Posterior Distortion from Hypothesis-Dependent Validity Gating in Runtime Diagnosis

Andres Felipe Peña Avila
SWORT LLC, Bethesda, MD, USA
andrfp@swort.net

September 2026

Abstract

A runtime validity gate can be treated as a binary diagnostic observation: its predicate is evaluated on observables whose distribution may depend on the fault state, so its outcome induces a hypothesis-dependent likelihood ratio. Monitors that use the gate only to select which evidence channels to score, and then rank hypotheses over the survivors, discard that detector's output. We give an exact account of what is discarded.

Writing q^{omit} for the distribution such a monitor produces and p for the Bayesian posterior given the full observation, the two differ by an exponential tilt in the gate likelihood, and every pairwise log odds shifts by the gate's log-likelihood ratio for that pair. The correction vanishes for all pairs precisely when the gate likelihood is constant across hypotheses, which for a binary gate is conditional independence of gate and hypothesis given context. The identity does not assume the surviving evidence is independent of the gate.

Let R be the range of the gate log-likelihood, the maximal pairwise log-likelihood ratio the gate outcome induces. Writing $A_R = R/(1-e^{-R})$, the Kullback-Leibler divergence is bounded by $A_R - 1 - \log A_R$ and the total variation by $\tanh(R/4)$; both are attained, and both are the corresponding cases of a bounded-likelihood-ratio theorem for which we claim nothing. Distortion does not determine decision: the leader survives the omission exactly when the gate's pairwise log-likelihood ratio stays below the pre-gate decision margin, and we exhibit arbitrarily small divergence with a changed decision and arbitrarily large divergence without one. Separately, admitting the gate as a signature coordinate can only refine the declared identifiability partition; the probabilistic and structural analyses are parallel and neither implies the other.

The component results are established and we claim priority for none of them individually. The contribution is a single connected account, organised by the observation that a gate outcome is itself a diagnostic observation, of what a diagnostic monitor gives up by treating it as context. Appendices extend the account to smoothed and tempered scoring, interval certificates under gate-model uncertainty, conditionally dependent gates, and source-attribution error. A worked eight-hypothesis configuration shows a $45:1$ gate moving nominal posterior mass from 0.860 to 0.257 while reversing one pairwise ranking by only 0.045 nats.

Keywords: validity gating; informative censoring; nonignorable missingness; fault signature matrix; structural isolability; runtime diagnosis.

Declarations. *Funding:* This work was supported internally by SWORT LLC. *Competing interests:* The author is the founder of and has an ownership interest in SWORT LLC. *Data availability:* No empirical data were used; every numerical value is generated from the stated model assumptions by the reproduction script `reproduce.py`, supplied as supplementary material. *Code availability:* The reproduction package (`reproduce.py`, `srcattr_controls.py`, and their JSON outputs) is supplied as supplementary material; no permanent public archive is claimed. *Author contribution:* Single author.

1 Introduction and exact problem

Consider a monitor that guards a residual with a validity predicate. When the predicate fails, the residual is meaningless and the monitor deactivates it. The monitor then diagnoses using the channels that remain. This is standard and defensible practice.

The difficulty arises when the predicate is not diagnostically inert. If a fault under diagnosis is itself likely to trip the guard, then the guard's state is an observation about that fault, and deactivating the guarded channels while conditioning on the survivors discards it. The discarded quantity is exactly computable, and this paper computes it.

Example 1 (Before any notation)**.** Two hypotheses: a receiver is nominal, or it is jammed. One residual compares receiver velocity against inertial velocity. One guard: the receiver validity flag. Jamming trips the flag with probability 0.9; nominal operation trips it with probability 0.02. When the flag trips, the residual is deactivated, so no residual evidence exists. The context-conditioned monitor

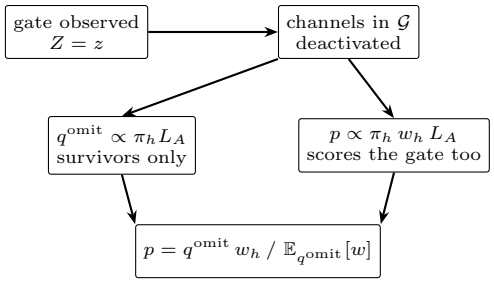


Figure 1. The omission. A gate outcome z is observed and used to deactivate channels. The omitted-gate score ranks over the survivors; the gate-aware posterior additionally scores the event $Z = z$. Both use z to choose the active set; only p scores it. Proposition 1 gives the exact relation.

now has no active evidence and returns its prior. The gate-aware monitor observes an event that is 45 times more likely under jamming and multiplies the prior odds accordingly. The first monitor discarded the only evidence in the epoch, in the epoch where that evidence was most informative.

The organising fact. A validity gate can be treated as a binary diagnostic observation. Its predicate is evaluated on observables whose distribution may depend on the fault state, so the gate's outcome induces a hypothesis-dependent likelihood ratio, exactly as a binary test statistic would. The quantity this paper calls the endogeneity range,

$$R = \max_h \log p(Z = z \mid h) - \min_h \log p(Z = z \mid h),$$

is exactly the maximal pairwise log-likelihood ratio induced by the gate outcome. A monitor that treats the gate as pure context, using it only to select which channels to score, discards that observation. We do not claim this observation as new; it is the reason the rest of the paper exists, and every result below is a consequence of it.

Four questions follow, each the corresponding question one would ask of any diagnostic observation that was dropped from scoring. What does its output do to the posterior when restored (Section 4)? How far can a detector with log-likelihood range R move a distribution (Section 5)? Does admitting it as a signature coordinate change which hypotheses are structurally separable (Section 6)? And when does it actually change the reported diagnosis, which is a different question from how far it moves the distribution (Section 7)? A worked configuration follows (Section 8). Practical smoothed scoring, interval certificates under gate-model uncertainty, multiple dependent gates, and source-attribution error are treated in the appendices; each is a consequence of the same fact but none is needed for the main line.

What this paper is. The component mathematical tools are established: likelihood reweighting of a posterior, bounded-likelihood-ratio divergence bounds, argmax stability, and coordinate refinement of equality partitions. Section 2 identifies where each lives. The contribution is the connected account: that a validity gate outcome is a diagnostic observation with a hypothesis-dependent likelihood ratio, that context conditioning discards it, and that the omission has an exact probabilistic cost, an exact decision-change criterion, and a separate structural consequence, all stated in the signature-matrix language the fault detection and isolation community uses. The status of every claim is recorded in Appendix F.

2 Closest prior theory

Diagnosis under unreliable and fault-dependent observation is well established in the discrete-event-systems literature. Dong, Yin and Li [1] give a uniform framework for diagnosis with unreliable sensors in which the observation map $O : Q \times \Sigma \to 2^{\Delta \cup \{\epsilon\}}$ is *state dependent*, and the verification construction augments the state with a fault label. Since the fault is encoded in the state, their framework permits observation loss that depends on the fault itself, and sensor constraints are arbitrary linear temporal logic formulas. Their setting encompasses the earlier robust-diagnosability results of Carvalho et al. on intermittent [2] and permanent [3] loss of observations, and Takai's general framework for sensor failures [4]. **Priority for the general mechanism belongs to that line and we assert none.** In the literature reviewed for this work we did not identify an equivalent treatment that combines the gate-likelihood posterior correction with fault-signature partition analysis in the form given here; the searches conducted were targeted rather than exhaustive, and absence from them is not evidence of absence from the literature.

Bayesian diagnosis over signature matrices. Residual-based Bayesian network methods already combine signature-matrix structure with posterior inference over observed zero and non-zero residuals [12]. The apparatus of ranking hypotheses probabilistically against a structural signature is therefore established, and scoring an observed indicator is ordinary Bayesian practice. What this paper offers is the explicit specialisation to validity-gated runtime diagnosis and its connection to structural signatures, robust decision effects and smoothed ranking.

Informative censoring. Gate conditioning is missingness that is not at random: the missingness indicator depends on the latent quantity of interest. Little and Rubin [6] classify the mechanisms and Manski [7] develops partial identification. The cited work establishes the broader mechanism; this paper contributes a runtime-diagnosis specialization and an accompanying audit framework.

Signature matrices and isolability. The column-equality criterion for structural isolability is due to this community [8, 9], with sensor placement [10] and multi-mode signatures [11]. We take the apparatus unchanged and evaluate it on the active channel set.

Runtime monitoring of perception and autonomy stacks. Antonante et al. [13] treat diagnostic graphs and diagnosability for autonomy monitors. Graph-structured runtime monitoring is theirs and others'.

3 Formal setup

Fix one decision epoch. Let $\mathcal{H} = \{H_0, \dots, H_{m-1}\}$ be finite and mutually exclusive with H_0 nominal and priors $\pi_h > 0$.

Assumption 1 (Binary gate)**.** The validity gate is binary, $Z \in \{0, 1\}$, with z the observed value. This matches the runtime setting, in which a predicate either holds or fails, and it is assumed throughout.

Let $\mathcal{G}$ be the channels deactivated at $Z = z$ and $A(z)$ the active set, with e_A the complete active-evidence vector. Let C denote the operating context, which may include operating domain, flight phase, environment, sensor geometry, estimator mode and mission regime. Define the *joint active-evidence likelihood*

$$L_A(h; z, C) \coloneqq p(e_A \mid Z = z,\, H = h,\, C) > 0, \qquad (1)$$

conditioned on the observed gate value. We do **not** assume that the active channels factorise, nor that e_A is independent of Z given H. Write $\pi_h(C)$ for the prior and

$$w_h(C) \coloneqq p(Z = z \mid H = h,\, C) > 0. \qquad (2)$$

Positivity of L_A and w_h is assumed throughout so that all log terms are finite. Since C is fixed within an epoch we suppress it below, but it should not be read as a global constant: gate likelihoods are context-dependent and a value estimated in one operating domain need not transfer to another. Section 8 exhibits two domains with materially different w.

Context conditioning, which is common practice, forms

$$q^{\text{omit}}(h \mid e_A, z, C) \;\propto\; \pi_h(C)\, L_A(h; z, C), \qquad (3)$$

and the *gate-aware posterior*, which is the correct Bayesian posterior for the full observation $(e_A, Z = z)$, is

$$p(h \mid e_A, z, C) \;\propto\; \pi_h(C)\, w_h(C)\, L_A(h; z, C). \qquad (4)$$

We write q^{omit} for (3) and p for (4). The first is a scoring rule that omits an available observation. It is not the Bayesian posterior for the full observation under the declared generative model, since the monitor has in fact observed $Z = z$ and used that fact to select $A(z)$. Calling it a posterior would misdescribe it. Equivalently, in log-score form, $\ell_h^{\text{ev}} = \ell_h^{\text{ctx}} + \log w_h$ with $\ell_h^{\text{ctx}} = \log \pi_h + \log L_A(h; z, C)$. The two differ only by the gate factor.

Definition 1 (Diagnostic exogeneity)**.** Z is *diagnostically exogenous* for $\mathcal{H}$ in context C if $Z \perp \mathcal{H} \mid C$.

Lemma 1 (Equivalence under Assumption 1)**.** *For binary Z, exogeneity holds if and only if w_h is constant in h.*

Proof. Constancy of $p(Z = z \mid h)$ forces constancy of $p(Z = 1 - z \mid h) = 1 - p(Z = z \mid h)$, which is conditional independence; the converse is immediate. □

Remark 1 (Why binary matters). The equivalence is genuinely specific to the binary case. For multivalued Z, constancy of $p(Z = z \mid h)$ at the single observed value z does not imply independence: with $\mathcal{H} = \{h_0, h_1\}$ and $Z \in \{0, 1, 2\}$, the rows $(0.2, 0.5, 0.3)$ and $(0.2, 0.1, 0.7)$ agree at $z = 0$ while $Z \not\perp \mathcal{H}$. All results below use only w_h at the observed z and therefore hold verbatim for multivalued Z; what requires Assumption 1 is Lemma 1, and hence the reading of Corollary 1 as a statement about conditional independence rather than about the single event $\{Z = z\}$.

Remark 2 (Two distinct gate quantities). Two separate quantities are attached to the gate and must not be conflated. $w_h = p(Z = z \mid H = h, C)$ is a *probabilistic* emission likelihood entering (5). $S_{Z,h} \in \{0, 1\}$ below is a *structural* signature entry recording whether h predicts the observed gate state. They are logically independent: a gate may be strongly endogenous, with large R and large λ_{ab}, while $S_{Z,a} = S_{Z,b}$ for the pair of interest, in which case the correction shifts the posterior but the gate row adds no structural separation; conversely $S_{Z,a} \neq S_{Z,b}$ is possible with λ_{ab} small. Sections 4 and 5 concern w; Section 6 concerns S_Z only. No result here derives one from the other.

Definition 2 (Signature and partition)**.** $S_{j,h}(C) \in \{0, 1\}$ records whether h predicts deviation on channel j. Write $h_a \sim^{\text{ctx}} h_b$ when signature columns agree on $A(z)$, and $h_a \sim^{\text{ev}} h_b$ when they agree on $A(z) \cup \{Z\}$, with induced partitions Π^{ctx} and Π^{ev}.

4 The exact gate tilt

Proposition 1 (Tilt identity)**.** *For every h, $\ell_h^{\text{ev}} - \ell_h^{\text{ctx}} = \log w_h$, hence*

$$p(h) = \frac{q^{\text{omit}}(h)\, w_h}{\mathbb{E}_{q^{\text{omit}}}[w]}, \qquad (5)$$

and for any pair,

$$\log \frac{p(h_a)}{p(h_b)} = \log \frac{q^{\text{omit}}(h_a)}{q^{\text{omit}}(h_b)} + \lambda_{ab}, \quad \lambda_{ab} = \log \frac{w_a}{w_b}. \qquad (6)$$

Proof. By (3) and (4) the two unnormalised scores differ by the single factor w_h, since $\pi_h(C)$ and $L_A(h; z, C)$ appear identically in both. Normalising and dividing numerator and denominator by $\sum_{h'} \pi_{h'} L_A(h'; z, C)$ gives (5). Taking the ratio for a pair cancels $\mathbb{E}_{q^{\text{omit}}}[w]$. □

Remark 3 (The identity does not assume gate-residual independence). $L_A(h; z, C)$ is arbitrary and may depend on z in any way. It appears in both (3) and (4) and therefore cancels exactly in (5), so Proposition 1 holds with no assumption that the active evidence is independent of the gate given the hypothesis. This was verified symbolically and, as a check, numerically over 2×10^5 random (π, L_A, w) draws with maximum deviation 5.6×10^{-16}.

Distinct modelling errors must be separated. The first concerns the score; the remainder are species of a single

genus, *misspecification of the active-evidence likelihood*, differing in which latent variable is mishandled.

Error A, omitted gate evidence. The monitor observes $Z = z$ but excludes $w_h = p(Z = z \mid H = h, C)$ from the hypothesis score. This is exactly the error quantified in this paper.

Error B, misspecification with respect to the gate. The monitor uses $p(e_A \mid H = h, C)$ where the correct conditional model is $p(e_A \mid Z = z, H = h, C)$. Residual magnitude and gate state may both depend on signal-to-noise ratio, geometry, estimator health, dynamics or regime even after conditioning on the hypothesis.

Error C, misspecification with respect to source identity. Every result above presumes each active observation is correctly attributed to its declared source. An implementation that selects among physical sources, for example through a multiplexer, can produce a measurement that is numerically valid, physically real, in range, and fresh, while originating from a source other than the declared one. This violates the attribution assumption without producing a missing value, an invalid value, or an implausible one, so no gate and no plausibility check need fire. Remark 20 gives the corresponding reweighting.

Errors B and C are siblings rather than instances of one another: B misspecifies the conditional law given the *gate* variable Z, C misspecifies it given the latent *source* variable S. **This paper addresses neither.** All results isolate Error A relative to a baseline (3) that already models the active evidence correctly, conditional on the observed gate and context and under correct source attribution. A monitor committing B or C as well incurs further distortion not quantified here.

Lemma 2 (Scale invariance)**.** *For any $c > 0$, tilting q^{omit} by w and by cw gives the same p, and λ_{ab} and R are unchanged. Consequently any positive finite weight vector may be rescaled by $c \le 1/\max_h w_h$ to obtain admissible binary-gate event probabilities $w_h \in (0, 1]$ without altering any quantity in this paper.*

Proof. The factor c cancels between numerator and $\mathbb{E}_{q^{\text{omit}}}[cw]$ in (5), and it cancels in every ratio w_a/w_b, hence also in $R = \max_h \log w_h - \min_h \log w_h$. □

Remark 4. All numerical constructions below are stated with $w_h \in (0, 1]$, as Lemma 2 permits; where a construction is more readable in unnormalised form we give the rescaled probabilities alongside.

Corollary 1 (Exogeneity is exactly the null correction)**.** *$\lambda_{ab} = 0$ for every pair if and only if w_h is constant in h, in which case $p = q^{\text{omit}}$. Under Assumption 1 this is exactly exogeneity in the sense of Definition 1, by Lemma 1.*

Proof. If w is constant then $\lambda_{ab} = 0$ and (5) gives $p = q^{\text{omit}}$. Conversely $\lambda_{ab} = 0$ for all pairs forces $w_a = w_b$ throughout. □

Corollary 1 is the licensing condition for current practice: context conditioning is exactly correct under exogeneity and wrong by exactly λ_{ab} otherwise. Gates guarding a sensor against the conditions that sensor's own faults produce can be endogenous.

Algorithm 1: Gate-aware scoring for one epoch. Omitting the third term in the loop recovers q^{omit}.

Input: π, C, z, e_A; models $w_h(C)$ and $L_A(h; z, C)$
Output: p, $\{\lambda_{ab}\}$, R
foreach *hypothesis* h **do**
  $\ell_h \leftarrow \log \pi_h(C) + \log L_A(h; z, C) + \log w_h(C)$
end
$p(h) \leftarrow \exp(\ell_h) \,/\, \sum_{h'} \exp(\ell_{h'})$
$R \leftarrow \max_h \log w_h - \min_h \log w_h$
foreach *pair* (a, b) **do**
  $\lambda_{ab} \leftarrow \log w_a - \log w_b$
end

Table 1. Audit output for the worked configuration of Section 8, gate Z = receiver validity flag, $z = 0$. Pre-gate margin $\Delta = \log(0.86/0.02) = 3.7612$ nats for every fault against nominal. For this worked configuration, which has no active residual evidence, a nominal-versus-fault pair is decision-critical when the gate correction alone exceeds the pre-gate pairwise margin, $\lambda_{h0} > \Delta$. Emission probabilities are model assumptions.

	Open (R=3.807)		Denied (R=3.401)	
h vs nominal	λ_{h0}	$> \Delta$	λ_{h0}	$> \Delta$
degraded	+3.807	yes	+0.493	no
interference	+3.807	yes	+0.493	no
spoof	+0.405	no	−2.909	no
fault 4	+3.401	no	+0.087	no
faults 5–7	0	no	0	no
structural $S_{Z,h} \neq S_{Z,0}$	declared per hypothesis; not implied by $\lambda_{h0} \neq 0$ (Rem. 2)			

Diagnostic gate audit. The results above support a short practitioner-facing check, per validity gate: (i) identify which hypotheses change the distribution of the gate; (ii) model the active evidence conditional on the observed gate and operating context, per (1); (iii) include the gate likelihood in hypothesis ranking when it is informative (Proposition 1, Corollary 1); (iv) include the gate state in structural isolability analysis (Proposition 5); (v) for multiple gates use the joint likelihood unless conditional factorisation is justified (Proposition 12, Appendix C); (vi) compare the gate correction against the prior log-odds barrier before assuming it can change the reported diagnosis (Section 8). Each item is supported by a result in this paper and by nothing beyond them. Gate-emission models may be estimated or bounded using fault injection, hardware-in-the-loop testing, calibrated simulation, labeled operational data, analytical fault models, or structured expert elicitation; the correction is conditional on the adequacy of that model.

Table 1 shows what the audit produces on the worked configuration. In open sky the gate is decision-critical for exactly two of seven faults, and by a margin of only 0.045 nats; in the denied domain it is decision-critical for none. The same gate, the same threshold, and the same priors

give opposite audit verdicts in the two contexts, which is the practical content of writing $w_h(C)$ rather than w_h.

Remark 5 (Gate-model uncertainty maps directly to the correction). The gate emission probabilities are modelled quantities. If estimates $\hat{w}_a, \hat{w}_b$ are used in place of the true w_a, w_b, the error in the gate contribution to the pairwise log odds is exactly $\hat{\lambda}_{ab} - \lambda_{ab} = \log(\hat{w}_a/\hat{w}_b) - \log(w_a/w_b)$, since by (6) the gate enters additively. If only positive intervals $w_a \in [\underline{w}_a, \overline{w}_a]$ and $w_b \in [\underline{w}_b, \overline{w}_b]$ are known, then by monotonicity

$$\log \frac{\underline{w}_a}{\overline{w}_b} \;\le\; \lambda_{ab} \;\le\; \log \frac{\overline{w}_a}{\underline{w}_b}, \tag{7}$$

an interval of width $\log(\overline{w}_a/\underline{w}_a)+\log(\overline{w}_b/\underline{w}_b)$. Uncertainty in the gate-emission model therefore maps directly, and additively, into uncertainty in the log-odds correction. For example $w_a \in [0.80, 0.95]$ and $w_b \in [0.01, 0.05]$ give $\lambda_{ab} \in [2.773, 4.554]$, a width of 1.781 nats. No robust-inference procedure is proposed here; this is interval arithmetic on (6).

5 Divergence characterisation

An observation whose log-likelihood ratio is confined to a range of width R can move a distribution only so far. This section gives the exact limit.

Let $R \coloneqq \max_h \log w_h - \min_h \log w_h$. Then $R = 0$ iff w_h is constant in h, which under Assumption 1 is exogeneity in the sense of Definition 1 by Lemma 1; for a multivalued gate it is only constancy at the observed value z.

Proposition 2 (Identity and range bound)**.**

$$\mathrm{KL}(p\|q^{\mathrm{omit}}) = \mathop{\mathbb{E}}_{p}[\log w] - \log \mathop{\mathbb{E}}_{q^{\mathrm{omit}}}[w] \le R.$$

Proof. By (5), $\log(p/q^{\mathrm{omit}}) = \log w - \log \mathbb{E}_{q^{\mathrm{omit}}}[w]$; take expectation under p. For the bound, $\mathbb{E}_p[\log w] \le \max_h \log w_h$ and $\log \mathbb{E}_{q^{\mathrm{omit}}}[w] \ge \min_h \log w_h$. □

Let $\Lambda(t) = \log \mathbb{E}_{q^{\mathrm{omit}}}[w^t]$ be the cumulant generating function of $X = \log w$ under q^{omit}.

Lemma 3. *$\Lambda(0) = 0$, $\Lambda'(t) = \mathbb{E}_t[X]$ and $\Lambda''(t) = \mathrm{Var}_t(X)$ under the tilt $p_t \propto q^{\mathrm{omit}} w^t$; $p_1 = p$ and $\mathrm{KL}(p\|q^{\mathrm{omit}}) = \Lambda'(1) - \Lambda(1)$.*

Proposition 3 (Sharp range-only bound)**.** *For $R > 0$ let $A_R \coloneqq R/(1 - e^{-R})$, extended continuously by $A_0 \coloneqq 1$ so that the bound reads 0 at $R = 0$, consistent with $\mathrm{KL} = 0$ under exogeneity. Then*

$$\mathrm{KL}(p\|q^{\mathrm{omit}}) \;\le\; A_R - 1 - \log A_R, \tag{8}$$

and the bound is attained. The legacy bounds $\mathrm{KL} \le R$ and $\mathrm{KL} \le R^2/8$ remain valid, and their minimum is strictly looser than (8) for every $R > 0$.

Extremizer. *Equality holds at the two-point configuration with likelihood ratios $b_{\min} = 0.256860$ and $b_{\max} = 11.558678$ at $R = 3.806662$, carried by masses $q^{\mathrm{omit}} = (0.934246, 0.065754)$, which satisfy the normalisation $\mathbb{E}_{q^{\mathrm{omit}}}[w] = 1$; in admissible probability form, by Lemma 2 with $c = 1/b_{\max}$, this is $w = (0.022222, 1)$. Direct evaluation gives $\mathrm{KL} = 1.533952$, matching (8).*

Proof. By (5) the likelihood ratio $\mathrm{d}p/\mathrm{d}q^{\mathrm{omit}} = w/\mathbb{E}_{q^{\mathrm{omit}}}[w]$ lies in $[b_{\min}, b_{\max}]$ with $b_{\max}/b_{\min} = e^R$ and, since its mean under q^{omit} is one, $b_{\min} \le 1 \le b_{\max}$. Rukhin's theorem [5] bounds any f-divergence with f bowl-shaped and minimised at 1 by the two-point interpolation

$$D_f \;\le\; \frac{b_{\max} - 1}{b_{\max} - b_{\min}} f(b_{\min}) + \frac{1 - b_{\min}}{b_{\max} - b_{\min}} f(b_{\max}),$$

which is tight. Taking $f(t) = t \log t$ and maximising the right-hand side over $b_{\min} \in (0, 1]$ subject to $b_{\max} = b_{\min} e^R \ge 1$ gives (8). For the quadratic form, $X = \log w$ lies in an interval of width R under every exponential tilt p_t, so $\Lambda''(t) = \mathrm{Var}_t(X) \le R^2/4$ by Popoviciu's inequality, and $\Lambda'(1) - \Lambda(1) = \int_0^1 \int_s^1 \Lambda''(u)\, du\, ds \le R^2/8$. □

Proposition 4 (Total variation, sharp)**.**

$$\mathrm{TV}(p, q^{\mathrm{omit}}) \;\le\; \tanh(R/4), \tag{9}$$

and the bound is attained, at $q^{\mathrm{omit}} = (\sigma(R/2), 1 - \sigma(R/2))$ with $\log w = (0, R)$, where σ is the logistic function.

Proof. Two routes give the same constant. Directly: let $A = \{h : p(h) > q^{\mathrm{omit}}(h)\}$ and merge A and its complement into single atoms. Total variation equals the positive part $p(A) - q^{\mathrm{omit}}(A)$ and is therefore preserved exactly by this particular merge; the merged gate values are q^{omit}-conditional means of w and so lie in $[\min_h w_h, \max_h w_h]$, whence the merged range is at most R. This reduces the problem to two atoms, $q^{\mathrm{omit}} = (a, 1 - a)$ with $w = (1, t)$, $t = e^R$, for which $\mathrm{TV} = a(1-a)(t-1)/(t - a(t-1))$. Setting the derivative in a to zero gives $a^2(t-1) - 2at + t = 0$, whose root in $(0, 1)$ is $a^\star = \sqrt{t}/(\sqrt{t}+1) = \sigma(R/2)$; substituting, $t - a^\star(t-1) = \sqrt{t}$ and $\mathrm{TV}(a^\star) = (\sqrt{t}-1)/(\sqrt{t}+1) = \tanh(R/4)$. Since $\tanh(R/4)$ is increasing in R, a merged range below R only lowers the value. Alternatively, (9) is Rukhin's bound with $f(t) = |t - 1|/2$, maximised over admissible $(b_{\min}, b_{\max})$. □

Remark 6 (Both bounds are instances of one 1997 theorem). Propositions 3 and 4 are the Kullback-Leibler and total-variation cases of Rukhin's bounded-likelihood-ratio f-divergence theorem [5], specialised to the range constraint that the gate imposes. We verified this numerically: maximising Rukhin's expression over admissible $(b_{\min}, b_{\max})$ with $b_{\max}/b_{\min} = e^R$ reproduces $A_R - 1 - \log A_R$ and $\tanh(R/4)$ to six decimals at $R \in \{0.5, 1, 2, 3.806662, 8\}$. They are stated here because they are sharp, because the weaker alternatives commonly reached for are misleading, and because a practitioner needs the constant in terms of R. At $R = 3.806662$, (8) gives 1.533952 against $R^2/8 = 1.811335$, and (9) gives 0.740536 against $R/4 = 0.951666$ and $1 - e^{-R} = 0.977778$; the exact values in that configuration are $\mathrm{KL} = 1.449211$ and $\mathrm{TV} = 0.656418$.

6 Structural consequence

Proposition 5 (Refinement)**.** Π^{ev} *refines* Π^{ctx}*, strictly if and only if some pair satisfies* $h_a \sim^{\mathrm{ctx}} h_b$ *with* $S_{Z,a} \neq S_{Z,b}$*.*

Proof. The gate-aware column is $\sigma_h^{\mathrm{ev}} = (\sigma_h^{\mathrm{ctx}}, S_{Z,h})$, so equality of ev-columns implies equality of ctx-columns. Hence $\sim^{\mathrm{ev}} \subseteq \sim^{\mathrm{ctx}}$, which is refinement. Strictness requires a pair in the difference, which is the stated condition. □

Remark 7 (This is the general coordinate fact, and a distinction that matters). Proposition 5 is the elementary observation that appending a coordinate *uniformly to every column* refines an equality partition.

It must not be confused with the superficially similar operation of deleting signature entries *per hypothesis*, as arises when one models a fault silencing its own evidence. That operation is column-dependent and does **not** refine: two hypotheses with identical columns can separate when only one is silenced. The smallest explicit witness has two hypotheses and one channel. Let $S_{1,a} = S_{1,b} = 1$, so the columns are equal and the pair is merged. Silence only h_a's entry, giving $S_{1,a}^{\mathrm{eff}} = 0$ and $S_{1,b}^{\mathrm{eff}} = 1$: the columns are now unequal and the pair separates. Column-dependent deletion has refined the partition. No uniform appending can do this, since appending a common coordinate to equal columns leaves them equal. A randomised check over 2×10^5 draws found 0 violations of Proposition 5 and 110,819 violations of the corresponding statement for column-dependent deletion. Any argument by entrywise domination is therefore invalid for the latter and is not used here.

Corollary 2 (One-directional interpretation)**.** *Treating an endogenous gate as pure context can only coarsen the identifiability partition. Context conditioning may report ambiguity that does not exist and can never report separability that does not exist.*

Proposition 6 (Nominal absorption, existential)**.** H_0 *shares a* Π^{ctx} *block with a fault* h *if and only if* $S_{j,0} = S_{j,h}$ *for every* $j \in A(z)$*. In particular, if* $\mathcal{G} \supseteq \{j : S_{j,0} \neq S_{j,h}\}$ *then absorption occurs, and admitting* Z *separates the pair if and only if* $S_{Z,0} \neq S_{Z,h}$*.*

Proof. Direct from Definition 2 and Proposition 5. □

Remark 8 (Scope). Proposition 6 is conditional. It states when absorption occurs, not that endogenous gating implies it. A gate may be strongly endogenous, with large R and large λ, and cause no coarsening at all if the channels it removes carry no distinguishing power. Endogeneity strength and structural severity are independent and both must be reported.

Remark 9 (Asymmetric safety consequence). Coarsening is conservative for isolation authority, since a monitor reporting a class has declined to commit. It is anti-conservative for detection when the class contains H_0, because a monitor resolving the class by prior mass, and H_0 usually carries the dominant prior, reports nominal during a fault.

7 When the gate can change the reported diagnosis

An observation can move a distribution a great deal without changing which hypothesis leads, and can change the leader while moving it hardly at all. Divergence measures how far the two distributions are; it does not say whether the reported hypothesis changes. This section gives the exact decision condition. Every statement is a rearrangement of (6); they are collected because the decision question, not the divergence question, is what a monitor designer asks.

Let $h^\star$ be a maximiser of q^{omit} and write $\Delta := \min_{j \neq h^\star} \log\left(q^{\mathrm{omit}}(h^\star)/q^{\mathrm{omit}}(j)\right)$ for the pre-gate decision margin, so $\Delta > 0$ exactly when $h^\star$ is the unique maximiser.

Proposition 7 (Exact decision condition)**.** $h^\star$ *remains a maximiser of* p *if and only if*

$$\log \frac{q^{\mathrm{omit}}(h^\star)}{q^{\mathrm{omit}}(j)} \;\geq\; \log \frac{w_j}{w_{h^\star}} \qquad \text{for every } j \neq h^\star, \tag{10}$$

and remains the unique *maximiser if and only if every inequality is strict. If* q^{omit} *has several maximisers,* (10) *applied to each determines which survive.*

Proof. Immediate from (6): the post-gate pairwise log odds is the pre-gate log odds plus $\lambda_{h^\star j} = \log(w_{h^\star}/w_j)$, and it is non-negative exactly under (10). □

Corollary 3 (Margin versus range)**.** *If* $R < \Delta$ *then* $h^\star$ *remains the unique maximiser under every gate likelihood with range at most* R*. If* $R = \Delta$ *a tie is attainable but no strict reversal is possible. If* $R > \Delta$ *there is a positive* w *with range* R *under which* $h^\star$ *is not a maximiser.*

Proof. $\log(w_j/w_{h^\star}) \leq R$ always, so $R < \Delta$ makes every inequality in (10) strict. At $R = \Delta$ equality is achieved by $w_{h^\star} = \min_h w_h$ and $w_{j_0} = \max_h w_h$ at the nearest competitor j_0. For $R > \Delta$ set $w_{j_0} = e^R$ and $w_h = 1$ otherwise, giving post-gate log odds $\Delta - R < 0$. □

Remark 10 (This is argmax stability, and we say so). Corollary 3 is the standard statement that an argmax survives an additive perturbation of the scores whose range is below the decision margin, instantiated with the perturbation $\log w$. It is a direct corollary of Proposition 7. We name no new quantity for Δ.

Remark 11 (Distortion does not determine decision change). The two questions are incomparable in both directions, and small examples settle it.

Arbitrarily small distortion, changed decision. Take $q^{\mathrm{omit}} = (\frac{1}{2} + \delta, \frac{1}{2} - \delta)$ and $\log w = (0,\, 2.05 \log \frac{1/2+\delta}{1/2-\delta})$. The maximiser flips for every $\delta > 0$, while at $\delta = 10^{-4}$ we have $\mathrm{TV} = 2.05 \times 10^{-4}$ and $\mathrm{KL} = 8.41 \times 10^{-8}$, both $\to 0$.

Arbitrarily large distortion, unchanged decision. Take $q^{\mathrm{omit}} = (1-\varepsilon, \varepsilon)$ and $w = (1, (1-\varepsilon)/(2\varepsilon))$. The maximiser never changes, while $\mathrm{KL} = 0.905,\ 2.434,\ 5.504,\ 10.109$ at $\varepsilon = 10^{-2}, 10^{-4}, 10^{-8}, 10^{-14}$.

Divergence must therefore be compared against the pre-gate margin Δ, not read as a decision effect on its own. This is the practical reason Section 5 alone does not answer the designer's question.

8 Reproducible numerical example

This section verifies the algebra on one configuration. It is **verification, not validation**: it establishes that the identities and bounds evaluate as stated, and nothing about operational frequency.

Eight mutually exclusive hypotheses with $\pi_{\text{nominal}} = 0.86$ and seven fault priors of 0.02; $\alpha_s = 0.06$, $\eta = 1.5$, $f = 25\,\text{Hz}$. The gate is a receiver validity flag with $z = 0$ meaning validity lost.

Table 2. Gate emission probabilities $p(Z = 0 \mid H)$. **These are model assumptions, not measured quantities.** They were not estimated from any flight corpus.

Domain	nominal	degraded	interference	spoof
Open sky	0.02	0.90	0.90	0.03
Denied	0.55	0.90	0.90	0.03

The remaining four hypotheses take 0.60, 0.02, 0.02, 0.02 in the open-sky row and 0.60, 0.55, 0.55, 0.55 in the denied row. A spoof that maintains a plausible fix does not drop the flag, which is why the spoof column is low in both; this is a property of the modelled hypothesis, not of spoofing in general.

In this configuration the gate deactivates every GNSS-derived channel and no other channel bears on the hypotheses, so there is no active residual evidence in the epoch. Consequently $L_A(h; z, C)$ is constant in h and cancels identically, leaving q^{omit} equal to the prior. The example therefore isolates the gate factor and does not depend on any modelling of the active evidence.

Table 3. Divergence quantities, with q^{omit} equal to the prior in an epoch with no active residual evidence. Sharp bounds are Propositions 3 and 4; $R^2/8$ is shown as a legacy comparison. All bounds hold in both rows.

		Kullback-Leibler			Total variation	
Domain	R	sharp	$R^2/8$	actual	sharp	actual
Open sky	3.807	1.534	1.811	1.449	0.741	0.656
Denied	3.401	1.260	1.446	0.022	0.691	0.027

The contrast is the point. In open sky, losing validity is strong evidence against nominal and $p(H_0)$ falls from 0.860 to 0.257. In a denied domain the same event carries almost nothing, KL = 0.022. The ranges are similar, 3.807 against 3.401, which shows that R is a worst-case bound over pairs rather than a summary of realised divergence.

Transient. The contraction factor is 0.94 per tick. The exact discrete e-folding time is $\tau = -1/\log(1 - \alpha_s) = 16.162$ ticks (0.646 s); the familiar $1/\alpha_s = 16.667$ ticks (0.667 s) is the small-α_s approximation to it. Continuous settling indices to 90, 95 and 99 percent are 37.213, 48.416 and 74.427; since updates occur at discrete ticks, the first ticks meeting those tolerances are 38, 49 and 75, that is 1.52 s, 1.96 s and 3.00 s.

The prior dominates the gate. For the interference hypothesis against nominal in open sky, $\delta\ell_0^{\text{ctx}} = \log(0.02/0.86) = -3.761$ and $\lambda = \log(0.90/0.02) = +3.807$, a likelihood ratio of 45 : 1. The gate-aware steady state is $D_\infty = +0.045462$, and at $\eta = 1.5$ the reported log odds are 0.068194, an odds ratio of 1.070573. The upward case of Proposition 10 applies, since $\eta D_0 = -5.6418 < L^\star = 0 < \eta D_\infty = 0.068194$. Equation (14) gives $n^\star = 71.557$ ticks, a continuous crossing time of 2.862 s; the first tick at which the implementation first satisfies the target is $n_{\text{hit}} = 72$, that is $T_{\text{hit}} = 2.880\,\text{s}$, which we confirmed by direct simulation of (12). For any target above an odds ratio of 1.070573, case 3 applies and the crossing time is infinite.

A gate likelihood ratio of 45 to 1, against these priors, leaves this pair close to indifference. It is important to state precisely what does and does not move. The gate substantially changes the *posterior distribution*: $p(H_0)$ falls from the prior 0.860 to 0.256716 in the open-sky row. What it barely changes is the *pairwise ordering* of interference against nominal, whose tempered steady-state odds reach only 1.070573. A correction of 3.807 nats against a prior barrier of 3.761 nats leaves a margin of 0.045 nats, so the pair reverses but only just. Admitting the gate therefore reshapes the posterior materially while leaving this particular pairwise ranking near indifference. The general condition is

$$\log \frac{\pi_0}{\pi_h} < \lambda_{h0} - \frac{L^\star}{\eta}, \tag{11}$$

which at $L^\star = 0$ reads $\pi_0/\pi_h < w_h/w_0$. Here $\pi_0/\pi_h = 43$ against $w_h/w_0 = 45$, a margin of about five percent, which is exactly why the pairwise reversal is so slight. This is an exact numerical illustration of the algebra, not an empirical finding.

Remark 12 (Design consequence). Priors chosen by engineering judgement silently set the evidential barrier every gate and residual must clear. A nominal prior of 0.86 against seven faults at 0.02 imposes $\log 43 = 3.761$ nats on any single piece of evidence attempting to overturn nominal. We recommend reporting priors alongside the barrier they imply and checking (11) per gate at design time.

9 Scope and limitations

1. **Priority.** Diagnosis under fault-dependent observation loss is well established in the discrete-event-systems literature [1, 2, 3, 4]. Informative censoring is [6, 7]. The signature-matrix criterion is [8, 9]. We claim priority for none.

2. **Component results are established.** Propositions 1, 9 and 10 are likelihood reweighting and affine recursion; Proposition 3 is the sharp bounded-likelihood-ratio specialisation of Rukhin's theorem [5],

with Hoeffding's lemma giving only the weaker $R^2/8$ comparison; Proposition 5 is the general coordinate fact of Remark 7. Appendix F records the status of each.

3. **Single epoch and fixed gate.** Sections 4 to 6 treat one epoch at a fixed gate value and Appendix B assumes the gate holds. Gates switching during accumulation are not covered.

4. **Structural, not statistical.** Binary signatures express structural isolability only. A structurally merged pair may separate on magnitude or dispersion.

5. **Emission probabilities are assumptions.** Table 2 is modelling, not measurement. No quantity in this paper was estimated from flight data, and the qualitative conclusions depend on the ordering of w_h rather than its values.

6. **Source attribution is assumed, not established.** Every result presumes each active observation is correctly attributed to its declared source (Error C, Remarks 3 and 20). Correct readback of a selection register can expose some selection errors, but neither a readback nor a plausible in-range value alone establishes physical source identity, and an attribution taken as given is not thereby verified. Nothing in this paper detects, bounds, or corrects a source-attribution error; (21) only names the reweighting such a correction would require, and that reweighting needs an attribution model the implementation may not have.

7. **Companion systems paper.** A companion measurement study reports implementation-level characterisation of this mechanism in an implemented runtime diagnostic layer. It makes no claim of field deployment or certified flight use. It likewise treats fault-dependent observation loss as established prior art [1] and claims only measurement. Neither paper claims discovery of fault-dependent observation loss, of signature matrices, of informative censoring, or of the coordinate-refinement fact.

8. **Operational frequency unmeasured.** Nothing here says how often endogenous gates close in operation. A phenomenon that is structurally guaranteed but rare is a different engineering problem from one that is common.

9. **Not established here.** We do not establish, and this paper must not be read as implying: any universal statement that faults silence their own evidence; any irreparability result quantifying over all possible additional evidence; partition monotonicity under fault-dependent or column-dependent masks (see Remark 7); novelty for coherent-system or minimal-cut-set methods applied to diagnosis; or novelty for Kullback-Leibler diagnosability measures or distributionally robust fault detection.

10 Conclusion

A validity gate outcome is a diagnostic observation with a hypothesis-dependent likelihood ratio, and a monitor that uses it only to choose which channels to score has thrown that observation away. Restored, the reading tilts the posterior by the gate likelihood, shifts every pairwise log odds by the gate's log-likelihood ratio for that pair, and vanishes exactly when that ratio is one for every pair. An observation with log-likelihood range R can move the distribution at most $A_R - 1 - \log A_R$ in Kullback-Leibler and $\tanh(R/4)$ in total variation, both attained; whether it changes the reported hypothesis is a separate question, settled by comparing R against the pre-gate decision margin. Admitting the gate as a signature coordinate can only refine the declared partition, by an argument that uses no probability at all. On the configuration examined, a $45:1$ gate against a $43:1$ prior reshapes the posterior substantially while reversing the ranking it targets by 0.045 nats, which is the general point in one number. The component mathematics is established; the contribution is the account.

A Interval certificates and decision stability under gate-model uncertainty

Proposition 8 (Interval certificate)**.** *Suppose only $w_h \in [l_h, u_h]$ is known with $0 < l_h \le u_h$, and that these constraints are independent, so the admissible set is the product box. Then $h^\star$ maximises p for* every *admissible w if and only if $q^{\mathrm{omit}}(h^\star)\, l_{h^\star} \ge \max_{j \ne h^\star} q^{\mathrm{omit}}(j)\, u_j$, and for* no *admissible w if and only if $q^{\mathrm{omit}}(h^\star)\, u_{h^\star} < \max_{j \ne h^\star} q^{\mathrm{omit}}(j)\, l_j$. Otherwise the outcome is model-dependent.*

Proof. Sufficiency of each condition is immediate. For necessity, the box structure lets each coordinate be set independently: taking $w_{h^\star} = l_{h^\star}$ and $w_j = u_j$ realises the worst case for $h^\star$, and $w_{h^\star} = u_{h^\star}$ with $w_j = l_j$ realises the best. **Necessity fails without the product structure**: if the coordinates are coupled, those corner assignments need not be admissible and the trichotomy gives only sufficient conditions. □

B Smoothed and tempered implementations

Deployed monitors often smooth and temper rather than form a posterior:

$$z_{h,k} = (1 - \alpha_s) z_{h,k-1} + \alpha_s \ell_{h,k}, \quad p_k(h) = \frac{e^{\eta z_{h,k}}}{\sum_{h'} e^{\eta z_{h',k}}}. \tag{12}$$

Proposition 9 (Steady state)**.** *If the gate holds at $Z = z$ and scores are constant at ℓ^{ev}, then $\log \frac{p_k(h_a)}{p_k(h_b)} \to \eta(\delta \ell^{\mathrm{ctx}}_{ab} + \lambda_{ab})$ geometrically with factor $(1 - \alpha_s)$.*

Proof. $D_k = z_{a,k} - z_{b,k}$ obeys an affine contraction with factor $(1-\alpha_s)$ and fixed point $\delta\ell^{\mathrm{ev}}_{ab}$; the tempered log odds are ηD_k. □

Remark 13 (The smoothed output is not generally a posterior). Two distinct departures from Bayesian conditioning are present in (12) and must not be conflated.

At constant-score steady state. If $z_{h,k} \to \ell_h$ and ℓ_h is the ordinary Bayes log score up to a common normaliser, then $p_\infty(h) \propto \exp(\eta\ell_h) \propto [p_{\mathrm{Bayes}}(h)]^\eta$, a normalised power transform of the corresponding Bayes posterior. We avoid the term "power posterior" because in the Bayesian literature it denotes a posterior formed by raising the *likelihood* to a power, a related but distinct construction; here the exponent is applied to the normalised posterior itself. In this special setting $\eta = 1$ recovers the corresponding Bayes posterior, and $\eta > 1$ over-weights the gate evidence and equally over-weights the prior.

During transients. Away from steady state, $z_{h,k}$ is an exponentially weighted average of the score history rather than the current epoch's score. Consequently $\mathrm{softmax}(z_k)$ is **not** in general the Bayes posterior for the current epoch even when $\eta = 1$.

Nothing in this construction licenses interpreting p_k as a calibrated probability of fault without a separate calibration argument.

Proposition 10 (Upward threshold crossing time)**.** *Let $D_k = z_{a,k} - z_{b,k}$, let $D_\infty = \delta\ell^{\mathrm{ev}}_{ab}$ be the fixed point under constant scores, and let D_0 be the initial value. Then*

$$D_k = D_\infty + (1-\alpha_s)^k (D_0 - D_\infty). \tag{13}$$

Assume the upward *case $D_\infty > D_0$, and let $L^\star$ be a target in tempered log odds. Exactly one of the following holds.*

1. *If $L^\star \le \eta D_0$, the target is already satisfied at $k = 0$; this includes the boundary $L^\star = \eta D_0$.*
2. *If $\eta D_0 < L^\star < \eta D_\infty$, the real-valued crossing index is*

$$n^\star = \frac{\log\left[\dfrac{D_\infty - L^\star/\eta}{D_\infty - D_0}\right]}{\log(1-\alpha_s)}, \tag{14}$$

   the first tick satisfying the target is $n_{\mathrm{hit}} = \lceil n^\star \rceil$, and $T_{\mathrm{hit}} = n_{\mathrm{hit}}/f$.
3. *If $L^\star \ge \eta D_\infty$, the target is never reached and the crossing time is infinite; this includes the boundary $L^\star = \eta D_\infty$, which is approached but not attained since $(1-\alpha_s)^k > 0$ for all finite k.*

The three cases are exhaustive and mutually exclusive. In case 2 the bracketed ratio lies in $(0,1)$, so $n^\star > 0$ and $n_{\mathrm{hit}} = \lceil n^\star \rceil \ge 1$; the degenerate value $n^\star = 0$ arises only at the case 1 boundary and is excluded there. The downward case $D_\infty < D_0$ is symmetric and is not used here; the proposition is scoped to upward crossing.

Proof. (13) follows from the affine contraction of Proposition 9. Under $D_\infty > D_0$, D_k increases monotonically to D_∞, so $\eta D_k \ge L^\star$ first holds at the smallest integer k with $(1-\alpha_s)^k \le (D_\infty - L^\star/\eta)/(D_\infty - D_0)$. Case 2 gives a bracketed ratio in $(0,1)$, and taking logarithms and dividing by $\log(1-\alpha_s) < 0$ yields (14); the ceiling follows because updates occur at discrete ticks. Case 1 makes the ratio at least one and Case 3 makes it non-positive, so no finite k exists in Case 3. □

Remark 14 (Continuous versus discrete). $n^\star/f$ is a continuous interpolation and is reported as such; $T_{\mathrm{hit}} = \lceil n^\star \rceil / f$ is the time at which the implementation first satisfies the target.

Remark 15 (Structural versus statistical equivalence). A statistically equivalent pair, meaning identical conditional observation distributions on all available channels, has infinite reduction time. A structurally equivalent pair may still separate over time, since binary signatures cannot express a difference carried by magnitude or dispersion. The infinite case is conditional on a fixed context and observation model; a regime change reactivating a gated channel, or a chosen input in the sense of active diagnosis [14], can make it finite.

C Dependent multiple gates

Proposition 11 (Refinement survives joint admission)**.** *Admitting $Z_1, \ldots, Z_K$ refines Π^{ctx} regardless of dependence among the gates.*

Proof. The columns gain K appended coordinates; equality of the extended columns implies equality of the first $|A(z)|$ coordinates. The argument is structural and uses no probabilistic assumption. □

Proposition 12 (Product-form gate correction fails under conditional dependence)**.** *Write $w_h^{(1:K)} = p(Z_1 = z_1, \ldots, Z_K = z_K \mid h)$.*

1. *If the gates are conditionally independent given H, that is $w_h^{(1:K)} = \prod_i p(Z_i = z_i \mid h)$ for every h, then the pairwise corrections add: $\lambda_{ab}^{(1:K)} = \sum_i \lambda_{ab}^{(i)}$.*
2. *If that factorisation fails, the summed marginal correction is in general not equal to the true joint correction. For the family of K conditionally comonotone gates constructed below, the discrepancy $|\sum_i \lambda_{ab}^{(i)} - \lambda_{ab}^{(1:K)}|$ grows linearly in K and is therefore unbounded as $K \to \infty$.*

Proof. Part 1 is factorisation followed by Proposition 1.

For part 2, take $\mathcal{H} = \{H_0, H_1\}$ with equal priors and K gates each having marginals $p(Z_i = 0 \mid H_0) = 0.2$ and $p(Z_i = 0 \mid H_1) = 0.8$. Suppose the gates are conditionally comonotone given H, so that all close together and $w_h^{(1:K)} = \min_i p(Z_i = 0 \mid h)$, giving $(0.2, 0.8)$ independently of K. Then $\lambda_{ab}^{(1:K)} = \log 4 = 1.386$ nats for every K, whereas each marginal contributes $\log 4$ and the sum is $K \log 4$. The discrepancy is $(K-1)\log 4$, linear in K.

At $K = 2$ it is $\log 4 = 1.386$ nats, so the summed form gives 2.773 against a true 1.386, exactly double. □

Remark 16 (Scope). The mathematical condition is failure of conditional factorisation, not causal structure. Shared upstream causes such as a common power rail or clock domain are one mechanism producing such dependence and motivate the comonotone construction, but they are an illustration rather than the hypothesis. No claim is made that the discrepancy is unbounded for fixed finite K; the growth statement is in K.

Proposition 13 (Dependence-robust envelope from marginals)**.** *Suppose only the gate marginals $p_{i,h} = P(Z_i = z_i \mid H = h, C)$ are modelled, with $0 < p_{i,h} < 1$.* ***Assume the conditional dependence structures are unconstrained and may be chosen independently for each hypothesis****; the sharpness claims below depend on this and fail without it. Then the joint gate likelihood W_h is confined to the Frechet-Hoeffding interval for an intersection,*

$$L_h \;\le\; W_h \;\le\; U_h,$$
$$L_h = \max\Big(0,\; \textstyle\sum_i p_{i,h} - (K-1)\Big), \qquad U_h = \min_i p_{i,h}, \tag{15}$$

both endpoints attainable. Finiteness of the pairwise correction bounds is directional:

- *a finite* lower *bound on λ_{ab} exists iff $L_a > 0$, and is then $\log(L_a/U_b)$;*
- *a finite* upper *bound exists iff $L_b > 0$, and is then $\log(U_a/L_b)$;*
- *a finite* two-sided *bound exists iff $L_a > 0$ and $L_b > 0$, giving*

$$\log \frac{L_a}{U_b} \;\le\; \lambda_{ab} \;\le\; \log \frac{U_a}{L_b}, \tag{16}$$

with both endpoints jointly attainable under the independence-of-structures assumption above.

If $L_a = 0$ and $L_b > 0$ then λ_{ab} can be made arbitrarily negative while $\log(U_a/L_b)$ still bounds it above. If $L_a > 0$ and $L_b = 0$ it can be made arbitrarily positive with $\log(L_a/U_b)$ still bounding it below. If $L_a = L_b = 0$ the marginals alone give no finite bound in either direction.

Proof. (15) is the classical Frechet-Hoeffding bound for the intersection of K events with given marginals, and both endpoints are attained by admissible event configurations. For $K = 2$ the upper endpoint is attained by the comonotone coupling and the lower by the countermonotone one; for $K > 2$ we do not use that terminology, and refer instead to an admissible event construction attaining the Frechet lower bound, which the classical theorem provides. The directional statements follow by monotonicity of $\log(\cdot/\cdot)$ in each argument together with the independent choice of the two couplings; when $L_h = 0$ the interval for W_h reaches zero, so $\log W_h$ has no finite lower bound. □

Corollary 4 (Dependence-robust decision certificate)**.** *Let $h^\star$ maximise q^{omit}. If for every $j \neq h^\star$*

$$\log \frac{q^{\mathrm{omit}}(h^\star)}{q^{\mathrm{omit}}(j)} + \log \frac{L_{h^\star}}{U_j} \;>\; 0,$$

then $h^\star$ remains the unique maximiser under every admissible dependence structure. If $L_{h^\star} = 0$ no such certificate exists. The converse certificate must be stated carefully: the condition $\log(q^{\mathrm{omit}}(h^\star)/q^{\mathrm{omit}}(j)) + \log(U_{h^\star}/L_j) < 0$ for some j shows only that that *competitor can win, which is weaker than showing $h^\star$ loses under every admissible structure.*

Remark 17 (What the marginals can and cannot certify). Three cases, all with $K = 2$.

Independence is safe. Marginals $(0.9, 0.9)$ give $W \in [0.8, 0.9]$, a narrow envelope containing the product value 0.81; the correction is well determined whatever the dependence.

The product form is decision-unsafe. Under $H = a$ take marginals $(0.5, 0.5)$ and under $H = b$ take $(0.25, 0.25)$. The product form gives $\lambda_{ab} = \log 4 = 1.386$, which with a pre-gate margin of $\Delta = \log 1.5 = 0.405$ keeps a the winner. But $L_a = 0$, so an admissible countermonotone structure under $H = a$ drives W_a toward zero and $\lambda_{ab} \to -\infty$: at $W_a = 0.05$ against $W_b = 0.25$ the post-gate log odds is already -1.204 and b wins. The same per-gate marginal specification therefore admits dependence structures placing the decision on opposite sides of the boundary. Note the marginals are not identical *across* hypotheses here; what is held fixed is the marginal specification, and what varies is the dependence.

Marginally healthy, jointly unidentified. Two gates each with marginal 0.5 give $L = \max(0, 1 - 1) = 0$. Each gate looks unremarkable, yet the observed joint event can have probability arbitrarily close to zero, and the joint gate evidence is unidentified from marginals alone.

The exact statement is this. From gate marginals alone, a strictly positive lower bound on the joint observed-gate likelihood W_h exists if and only if $\sum_i p_{i,h} > K - 1$. When that fails, W_h may approach zero under an admissible dependence structure, so the marginals alone provide no finite lower log-likelihood contribution for that hypothesis, and consequently no finite lower bound on any pairwise correction in which it appears as the numerator. Bounding the correction in the other direction requires the same condition for the denominator hypothesis. A designer wanting a two-sided guarantee from marginals must therefore verify $\sum_i p_{i,h} > K - 1$ for *both* hypotheses in the pair; otherwise information beyond the marginals, such as a dependence model, is required.

Remark 18 (Attribution). Proposition 13 is the Frechet-Hoeffding bound followed by (6), and Corollary 4 is interval dominance applied to the resulting envelope. The underlying mathematics is Frechet-Hoeffding plus Bayes plus interval dominance, all standard in probability bounds analysis and imprecise probability. They are included be-

cause Proposition 12 otherwise leaves a designer with a counterexample and no remedy.

Remark 19. We do not conjecture a general inequality relating joint and summed corrections. Association inequalities suggest positive dependence should attenuate, but the ordering of $w^{(i)}$ across hypotheses need not be common across gates, and we have neither a proof nor a counterexample in the general case.

D Source attribution as a second reweighting

Remark 20 (Source attribution as a second reweighting). Error C admits the same treatment as Error A, which is the point of recording it here rather than only as a caveat. Let S be the actual physical-source random variable, $\tilde{S}$ the declared-label random variable, $\tilde{s}$ the observed declared-label value, and e the observation. Retain the unconditional prior $\pi_h = P(H = h)$ used throughout, and write the *label likelihood*

$$r_h(\tilde{s}) \coloneqq p(\tilde{S} = \tilde{s} \mid H = h), \tag{17}$$

which a correctly written generative model must carry because the declared label is itself an observation. An implementation that takes the declared label at face value scores

$$q^{\text{label}}(h) \;\propto\; \pi_h \, r_h(\tilde{s}) \, p(e \mid S = \tilde{s},\, \tilde{S} = \tilde{s},\, h), \tag{18}$$

whereas under the declared generative model the correct likelihood marginalises over the physical source,

$$\begin{aligned} L_{\text{src}}(h; \tilde{s}) = \sum_s p(e \mid S = s,\, \tilde{S} = \tilde{s},\, h) \\ \times\; p(S = s \mid \tilde{S} = \tilde{s},\, h), \end{aligned} \tag{19}$$

and the correct score is

$$p(h \mid e, \tilde{S} = \tilde{s}) \;\propto\; \pi_h \, r_h(\tilde{s}) \, L_{\text{src}}(h; \tilde{s}). \tag{20}$$

Neither (19) nor (20) makes an independence assumption. Assume $p(e \mid S = \tilde{s}, \tilde{S} = \tilde{s}, h) > 0$ and $r_h(\tilde{s}) > 0$ for every h, and that the attribution model $p(S = s \mid \tilde{S} = \tilde{s}, h)$ is declared rather than left implicit. Setting

$$\rho_h(\tilde{s}) = \frac{L_{\text{src}}(h; \tilde{s})}{p(e \mid S = \tilde{s},\, \tilde{S} = \tilde{s},\, h)}$$

gives

$$p(h \mid e, \tilde{S} = \tilde{s}) \;\propto\; q^{\text{label}}(h) \, \rho_h(\tilde{s}), \tag{21}$$

with pairwise correction $\log(\rho_a(\tilde{s})/\rho_b(\tilde{s}))$, by the same argument as Proposition 1.

On the label factor. The factor $r_h(\tilde{s})$ appears in both (18) and (20) and therefore cancels in the ratio (21): the source correction ρ_h is unaffected by it. It does not cancel from either score individually. An implementation that omits $r_h(\tilde{s})$ *and* assumes correct attribution commits two distinct errors, and (21) repairs only the second. If $\tilde{S} \perp H$, then $r_h(\tilde{s})$ is constant in h and drops out of both normalised scores; this is a defensible assumption when the declared label is software-commanded and genuinely exogenous to the fault state. It must not be assumed when label generation, channel selection, multiplexing, routing, or bus behaviour can depend on the hypothesis. A label-conditioned prior $\pi_h^{\tilde{s}} \coloneqq P(H = h \mid \tilde{S} = \tilde{s})$ is an equivalent convention; we do not use it, to avoid mixing it with the unconditional π_h elsewhere in the paper.

Optional simplification. If the measurement depends on the declared label only through the physical source, $e \perp\!\!\!\perp \tilde{S} \mid (S, H)$, then $p(e \mid S = s, \tilde{S} = \tilde{s}, h) = p(e \mid S = s, h)$ and (19) reduces to $L_{\text{src}}(h; \tilde{s}) = \sum_s p(e \mid S = s, h)\, p(S = s \mid \tilde{S} = \tilde{s}, h)$. This holds when the label is a routing artefact that does not alter the physics of the measurement. It is an assumption, stated separately, and nothing above requires it.

Controls. Five deterministic checks accompany the reproduction artifact: an exogenous label (constant r_h) cancels from the normalised posterior to machine precision; a hypothesis-dependent label changes the posterior by up to 0.18 in total mass while (21) continues to hold to 10^{-17}; perfect attribution gives $\rho_h = 1$ for every h; misattribution gives a nontrivial pairwise correction (-0.671 nats on the worked instance); and with no declared attribution model, ρ_h is unidentified, ranging over the model class rather than defaulting to one.

Three points bound what this says. First, (21) is ordinary self-normalised likelihood reweighting, as (5) is. Second, it is not solved by the gate correction: w_h and $\rho_h(\tilde{s})$ are reweightings with respect to different latent variables, and admitting the gate leaves $\rho_h(\tilde{s})$ untouched. Third, and most restrictive, it requires a declared model for $p(S = s \mid \tilde{S}, h)$, which an implementation that assumes correct attribution does not possess. Where no such model is declared, (21) identifies the missing object rather than supplying it.

E A recorded failed sharpening

Remark 21 (Failed sharpening, recorded). Neither $\text{Var}_{q^{\text{omit}}}(\log w)$ nor $\text{Var}_p(\log w)$ bounds the divergence. An explicit two-atom witness suffices: take $q^{\text{omit}} = (0.993, 0.007)$ and $w = (e^{-9.8}, 1)$, that is $w \approx (5.5 \times 10^{-5}, 1)$ with $R = 9.8$ (equivalent to $\log w = (-4.9, 4.9)$ by Lemma 2). Then $p = (0.007805, 0.992195)$ and

$$\begin{aligned} &\text{KL} = 4.877523 \;>\; \text{Var}_{q^{\text{omit}}}(\log w) = 0.667574, \\ &\text{KL} \;>\; \text{Var}_p(\log w) = 0.743724. \end{aligned}$$

Both endpoint variances fail on the same configuration. Each is evaluated at a single point of the tilt path while the divergence integrates Λ'' along the whole path.

F Claim status

Every numbered result and substantive remark carries one primary status. *Proved, known* means the statement is established in the cited literature and re-proved here for completeness; *proved, application* means a known result specialised to this setting; *corollary of* names the theorem it follows from. The results below are classified by their mathematical provenance. The paper's contribution is their unified specialization to validity-gated runtime diagnosis and the resulting design-audit framework, rather than priority for the component tools. No entry is labelled novel.

Result	Status
Prop. 1	Proved, known
Prop. 2	Proved, known
Prop. 3	Proved, tight; corollary of [5]
Prop. 4	Proved, tight; corollary of [5]
Prop. 5	Proved, known
Prop. 6	Proved, application
Prop. 7	Proved, known
Prop. 8	Proved, known (interval dominance)
Prop. 9	Proved, known
Prop. 10	Proved, application (implementation result)
Prop. 11	Proved, known
Prop. 12	Proved (part 1); by construction (part 2)
Prop. 13	Proved, classical (Fréchet-Hoeffding)
Lem. 1	Proved, known (binary gate)
Cor. 1	Proved, known
Cor. 2	Proved, application
Cor. 3	Proved, direct corollary
Cor. 4	Proved, known (interval dominance)
Rem. 1	Counterexample, scope
Rem. 2	Scope (structural versus probabilistic)
Rem. 3	Scope clarification (Errors A, B, C)
Rem. 5	Proved, interval arithmetic
Rem. 6	Attribution
Rem. 7	Explicit counterexample
Rem. 10	Attribution (argmax stability)
Rem. 11	Explicit constructions
Rem. 13	Scope clarification
Rem. 16	Scope clarification
Rem. 17	Explicit constructions
Rem. 18	Attribution
Rem. 20	Proved, known; deterministic controls
Rem. 21	Explicit counterexample
Sec. 8	Numerical verification

G Reproducibility

Every numerical value in the paper regenerates from `reproduce.py` (NumPy, float64, deterministic). The source-attribution controls of Appendix D regenerate from `srcattr_controls.py`. Both scripts and their JSON outputs form the supplementary reproduction package, whose manifest records interpreter and library versions, seeds, sample counts, and a SHA-256 digest of each shipped script and data file.

Four kinds of check appear in the paper and are labelled as such throughout. *Symbolic identities* (the tilt, the pairwise correction, the crossing formula) hold by algebra and are confirmed by direct evaluation. *Numerical verification* evaluates closed-form quantities at the declared parameters, for example the divergence identity to 10^{-12} and the discrete crossing tick by simulation of (12). *Randomised falsification searches* test an inequality over random configurations with a fixed seed: the tilt cancellation of L_A over 2×10^5 draws (maximum deviation 5.6×10^{-16}), the total-variation bound over 4×10^5 (zero violations, maximum ratio 1.000000), the refinement direction over 2×10^5 (0 violations for uniform appending, 110,819 for column-dependent deletion). A randomised search that finds no violation is evidence, not proof; the propositions rest on their proofs. *Empirical validation* does not occur in this paper.

References

[1] W. Dong, X. Yin, and S. Li, "A uniform framework for diagnosis of discrete-event systems with unreliable sensors using linear temporal logic," *IEEE Trans. Autom. Control*, vol. 69, no. 1, pp. 145–160, 2024.

[2] L. K. Carvalho, J. C. Basilio, and M. V. Moreira, "Robust diagnosis of discrete-event systems against intermittent loss of observations," *Automatica*, vol. 48, no. 9, pp. 2068–2078, 2012.

[3] L. K. Carvalho, M. V. Moreira, J. C. Basilio, and S. Lafortune, "Robust diagnosis of discrete-event systems against permanent loss of observations," *Automatica*, vol. 49, no. 1, pp. 223–231, 2013.

[4] S. Takai, "A general framework for diagnosis of discrete event systems subject to sensor failures," *Automatica*, vol. 129, 109669, 2021.

[5] A. L. Rukhin, "Information-type divergence when the likelihood ratios are bounded," *Applicationes Mathematicae*, vol. 24, no. 4, pp. 415–423, 1997.

[6] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Wiley, 2019.

[7] C. F. Manski, *Partial Identification of Probability Distributions*. Springer, 2003.

[8] J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker, 1998.

[9] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Diagnosis and Fault-Tolerant Control*, 3rd ed. Springer, 2016.

[10] M. Krysander and E. Frisk, "Sensor placement for fault diagnosis," *IEEE Trans. Syst., Man, Cybern. A*, vol. 38, no. 6, pp. 1398–1410, 2008.

[11] M. Krysander and F. Hashemniya, "Test selection for diagnosing multimode systems," in *35th Int. Conf. Principles of Diagnosis and Resilient Systems (DX 2024)*, OASIcs vol. 125, pp. 28:1–28:14, 2024. doi:10.4230/OASIcs.DX.2024.28

[12] M. A. Atoui and A. Cohen, "Coupling data-driven and model-based methods to improve fault diagnosis," *Computers in Industry*, vol. 128, 103401, 2021.

[13] P. Antonante, H. Nilsen, and L. Carlone, "Monitoring of perception systems: deterministic, probabilistic, and learning-based fault detection and identification," *Artificial Intelligence*, vol. 325, 103998, 2023.

[14] R. Nikoukhah, "Guaranteed active failure detection and isolation for linear dynamical systems," *Automatica*, vol. 34, no. 11, pp. 1345–1358, 1998.