sample stability of NITA

This controlled independent-unit study compares empirical uniform error and checkpoint-selection regret with the form of Theorem 3.8, separately from the pooled streaming evaluation in Appendix D.1.

Use full-box NITA ($x_0 = 0$) with the binary i.i.d. QCD checkpoints, training-derived scales, normalized grid, restriction horizons, and operating box fixed before validation. Two independent validation pools of 10^5 independently generated records separate reference evaluation from selection. For each coordinate and checkpoint–threshold pair, each record contributes one event-time/follow-up pair. The theorem-level interpretation additionally requires event-time/follow-up independence and positive population follow-up coverage at every time before each H_ι, as specified in Appendix A.3. This coverage condition allows an empty risk set in a realized subsample, but not a restriction horizon extending beyond the population follow-up support. The first pool gives each checkpoint's KM reference $\mathrm{NITA}_m^{\mathrm{ref}}$. Under these KM conditions, it is a Monte Carlo proxy for restricted-mean population grid-NITA. From the second pool, subsample n independent records, with

$$n \in \{100, 200, 500, 1000, 2000, 5000, 10000, 20000\},$$

recompute empirical NITA for every checkpoint, and select $\widehat{m}_n \in \arg\max_m \widehat{\mathrm{NITA}}_m^{(n)}$. Raw thresholds, horizons, and the operating box are never re-estimated from a subsample, while KM counts, survival, and restricted means are recomputed. Average over ten repetitions.

We report two empirical stability quantities, compared with the form of Theorem 3.8 under its assumptions. First, we measure

$$\mathrm{Err}(n) := \max_m \left| \widehat{\mathrm{NITA}}_m^{(n)} - \mathrm{NITA}_m^{\mathrm{ref}} \right|.$$

Second, to illustrate the selector consequence, we measure the reference-NITA regret

$$\mathrm{Regret}(n) := \mathrm{NITA}_{m^\star}^{\mathrm{ref}} - \mathrm{NITA}_{\widehat{m}_n}^{\mathrm{ref}}, \qquad m^\star \in \arg\max_m \mathrm{NITA}_m^{\mathrm{ref}}.$$

The regret quantity evaluates the checkpoint selected from the n-sample empirical NITA on an independent high-sample reference pool.

Figure 17 shows approximate $n^{-1/2}$ uniform-error decay and decreasing reference regret. The fitted curve is a rate guide, not the theorem's constant or a verification of the censoring and coverage assumptions. The high-sample reference remains a Monte Carlo approximation. These results concern the independent-unit design and do not certify the separate pooled streaming estimator. The complete-follow-up special case and KM assumptions are those in Appendix A.3.

E Statistical tests

E.1 Two-sided tests on run-level Spearman coefficients

For each experimental condition, we trained the model with R independent random seeds. Within each run r, we computed Spearman's rank correlation coefficient ρ_r (Spearman, 1904) across saved checkpoints between cumulative-trajectory SINE_Σ^2 and each validation criterion: high-threshold

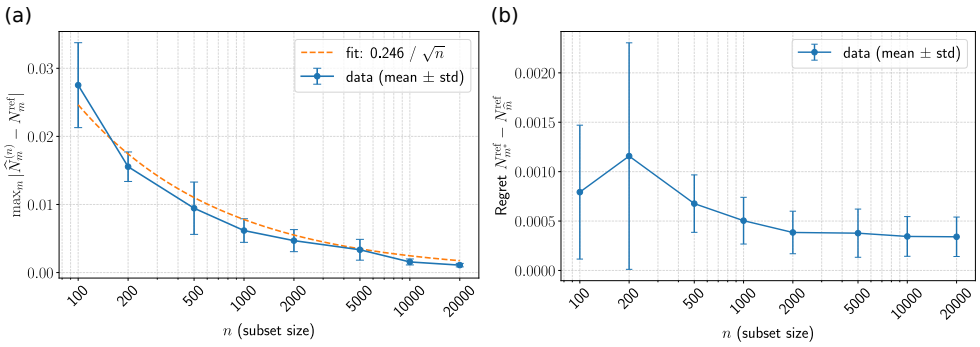


Figure 17: **Empirical finite-sample stability of full-box NITA**, compared with the error and selection-regret quantities in Theorem 3.8. This controlled independent-unit study is separate from the pooled streaming benchmark; the additional KM conditions for the theorem comparison are stated in Appendix D.14. We fix the learned checkpoints, training-derived scales, normalized grid, operating box ($x_0 = 0$), and KM restriction horizons before validation. A high-sample reference $\mathrm{NITA}_m^{\mathrm{ref}}$ is computed for each checkpoint on an independent validation pool of 10^5 records. Empirical NITA $\widehat{\mathrm{NITA}}_m^{(n)}$ is then computed from n independent records sampled from a second independent validation pool. Each contributes one event-time/follow-up pair per coordinate and checkpoint–threshold pair. Error bars show mean $\pm$ one standard deviation over 10 repetitions. (a) The uniform empirical-NITA error $\max_m |\widehat{\mathrm{NITA}}_m^{(n)} - \mathrm{NITA}_m^{\mathrm{ref}}|$ decays approximately as $n^{-1/2}$, consistent with the concentration rate in Theorem 3.8. The dashed curve is a fitted rate guide $0.246/\sqrt{n}$, not the conservative concentration bound. (b) The reference-NITA regret is $\mathrm{NITA}_{m^\star}^{\mathrm{ref}} - \mathrm{NITA}_{\widehat{m}_n}^{\mathrm{ref}}$. The regret curve is not expected to follow an exact $n^{-1/2}$ law because checkpoint selection is an argmax over a discrete and potentially near-tied checkpoint set. Theorem 3.8 bounds regret through the uniform NITA error rather than asserting equality of rates.

NITA, CE, or local Acc. While synthetic data provide the true LLR reference, real-world data use the training-fitted KDE pseudo-LLR reference described in Appendix D.5. We report the median run-level coefficient for each criterion in Table 1.

Because checkpoints within a run are temporally correlated, runs (seeds), rather than individual checkpoints, are the independent replicates for inference. All tests on run-level Spearman coefficients are two-sided. We apply the Wilcoxon signed-rank test (Wilcoxon, 1945) against a distribution centered at zero. The hypotheses can be written as

$$H_0 : \mathrm{median}(\rho_r) = 0, \qquad H_1 : \mathrm{median}(\rho_r) \neq 0.$$

The signed-rank test is not a distribution-free test of the median without the symmetry assumption. The accompanying two-sided sign test assesses the balance of positive and negative run-level coefficients without requiring symmetry.

For NITA and local Acc, which are maximized, negative correlation with SINE supports the desired ordering, whereas positive correlation indicates the opposite ordering. For CE, which is minimized, positive correlation with SINE supports the desired ordering, whereas negative correlation indicates the opposite ordering. Thus, a significant association alone does not establish that a criterion selects checkpoints in the direction of smaller SINE. The same two-sided testing convention applies to correlation analyses in the full-box ablations and noise stress tests.

E.2 One-sided median test of delay overheads

Let $d \in \{1, \ldots, D\}$ index datasets ($D = 12$), $l \in \{1, \ldots, L\}$ index evaluation criteria/levels, m index models, and r index repeated runs. Denote by $x_{d,r,m,l} \geq 0$ the scalar delay value of model m on dataset d at level l in repetition r: the KM-based operating coordinate for correct-class confirmation lag for CUSUM datasets (including the finite-stream convention in Appendix D.1.1) and the mean hitting time (MHT) for SPRT datasets. We take model b (`nita`) as the baseline.

For QCD, the zero-indexed lag is the first correct-class all-rivals confirmation within the post-change interval minus the labeled changepoint. It can be zero, including when the row already exceeds the threshold at that interval's start. Intervals without correct-class confirmation contribute censored follow-up to KM, as in Appendix D.1. To avoid a zero denominator without removing any run-level pair, we use a task-dependent one-step stabilization:

$$s_d := \begin{cases} 1, & \text{if dataset } d \text{ is a QCD/CUSUM task,} \\ 0, & \text{if dataset } d \text{ is an SPRT task.} \end{cases}$$

For each non-baseline model $m \neq b$, dataset d, repetition r, and level l, we compute

$$\Delta_{d,r,m,l} = \frac{(x_{d,r,m,l} + s_d) - (x_{d,r,b,l} + s_d)}{x_{d,r,b,l} + s_d} = \frac{x_{d,r,m,l} - x_{d,r,b,l}}{x_{d,r,b,l} + s_d}.$$

Positive values ($\Delta_{d,r,m,l} > 0$) indicate slower baselines than NITA. We aggregate within each dataset by the median across all available runs: $\widetilde{\Delta}_{d,m,l} := \mathrm{median}\{\Delta_{d,r,m,l} : r \in \mathcal{R}_{d,m,l}\}$, where $\mathcal{R}_{d,m,l}$ is the set of available repetitions for dataset d, model m, and level l. We then form the dataset-level sample $\widetilde{\mathcal{S}}_{m,l} = \{\widetilde{\Delta}_{d,m,l}\}_{d \in \mathcal{D}_{m,l}}$, where $\mathcal{D}_{m,l}$ is the set of datasets with available run-level results.

For each level l and model $m \neq b$, we apply a one-sided Wilcoxon signed-rank test to $\widetilde{\mathcal{S}}_{m,l}$ to test

$$H_0 : \mathrm{median}(\widetilde{\Delta}_{m,l}) = 0 \quad \text{vs.} \quad H_1 : \mathrm{median}(\widetilde{\Delta}_{m,l}) > 0.$$

The median interpretation assumes a symmetric location model for the dataset-level differences. The one-sided alternative tests whether the baseline is slower than NITA, corresponding to positive median overhead. We report the number of contributing datasets $|\mathcal{D}_{m,l}|$, the sample median of $\widetilde{\mathcal{S}}_{m,l}$, the raw p-value, and the Holm–Bonferroni adjusted p-value.

Violin plots in the main paper display the pooled run-level stabilized relative differences for visualization, but inferential p-values are computed from the dataset-level summaries above so that each dataset contributes equally.

F Supplementary Discussion

F.1 The affine-noise model: theoretical role and empirical evidence

Role of the affine-noise model. Assumption 3.4 isolates a mechanism connecting projective estimation error to sequential operating performance. In the family $g_\sigma = \alpha\lambda + \sigma\xi$, the signal and nuisance laws are held fixed while the nuisance magnitude varies. Population orthogonality gives an explicit expression for SINE (Eq. (13)), independence factors the pre-change exponential moment in Eq. (42) (Lemma B.12), and centering leaves the post-change drift unchanged. Theorem 3.6 combines these properties with the high-threshold conditions of Assumption 3.5 to establish SINE-NITA antitone on common evaluation regions. Therefore, the model provides a tractable explanation under sufficient conditions, rather than a distributional requirement that every neural checkpoint must satisfy.

What the stress tests establish. Appendix D.13 directly examines whether the observed SINE-NITA antitone persists when the nuisance departs from the independent, homoskedastic construction. Figs. 14 varies temporal correlation while keeping marginal nuisance variance fixed, including a perfectly correlated samples. Fig. 15 and 16 redistribute nuisance variance toward stronger and weaker evidence regions, respectively, while normalizing the global perturbation scale. These experiments thus change the temporal organization or signal-dependent variance allocation of the nuisance, rather than altering its magnitude only. Within each tested structure, the reported rank associations across nuisance magnitude remain strongly negative, although the SINE-NITA curves change shape. Therefore, temporal independence and constant local nuisance variance are not prerequisites for the observed relationship in these experiments.

Relation to the checkpoint selection experiments. The controlled perturbations isolate sensitivity to specified violations of the assumptions, while the checkpoint selection experiments assess practical

selection without imposing a common affine noise family. Table 1 reports consistently negative SINE-NITA associations, and Fig. 4 evaluates the held-out performance of the selected checkpoints. The full-box and backbone ablations further examine whether the observed relationship depends on the high-threshold restriction or one particular architecture. These experiments provide complementary evidence: the perturbation studies examine particular assumptions, whereas the checkpoint studies evaluate the proposed validation criterion on trained estimators. Together, they support empirical applicability beyond the principal theoretical model. Identifying weaker sufficient conditions that explain this broader behavior is a direction for extending the theory.

Which results depend on this model? Assumption 3.4 supports the SINE-NITA antitone ordering: it is not required to define or compute NITA. The exact decision equivalence under threshold rescaling in Proposition 2.1 does not require the affine-noise model. Theorem 3.8 separately bounds checkpoint selection regret relative to the best population grid-NITA under its stated validation sampling, and when applicable, censoring assumptions, without requiring affine noise or high threshold asymptotics. Therefore, applying NITA does not require fitting an affine noise model or estimating its nuisance distribution. These results address distinct aspects of the framework: the learning target, its connection to operating performance, and finite-validation selection.

Likelihood process motivation. Two likelihood-process decompositions motivate the signal-plus-noise picture: conditional LLRs have centered martingale residuals, and constant-drift Gaussian diffusions admit an exact Brownian LLR representation (Ito, 1944; Girsanov, 1960; Øksendal, 2014). Note that neither construction verifies independence of a learned residual, which remains an assumption of the main theorem. The continuous-time example below is motivational, not an ingredient of the discrete-time proof.

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P}_1)$ be a filtered probability space, and suppose that under two regimes $y \in \{0, 1\}$ the conditional densities of X_t given $\mathcal{F}_{t-1}$ are

$$p_y(\cdot \mid \mathcal{F}_{t-1}).$$

Define the one-step conditional log-likelihood-ratio increment by

$$\lambda^t := \log \frac{p_1(X_t \mid \mathcal{F}_{t-1})}{p_0(X_t \mid \mathcal{F}_{t-1})}. \tag{191}$$

Proposition F.1 (Predictable information drift plus martingale residual)**.** *Assume* $p_1(\cdot \mid \mathcal{F}_{t-1}) \ll p_0(\cdot \mid \mathcal{F}_{t-1})$ *and vice versa almost surely. Then:*

1. *Under* $\mathbb{P}_1$,

$$\lambda^t = I_t^{10} + \Delta M_t, \qquad I_t^{10} := \mathbb{E}_1[\lambda^t \mid \mathcal{F}_{t-1}] = D_{\mathrm{KL}}\Big(p_1(\cdot \mid \mathcal{F}_{t-1}) \,\Big\|\, p_0(\cdot \mid \mathcal{F}_{t-1})\Big), \tag{192}$$

where (ΔM_t) *is a martingale-difference sequence:*

$$\mathbb{E}_1[\Delta M_t \mid \mathcal{F}_{t-1}] = 0.$$

2. *Under* $\mathbb{P}_0$,

$$\lambda^t = -I_t^{01} + \Delta \widetilde{M}_t, \qquad I_t^{01} := D_{\mathrm{KL}}\Big(p_0(\cdot \mid \mathcal{F}_{t-1}) \,\Big\|\, p_1(\cdot \mid \mathcal{F}_{t-1})\Big), \tag{193}$$

where $(\Delta \widetilde{M}_t)$ *is a martingale-difference sequence under* $\mathbb{P}_0$.

Consequently, the accumulated LLR process

$$\Lambda_t := \sum_{s=1}^{t} \lambda^s$$

admits the decomposition

$$\Lambda_t = \sum_{s=1}^{t} I_s^{10} + M_t \quad \textit{under } \mathbb{P}_1, \qquad \Lambda_t = -\sum_{s=1}^{t} I_s^{01} + \widetilde{M}_t \quad \textit{under } \mathbb{P}_0,$$

for martingales (M_t) *and* $(\widetilde{M}_t)$.

Proof. Conditioning on $\mathcal{F}_{t-1}$ gives $\mathbb{E}_1[\lambda^t \mid \mathcal{F}_{t-1}] = I_t^{10}$ and $\mathbb{E}_0[\lambda^t \mid \mathcal{F}_{t-1}] = -I_t^{01}$ by the conditional-density definitions of KL divergence. Subtract each conditional mean to obtain the stated martingale differences, then sum their increments. □

These decompositions motivate centered fluctuations. The dependent-score analysis is in Appendix B.9.

Diffusion special case via Girsanov's theorem. The most transparent example is a drift-shifted Gaussian Itô diffusion. Let $(X_t)_{t\geq 0}$ be real-valued and suppose

$$H_0 \text{ (under } \mathbb{P}_0) : \qquad dX_t = \mu_0\, dt + \sigma_X\, dW_t^{(0)}, \tag{194}$$

$$H_1 \text{ (under } \mathbb{P}_1) : \qquad dX_t = \mu_1\, dt + \sigma_X\, dW_t^{(1)}, \tag{195}$$

where $\mu_0 \neq \mu_1$ and $\sigma_X > 0$ is shared. Let

$$\theta_{\text{eff}} := \frac{\mu_1 - \mu_0}{\sigma_X}, \qquad \Lambda_t := \log \left.\frac{d\mathbb{P}_1}{d\mathbb{P}_0}\right|_{\mathcal{F}_t}.$$

Proposition F.2 (Canonical drift-plus-Brownian form of the diffusion LLR)**.** *Under the diffusion model above, the accumulated LLR satisfies*

$$d\Lambda_t = -\frac{1}{2}\theta_{\text{eff}}^2\, dt + \theta_{\text{eff}}\, dW_t^{(0)} \qquad \textit{under } \mathbb{P}_0, \tag{196}$$

and

$$d\Lambda_t = \frac{1}{2}\theta_{\text{eff}}^2\, dt + \theta_{\text{eff}}\, dW_t^{(1)} \qquad \textit{under } \mathbb{P}_1. \tag{197}$$

Thus the true LLR process decomposes exactly into a constant information drift plus a martingale noise term; in the diffusion case, that noise term has independent increments.

Proof. Girsanov's exponential likelihood gives $\Lambda_t = \theta_{\text{eff}} W_t^{(0)} - \frac{1}{2}\theta_{\text{eff}}^2 t$, which yields Eq. (196). Under $\mathbb{P}_1$, $dW_t^{(0)} = \theta_{\text{eff}}\, dt + dW_t^{(1)}$, and substituting gives Eq. (197). □

The diffusion identity concerns the true LLR. The learned-score independence assumption and its structured empirical violations remain distinct (Appendices B.4 and D.13).

F.2 High-threshold and full-box NITA: theoretical motivation and practical use

Why use high-threshold NITA in the main experiments? Theorem 3.6 establishes the SINE–NITA ordering on sufficiently high, common non-saturated comparison regions. We therefore use high-threshold NITA in the main experiments to focus checkpoint selection on the operating region motivated by this theory. The empirical protocol in Appendix D.12.1 implements this focus through the upper half of the attainable ARL or terminal-accuracy interval measured by a training-only pilot. The corresponding x_0, threshold subset, and common normalizer are frozen before validation. The full-box ablation then tests whether this restriction is needed for effective checkpoint selection, while keeping the training runs, candidate checkpoints, threshold grid, and operating-coordinate estimators unchanged.

Full-box NITA simplifies validation. Full-box NITA integrates the entire operating envelope within $[0, X^*] \times [0, Y^*]$, using the same box for all checkpoints. It requires no additional identification of a common high-threshold subregion. Locating a suitable high-threshold region can require preliminary sweeps before the final comparison region is fixed. Full-box NITA avoids this additional region-selection step while preserving the common normalization needed for meaningful checkpoint comparison. Its practical appeal is therefore not merely that it includes more operating points. It also reduces the evaluation choices required to use NITA as a validation criterion.

Evidence beyond the high-threshold region. The full-box ablation evaluates whether including the remaining operating range preserves the observed SINE–NITA association and checkpoint-selection utility. Table 6 shows statistically significant negative SINE–NITA associations for both variants in all twelve settings. Figure 13 reports the corresponding full-box experiments. These results support full-box NITA as a practical alternative without first locating a high-threshold comparison region.

A complementary result is the improvement at the 50% test operating point in Fig. 4c. Here, checkpoints selected using high-threshold NITA also perform well at a moderate operating point. This tests transfer across operating regions, whereas the full-box ablation changes the checkpoint-selection criterion itself. The pointwise test comparisons provide evidence beyond area ordering, which summarizes performance rather than providing guarantee at every operating point.

Implications for theory and practice. Together, these findings suggest that the high-threshold restriction is a sufficient condition for the present SINE–NITA theorem rather than a prerequisite for useful empirical checkpoint selection. Theorem 3.8 separately establishes finite-validation reliability for a fixed grid-NITA criterion under its sampling assumptions, without requiring high-threshold asymptotics. An important next step is to establish conditions for SINE–NITA ordering on finite-threshold and full-box regions, together with performance bounds at specified operating points. Such results would explain more fully when the simpler full-box protocol preserves the relationship between projective estimation error and sequential performance.

F.3 Why LSEL does not resolve the CE-validation issue?

For the consistent pairwise LLR matrices used in our multiclass experiments, the log-sum-exp loss (LSEL) is introduced in Miyagawa & Ebihara (2021). However, it is essentially CE, written in LLR coordinates. The identity below explains why replacing CE by LSEL does not change the distinction between training loss and sequential validation results.

Consistent pairwise-difference matrices. For a fixed sample prefix $X^{(1,t)}$, let

$$\widehat{\mathbf{\Lambda}}^{1:t} = \left(\widehat{\Lambda}^{1:t}_{k\ell}\right)_{k,\ell\in[K]} \in \mathbb{R}^{K\times K}$$

denote an estimated accumulated LLR matrix. In (Miyagawa & Ebihara, 2021), the LLR matrix is assumed to satisfy not only antisymmetry,

$$\widehat{\Lambda}^{1:t}_{k\ell} = -\widehat{\Lambda}^{1:t}_{\ell k},$$

but also the transitivity constraint

$$\widehat{\Lambda}^{1:t}_{k\ell} + \widehat{\Lambda}^{1:t}_{\ell m} = \widehat{\Lambda}^{1:t}_{km} \qquad \forall k,\ell,m \in [K].$$

Equivalently, there exists a score vector $s^{(t)} = (s^{(t)}_1, \dots, s^{(t)}_K) \in \mathbb{R}^K$ such that

$$\widehat{\Lambda}^{1:t}_{k\ell} = s^{(t)}_k - s^{(t)}_\ell \qquad \forall k,\ell \in [K]. \tag{198}$$

Proposition F.3 (LSEL is equivalent to class-balanced CE on a consistent LLR matrix)**.** *Fix a prefix $X^{(1,t)}$ and a true class label $y \in [K]$. Assume the estimated LLR matrix $\widehat{\mathbf{\Lambda}}^{1:t}$ satisfies Eq. (198). Define the row-mean logits*

$$r^{(t)}_k := \frac{1}{K}\sum_{\ell=1}^{K} \widehat{\Lambda}^{1:t}_{k\ell}, \qquad k \in [K]. \tag{199}$$

Then LSEL,

$$\ell_{\mathrm{LSEL}}(y, \widehat{\mathbf{\Lambda}}^{1:t}) := \log\left(1 + \sum_{\ell\neq y} e^{-\widehat{\Lambda}^{1:t}_{y\ell}}\right), \tag{200}$$

is exactly equal to multiclass cross-entropy on the posterior obtained from the row-mean logits:

$$\ell_{\mathrm{LSEL}}(y, \widehat{\mathbf{\Lambda}}^{1:t}) = -\log\left(\frac{e^{r^{(t)}_y}}{\sum_{k=1}^{K} e^{r^{(t)}_k}}\right). \tag{201}$$

Thus, after averaging over classes, prefixes, and samples with the same outer weighting as in the empirical LSEL, LSEL is exactly a class-balanced CE objective.

Proof. Under Eq. (198), $\ell_{\mathrm{LSEL}} = -s_y^{(t)} + \log \sum_k e^{s_k^{(t)}}$. The row means satisfy $r_k^{(t)} = s_k^{(t)} - \bar{s}^{(t)}$, where $\bar{s}^{(t)} = K^{-1} \sum_\ell s_\ell^{(t)}$. Softmax is unchanged by this common shift, proving Eq. (201). Applying identical outer weights preserves the equality. □

Implication for checkpoint selection. Our multiclass LLR estimators satisfy Eq. (198). Thus LSEL can remain a valid training loss. However, its equivalence to CE invites the same problem: because CE/LSEL are per-time-step, local evaluation criteria, they does not establish a checkpoint-ordering guarantee for sequential test performance, under finite training data and computational resources.

F.4 Scope of the positive-projective learning target

P_+LLR provides a decision-sufficient learning target for calibrated SPRT and CUSUM. Recovering the true LLR up to one global positive scale preserves stopping times and decisions when the thresholds are rescaled by the same factor (Proposition 2.1). Removing this scale from the estimand resolves the incompatibility between scale-invariant operating performance and exact-scale LLR estimation error (Corollary 2.2).

Validation compares performance across a common threshold sweep, rather than relying on agreement at one operating point. For checkpoint m, we use $h_{m,j} = s_m \theta_j$, where θ_j belongs to the shared normalized grid and $s_m > 0$ is fixed from training data before validation. This construction allows one positive rescaling throughout the sweep. A nonlinear transformation, such as cubing the cumulative score and its SPRT boundaries, instead requires a nonlinear threshold correspondence and does not preserve the same indexed sweep.

The connection from operating performance to projective error comes from the stated model assumptions. In the binary i.i.d. affine independent-noise family, Theorem 3.6 shows that increasing SINE decreases the effective slope of the population operating frontier. The full frontier therefore distinguishes nuisance levels within this family, and smaller SINE yields larger NITA on suitable common non-saturated high-threshold regions. This establishes a conditional relationship between observable performance and the P_+LLR estimation target.

For finite validation data, Theorem 3.8 separately bounds checkpoint-selection error relative to the best population grid-NITA among the available candidates under its sampling assumptions. The candidates does not necessarily recover the true LLR exactly. This is the practical role of the formulation: to select among imperfect checkpoints when data and network capacity are limited. In our P_+LLR learning theory, P_+LLR defines the learning target, SINE measures its oracle error, and NITA provides an observable validation criterion without requiring the true LLR.

F.5 Streaming checkpoint selection and transfer to first-alarm CUSUM

Our study is motivated by the practical problem of selecting learned LLR checkpoints for sequential tests on finite streams with multiple changepoints. The main QCD benchmark in Appendix D.1 processes each record continuously, carrying network state and evidence history across regime changes. Labels identify scoring intervals without resetting the scores, and KM-based operating coordinates incorporate right-censored observations using the finite-stream convention in Appendix D.1.1. This design enables checkpoint comparison on available labeled streams without reconstructing separate long single-regime evaluation sequences.

Controlled reliability and practical applicability. The evaluations address complementary questions. Appendix D.14 examines how finite validation affects NITA estimation and checkpoint selection using independent units and frozen evaluation choices. Its interpretation through Theorem 3.8 concerns population restricted-mean grid-NITA under the stated censoring and follow-up coverage conditions. In contrast, the main streaming benchmark asks whether NITA selects useful checkpoints under practical observation constraints, including inherited state, pooled transitions, and the specified flat-tail convention. The SINE associations and held-out checkpoint comparisons assess empirical usefulness under these constraints, rather than extending the estimation guarantee to unsupported event-time tails. This separation examines both finite-validation reliability and applicability beyond the sufficient conditions of the analysis. it does not restrict all experiments to one theoretical observation model.

After a change, new evidence can favor the current regime while the accumulated LLR still reflects the preceding regime. Predictions based on that accumulated evidence can therefore incur transient errors even when the LLR trajectory is responding correctly to the change. Local Acc and CE evaluate predictive correctness and confidence at individual times, which does not necessarily rank checkpoints by their sequential test performance. NITA instead evaluates the resulting delay vs. false-alarm tradeoff under the specified protocol. The streaming benchmark thus tests the practical value of validating learned LLR evidence through its sequential operating curve.

Exact recovery of $\mathrm{P}_+\mathrm{LLR}$, with corresponding threshold rescaling, preserves the stopping times and decisions of the true-LLR test (Proposition 2.1). It consequently preserves that test's optimality guarantees under their stated conditions. The invariance is pathwise: it does not require a single-changepoint setting, high-threshold asymptotics, or the affine independent noise model. Appendix B.1 gives the corresponding invariance for the all-rivals rule. Exact $\mathrm{P}_+\mathrm{LLR}$ recovery therefore inherits any optimality guarantee established for the corresponding true-LLR test. Our experiments address the practical question of selecting among imperfect learned checkpoints using finite streaming validation data. This question is distinct from re-establishing classical SPRT/CUSUM optimality experimentally.

To complement the streaming benchmark, we evaluate the original frozen checkpoint selections using the classical binary CUSUM rule. Each run stops at its first alarm, and performance is measured by ARL under the no-change law and immediate-change ADD_1, with thresholds calibrated on validation data. This controlled evaluation uses the stopping rule and operating criteria of the main theory to test whether the checkpoint-selection benefit extends beyond the main streaming experiments.

Frozen selections and validation-calibrated thresholds. We freeze the original checkpoints selected by NITA, $\widehat{J}$, local Acc, and CE (10 checkpoints per selector). In one binary QCD setting, each checkpoint supplies learned increments to Page's zero-initialized recursion in Eq. (3). Evaluation ends at the first alarm, rather than searching a labeled post-change interval for a later confirmation. The operating quantities are $\mathrm{ARL} = \mathbb{E}_\infty[\tau]$ and $\mathrm{ADD}_1 = \mathbb{E}_{\nu=1}[\tau - 1]$, as defined in Eq. (4). For each checkpoint, thresholds are calibrated on validation data to target ARLs of 15, 20, and 25, then held fixed for test evaluation. Neither checkpoint selection nor threshold calibration uses the test data.

Table 7: **Transfer to classical binary first-alarm CUSUM.**

	Target ARL 15		Target ARL 20		Target ARL 25	
Selector	ARL	ADD_1	ARL	ADD_1	ARL	ADD_1
NITA	14.986	0.162	20.026	0.230	24.980	0.305
$\widehat{J}$	14.940	0.165	20.012	0.232	24.896	0.306
Local Acc	15.208	0.879	20.013	1.307	24.918	1.843
CE	15.193	0.875	20.039	1.308	24.962	1.844

Results and interpretation. Table 7 shows average test ARL and ADD_1 results on the 10 trials. Original checkpoints are frozen, and thresholds are calibrated on validation data and applied unchanged to test data. Central test estimates are shown in observation steps. Higher ARL and lower ADD_1 are preferred. For NITA the largest discrepancy is 0.13%. At these closely matched false-alarm levels, NITA's reported delays are smallest among the selectors, and especially 81.5–83.5% lower than those of CE and local Acc.

The result supports transfer of the selection benefit by NITA beyond the confirmation benchmark to binary first-alarm CUSUM, with thresholds calibrated before test evaluation. While Theorem 3.6 explains the projective surrogate relationship under its assumptions, this experiment assesses its practical checkpoint-selection utility. This transfer evidence is complementary to the independent-unit reliability study, not an extension of Theorem 3.8 to the pooled streaming estimator.

F.6 Novelty, impact, and the future of $\mathrm{P}_+\mathrm{LLR}$ learning

Our contribution is to establish the *relationship* between the learning target and the operating performance of sequential tests, enabling principled checkpoint selection. From a practitioner's perspective, selecting a checkpoint after training is an essential step toward product deployment, yet

before this work there was no principled way to validate learned LLRs across a wide range of datasets and tasks. Our theory establishes when an observable operating-curve metric can serve as a validation surrogate for LLR estimation error. Historically, operating-curve performance was not thought to reveal LLR estimation quality because of its scale invariance, as Moustakides & Basioti (2019) illustrate. With a proper evaluation protocol, we turn the operating curve into a proper validation metric for monitoring estimation error by shifting the estimand from the true LLR to $\mathrm{P}_{+}\mathrm{LLR}$.

The main novelty does not reside in *each ingredient* of the theory. For example, NITA is based on a conventional operating curve, which is related to hypervolume (Zitzler et al., 2007). SINE uses scale-invariant projection, a construction also underlying SI-SDR (Roux et al., 2018). There are also task-specific precedents for selecting models through sequential performance. Wang et al. (2016) use validation accuracy–earliness area for early stopping. Achenchabe et al. (2021) combine misclassification and decision-delay costs for early classification and use average decision cost for validation-based hyperparameter selection. In QCD, Cooper & Meyn (2024) analyze mismatched CUSUM optimization and recognize score–threshold rescaling invariance. Thus, each individual ingredient has a precedent. Accordingly, we do not claim strong novelty in the *form* of each ingredient.

The theoretical findings established in this paper is summarized as the following relationship. Corollary 2.2 identifies the incompatibility between exact-scale LLR error and scale invariant operating curve. Formulating the target as $\mathrm{P}_{+}\mathrm{LLR}$ removes this incompatibility, and Theorem 3.6 establishes the SINE-NITA antitone under explicit sufficient conditions. Corollary 3.7 quantifies the local effective slope loss, while Theorem 3.8 separately establishes finite validation reliability for the grid-NITA criterion. Thus, the contribution is not simply to use an operating curve area, but to establish when such an observable criterion is ordered with projective estimation error and how accurately it can support checkpoint selection.

The empirical contribution is complementary. Learned checkpoint comparisons demonstrate practical selection benefits, while the structured noise, full-box, and backbone experiments examine whether the observed relationship persists beyond particular conditions of the main analysis. The preserved ordering under these changes provides evidence that applicability extends beyond a sufficient-condition model. Future work can use these findings to identify weaker sufficient conditions and develop finite threshold ordering results, extending the explanation of behavior already observed in the present experiments.