DIRECTIONAL INFORMATION IN TEMPORALLY AGGREGATED HAWKES PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Timestamp aggregation removes event order but can preserve directional information in event counts. We characterize this distinction in a stationary, mean-matched, feed-forward Hawkes family observed through independent windows. For multiple bins, we derive the exact first-order directional score and its finite-window information. For kernels with finite mean lag μ_h, information per unit exposure decays as $2\mu_h^2/\Delta^3$ when the long-window limit precedes the coarse-resolution limit. With one bin, the first-order score vanishes, but a nonzero second-order tangent combines excess dispersion with a mixed third-degree count moment. We derive the exact Gaussian testing limit at excitation strength $a_N = \gamma N^{-1/4}$ and a statistic attaining it. The mixed term supplies a calculable fraction of local information absent from dispersion alone. Both tests retain local efficiency after estimating the common rate. We characterize which count summaries preserve the tangents and which discard their distinctions. Paired count-law experiments quantify the mixed term's contribution and show slow finite-sample convergence. Repeated calibration experiments assess the grid-specific assertion control given by test inversion. Misspecification experiments show substantial failures outside that grid. The results separate temporal ordering information, higher-order count information, and the scope of calibrated reporting.

1 INTRODUCTION

Temporal aggregation removes within-bin event order without necessarily removing count dependence. A model can therefore predict counts accurately while assigning their dependence to a lag kernel that the observations do not determine. Accepting tied timestamps does not resolve this inferential problem. The relevant questions concern the contrasts retained by the observation map, the signal strength needed to recover them, and the uncertainty of the resulting conclusions.

Aggregation does not generally destroy Hawkes identifiability. Bayesian inference establishes identifiability under explicit specifications (Zhou and Papadogeorgou, 2026), and spectral methods estimate parameters consistently from counts (Cheysson and Lang, 2022). Integrated cumulants can recover interaction structure without estimating complete kernel shapes (Achab et al., 2017). Coarse-grained Hawkes models preserve intra-bin second-order structure (Koyama, 2025). These results rule out a blanket claim of directional impossibility. The question is instead the *order and magnitude* of directional information in a specified aggregated experiment.

We study a stationary feed-forward family in which one Poisson component excites another. The reverse model exchanges their roles. Both marginal mean rates are equal under both orientations. Mean matching removes a direct rate clue but does not make the coordinates exchangeable. Cross-bin temporal antisymmetry distinguishes direction at order a, whereas dispersion and a mixed third-degree count moment distinguish it at order a^2. Here a is the excitation mass. The two sources of information give different local sample-size scales.

We identify both directional tangents and their exact testing constants. For multiple bins, the first-order information has an explicit finite-window expression and a cubic coarse-resolution law. For a single bin, the second-order tangent is nonzero and gives a test that attains the fourth-root local limit. The mixed count moment is essential to its full information constant. We quantify that contribution, give a summary that preserves it, and construct two datasets that the original count summaries cannot distinguish. Estimating the common mean rate costs no local efficiency in either experiment.

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.

Relation to existing inference methods. Monte Carlo expectation–maximization integrates over latent event times (Shlomovich et al., 2022a;b). Particle Markov chain Monte Carlo provides likelihood-based inference for discretely observed multivariate processes (Lambe et al., 2026b), and mean-behavior approximations address interval-censored observations (Rizoiu et al., 2022). Lambe et al. (2026a) develop neural inference from simulated count summaries, including uncertainty quantification. Transformer Hawkes provides flexible event representations (Zuo et al., 2020), while interval-censored Transformer Hawkes addresses coarsened observations (Kong et al., 2023). Synchronization noise defines a distinct observation mechanism (Trouleau et al., 2019). Aggregation kernels, neural inverse estimation, and simulation-based confidence sets (Dalmasso et al., 2020) are therefore established ingredients. Our contribution is the exact observed-Hawkes directional tangents, their stationary-boundary integrals, and their attainable testing constants. Nonstandard rates under singular information are classical (Rotnitzky et al., 2000). The fourth-root rate alone is not new.

The learning study compares neural, linear, moment, and coarse-grained decoders under matched observations and calibration budgets. We ask which directional and temporal assertions a specified candidate family supports. Repeated calibration measures on-grid reporting, and omitted-parameter challenges test its limits. A correct directional prediction and a correct temporal envelope remain distinct outcomes even when no decoder dominates.

2 THE DIRECTIONAL EXPERIMENT

Let $h\colon (0,\infty) \to [0,\infty)$ be bounded with $\int h = 1$. For $\nu > 0$ and $0 \le a \le \bar{a} < 1$, define the forward stationary process by

$$\lambda_1(t) = \nu, \qquad \lambda_2(t) = \nu(1-a) + a \int_{(-\infty,t)} h(t-s)\, dN_1(s). \tag{1}$$

The reverse model exchanges the two coordinates. The source is Poisson on $\mathbb{R}$, and the target is conditionally Poisson given that source. The reproduction matrix is nilpotent. The restriction $a < 1$ ensures a positive target baseline, not a recurrent-Hawkes stability threshold. Both stationary mean rates are ν. Direction refers to predictive dependence in this family, not an intervention-based causal effect.

We observe N independent stationary windows of length $T = m\Delta$ through counts $X_j = N_1(B_j)$ and $Y_j = N_2(B_j)$ in $B_j = [j\Delta, (j+1)\Delta)$, $0 \le j < m$. Write $Q^T_{a,\pm,\Delta} = (C_\Delta)_\# P^T_{a,\pm}$ for the count laws. At $a = 0$, all counts are independent $\mathrm{Pois}(\lambda)$ with $\lambda = \nu\Delta$. Set $x_j = X_j - \lambda$ and $y_j = Y_j - \lambda$. The equal-prior optimal testing risk is R^*_N. Table 1 summarizes the regimes studied below. All constructions include the unobserved stationary past.

Table 1: The two local experiments and their observation cost. In each local limit, the kernel and physical window are fixed, and the common rate is known or estimated. N counts independent stationary windows, not arbitrary adjacent segments.

Question	Statistic and regime	Information and exposure
Cross-bin direction	$D_{h,\Delta}$, $m \ge 2$, $J > 0$; $a_N = \gamma N^{-1/2}$	J per window, J/T per unit time; total exposure NT
Single-bin direction	$D^{(2)}$, $m = 1$; $a_N = \gamma N^{-1/4}$	$J^{(2)}$ for the second-order contrast per window; total exposure NT
What a decoder reports	Independent rank inversion on a correct finite grid	Marginal coverage; wrong assertions divided by all datasets

Independent repetitions and a single long stream have different collection costs. With gaps of length g, observing N windows along one stream takes $NT + (N-1)g$ time. Equation (21) quantifies the independence approximation. The finite-sample envelopes in Section 5 are separate from the local limits and may be vacuous.

3 FIRST-ORDER DIRECTIONAL INFORMATION

For $d \geq 1$, define integrated lag weights

$$A_d = \int_0^\Delta \int_{d\Delta}^{(d+1)\Delta} h(t-s)\,dt\,ds, \qquad A_0 = \int_0^\Delta (\Delta - u)h(u)\,du. \tag{2}$$

These weights are established aggregation quantities (Koyama, 2025). Their exact implication for the directional tangent is as follows.

Theorem 1 (Projected directional score)**.** *The two count laws are right-differentiable in quadratic mean at $a = 0$. Their score difference and its null variance are*

$$D_{h,\Delta} = \frac{1}{\nu\Delta^2}\sum_{j<k} A_{k-j}(x_j y_k - y_j x_k), \qquad J_{\Delta,T}(h) = \frac{2}{\Delta^2}\sum_{d=1}^{m-1}(m-d)A_d^2. \tag{3}$$

The same-bin term $A_0 \sum_j x_j y_j/(\nu\Delta^2)$ occurs in each individual score and cancels in their difference. The exact-time directional information is

$$J_{0,T}(h) = 2\int_0^T (T-u)h(u)^2\,du. \tag{4}$$

Conditioning the fine-time null score on counts averages event locations uniformly within each bin. The directional difference retains only cross-bin pairs. Orthogonality of centered Poisson products gives (3). Appendix C proves the projection identity, including the contribution of the hidden past.

For $h_\tau(u) = \tau^{-1}e^{-u/\tau}\mathbf{1}\{u > 0\}$ and $r = e^{-\Delta/\tau}$,

$$A_0 = \Delta - \tau(1-r), \qquad A_d = \tau(1-r)^2 r^{d-1},$$
$$J_{\Delta,T}(h_\tau) = \frac{2\tau^2(1-r)^4}{\Delta^2}\frac{(m-1) - mr^2 + r^{2m}}{(1-r^2)^2}, \qquad J_{0,T}(h_\tau) = \frac{T}{\tau} - \frac{1-e^{-2T/\tau}}{2}. \tag{5}$$

The finite sum in (3) is numerically preferable near $r = 1$. At fixed Δ, τ, as $T = m\Delta \to \infty$,

$$\frac{J_{\Delta,T}(h_\tau)}{T} \longrightarrow \frac{1}{\tau}\Psi(\Delta/\tau), \qquad \Psi(z) = \frac{2(1-e^{-z})^3}{z^3(1+e^{-z})}. \tag{6}$$

Thus $\Psi(z) \to 1$ at fine resolution and $\Psi(z) \sim 2z^{-3}$ at coarse resolution.

Corollary 2 (Cubic decay for finite-mean kernels)**.** *If $\mu_h = \int_0^\infty uh(u)\,du < \infty$, then*

$$\mathcal{I}_\Delta(h) := \lim_{m\to\infty}\frac{J_{\Delta,m\Delta}(h)}{m\Delta} = \frac{2}{\Delta^3}\sum_{d\geq 1} A_d^2 \sim \frac{2\mu_h^2}{\Delta^3} \quad (\Delta \to \infty). \tag{7}$$

Since $A_1 \to \mu_h$ and $\sum_{d\geq 1} A_d \to \mu_h$, the sum of squared weights converges to μ_h^2. For a kernel supported on $(0, L]$, the rate is exactly $2\mu_h^2/\Delta^3$ once $\Delta \geq L$. The long-window limit precedes the coarse-resolution limit. At fixed m, the exponential information instead satisfies $J_{\Delta,m\Delta} \sim 2(m-1)\tau^2/\Delta^2$. A single bin has $J = 0$ exactly.

Theorem 3 (Attaining the first-order limit)**.** *For every admissible a,*

$$\mathbb{E}_{a,\pm} D_{h,\Delta} = \pm aJ/2, \qquad \mathrm{Var}_{a,\pm}(D_{h,\Delta}) \leq C_a J, \qquad C_a = 1 + \frac{(a+a^2)\|h\|_\infty}{2\nu} + 2a^2. \tag{8}$$

For $a > 0, J > 0$, majority sign of b independent group means has each directional error at most δ if b is odd, $b \geq 8\log(1/\delta)$, each group contains $k \geq \lceil 16C_a/(a^2J)\rceil$ windows, and $N \geq bk$.

Corollary 4 (Exact first-order local efficiency)**.** *For fixed ν, h, T, Δ with $J > 0$ and $a_N = \gamma/\sqrt{N}$, both optimal risk and the risk of* $\mathrm{sign}\sum_i D_{h,\Delta}^{(i)}$ *converge to*

$$\Phi(-\gamma\sqrt{J}/2). \tag{9}$$

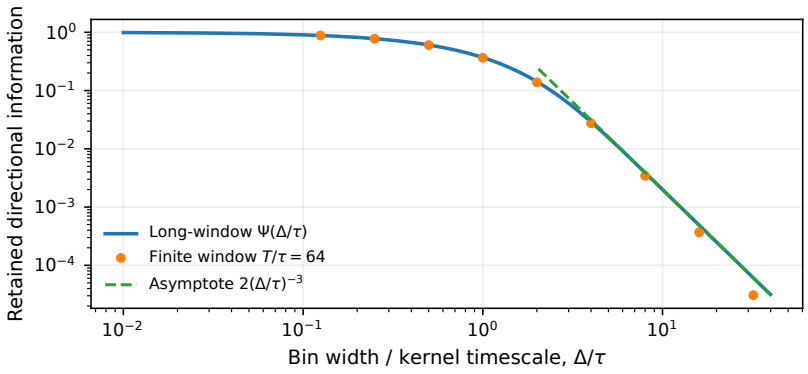


Figure 1: First-order directional information retained by binning. The solid curve is the exact long-window transfer $\Psi(\Delta/\tau)$, markers use $J_{\Delta,T}/J_{0,T}$ at $T/\tau = 64$, and the dashed curve is the coarse asymptote. This first-order loss does not include the second-order single-bin information in Theorem 5.

For exponential kernels, the score requires $O(m)$ time and constant working memory per window. Initialize $R_0^x = R_0^y = 0$ and update $R_{k+1}^x = rR_k^x + x_k$ and $R_{k+1}^y = rR_k^y + y_k$. Then compute

$$D_{h_\tau,\Delta} = \frac{\tau(1-r)^2}{\nu\Delta^2} \sum_{k=0}^{m-1} (y_k R_k^x - x_k R_k^y). \tag{10}$$

Appendices E and F prove the nonzero-signal moment bounds and local optimality without assuming independence between bins.

4 Second-order direction survives a single bin

Set $\Delta = T$, $m = 1$, and write X, Y for the observed totals. Define

$$p_T(s) = \int_0^T h(t-s)\mathbf{1}\{s < t\}\, dt, \qquad \beta_2 = \int_{\mathbb{R}} p_T(s)^2 ds, \qquad \chi_2 = \int_0^T p_T(s)^2 ds. \tag{11}$$

Let $\lambda = \nu T$, $c_1(z) = z - \lambda$, and $c_2(z) = (z-\lambda)^2 - z$. These centered Poisson polynomials distinguish excess dispersion from ordinary Poisson variability. The directional second-order statistic is

$$\begin{aligned} D^{(2)} = {} & \frac{\nu\beta_2}{2\lambda^2}\{c_2(Y) - c_2(X)\} \\ & + \frac{\nu\chi_2}{2\lambda^3}\{c_1(X)c_2(Y) - c_1(Y)c_2(X)\}. \end{aligned} \tag{12}$$

Write $V = c_2(Y) - c_2(X)$ and $M = c_1(X)c_2(Y) - c_1(Y)c_2(X)$. The dispersion contrast V compares marginal excess dispersion. The mixed contrast M measures how one coordinate's excess count accompanies the other's excess dispersion. The coefficient β_2 includes parents before the window, whereas χ_2 includes only parents whose counts are observed.

Theorem 5 (The single-bin second-order tangent)**.** *For the bounded normalized kernel family,*

$$J^{(2)} := \mathbb{E}_0(D^{(2)})^2 = \frac{\beta_2^2}{T^2} + \frac{\chi_2^2}{\nu T^3} > 0, \qquad \mathbb{E}_{a,\pm} D^{(2)} = \pm a^2 J^{(2)}/2. \tag{13}$$

The count square-root densities satisfy $\sqrt{p_a^+} - \sqrt{p_a^-} = (a^2/2)D^{(2)}\sqrt{p_0} + O_{L^2}(a^3)$. *For fixed* ν, h, T *and* $a_N = \gamma N^{-1/4}$, *both optimal risk and the risk of* $\operatorname{sign}\sum_i D_i^{(2)}$ *converge to*

$$\Phi(-\gamma^2\sqrt{J^{(2)}}/2). \tag{14}$$

Appendix K derives the tangent from the stationary count generating function and bounds the third square-root-density derivative. The likelihood proof compares the two moving alternatives through their symmetric mixture. They are *not* contiguous to the product independent-process null at the fourth-root scale, so a null-based likelihood argument would not apply.

For the exponential family, $\beta_2 = T - \tau(1-r)$ and $\chi_2 = T - 2\tau(1-r) + \tau(1-r^2)/2$, now with $r = e^{-T/\tau}$. More generally $J^{(2)} \to 1$ as $T \to \infty$ at fixed ν. The number of independent long bins needed at fixed local error is therefore of order a^{-4}, rather than infinite. Each bin costs exposure T. This replicate count does not establish a benefit at fixed total observation time. The exact information split is

$$\rho_V = \frac{\beta_2^2}{\beta_2^2 + \chi_2^2/(\nu T)}, \qquad \rho_M = 1 - \rho_V. \tag{15}$$

The fraction ρ_V measures dispersion's squared local efficacy relative to $D^{(2)}$, not its finite-sample accuracy ratio. At $T = \tau = 1$ and expected counts $\nu T = 0.25, 1, 4$, the mixed term contributes respectively 45.51%, 17.27%, and 4.96%. Figure 8 gives the dependence on count intensity and window length.

A distinction absent from marginal dispersion. At $\nu = T = \tau = 1$, take four independent windows with source counts $X = (0, 1, 2, 3)$. The two target arrays $Y_A = (0, 2, 3, 1)$ and $Y_B = (0, 3, 1, 2)$ have identical marginal count moments and $\sum_i X_i Y_i = 11$. Both have $\sum_i V_i = 0$, but $\sum_i M_i = 2$ and -2, respectively. Consequently $\sum_i D_i^{(2)} = \pm 0.16809$. All nine aggregate decoder summaries in (89) coincide on these datasets. Adding the mixed summary preserves the distinction (Appendix S). This is an algebraic decision example, not a reliability claim for four windows.

Theorem 6 (Adaptation to an unknown common rate)**.** *Let* $\hat{\nu}_N = \max\{N^{-1}, \sum_{i,j}(X_{ij} + Y_{ij})/(2NT)\}$. *Replacing* ν *everywhere in* $D_{h,\Delta}$ *by* $\hat{\nu}_N$ *retains* (9)*. For the single-bin experiment, replacing* ν *everywhere in* $D^{(2)}$ *by the same estimate retains* (14)*. The assertions are pointwise in a fixed positive common rate, with kernel and design fixed.*

Both directional statistics are orthogonal to the null rate score. Appendix L proves that their plug-in differences, after summation and $N^{-1/2}$ scaling, are $o_P(1)$. This rate adaptation concerns the score tests, not recalibration of a fixed-rate decoder.

5 Finite-sample bounds and their applicability

Define root Hellinger distance by $H^2(P, Q) = \int(\sqrt{p} - \sqrt{q})^2$. Set $c = 1 - \bar{a}$ and

$$M_4 = \nu \int h^4 + 3\nu^2 \left(\int h^2\right)^2, \qquad K = \frac{1}{2}\sqrt{M_4\left(\frac{3T^2}{c^2\nu^2} + \frac{T}{c^3\nu^3}\right)}. \tag{16}$$

For h_τ, $M_4 = \nu/(4\tau^3) + 3\nu^2/(4\tau^2)$.

Theorem 7 (An explicit first-order finite-sample envelope)**.** *With* $L = (a\sqrt{J}/2 - a^2K)_+$ *and* $U = a\sqrt{J}/2 + a^2K$*,*

$$L \le H(Q^T_{a,+,\Delta}, Q^T_{a,-,\Delta}) \le \min\{U, \sqrt{2}\}, \tag{17}$$

$$\frac{1 - \sqrt{1 - [1 - U^2/2]_+^{2N}}}{2} \le R_N^* \le \frac{1}{2}e^{-NL^2/2}. \tag{18}$$

Also $R_N^* \ge (1 - \sqrt{N}U)_+/2$*. The sufficient condition* $aK \le \sqrt{J}/4$ *makes* H *comparable to* $a\sqrt{J}$ *within fixed constants.*

The proof controls observed root-density curvature without summing latent orders. The bound holds at every admissible amplitude but can be uninformative. At $\nu = \tau = 1$, $T = 64$, and $a = \bar{a} = 0.2$, it gives only $0 \le R_N^* \le 1/2$ at all four experimental resolutions, regardless of N (Table 2). The local information law consequently gives no sharp finite-sample guarantee in that slice. A nonvacuous example at the same T is $\Delta = 0.25$, $a = 0.01$, $\bar{a} = 0.2$, $N = 1000$, for which (18)

Table 2: Applicability of the first-order envelope for $\nu = \tau = 1$, $T = 64$, and fixed $\bar{a} = 0.2$. The last column is a *sufficient*, not necessary, amplitude threshold. The envelope at $a = 0.2$ is vacuous in every row.

Δ	J	K	$aK/\sqrt{J}$ at $a = .2$	$\sqrt{J}/(4K)$
0.25	49.34922	69.5072	1.979	0.025267
1	23.20824	69.5072	2.886	0.017327
4	1.74191	69.5072	10.533	0.004747
16	0.02344	69.5072	90.804	0.000551

gives $0.0445 \leq R_N^* \leq 0.3362$. The score error on 2000 independent datasets is 0.1245 ± 0.0074 (one standard error).

The score moments in Theorem 3 provide a separate finite-sample upper bound. Total-window dispersion, whose mean and exact variance are given in Appendix I, supplies another bound that does not deteriorate when the same window is coarsened. Proposition 14 chooses between these two tests using their explicit mean-to-variance ratios. Appendix K also gives a finite second-order Hellinger remainder. For $T = 64$, $\nu = \tau = 1$, $a = 0.2$, and error target 0.05, 25 groups suffice. At $\Delta = 0.25$, the construction uses 10 windows per group, hence 250 windows and 16,000 time units of observed exposure. At $\Delta = 16$, it selects dispersion and requires 10,339 windows per group, hence 258,475 windows and 16,542,400 exposure units. These sufficient counts are conservative. They neither estimate optimal sample size nor predict the decoder study's performance. Figure 4 shows the per-group quantities.

6 Reporting only supported distinctions

Let Θ be a specified finite parameter grid. A frozen decoder assigns scores $q_\theta(Z)$ to a dataset of N count windows. For each θ, independently simulate B calibration datasets of the same design, and form

$$p_\theta(Z) = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{q_\theta(Z_{\theta b}) \leq q_\theta(Z)\}}{B + 1}, \qquad \mathcal{C}_\alpha(Z) = \{\theta : p_\theta(Z) > \alpha\}. \tag{19}$$

High scores indicate compatibility. The score may be a neural probability, a moment discrepancy with its sign reversed, or a coarse-grained objective. The principle is test inversion, not an assumption that neural probabilities are calibrated (Dalmasso et al., 2020).

Theorem 8 (Finite-grid assertions and a support guard)**.** *Conditional on independent training, for every $\theta \in \Theta$,*

$$\mathbb{P}_\theta\{\theta \in \mathcal{C}_\alpha(Z)\} \geq 1 - \alpha. \tag{20}$$

The probability includes the observation and calibration bank. Every assertion shared by a nonempty $\mathcal{C}_\alpha$ is simultaneously correct with probability at least $1 - \alpha$. If two independently trained or training-fixed compatibility scores are each rank-calibrated at $\alpha/2$, the intersection of their candidate sets has the same guarantee. The two calibration tests need not be independent.

A direction is reported only when all surviving candidates agree. A nonnull timescale is reported only when the set is a single nonnull candidate. For $F_\theta(u) = \sigma a(1 - e^{-u/\tau})$, the minimum and maximum over a nonempty set give a simultaneous cumulative-kernel envelope for all $u \geq 0$. Empty sets yield no assertion. A wrong report has *unconditional* probability at most α; the conditional error among reports can exceed α. Appendix J proves these claims and states an explicit total-variation misspecification extension.

A normalized decoder score can favor a candidate even when every candidate fits poorly. We examine a guarded rule to test for this failure. Its second score is minus the candidate-specific Mahalanobis distance in the nine observed summaries, with means and regularized covariances fitted using training simulations only. Intersecting the two $\alpha/2$ rank sets preserves on-grid validity by Theorem 8. The absolute-fit check retains the on-grid guarantee but supplies no universal off-grid certificate.

Independence of the observed windows requires separate treatment. For windows separated by a gap g in the same exponential feed-forward trajectory, Theorem 15 proves a total-variation discrepancy from independent windows of at most

$$\epsilon_g = \min\{1, 2(N-1)\nu a\tau e^{-g/\tau}(1 - e^{-T/\tau})\}. \tag{21}$$

An independent-bank report at level α_0 then has unconditional error at most $\alpha_0 + \epsilon_g$. A gap and a stricter α_0 can be chosen from known parameter bounds. In the diagnostic with $N = 32$, $T = 8$, $\nu = 1$, $a = 0.6$, and $\tau = 8$, gap $g = 96$ permits a useful correction but costs 3232 collection-time units for 256 observed-window units. Gap 64 still gives $\epsilon_g > 0.05$. The coupling is specific to the feed-forward family.

7 Empirical comparisons and specification tests

All experimental observations are synthetic. The feed-forward exponential simulations include pre-window offspring exactly. The other stress samplers have explicit prehistory-truncation bounds. Training, calibration, and testing use disjoint paths. Appendices Q, R, and T give the protocols, numerical methods, and complete results.

Local testing risk and a likelihood reference. For the first-order setting $\Delta = 4$, $T = 64$, $\nu = \tau = 1$, and $\gamma\sqrt{J}/2 = 1.5$, we increase N from 1024 to 16384 and repeat with three independent simulation seeds. Table 3 reports 1200 independent test datasets per N and includes the estimated-rate statistic. The risk limit 0.0668 is not a finite-N identity. Across a separate 45-cell study with $\nu \in \{0.25, 0.5, 1, 2, 4\}$ and three amplitudes, paired comparisons assess the cost of estimating the rate.

For the single-bin experiment, positive compound-Poisson coefficient recursion gives a numerical count-likelihood reference. A multinomial histogram of count pairs then generates a dataset of N independent windows without simulating every latent event. We evaluate $T = 1$, $\nu \in \{0.25, 1, 4\}$ and $T = 8$, $\nu = 1$, each at $N \in \{4096, 65536\}$, with $\tau = 1$ and a_N chosen so that $\gamma^2\sqrt{J^{(2)}}/2 = 1$. Each cell uses 20,000 fresh datasets, paired across methods. Quadrature doubling, a separate Fourier calculation, and count-tail bounds check the numerical law (Appendix T).

Table 3: First-order local-risk comparison. Entries are percentages, with Monte Carlo standard errors from 1200 independent datasets per row. Both methods use the same datasets. The limiting error is 6.68%.

N	a_N	Known rate	Estimated rate
1024	0.07103	5.58 (0.66)	5.58 (0.66)
4096	0.03552	6.42 (0.71)	6.42 (0.71)
16384	0.01776	6.67 (0.72)	6.67 (0.72)

Table 4: The mixed count term at $T = \tau = 1$. D, V, and L denote second-order score, dispersion, and numerical likelihood. Risks are percentages from 20,000 paired datasets per row. Differences and their paired standard errors (parentheses) are percentage points. Negative differences favor D. The efficient local limit is 15.87%.

νT	N	R_D	R_V	R_L	$R_D - R_V$ (SE)	$R_D - R_L$ (SE)
0.25	4,096	22.70	23.81	21.99	−1.11 (0.29)	+0.71 (0.18)
0.25	65,536	19.43	23.18	19.19	−3.75 (0.29)	+0.24 (0.13)
1	4,096	18.15	18.20	17.66	−0.05 (0.20)	+0.49 (0.13)
1	65,536	16.81	18.48	16.82	−1.67 (0.21)	−0.01 (0.09)
4	4,096	16.88	17.04	16.91	−0.17 (0.14)	−0.03 (0.09)
4	65,536	15.90	16.62	15.99	−0.73 (0.15)	−0.10 (0.06)

At $\nu T = 1$, $N = 65536$, the score reduces error relative to dispersion by 1.67 (paired standard error 0.21) percentage points. At $N = 4096$, the paired difference is smaller and remains unresolved

relative to its Monte Carlo uncertainty. At $\nu T = 0.25$, the larger mixed-information fraction is accompanied by a larger error separation, but the score remains above its local limit. Finite amplitude need not preserve the asymptotic efficiency ordering. The complete count-law study and the event-sampler comparison, including cells favoring dispersion, appear in Appendices T and R.

Repeated calibration with matched observations. The seven-candidate grid contains a null and the six combinations of $a = 0.2$, $\tau \in \{0.5, 1, 2\}$, and the two directions, with $\nu = 1$, $T = 64$, and $N = 64$. Each fitted method has access to the same 800 training datasets per candidate. We freeze one training seed and generate ten independent calibration/test banks, each with 199 calibration and 200 test datasets per candidate. The neural and linear decoders use nine summaries. For every dictionary kernel, these contain $\sum_i D^{(i)}_{h,\Delta}$ up to a known normalization, so the corresponding binary first-order test is representable. At one bin they retain dispersion but not the mixed statistic, as the collision in Section 4 demonstrates. Appendix S gives a tenth summary and exact linear readout of $\sum_i D^{(2)}_i$. The frozen nine-feature decoders do not use that additional feature, and no local optimality claim applies to them. The moment comparator fits regularized Gaussian summaries, while the coarse-grained quadratic comparator uses the specified candidate parameters and requires no fitting simulations. Every method sees the same counts and calibration datasets. Score-only and dispersion-only ablations appear in the appendix. The guarded neural rule uses no additional simulations.

Table 5: Repeated-bank reporting at $N = 64$. Percentages pool ten independent banks, with 14,000 test datasets per resolution. "Cov." is candidate coverage, "Empty" the empty-set rate, "Report" a unanimous nonempty directional report, "Wrong/all" its unconditional error, and "Wrong/report" error conditional on reporting. Direction includes the null. Full classwise coverage and bank variation are in the appendix.

Δ	Decoder	Cov.	Empty	Report	Wrong/all	Wrong/report
0.25	Neural	94.97	0.33	99.67	0.53	0.53
0.25	Moments	95.02	0.39	99.59	0.46	0.46
0.25	CG quadratic	94.81	1.61	98.39	0.04	0.04
0.25	Guarded neural	95.38	1.97	97.69	0.41	0.42
1	Neural	95.00	0.64	99.34	0.26	0.26
1	Moments	94.84	0.55	99.44	0.35	0.35
1	CG quadratic	94.61	0.77	99.23	0.17	0.17
1	Guarded neural	95.14	1.84	98.11	0.38	0.39
4	Neural	94.86	0.01	61.99	2.51	4.06
4	Moments	94.90	0.00	61.69	2.17	3.52
4	CG quadratic	94.93	0.00	61.69	2.19	3.54
4	Guarded neural	95.15	1.33	53.22	1.52	2.86
16	Neural	94.91	0.00	14.11	1.57	11.14
16	Moments	94.86	0.00	12.24	1.05	8.58
16	CG quadratic	94.96	0.00	14.86	1.61	10.87
16	Guarded neural	95.64	0.78	10.65	0.85	7.98

Table 5 reports each denominator separately. Coarsening reduces directional reporting, and singleton timescale reports disappear while some directional evidence remains. The non-neural comparators remain competitive. Each repetition changes calibration and test simulations while holding training fixed. Calibration scores, test scores, and candidate sets for all ten banks are retained.

Misspecification and abstention. We separately vary the true rate, amplitude, and timescale, replace the exponential kernel by an Erlang kernel or an exponential mixture, and introduce feedback. Models and calibration banks remain frozen. Table 6 gives a selected $\Delta = 1$ slice. Appendix R reports all cases and resolutions. Outside the grid, exact-candidate coverage is undefined. We instead evaluate directional assertions and whether the reported envelope contains the true kernel on 257 finite lags and its limiting amplitude. The latter is an empirical evaluated-lag criterion, not a simultaneous continuous-parameter theorem. In the bidirectional case there is no true one-way label, so directional accuracy is not assigned.

Table 6: Frozen-decoder challenges at $\Delta = 1$, $N = 64$, with 800 test datasets per case, balanced across orientations. "Base" is neural rank inversion and "Guard" its two-test version. Rates and amplitudes differ from training only where shown. These are misspecification diagnostics, not off-grid coverage guarantees.

True setting	Rule	Report	Empty	Wrong/all	Envelope
$\nu = 2$	Base	100.00	0.00	50.00	0.00
$\nu = 2$	Guard	0.00	100.00	0.00	0.00
$a = 0.05$	Base	85.00	14.75	66.50	0.25
$a = 0.05$	Guard	92.50	7.12	74.00	0.38
$\tau = 0.75$	Base	100.00	0.00	0.00	30.00
$\tau = 0.75$	Guard	98.25	1.75	0.00	43.50
Erlang-2	Base	100.00	0.00	0.00	0.00
Erlang-2	Guard	90.00	10.00	0.00	0.00
Feedback	Base	99.50	0.00	–	–
Feedback	Guard	94.62	5.25	–	–

Table 6 separates directional prediction from temporal reporting. Accurate directional point predictions need not yield correct temporal envelopes. A nonempty confidence set can also fail to detect an omitted candidate. The absolute-fit guard rejects every dataset in the rate-doubling challenge, but worsens weak-amplitude false reports from 66.5% to 74.0%. It cannot make an incomplete parameter grid correct. A separate dependent-window experiment applies the gap correction in (21); its coverage statement is marginal over independent calibration, not conditional on its realized bank. Count forecasting and random-jitter comparisons are retained only as secondary diagnostics in the appendix. No forecasting superiority is claimed.

8 Discussion

The two directional experiments require different summaries. Cross-bin antisymmetry carries the root-N signal. With only one bin, excess dispersion and a mixed count moment carry the fourth-root signal. Their information split identifies a distinction lost when the decoder averages only the original nine summaries. One additional coordinate preserves that distinction. Common-rate estimation leaves both local testing constants unchanged.

The results concern the stated feed-forward family. A sharp many-bin transition, unequal unknown rates, continuous kernel uncertainty, and external-data validation remain outside their scope. The reporting experiments delineate the limits of a specified candidate family rather than solve off-grid inference. Directional accuracy, temporal specificity, and calibrated assertions are separate outcomes. Decoder expressiveness alone cannot justify a distinction absent from the retained summaries or unsupported by the calibration design.

STATEMENT ON AI USE

Generative AI was used for proofreading, including grammatical correction, text enhancement, and general “polishing". The authors assume full responsibility for the correctness and originality of the research.

EXPERIMENTAL MATERIALS AND SCOPE

The supplementary archive contains simulation and inference code, fitted decoder weights, numerical outputs, and scripts that generate the figures and tables. All experimental data are synthetic. The appendices specify the data splits, seeds, model configurations, numerical checks, and evaluation denominators. Direction must not be interpreted as an intervention-based causal conclusion.

REFERENCES

Massil Achab, Emmanuel Bacry, Stéphane Gaïffas, Iacopo Mastromatteo, and Jean-François Muzy. Uncovering causality from multivariate Hawkes integrated cumulants. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1–10, 2017. PMLR 70, achab17a.

Felix Cheysson and Gabriel Lang. Spectral estimation of Hawkes processes from count data. *The Annals of Statistics*, 50(3):1722–1746, 2022. doi:10.1214/22-AOS2173.

Niccolò Dalmasso, Rafael Izbicki, and Ann B. Lee. Confidence sets and hypothesis testing in a likelihood-free inference setting. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2323–2334, 2020. PMLR 119, dalmasso20a.

Alan G. Hawkes and David Oakes. A cluster process representation of a self-exciting process. *Journal of Applied Probability*, 11(3):493–503, 1974. doi:10.2307/3212693.

Quyu Kong, Pio Calderon, Rohit Ram, Olga Boichak, and Marian-Andrei Rizoiu. Interval-censored Transformer Hawkes: Detecting information operations using the reaction of social systems. In *Proceedings of the ACM Web Conference 2023*, pp. 1813–1821, 2023. doi:10.1145/3543507.3583481.

Shinsuke Koyama. Coarse-grained Hawkes processes. *Entropy*, 27(6):555, 2025. doi:10.3390/e27060555.

Jason J. Lambe, Feng Chen, Tom Stindl, and Tsz-Kit Jeffrey Kwan. Neural networks for parameter estimation of the discretely observed Hawkes process. *arXiv preprint arXiv:2506.01258v3*, 2026a. arXiv:2506.01258v3.

Jason J. Lambe, Feng Chen, Tom Stindl, and Tsz-Kit Jeffrey Kwan. Parametric inference for the discretely observed multivariate Hawkes process using particle Markov Chain Monte Carlo. *arXiv preprint arXiv:2503.18351v3*, 2026b. arXiv:2503.18351v3.

Marian-Andrei Rizoiu, Alexander Soen, Shidi Li, Pio Calderon, Leanne J. Dong, Aditya Krishna Menon, and Lexing Xie. Interval-censored Hawkes processes. *Journal of Machine Learning Research*, 23(338):1–84, 2022. JMLR 23, 21-0917.

Andrea Rotnitzky, David R. Cox, Matteo Bottai, and James Robins. Likelihood-based inference with singular information matrix. *Bernoulli*, 6(2):243–284, 2000. `https://projecteuclid.org/journals/bernoulli/volume-6/issue-2/Likelihood-based-inference-with-singular-information-matrix/bj/1081788028.full`.

Leigh Shlomovich, Edward A. K. Cohen, Niall Adams, and Lekha Patel. Parameter estimation of binned Hawkes processes. *Journal of Computational and Graphical Statistics*, 31(4):990–1000, 2022a. doi:10.1080/10618600.2022.2050247.

Leigh Shlomovich, Edward A. K. Cohen, and Niall Adams. A parameter estimation method for multivariate binned Hawkes processes. *Statistics and Computing*, 32:98, 2022b. doi:10.1007/s11222-022-10121-2.

William Trouleau, Jalal Etesami, Matthias Grossglauser, Negar Kiyavash, and Patrick Thiran. Learning Hawkes processes under synchronization noise. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6325–6334, 2019. PMLR 97, trouleau19a.

Lingxiao Zhou and Georgia Papadogeorgou. Bayesian inference for aggregated Hawkes processes. *Bayesian Analysis*, 21(3):1247–1275, 2026. doi:10.1214/25-BA1516.

Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer Hawkes process. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11692–11702, 2020. PMLR 119, zuo20a.

A NOTATION AND THE STATIONARY MODEL

Table 7: Notation used in the proofs. A single observed window is the unit of independence.

Symbol	Meaning
ν	Common stationary event rate of both coordinates.
$a, \bar{a}, c$	Excitation mass, upper mass bound, and $c = 1 - \bar{a} > 0$.
h, h_τ	A normalized causal kernel and the exponential kernel with timescale τ.
T, Δ, m, N	Window length, bin width, bins per window, and independent windows.
B_j, X_j, Y_j	Bin j and its two observed counts.
λ, x_j, y_j	Null bin mean $\nu\Delta$ and centered counts.
A_d	Integrated lag weight in (2).
S_+, S_-	Right scores at the independent-process null for the two branches.
$D_{h,\Delta}, J_{\Delta,T}$	Observed directional score and its null variance.
$J_{0,T}$	Directional information from exact event times inside the window.
M_4, K, C_a	Shot-noise fourth moment, root-density curvature bound, and score-variance factor.
$G_{\Delta,T}$	Null Gram matrix of an exponential score dictionary.
$\Theta, B, \mathcal{C}_\alpha$	Candidate grid, calibration datasets per candidate, and confidence set.

All integrals over a kernel without limits are over $(0, \infty)$. The assumptions $h \geq 0$, $\int h = 1$, and $\|h\|_\infty < \infty$ imply $h \in L^p$ for every finite $p \geq 1$. A stationary source Poisson random measure of intensity $\nu\, ds$ on $\mathbb{R}$ has shot-noise expectation ν. The shot noise is therefore almost surely finite at each fixed time. Given that source, superpose an independent target baseline process of rate $\nu(1-a)$ and independent Poisson offspring with intensity $ah(t-s)\, dt$ from each source point s. There are no target-to-source or target-to-target offspring. This directly constructs the stationary law (1).

Its target mean rate is $\nu(1-a) + a\nu \int h = \nu$. The construction specializes the Poisson cluster representation (Hawkes and Oakes, 1974) to a feed-forward model with no descendants beyond the first generation. Conditioning on the source then gives a Poisson target process. The target is not generally Poisson after the source has been marginalized.

Let $M(ds) = N_1(ds) - \nu\, ds$ be the compensated source measure. For integrable deterministic f, define $I_1(f) = \int f\, dM$. The identities used below follow by differentiating the Poisson generating functional,

$$\mathbb{E} \exp\{tI_1(f)\} = \exp\left\{\nu \int (e^{tf(s)} - 1 - tf(s))\, ds\right\}, \tag{22}$$

when the expression exists in a neighborhood of zero, and otherwise by bounded truncation for the finite moments in question. In particular,

$$\begin{aligned} \mathbb{E} I_1(f) I_1(g) &= \nu \int fg, \\ \mathbb{E} I_1(f) I_1(g) I_1(k) &= \nu \int fgk, \\ \mathbb{E} I_1(f)^4 &= \nu \int f^4 + 3\nu^2 \left(\int f^2\right)^2. \end{aligned} \tag{23}$$

All subsequent interchanges are justified by these finite moment bounds, Tonelli's theorem for nonnegative kernels, or L^2 approximation by bounded, compactly supported integrands.

B DIFFERENTIABILITY AND AN EXPLICIT CURVATURE BOUND

We first establish the regularity needed for score projection and the finite-sample testing bound. Augment the observation by the entire source process, including its negative-time history. The

distribution of this source does not depend on a. Conditional on it, set

$$Z(t) = \int_{(-\infty,t)} h(t-s)\, dM(s), \qquad \lambda_a(t) = \nu + aZ(t), \quad 0 \le t \le T. \tag{24}$$

Because $Z(t) \ge -\nu$, $\lambda_a(t) \ge c\nu$. The conditional likelihood of the target relative to rate-ν Poisson observation on $[0,T]$ is

$$L_a = \exp\left\{-a\int_0^T Z(t)\, dt\right\} \prod_{t\in N_2\cap[0,T]} \left(1 + \frac{aZ(t)}{\nu}\right). \tag{25}$$

The source has no likelihood contribution. The complete score and observed negative second log derivative are

$$S_a = \int_0^T q_a(t)\{dN_2(t) - \lambda_a(t)\, dt\}, \qquad W_a = \int_0^T q_a(t)^2\, dN_2(t), \qquad q_a(t) = \frac{Z(t)}{\lambda_a(t)}. \tag{26}$$

For a complete-data square-root density f_a, differentiation gives

$$\frac{f_a'}{f_a} = \frac{1}{2}S_a, \qquad \frac{f_a''}{f_a} = \frac{1}{4}(S_a^2 - 2W_a). \tag{27}$$

Conditional on the source, define

$$A = \int_0^T \frac{Z(t)^2}{\lambda_a(t)}\, dt, \qquad C = \int_0^T \frac{Z(t)^4}{\lambda_a(t)^3}\, dt. \tag{28}$$

These symbols A, C are local to this proof and are not the bin weights A_d.

Lemma 9 (Complete and marginal root-density curvature)**.** *Every marginal observation of this augmented experiment, including exact event times in $[0,T]$ and binned counts, has a square-root density g_a whose L^2 first derivative is absolutely continuous on $[0,\bar{a}]$, with one-sided endpoint derivatives. Its almost-everywhere second derivative satisfies*

$$\operatorname*{ess\,sup}_{0\le a\le\bar{a}} \|g_a''\|_2 \le K, \qquad \|g_a - g_0 - ag_0'\|_2 \le \frac{a^2K}{2}, \tag{29}$$

where K is defined in (16)*.*

Proof. Conditional Poisson moments imply

$$\mathbb{E}(S_a^4 \mid N_1) = 3A^2 + C, \quad \mathbb{E}(S_a^2 W_a \mid N_1) = A^2 + C, \quad \mathbb{E}(W_a^2 \mid N_1) = A^2 + C. \tag{30}$$

For example, $W_a = A + I_1(q_a^2)$ under the conditional target intensity measure, and the mixed third centered moment is $\int q_a^4 \lambda_a = C$. Substituting (30) into (27) gives the exact identity

$$\|f_a''\|_2^2 = \frac{3\mathbb{E}A^2 + \mathbb{E}C}{16}. \tag{31}$$

The stationary fourth moment of $Z(t)$ is M_4 by (23). Cauchy–Schwarz and the positive lower bound on intensity yield

$$\mathbb{E}A^2 \le \frac{T^2 M_4}{c^2\nu^2}, \qquad \mathbb{E}C \le \frac{TM_4}{c^3\nu^3}. \tag{32}$$

Let Y denote any marginal observation. Conditional expectations in the next display are under the parameter a. Differentiating $g_a(y)^2 = \int f_a(y,w)^2\, dw$ gives

$$\frac{g_a''(Y)}{g_a(Y)} = \mathbb{E}_a\left[\frac{f_a''}{f_a}\,\middle|\, Y\right] + \frac{1}{4}\operatorname{Var}_a(S_a \mid Y). \tag{33}$$

Conditional Jensen's inequality gives

$$\begin{aligned} \|g_a''\|_2 &\le \|f_a''\|_2 + \frac{1}{4}\sqrt{\mathbb{E}_a[\operatorname{Var}_a(S_a \mid Y)^2]} \\ &\le \frac{1}{4}\sqrt{3\mathbb{E}A^2 + \mathbb{E}C} + \frac{1}{4}\sqrt{\mathbb{E}_a S_a^4} \\ &= \frac{1}{2}\sqrt{3\mathbb{E}A^2 + \mathbb{E}C} \le K. \end{aligned} \tag{34}$$

To justify differentiation of the marginal densities, we establish the integrated derivative identities first. Use the complete null probability measure as the dominating measure, so $f_a = \sqrt{L_a}$. For almost every complete observation, $\int_0^T |Z(t)|\,dt < \infty$ and there are finitely many target events, at all of which Z is finite. The lower bound $\lambda_a \geq c\nu$ makes $a \mapsto f_a$ twice continuously differentiable pointwise on the compact parameter interval. In addition to (31),

$$\|f_a'\|_2^2 = \frac{1}{4}\mathbb{E}A \leq \frac{T}{4c}\int h^2.$$

The first two derivatives are jointly measurable and uniformly bounded in L^2. Their Bochner integrals therefore exist. The pointwise fundamental theorem and Fubini's theorem identify these integrals with the differences of f_a and f_a' in L^2. Consequently f_a has an absolutely continuous L^2 first derivative.

By Cauchy–Schwarz and the preceding bounds, the complete densities f_a^2 have first and second derivatives $2f_af_a'$ and $2(f_a')^2 + 2f_af_a''$ in L^1. Conditional expectation under the null preserves these L^1 identities for the marginal densities $p_a(Y) = \mathbb{E}_0[f_a^2 \mid Y]$. These densities are strictly positive on the entire compact interval almost surely. Indeed, the pointwise minimum $\inf_a f_a^2$ is positive, and its conditional expectation gives a positive lower bound for p_a. A finite upper envelope follows from the integrability of $\sup_a f_a^2$, which is bounded in expectation using $f_a = f_0 + \int_0^a f_u'\,du$ and Cauchy–Schwarz. Thus the scalar chain rule for $g_a = \sqrt{p_a}$ gives (33) for almost every parameter and marginal observation. The first derivative also satisfies $g_a'/g_a = \mathbb{E}_a[S_a \mid Y]/2$. Conditional Jensen gives its uniform L^2 bound, while (34) bounds the second derivative. Applying Fubini again upgrades the scalar fundamental theorem to the L^2 identity and gives an absolutely continuous first derivative, including its one-sided trace at zero.

The fundamental theorem for such Hilbert-space curves therefore gives

$$g_a - g_0 - ag_0' = \int_0^a (a-u)g_u''\,du.$$

Minkowski's inequality proves (29). In particular the remainder is $O(a^2)$ in L^2, which is stronger than the $o(a)$ remainder required for right differentiability in quadratic mean. □

The source augmentation is different for the forward and reverse branches, but both have the same curvature bound. Each observed branch is dominated by the same independent-process observation law at $a = 0$, so their square-root expansions can be compared in a common L^2 space. The bound is uniform over all marginalizations. This uniformity gives no favorable Δ dependence and can make the bound conservative.

C Proof of the projected-score and transfer formulas

C.1 Projection from the stationary complete experiment

At $a = 0$, source and target are independent Poisson processes. The complete forward score from (26) is

$$S_{h,+}^c = \frac{1}{\nu}\int_0^T \left\{\int_{(-\infty,t)} h(t-s)\,dM_1(s)\right\} dM_2(t). \tag{35}$$

The marginal score is the conditional expectation of the complete score. This follows by writing the observed density as the integral of the complete density, differentiating once, and dividing by the observed density. At the null, differentiation is justified by Lemma 9 or by the L^2 first-derivative bound implicit in its proof.

Given exact event times of both coordinates in $[0, T]$, the part of (35) involving source times $s < 0$ has conditional expectation zero. Indeed, the compensated source past is independent of the entire observed null window. Thus the exact-time marginal score is

$$S_{h,+}^0 = \frac{1}{\nu}\int_{0<s<t<T} h(t-s)\,dM_1(s)\,dM_2(t). \tag{36}$$

The reverse score exchanges M_1, M_2. The strict inequalities do not affect the integrals because the processes almost surely have no deterministic-time or cross-coordinate ties under the continuous model.

Conditional on a Poisson count in a fixed bin, event locations are independent and uniform in that bin. Counts from the two null coordinates are independent. For $j < k$, conditioning the contribution over $B_j \times B_k$ gives

$$\mathbb{E}_0\left[\int_{B_j}\int_{B_k} h(t-s)\,dM_1(s)\,dM_2(t)\,\middle|\,X,Y\right] = \frac{A_{k-j}}{\Delta^2}x_j y_k.$$

For $j = k$, the integration domain is the causal triangle and the same computation gives $A_0 x_j y_j/\Delta^2$. For $j > k$ the contribution vanishes. Consequently,

$$\begin{aligned} S^{\Delta}_{h,+} &= \frac{1}{\nu\Delta^2}\left(A_0\sum_j x_j y_j + \sum_{j<k} A_{k-j}x_j y_k\right), \\ S^{\Delta}_{h,-} &= \frac{1}{\nu\Delta^2}\left(A_0\sum_j x_j y_j + \sum_{j<k} A_{k-j}y_j x_k\right). \end{aligned} \tag{37}$$

Their difference proves the score formula in Theorem 1.

C.2 Orthogonality and information

Under Q_0, the x_j and y_k are independent centered variables with variance $\lambda = \nu\Delta$. In particular,

$$\mathbb{E}_0[(x_j y_k)(x_\ell y_v)] = \lambda^2 \mathbf{1}\{j = \ell,\ k = v\}.$$

Applying this identity to (3) proves

$$\mathbb{E}_0 D^2_{h,\Delta} = \frac{2}{\Delta^2}\sum_{j<k} A^2_{k-j} = \frac{2}{\Delta^2}\sum_{d=1}^{m-1}(m-d)A_d^2.$$

The same calculation also gives

$$\mathbb{E}_0(S^{\Delta}_{h,+})^2 = \mathbb{E}_0(S^{\Delta}_{h,-})^2 = m\frac{A_0^2}{\Delta^2} + \frac{J_{\Delta,T}}{2}, \qquad \mathbb{E}_0 S^{\Delta}_{h,+}S^{\Delta}_{h,-} = m\frac{A_0^2}{\Delta^2}. \tag{38}$$

The same-bin component therefore retains first-order information about excitation without distinguishing its orientation. The directional dispersion effect in Proposition 12 instead appears quadratically in a.

For the fine-time score difference, the two causal triangles in the product space $(s,t) \in [0,T]^2$ are disjoint. The isometry for independent compensated Poisson measures yields

$$J_{0,T} = 2\int_{0<s<t<T} h(t-s)^2\,ds\,dt = 2\int_0^T (T-u)h(u)^2\,du.$$

Conditional expectation of the fine-time score gives $0 \le J_{\Delta,T} \le J_{0,T}$. The preceding calculation identifies the resolution dependence that this inequality alone leaves unspecified.

C.3 Exponential evaluation and the order of limits

For $r = e^{-\Delta/\tau}$ and $d \ge 1$, direct integration gives

$$\begin{aligned} A_d &= \frac{1}{\tau}\left(\int_0^{\Delta} e^{s/\tau}\,ds\right)\left(\int_{d\Delta}^{(d+1)\Delta} e^{-t/\tau}\,dt\right) \\ &= \tau(1-r)^2 r^{d-1}. \end{aligned}$$

The same-bin integral is $A_0 = \Delta - \tau(1-r)$. For $z = r^2$,

$$\sum_{d=1}^{m-1}(m-d)z^{d-1} = \frac{(m-1) - mz + z^m}{(1-z)^2}. \tag{39}$$

One proof is to write $m - d = \sum_{v=d}^{m-1} 1$, exchange the finite sums, and evaluate the resulting geometric series. Substitution yields (5). Integration of $2(T-u)\tau^{-2}e^{-2u/\tau}$ over $[0,T]$ gives its exact-time counterpart.

For fixed $r \in (0,1)$,

$$\frac{1}{m}\sum_{d=1}^{m-1}(m-d)r^{2(d-1)} \longrightarrow \frac{1}{1-r^2}.$$

Using $T = m\Delta$ proves (6). The Taylor expansion $1 - e^{-z} = z + O(z^2)$ gives $\Psi(z) \to 1$, and $e^{-z} \to 0$ gives $\Psi(z) \sim 2z^{-3}$. These are iterated limits. With m held fixed while $\Delta/\tau \to \infty$, instead

$$J_{\Delta,m\Delta}(h_\tau) \sim \frac{2(m-1)\tau^2}{\Delta^2}.$$

The two statements agree once the changing exposure $T = m\Delta$ is accounted for. Neither is a uniform approximation over all values of T/Δ.

C.4 Finite-mean kernels and the cubic exponent

For each fixed $\Delta > 0$, Tonelli's theorem gives

$$\sum_{d\geq 1} A_d = \int_0^\Delta \int_\Delta^\infty h(t-s)\,dt\,ds = \int_0^\infty \min\{u,\Delta\}h(u)\,du \leq \Delta. \tag{40}$$

Thus $\sum_{d\geq 1} A_d^2 < \infty$. Rewrite the information formula as

$$\frac{J_{\Delta,m\Delta}(h)}{m\Delta} = \frac{2}{\Delta^3}\sum_{d\geq 1}(1-d/m)_+ A_d^2.$$

Dominated convergence proves the first equality in (7). If $\mu_h < \infty$, monotone convergence in (40) gives $\sum_{d\geq 1} A_d \to \mu_h$ as $\Delta \to \infty$. The adjacent-bin weight is

$$A_1 = \int_0^\Delta uh(u)\,du + \int_\Delta^{2\Delta}(2\Delta - u)h(u)\,du. \tag{41}$$

Its first term tends to μ_h, and its second is between zero and $\int_\Delta^\infty uh(u)\,du$, which tends to zero. Since all weights are nonnegative,

$$A_1^2 \leq \sum_{d\geq 1} A_d^2 \leq \left(\sum_{d\geq 1} A_d\right)^2.$$

Both bounds tend to μ_h^2. Substitution proves Corollary 2. The mean is positive because a normalized Lebesgue density on $(0,\infty)$ cannot concentrate all its mass at zero.

For a kernel supported in $(0, L]$, the same proof gives an exact identity once $\Delta \geq L$. Namely, $A_1 = \mu_h$ and $A_d = 0$ for $d \geq 2$, so $\mathcal{I}_\Delta(h) = 2\mu_h^2/\Delta^3$. These calculations concern the directional tangent of (1); they make no claim for arbitrary recurrent multivariate Hawkes families or infinite-mean kernels.

D Proof of the finite-sample testing bounds

Let $g_{a,+}, g_{a,-}$ be square-root densities of the two count laws with respect to a common dominating measure. The null density is common. By Lemma 9 and (37),

$$g_{a,\pm} = g_0 + \frac{a}{2}S_{h,\pm}^\Delta g_0 + R_{a,\pm}, \qquad \|R_{a,\pm}\|_2 \leq \frac{1}{2}a^2K.$$

Subtracting gives

$$g_{a,+} - g_{a,-} = \frac{a}{2} D_{h,\Delta} g_0 + R_{a,+} - R_{a,-}.$$

The first term has norm $a\sqrt{J}/2$, and the remainder has norm at most $a^2 K$. The triangle and reverse triangle inequalities prove (17), including its positive-part lower bound. All probability densities have root Hellinger distance at most $\sqrt{2}$.

For two laws P, Q, write their affinity as

$$\rho(P,Q) = \int \sqrt{pq} = 1 - \frac{1}{2} H^2(P,Q).$$

The optimal equal-prior error is $R^*(P,Q) = (1 - \mathrm{TV}(P,Q))/2$. Cauchy–Schwarz gives

$$\begin{aligned} \mathrm{TV}(P,Q) &= \frac{1}{2} \int |\sqrt{p} - \sqrt{q}|(\sqrt{p} + \sqrt{q}) \\ &\le \frac{1}{2} \sqrt{(2-2\rho)(2+2\rho)} = \sqrt{1-\rho^2}. \end{aligned}$$

Also $\min\{p,q\} \le \sqrt{pq}$, so $R^*(P,Q) \le \rho/2$. Affinity is multiplicative across independent observations. Consequently

$$\frac{1 - \sqrt{1-\rho^{2N}}}{2} \le R_N^* \le \frac{1}{2} \rho^N. \tag{42}$$

Using $H \le U$ gives $\rho \ge [1 - U^2/2]_+$ in the lower bound. Using $H \ge L$ and $1 - u \le e^{-u}$ gives the upper bound. Finally $\mathrm{TV}(P,Q) \le H(P,Q)$ and

$$H^2(P^{\otimes N}, Q^{\otimes N}) = 2(1 - \rho^N) \le N H^2(P,Q)$$

prove $R_N^* \ge (1 - \sqrt{N} U)_+/2$.

Under $aK \le \sqrt{J}/4$, the bounds on L, U in the main text follow. For example, if $Na^2 J \le 1/9$, then $\sqrt{N} U \le 1/4$, and every test has equal-prior error at least $3/8$. Conversely $R_N^* \le \frac{1}{2} \exp(-Na^2 J/32)$ in this regime. These are explicit constants for constant-error separation. The constructive upper bound in the next section applies without imposing the curvature condition.

E Proof of the score moments and constructive recovery

E.1 Conditional target counts

For each target bin define

$$p_k(s) = \int_{B_k} h(t-s) \mathbf{1}\{s < t\} \, dt, \qquad Z_k = \int_{\mathbb{R}} p_k(s) \, dM(s). \tag{43}$$

Conditional on the complete source, the target counts in distinct bins are independent and

$$\mathbb{E}(Y_k \mid N_1) = \lambda + a Z_k, \qquad \mathrm{Var}(Y_k \mid N_1) = \lambda + a Z_k. \tag{44}$$

The centered source count is $x_j = I_1(\mathbf{1}_{B_j})$. Hence

$$\mathrm{Cov}(X_j, Y_k) = a\nu \int_{B_j} p_k(s) \, ds = \begin{cases} a\nu A_{k-j}, & j < k, \\ a\nu A_0, & j = k, \\ 0, & j > k. \end{cases} \tag{45}$$

Substituting this into the definition of $D_{h,\Delta}$ proves $\mathbb{E}_{a,+} D_{h,\Delta} = aJ/2$. Exchanging labels reverses the sign.

E.2 A SKEW-MATRIX REPRESENTATION

Let B now denote the $m \times m$ skew-symmetric matrix

$$B_{jk} = \begin{cases} A_{k-j}/(\nu\Delta^2), & j < k, \\ -A_{j-k}/(\nu\Delta^2), & j > k, \\ 0, & j = k. \end{cases} \tag{46}$$

This matrix is unrelated to the calibration size B used later. Then $D = x^\top By$ and

$$J = \nu^2\Delta^2 \|B\|_F^2 . \tag{47}$$

Write $y = aZ + \varepsilon$, where $\mathbb{E}(\varepsilon \mid N_1) = 0$. By total variance,

$$\mathrm{Var}(D) = \mathbb{E}\left[\sum_k (x^\top B)_k^2(\lambda + aZ_k)\right] + a^2 \mathrm{Var}(x^\top BZ). \tag{48}$$

The first part involving λ is exactly J. For its remaining part, the Poisson third-moment identity gives

$$\mathbb{E}[x_j x_\ell Z_k] = \nu \mathbf{1}\{j = \ell\} \int_{B_j} p_k(s)\, ds.$$

The diagonal $B_{kk} = 0$ removes same-bin terms. Causality removes $j > k$. If $A_{\max} = \max_{1 \le d < m} A_d$, then

$$\begin{aligned} a\sum_k \mathbb{E}[(x^\top B)_k^2 Z_k] &= a\nu \sum_{j<k} B_{jk}^2 A_{k-j} \\ &\le \frac{a\nu A_{\max}}{2} \|B\|_F^2 \le \frac{a\|h\|_\infty}{2\nu} J, \end{aligned} \tag{49}$$

because $A_{\max} \le \Delta^2 \|h\|_\infty$.

E.3 BOUNDING THE QUADRATIC SOURCE TERM

Define

$$f(s,t) = \sum_{j,k} B_{jk} \mathbf{1}_{B_j}(s) p_k(t).$$

The Poisson product formula yields

$$x^\top BZ = I_2(\mathrm{sym}\, f) + I_1(f_{\mathrm{diag}}) + \nu \int f(s,s)\, ds, \qquad f_{\mathrm{diag}}(s) = f(s,s). \tag{50}$$

Here I_2 is the compensated second-order Poisson integral with the diagonal excluded, and $\mathrm{sym}\, f(s,t) = (f(s,t) + f(t,s))/2$. To obtain (50), expand $I_1(u)I_1(v)$ as the sum over distinct points and the sum over equal points, center each term, and extend linearly. The first- and second-order centered integrals are orthogonal. Their isometries therefore give

$$\mathrm{Var}(x^\top BZ) = 2\nu^2 \|\mathrm{sym}\, f\|_{L^2(\mathbb{R}^2)}^2 + \nu \|f_{\mathrm{diag}}\|_{L^2(\mathbb{R})}^2 . \tag{51}$$

For fixed j, $\sum_k B_{jk} p_k$ is the convolution, up to reflection, of h with $g_j = \sum_k B_{jk} \mathbf{1}_{B_k}$. Since $\|h\|_1 = 1$, the L^2 convolution contraction gives

$$\|f\|_2^2 = \Delta \sum_j \left\| \sum_k B_{jk} p_k \right\|_2^2 \le \Delta \sum_j \|g_j\|_2^2 = \Delta^2 \|B\|_F^2 . \tag{52}$$

The contraction itself follows from Jensen's inequality applied to the probability density h, followed by Fubini. Symmetrization cannot increase the L^2 norm.

For $s \in B_j$, only $k > j$ contributes to $f(s,s)$, because $B_{jj} = 0$ and $p_k(s) = 0$ for $k < j$. Moreover $\sum_{k>j} p_k(s) \le 1$. Weighted Cauchy–Schwarz gives

$$f(s,s)^2 \le \sum_{k>j} B_{jk}^2 p_k(s).$$

After integration,

$$\|f_{\mathrm{diag}}\|_2^2 \le \sum_{j<k} B_{jk}^2 A_{k-j} \le \frac{A_{\max}}{2} \|B\|_F^2 . \tag{53}$$

Combining (51)–(53) yields

$$\mathrm{Var}(x^\top BZ) \le \left(2 + \frac{\|h\|_\infty}{2\nu}\right) J.$$

Together with (48) and (49), this proves the variance bound in Theorem 3. The reverse variance agrees by label exchange. The full stationary source is included throughout, so the variance bound accounts for hidden-past variation.

E.4 Median-of-means sign recovery

A group mean based on k independent windows has mean $aJ/2$ under the forward law and variance at most $C_a J/k$. Chebyshev's inequality implies

$$\mathbb{P}_{a,+}\{\overline{D} \le 0\} \le \frac{4C_a}{ka^2 J}.$$

Under the reverse law the same bound holds with the opposite inequality. Choose $k \ge 16C_a/(a^2 J)$ so that a group has error probability at most $1/4$. The error indicators of b disjoint groups are independent. Hoeffding's inequality for their sum gives

$$\mathbb{P}\{\text{at least half the groups err}\} \le \exp\{-2b(1/2 - 1/4)^2\} = e^{-b/8}.$$

An odd $b \ge 8\log(1/\delta)$ proves the theorem. Any observations left over after forming the groups can be discarded. The moment assumptions needed here are already established in the preceding calculation.

For the exponential kernel, $R_k^x = \sum_{j<k} r^{k-j-1} x_j$ and the analogous identity for R_k^y follow by induction. Substituting $A_{k-j} = \tau(1-r)^2 r^{k-j-1}$ proves (10) and its complexity statement.

F Proof of local asymptotic efficiency

The boundary $a = 0$ requires a one-sided likelihood expansion. An information inequality alone would not establish attainment. Both alternatives approach the boundary from the right, so only the right derivative proved above is required.

For a single branch, let S be its null score, $I = \mathbb{E}_0 S^2$, and $a_N = \gamma/\sqrt{N}$. Use Q_0 as reference measure and write

$$\sqrt{dQ_{a_N}/dQ_0} = 1 + u_N, \qquad u_N = \frac{\gamma}{2\sqrt{N}} S + r_N, \qquad \|r_N\|_{L^2(Q_0)} = O(N^{-1}).$$

For independent null observations $Z_1, \dots, Z_N$, the following three facts give the expansion. First,

$$\sum_{i=1}^N [r_N(Z_i) - \mathbb{E}_0 r_N] = o_{Q_0}(1)$$

by its variance bound $N\|r_N\|_2^2 = O(N^{-1})$. Second,

$$\sum_{i=1}^N u_N(Z_i)^2 \longrightarrow \gamma^2 I/4$$

in probability, by the law of large numbers for S^2 and Cauchy–Schwarz for the r_N terms. Third, $\max_i |u_N(Z_i)| \to 0$ in probability. Indeed, $\sqrt{N}u_N \to \gamma S/2$ in L^2, so its squared tails are uniformly integrable, and for each fixed $\varepsilon > 0$,

$$N\mathbb{P}_0\{|u_N| > \varepsilon\} \le \varepsilon^{-2}\mathbb{E}_0[(\sqrt{N}u_N)^2 \mathbf{1}\{|\sqrt{N}u_N| > \varepsilon\sqrt{N}\}] \to 0.$$

Normalization of the density implies $2\mathbb{E}_0 u_N + \mathbb{E}_0 u_N^2 = 0$. Expanding $2\log(1+u_N)$ and using the three facts therefore gives

$$\log \frac{dQ_{a_N}^{\otimes N}}{dQ_0^{\otimes N}} = \frac{\gamma}{\sqrt{N}} \sum_{i=1}^{N} S(Z_i) - \frac{\gamma^2 I}{2} + o_{Q_0}(1). \tag{54}$$

The logarithmic Taylor remainder is bounded on the event $\max_i |u_N| \le 1/2$ by a constant times $\max_i |u_N| \sum_i u_N^2$, which tends to zero.

Apply (54) jointly to $S_+ = S_{h,+}^{\Delta}$ and $S_- = S_{h,-}^{\Delta}$. Label symmetry gives equal individual information I, while (38) gives

$$\mathrm{Var}_0(S_+ - S_-) = J, \qquad \mathrm{Cov}_0(S_+ - S_-, S_+) = J/2.$$

The multivariate central limit theorem and the joint likelihood expansion imply that under the forward local law,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} D(Z_i) \implies \mathcal{N}(\gamma J/2, J). \tag{55}$$

This change of measure follows by multiplying bounded test functions of the joint Gaussian limit by the likelihood ratio in (54). The limiting exponential tilt has expectation one. The nonnegative finite-sample likelihood ratios also have expectation one, so convergence to that limit gives uniform integrability and validates the change of measure. Exponential tilting of a centered joint Gaussian by γS_+ shifts the mean of D by $\gamma \,\mathrm{Cov}_0(D, S_+) = \gamma J/2$. It leaves its variance unchanged. The reverse local law shifts the mean by $-\gamma J/2$.

Subtracting the two expansions (54) cancels the equal quadratic terms. The log likelihood ratio between the local forward and reverse laws is

$$\frac{\gamma}{\sqrt{N}} \sum_{i} D(Z_i) + o_{Q_0}(1).$$

The same remainder is negligible under either local law by the preceding change-of-measure argument. Since $J > 0$, the limiting Gaussian has no atom at zero. Both the optimal likelihood-ratio decision and the score-sign decision therefore have limiting error $\Phi(-\gamma\sqrt{J}/2)$ under each orientation. Their equal-prior risks converge to the same value, proving Corollary 4.

G Directional score geometry and its rank collapse

Let $\tau_1, \ldots, \tau_q$ be a fixed exponential dictionary with $q \ge 2$. Write $v_d = (A_d(\tau_\ell))_{\ell=1}^{q}$. The null Gram matrix of its directional scores is

$$G_{\Delta,T} = \frac{2}{\Delta^2} \sum_{d=1}^{m-1} (m-d) v_d v_d^{\top}. \tag{56}$$

Proposition 10 (Directional rank collapse)**.** *Assume $m \ge 2$. If $\tau_{\max} = \max_\ell \tau_\ell$ and $r_{\max} = e^{-\Delta/\tau_{\max}}$, then*

$$\frac{\lambda_2(G_{\Delta,T})}{\lambda_1(G_{\Delta,T})} \le \frac{r_{\max}^2}{1 - r_{\max}^2}. \tag{57}$$

For $m = 2$, the Gram matrix has rank one exactly. For $m = 1$, $G_{\Delta,T} = 0$ and the eigenvalue ratio is undefined.

At coarse resolution, a leading directional combination can retain information while orthogonal combinations have exponentially less. The statement concerns the directional tangent Gram matrix, not global identification of all kernel parameters. Uncertainty about detailed timescales can therefore grow while evidence of dependence remains.

Proof. For two kernels h, g, orthogonality of the Poisson products used in Appendix C gives

$$\mathbb{E}_0 D_{h,\Delta} D_{g,\Delta} = \frac{2}{\Delta^2} \sum_{d=1}^{m-1} (m-d) A_d(h) A_d(g).$$

This proves (56). The leading $d = 1$ term has rank one and nonzero eigenvalue $2(m-1)\|v_1\|^2/\Delta^2$. Thus

$$\lambda_1(G) \geq \frac{2(m-1)}{\Delta^2} \|v_1\|^2 . \tag{58}$$

The min–max principle, evaluated on $v_1^{\perp}$, gives

$$\lambda_2(G) \leq \left\| \frac{2}{\Delta^2} \sum_{d=2}^{m-1} (m-d) v_d v_d^{\top} \right\|_{\mathrm{op}} \leq \frac{2(m-2)}{\Delta^2} \sum_{d=2}^{\infty} \|v_d\|^2 .$$

For each dictionary element, $A_d(\tau_\ell) = A_1(\tau_\ell) r_\ell^{d-1}$ with $r_\ell = e^{-\Delta/\tau_\ell} \leq r_{\max}$. Consequently

$$\sum_{d=2}^{\infty} \|v_d\|^2 = \sum_{\ell} A_1(\tau_\ell)^2 \frac{r_\ell^2}{1-r_\ell^2} \leq \|v_1\|^2 \frac{r_{\max}^2}{1-r_{\max}^2}.$$

Dividing by (58) proves the slightly stronger factor $(m-2)/(m-1)$ times the right side of (57). Dropping that factor gives the stated bound. For $m = 2$ there is no remainder, so the matrix has rank one. Repeated dictionary timescales do not invalidate the bound.

The result concerns the unstandardized score Gram matrix. Rescaling its coordinates changes the eigenvalue ratio, so the bound is not invariant to arbitrary feature normalization. In the specified coordinates, removing the first lag-weight vector leaves directional energy bounded by the displayed exponential factor relative to that leading component.

H Proof of the random-jitter identities

Let D^{jit} be the fine-time score after placing every observed event independently and uniformly inside its recorded bin. Average L independent placements conditional on the counts.

Proposition 11 (Jitter efficiency)**.** *Under the null,*

$$\mathbb{E}_0[D^{\mathrm{jit}} \mid C_\Delta] = D_{h,\Delta}, \qquad \mathrm{Var}_0(\overline{D}_L^{\mathrm{jit}}) = J_{\Delta,T} + \frac{J_{0,T} - J_{\Delta,T}}{L}. \tag{59}$$

The local efficacy of the averaged-jitter statistic is $J_{\Delta,T}^2/[J_{\Delta,T} + (J_{0,T} - J_{\Delta,T})/L]$.

A single random placement therefore loses a factor $J_{\Delta,T}/J_{0,T}$ relative to the projected score. To retain a fixed fraction of its efficacy requires L of order $J_{0,T}/J_{\Delta,T}$ when this ratio is large. Uniform jitter is a randomization rule, not the correct latent-time posterior under excitation.

Proof. At the null, the conditional law of event positions given all bin counts is exactly the product of uniform within-bin laws. Hence a fresh jittered window has the fine-time null distribution and

$$\mathbb{E}_0[D^{\mathrm{jit}} \mid C_\Delta] = D_{h,\Delta}, \qquad \mathrm{Var}_0(D^{\mathrm{jit}}) = J_{0,T}.$$

The law of total variance gives

$$\mathbb{E}_0 \,\mathrm{Var}_0(D^{\mathrm{jit}} \mid C_\Delta) = J_{0,T} - J_{\Delta,T}.$$

Conditional independence of the L placements divides only this latter variance by L. The conditional mean remains $D_{h,\Delta}$. A second application of total variance proves (59).

The conditional expectation is also an algebraic identity for every fixed count array under the chosen uniform randomization. Thus, even when the observed counts are generated at $a > 0$,

$$\mathbb{E}_{a,\pm} \overline{D}_L^{\mathrm{jit}} = \mathbb{E}_{a,\pm} D_{h,\Delta} = \pm a J_{\Delta,T}/2.$$

For the experiment augmented by independent random seeds, the score of the observed data remains $S^{\Delta}_{h,\pm}$. The same local Gaussian tilting argument used in Appendix F therefore gives a limiting standardized mean shift

$$\frac{\gamma J_{\Delta,T}}{2\sqrt{J_{\Delta,T} + (J_{0,T} - J_{\Delta,T})/L}}$$

for the sum of averaged-jitter scores. The squared numerator-to-variance factor is the stated local efficacy. To attain at least a fraction $\eta \in (0,1)$ of the projected-score efficacy, it is sufficient and necessary at this local-variance level that

$$L \geq \frac{\eta}{1-\eta}\left(\frac{J_{0,T}}{J_{\Delta,T}} - 1\right).$$

This expression is meaningful when $J_{\Delta,T} > 0$. A model with a single observed bin has no first-order directional score to recover by uniform randomization.

For numerical evaluation of the fine exponential score, sorting the two event streams gives the pair sum plus exact boundary terms,

$$D_{h_\tau,0} = \frac{1}{\nu\tau}\left(\sum_{s\in N_1,\ t\in N_2,\ s<t} e^{-(t-s)/\tau} - \sum_{t\in N_2,\ s\in N_1,\ t<s} e^{-(s-t)/\tau}\right) + \sum_{s\in N_1}\left(e^{-(T-s)/\tau} - e^{-s/\tau}\right) - \sum_{t\in N_2}\left(e^{-(T-t)/\tau} - e^{-t/\tau}\right), \tag{60}$$

where every event in the sums lies in $[0,T]$. The boundary terms arise from the compensated measures and cannot be omitted in a finite-window comparison.

I Exact single-bin dispersion moments

For one bin, let $\lambda = \nu\Delta$ and

$$V = (Y-\lambda)^2 - Y - \big[(X-\lambda)^2 - X\big]. \tag{61}$$

Proposition 12 (A surviving second-order channel)**.** *For the exponential model,*

$$\mathbb{E}_{a,\pm} V = \pm a^2 \nu\{\Delta - \tau(1-e^{-\Delta/\tau})\}. \tag{62}$$

At fixed ν, τ, a, as $\Delta \to \infty$,

$$\frac{\mathbb{E}_{a,+} V}{\nu\Delta} \to a^2, \qquad \frac{\mathrm{Var}_{a,\pm} V}{(\nu\Delta)^2} \to 4 + 2a^4. \tag{63}$$

An exact finite-Δ variance is given in (72)*. For fixed error probability, a median-of-means sign test based on independent single-bin windows needs $O(a^{-4})$ replicates in the coarse-bin limit.*

Direction can therefore remain recoverable without within-bin ordering. The single-bin first-order information is zero, but the dispersion contrast has mean separation quadratic in a. The phenomenon is consistent with the second-order count structure retained by coarse-grained Hawkes models (Koyama, 2025). Our calculation quantifies it and makes the scope of the local temporal law explicit.

Exact finite-bin calculation. Let $p(s) = \int_0^{\Delta} h_\tau(t-s)\mathbf{1}\{s<t\}\,dt$ and $B_k = \int_{\mathbb{R}} p(s)^k\,ds$. Directly,

$$p(s) = \begin{cases} (1-r)e^{s/\tau}, & s<0, \\ 1 - e^{-(\Delta-s)/\tau}, & 0 \leq s \leq \Delta, \\ 0, & s > \Delta, \end{cases} \qquad r = e^{-\Delta/\tau}. \tag{64}$$

In particular,

$$B_k = \Delta + \tau\left\{\frac{(1-r)^k}{k} + \sum_{j=1}^{k}(-1)^j\binom{k}{j}\frac{1-r^j}{j}\right\}, \quad B_1 = \Delta, \quad B_2 = \Delta - \tau(1-r). \tag{65}$$

The first term inside braces integrates the negative-time part. Expanding $(1-e^{-u/\tau})^k$ and integrating u from zero to Δ gives the second part.

I.1 MARGINAL TARGET CUMULANTS

Given the source, Y is Poisson with random mean

$$\Lambda = \nu(1-a)\Delta + a\int p(s)\,dN_1(s).$$

Its unconditional cumulant generating function is

$$K_Y(t) = \nu(1-a)\Delta(e^t-1) + \nu\int_{\mathbb{R}}\left[\exp\{ap(s)(e^t-1)\}-1\right]ds. \tag{66}$$

With $\lambda = \nu\Delta$, its first four cumulants are

$$\begin{aligned}
\kappa_1(Y) &= \lambda,\\
\kappa_2(Y) &= \lambda + a^2\nu B_2,\\
\kappa_3(Y) &= \lambda + 3a^2\nu B_2 + a^3\nu B_3,\\
\kappa_4(Y) &= \lambda + 7a^2\nu B_2 + 6a^3\nu B_3 + a^4\nu B_4.
\end{aligned} \tag{67}$$

The source count X is Poisson with mean λ. Put $C_X = (X-\lambda)^2 - X$ and $C_Y = (Y-\lambda)^2 - Y$, so $V = C_Y - C_X$. Since $\mathbb{E}C_X = 0$ and $\mathbb{E}C_Y = \kappa_2(Y) - \lambda$, (62) follows.

For a variable W of mean λ, let $z = W - \lambda$. Expanding $(z^2 - z - \lambda)^2$ and using $\mathbb{E}z^4 = \kappa_4 + 3\kappa_2^2$, $\mathbb{E}z^3 = \kappa_3$, gives

$$\mathrm{Var}((W-\lambda)^2 - W) = \kappa_4 - 2\kappa_3 + \kappa_2 + 2\kappa_2^2. \tag{68}$$

Thus $\mathrm{Var}(C_X) = 2\lambda^2$ and, writing $M = a^2\nu B_2$,

$$\mathrm{Var}(C_Y) = 2(\lambda+M)^2 + 2a^2\nu B_2 + 4a^3\nu B_3 + a^4\nu B_4. \tag{69}$$

I.2 THE CROSS TERM

The covariance between C_X and C_Y must also be included. Their joint cumulant generating function is

$$K_{X,Y}(s,t) = \nu(1-a)\Delta(e^t-1) + \nu\int_{\mathbb{R}}\left[\exp\{s\mathbf{1}_{[0,\Delta]}(u) + ap(u)(e^t-1)\}-1\right]du. \tag{70}$$

Write κ_{jk} for its mixed cumulant with j derivatives in s and k in t. Because $\mathbf{1}_{[0,\Delta]}^j = \mathbf{1}_{[0,\Delta]}$ for every positive integer j,

$$\kappa_{21} = \kappa_{11}, \qquad \kappa_{22} = \kappa_{12}, \qquad \kappa_{11} = a\nu\int_0^\Delta p(u)\,du = a\nu B_2.$$

The centered fourth mixed moment satisfies $\mathrm{Cov}((X-\lambda)^2,(Y-\lambda)^2) = \kappa_{22} + 2\kappa_{11}^2$. Expanding $\mathrm{Cov}(C_X, C_Y)$ then cancels the remaining mixed cumulants and gives

$$\mathrm{Cov}(C_X, C_Y) = \kappa_{22} - \kappa_{21} - \kappa_{12} + \kappa_{11} + 2\kappa_{11}^2 = 2(a\nu B_2)^2. \tag{71}$$

Combining (69) and (71) yields the exact variance

$$\boxed{\mathrm{Var}(V) = 2(\lambda+M)^2 + 2\lambda^2 - 4(a\nu B_2)^2 + 2a^2\nu B_2 + 4a^3\nu B_3 + a^4\nu B_4.} \tag{72}$$

The reverse law changes the sign of V without changing its variance.

For each fixed k, $B_k/\Delta \to 1$ as $\Delta \to \infty$. Hence $M/\lambda \to a^2$. Dividing (72) by λ^2 gives

$$2(1+a^2)^2 + 2 - 4a^2 = 4 + 2a^4,$$

because the remaining terms have order Δ rather than Δ^2. This proves (63). For N independent single-bin windows, Chebyshev's inequality bounds the mean-sign error by $\mathrm{Var}(V)/(N(\mathbb{E}V)^2)$. A median-of-means version improves the dependence on the desired error probability to a logarithm exactly as in Appendix E. Since the variance-to-squared-mean ratio tends to $(4+2a^4)/a^4$, the required number of replicates is $O(a^{-4})$ at fixed confidence in the coarse-bin limit.

This establishes a constructive direction-recovery channel. It does not assert convergence of the Hawkes likelihood or its Kullback–Leibler divergence to a Gaussian likelihood. A count central limit theorem by itself would not justify such a divergence statement.

J CONFIDENCE-SET VALIDITY AND ADAPTIVE REPRESENTATIONS

J.1 FINITE-GRID TEST INVERSION

Condition on the entire independent training procedure, including all choices of network architecture, feature normalization, and stopping time. Fix the true $\theta \in \Theta$. The $B + 1$ discrepancies

$$t_\theta(Z),\ t_\theta(Z_{\theta 1}), \ldots, t_\theta(Z_{\theta B})$$

are identically distributed and exchangeable, because the observed and calibration datasets have the same window count, observation resolution, and parameter. In the absence of ties, the rank of the observed discrepancy in descending order is uniform on $\{1, \ldots, B + 1\}$. With ties, counting every calibration discrepancy at least as large as the observed value makes the resulting rank conservative. One can formally add independent continuous tie-breakers and note that the nonrandomized rank is no smaller than the corresponding randomized rank. Thus

$$\mathbb{P}_\theta\{p_\theta(Z) \le \alpha\} \le \frac{\lfloor \alpha(B+1) \rfloor}{B+1} \le \alpha. \tag{73}$$

Coverage requires only the rank at the true parameter. There is no union bound over Θ in (20). Calibration simulations at distinct candidate values may be independent, as in the experiments, but their mutual independence is not needed for this particular inclusion guarantee. Independence from training and exchangeability with the observed data at the true candidate are needed.

If a nonempty confidence set has one common direction, then on the event $\theta \in \mathcal{C}_\alpha$ the common direction must be correct. Hence a false directional report is contained in the event $\theta \notin \mathcal{C}_\alpha$, which proves its unconditional error control. A confidence set can be empty, even under a correctly specified model. The argument requires nonemptiness. The implementation checks it before reporting agreement among candidates.

J.2 SIMULTANEOUS KERNEL AND HIERARCHY REPORTS

For any family of functions $F_\theta : \mathcal{U} \to \mathbb{R}$, on the event $\theta \in \mathcal{C}_\alpha$,

$$\min_{\vartheta \in \mathcal{C}_\alpha} F_\vartheta(u) \le F_\theta(u) \le \max_{\vartheta \in \mathcal{C}_\alpha} F_\vartheta(u) \qquad \text{for every } u \in \mathcal{U}.$$

The implication holds on one event for every u, so no gridwise union bound is needed. Taking $F_\theta(u) = \sigma a(1 - e^{-u/\tau})$ proves the envelope statement. Taking other candidate-defined functionals gives analogous simultaneous reports.

More generally, let $\mathcal{P}_1, \ldots, \mathcal{P}_R$ be a prespecified hierarchy of successively finer partitions of Θ. Given a nonempty $\mathcal{C}_\alpha$, select any level r such that some cell $A_r \in \mathcal{P}_r$ contains the entire confidence set. This level can be selected from the data, including by maximizing r. On the truth-in-set event, $\theta \in A_r$. Therefore the probability of returning any false cell in this hierarchy is at most α. For the experimental grid, the first useful partition distinguishes null, forward, and reverse, and the finer partition distinguishes individual non-null timescales. A report of a singleton null does not identify a meaningful null timescale.

These statements justify a set-valued output representation and adaptive reporting. They do not prove that zeroing arbitrary neural hidden units leaves statistical information unchanged. Optimizing the discrepancy after viewing the test observation also requires accounting for that additional selection.

J.3 INTERSECTING A RELATIVE SCORE AND AN ABSOLUTE-FIT TEST

Let $\mathcal{C}^{\mathrm{rel}}_{\alpha/2}$ use the decoder compatibility score, and let $\mathcal{C}^{\mathrm{abs}}_{\alpha/2}$ use a second compatibility score trained without the calibration data. Each individual rank inversion misses the true grid candidate with probability at most $\alpha/2$. The guarded set is their intersection. Its miss event is the union of the two individual miss events, so its probability is at most α by the union bound. The two tests may use the same calibration bank, and no independence between their scores is required. A unanimous report from the intersection is valid whenever the true candidate remains in it. An empty intersection gives no report.

This intersection is not necessarily a subset of the original relative-score set at level α, because both tests in the intersection use the less stringent rejection level $\alpha/2$. It can therefore report more often, or make more off-grid errors, in some settings. The construction preserves the stated on-grid marginal guarantee, not a uniform robustness guarantee. The weak-amplitude challenge in Appendix R illustrates that distinction.

J.4 A LIMITED ROBUSTNESS EXTENSION

Suppose the actual dataset law is P_* and some candidate law $Q_{\theta_0}^{\otimes N}$ satisfies

$$\mathrm{TV}(P_*, Q_{\theta_0}^{\otimes N}) \le \eta.$$

Use the same grid-generated calibration banks and trained decoder under both experiments. Their joint laws differ only through the observed dataset, so their total variation distance is at most η. It follows that

$$\mathbb{P}_*\{\theta_0 \in \mathcal{C}_\alpha\} \ge 1 - \alpha - \eta.$$

If a target functional F_* also satisfies $\sup_u |F_*(u) - F_{\theta_0}(u)| \le \varepsilon$, enlarging both sides of the confidence envelope by ε gives simultaneous coverage at least $1 - \alpha - \eta$. Here η is a discrepancy for the entire N-window dataset. A small per-window discrepancy need not remain small after taking products. This extension is not an off-grid validity guarantee without a justified bound on η and ε.

K THE SECOND-ORDER DIRECTIONAL EXPERIMENT

This section concerns a single observed bin, $m = 1$ and $\Delta = T$. The window is held fixed in the local-signal limit. Write p_a^+ and p_a^- for the two count probability mass functions and p_0 for their common independent-Poisson null. All derivatives below are right derivatives at zero when evaluated at the endpoint.

K.1 A THIRD ROOT-DENSITY DERIVATIVE WITH A FINITE BOUND

The following extension of Lemma 9 supplies the remainder needed for a second-order comparison. Define

$$\begin{aligned}
M_6 &= \nu \int h^6 + 15\nu^2 \left(\int h^4\right)\left(\int h^2\right) + 10\nu^2 \left(\int h^3\right)^2 + 15\nu^3 \left(\int h^2\right)^3, \\
A_6 &= M_6 \left\{ \frac{T}{(c\nu)^5} + \frac{25T^2}{(c\nu)^4} + \frac{15T^3}{(c\nu)^3} \right\}, \\
B_3 &= M_6 \left\{ \frac{T}{(c\nu)^5} + \frac{3T^2}{(c\nu)^4} + \frac{T^3}{(c\nu)^3} \right\}, \\
C_3 &= M_6 \left\{ \frac{T}{(c\nu)^5} + \frac{T^2}{(c\nu)^4} \right\}, \qquad K_3 = \frac{13}{8}\sqrt{A_6} + \frac{9}{4} A_6^{1/6} B_3^{1/3} + \sqrt{C_3}.
\end{aligned} \tag{74}$$

Here $c = 1 - \bar{a}$. The symbols B_3, C_3 in (74) are derivative bounds, not the exposure integrals B_k of Appendix I.

Lemma 13 (Third-order root-density control)**.** *For either directional branch and every marginal observation of the augmented stationary experiment, the second L^2 derivative of its square-root density is absolutely continuous, and*

$$\operatorname*{ess\,sup}_{0 \le a \le \bar{a}} \|g_a'''\|_2 \le K_3, \qquad \left\| g_a - g_0 - a g_0' - \frac{a^2}{2} g_0'' \right\|_2 \le \frac{a^3 K_3}{6}. \tag{75}$$

Proof. Use the complete experiment and S_a, W_a, q_a from (26). Put $V_a = \int_0^T q_a(t)^3 \, dN_2(t)$ and $H_k = \int_0^T |q_a(t)|^k \lambda_a(t) \, dt$. Conditional on the source, the sixth centered Poisson moment gives

$$\mathbb{E}(S_a^6 \mid N_1) \le H_6 + 15 H_4 H_2 + 10 H_3^2 + 15 H_2^3.$$

Also $\mathbb{E}(W_a^3 \mid N_1) = H_6 + 3H_4H_2 + H_2^3$ and $\mathbb{E}(V_a^2 \mid N_1) \le H_6 + H_3^2$. Since $\lambda_a \ge c\nu$,

$$H_k \le (c\nu)^{1-k} \int_0^T |Z(t)|^k \, dt.$$

The centered sixth moment of the stationary shot noise is M_6, by the Poisson cumulant formula. Hölder's inequality gives

$$\mathbb{E}H_2^3 \le \frac{T^3 M_6}{(c\nu)^3}, \quad \mathbb{E}H_4H_2,\ \mathbb{E}H_3^2 \le \frac{T^2 M_6}{(c\nu)^4}, \quad \mathbb{E}H_6 \le \frac{TM_6}{(c\nu)^5}.$$

Thus $\mathbb{E}S_a^6 \le A_6$, $\mathbb{E}W_a^3 \le B_3$, and $\mathbb{E}V_a^2 \le C_3$, uniformly in a. These constants are finite because a bounded normalized kernel belongs to every L^k, $k \ge 1$.

For an observed marginal Y, define

$$R_1 = \mathbb{E}_a[S_a \mid Y], \quad R_2 = \mathbb{E}_a[S_a^2 - W_a \mid Y], \quad R_3 = \mathbb{E}_a[S_a^3 - 3S_aW_a + 2V_a \mid Y].$$

Differentiating the marginal density three times and then taking its square root yields

$$\frac{g_a'''}{g_a} = \frac{1}{2}R_3 - \frac{3}{4}R_1R_2 + \frac{3}{8}R_1^3. \tag{76}$$

Conditional Jensen, Minkowski, and Hölder inequalities imply

$$\begin{aligned} \|R_3\|_{L^2(P_a)} &\le \sqrt{A_6} + 3A_6^{1/6}B_3^{1/3} + 2\sqrt{C_3}, \\ \|R_1R_2\|_{L^2(P_a)} &\le \sqrt{A_6} + A_6^{1/6}B_3^{1/3}, \\ \|R_1^3\|_{L^2(P_a)} &\le \sqrt{A_6}. \end{aligned}$$

Substitution gives K_3.

The differentiations are justified by extending the argument of Lemma 9 by one derivative. The complete square-root derivative satisfies

$$f_a'''/f_a = S_a^3/8 - 3S_aW_a/4 + V_a,$$

whose L^2 norm is uniformly finite by the bounds just proved. Consequently the complete density has an L^1 third derivative, using $(f_a^2)''' = 2f_af_a''' + 6f_a'f_a''$ and Cauchy–Schwarz. Conditional expectation preserves its integrated L^1 derivative identities. Positivity of marginal densities on the compact parameter interval permits the scalar square-root chain rule. The uniform bound in (76), followed by Fubini, upgrades the integrated identities to L^2 identities. In particular g_a'' has an absolutely continuous representative and a trace at zero. Taylor's integral remainder is $\int_0^a (a-u)^2 g_u'''/2 \, du$, giving (75). □

K.2 Identifying the second directional derivative

Let $p_T(s) = \int_0^T h(t-s)\mathbf{1}\{s < t\} \, dt$, $\beta_k = \int_{\mathbb{R}} p_T(s)^k \, ds$, and $\chi_k = \int_0^T p_T(s)^k \, ds$. Then $0 \le p_T \le 1$, $\beta_1 = T$, and $\chi_1 = A_0$. For $u = z_x - 1$, $v = z_y - 1$, the joint probability generating function has logarithm

$$\log G_a^+(z_x, z_y) = \lambda(u+v) + a\nu\chi_1 uv + \frac{a^2\nu}{2}v^2\{\beta_2 + \chi_2 u\} + O(a^3). \tag{77}$$

The expansion follows by separating source locations inside $[0, T]$ from those before zero in the exact cluster generating function

$$\log G_a^+ = \nu(1-a)T(z_y - 1) + \nu \int_{\mathbb{R}} \left[z_x^{\mathbf{1}_{[0,T]}(s)} \exp\{ap_T(s)(z_y - 1)\} - 1 \right] ds. \tag{78}$$

The integral is finite. For fixed z_x, z_y in a bounded neighborhood of one, all its first three derivatives are dominated using $p_T^k \le p_T$ and $\int p_T = T$. Equation (77) can therefore be differentiated and coefficients identified. The same derivatives also follow from the density regularity in Lemma 13.

For $Z \sim \mathrm{Pois}(\lambda)$, set $c_1(Z) = Z - \lambda$ and $c_2(Z) = (Z-\lambda)^2 - Z$. The coefficient of z^k in $(z-1)\exp\{\lambda(z-1)\}$ is $\mathbb{P}(Z=k)c_1(k)/\lambda$, and its counterpart for $(z-1)^2$ is $\mathbb{P}(Z=k)c_2(k)/\lambda^2$. Expanding G_a^+ rather than just its logarithm gives

$$\frac{(p_a^+)''|_0}{p_0} = \frac{\nu\beta_2}{\lambda^2}c_2(Y) + \frac{\nu\chi_2}{\lambda^3}c_1(X)c_2(Y) + \frac{(\nu\chi_1)^2}{\lambda^4}c_2(X)c_2(Y). \tag{79}$$

The last term comes from the squared first log-generating-function derivative. It is symmetric under exchanging X, Y and cancels between the two orientations. Their first density derivatives are identical. Hence

$$(p_+)''|_0 - (p_-)''|_0 = 2D^{(2)}p_0, \qquad (\sqrt{p_+})''|_0 - (\sqrt{p_-})''|_0 = D^{(2)}\sqrt{p_0}.$$

Taylor expansion and Lemma 13 now give the quantitative root-density comparison

$$\left\| \sqrt{p_a^+} - \sqrt{p_a^-} - \frac{a^2}{2}D^{(2)}\sqrt{p_0} \right\|_2 \le \frac{a^3 K_3}{3}. \tag{80}$$

This proves the expansion asserted in Theorem 5. It also gives a finite-sample Hellinger envelope with

$$L_2 = (a^2\sqrt{J^{(2)}}/2 - a^3K_3/3)_+, \qquad U_2 = a^2\sqrt{J^{(2)}}/2 + a^3K_3/3.$$

Substituting L_2, U_2 into (18) is valid for every $a \le \bar{a}$. The envelope is explicit, but its conservative constant need not give useful bounds at the experimental amplitudes. For example, $aK_3 \le 3\sqrt{J^{(2)}}/4$ suffices to compare H to $a^2\sqrt{J^{(2)}}$ within fixed factors.

K.3 Information and exact means

Under p_0, $\mathbb{E}c_1 = \mathbb{E}c_2 = \mathbb{E}c_1c_2 = 0$, $\mathbb{E}c_1^2 = \lambda$, and $\mathbb{E}c_2^2 = 2\lambda^2$. Independence of X, Y therefore gives

$$\mathrm{Var}_0(c_2(Y) - c_2(X)) = 4\lambda^2,$$
$$\mathrm{Var}_0(c_1(X)c_2(Y) - c_1(Y)c_2(X)) = 4\lambda^3.$$

The two displayed variables are orthogonal. These identities prove the formula for $J^{(2)}$.

The exact mean follows from the target excess variance, which is $a^2\nu\beta_2$, so $\mathbb{E}_+[c_2(Y) - c_2(X)] = a^2\nu\beta_2$. Since both means equal λ,

$$\mathbb{E}_+[c_1(X)c_2(Y) - c_1(Y)c_2(X)] = \kappa_{12} - \kappa_{21} = a^2\nu\chi_2.$$

Indeed the joint cumulant formula (78) gives $\kappa_{12} = a\nu\chi_1 + a^2\nu\chi_2$ and $\kappa_{21} = a\nu\chi_1$. Substitution yields

$$\mathbb{E}_{a,\pm}D^{(2)} = \pm\frac{a^2J^{(2)}}{2} \tag{81}$$

for every admissible a, not merely to leading order.

K.4 Sharp testing risk at the fourth-root scale

Put $m_a = (p_a^+ + p_a^-)/2$ and $u_a = (p_a^+ - p_a^-)/(p_a^+ + p_a^-)$. Then $|u_a| \le 1$, $\mathbb{E}_{m_a}u_a = 0$, and

$$\frac{u_a\sqrt{m_a}}{a^2} = \frac{\sqrt{p_a^+} - \sqrt{p_a^-}}{a^2}\,\frac{\sqrt{p_a^+} + \sqrt{p_a^-}}{\sqrt{2(p_a^+ + p_a^-)}} \longrightarrow \frac{1}{2}D^{(2)}\sqrt{p_0} \quad \text{in } L^2.$$

The second factor is bounded by one and converges pointwise to one. The L^2 convergence follows from (80), truncation of its fixed L^2 limit, and bounded convergence. In particular $\mathbb{E}_{m_a}u_a^2/a^4 \to J^{(2)}/4$. The same truncation argument shows, for $k = 3, 4$, $\mathbb{E}_{m_a}|u_a|^k = o(a^4)$ and the corresponding Lindeberg condition. One may verify the truncation directly on finite subsets of the count lattice, where p_0 is everywhere positive, and control the complement by the preceding L^2 convergence.

Set $a_N = \gamma N^{-1/4}$ and $\kappa = \gamma^2\sqrt{J^{(2)}}/2$. For independent observations under m_{a_N}, the triangular-array central limit theorem and the fourth-moment bound give

$$\sum_{i=1}^N u_{a_N}(Z_i) \Rightarrow N(0, \kappa^2), \qquad \sum_{i=1}^N u_{a_N}(Z_i)^2 \longrightarrow \kappa^2 \quad \text{in probability.}$$

Moreover $\max_i |u_{a_N}(Z_i)| \to 0$ in probability. Taylor expansion of $\log(1 \pm u)$ is valid on that event, and the sum of the absolute cubic remainders converges to zero in probability. Thus

$$\log \frac{(p^{\pm}_{a_N})^{\otimes N}}{m^{\otimes N}_{a_N}} = \pm \sum_i u_{a_N}(Z_i) - \kappa^2/2 + o_{m^{\otimes N}_{a_N}}(1).$$

The limiting likelihood ratios are positive lognormal variables with mean one. Truncation of the likelihood ratios therefore yields contiguity and the usual Gaussian change of measure. Under the forward law, the forward-versus-reverse log likelihood ratio tends to $N(2\kappa^2, 4\kappa^2)$, with the signs reversed under the reverse law. Its threshold-zero test has limiting risk $\Phi(-\kappa)$, which is the optimal equal-prior risk.

It remains to show that the stated statistic attains this risk. Its mean is exactly (81). Its variance converges to $J^{(2)}$ as $a \to 0$. Uniform moments of order greater than two follow because $D^{(2)}$ is a degree-three polynomial in counts, and these counts have uniformly bounded moments of every fixed order on a compact amplitude interval. This can be seen directly from (78) and $p_T \le 1$. A triangular-array Lyapunov theorem gives

$$\frac{1}{\sqrt{N}} \sum_i D^{(2)}(Z_i) \Rightarrow N(\pm\gamma^2 J^{(2)}/2, J^{(2)}).$$

The sign test consequently attains $\Phi(-\kappa)$. This proof compares two moving directional alternatives through their moving symmetric mixture. It does not incorrectly assert contiguity of either product alternative to the independent-process null at the fourth-root scale.

K.5 Exponential evaluation and the coarse-bin limit

For h_τ, direct integration yields

$$\beta_2 = T - \tau(1-r), \qquad \chi_2 = T - 2\tau(1-r) + \frac{\tau}{2}(1-r^2), \qquad r = e^{-T/\tau}.$$

The first quantity is the B_2 of Appendix I. The second excludes parents before zero, which accounts for the difference between the two integrals. More generally, for any normalized causal kernel, $\chi_2 = \int_0^T F(u)^2\,du$ with $F(u) = \int_0^u h$. Since $F(u) \to 1$, Cesàro averaging gives $\chi_2/T \to 1$. Also $\chi_2 \le \beta_2 \le \int p_T = T$, so $\beta_2/T \to 1$. At fixed $\nu > 0$, therefore, $J^{(2)} \to 1$ as $T \to \infty$. This limit is per independent bin. At the second-order parameterization, information per unit observed time is of order $1/T$. A larger T therefore incurs a larger exposure cost.

The dispersion statistic alone has limiting mean-to-standard-deviation shift $\gamma^2\beta_2/(2T)$ at the same local scale. Its squared relative efficacy is

$$\frac{\beta_2^2/T^2}{J^{(2)}}.$$

The mixed third-degree term can matter at small expected count νT. It becomes negligible relative to the dispersion term for long bins at a fixed rate. Neither conclusion identifies kernel shape from second-order information alone.

L Estimating the common mean rate without local efficiency loss

At the null, the score for the common rate is

$$W_\nu = \frac{1}{\nu} \sum_{j=0}^{m-1} (x_j + y_j), \qquad \mathbb{E}_0 W_\nu^2 = 2T/\nu.$$

The directional score in (3) is a sum of products of independent centered Poisson counts. Hence $\mathbb{E}_0 D_{h,\Delta} W_\nu = 0$. For one bin, the orthogonality of the first two Poisson polynomials used above also gives $\mathbb{E}_0 D^{(2)} W_\nu = 0$. The plug-in assertion requires the following additional argument.

For any positive argument v, expanding the centered products in the first-order statistic gives

$$D_{h,\Delta}(v) = \frac{\mathsf{B}}{v} - \mathsf{L},$$
$$\mathsf{B} = \Delta^{-2} \sum_{j<k} A_{k-j}(X_j Y_k - Y_j X_k),$$
$$\mathsf{L} = \Delta^{-1} \sum_{j<k} A_{k-j}(X_j + Y_k - Y_j - X_k). \tag{82}$$

Stationarity and equal means imply $\mathbb{E}\mathsf{L} = 0$. Along $a_N = \gamma/\sqrt{N}$, the exact score mean gives $\mathbb{E}\mathsf{B} = \pm\nu a_N J/2$. Uniform count moments imply $\sum_i \mathsf{B}_i/\sqrt{N} = O_{P_{a_N,\pm}}(1)$. The pooled rate estimate is unbiased before its vanishing lower truncation and has variance $O(1/N)$, because windows are independent and their fixed-length total counts have uniformly bounded variance. Thus $\hat{\nu}_N - \nu = O_P(N^{-1/2})$, and

$$\frac{1}{\sqrt{N}} \sum_i \{D_i(\hat{\nu}_N) - D_i(\nu)\} = (\hat{\nu}_N^{-1} - \nu^{-1}) \frac{\sum_i \mathsf{B}_i}{\sqrt{N}} = o_P(1).$$

The local risk result now follows from Corollary 4 and continuity of the limiting normal distribution at zero. Data reuse in estimating ν does not require a sample split for this argument.

For the second-order statistic, the analogous identity is a linear combination of v^{-2}, v^{-1}, and a constant, with count-polynomial coefficients. To see this, set $d = Y - X$ and note

$$c_2(Y; vT) - c_2(X; vT) = Y^2 - Y - X^2 + X - 2vTd,$$
$$c_1(X; vT)c_2(Y; vT) - c_1(Y; vT)c_2(X; vT) = d\{XY - vT(X + Y) + (vT)^2 + vT\}.$$

Each coefficient is antisymmetric under label exchange and has uniformly bounded variance near $a = 0$. Its null expectation is zero. Its first derivative in a is also zero because the single-bin first density derivative is symmetric. Expectations of these degree-at-most-three count polynomials are polynomials in a, by (78), so each coefficient has expectation $O(a^2)$. At $a_N = \gamma N^{-1/4}$ their sums divided by $\sqrt{N}$ are therefore $O_P(1)$. Multiplication by $\hat{\nu}_N^{-k} - \nu^{-k} = O_P(N^{-1/2})$, $k = 1, 2$, proves the second-order plug-in difference is $o_P(1)$ as well. The optimal risk in Theorem 5 is retained.

The result is pointwise in a fixed positive common rate, with the kernel and design fixed. It does not establish finite-sample coverage with an estimated rate, adaptivity to two different unknown marginal rates, or efficient recovery of a continuous unknown timescale. Those are different nuisance problems.

M A FINITE-SAMPLE CHOICE BETWEEN THE TWO INFORMATION CHANNELS

The total counts of a length-T window are available at every resolution. Let V_T be (61) formed from these total counts, replacing Δ by T, and put

$$d_T = a^2 \nu \{T - \tau(1 - e^{-T/\tau})\}, \qquad v_T = \mathrm{Var}(V_T),$$

where v_T is the exact expression (72) with $\Delta = T$. Define

$$e_D = \frac{a^2 J}{4C_a}, \qquad e_V = \frac{d_T^2}{v_T}. \tag{83}$$

These are respectively a lower bound on the squared mean-to-standard-deviation ratio of D, and that exact ratio for V_T. The specified binary experiment has the same a, h, ν, T under both orientations, so choosing the larger ratio uses no unknown directional label.

Proposition 14 (A constructive two-channel envelope)**.** *For the specified exponential binary experiment and $a > 0$, select D if $e_D \geq e_V$ and V_T otherwise. For odd $b \geq 8\log(1/\delta)$ and $k \geq \lceil 4/\max(e_D, e_V) \rceil$, majority sign of b independent group means, each using k windows, has each directional error at most δ whenever $N \geq bk$.*

Proof. Both selected statistics have positive mean in the forward model and its negative in the reverse model. The variance-to-squared-mean ratio is at most $1/\max(e_D, e_V)$. Chebyshev bounds the error of each group by $1/4$. Independence of the groups and the same Hoeffding argument as Appendix E give $e^{-b/8} \le \delta$. □

This bound remains valid even when the Hellinger remainder makes the corresponding lower bound uninformative. It supplies a sufficient sample size, not a matching nonlocal minimax theorem. The selection is a design rule for specified nuisance parameters. For unknown candidates, the learned or moment decoder in Section 6 can combine both channels, but its finite-grid reporting guarantee comes from independent calibration, not from substituting fitted parameters into Proposition 14.

The design calculation holds T fixed as Δ varies. The V_T channel is therefore unchanged by coarsening the same window, whereas the first-order information $J_{\Delta,T}$ decreases. For $\tau \ll \Delta \ll T$ and weak a, their explicit efficacies are approximately $a^2 T\tau^2/(2\Delta^3)$ and $a^4/4$, respectively. Equating these expressions suggests $\Delta^3 \approx 2T\tau^2/a^2$ when both approximations apply. This comparison of two attainable statistics is not a lower bound for all information in a many-bin experiment.

N Dependent windows and a quantitative calibration correction

Consider consecutive length-T windows separated by a gap $g \ge 0$ along one stationary trajectory. Let $\widetilde{Q}^{(N)}$ denote their joint count law, and let $Q^{\otimes N}$ be the product of their stationary marginal laws under the same parameters.

Theorem 15 (Gap coupling for the feed-forward family)**.** *For every admissible kernel,*

$$\mathrm{TV}(\widetilde{Q}^{(N)}, Q^{\otimes N}) \le \min\left\{1,\ 2(N-1)\nu a \int_g^{g+T} \overline{H}(u)\,du\right\}, \qquad \overline{H}(u) = \int_u^\infty h(v)\,dv. \quad (84)$$

For $h = h_\tau$, the right side equals

$$\epsilon_g = \min\{1,\ 2(N-1)\nu a\tau e^{-g/\tau}(1 - e^{-T/\tau})\}. \quad (85)$$

With an independent calibration bank, a rank-inverted confidence set at level α_0 consequently has coverage at least $1 - \alpha_0 - \epsilon_g$ under the dependent observation law. The same upper bound $\alpha_0 + \epsilon_g$ applies to any wrong nonempty unanimous report.

Proof. Let $s_i = (i-1)(T+g)$ and $e_i = s_i + T$. Keep the original source and offspring construction for the first window. For each $i \ge 2$, retain the source points in $[e_{i-1}, e_i)$ and their offspring observed in window i. Replace all source points before e_{i-1}, and their offspring in window i, by an independent stationary Poisson source with the same rate. Keep independent target baselines on the observed windows. In this new construction every window has its original stationary marginal distribution.

The newly constructed windows are independent. Their retained source blocks $[e_{i-1}, e_i)$ are disjoint, their newly supplied pasts are independent, and their baseline and offspring randomness are independent across these blocks. No retained source point contributes to two newly constructed windows, because the next window replaces sources preceding its own cutoff.

The source-bin counts agree in each original and newly constructed window. A discrepancy in target counts can occur only through offspring in window i whose source precedes e_{i-1} in one of the two constructions. The expected number of such offspring in each construction is

$$\nu a \int_{-\infty}^{e_{i-1}} \int_{s_i}^{s_i+T} h(t-s)\,dt\,ds = \nu a \int_g^{g+T} \overline{H}(u)\,du.$$

The probability of at least one discrepancy is bounded by twice this mean. Summing over $i = 2, \dots, N$ proves (84) by the coupling inequality. Integrating the exponential tail gives (85). Finally, attach the same independent calibration randomness to both coupled observations. Total variation cannot increase under this augmentation or under a reporting rule. Transfer the independent-window coverage and wrong-report bounds of Theorem 8 to obtain the claimed correction. □

When $\epsilon_g < \alpha$, using $\alpha_0 = \alpha - \epsilon_g$ preserves a target unconditional error bound α. A known parameter envelope may be used to upper bound ϵ_g before observing the data. When the bound is one, it gives no useful correction and provides no justification for treating adjacent windows as independent. It concerns this feed-forward cluster construction, not recurrent Hawkes processes or arbitrary dependent event streams. Elapsed collection time includes the discarded gaps. For example, $N = 32$, $T = 8$, $\nu = 1$, $a = 0.6$, $\tau = 8$, and $g = 96$ give the small bound reported in Table 31; adjacent windows in the same design have only the trivial bound.

O Conditional count forecasting across resolutions

Let the source be observed through $X_0, \ldots, X_{m-1}$ on $[0, T]$, and consider a future interval $[T, T+H]$. Under the forward candidate, the mean source count is νH. The target count conditional on the source-bin counts has mean

$$m_{2,\theta}(X) = \nu H + a \sum_{j=0}^{m-1} w_j(\Delta, H, \tau)(X_j - \nu\Delta), \tag{86}$$

where

$$w_j(\Delta, H, \tau) = \frac{\tau(1 - e^{-\Delta/\tau})(1 - e^{-H/\tau})}{\Delta} e^{-(m-j-1)\Delta/\tau}. \tag{87}$$

For the reverse candidate, exchange the roles of X and Y and the target coordinate.

To prove the formula, the expected number of offspring in $[T, T + H]$ from a source event at $s < T$ is

$$a \int_T^{T+H} h_\tau(t - s)\, dt = a(1 - e^{-H/\tau})e^{-(T-s)/\tau}.$$

Given the source count in bin j, its source times are uniform there because the source itself is Poisson. Averaging the last display over $s \in B_j$ gives aw_j. Source events before zero and after T are independent of the observed source-bin counts. Their expected contribution, together with the target baseline, is the stationary mean νH minus the unconditional contribution of the observed source bins. Subtracting $\nu\Delta$ from each X_j therefore gives (86).

This conditions only on the designated source-bin counts. The target history generally contains additional information about latent source timing and offspring activity, so (86) is not the full-history conditional mean. It is consequently a reference functional, not a universal lower bound on prediction MSE.

For any new history, the minimum and maximum of the candidate mean (86) over $\mathcal{C}_\alpha$ cover the true candidate-defined mean whenever the true candidate is in the set. This holds simultaneously over histories and horizons as a deterministic consequence of truth-in-set coverage. It is not a prediction interval for a realized future count. The empirical predictors are described in Appendix Q.

P An exact stationary sampler without burn-in

The finite-window sampler exploits the absence of second-generation offspring. First draw source points from a rate-ν Poisson process on $[0, T]$. Each source point independently generates a Poisson number of daughters with mean a and independent exponential delays of mean τ. Keep daughters in $[0, T]$. Add an independent target baseline process of rate $\nu(1 - a)$. It remains to include daughters in the observation window whose source lies before zero.

For a source at $s < 0$, the number of daughters in $[0, T]$ is Poisson with mean

$$v(s) = ae^{s/\tau}(1 - e^{-T/\tau}). \tag{88}$$

Such a source contributes at least one observed daughter with probability $1 - e^{-v(s)}$. Therefore active pre-window source points form a thinned Poisson process of intensity $\nu(1 - e^{-v(s)})\, ds$. Since $1 - e^{-v} \le v$, use the dominating intensity $\nu v(s)\, ds$, whose total mass is

$$\nu a\tau(1 - e^{-T/\tau}).$$

The algorithm is as follows. Draw that Poisson number of candidate source points. For each candidate draw $U \sim \mathrm{Unif}(0,1)$, set $s = \tau \log U$ and $v = aU(1 - e^{-T/\tau})$, and accept it with probability $(1 - e^{-v})/v$. Given acceptance, draw a $\mathrm{Pois}(v)$ count conditioned to be positive. Place these daughters independently with density

$$t \longmapsto \frac{e^{-t/\tau}}{\tau(1 - e^{-T/\tau})} \mathbf{1}\{0 < t < T\}.$$

This density does not depend on the negative-time parent location, because the exponential factor $e^{s/\tau}$ cancels on conditioning a daughter to fall in the window.

Poisson thinning proves that the accepted parents have exactly the desired active-parent intensity. The zero-truncated Poisson count and conditional daughter times are exactly the marks of an active parent in the original process. Superposing these contributions with those from in-window sources and the baseline therefore gives the stationary target restriction to $[0, T]$. The construction simulates the relevant negative-time contribution exactly with finitely many draws, without a finite-history approximation. Binning these generated event times applies the observation map C_Δ exactly, up to ordinary floating-point arithmetic.

The expected number of candidate pre-window parents is at most $\nu a \tau$. An independent repetition generates each independent window. For paired-resolution experiments, the same generated path is binned first at the finest required width and then summed into larger bins. No invented within-bin order enters any non-jittered statistic.

Q Experimental protocol and supplementary results

This section documents the fixed-bank experiment. Appendix R describes the independently repeated calibration and specification challenges. The training-seed variation reported here is distinct from variation across independent calibration banks.

Q.1 Feature map, training, and independence

The nine per-window features of the full decoder are

$$\begin{aligned}
f_1 &= \frac{\sum_j x_j}{\sqrt{\nu T}}, & f_2 &= \frac{\sum_j y_j}{\sqrt{\nu T}}, \\
f_3 &= \frac{\sum_j (x_j^2 - X_j)}{\lambda\sqrt{2m}}, & f_4 &= \frac{\sum_j (y_j^2 - Y_j)}{\lambda\sqrt{2m}}, \\
f_5 &= \frac{\sum_j x_j y_j}{\lambda\sqrt{m}}, & f_6 &= \frac{\sum_{j=0}^{m-2} (x_j y_{j+1} + y_j x_{j+1})}{\lambda\sqrt{2(m-1)}}, \\
f_{6+\ell} &= \frac{D_{h_{\tau_\ell},\Delta}}{\sqrt{J_{\Delta,T}(h_{\tau_\ell})}}, \quad \ell = 1,2,3, \quad (\tau_1,\tau_2,\tau_3) = (0.5,1,2). & & \qquad (89)
\end{aligned}$$

For $m = 1$, set $f_6 = 0$ and all three directional-score features to zero rather than evaluating a zero denominator. The input for a dataset of N independent windows is $N^{-1/2} \sum_i f^{(i)}$. Appendix S characterizes its retained tangents and the missing mixed coordinate. The "no directional scores" ablation uses $f_1, \ldots, f_6$. It is not an unordered-count-only method because f_6 retains symmetric adjacent-bin structure. The full decoder is a $9 \to 32 \to 32 \to 7$ multilayer perceptron with tanh nonlinearities and softmax output. The ablated decoder changes only the input dimension to six. The linear comparator is a $9 \to 7$ affine map followed by softmax.

The seven candidate labels are ordered as null, forward with $\tau = 0.5, 1, 2$, and reverse with $\tau = 0.5, 1, 2$. Non-null candidates have $a = 0.2$, all candidates have $\nu = 1$, and the null has no meaningful timescale. There are 800 training, 499 calibration, and 500 test datasets per candidate. Each maximal dataset has 128 independent windows of length 64. Smaller datasets use their first 16 or 64 windows. Bins are coarsened from width 0.25 to 1, 4, and 16. This pairing reduces comparison noise without making the windows within any one dataset dependent.

For each resolution, sample size, architecture, and training seed, a random 80% of the training datasets are used for optimization and the remaining 20% for validation. Feature means and standard deviations are fitted on the optimization subset only. Standard deviations are lower-bounded at 10^{-5}. Optimization uses AdamW with learning rate 0.003, weight decay 0.001, batches of 256, and at most 200 epochs. Training stops after 25 epochs without a validation cross-entropy improvement of at least 10^{-5}. The best validation checkpoint is restored. The three training seeds are 7000, 7001, and 7002.

Training, calibration, and test simulation seeds start at 11000001, 22000001, and 33000001. For candidate index k and chunk starting at dataset index s, the seed is the split seed plus $100000k + s$. Chunks have at most 12 datasets and each dataset has 128 simulated windows. This scheme makes the raw count data used by the non-neural comparator identical to the count data underlying the learned features. No calibration or test dataset is used for optimization, feature scaling, early stopping, or architecture selection. The training/calibration/test study uses 1,611,904 distinct maximal-window simulations before the independent moment and forecasting checks.

All rank sets use $\alpha = 0.05$ and $B = 499$. For numerical stability the implementation ranks the candidate probability directly, counting calibration probabilities no larger than the observed probability. This is equivalent to the upper-tail rank of $-\log \pi_\phi$ and avoids taking the logarithm of a numerically zero probability. The smallest nonzero rank is $1/500$.

The tabulated seed averages use the same held-out datasets and calibration banks across the three fitted networks. They are not three independent repetitions of the entire simulation experiment. Conditional on a fixed fitted network and calibration bank, test outcomes are independent within each true-candidate block. However, the exact coverage theorem averages over calibration randomness, so a realized bank can have test coverage below or above 95%. Empirical deviations near that level should not be mistaken for violations of the theorem or for independent evidence of conditional-on-bank validity.

Q.2 Coarse-grained quadratic comparator

We specialize the quadratic objective of Koyama (2025) to the finite candidate grid, using the known stationary rate rather than estimating it. Let $b_d = A_d/\Delta$. For a forward candidate define the filtered source

$$q_k = a\left(b_0 x_k + \sum_{j<k} b_{k-j} x_j\right).$$

Up to a parameter-independent constant, its score is

$$\ell_{\mathrm{CG},+}(X, Y) = \frac{1}{2\lambda} \sum_k \{2y_k q_k - q_k^2\}. \tag{90}$$

The reverse score exchanges x, y. The triangular within-bin coefficient matrix has determinant one after subtraction from the identity, so the log-determinant term in the general objective is zero here. The null score is zero. Unobserved pre-window centered counts are set to their zero mean in this comparator's finite-window filtering rule. The true data and calibration simulations remain stationary.

We sum (90) over all N windows, apply softmax over the seven candidates, and use exactly the same rank-set construction. This produces a calibrated finite-grid comparator despite approximation in its score. It is not an implementation of latent-time MC-EM, exact aggregated likelihood, or a full unconstrained fit of the general coarse-grained model.

Q.3 All decoding settings

Tables 8 and 9 give the full decoder and coarse-grained comparator at every (Δ, N) setting. All percentages are empirical frequencies. "Exact" denotes exact seven-candidate classification, "Dir." denotes classification into null/forward/reverse, and "Wrong" is an unconditional wrong-report frequency. A nonempty confidence set with more than one timescale can still provide a directional report.

Table 8: Full neural decoder across all settings, averaged over three training seeds.

Δ	N	Exact	Dir.	Cover	Report	Wrong	$\lvert\mathcal{C}\rvert$
0.25	16	70.96	92.72	94.44	81.67	1.57	1.89
	64	93.67	99.44	95.31	99.30	0.41	1.06
	128	97.40	99.90	94.56	96.36	0.05	0.96
1	16	63.50	91.36	95.34	77.90	1.91	2.39
	64	87.44	99.52	94.72	99.31	0.37	1.25
	128	95.08	99.95	94.37	98.77	0.02	1.00
4	16	39.30	69.29	94.87	8.98	0.55	4.76
	64	60.34	87.09	94.82	59.88	2.28	2.74
	128	73.38	94.00	95.53	81.58	1.08	1.97
16	16	24.21	50.72	95.39	4.91	1.42	6.02
	64	32.83	64.88	95.18	12.87	1.04	5.06
	128	37.91	71.26	94.81	25.22	2.13	4.55

Table 9: Coarse-grained finite-grid comparator across all settings.

Δ	N	Exact	Dir.	Cover	Report	Wrong	$\lvert\mathcal{C}\rvert$
0.25	16	77.06	95.71	95.31	91.29	1.03	1.68
	64	95.91	99.89	95.46	98.89	0.03	1.00
	128	99.11	100.00	94.97	95.17	0.00	0.95
1	16	66.89	92.71	95.43	81.49	1.63	2.24
	64	90.51	99.80	95.37	99.11	0.14	1.18
	128	96.71	100.00	95.03	97.97	0.00	0.98
4	16	40.31	69.66	94.40	13.91	0.97	4.65
	64	60.57	87.29	94.54	60.69	2.37	2.73
	128	73.86	93.91	95.34	80.91	0.94	1.94
16	16	25.71	51.11	95.34	6.00	1.83	5.98
	64	33.43	65.71	94.91	15.94	1.54	5.02
	128	37.43	70.97	94.26	27.00	2.57	4.51

Q.4 Softmax credible sets are a different object

For comparison only, sort the seven learned probabilities in descending order and retain the smallest prefix whose cumulative probability is at least 0.95. These nominal 95% softmax credible sets are neither model-based Bayesian posteriors nor finite-sample confidence sets. Table 10 reports their observed behavior. In these runs they are generally more conservative than the rank sets. The results do not support a general claim that uncalibrated neural scores undercover.

Table 10: Full neural decoder at $N = 128$ under rank inversion and softmax-prefix reporting. Coverage and report rates are percentages. Values average three training seeds.

Δ	Set construction	Cover	Report	Wrong	$\lvert\mathcal{C}\rvert$
0.25	Rank inversion	94.56	96.36	0.05	0.96
	Softmax prefix	99.41	99.63	0.02	1.08
1	Rank inversion	94.37	98.77	0.02	1.00
	Softmax prefix	98.89	99.85	0.01	1.19
4	Rank inversion	95.53	81.58	1.08	1.97
	Softmax prefix	98.28	72.80	0.34	2.34
16	Rank inversion	94.81	25.22	2.13	4.55
	Softmax prefix	99.12	11.38	0.02	5.23

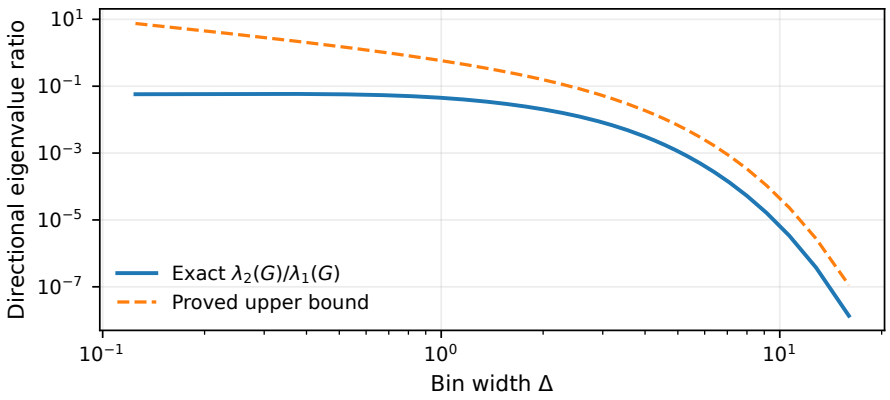


Figure 2: Directional-score rank collapse for $\tau \in \{0.5, 1, 2\}$ and $T = 64$. The exact ratio of the second to the first eigenvalue is compared with Proposition 10. Each point uses an integer number of bins. The bound concerns the unstandardized Gram matrix specified in (56).

Q.5 Moment checks, random jitter, and single-bin recovery

The exact score moment checks use 30,000 independent windows for each combination of $\Delta \in \{0.25, 1, 4, 16\}$ and $a \in \{0, 0.2, 0.6\}$, with $\nu = \tau = 1$ and $T = 64$. At the null, the exact score variances are 49.3492, 23.2082, 1.74191, and 0.0234375. The corresponding simulated variances are 49.5265, 22.9718, 1.72860, and 0.0234136. These simulations provide a numerical check of the formulas, not a proof. The result files contain the full means, mean standard errors, and variance bounds.

The jitter experiment generates 12,000 independent null windows per resolution, then makes 16 independent within-bin placements per window. Prefix averages of one, four, and sixteen placements are compared in Table 11. The independent-randomization unit for the variance is the window, not the individual placement. Boundary terms from (60) are included.

Table 11: Null random-jitter variance, 12,000 independent windows per resolution. Each cell gives the predicted variance followed by the simulated variance in parentheses. The exact projected-score variance is J.

Δ	J	$L = 1$	$L = 4$	$L = 16$
0.25	49.349	63.500 (63.204)	52.887 (52.713)	50.234 (49.721)
1	23.208	63.500 (63.282)	33.281 (33.653)	25.726 (25.841)
4	1.7419	63.500 (63.564)	17.181 (17.599)	5.602 (5.597)
16	0.023437	63.500 (63.279)	15.893 (16.111)	3.991 (3.970)

The one-bin experiment uses $a \in \{0.2, 0.6\}$, $\Delta \in \{1, 4, 16, 64\}$, and $N \in \{16, 64, 128\}$, with $\nu = \tau = 1$. There are 2,000 independent datasets for each amplitude and width. Within each dataset the smaller-N conditions use prefixes of 128 independent windows. The statistic reports forward when $\sum_i V_i > 0$ and otherwise reverse. Ties are counted as errors for the forward data, a conservative convention at very small counts. Label symmetry gives the corresponding reverse experiment. Table 12 reports all three sample sizes.

Table 12: Single-bin forward recovery. Values are percent correct with a binomial standard error in percentage points, based on 2,000 independent datasets. First-order directional information is zero for every row.

a	Δ	$N = 16$	$N = 64$	$N = 128$
0.2	1	45.95 (1.11)	50.25 (1.12)	49.80 (1.12)
	4	49.30 (1.12)	54.05 (1.11)	57.15 (1.11)
	16	51.15 (1.12)	54.05 (1.11)	55.40 (1.11)
	64	53.30 (1.12)	56.35 (1.11)	59.90 (1.10)
0.6	1	53.25 (1.12)	67.15 (1.05)	74.40 (0.98)
	4	68.30 (1.04)	83.30 (0.83)	93.05 (0.57)
	16	75.90 (0.96)	90.00 (0.67)	96.65 (0.40)
	64	76.95 (0.94)	91.10 (0.64)	97.55 (0.35)

Q.6 Forecast study

For each true candidate we generate 500 fresh stationary paths on $[0, 96]$, using simulation seed $9500001 + 1000k$. Histories lie in $[0, 64]$ and future horizons have lengths $H \in \{\Delta/2, \Delta, 2\Delta\}$. Counts are first recorded at step 0.125 and then coarsened. These histories and outcomes are independent of the dataset used to decode the candidate. Each test dataset's candidate probabilities or confidence set is paired with one independent forecasting history from the same true candidate.

Let $\boldsymbol{m}_\theta$ be the two-coordinate source-count mean from (86). The point predictor is $\sum_\theta \pi_\phi(\theta \mid Z)\boldsymbol{m}_\theta$. The confidence-average predictor is the uniform mean of $\boldsymbol{m}_\theta$ over a nonempty $\mathcal{C}_\alpha$. For an empty confidence set, it uses a uniform mean over the entire seven-candidate grid and issues no directional report. The constant-rate predictor is $(\nu H, \nu H)$. The source-count oracle uses the true candidate's $\boldsymbol{m}_\theta$.

For an observed future count vector $\boldsymbol{Y}^{\mathrm{fut}}$, normalized count MSE is

$$\frac{(\widehat{m}_1 - Y_1^{\mathrm{fut}})^2 + (\widehat{m}_2 - Y_2^{\mathrm{fut}})^2}{2\nu H}.$$

The mean-functional error replaces $\boldsymbol{Y}^{\mathrm{fut}}$ with the true candidate's source-count mean and omits the νH normalization. For neural methods, losses are first averaged across the three training seeds within each forecasting example. Means and standard errors are then computed over the 3,500 forecasting examples, conditional on the fitted decoders and shared calibration banks. Rows for different methods, horizons, and training seeds are paired, not independent. Table 13 reports $N = 128, H = \Delta$. All horizons and sample sizes appear in the result files.

Because methods use the same forecast examples, comparisons use the standard error of their paired loss difference, not the separate marginal error bars. Table 14 reports these differences without multiplicity adjustment. Neural–CG differences change sign across resolutions and are small on the normalized-MSE scale. Confidence-set averaging does not consistently improve the point predictor. The source-count reference is not the full-history Bayes predictor, and slight empirical improvements over it do not establish better estimation of the full conditional law. The study shows that similar count predictions can coexist with different levels of temporal identification and reporting specificity. It does not establish a forecasting gain from confidence-set averaging.

Table 13: Forecasting at $N = 128$ and $H = \Delta$. The count-MSE column gives a normalized mean and its test standard error. The functional column gives squared error to the true source-count mean, multiplied by 10^3. Neural-based rows average losses across the three fitted networks within each test example before computing its standard error.

Δ	Predictor	Count MSE (SE)	Functional $\times 10^3$
0.25	Constant rate	1.01571 (0.02983)	0.44359
	True source-count mean	1.01505 (0.02982)	0.00000
	Neural probability	1.01512 (0.02981)	0.00170
	Neural confidence set	1.01493 (0.02981)	0.01031
	CG probability	1.01508 (0.02982)	0.00039
	CG confidence set	1.01481 (0.02981)	0.01094
1	Constant rate	1.01371 (0.02073)	2.86090
	True source-count mean	1.01278 (0.02060)	0.00000
	Neural probability	1.01263 (0.02060)	0.01349
	Neural confidence set	1.01260 (0.02060)	0.03016
	CG probability	1.01274 (0.02060)	0.00738
	CG confidence set	1.01266 (0.02060)	0.04033
4	Constant rate	1.01714 (0.01806)	5.00740
	True source-count mean	1.01581 (0.01806)	0.00000
	Neural probability	1.01576 (0.01805)	0.48118
	Neural confidence set	1.01587 (0.01805)	0.55819
	CG probability	1.01584 (0.01805)	0.45690
	CG confidence set	1.01579 (0.01805)	0.55509
16	Constant rate	1.01075 (0.01767)	1.86709
	True source-count mean	1.01040 (0.01765)	0.00000
	Neural probability	1.01059 (0.01765)	1.02132
	Neural confidence set	1.01056 (0.01766)	1.07281
	CG probability	1.01059 (0.01766)	1.02461
	CG confidence set	1.01052 (0.01766)	1.09122

Table 14: Paired forecast-loss differences at $N = 128, H = \Delta$, multiplied by 10^4. Each entry is a mean difference followed by its paired standard error. Negative values favor the first method. "Prob." averages candidate means with decoder probabilities, and "set" averages over the confidence set.

Δ	NN prob. − CG prob.	NN set − NN prob.	NN prob. − constant
0.25	0.360 (0.438)	-1.908 (1.155)	-5.964 (10.514)
1	-1.095 (0.483)	-0.319 (0.821)	-10.827 (13.453)
4	-0.804 (0.639)	1.065 (1.090)	-13.811 (8.121)
16	0.003 (0.155)	-0.278 (0.436)	-1.642 (1.760)

Q.7 Local-alternative score-risk experiment

We separately test the known-kernel score under the local alternatives of Corollary 4. Fix $\nu = \tau = 1, T = 64$ and choose $\Delta \in \{0.25, 1, 4\}$, $N \in \{64, 256, 1024\}$, and $z \in \{0.25, 0.5, 1, 1.5\}$. For each (Δ, z), set $\gamma = 2z/\sqrt{J_{\Delta,T}}$ and $a_N = \gamma/\sqrt{N}$. The limiting equal-prior score risk is therefore $\Phi(-z)$ at every resolution. Each setting uses 1,000 independent datasets, and different settings use independent seeds. Under the forward model we average $\mathbf{1}\{\sum_i D_i < 0\} + \frac{1}{2}\mathbf{1}\{\sum_i D_i = 0\}$. Label exchange makes this equal to the equal-prior risk of random tie-breaking. The reported standard error is the sample standard deviation of this bounded loss divided by $\sqrt{1000}$.

Figure 3 and Table 15 compare the finite-N errors with the asymptotic benchmark. The comparison checks the score mechanism without assuming that every finite setting is close to the limiting regime. In particular, the coarse-resolution case $\Delta = 4, N = 1024, z = 1.5$ has empirical risk 0.046 with standard error 0.0066, versus the asymptotic value 0.0668. Finite-sample effects and Monte Carlo variation both contribute to deviations from the limiting risk. The raw output also records $a_N K/\sqrt{J}$ with $\bar{a} = 0.4$; many coarse settings do not satisfy the sufficient curvature condition in Section 5.

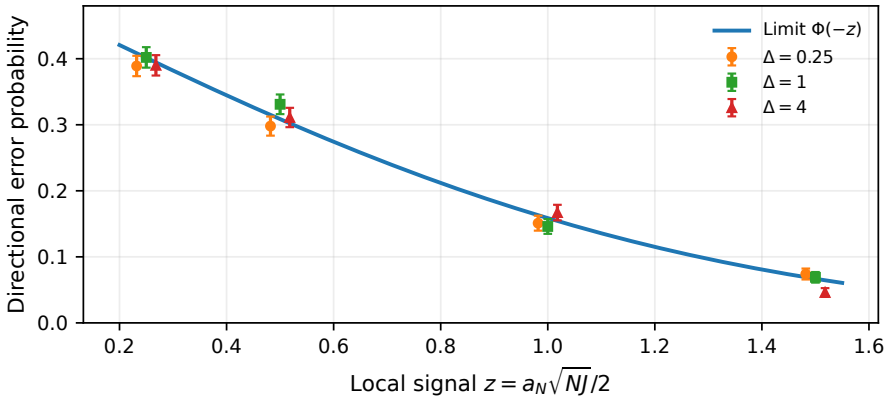


Figure 3: Known-kernel directional score risk under local alternatives at $N = 1024$. Markers show 1,000 independent datasets per setting, with one Monte Carlo standard error. The curve is the exact asymptotic risk $\Phi(-z)$, where $z = a_N\sqrt{NJ}/2$. Markers are offset slightly in the horizontal direction for legibility. Finite-sample deviations are not omitted.

Table 15: Local-alternative score error in percent. Each Monte Carlo estimate uses 1,000 independent datasets. Values in parentheses are one standard error, also in percentage points. The column R_∞ is $100\Phi(-z)$.

Δ	z	R_∞	$N = 64$	$N = 256$	$N = 1024$
0.25	0.25	40.13	40.80 (1.55)	39.30 (1.55)	38.90 (1.54)
	0.5	30.85	31.80 (1.47)	30.80 (1.46)	29.80 (1.45)
	1	15.87	16.60 (1.18)	16.90 (1.19)	15.10 (1.13)
	1.5	6.68	7.00 (0.81)	7.50 (0.83)	7.40 (0.83)
1	0.25	40.13	41.50 (1.56)	43.20 (1.57)	40.20 (1.55)
	0.5	30.85	30.30 (1.45)	33.00 (1.49)	33.10 (1.49)
	1	15.87	16.70 (1.18)	16.30 (1.17)	14.60 (1.12)
	1.5	6.68	6.90 (0.80)	5.80 (0.74)	6.90 (0.80)
4	0.25	40.13	40.90 (1.56)	38.90 (1.54)	39.00 (1.54)
	0.5	30.85	31.80 (1.47)	30.10 (1.45)	31.10 (1.46)
	1	15.87	15.90 (1.16)	15.70 (1.15)	16.70 (1.18)
	1.5	6.68	7.50 (0.83)	6.10 (0.76)	4.60 (0.66)

Q.8 Computational record and verification scope

The implementation includes the stationary sampler, exact information and moment formulas, linear-time score evaluation, neural training, independent rank calibration, the coarse-grained comparator, random-jitter experiments, forecast evaluation, and figure/table generation. Raw aggregate results and fitted decoder weights accompany the manuscript. Numerical checks also evaluate the Gram eigenvalue inequality at 100 random valid parameter configurations. The main neural experiment was run on CPU with a single computational thread per training process and took 116.1 seconds in the recorded environment. This is an execution record, not a hardware-normalized benchmark. Package versions and measured runtime are recorded with the results.

No application dataset, latent-time Bayesian fit, MC-EM implementation, or interval-censored Transformer benchmark was run for this study. Such comparisons would address broader empirical questions than the present restricted-family information theorem. The simulator-defined grid is not a validation of model fit to an external event stream. The finite-grid coverage claim applies only under its stated exchangeability and specification assumptions, and the local efficiency result applies to the specified known-kernel score, not to an arbitrary trained neural network.

R EXTENDED EXPERIMENTS AND SPECIFICATION CHALLENGES

The experiments in this section test separate implications of the theory. They are not pooled into a single benchmark score. Binary score tests use a specified kernel and amplitude and either a known or estimated common rate. The seven-candidate decoding experiment instead has an unknown discrete direction and timescale, but its training grid fixes the nonnull amplitude and rate. Frozen-grid challenges intentionally violate that specification. The numerical likelihood comparison is confined to the single-bin binary experiment. The comparisons answer different questions and do not establish a uniform advantage for any method.

R.1 RELATION TO THE CLOSEST METHODS

Table 16 locates the contribution relative to existing methods. Second-order statistics, neural regression on simulated summaries, and inversion of calibrated tests are established ingredients. Here we derive the first two directional derivatives and their exact local risk constants for a specified mean-matched experiment. The fourth-root result uses a mixed third-degree count statistic as well as dispersion. It complements integrated-cumulant approaches to identifying interaction structure.

Table 16: Closest methodological connections and the distinction made here. These are scope comparisons, not performance rankings.

Work	Established contribution	Distinction in this paper
Achab et al. (2017)	Second- and third-order integrated cumulants recover integrated interactions	Exact finite-window directional tangents and their local testing constants
Zhou and Papadogeorgou (2026)	Bayesian aggregated-process inference and specified-model identifiability	Quantifies local information loss without asserting global nonidentifiability
Rizoiu et al. (2022)	Mean-behavior approximations for interval-censored inference	Studies directional testing in the observed count law
Koyama (2025)	Coarse-grained dynamics including intra-bin covariance	Uses established aggregation weights to derive antisymmetric information
Lambe et al. (2026a)	Simulated-summary neural parameter estimation and quantile uncertainty	A restricted directional theory, not a first neural count-inference method
Dalmasso et al. (2020)	Likelihood-free confidence sets and hypothesis testing	Candidatewise rank calibration and a separately evaluated support guard
Rotnitzky et al. (2000)	Likelihood asymptotics with singular information	Observed Hawkes tangents and exact directional testing constants, not a new generic nonstandard-rate principle

R.2 FINITE-SAMPLE APPLICABILITY AND A WORKED DATA REGIME

The script `run_revision_theory.py` evaluates the explicit Hellinger envelope on 294 designs, with $T \in \{8, 64\}$, admissible widths from $\{0.25, 1, 4, 16\}$, seven amplitudes from 10^{-4} through 0.2, and six sample sizes from 64 through 10^6. The bound always uses the declared upper amplitude $\bar{a} = 0.2$. The file `finite_bound_region.csv` records $J, K, aK/\sqrt{J}$, the sufficient threshold, and both risk bounds. The recorded lower bound is retained even when it is numerically close to zero. Increasing N cannot make a per-window lower Hellinger envelope informative once its remainder exceeds the leading term.

A separate simulation uses $\nu = \tau = 1, T = 64, \Delta = 0.25, a = 0.01, N = 1000$. The bounds are $0.0445179 \le R_N^* \le 0.3362070$. On 2000 independent forward datasets, the score error is 0.1245 with Monte Carlo standard error 0.00738. Estimating the common rate produces the same decisions in this sample. The example gives an informative finite-sample bound and a numerical check of the score. The simulation does not prove the bounds on optimal risk. Seeds start at 941000000 and advance by the chunk's dataset index, with four datasets per chunk. The outputs are in `nonvacuous_experiment.csv`.

Figure 4 compares the two constructive group-size bounds at fixed total window length. Every point uses the exact dispersion variance, not its large-bin approximation. The vertical axis counts windows per group. For error target $\delta = 0.05$, choose $b = 25$ groups. At $\Delta = 0.25$, rounding gives $k = 10$ and $N = bk = 250$. At $\Delta = 16$, it gives $k = 10339$ and $N = 258475$. Their observed exposures are 16000 and 16542400, respectively. No gap cost is included because these are independent repetitions. These sufficient counts are not estimated optimal recovery boundaries.

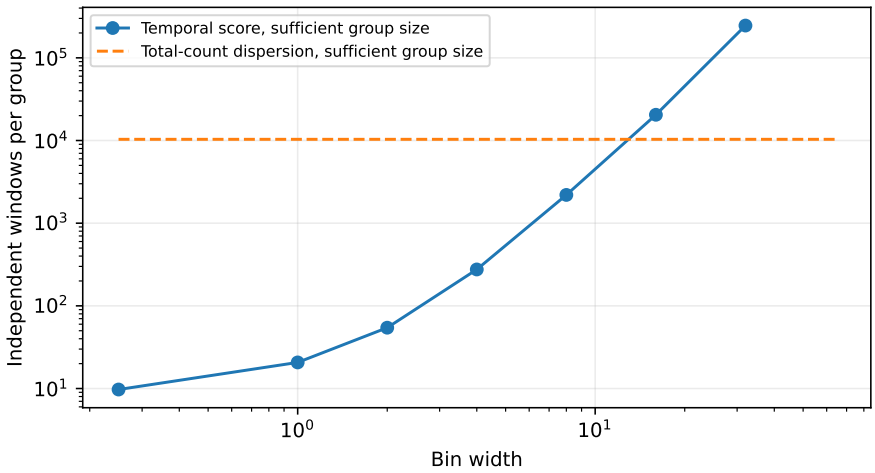


Figure 4: A fixed-exposure design comparison with $T = 64$, $\nu = \tau = 1$, and $a = 0.2$. Curves show $4/e_D$ and $4/e_V$, before rounding up. The temporal curve omits its infinite value at the single-bin endpoint. Total-count dispersion is available at every resolution of the same window. These are sufficient group-size bounds, not matching nonlocal sample-complexity rates.

R.3 First-order local risk with independent repetitions

The extension fixes $\Delta = 4, T = 64, \nu = \tau = 1$ and sets $a_N = 3/\sqrt{NJ}$ for $N = 1024, 4096, 16384$. Three independent runs contain 400 datasets each. Seed $850000000 + 1000000s + 10N + b$ uses run index $s \in \{0, 1, 2\}$ and chunk index b, with four datasets per chunk. Within a dataset, the known- and estimated-rate tests use the same windows. Their decisions coincide in this particular extension. The known-rate pooled errors are 5.583%, 6.417%, and 6.667%, compared with the limit 6.681%. Table 3 reports pooled Monte Carlo standard errors, not optimizer-seed variability. Appendix Q.7 also retains the original 4.6% result at $N = 1024$, separately from these independent runs.

R.4 Numerical count likelihood for a single bin

The feed-forward cluster law permits an exact analytic probability generating function, equation (78). For an exponential kernel, its two source contributions can be written

$$I_{\mathrm{in}}(z) = T \int_0^1 \mathrm{expm1}\{a(1 - e^{-Tu/\tau})(z-1)\}\, du,$$

$$I_{\mathrm{past}}(z) = \tau \int_0^1 \frac{\mathrm{expm1}\{a(1 - e^{-T/\tau})u(z-1)\}}{u}\, du,$$

$$\log G_a^+(z_x, z_y) = \nu T(z_x - 1) + \nu(1-a)T(z_y - 1) + \nu\{z_x I_{\mathrm{in}}(z_y) + I_{\mathrm{past}}(z_y)\}. \tag{91}$$

The removable singularity in the second integral is handled by its continuous integrand. We use 256-point Gauss–Legendre quadrature and evaluate the generating function on a 256×256 unit-circle grid. Fourier coefficient extraction gives a count probability array. Infinitesimal negative coefficients due to floating-point error are set to zero, and the array is normalized. The reverse probability is its transpose. Summing $\log p_a^+(X_i, Y_i) - \log p_a^+(Y_i, X_i)$ gives the numerical likelihood-ratio statistic. This is a restricted count-likelihood reference, not an implementation of MC-EM or particle MCMC.

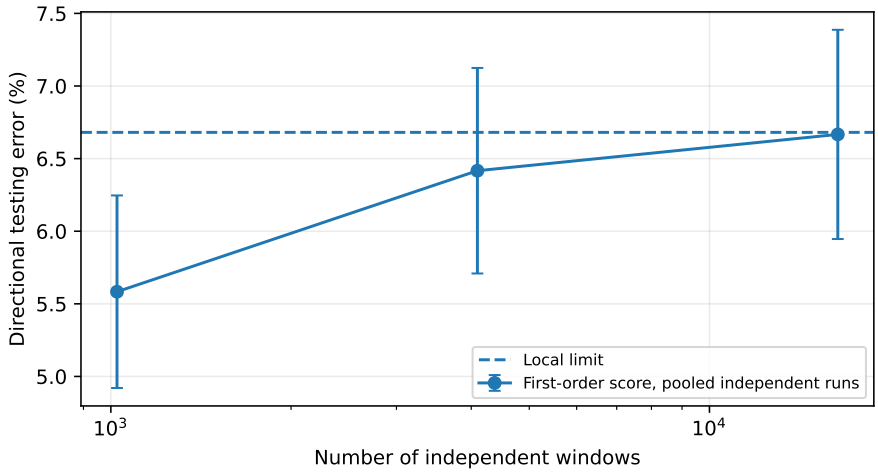


Figure 5: Independent extension of the coarse first-order local-risk check. Each point pools 1200 datasets, with one Monte Carlo standard error. The horizontal line is the limiting error, not a finite-sample identity.

Fourier extraction aliases counts above the array boundary. For $t > 0$, convexity of $e^{cp} - 1$ on $p \in [0, 1]$, together with $\int p_T = T$, gives

$$\log \mathbb{E}e^{tY} \le \nu T \left[(1-a)(e^t - 1) + e^{a(e^t-1)} - 1\right], \qquad \log \mathbb{E}e^{tX} = \nu T(e^t - 1). \tag{92}$$

Multiplying either moment bound by e^{-256t} bounds the corresponding count tail. We numerically choose $t \in [10^{-5}, 2]$ to tighten these valid bounds. Their sum bounds the total alias probability. Numerical optimization need not find the exact minimizer to retain validity of any evaluated Chernoff bound.

For $T = 64$, $a \in \{0.05, 0.2, 0.6\}$, and $\tau \in \{0.5, 1, 2\}$, the recorded overflow bounds range from 5.64×10^{-71} to 4.81×10^{-46}. Repeating the computation with 512 quadrature nodes and a 512×512 Fourier grid changes each probability in the original block by at most 2.01×10^{-16}. Removed negative mass is below 9.4×10^{-15}. The analytic tail bound and numerical quadrature check assess different errors. Neither gives an interval-arithmetic certificate for the entire likelihood array. All observed counts remained inside the smaller array.

Table 17 compares this likelihood statistic with dispersion, using 1000 independent datasets for each (a, τ, N) and $N \in \{16, 64, 128\}$. The seed is $900000000 + 1000000 i_a + 100000 i_\tau + 100N + b$, where b starts each chunk of eight datasets. Both methods are applied to the same counts. Ties receive half an error. The paired difference column reports dispersion minus likelihood error with its within-dataset standard error, not the sum of marginal standard errors.

Table 17: Single-bin count-likelihood reference at $T = 64$, $\nu = 1$. Errors and standard errors are percentage points. Parentheses in the last column are standard errors of paired differences. No nuisance parameter is fitted in this binary comparison.

a	τ	N	Likelihood	Dispersion	Disp. minus lik.
0.05	0.5	16	48.20 (1.58)	49.00 (1.58)	0.80 (0.55)
0.05	0.5	64	49.20 (1.58)	49.65 (1.58)	0.45 (0.47)
0.05	0.5	128	50.60 (1.58)	50.65 (1.58)	0.05 (0.52)
0.05	1	16	50.00 (1.58)	50.30 (1.58)	0.30 (0.57)
0.05	1	64	49.60 (1.58)	48.90 (1.58)	-0.70 (0.59)
0.05	1	128	50.50 (1.58)	50.95 (1.58)	0.45 (0.58)
0.05	2	16	51.50 (1.58)	51.55 (1.58)	0.05 (0.50)
0.05	2	64	48.50 (1.58)	48.80 (1.58)	0.30 (0.62)
0.05	2	128	48.90 (1.58)	48.55 (1.58)	-0.35 (0.61)
0.2	0.5	16	49.90 (1.58)	50.00 (1.58)	0.10 (0.48)
0.2	0.5	64	43.20 (1.57)	43.45 (1.57)	0.25 (0.60)
0.2	0.5	128	38.10 (1.54)	38.10 (1.54)	0.00 (0.55)
0.2	1	16	47.50 (1.58)	47.30 (1.58)	-0.20 (0.57)
0.2	1	64	44.60 (1.57)	44.80 (1.57)	0.20 (0.57)
0.2	1	128	41.90 (1.56)	42.95 (1.57)	1.05 (0.58)
0.2	2	16	46.50 (1.58)	46.10 (1.58)	-0.40 (0.47)
0.2	2	64	42.60 (1.56)	43.50 (1.57)	0.90 (0.50)
0.2	2	128	40.60 (1.55)	41.10 (1.56)	0.50 (0.54)
0.6	0.5	16	22.20 (1.31)	22.00 (1.31)	-0.20 (0.35)
0.6	0.5	64	6.80 (0.80)	7.10 (0.81)	0.30 (0.33)
0.6	0.5	128	2.70 (0.51)	2.60 (0.50)	-0.10 (0.22)
0.6	1	16	23.40 (1.34)	24.05 (1.35)	0.65 (0.36)
0.6	1	64	7.00 (0.81)	7.25 (0.82)	0.25 (0.32)
0.6	1	128	2.80 (0.52)	2.60 (0.50)	-0.20 (0.20)
0.6	2	16	24.10 (1.35)	24.05 (1.35)	-0.05 (0.30)
0.6	2	64	7.00 (0.81)	6.75 (0.79)	-0.25 (0.34)
0.6	2	128	2.70 (0.51)	2.70 (0.51)	0.00 (0.20)

R.5 THE SECOND-ORDER LOCAL EXPERIMENT AND NUMERICAL DERIVATIVES

For $T \in \{1, 8, 64\}$ and $\nu = \tau = 1$, let $\gamma = \{2/\sqrt{J^{(2)}}\}^{1/2}$, so the second-order limiting error is $\Phi(-1) = 0.158655$. For $N = 64, 256, 1024, 4096$, simulate $a_N = \gamma N^{-1/4}$ and 2000 independent datasets per cell. Compare $D^{(2)}$, dispersion, and the count-likelihood reference on the same data. Seeds are $910000000 + 10000000 i_T + 100N + b$, with chunks of eight datasets. The second-order statistic includes both terms in (12).

Dispersion has a different limiting error,

$$\Phi\left(-\sqrt{\frac{\beta_2^2/T^2}{J^{(2)}}}\right),$$

which is 0.181530, 0.171030, and 0.160502 at the three window lengths. Using 0.158655 as dispersion's limit in every row would omit the mixed third-degree contribution. The short-window amplitudes remain appreciable even at $N = 4096$, so finite-sample efficiency need not follow the ordering of limiting efficacies. For example, at $T = 1, N = 64$, dispersion has smaller empirical error than $D^{(2)}$. Conversely, at $T = 64, N = 4096$, the second-order and likelihood statistics both have empirical error 0.1590. The theorem is a pointwise limit, not a uniform numerical approximation across these rows.

An independent 200,000-window null sample checks the variance $J^{(2)}$ at each T. The seed is $920000000 + i_T$. A separate generating-function check compares $(p_a^+ - p_a^-)/a^2$ with $p_0 D^{(2)}$ at $a = 0.02, 0.01, 0.005$. Table 19 reports the square root of $\sum(\widehat{d} - p_0 D^{(2)})^2/p_0$ on cells with $p_0 > 10^{-10}$. The included null probability mass is stored alongside the error. Its decrease with a checks the density derivative numerically. The L^2 root-density expansion is proved separately.

Table 18: Single-bin fourth-root experiments. Each error uses 2000 independent datasets, with one standard error in parentheses. The $D^{(2)}$ and likelihood limits are 15.87%; the dispersion limits depend on T as stated in the text.

T	N	a_N	$D^{(2)}$	Dispersion	Likelihood
1	64	0.7862	23.00 (0.94)	21.05 (0.89)	20.75 (0.91)
1	256	0.5559	18.20 (0.86)	17.70 (0.84)	17.15 (0.84)
1	1024	0.3931	20.30 (0.90)	19.65 (0.88)	19.15 (0.88)
1	4096	0.2780	17.70 (0.85)	17.42 (0.85)	16.90 (0.84)
8	64	0.5210	17.05 (0.84)	16.95 (0.84)	16.15 (0.82)
8	256	0.3684	17.80 (0.86)	17.80 (0.85)	17.30 (0.85)
8	1024	0.2605	16.55 (0.83)	17.20 (0.84)	17.05 (0.84)
8	4096	0.1842	18.00 (0.86)	18.60 (0.87)	17.75 (0.85)
64	64	0.5020	16.55 (0.83)	16.55 (0.83)	16.05 (0.82)
64	256	0.3550	16.10 (0.82)	16.05 (0.82)	16.35 (0.83)
64	1024	0.2510	15.85 (0.82)	16.35 (0.83)	16.00 (0.82)
64	4096	0.1775	15.90 (0.82)	16.55 (0.83)	15.90 (0.82)

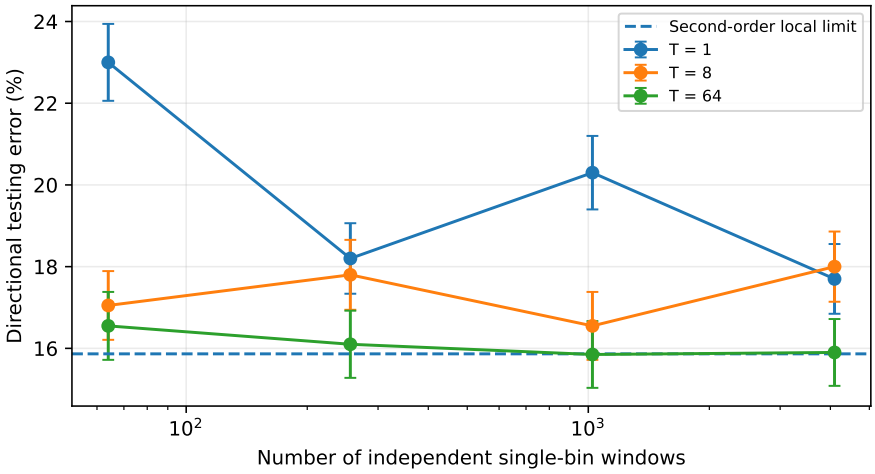


Figure 6: Second-order score risk along the fourth-root alternatives. Error bars are one Monte Carlo standard error. The short-window curves approach the limit slowly and retain visible finite-sample departures. Each line fixes a physical window length.

Table 19: Independent second-order moment and density-difference checks. The last three columns are weighted density-difference errors on the stated numerical support, not root Hellinger distances.

T	$J^{(2)}$	Null var.	Null mean	$a = .02$	$a = .01$	$a = .005$
1	0.163590	0.163177	-0.000182	0.006704	0.003351	0.001676
8	0.848235	0.845605	-0.002421	0.030449	0.015219	0.007609
64	0.983895	0.982576	-0.003051	0.034292	0.017140	0.008569

R.6 UNKNOWN COMMON RATES

The first-order nuisance study fixes $T = 64$, $N = 128$, and $\tau = 1$, and crosses five rates $\nu \in \{0.25, 0.5, 1, 2, 4\}$, three widths $\Delta \in \{1, 4, 16\}$, and three amplitudes $a \in \{0.05, 0.2, 0.6\}$. Every cell contains 1000 independent forward datasets. The estimated-rate statistic centers and normalizes with the same pooled estimate in Theorem 6; the simulation uses the true rate. The positive floor is N^{-1}, as in the theorem. Both statistics use the same data, so the relevant uncertainty on their difference is paired. Seeds are $860000000 + 1000000i_\nu + 100000i_\Delta + 10000i_a + b$, with chunks of eight datasets. Tables 20–22 report all 45 cells.

Table 20: Unknown-rate comparison at $a = 0.05$. Errors and paired estimated-minus-known differences are percentage points, with one standard error in parentheses. Rate RMSE is in events per unit time.

ν	Δ	Known	Estimated	Difference	RMSE($\hat{\nu}$)
0.25	1	10.00 (0.95)	10.00 (0.95)	0.00 (0.00)	0.0040
0.25	4	36.60 (1.52)	36.70 (1.52)	0.10 (0.17)	0.0041
0.25	16	47.80 (1.58)	48.20 (1.58)	0.40 (0.28)	0.0040
0.5	1	8.00 (0.86)	8.00 (0.86)	0.00 (0.00)	0.0056
0.5	4	36.70 (1.52)	36.80 (1.53)	0.10 (0.17)	0.0056
0.5	16	50.30 (1.58)	50.90 (1.58)	0.60 (0.24)	0.0056
1	1	10.60 (0.97)	10.60 (0.97)	0.00 (0.00)	0.0078
1	4	36.00 (1.52)	36.10 (1.52)	0.10 (0.17)	0.0079
1	16	46.80 (1.58)	47.70 (1.58)	0.90 (0.36)	0.0083
2	1	9.70 (0.94)	9.70 (0.94)	0.00 (0.00)	0.0110
2	4	35.00 (1.51)	35.20 (1.51)	0.20 (0.14)	0.0120
2	16	46.60 (1.58)	46.60 (1.58)	0.00 (0.28)	0.0112
4	1	8.50 (0.88)	8.50 (0.88)	0.00 (0.00)	0.0165
4	4	34.30 (1.50)	34.40 (1.50)	0.10 (0.10)	0.0171
4	16	44.80 (1.57)	45.50 (1.58)	0.70 (0.33)	0.0161

Table 21: Unknown-rate comparison at $a = 0.2$, with the same design and denominators as Table 20.

ν	Δ	Known	Estimated	Difference	RMSE($\hat{\nu}$)
0.25	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0045
0.25	4	8.00 (0.86)	8.00 (0.86)	0.00 (0.00)	0.0042
0.25	16	40.50 (1.55)	40.40 (1.55)	-0.10 (0.26)	0.0044
0.5	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0060
0.5	4	6.70 (0.79)	6.70 (0.79)	0.00 (0.00)	0.0060
0.5	16	43.90 (1.57)	44.00 (1.57)	0.10 (0.30)	0.0062
1	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0087
1	4	7.70 (0.84)	7.70 (0.84)	0.00 (0.00)	0.0087
1	16	44.00 (1.57)	44.40 (1.57)	0.40 (0.28)	0.0086
2	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0124
2	4	5.80 (0.74)	5.80 (0.74)	0.00 (0.00)	0.0123
2	16	44.40 (1.57)	44.50 (1.57)	0.10 (0.26)	0.0122
4	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0167
4	4	5.20 (0.70)	5.20 (0.70)	0.00 (0.00)	0.0176
4	16	44.10 (1.57)	43.60 (1.57)	-0.50 (0.22)	0.0164

Table 22: Unknown-rate comparison at $a = 0.6$. This is a finite-strength diagnostic, not an application of the local efficiency theorem at a fixed nonvanishing amplitude.

ν	Δ	Known	Estimated	Difference	RMSE($\hat{\nu}$)
0.25	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0052
0.25	4	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0053
0.25	16	31.00 (1.46)	30.90 (1.46)	-0.10 (0.10)	0.0052
0.5	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0075
0.5	4	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0073
0.5	16	32.40 (1.48)	32.20 (1.48)	-0.20 (0.24)	0.0072
1	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0103
1	4	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0105
1	16	29.70 (1.45)	29.50 (1.44)	-0.20 (0.20)	0.0108
2	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0150
2	4	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0147
2	16	30.10 (1.45)	30.30 (1.45)	0.20 (0.24)	0.0151
4	1	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0212
4	4	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.0207
4	16	30.40 (1.46)	30.90 (1.46)	0.50 (0.22)	0.0199

The second-order rate check fixes $T = \tau = 1$, uses $\nu \in \{0.25, 1, 4\}$ and $N \in \{256, 4096\}$, and chooses the fourth-root amplitude to keep $\gamma^2\sqrt{J^{(2)}}/2 = 1$. Each cell uses 1000 independent datasets. Seeds are $942000000 + 100000i_\nu + 10N + b$. The N^{-1} floor is used exactly. Table 23 reports paired differences. Small expected counts can make finite-sample errors substantially different from their limiting value; the purpose of this comparison is to isolate the effect of estimating the rate on identical data.

Table 23: Estimated-rate second-order score. Errors and paired estimated-minus-known differences are percentage points, with one standard error in parentheses. Each row has 1000 paired datasets.

ν	N	a_N	Known	Estimated	Difference
0.25	256	0.5008	28.20 (1.42)	28.00 (1.42)	-0.20 (0.28)
0.25	4096	0.2504	21.70 (1.30)	21.70 (1.30)	0.00 (0.00)
1	256	0.5559	20.70 (1.28)	21.20 (1.29)	0.50 (0.36)
1	4096	0.2780	19.40 (1.25)	19.50 (1.25)	0.10 (0.10)
4	256	0.5755	16.40 (1.17)	16.20 (1.17)	-0.20 (0.35)
4	4096	0.2878	17.30 (1.20)	17.40 (1.20)	0.10 (0.17)

R.7 Independent calibration banks and matched-budget comparisons

We use the saved $N = 64$ training features and freeze training seed zero in the archive, corresponding to optimization seed 7000. The banks use fresh observations independent of the earlier calibration and test data. Each of ten banks contains 199 calibration and 200 test datasets per candidate. Every dataset contains 64 independent windows. The seed for bank b is $810000000 + 1000000b$; within candidate k, chunks of eight datasets use this seed plus $100000k + s$, where s is the chunk's starting index. Continuous-time observations are shared across resolutions within each bank. The first 199 datasets per candidate are calibration and the remaining 200 are test. Banks are mutually independent.

Every learned comparator has access to the same 800 original training datasets per candidate. The neural and linear decoders retain their original optimization/validation split; the moment comparators use all 800 for their fixed closed-form fit. There is no additional hyperparameter search or larger simulation budget for the moment methods. Their inputs are subsets of the same nine summaries in (89). Scores-only uses features 7–9, dispersion-only uses features 3–4, and the combined moment method uses features 3–9. After dividing each selected feature by its pooled training standard deviation, fit candidate mean μ_k and covariance

$$\Sigma_k = 0.95\widehat{\Sigma}_k + 0.05\,\mathrm{diag}(\widehat{\Sigma}_k) + 10^{-5}I.$$

The compatibility probabilities are the softmax of $-\{(f-\mu_k)^\top \Sigma_k^{-1}(f-\mu_k) + \log\det\Sigma_k\}/2$. The shrinkage is fixed in advance and not selected using calibration or test data. The coarse-grained method uses its analytic candidate score and needs no training simulations. Its fitting computation is therefore not matched to that of a neural network.

For the guard, fit the same form of means and covariances using all nine features. Its compatibility score is the negative Mahalanobis distance alone, without normalization across candidates. Both neural and support scores are rank-calibrated at $\alpha/2 = 0.025$ using the same 199 simulations per candidate. Their sets are intersected. The unguarded methods use $\alpha = 0.05$. Training data are independent of all calibration and test datasets for both scores.

The repeated-bank table uses 14,000 test datasets per resolution, balanced over the seven candidates. Coverage is the fraction retaining the true candidate. Empty is the fraction with no retained candidates. Report is the fraction with a nonempty set agreeing on null, forward, or reverse. Wrong/all uses all test datasets as denominator. Wrong/report divides the total wrong reports by the total reports across the ten banks, rather than averaging ten conditional ratios with different denominators. A timescale report requires a singleton nonnull set. This last rate is unconditional over all datasets, including the null cases.

Table 24: All repeated-bank decoders. Param. is seven-candidate point accuracy. Timescale is the unconditional singleton nonnull reporting rate. Percentages pool ten independent banks; conditional error uses the pooled report denominator.

Δ	Decoder	Param.	Cov.	Report	Timescale	Wrong/report
0.25	Neural	92.91	94.97	99.67	77.68	0.53
0.25	Guarded neural	92.91	95.38	97.69	67.47	0.42
0.25	Linear	89.80	94.83	99.34	68.81	0.83
0.25	Moments	93.12	95.02	99.59	78.21	0.46
0.25	CG quadratic	95.56	94.81	98.39	82.58	0.04
0.25	Scores only	83.26	94.91	98.99	47.76	0.92
0.25	Dispersion only	21.44	94.84	2.94	0.02	25.73
1	Neural	86.20	95.00	99.34	57.04	0.26
1	Guarded neural	86.20	95.14	98.11	42.90	0.39
1	Linear	85.75	94.89	99.51	55.26	0.67
1	Moments	86.91	94.84	99.44	58.96	0.35
1	CG quadratic	88.96	94.61	99.23	66.48	0.17
1	Scores only	61.84	94.82	90.72	6.01	1.70
1	Dispersion only	24.62	94.68	5.21	0.00	16.87
4	Neural	60.76	94.86	61.99	3.66	4.06
4	Guarded neural	60.76	95.15	53.22	2.81	2.86
4	Linear	60.89	94.98	62.28	3.81	3.84
4	Moments	61.06	94.90	61.69	4.64	3.52
4	CG quadratic	61.26	94.93	61.69	4.97	3.54
4	Scores only	32.80	94.67	20.17	0.00	5.52
4	Dispersion only	21.83	95.11	4.35	0.00	16.91
16	Neural	31.44	94.91	14.11	0.00	11.14
16	Guarded neural	31.44	95.64	10.65	0.39	7.98
16	Linear	32.14	94.79	13.32	0.00	9.76
16	Moments	31.95	94.86	12.24	0.00	8.58
16	CG quadratic	32.66	94.96	14.86	0.00	10.87
16	Scores only	16.74	94.69	4.61	0.00	44.43
16	Dispersion only	18.26	95.20	5.00	0.00	30.86

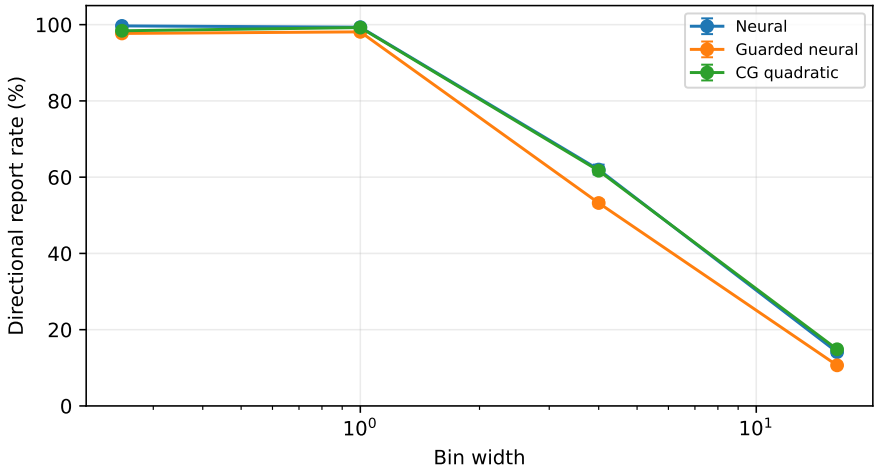


Figure 7: Repeated-bank directional reporting. Error bars use the sample standard deviation of the ten bank-level report frequencies divided by $\sqrt{10}$. They include calibration and test variability, conditional on fixed training.

Table 25: Standard errors across independent banks, in percentage points. Each entry is the sample standard deviation of ten bank-level frequencies divided by $\sqrt{10}$. These are not optimizer-seed standard errors or conditional-on-bank coverage guarantees.

Δ	Decoder	Cov. SE	Empty SE	Report SE	Wrong/all SE
0.25	Neural	0.18	0.05	0.05	0.07
0.25	Guarded neural	0.22	0.14	0.16	0.08
0.25	Linear	0.34	0.04	0.12	0.11
0.25	Moments	0.27	0.07	0.06	0.07
0.25	CG quadratic	0.25	0.21	0.21	0.02
0.25	Scores only	0.17	0.01	0.13	0.08
0.25	Dispersion only	0.22	0.00	0.25	0.13
1	Neural	0.17	0.08	0.09	0.04
1	Guarded neural	0.21	0.17	0.16	0.04
1	Linear	0.22	0.08	0.07	0.05
1	Moments	0.26	0.09	0.09	0.04
1	CG quadratic	0.25	0.08	0.08	0.03
1	Scores only	0.24	0.00	0.23	0.10
1	Dispersion only	0.22	0.00	0.52	0.11
4	Neural	0.22	0.01	1.30	0.12
4	Guarded neural	0.18	0.10	0.85	0.10
4	Linear	0.18	0.00	1.10	0.11
4	Moments	0.18	0.00	1.18	0.12
4	CG quadratic	0.20	0.00	0.92	0.15
4	Scores only	0.23	0.00	0.86	0.13
4	Dispersion only	0.20	0.00	0.36	0.12
16	Neural	0.33	0.00	0.76	0.27
16	Guarded neural	0.19	0.08	0.40	0.13
16	Linear	0.25	0.00	0.41	0.16
16	Moments	0.25	0.00	0.39	0.13
16	CG quadratic	0.36	0.00	0.62	0.18
16	Scores only	0.23	0.00	0.16	0.08
16	Dispersion only	0.24	0.00	0.42	0.18

At $\Delta = 16$, the unguarded neural rule reports direction on 14.107% of all datasets and makes wrong reports on 1.571%. Thus the pooled conditional error among reports is 11.139%, not below 5%. The guarded version has fewer reports, 10.650%, and conditional error 7.981%. Both are compatible

Table 26: Classwise coverage averaged over independent banks. Nonnull headings give direction and timescale. Each class uses 2000 test datasets across ten banks. Complete classwise outcomes for all methods are stored in `classwise_calibration.csv`.

Δ	Decoder	Null	$+.5$	$+1$	$+2$	$-.5$	-1	-2
0.25	Neural	95.8	94.0	96.2	94.1	94.8	94.7	95.3
0.25	Guarded neural	96.2	94.5	95.7	94.9	95.2	95.4	95.7
1	Neural	95.8	93.7	94.8	94.6	95.2	95.0	95.8
1	Guarded neural	96.5	93.8	96.3	94.7	94.8	94.9	95.0
4	Neural	94.4	94.8	94.9	94.8	94.7	96.0	94.5
4	Guarded neural	95.7	94.3	94.9	95.2	94.4	96.3	95.2
16	Neural	93.8	95.6	95.5	95.3	94.0	95.7	94.4
16	Guarded neural	94.8	96.1	95.5	96.4	95.2	96.3	95.2

with the unconditional theorem. Exact-class uncertainty and directional uncertainty are also different. Most methods make no singleton timescale reports at this resolution even though some directional reports remain.

R.8 Frozen-grid departures and their simulation accuracy

We keep all decoders and bank-zero calibration fixed. For each challenge, 400 forward and 400 label-reversed datasets are generated, with $N = 64$, $T = 64$, and coarsening widths $0.25, 1, 4, 16$. The scenarios are listed in Tables 27–30. Only the displayed parameter changes from $(\nu, a, \tau) = (1, 0.2, 1)$. The Erlang-2 kernel has shape two and scale $\tau/2$, hence mean τ. The mixture has equal weights on exponentials of means 0.25τ and 1.75τ, also with mean τ. Feedback has forward mass 0.2, reverse mass 0.05, common exponential mean lag one, and baselines chosen to keep both stationary mean rates one. It has no true one-way class.

The exponential feed-forward sampler is exact. For the Erlang and mixture challenges, source points are simulated on $[-L, T]$, with $L = 40\tau$ and 70τ, respectively. Omitted offspring from earlier sources have expected number at most

$$\nu a \int_L^\infty \overline{H}(u)\, du = \nu a \mathbb{E}[(U - L)_+]$$

per window. This is also a coupling bound on the count law's total variation. Multiply by 64 for a dataset. For the Erlang kernel the integral equals $\tau e^{-2L/\tau}(1 + L/\tau)$; for the mixture it is the corresponding weighted sum of exponential mean residual tails. These bounds are below 10^{-30} and 10^{-16} per dataset in the stated designs, respectively.

The feedback sampler starts immigrants at -100τ and follows their full Poisson offspring trees through time T. It is not described as an exact stationary sampler. Each parent has mean total offspring at most $\rho = 0.2$. Dominating by a scalar exponential-kernel branching process, the expected density of generation-one-and-later descendants at age u is at most $(\rho/\tau)e^{-(1-\rho)u/\tau}$. With total immigrant rate at most 2ν, the omitted event count per window is bounded by

$$\frac{2\nu\rho\tau}{(1-\rho)^2} e^{-(1-\rho)L/\tau}.$$

Multiplying by 64 bounds the dataset discrepancy. It is below 10^{-32} for this challenge. The construction keeps all simulated descendants before T and requires no arbitrary generation cutoff.

Challenge i uses seed $830000000 + 1000000i$, then adds $100000k + s$ for orientation k and chunk index s. The feature encoder still uses the training rate one in the rate challenges. This evaluates misspecification of a fixed decoder, separately from the statistically adapted score test of Theorem 6.

There is no exact true candidate outside the grid. Accordingly, exact-candidate accuracy and coverage are not reported for these challenges. For nonnull feed-forward alternatives, directional correctness compares against the true orientation; a null report is wrong. Kernel-envelope evaluation checks the true signed cumulative kernel at 257 equally spaced lags in $[0, 8]$ and its limiting amplitude. Empty sets fail this evaluated-lag coverage criterion and make no directional assertion. A successful

finite-lag check is not a proof of simultaneous off-grid coverage. For feedback, both directional error and this one-way-kernel coverage are undefined; the tables report only the frequency of purported one-way reports and empty sets. Singleton and report frequencies describe the output, not correctness, when the data have feedback.

The guard improves some challenges and worsens others. At $\Delta = 1$, $\nu = 2$, it rejects all datasets, whereas the unguarded neural rule reports on all of them with 50% wrong direction. At $a = 0.05$, however, guarded wrong reports increase to 74% of all datasets from 66.5%. The guard uses half-level constituent tests and need not be nested inside the original level-α set. A weak omitted alternative can resemble the on-grid null and evade an absolute-fit check. At an off-grid timescale of 0.75, directional classification is accurate but evaluated-lag envelope coverage remains only 30% for the base rule and 43.5% for the guard. Erlang and mixture kernels can fall outside every reported envelope even when direction is correctly recovered. These failures occur outside the grid guarantee and do not contradict the on-grid exchangeability proof.

Table 27: Frozen-grid challenges at $\Delta = 0.25$. Each case has 800 independent test datasets and a shared frozen calibration bank. All reported numbers are percentages. "Envelope" is the evaluated-lag criterion, and feedback entries without a one-way truth are marked –.

True setting	Rule	Report	Empty	Wrong/all	Envelope
On grid	Base	100.00	0.00	0.00	95.75
On grid	Guard	98.25	1.75	0.00	95.62
$\nu = 0.5$	Base	100.00	0.00	50.00	0.00
$\nu = 0.5$	Guard	0.00	100.00	0.00	0.00
$\nu = 2$	Base	100.00	0.00	0.00	0.00
$\nu = 2$	Guard	0.00	100.00	0.00	0.00
$a = 0.05$	Base	86.75	13.25	56.62	0.00
$a = 0.05$	Guard	92.00	6.62	60.62	1.38
$a = 0.1$	Base	94.12	5.88	5.62	0.00
$a = 0.1$	Guard	92.88	6.62	7.38	0.50
$a = 0.4$	Base	100.00	0.00	0.00	0.00
$a = 0.4$	Guard	4.88	95.12	0.00	0.00
$a = 0.6$	Base	100.00	0.00	0.00	0.00
$a = 0.6$	Guard	0.00	100.00	0.00	0.00
$\tau = 0.75$	Base	100.00	0.00	0.00	12.62
$\tau = 0.75$	Guard	97.00	3.00	0.00	27.75
$\tau = 1.5$	Base	99.88	0.12	0.00	32.00
$\tau = 1.5$	Guard	98.75	1.25	0.12	53.25
Erlang-2	Base	100.00	0.00	0.00	0.00
Erlang-2	Guard	89.50	10.50	0.00	0.00
Mixture	Base	100.00	0.00	0.00	0.00
Mixture	Guard	93.00	7.00	0.00	0.00
Feedback	Base	100.00	0.00	–	–
Feedback	Guard	90.00	10.00	–	–

Table 28: Frozen-grid challenges at $\Delta = 1$, with the same definitions as Table 27.

True setting	Rule	Report	Empty	Wrong/all	Envelope
On grid	Base	100.00	0.00	0.12	93.25
On grid	Guard	97.88	2.12	0.12	93.38
$\nu = 0.5$	Base	33.38	0.00	1.00	78.00
$\nu = 0.5$	Guard	0.00	100.00	0.00	0.00
$\nu = 2$	Base	100.00	0.00	50.00	0.00
$\nu = 2$	Guard	0.00	100.00	0.00	0.00
$a = 0.05$	Base	85.00	14.75	66.50	0.25
$a = 0.05$	Guard	92.50	7.12	74.00	0.38
$a = 0.1$	Base	87.00	12.25	13.38	0.75
$a = 0.1$	Guard	90.88	8.75	17.50	0.38
$a = 0.4$	Base	100.00	0.00	0.00	0.00
$a = 0.4$	Guard	0.62	99.38	0.00	0.00
$a = 0.6$	Base	100.00	0.00	0.00	0.00
$a = 0.6$	Guard	0.00	100.00	0.00	0.00
$\tau = 0.75$	Base	100.00	0.00	0.00	30.00
$\tau = 0.75$	Guard	98.25	1.75	0.00	43.50
$\tau = 1.5$	Base	99.88	0.12	0.12	50.38
$\tau = 1.5$	Guard	99.12	0.88	0.12	64.75
Erlang-2	Base	100.00	0.00	0.00	0.00
Erlang-2	Guard	90.00	10.00	0.00	0.00
Mixture	Base	99.62	0.12	0.38	0.00
Mixture	Guard	97.00	2.88	0.38	0.00
Feedback	Base	99.50	0.00	–	–
Feedback	Guard	94.62	5.25	–	–

Table 29: Frozen-grid challenges at $\Delta = 4$, with the same definitions as Table 27.

True setting	Rule	Report	Empty	Wrong/all	Envelope
On grid	Base	53.62	0.00	2.62	96.38
On grid	Guard	45.88	1.12	1.50	96.38
$\nu = 0.5$	Base	0.00	0.00	0.00	100.00
$\nu = 0.5$	Guard	0.00	100.00	0.00	0.00
$\nu = 2$	Base	0.12	0.00	0.00	100.00
$\nu = 2$	Guard	0.00	100.00	0.00	0.00
$a = 0.05$	Base	78.88	0.00	63.25	16.75
$a = 0.05$	Guard	63.88	1.25	57.25	26.62
$a = 0.1$	Base	67.38	0.00	30.25	29.88
$a = 0.1$	Guard	49.88	1.12	22.88	43.38
$a = 0.4$	Base	27.00	0.00	0.00	0.00
$a = 0.4$	Guard	9.12	88.75	0.00	0.00
$a = 0.6$	Base	16.25	0.00	0.00	0.00
$a = 0.6$	Guard	0.00	100.00	0.00	0.00
$\tau = 0.75$	Base	42.62	0.00	2.50	89.75
$\tau = 0.75$	Guard	35.50	1.00	1.75	89.38
$\tau = 1.5$	Base	74.75	0.00	1.25	93.12
$\tau = 1.5$	Guard	65.38	0.88	1.00	95.00
Erlang-2	Base	60.25	0.00	0.88	35.88
Erlang-2	Guard	52.62	0.88	0.75	42.62
Mixture	Base	50.00	0.00	2.12	0.00
Mixture	Guard	41.12	1.62	1.62	0.00
Feedback	Base	27.62	0.00	–	–
Feedback	Guard	26.88	4.38	–	–

Table 30: Frozen-grid challenges at $\Delta = 16$, with the same definitions as Table 27. Full results for linear, moment, score-only, dispersion-only, and coarse-grained decoders are supplied in `misspecification.csv`.

True setting	Rule	Report	Empty	Wrong/all	Envelope
On grid	Base	4.62	0.00	1.00	97.00
On grid	Guard	2.50	1.12	0.88	96.38
$\nu = 0.5$	Base	0.00	0.00	0.00	100.00
$\nu = 0.5$	Guard	0.00	100.00	0.00	0.00
$\nu = 2$	Base	0.00	0.00	0.00	100.00
$\nu = 2$	Guard	0.00	100.00	0.00	0.00
$a = 0.05$	Base	35.75	0.00	34.00	52.38
$a = 0.05$	Guard	30.88	0.62	30.38	57.75
$a = 0.1$	Base	14.88	0.00	12.50	75.88
$a = 0.1$	Guard	11.25	0.25	10.62	80.75
$a = 0.4$	Base	0.88	0.00	0.00	0.00
$a = 0.4$	Guard	11.88	32.62	1.88	0.00
$a = 0.6$	Base	0.75	0.00	0.00	0.00
$a = 0.6$	Guard	0.50	98.12	0.00	0.00
$\tau = 0.75$	Base	4.75	0.00	1.38	93.38
$\tau = 0.75$	Guard	3.25	0.75	0.88	94.62
$\tau = 1.5$	Base	4.00	0.00	1.00	96.75
$\tau = 1.5$	Guard	2.50	0.38	0.75	97.00
Erlang-2	Base	3.50	0.00	1.00	92.25
Erlang-2	Guard	2.50	0.75	0.88	95.00
Mixture	Base	4.38	0.00	1.00	0.00
Mixture	Guard	2.00	1.38	0.62	0.00
Feedback	Base	2.00	0.00	–	–
Feedback	Guard	1.88	1.12	–	–

R.9 DEPENDENT-WINDOW DIAGNOSTIC

This diagnostic isolates the observation-independence assumption from neural fitting. A fixed one-sided score test excludes the reverse candidate when the aggregate temporal score is too large under independent reverse simulations. Its calibration bank has 999 independent datasets, each with $N = 32$ windows of $T = 8$, and uses seed 880000001. Set $\nu = 1, a = 0.6, \tau = 8, \Delta = 4$. The test data are 2000 independent long trajectories in each gap condition. Each trajectory supplies 32 windows separated by $g \in \{0, 8, 32, 64, 96\}$. Every underlying process is reverse. Thus a forward report is false, and the test assesses one necessary component of safe directional reporting rather than a full neural confidence set.

For gap index i_g and chunk index s, test seeds are $890000000 + 100000i_g + s$. An independent-window test reference uses seed 880000003 with the same calibration bank. Table 31 records the nominal 5% test and the adjusted level $\alpha_0 = (0.05 - \epsilon_g)_+$. When the bound exceeds 0.05, the adjusted test never reports. At $g = 96$, the usable correction is $\epsilon_g = 0.001156$, giving $\alpha_0 = 0.048844$. The nominal and adjusted empirical false-report frequencies are 4.20% and 3.95%. The independent reference has 5.50% false reports on this realized calibration bank. None of these conditional-on-bank frequencies alone proves or contradicts the marginal guarantee.

Table 31: Dependent-window calibration diagnostic. Errors are percentages, with one test Monte Carlo standard error in parentheses for the nominal rule. Gaps and window lengths are measured in the same time units. An adjusted level of zero represents compulsory abstention.

Gap	ϵ_g	Unadjusted error	α_0	Adjusted error
0	1.000000	5.20 (0.50)	0.000000	0.00
8	1.000000	5.65 (0.52)	0.000000	0.00
32	1.000000	4.95 (0.49)	0.000000	0.00
64	0.063107	4.30 (0.45)	0.000000	0.00
96	0.001156	4.20 (0.45)	0.048844	3.95
Independent	0.000000	5.50 (0.51)	0.050000	5.50

R.10 FILES, CHECKS, AND UNTESTED CLAIMS

The code directory contains separate programs for finite-sample and nuisance studies, repeated-bank calibration, numerical count likelihoods, fourth-root experiments, and nonvacuous checks. The archive README gives their filenames and commands. They write CSV and compressed prediction arrays under `results/revision`. The figure and table generator reads those recorded outputs without refitting a model. The inference script supports both unguarded and guarded reporting on the documented grid. Training checkpoints are unchanged from the original study. The additional fits are the training-only Gaussian summary fits described above. No neural network is retrained on the challenges.

The experiments do not establish robustness to an arbitrary rate function, unequal marginal means, unknown continuous kernel families, inhibition, or general feedback. No real event dataset, general latent-time Bayesian estimator, or interval-censored Transformer training run is included. The count-likelihood reference is numerical and restricted to one bin. No neural superiority or forecasting advantage follows from these results. The unresolved many-bin joint-strength transition is not replaced by the two-channel sufficient bound. These restrictions delimit the empirical evidence.

S INFORMATION RETAINED BY THE COUNT SUMMARIES

This section connects the binary testing results to the summaries used by the decoder. Representing an attaining statistic and learning its decision rule are separate requirements. No local optimality claim is made for the frozen seven-class networks.

S.1 THE LOCAL CONTRIBUTION OF THE MIXED COUNT MOMENT

Let $V = c_2(Y) - c_2(X)$ and $M = c_1(X)c_2(Y) - c_1(Y)c_2(X)$ in the single-bin experiment. Appendix K establishes

$$\mathbb{E}_{a,\pm}(V, M) = \pm a^2 \nu(\beta_2, \chi_2), \qquad \mathrm{Cov}_0(V, M) = \begin{pmatrix} 4\lambda^2 & 0 \\ 0 & 4\lambda^3 \end{pmatrix}.$$

For fixed weights $w = (w_V, w_M)$ with positive directional mean, the sign of $\sum_i (w_V V_i + w_M M_i)$ has limiting risk

$$\Phi\left(-\frac{\gamma^2 \nu(w_V \beta_2 + w_M \chi_2)}{\sqrt{4\lambda^2 w_V^2 + 4\lambda^3 w_M^2}} \right), \qquad a_N = \gamma N^{-1/4}. \tag{93}$$

The exact means above determine the local shift. Variance continuity and the uniform higher count moments used in Theorem 5 give the triangular-array central limit theorem for this fixed polynomial. A nonzero weight vector has positive null variance, so threshold zero is a continuity point of the limit.

Cauchy–Schwarz in this diagonal covariance metric shows that the largest squared shift in (93) is $\gamma^4 J^{(2)}/4$. It is attained by weights proportional to $(\nu\beta_2/\lambda^2, \nu\chi_2/\lambda^3)$, which gives $D^{(2)}$. Setting $w_M = 0$ instead gives squared shift $\gamma^4 \beta_2^2/(4T^2)$. The ratio is ρ_V in (15). The statement concerns the efficacy of the summed dispersion contrast. It does not characterize arbitrary nonlinear procedures based on dispersion statistics.

At a fixed kernel and window shape, write $q_T = \chi_2/\beta_2$ and $\lambda = \nu T$. Then

$$\rho_M = \frac{q_T^2}{\lambda + q_T^2}.$$

It is decreasing in expected count, and is at least one half exactly when $\lambda \le q_T^2$. At $T = \tau = 1$, $q_T^2 = 0.208775$. The rates displayed in Figure 8 index separate fixed-rate local experiments; no uniform limit in which ν also approaches zero is asserted.

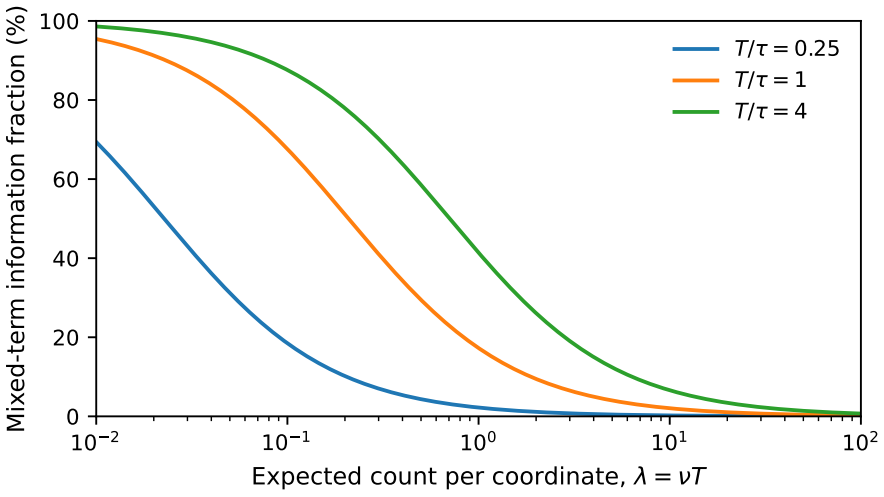


Figure 8: Fraction ρ_M of second-order local information supplied by the mixed term. Each curve fixes T/τ and varies the expected count $\lambda = \nu T$. The formula uses the stationary exposure integrals, not simulated errors. The corresponding comparison of summed statistics is (93).

S.2 AN EXACT SUMMARY-TO-SCORE MAP

For a dataset of N windows, let $F = N^{-1/2} \sum_i f^{(i)}$, where f is the nine-dimensional feature map in (89). At multiple bins and every dictionary kernel with $J_\ell > 0$,

$$\frac{1}{\sqrt{N}} \sum_i D^{(i)}_{h_{\tau_\ell}, \Delta} = \sqrt{J_\ell}\, F_{6+\ell}. \tag{94}$$

Thus its binary first-order sign test is a linear readout of the decoder input. The readout follows algebraically, without training. A trained seven-class decoder need not implement it.

At $m = 1$, set $f_6 = f_7 = f_8 = f_9 = 0$. Add one coordinate

$$F_{10} = \frac{1}{2\lambda^{3/2}\sqrt{N}} \sum_i \{c_1(X_i)c_2(Y_i) - c_1(Y_i)c_2(X_i)\}. \tag{95}$$

It has unit variance at the null. Since $\sum_i V_i = \lambda\sqrt{2N}(F_4 - F_3)$, the second-order readout is

$$\frac{1}{\sqrt{N}} \sum_i D_i^{(2)} = \frac{\nu\beta_2}{\sqrt{2}\lambda}(F_4 - F_3) + \frac{\nu\chi_2}{\lambda^{3/2}} F_{10}. \tag{96}$$

Substituting the feature normalizations into the definitions proves the identities. Equations (94) and (96) therefore recover the attaining statistics exactly, under the corresponding observation designs. They do not make the summaries sufficient for the full likelihood at nonzero amplitude, for arbitrary kernels, or for a many-bin second-order experiment.

The missing coordinate cannot in general be reconstructed after the original summaries have been averaged. At $N = 4$, $\lambda = 1$, take the two datasets in Section 4. Both have

$$\sum X_i = \sum Y_i = 6, \qquad \sum X_i^2 = \sum Y_i^2 = 14, \qquad \sum X_i Y_i = 11.$$

These quantities determine $F_1, \ldots, F_5$, and the other original coordinates vanish. Nevertheless, the mixed sums equal 2 and -2. The two datasets are in the support of both orientations because the source rate and target baseline are positive. Hence no deterministic function of the nine aggregate summaries can reconstruct $\sum_i D_i^{(2)}$ on all possible datasets. This finite-sample collision is not a minimax or asymptotic lower bound for every nonlinear decoder of those summaries.

A separate example illustrates a strict reversal rather than a dispersion tie. With $N = 2$, $X = (0, 0)$ and $Y = (4, 3)$ at the same parameters, $\sum V_i = 4$ and $\sum M_i = -11$. The dispersion rule favors the forward orientation, whereas

$$\sum_i D_i^{(2)} = 2\beta_2 - \tfrac{11}{2}\chi_2 = -0.188743$$

favors the reverse orientation. This is a count-polynomial calculation, not an assurance of reliable inference from two windows.

The function `single_bin_summaries` in `count_law.py` implements the enlarged summary. The existing checkpoints retain their nine-feature input. The additional feature makes the decision rule representable but does not show that a network has learned it. A new fitted decoder would require independent fitting and calibration before invoking Theorem 8.

T PAIRED NUMERICAL EXPERIMENTS FOR THE MIXED TERM

The experiment isolates the contribution of M in the fixed-kernel binary problem. Its design was written to `paired_design.json` before sampling. We use $(T, \nu) \in \{(1, 0.25), (1, 1), (1, 4), (8, 1)\}$, $\tau = 1$, and $N \in \{4096, 65536\}$. In each cell, $a_N = \{2/\sqrt{J^{(2)}}\}^{1/2} N^{-1/4}$, so the efficient local risk is $\Phi(-1)$. Each of the eight cells contains 20,000 independently generated datasets, separate from both the earlier event-sampler datasets and the referee's numerical checks. Within a dataset, all three rules see exactly the same count observations. The complete design has 160,000 datasets, with separate seeds across cells.

T.1 POSITIVE-COEFFICIENT EVALUATION OF THE COUNT LAW

Let b_k be the offspring-count probability for one source point uniformly distributed in the observed window,

$$b_k = \frac{1}{T} \int_0^T e^{-a p_T(s)} \frac{\{a p_T(s)\}^k}{k!}\, ds, \qquad k \geq 0.$$

The baseline and parents before zero contribute an independent count with compound-Poisson jump rates

$$r_k = \mathbf{1}\{k=1\}\nu(1-a)T + \nu\int_{-\infty}^{0} e^{-ap_T(s)}\frac{\{ap_T(s)\}^k}{k!}\,ds, \qquad k \geq 1.$$

Their total nonzero-jump rate is $r = \nu(1-a)T + \nu\int_{-\infty}^{0}(1-e^{-ap_T(s)})\,ds$. If q_k is this independent background count's probability mass, coefficient comparison in $\sum_k q_k z^k = \exp\{-r + \sum_{j\geq 1} r_j z^j\}$ gives

$$q_0 = e^{-r}, \qquad q_k = \frac{1}{k}\sum_{j=1}^{k} j r_j q_{k-j}. \tag{97}$$

Conditional on $X = x$, the x observed source locations are independent uniforms, so

$$p_a^+(x,y) = \mathrm{Pois}(\nu T)(x)\,[q * b^{*x}]_y. \tag{98}$$

The implemented recursion therefore follows directly from the cluster model. Every coefficient is nonnegative, and only coefficients through y enter the recursion for probability y. Truncating arrays at L does not fold the upper tail back onto lower counts. For the exponential kernel, the substitution $u = e^{s/\tau}$ makes the past integral a finite integral on $[0,1]$ with weight τ/u. The integrands used for positive jumps have finite endpoint limits. We evaluate both past and within-window integrals by 256-node Gauss–Legendre quadrature.

For a separate numerical check, we evaluate the generating function in Appendix R on a doubled Fourier grid with 512 quadrature nodes. We also double the positive-recursion array and quadrature sizes. Differences and moment checks are recorded for every cell, including those omitted from the main table. This is floating-point quadrature and coefficient extraction, not a verified-arithmetic likelihood enclosure.

T.2 Truncation and sampling

For $t > 0$ put $v = e^t - 1$. The source count has log moment generating function $\nu T v$. For the target, convexity of e^{avp} on $0 \leq p \leq 1$ and $\int p_T = T$ give

$$\log \mathbb{E}e^{tY} \leq \nu T\{(1-a)v + e^{av} - 1\}.$$

Consequently the sum of the two Chernoff bounds

$$\eta_L = \min\left\{1, \inf_{t>0} e^{-tL+\nu T(e^t-1)} + \inf_{t>0} e^{-tL+\nu T\{(1-a)(e^t-1)+e^{a(e^t-1)}-1\}}\right\} \tag{99}$$

bounds the probability that either count is at least L. The implementation uses positive numerically selected t, whose objective values remain upper bounds even without certified minimization. Starting at $L = 32$, it doubles L until $N\eta_L < 10^{-10}$. It retains the symmetric set of lattice cells with $p_a^+(x,y) + p_a^+(y,x) > 10^{-22}$, and records the additional omitted mass. Symmetric support permits the reverse likelihood to be evaluated on the same lattice. Probability masses are then normalized on this support.

For a dataset of N independent windows, the histogram of count pairs is multinomial with this normalized mass vector. We sample that histogram and take its inner products with the per-pair values of $D^{(2)}$, V, $\log\{p_a^+(x,y)/p_a^+(y,x)\}$, and M. This has the same distribution as generating and aggregating N independent pairs from the numerical law. It avoids simulating N latent event sequences without replacing the dataset statistic by a Gaussian draw. A product coupling bounds truncation error by N times the omitted per-pair probability; numerical quadrature error is assessed separately by the comparisons above.

Across the eight cells, the largest absolute probability discrepancy after doubling the quadrature and lattice is 1.30×10^{-16}, and the largest discrepancy from the separate Fourier calculation is 1.58×10^{-16}. The largest N-scaled sum of the analytic overflow bound and discarded lattice mass is 8.90×10^{-11}. The largest residual in the exact mean identity for $D^{(2)}$ is 1.18×10^{-15}. The computed discarded-mass and Chernoff quantities do not enclose all floating-point error. The implemented likelihood remains a numerical reference, despite agreement between the two constructions.

T.3 PAIRED RESULTS AND FINITE-STRENGTH BEHAVIOR

For method k and dataset b, the stored bounded loss is $e_{bk} = \mathbf{1}\{S_{bk} < 0\} + \frac{1}{2}\mathbf{1}\{S_{bk} = 0\}$ under the forward law. Label exchange gives the corresponding equal-prior risk of random tie-breaking. The estimate is $\bar{e}_k$, with standard error $s(e_{\cdot k})/\sqrt{B}$. For two methods, we report $\overline{e_k - e_l}$ and $s(e_{\cdot k} - e_{\cdot l})/\sqrt{B}$, not a sum of marginal uncertainties. These are Monte Carlo standard errors conditional on the numerical law and fixed design, without multiplicity adjustment.

Table 32: All paired count-law experiments. Parentheses in the risk columns are marginal Monte Carlo standard errors. Errors are percentages. Each row uses 20,000 independent datasets. The efficient limit is 15.87%.

T	ν	N	a_N	R_D (SE)	R_V (SE)	R_L (SE)
1	0.25	4,096	0.2504	22.70 (0.30)	23.81 (0.30)	21.99 (0.29)
1	0.25	65,536	0.1252	19.43 (0.28)	23.18 (0.30)	19.19 (0.28)
1	1	4,096	0.2780	18.15 (0.27)	18.20 (0.27)	17.66 (0.27)
1	1	65,536	0.1390	16.81 (0.26)	18.48 (0.27)	16.82 (0.26)
1	4	4,096	0.2878	16.88 (0.26)	17.04 (0.27)	16.91 (0.27)
1	4	65,536	0.1439	15.90 (0.26)	16.62 (0.26)	15.99 (0.26)
8	1	4,096	0.1842	16.34 (0.26)	17.11 (0.27)	16.23 (0.26)
8	1	65,536	0.0921	16.73 (0.26)	17.19 (0.27)	16.60 (0.26)

Table 33: Paired error differences for Table 32, in percentage points. Parentheses are standard errors of within-dataset differences. D, V, and L denote second-order score, dispersion, and numerical likelihood. The last column is dispersion's local-limit risk in percent.

T	ν	N	ρ_M (%)	$R_D - R_V$ (SE)	$R_D - R_L$ (SE)	$R_{V,\infty}$ (%)
1	0.25	4,096	45.51	−1.11 (0.29)	+0.71 (0.18)	23.02
1	0.25	65,536	45.51	−3.75 (0.29)	+0.24 (0.13)	23.02
1	1	4,096	17.27	−0.05 (0.20)	+0.49 (0.13)	18.15
1	1	65,536	17.27	−1.67 (0.21)	−0.01 (0.09)	18.15
1	4	4,096	4.96	−0.17 (0.14)	−0.03 (0.09)	16.48
1	4	65,536	4.96	−0.73 (0.15)	−0.10 (0.06)	16.48
8	1	4,096	9.73	−0.77 (0.17)	+0.10 (0.10)	17.10
8	1	65,536	9.73	−0.45 (0.17)	+0.14 (0.07)	17.10

At $\nu T = 1$, the large-N score–dispersion separation is resolved, whereas the smaller-N difference remains within its Monte Carlo uncertainty. At $\nu T = 0.25$, the mixed term accounts for 45.51% of the local information, but finite-strength convergence is slower. The $T = 8$ cells show that a larger sample need not produce a monotone sequence of observed errors in independent Monte Carlo runs. The experiments do not establish that either statistic dominates at every nonzero amplitude. The event-sampler table also includes cells in which dispersion has lower observed error.

The seeds are $960000000 + 10000i + j$, with i indexing the four (T, ν) pairs in the order above and $j \in \{0, 1\}$ indexing N. NumPy's default random generator samples chunks of 256 histograms. The polynomial score is formed from the summed V and M components, preserving the algebraic tie when both are zero. The archive retains each dataset's four summed statistics and three losses, together with the support, probabilities, weights, and design. The program `run_round2_paired.py` generates the complete study. No network is trained on these evaluation datasets.

U COMPLETE BANK RECORDS AND VERIFICATION

The ten-bank study was rerun from the original seeds using the same frozen checkpoints and training summaries. Every bank's calibration features and scores, test features and scores, absolute-fit scores, labels, and candidate sets are now saved. All 280 aggregate rows, 1960 candidatewise rows, and 336 challenge rows match the preceding record exactly in the recorded environment. This is a seeded rerun of the same experiment, not ten additional independent banks. Original bank-zero keys are retained for compatibility with the inference and earlier checking programs.

The lightweight checker `verify_round2.py` requires NumPy and SciPy but neither PyTorch nor Numba. It reconstructs the rank sets from each bank's calibration and test scores, including tied scores and the intersection of the two half-level tests. It recomputes every aggregate and candidatewise reporting metric. It also checks the paired losses and their uncertainty, the count-law calculations, null information identities, the summary collision and readout, the one-bin zero Gram matrix, and the multi-bin rank inequality. Saved checkpoint hashes are checked separately from evaluating their predictions. The framework-based re-forward check covers all 108 original checkpoints. Their largest probability difference from the stored float32 outputs is 8.94×10^{-7} in the present environment, with no changed classifications or candidate sets. The checker declares an absolute probability tolerance of 2×10^{-6} while requiring identical decisions and sets. This numerical tolerance does not alter any saved probability or reported result.