AGGREGATION/OPTIMIZATION STABILITY GAP IN NEURO-SYMBOLIC INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Neuro-symbolic systems compute two structurally different kinds of query over the same representation: *aggregation* (marginalizing or averaging over hypotheses) and *optimization* (selecting a single best witness). We show these obey different stability laws. Aggregation is provably Lipschitz-continuous under belief perturbations, for any bounded-weight aggregation and any Lipschitz neural scorer. Optimization is not: we construct, explicitly, a neural scorer of arbitrary smoothness under which an arbitrarily small input perturbation produces an arbitrarily large change in the selected witness. We show both the magnitude and frequency of an optimization failure. We instantiate this result on three structurally unrelated representations – model-based diagnosis, abductive explanation, and constrained route planning under belief uncertainty – and validate it empirically on all three, including a controlled, timed comparison of aggregation and optimization on identical trained representations. We further show optimization-side intractability is not inevitable: two independent mechanisms, structural (bounded treewidth) and algorithmic (randomized restart), recover practical tractability.

1 INTRODUCTION

Existing neuro-symbolic taxonomies typically classify systems by where neural and symbolic computation occur (Bhuyan et al., 2024; Hitzler et al., 2022; Velasquez et al., 2026). We argue a second, orthogonal axis matters at least as much: *what kind of query* is being computed – aggregation or optimization – since the same representation can support both, with radically different computational and statistical behavior. Aggregation-mode inference computes beliefs, expectations, or weighted sums over explanations. Optimization-mode inference selects a single best explanation, witness, or action. These two modes often exhibit fundamentally different computational complexity even when implemented over the same representation. While existing neurosymbolic frameworks classify systems according to the interaction of neural and symbolic components, they generally do not distinguish aggregation-oriented and optimization-oriented inference as separate computational regimes. We argue that this distinction is complexity-theoretically fundamental because the same representation may support tractable aggregation while rendering optimization intractable.

Sum-product networks make this concrete already: computing a marginal probability is polynomial-time by construction, while computing the most probable explanation (MAP) over the identical network is NP-hard (Peharz, 2015; Mei et al., 2018; Conaty et al., 2017) – same representation, same weights, two queries, two complexity classes. This is not an isolated curiosity of one representation; it is an instance of a general phenomenon.

We formalize this property using Theorem 1: aggregation is Lipschitz-stable; optimization, even composed with an arbitrarily smooth scorer, need not be. We give an explicit worst-case construction proving this, then show it is realized concretely by three structurally unrelated representations, with empirical validation on all three. We formally show special cases of when stability can be recovered in Theorem 2. We extend our results to show how often these issues arise in Theorems 3 and 4.

Contributions.

- A stability theorem for aggregation and optimization (Theorem 1) with an explicit worst-case construction, instantiated on 3 structurally unrelated representations, model-based di-

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.

agnosis, abductive explanation, constrained routing, which show that Theorem 1 is not domain-specific (Section 5.2).

- A frequency theory (Theorems 3–4) converting the worst-case result into a margin-based prediction.
- Six hypotheses designed to empirically validate the theoretical results: (H1-H3 for Theorem 1, Corollary 2; H4 for Theorem 2) and tested on a trained representation (Sections 4, 7); H5 and H6 for Theorems 3–4 (Section 8).

2 Related Work

Due to space constraints, we provide a brief discussion of related work here; Appendix B presents a detailed comparison with differentiable optimization, decision-focused learning, probabilistic circuits, model-based diagnosis, abductive explanation, and neuro-symbolic reasoning literature. It is important to note that this paper, in particular Theorem 1, is not competing with Gumbel-Softmax or differentiable combinatorial optimization Jang et al. (2017); Vlastelica et al. (2020); it studies witness-set sensitivity and query-dependent stability instead.

Formal explanations. The abductive/contrastive explanation literature independently establishes the duality Corollary 3 uses: AXp is a minimal hitting set of CXps, and vice versa (Ignatiev et al., 2021), underlying enumeration algorithms such as MARCO (Liffiton et al., 2016), which Appendix F implements directly.

Tractable probabilistic models. Sum-product networks and probabilistic circuits are explicitly designed for tractable marginal inference (Peharz, 2015); MAP inference over the identical network is NP-hard (Mei et al., 2018; Conaty et al., 2017) – the query-separation phenomenon Theorem 1 generalizes.

Semirings and message passing. We distinguish the generalized distributive law (Aji & McEliece, 2000) from our results in Remarks 1 and 4.

Structural tractability. Courcelle's theorem Courcelle (1990) is the standard route from MSO-expressible queries to FPT algorithms on bounded-treewidth structures; constant factors in this route can be large, and recent work sharpens known lower bounds on exactly how large they must be in the worst case for certain problem classes – a caveat Proposition 1's own achievability result inherits and does not evade.

3 Preliminaries

An inference task over weighted hypotheses H (weights $w(h) \geq 0$) is *aggregative* if the answer combines contributions from many hypotheses – $A = \sum_h f(h)w(h)$, e.g. marginal probability, expectation, partition function – and *optimization-based* if the answer is a single extremal hypothesis, $h^* = \arg\max_h w(h)$, e.g. MAP, MPE, best plan. Aggregation preserves uncertainty; optimization collapses it into one witness, and only the optimum matters.

Remark 1 (Relation to the generalized distributive law)**.** *This split has a standard algebraic description in the message-passing literature: given a factored objective $F(x) = \bigotimes_i f_i(x_{S_i})$, computing $\bigoplus_{x_2,\ldots,x_n} F(x)$ with $(\oplus, \otimes) = (+, \times)$ is the* sum-product *semiring reduction (marginals, expectations), and with $(\oplus, \otimes) = (\max, \times)$ is the* max-product *semiring reduction (MAP/MPE value) – the generalized distributive law (Aji & McEliece, 2000), which recovers belief propagation, the junction-tree algorithm, Viterbi, and the FFT as special cases via choice of semiring. Our aggregation/optimization split is the same conceptual distinction, and the junction-tree connection already noted for bounded-treewidth structures (Remark 4 makes this precise for the structural-achievability route below) is a direct instance of it. One difference worth being precise about: the classical framework computes an optimal* value *under $(\max, \times)$, with the optimal* assignment *recovered by a separate traceback step; $\Psi_L(K)$ (Definition 1 below) computes the* set *of minimal witnesses directly, not a scalar reduction. The two are the same distinction at the conceptual level, not literally the same computational object.*

We study architectures of the shape $X \xrightarrow{f_\theta} p \xrightarrow{\tau} P \xrightarrow{\Psi} \Delta$: a neural scorer f_θ (assumed L-Lipschitz) produces soft confidences $p \in [0,1]^m$ over m candidate predicates; a discretization step τ produces a predicate set P; a decision operator Ψ maps P to a solution Δ. *Aggregation-mode* queries compute $F_{\text{agg}}(p) = \sum_i w_i p_i$ for bounded weights, never invoking Ψ. *Optimization-mode* queries compute $\Psi(\tau(f_\theta(x)))$ directly – a discrete, combinatorial selection.

Definition 1 (Optimization-mode inference)**.** *Given interpretations Ω, a satisfaction predicate $\mu : L \times \Omega \to \{0,1\}$ over a predicate family L, and a target family $\mathcal{K} \subseteq L$ (derived from P above via τ), optimization-mode inference computes*

$$\Psi_L(\mathcal{K}) = \{\, \omega \in \Omega : \mu(\kappa, \omega) = 1 \; \forall \kappa \in \mathcal{K},\ \omega \text{ inclusion-minimal with this property} \,\}.$$

This is general enough to state every result below without reference to any one domain; Section 5.2 instantiates $(\Omega, \mu, \mathcal{K})$ three separate ways.

Definition 2 (Distance metrics)**.** *$d_P(P_1, P_2)$ is the symmetric-difference cardinality between two predicate families, $|P_1 \diamond P_2|$. $d_\Delta(\Delta_1, \Delta_2)$ is the symmetric-difference cardinality between two solution sets, $|\Delta_1 \diamond \Delta_2|$ (each Δ_i itself a family of witnesses, e.g. the set of minimal covers $\Psi(P_i)$). Decision sensitivity is $S_\Psi = \sup_{P_1 \neq P_2} d_\Delta(\Psi(P_1), \Psi(P_2))/d_P(P_1, P_2)$: the worst-case ratio of solution-set change to predicate-set change.*

4 Hypotheses

The results in this paper support six precise, testable hypotheses about the aggregation/optimization distinction of Section 3.

H1 (Query separation). Aggregation queries require asymptotically less computation than optimization queries on the same representation, as the relevant uncertainty problem size/number of required waypoints (K) grows.

H2 (Complexity transition). For optimization queries, computational cost grows exponentially with K, whereas aggregation queries exhibit sub-exponential growth.

H3 (Stability). Aggregation outputs vary smoothly (Lipschitz-continuously) under belief perturbations; optimization outputs can change discontinuously at decision boundaries, with the size of the discontinuity governed by margin, not by perturbation size.

H4 (Recoverable tractability). Optimization-side intractability is not inevitable: it can be mitigated either structurally (bounded-treewidth, MSO_1-expressible queries) or algorithmically (randomized restart, conditional on a favorable search landscape).

H5 (Margin CDF Predicts Optimization Instability). The empirical instability frequency is upper-bounded by the empirical margin CDF.

H6 (Impact of Temperature Scaling on Calibration and Instability). Showing that instability is governed by the interaction of margin and the calibration error δ.

H1–H3 follow from Theorem 1 (Section 5); H4 follows from the achievability results of Theorem 2 in Section 5.3. Section 7 tests H1–H4 directly. H5 and H6 follow Theorem 3.

5 Theory

5.1 A General Stability Theorem

Theorem 1 (Aggregation-Optimization Stability)**.** *Let $X \xrightarrow{f_\theta} p \xrightarrow{\tau} P \xrightarrow{\Psi} \Delta$ be the pipeline of Section 3, with f_θ L-Lipschitz, τ a fixed per-candidate discretization threshold, and Ψ a minimal-cover operator.*

(a) Aggregation is stable. *For $F_{\text{agg}}(p) = \sum_{i=1}^m w_i p_i$ with $|w_i| \leq M$, $|F_{\text{agg}}(p) - F_{\text{agg}}(q)| \leq M\|p - q\|_1$ for all $p, q \in [0,1]^m$, and $F_{\text{agg}} \circ f_\theta$ is ML-Lipschitz in x.*

(b) Optimization need not be stable, even composed with a Lipschitz scorer. *For every $L > 0$, $\varepsilon > 0$, and target magnitude M', there exist an L-Lipschitz f_θ, a threshold τ, a minimal-cover operator Ψ, and x, x' with $\|x - x'\| \leq \varepsilon$, such that $d_\Delta(\Psi(\tau(f_\theta(x))), \Psi(\tau(f_\theta(x')))) \geq M'$.* $\square$

We present all proofs in Appendix E. Theorem 1(a) is H3's stable direction; part (b) is H3's discontinuous direction, and its scaling consequence over K (the size of the witnessing instance) is H1 and H2. Both parts are, as stated, narrower than the theorem's title suggests – part (a) is proved only for linear bounded-weight sums, and part (b) only for minimal-cover Ψ. Neither restriction is essential, and both are generalized precisely below rather than left as an unstated gap.

5.1.1 WHAT ACTUALLY MAKES AGGREGATION STABLE

Lemma 1 (Bounded gradient implies Lipschitz stability)**.** *Let $F : [0,1]^m \to \mathbb{R}$ be differentiable with $\|\nabla F\|_\infty \le M$ everywhere on the domain. Then $|F(p) - F(q)| \le M\|p - q\|_1$ for all $p, q \in [0,1]^m$.*

This drives both stability results, rather than any property specific to sums or to products.

Corollary 1 (Theorem 1(a), as an instance)**.** *For $F_{\text{agg}}(p) = \sum_i w_i p_i$ with $|w_i| \le M$: $\partial F_{\text{agg}}/\partial p_i = w_i$, so $\|\nabla F_{\text{agg}}\|_\infty \le M$, and Lemma 1 gives Theorem 1(a) directly.*

Corollary 2 (Product aggregation is also stable)**.** *For $g(p) = \prod_{i=1}^n (1 - p_i)$ on $p \in [0,1]^n$ (Instantiation 3's aggregation query): $\partial g/\partial p_j = -\prod_{i \ne j}(1 - p_i) \in [0,1]$ for every j (each factor $(1 - p_i) \in [0,1]$, so their product is too), giving $\|\nabla g\|_\infty \le 1$, and Lemma 1 gives $|g(p) - g(q)| \le \|p - q\|_1$.*

Remark 2 (Why not every aggregation is stable)**.** *Lemma 1 is the right level of generality, not the maximal one: it is false that* every *aggregation over many hypotheses is Lipschitz. $F(p) = \sum_i \mathbb{1}[p_i > 0.5]$ (count how many exceed threshold) combines contributions from many hypotheses in the same informal sense sums and products do, but is a step function, discontinuous at every $p_i = 0.5$ – unbounded gradient, no Lipschitz bound at any constant. Bounded-gradient aggregations (sums, products, and by the same argument min/max) are stable; threshold-based aggregations are not, for essentially the same reason optimization is not (Section 5.1.2 below) – discretization is the common failure point in both directions.*

5.1.2 WHAT ACTUALLY MAKES OPTIMIZATION UNSTABLE

Theorem 1(b)'s construction and its proof (Appendix E.1) never use minimal-cover structure directly; they use only that Ψ admits a witnessing pair of predicate sets one flip apart with arbitrarily large solution-set change.

Definition 3 (Single-flip unbounded sensitivity)**.** *Ψ has* single-flip unbounded sensitivity *if for every target M', there exist predicate sets P_1, P_2 with $d_P(P_1, P_2) = 1$ and $d_\Delta(\Psi(P_1), \Psi(P_2)) \ge M'$.*

Proposition 1 (Theorem 1(b) generalizes beyond minimal-cover operators)**.** *Theorem 1(b) holds verbatim for any Ψ with single-flip unbounded sensitivity, minimal-cover or not – the proof in Appendix E.1 uses only the witnessing pair Definition 3 supplies, nothing about minimal covers specifically. Minimal-cover operators are a* proven witness class*, not the theorem's true scope: Theorem 5 (Appendix D) establishes exactly Definition 3 for them, via the k-pair construction.* □

Whether Instantiation 3's routing operator has single-flip unbounded sensitivity is not established here; nevertheless, Section 7 provides some suggestive evidence of unboundedness.

Remark 3. *Part (b) is a worst-case existence construction, not a claim about typical scorers – the point is that no amount of smoothness in f_θ* rules out *the discontinuity, not that it is typical. Section 7 tests how this manifests on real, trained (not adversarially-constructed) scorers.*

5.2 INSTANTIATIONS

To show the generality of this approach, we show that Theorem 1(b) is applicable to three structurally unrelated representations, which instantiate Definition 1's $(\Omega, \mu, \mathcal{K})$ and realize it.

Instantiation 1: model-based diagnosis. $\Omega = 2^U$ for component universe U, μ = "hits every conflict," $\mathcal{K}$ = the reported conflict family. Reiter diagnosis, structural diagnosis, MUS-based repair, and geometric MBD are recovered as instances by the same argument.

Instantiation 2: abductive explanation. $\Omega = 2^F$ for feature set F, μ = the hitting predicate, $\mathcal{K}$ = the contrastive-explanation family.

Corollary 3 (Abductive explanation as an instance of Definition 1)**.** *Taking $(\Omega, \mu, \mathcal{K})$ as above, Definition 1 gives $\Psi_L(\mathcal{K}) = \{$abductive explanations (AXps)$\}$. This is a direct corollary of an independently-established duality in the explainable-AI literature (AXp is a minimal hitting set of CXps, and vice versa), proved without reference to diagnosis at all.*

Instantiation 3: constrained routing under belief uncertainty. Ω = waypoint orderings, μ = route feasibility under a belief-induced blocked-cell set, $\mathcal{K}$ = the waypoint-visitation requirement. Its aggregation query ($\prod_{c \in \text{route}}(1 - p_c)$, a literal product, stable by Corollary 2) and optimization query (exact waypoint-ordering search) are computed over an identical trained representation, giving H1/H2 a direct, timed, controlled test that the other two instantiations cannot provide alone. *Unlike Instantiations 1 and 2, this Ψ is not a minimal-cover operator* – route feasibility under a blocked-cell set is a different combinatorial structure. By Proposition 1, Theorem 1(b) would apply directly if this Ψ has single-flip unbounded sensitivity (Definition 3); whether it does is not established here (Section 5.1.2). Section 7's H3 test on this instantiation is accordingly framed as evidence that the same underlying mechanism – margin governs whether a discontinuity is reachable at all – recurs in a structurally different, non-minimal-cover operator family, not as a direct corollary of Theorem 1(b).

5.3 Recovering Tractability

Optimization hardness is not inevitable – the theoretical basis for H4. **Structural**: if the optimization query is expressible in MSO_1 over a bounded-treewidth incidence graph, it is fixed-parameter tractable (Courcelle's theorem). **Algorithmic**: randomized-restart search achieves the target witness at cost logarithmic in target confidence, conditional on the search landscape's basin probability. Neither is unconditional; Section 7 validates both on real data and shows the algorithmic route's conditionality directly, via an instance-dependent gap.

Remark 4 (The structural route is a joint-tractability mechanism, not a separating one)**.** *Bounded treewidth's role here is worth distinguishing carefully from Theorem 1's own separation result, in the vocabulary of Remark 1: bounded-treewidth structures make* both *the sum-product (aggregation) and max-product (optimization) semiring reductions tractable, via the same variable-elimination/junction-tree algorithm, with only the semiring operation swapped from $(+, \times)$ to $(\max, \times)$ (Aji & McEliece, 2000). So structural achievability recovering tractability for optimization here says nothing about separating aggregation from optimization complexity – it is a fundamentally different mechanism than the one Theorem 1 characterizes, one that collapses the two query types together rather than distinguishing them.*

Theorem 2 (Recovering tractability)**.** *Optimization-mode inference need not exhibit practical intractability. Tractability can be recovered by either:*

1. *Structural restriction: when the query is MSO_1-expressible over a bounded-treewidth incidence graph, inference is fixed-parameter tractable.*
2. *Algorithmic amplification: when local search succeeds with probability $p > 0$ per restart, $O(\log(1/\gamma)/p)$ restarts suffice to reduce the failure probability below γ.*

Consequently, the worst-case instability and hardness phenomena identified in Theorem 1 do not preclude efficient inference on important structured subclasses or under favorable search landscapes.

6 Instability Frequency Analysis

Theorem 1 establishes a worst-case separation: optimization-mode inference can exhibit arbitrarily large output changes under arbitrarily small perturbations. As stated, however, Theorem 1 is an existence result. It characterizes the *magnitude* of possible failures but not their expected frequency.

The results below connect optimization instability to three measurable statistical quantities: (1) decision margins, (2) predictor calibration, and (3) perturbation size. The resulting theory converts Theorem 1 from a worst-case statement into a quantitative prediction about when instability should occur in practice.

6.1 Margin-Based Instability Frequency

Recall Proposition 2: if a predicate's margin exceeds $L\varepsilon$, then no perturbation of radius ε can change its membership. Let $\mathcal{E}(x)$ denote the set of belief-essential predicates at state x, i.e. predicates whose membership can alter the optimization output. For each essential predicate $S \in \mathcal{E}(x)$ define the margin $m_S(x) = |p_S(x) - \tau_S|$, and define the minimum essential margin $M(x) = \min_{S \in \mathcal{E}(x)} m_S(x)$.

The optimization instability event is $I = \left\{ \Psi(\tau(f_\theta(x))) \neq \Psi(\tau(f_\theta(x'))) \right\}$, where $\|x - x'\| \leq \varepsilon$.

Theorem 3 (Margin controls instability probability)**.** *We assume the optimization pipeline of Section 3. We further assume: (1) f_θ is L-Lipschitz; (2) $\|x - x'\| \leq \varepsilon$; and (3) optimization outputs can change only through changes in belief-essential predicates. Then* $\Pr(I) \leq \Pr(M \leq L\varepsilon)$. *Equivalently, if $F_M(t) = \Pr(M \leq t)$ denotes the CDF of the minimum essential margin,* $\Pr(I) \leq F_M(L\varepsilon)$.

Theorem 3 identifies the exact quantity governing instability frequency: not the optimization operator itself, but the amount of probability mass concentrated near decision boundaries. Note that outputs change only when an essential predicate flips.

6.2 Calibration-Aware Stability

Theorem 3 assumes only a margin distribution. Calibration permits a stronger statement.

Let $\eta_S(x) = \Pr(S = 1 \mid x)$ denote the true conditional probability for predicate S. Suppose the predictor satisfies a uniform calibration bound $|p_S(x) - \eta_S(x)| \leq \delta$ for every predicate. Define the true minimum essential margin $M^\star(x) = \min_{S \in \mathcal{E}(x)} |\eta_S(x) - \tau_S|$.

Theorem 4 (Calibration controls instability frequency)**.** *Assume $|p_S(x) - \eta_S(x)| \leq \delta$ for every essential predicate. Then* $\Pr(I) \leq \Pr(M^\star \leq L\varepsilon + \delta)$.

This result decomposes instability into two distinct effects: (1) geometric sensitivity ($L\varepsilon$), and (2) probability estimation error (δ). Even when optimization sensitivity is intrinsically discontinuous (Theorem 1), improved calibration reduces the measure of states close enough to threshold for that discontinuity to be reached.

Remark 5 (Improved calibration's meaning)**.** *δ is a* worst-case *bound on $|p_S(x) - \eta_S(x)|$ against the* true *conditional probability η_S, not the same quantity as an average-case calibration diagnostic such as Expected Calibration Error (ECE). A procedure that lowers ECE need not lower δ, and vice versa: ECE averages miscalibration over a data distribution, while δ bounds it uniformly over every essential predicate encountered. Section 8.3 reports an empirical test of this distinction directly, rather than assuming the two coincide.*

6.3 Power-Law Margin Regimes

We use the standard assumption of a margin-tail condition to prove power-law instability properties.

Definition 4 (Margin-tail exponent)**.** *The essential-margin distribution satisfies exponent $\alpha > 0$ if* $\Pr(M^\star \leq t) \leq Ct^\alpha$ *for some constant C and sufficiently small t.*

Corollary 4 (Power-law instability bound)**.** *If* $\Pr(M^\star \leq t) \leq Ct^\alpha$ *for sufficiently small t, then, provided $L\varepsilon + \delta$ is itself sufficiently small for the tail condition to apply,* $\Pr(I) \leq C(L\varepsilon + \delta)^\alpha$.

The exponent α quantifies how much probability mass accumulates near optimization decision boundaries. Large-margin systems (α large) exhibit substantially fewer optimization failures than systems whose beliefs frequently concentrate near thresholds.

7 Empirical Validation

All four hypotheses of Section 4 are tested on a single representation: a grid environment with a trained neural hazard predictor Φ, where the aggregation query (expected route safety) and optimization query (waypoint-constrained routing) are computed over the identical belief state. This

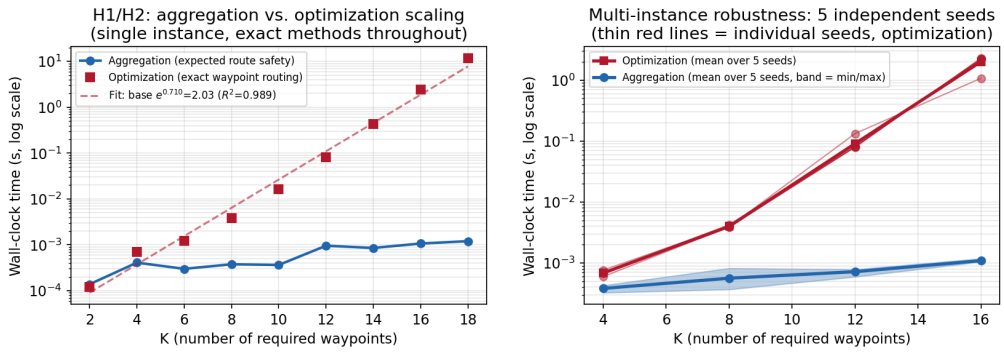

Figure 1: Left: aggregation vs. optimization wall-clock, single instance, both exact. Log-linear fit gives base 2.03 (R^2=0.989) for optimization's growth against a predicted base of 2.0; aggregation's fit gives base 1.13. Right: five independent seeds, same qualitative divergence in every instance.

provides a natural aggregation-mode counterpart to its optimization query on the same representation, and the only one that can test H1/H2 directly rather than by analogy.[1]

7.1 H1/H2: AGGREGATION VS. OPTIMIZATION

Figure 1 (left) tests H1 and H2, stated in Section 4, directly and quantitatively: aggregation and optimization computed over the identical trained representation, swept over instance-size/uncertainty parameter K. Aggregation time stays within a factor of 10 across the entire range tested (K=2 to 18); optimization grows from parity at K=2 to roughly 10^4–$10^5\times$ slower by K=18, tracking 2^K growth closely – a log-linear fit gives slope 0.710 against a predicted $\log 2 = 0.693$ (R^2=0.989), while aggregation's own fit gives a slope under a fifth as large.[2] H1 and H2 are therefore supported quantitatively, not only qualitatively, and Figure 1 (right) confirms the divergence replicates across five independently-trained instances with no exceptions.

7.2 H3: TESTING WHETHER THE SAME MECHANISM RECURS OUTSIDE MINIMAL COVERS

Instantiation 3's Ψ is not a minimal-cover operator (Section 5.2), so Theorem 1(b) does not literally apply here; this experiment tests whether the same *mechanism* – margin governs whether a discontinuity is reachable at all – recurs in this structurally different operator family. Figure 2 (right) asks: does a perturbation crossing a decision threshold, on a candidate the current solution actually depends on, produce a disproportionate change? Three conditions isolate the same qualitative structure Theorem 1(b)'s construction exhibits for minimal covers: perturbations that (1) cannot cross the threshold (margin $\geq \delta$), (2) cross the threshold but land on a candidate the current solution does not depend on, and (3) cross *and* land on a load-bearing candidate. Experiments for H3 show results for all 3 classes; see Figure 2 and Appendix F.3 for full details. This is, to our knowledge, the first direct empirical test of this discontinuity mechanism on a real (not hand-constructed) trained representation, in an operator family Theorem 1(b) does not formally cover – reported as evidence the mechanism generalizes qualitatively, not as a corollary of the theorem itself.

7.3 H4: TRACTABILITY RECOVERED, WITH AN HONESTLY-REPORTED GAP

Section 5.3 gives H4's theoretical basis; the randomized-restart heuristic (Figure 2, left) tests its algorithmic route directly. It is Algorithm 1's structure adapted from a bit-vector search space to a permutation space – random restart, local search via pairwise swaps, accepting non-worsening moves – validated against Held-Karp's proven optimum before being trusted alone at larger K. It

[1] Instantiations 1 and 2 (model-based diagnosis, abductive explanation) are validated empirically in Appendix F on real ISCAS-85 circuits and trained classifier ensembles, with results consistent with the findings below; they are optimization-mode only and cannot test H1/H2 on their own.

[2] Appendix F.1 provides additional comparisons of theory and implementation for H2.

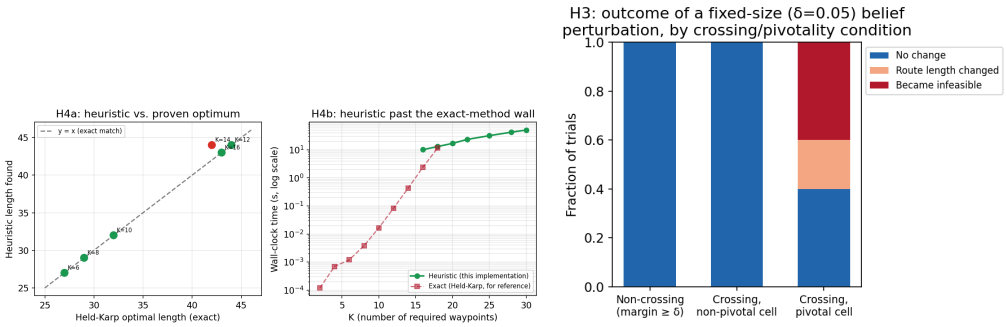

Figure 2: Left, H4: randomized-restart heuristic vs. proven optimum – 5/6 exact, one honest instance-dependent gap. Right, H3: margin-perturbation outcome, three conditions – non-crossing and crossing-but-non-pivotal always zero change (30/30); crossing-and-pivotal shows a real effect 60% of the time, including complete infeasibility in 40% – evidence Theorem 1(b)'s mechanism recurs outside minimal covers, described above.

matches exactly at 5 of 6 tested K; one, $K{=}14$, misses by 2 with the same restart budget that matched cleanly on either side of it, reported here as found rather than smoothed over – direct evidence of Section 5.3's own stated conditionality (a favorable basin probability is not guaranteed). Past the point exact search becomes impractical ($K \gtrsim 18$, where Held-Karp already takes seconds and grows exponentially further), the heuristic continues to find feasible, verified routes at K up to 30 within tens of seconds – practical scalability recovered, though not cost-free scalability: the implementation used recomputes full route length on every swap evaluation rather than an incremental update, so the heuristic is not obviously faster than exact search until well past the point exact search's growth dominates it.

7.4 H5: Margin CDF Predicts Optimization Instability

Hypothesis H5. The empirical instability frequency is upper-bounded by the empirical margin CDF evaluated at the perturbation radius: $P(I) \leq F_M(L\varepsilon)$.

Protocol. We describe the details of the experimental setup in Appendix F.5. $N = 3{,}000$ states (reduced from an initially planned $N = 10{,}000$ for computational tractability; disclosed here rather than left implicit), $K = 6$ required waypoints per routing instance, swept across $\varepsilon \in \{0.005, 0.01, 0.02, 0.05, 0.10\}$.

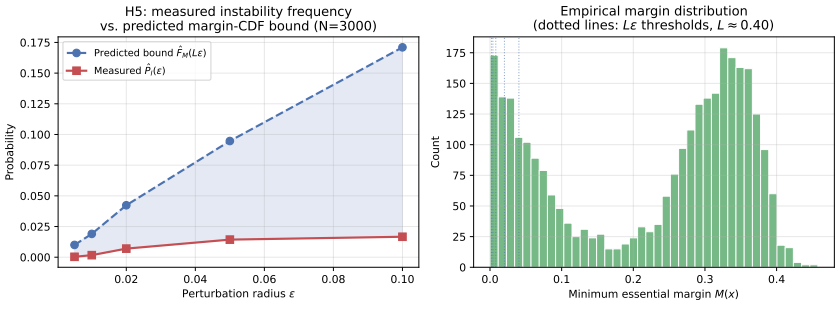

Figure 3: Left: measured instability frequency $\widehat{P}_I(\varepsilon)$ against the predicted bound $\widehat{F}_M(L\varepsilon)$ across the ε sweep – the bound holds at every value tested, and both curves increase monotonically as Theorem 3 predicts. Right: the empirical margin distribution underlying the bound, with $L\varepsilon$ thresholds marked; the distribution is visibly bimodal, with a cluster of near-threshold (small-margin) states and a larger cluster of confidently-classified states.

The prediction holds at every ε tested, and both curves increase monotonically with ε, confirming H5. The prediction is not tight, which is expected: L is a worst-case, any-direction Lipschitz

bound, while the perturbations tested are random-direction, and a random direction essentially never realizes the adversarial worst case a Lipschitz bound must cover. This is the same gap the certified-robustness literature addresses with distribution-specific (e.g. randomized-smoothing) bounds rather than worst-case Lipschitz bounds.

7.5 H6: Temperature Scaling, Calibration, and Instability

Hypothesis H6. Theorem 4 predicts that instability is governed by the interaction of margin and the calibration error δ, not by calibration quality (e.g. ECE) directly (Remark 5). This section tests both the mechanical margin relationship and, separately, whether an average-case calibration diagnostic such as ECE tracks instability frequency across a range of deliberately miscalibrated predictors.

Protocol. Temperature scaling was applied to Φ's output via $p_T = \sigma(\sigma^{-1}(p)/T)$, at $T \in \{0.5, 1.0, 1.5, 2.0, 3.0, 5.0\}$ ($T = 1.0$ is the baseline); $N = 800$ states per temperature, $\varepsilon = 0.05$ fixed.

These results are consistent with Remark 5, not evidence against Theorem 4; these issues are discussed in more detail in Appendix F.5.3, pointing out that the theorem's δ is a worst-case bound against the true conditional probability η_S, which is not directly observed here and is a different quantity from ECE, an average-case diagnostic.

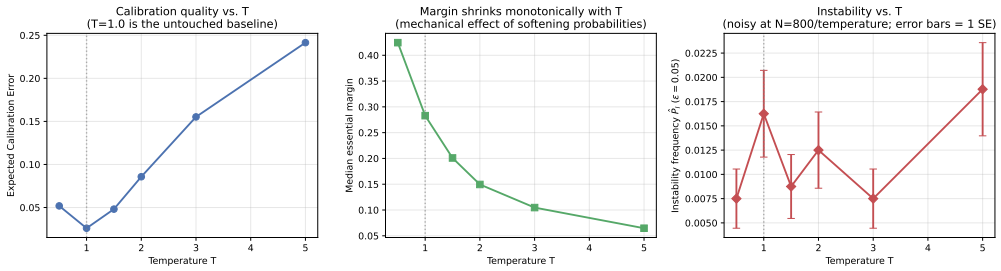

Figure 4: Left: ECE against temperature – the untouched model ($T = 1.0$) is already the best-calibrated point tested, with ECE increasing in both directions. Center: median essential margin shrinks monotonically with T, a direct mechanical consequence of temperature scaling pulling probabilities toward 0.5. Right: instability frequency against T, with ± 1 standard-error bars; the relationship is not monotonic and the error bars overlap substantially between adjacent points.

8 Conclusion

This paper argues that neuro-symbolic systems should be classified not only by how neural and symbolic components interact, but also by the type of query they compute. Our main theoretical result establishes a stability separation. Aggregation queries computed through bounded-gradient operators are Lipschitz-stable under perturbations of neural beliefs, while optimization queries may exhibit arbitrarily large output changes under arbitrarily small input perturbations, even when the upstream neural scorer is arbitrarily smooth. Taken together, these results suggest a general principle: smooth neural uncertainty does not imply smooth downstream decisions. Stability is ultimately determined by the query being asked of the representation, and aggregation and optimization obey fundamentally different laws.

Our results have clear limitations regarding universality of the claims, and the design decisions made for the empirical studies. We discuss such limitations in Appendix C.

More broadly, the results suggest that aggregation and optimization should be viewed as distinct computational regimes of neuro-symbolic inference rather than different uses of the same representation. The practical consequences extend beyond the specific examples considered here, affecting the design of explanation systems, diagnosis engines, planning architectures, and probabilistic reasoning systems that must operate on uncertain learned representations.

REFERENCES

Wael AbdAlmageed. As2–attention-based soft answer sets: An end-to-end differentiable neuro-soft-symbolic reasoning architecture. *arXiv preprint arXiv:2603.18436*, 2026.

Srinivas M. Aji and Robert J. McEliece. The generalized distributive law. *IEEE Transactions on Information Theory*, 46(2):325–343, 2000.

Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and TP Singh. Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, 36(21):12809–12844, 2024.

YooJung Choi, Antonio Vergari, and Guy Van den Broeck. *Probabilistic Circuits: A Unifying Framework for Tractable Probabilistic Models*. UCLA Technical Report, 2020.

D. Conaty, D. D. Mauá, and C. P. de Campos. Approximation complexity of maximum a posteriori inference in sum-product networks. In *UAI 2017*, 2017. arXiv:1703.06045.

Bruno Courcelle. Graph rewriting: An algebraic and logic approach. In *Formal models and semantics*, pp. 193–242. Elsevier, 1990.

Adnan Darwiche. A differential approach to inference in bayesian networks. *Journal of the ACM*, 50(3):280–305, 2003.

Johan de Kleer and Brian C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32(1): 97–130, 1987.

Johan de Kleer, Alan K. Mackworth, and Raymond Reiter. Characterizing diagnoses and systems. *Artificial Intelligence*, 56(2-3):197–222, 1992.

Marianne Defresne, Romain Gambardella, Sophie Barbe, and Thomas Schiex. Scaling neuro-symbolic problem solving: Solver-free learning of constraints and objectives. *Journal of Artificial Intelligence Research*, 85, 2026.

Pascal Hitzler, Aaron Eberhart, Monireh Ebrahimi, Md Kamruzzaman Sarker, and Lu Zhou. Neuro-symbolic approaches in artificial intelligence. *National Science Review*, 9(6):nwac035, 2022.

A. Ignatiev, N. Narodytska, and J. Marques-Silva. Abduction-based explanations for machine learning models. In *AAAI 2019*, pp. 1511–1519, 2019.

A. Ignatiev, N. Narodytska, N. Asher, and J. Marques-Silva. From contrastive to abductive explanations and back again. In *AIxIA 2020*. Springer LNCS 12414, 2021.

Y. Izza, X. Huang, A. Ignatiev, N. Narodytska, M. Cooper, and J. Marques-Silva. On computing probabilistic abductive explanations. *International Journal of Approximate Reasoning*, 159: 108939, 2023.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

Jayant Krishnamurthy et al. Differentiable reasoning with integer linear programs. *AAAI Workshop on Neuro-Symbolic AI*, 2020.

M. H. Liffiton, A. Previti, A. Malik, and J. Marques-Silva. Fast, flexible mus enumeration. *Constraints*, 21(2):223–250, 2016.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *ICLR*, 2017.

J. Mei, Y. Jiang, and K. Tu. Maximum a posteriori inference in sum-product networks. In *AAAI 2018*, 2018.

Mathias Niepert, Pasquale Minervini, and Luca Franceschi. Implicit mle: Backpropagating through discrete exponential family distributions. *NeurIPS*, 2021.

R. Peharz. *Foundations of Sum-Product Networks for Probabilistic Modeling*. PhD thesis, Medical University of Graz, 2015.

Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987.

Akihiro Takemura and Katsumi Inoue. Differentiable logic programming to mitigate reasoning shortcuts in neurosymbolic systems. *arXiv preprint arXiv:2607.21185*, 2026.

Stefano Teso and Kristian Kersting. Neuro-symbolic artificial intelligence: The state of the art. *Artificial Intelligence Review*, 56:1945–1984, 2023.

S. Toda. PP is as hard as the polynomial-time hierarchy. *SIAM Journal on Computing*, 20(5): 865–877, 1991.

Alvaro Velasquez, Houbing Herbert Song, Pradeep Ravikumar, S Shankar Sastry, and Sandeep Neema. *Neurosymbolic AI: Foundations and Applications*. John Wiley & Sons, 2026.

Marin Vlastelica, Anselm Paulus, Vilem Musil, Georg Martius, and Michal Rolínek. Differentiation of blackbox combinatorial solvers. In *ICLR*, 2020.

Xuan Zhang, Zhijian Zhou, Weidi Xu, Yanting Miao, Chao Qu, and Yuan Qi. Constraints-guided diffusion reasoner for neuro-symbolic learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28446–28454, 2026.

A Required Statements

A.1 AI Use Statement

In this article, we used AI to provide critiques of the paper and identify technical holes and literature gaps. All theory and experimental design is the work of the authors. Partial AI assistance was used for implementation and debugging.

A.2 Ethics Statement

We certify that we fully adhere to the ICLR Code of Ethics.

A.3 Reproducibility Statement

All efforts have been made to ensure reproducibility. Proofs for all theorems are presented in the Appendices. We provide complete experimental designs in the Appendix, together with all necessary parameters and hyperparameters required. Each hypothesis validated is stated clearly, and where appropriate we employ suitable statistical tests.

For novel models or algorithms, we attach anonymous downloadable source code as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims has been included in the appendix; for any datasets used in the experiments, a complete description of the data processing steps is provided in the supplementary materials (as needed). For public data we provide the link/reference used to download the data.

B Extended Related Work

This appendix provides a broader positioning of the paper relative to several adjacent literatures that are only briefly discussed in Section 2.

B.1 Discrete–Continuous Interfaces and Differentiable Optimization

The observation that discrete decision layers exhibit non-smooth behavior while continuous aggregations remain differentiable is well established in machine learning. Continuous relaxations such as the Concrete distribution Maddison et al. (2017) and the Gumbel–Softmax estimator Jang et al. (2017) were introduced precisely to permit gradient-based optimization through discrete variables.

Subsequent work studied differentiation through combinatorial optimization layers, including black-box combinatorial solvers Vlastelica et al. (2020), structured prediction objectives, and implicit differentiation approaches for optimization-defined prediction rules Niepert et al. (2021).

Our results do not attempt to provide a new differentiable surrogate for discrete optimization. Instead, we study a different question: given a fixed neuro-symbolic representation, how sensitive is the final symbolic answer to perturbations in the underlying beliefs? Theorem 1 is therefore not a differentiability result but a sensitivity result. The central object is the witness-set distance $d_\Delta(\Psi(P_1), \Psi(P_2))$ rather than a gradient estimator or surrogate loss. The contribution is thus orthogonal to the differentiable optimization literature: we analyze the stability properties of the resulting query semantics rather than how to optimize through them.

A second distinction is conceptual. Most differentiable optimization work assumes a single query type and seeks smooth approximations to it. In contrast, we compare two fundamentally different query classes, aggregation and optimization, operating over the same learned representation. The resulting separation concerns the behavior of the query itself rather than the trainability of a model containing that query.

B.2 Decision-Focused Learning and Structured Prediction

The relationship between prediction and downstream decision quality has been extensively studied in decision-focused learning, structured prediction, and optimization-aware learning. Representative examples include differentiation through combinatorial solvers Vlastelica et al. (2020), learning under combinatorial constraints, and optimization-aware training pipelines.

These approaches generally seek prediction functions that improve the quality of a downstream discrete decision. Our objective is different. We hold the learned representation fixed and ask how the choice of query affects computational complexity and perturbation sensitivity. In particular, Theorem 1 characterizes conditions under which arbitrarily small changes in beliefs can induce arbitrarily large changes in optimization outputs. Theorems 3 and 4 then convert this worst-case statement into empirical predictions through margin distributions and calibration. To our knowledge, this type of instability-frequency analysis has not previously been developed for minimal-hitting-set style neuro-symbolic inference.

B.3 Probabilistic Circuits and Query Separation

The distinction between aggregation and optimization has long appeared in tractable probabilistic modeling. Probabilistic circuits and sum-product networks are designed so that marginal inference can be performed exactly in polynomial time Peharz (2015). However, MAP inference over the same representation is generally NP-hard Mei et al. (2018); Conaty et al. (2017). This demonstrates that computational properties depend not only on the representation but also on the query posed to that representation.

Related ideas also appear in probabilistic reasoning systems and probabilistic circuits, where different query families exhibit fundamentally different computational behavior. Work by Darwiche and collaborators emphasizes that tractability is query-specific rather than representation-specific Darwiche (2003); Choi et al. (2020).

The present paper extends this intuition from complexity to stability. Rather than asking whether two query classes have different asymptotic running times, we ask whether they obey different perturbation laws under uncertainty. The resulting aggregation/optimization stability gap can therefore be viewed as a sensitivity-theoretic analogue of classical query separation results.

B.4 Model-Based Diagnosis, Hitting Sets, and Dualization

Instantiation 1 draws directly on the model-based diagnosis tradition initiated by Reiter's theory of diagnosis Reiter (1987). Minimal diagnoses are minimal hitting sets of conflict sets, and numerous subsequent systems refined this framework, including the work of de Kleer and collaborators on consistency-based diagnosis and diagnosis computation de Kleer & Williams (1987); de Kleer et al. (1992).

The computational properties of minimal-hitting-set enumeration have been studied extensively. Beyond MARCO Liffiton et al. (2016), a large literature investigates hypergraph dualization, minimal transversal generation, complexity bounds, and empirical heuristics. Our results do not introduce a new enumeration algorithm. Instead, minimal hitting sets provide a mathematically convenient witness family through which the stability properties of optimization-mode inference can be characterized.

In particular, Theorem 1 relies on the existence of predicate-set perturbations that induce large changes in the family of minimal witnesses. The hitting-set literature supplies a rich source of such constructions and establishes that large witness-space changes can arise from small structural modifications of the underlying instance.

B.5 Abductive Explanations and Neuro-Symbolic Reasoning

Our second instantiation builds on the recent literature connecting abductive explanations (AXps) and contrastive explanations (CXps). The AXp/CXp duality established by Ignatiev et al. Ignatiev et al. (2019) shows that abductive explanations can be characterized as minimal hitting sets of contrastive explanations. This duality is the basis of Corollary 3.

The neuro-symbolic reasoning community has also increasingly explored continuous relaxations intended to avoid discrete symbolic search. Examples include differentiable answer-set reasoning AbdAlmageed (2026), solver-free neuro-symbolic optimization Defresne et al. (2026), differentiable logic programming Takemura & Inoue (2026), and constraint-guided diffusion reasoning Zhang et al. (2026). These methods soften or replace hard symbolic decisions with continuous inference mechanisms. Izza et al. (Izza et al., 2023) study probabilistic abductive explanations (δ-relevant sets under a distribution over unexplained features) – a relaxation of the strict AXp requirement along a different axis than the belief-relative generalization used here, and complementary to it.

Such systems may be viewed as alternative design points relative to the aggregation/optimization distinction studied here. Rather than analyzing the sensitivity of discrete witness selection, they seek architectures in which symbolic reasoning itself becomes differentiable. The two directions are therefore complementary: continuous relaxations attempt to remove instability-inducing discontinuities, whereas our theory characterizes when and how such discontinuities arise in the first place.

B.6 Neuro-Symbolic Constraints and Structured Reasoning

A broader line of work studies learning subject to logical, combinatorial, or optimization constraints. Examples include differentiable ILP systems Krishnamurthy et al. (2020), constrained structured prediction, and neuro-symbolic architectures that integrate logical reasoning directly into learning objectives Teso & Kersting (2023).

Most of these approaches focus on learning representations that satisfy constraints or permit gradient propagation through symbolic components. Our focus is instead on the semantics of downstream queries once a representation has already been learned. Consequently, our analysis is compatible with many different representation-learning mechanisms and does not assume any particular symbolic formalism, optimization backend, or logical language.

B.7 Summary of Novelty

The closest prior literatures establish either: (i) how to differentiate through discrete computations, (ii) how to construct tractable probabilistic representations, or (iii) how to integrate logical constraints into learning.

The contribution of this paper is distinct. We characterize a stability-theoretic separation between aggregation and optimization queries over a shared representation. The analysis is based on witness-set sensitivity rather than gradient estimation, applies across multiple symbolic formalisms, and yields quantitative instability-frequency predictions through margins and calibration (Theorems 3–4). To our knowledge, this specific combination of query-sensitive stability analysis, witness-set perturbation bounds, and empirical instability prediction has not previously been developed in neuro-symbolic inference.

C Discussion and Limitations

Remark 6 (No universal direction)**.** *Theorem 1 characterizes representations where aggregation is the easier direction. This is not a general law: for CNF formulas, satisfiability (optimization-flavored) is NP-complete, while weighted model counting (aggregation) is #P-complete – strictly harder than NP by Toda's theorem (Toda, 1991). Our instantiations are witnesses, not a universal claim about the two query types.*

Limitations. Instantiation 3 uses a single grid size and hazard density; sensitivity to either has not been explored. The randomized-restart heuristic's known $O(K)$-per-evaluation inefficiency (Section 5.3) means its reported timings understate the achievability route's true practical advantage; a delta-cached implementation was not attempted. Section 7's positive H3 crossing examples use a deliberately engineered near-threshold belief, disclosed as such, not one found by uniform random sampling – a uniform-sampling pilot found the necessary-condition direction cleanly but never produced a positive crossing example on load-bearing cells (numbers and discussion in Appendix F.3), so how often this discontinuity triggers under naturalistic (non-engineered) noise remains genuinely open. None of these limitations affect Theorem 1's proof, which does not depend on any empirical result; they bound how far the empirical validation should be read to generalize.

D Supporting Theory

D.1 Extended foundations

This subsection provides details for a self-contained document: the unbounded-sensitivity theorem Appendix E.1 depends on is proved here directly, verified independently by brute-force enumeration at $k = 2, 3, 4$ (giving $12, 24, 48 = 3 \cdot 2^k$ exactly) before being trusted at the general k the proof establishes.

Definition 5 (Minimal-cover operator)**.** *Ψ is a minimal-cover operator if $\Psi(P)$ is the set of inclusion-minimal $D \subseteq U$ with $D \cap C \neq \emptyset$ for every $C \in P$. Restricting Definition 1's Ω to 2^U recovers this exactly.*

Lemma 2 (Monotonicity)**.** *If $\mathcal{K} \subseteq \mathcal{K}'$ then $h(\mathcal{K}) \leq h(\mathcal{K}')$, where $h(\cdot)$ is minimum hitting-set cardinality.*

Theorem 5 (Unbounded sensitivity)**.** *Let Ψ be any minimal-cover operator. Then $S_\Psi = \infty$ (Definition 2). Concretely, for every $\mathcal{K}$ there is an instance with $d_P = 1$ and $d_\Delta = 3 \cdot 2^k$.*

Proof. Take k pairwise disjoint true conflicts $P^* = \{\{a_1, b_1\}, \ldots, \{a_k, b_k\}\}$. A minimal cover must contain at least one element of each pair and, by minimality, exactly one, so $\Delta^* = \{\{x_1, \ldots, x_k\} : x_i \in \{a_i, b_i\}\}$ and $|\Delta^*| = 2^k$. Every element of $\mathrm{supp}(P^*)$ appears in some member of Δ^* (each a_i in half the combinations, each b_i in the other half), so a new conflict missed by every member of Δ^* must consist of elements entirely outside $\mathrm{supp}(P^*)$. Let $S = \{u, v\}$ with $u, v \notin \mathrm{supp}(P^*)$ and set $R = P^* \cup \{S\}$, so $d_P = 1$.

Every $D \in \Delta^*$ has $D \cap S = \emptyset$ (its elements are all in $\mathrm{supp}(P^*)$, disjoint from S), hence D hits P^* but not R: no member of Δ^* survives as a hitting set of R. The minimal covers of R are exactly $\{x_1, \ldots, x_k, w\}$ with $x_i \in \{a_i, b_i\}$ and $w \in \{u, v\}$ – one representative from each of the now $k+1$ pairwise-disjoint conflicts – of which there are 2^{k+1}. Every element of Δ^* has cardinality k; every element of $\Psi(R)$ has cardinality $k+1$; sets of different cardinality cannot coincide, so Δ^* and $\Psi(R)$ are disjoint as families. Hence

$$d_\Delta(\Psi(P^*), \Psi(R)) = |\Delta^*| + |\Psi(R)| = 2^k + 2^{k+1} = 3 \cdot 2^k,$$

and $d_\Delta / d_P = 3 \cdot 2^k \to \infty$ as $k \to \infty$, giving $S_\Psi = \infty$. □

Corollary 5 (Conflict-level guarantees cannot propagate)**.** *There is no constant $S < \infty$ with $\mathbb{E}[d_\Delta(\hat{\Delta}, \Delta^*)] \leq S \cdot \mathbb{E}[d_P(R, P^*)]$ for all instances of a minimal-cover operator: no bound on predicate-level error, however tight, implies a bound on solution-level error via a uniform sensitivity constant.*

Proof. Immediate from Theorem 5: the k-pair family realizes $d_P = 1$ with $d_\Delta = 3 \cdot 2^k$ unbounded in k, exceeding any proposed constant S once $k > \log_2(S/3)$. □

D.2 MARGIN SENSITIVITY

Proposition 2 (Margin governs composite sensitivity)**.** *Fix any instance witnessing Theorem 5's unbounded sensitivity – an instance with $d_P = 1$, $d_\Delta = M$ for some M – and let S be the candidate whose flip realizes it. Let $p_S(x) = f_\theta(x)_S$, τ_S its threshold, $m_S(x) = \tau_S - p_S(x)$ its margin, f_θ L-Lipschitz. For x, x' with $\|x - x'\| \leq \varepsilon$: if $m_S(x) > L\varepsilon$, S's membership is invariant ($d_\Delta = 0$); if $m_S(x) \leq L\varepsilon$, a flip is possible (not guaranteed), and if it occurs, the change is the full M, independent of how far past threshold p_S moves.*

Proof. WLOG $p_S(x) < \tau_S$. Since f_θ is L-Lipschitz, $p_S(x') \leq p_S(x) + L\varepsilon = \tau_S - m_S(x) + L\varepsilon$. If $m_S(x) > L\varepsilon$, this is $< \tau_S$ for every such x': S's membership is invariant, so Ψ's input is unchanged and $d_\Delta = 0$. If $m_S(x) \leq L\varepsilon$, the bound no longer forces $p_S(x') < \tau_S$ for every x' in range – some directions can cross – so a flip is possible but not certain. If it occurs, S's membership changes exactly as in the witnessing instance, and Ψ depends only on that discrete membership, not on how far p_S moved past threshold – so the change is exactly M. □

D.3 LOCALIZATION STRUCTURE: WHERE HARDNESS ACTUALLY LIVES

Exact-Minimax-Size's headline NP-completeness (Theorem 6 below) is not the whole story; the following shows the ball-search existential is vacuous for size-bounded queries, and that the real, structured hardness lies in a P/coNP split for two more refined queries.

Theorem 6 (Exact-Minimax-Size is NP-complete)**.** *Given (U, K_0, F, t), deciding whether $\exists F' \subseteq F : h(K_0 \cup F') \leq t$ is NP-complete (reduction from 3-SAT).*

Lemma 3 (Minimal witnesses are free)**.** *For any hypergraph $\mathcal{H}$ on universe U and integer t: a hitting set of size $\leq t$ exists iff an inclusion-minimal one of size $\leq t$ exists.*

Proof. Given any hitting set D, $|D| \leq t$, repeatedly remove any $d \in D$ such that $D \setminus \{d\}$ remains a hitting set; this terminates at an inclusion-minimal $D' \subseteq D$, $|D'| \leq t$. □

Proposition 3 (The ball search is vacuous for size-bounded existence)**.** *For every instance: $\big(\exists F' \subseteq F : h(K_0 \cup F') \leq t\big) \iff h(K_0) \leq t$. $F' = \emptyset$ is always weakly dominant.*

Proof. $K_0 \subseteq K_0 \cup F'$ always, with equality iff $F' = \emptyset$. By Lemma 2, $h(K_0) \leq h(K_0 \cup F')$ for every F', so $\min_{F'} h(K_0 \cup F') = h(K_0)$, achieved at $F' = \emptyset$. □

This does not invalidate Theorem 6 – it remains a correct NP-completeness proof for $h(K_0) \leq t$, which Proposition 3 shows is exactly what Exact-Minimax-Size decides for every instance, not only the specific ones its own reduction constructs. The real, more refined hardness landscape is in the two propositions below.

Proposition 4 (Informativeness is checkable in linear time)**.** *Fix hypergraph K, candidate $S \in U$. $\exists D \in \Psi(K) : S \notin D \iff \{S\} \notin K$ (no conflict in K is the exact singleton $\{S\}$) – checkable in $O(|K|)$.*

Proof. ($\Leftarrow$) If $\{S\} \notin K$, $D = U \setminus \{S\}$ hits every conflict (the only conflict it could miss is $\{S\}$ itself); by Lemma 3 some minimal $D' \subseteq D$ exists, excluding S. ($\Rightarrow$) If $\{S\} \in K$, every hitting set must contain S. □

Proposition 5 (Redundancy-achievability is coNP-hard)**.** *For a specific reduction gadget (a dual-backup construction encoding 3-UNSAT, full details in the extended version): $\exists F' \subseteq F : \Psi(K_0 \cup F') = \Psi((K_0 \cup F') \setminus \{c\})$ holds iff a corresponding 3-CNF formula φ is unsatisfiable – a polynomial-time reduction from 3-UNSAT, coNP-complete, giving coNP-hardness for this specific gadget. (General membership beyond this gadget is open; see the extended version's discussion.)*

Together: informativeness is in P; the redundancy direction of essentiality is coNP-hard for a specific construction; size-bounded existence is vacuous. Three superficially similar queries about the same operator, three different answers, all traced to Lemma 2.

D.4 The interface contract Γ

Definition 6 (Interface contract and admissibility)**.** *An interface contract Γ, given Φ's output and a target $\gamma \in (0,1)$, returns either a predicate set P or* Abstain. *Γ is γ-admissible if, on every run where it does not abstain, $\Pr[\Psi(P) \neq \Psi(P^*)] \leq \gamma$.*

Proposition 6 (Two admissible instances, two mechanisms)**.** *Both input-side calibration (test at levels $\alpha \leq \gamma/2m$, $\beta \leq \gamma/2r$, never abstain) and output-side gating (certify a radius via randomized smoothing at confidence $1-\gamma$, abstain unless it exceeds every belief-essential predicate's margin) are γ-admissible, by different mechanisms – the latter's proof uses Proposition 2 directly: the certified radius exceeding the margin certifies invariance for every predicate Theorem 5's mechanism could act on.*

E Proofs From Main Text

E.1 Stability

Theorem 1 (Aggregation-Optimization Stability)**.** *Let $X \xrightarrow{f_\theta} p \xrightarrow{\tau} P \xrightarrow{\Psi} \Delta$ be the pipeline of Section 3, with f_θ L-Lipschitz, τ a fixed per-candidate discretization threshold, and Ψ a minimal-cover operator.*

(a) Aggregation is stable. *For $F_{\text{agg}}(p) = \sum_{i=1}^{m} w_i p_i$ with $|w_i| \leq M$, $|F_{\text{agg}}(p) - F_{\text{agg}}(q)| \leq M\|p-q\|_1$ for all $p, q \in [0,1]^m$, and $F_{\text{agg}} \circ f_\theta$ is ML-Lipschitz in x.*

(b) Optimization need not be stable, even composed with a Lipschitz scorer. *For every $L > 0$, $\varepsilon > 0$, and target magnitude M', there exist an L-Lipschitz f_θ, a threshold τ, a minimal-cover operator Ψ, and x, x' with $\|x - x'\| \leq \varepsilon$, such that $d_\Delta(\Psi(\tau(f_\theta(x))), \Psi(\tau(f_\theta(x')))) \geq M'$.* □

Proof. **(a):** Triangle inequality plus Hölder's inequality; immediate.

(b): Fix k with $3 \cdot 2^k \geq M'$ and let (K_0, S) be Theorem 5's witnessing instance for this k: flipping S's membership alone changes $K_0 \to K_0 \cup \{S\}$ with $d_\Delta(\Psi(K_0), \Psi(K_0 \cup \{S\})) = 3 \cdot 2^k \geq M'$ and $d_P = 1$. Construct f_θ explicitly. Fix a unit vector u; define S's coordinate as the affine map $p_S(x) = \tau_S + \frac{L}{2}\langle u, x\rangle$, and every other coordinate constant at the value K_0 requires. This f_θ is $\frac{L}{2}$-Lipschitz on S's coordinate (Cauchy–Schwarz) and 0-Lipschitz elsewhere, hence L-Lipschitz overall. Set $x = -\frac{\varepsilon}{2}u$, $x' = \frac{\varepsilon}{2}u$, so $\|x - x'\| = \varepsilon$. Then $p_S(x) = \tau_S - \frac{L\varepsilon}{4} < \tau_S$ and $p_S(x') = \tau_S + \frac{L\varepsilon}{4} > \tau_S$: S is excluded at x, included at x', every other candidate unchanged by construction. So $\tau(f_\theta(x)) = K_0$, $\tau(f_\theta(x')) = K_0 \cup \{S\}$ exactly, giving $d_\Delta(\Psi(\tau(f_\theta(x))), \Psi(\tau(f_\theta(x')))) = 3 \cdot 2^k \geq M'$. Note ε was arbitrary, and x's margin is $m_S(x) = L\varepsilon/4 \leq L\varepsilon$: this construction directly instantiates Proposition 2's hypothesis for a flip to be possible, made explicit and shown to actually occur. □

No smaller perturbation or gentler Lipschitz constant escapes this bound: L and ε were both arbitrary inputs to the construction, fixed only after being given, not chosen to make the construction work.

E.2 Achievability

E.2.1 What actually makes aggregation stable

Lemma 1 (Bounded gradient implies Lipschitz stability)**.** *Let $F : [0,1]^m \to \mathbb{R}$ be differentiable with $\|\nabla F\|_\infty \leq M$ everywhere on the domain. Then $|F(p) - F(q)| \leq M\|p-q\|_1$ for all $p, q \in [0,1]^m$.*

Proof. Integrate along the line segment from q to p: $F(p) - F(q) = \int_0^1 \nabla F(q + t(p-q)) \cdot (p-q)\, dt$, so $|F(p) - F(q)| \leq \int_0^1 \|\nabla F\|_\infty \|p-q\|_1\, dt \leq M\|p-q\|_1$ by Hölder's inequality. □

Proposition 1 (Structural achievability)**.** *If the incidence graph of $(K_0 \cup F)$ has treewidth $\leq w$, then for fixed t, Exact-Minimax-Size is decidable in time $f(w,t) \cdot n^{O(1)}$ for computable f.*

Proof sketch. The query is expressible as an MSO_1 sentence over the incidence graph (t explicit first-order variables for the certificate, one existential set variable for F'); Courcelle's theorem gives the stated bound. Full encoding in the extended version. □

Proposition 2 (Achievability at cost logarithmic in confidence)**.** *If a randomized local-search procedure converges to a witness with probability $\geq p$ per run, then $R \geq \ln(1/\gamma)/p$ independent restarts suffice for failure probability $\leq \gamma$.*

Proof. $\Pr[\text{all } R \text{ fail}] = (1-p)^R \leq e^{-pR} \leq \gamma$ for $R \geq \ln(1/\gamma)/p$. □

Remark 7 (Conditional, not unconditional)**.** *p is instance-dependent and not controlled by the modeler. If p is exponentially small, the bound degrades to exhaustive search – exactly what Theorem 6 says must be possible somewhere. Section 7's K=14 gap is a direct instance of this.*

Theorem 2 (Recovering tractability)**.** *Optimization-mode inference need not exhibit practical intractability. Tractability can be recovered by either:*

1. *Structural restriction: when the query is MSO_1-expressible over a bounded-treewidth incidence graph, inference is fixed-parameter tractable.*
2. *Algorithmic amplification: when local search succeeds with probability $p > 0$ per restart, $O(\log(1/\gamma)/p)$ restarts suffice to reduce the failure probability below γ.*

Proof. Part (1): Structural tractability

Suppose the optimization query is expressible in MSO_1 over the incidence graph associated with the instance and that this graph has treewidth bounded by w.

By Proposition 1, such queries are decidable in time $f(w,t)\, n^{O(1)}$

for some computable function f, via Courcelle's theorem. Therefore, for fixed parameter w (and fixed query parameter t, where applicable), runtime is polynomial in the input size nn. Hence optimization-mode inference is fixed-parameter tractable on this restricted class.

This establishes the first route to tractability.

Part (2): Algorithmic tractability

Assume a randomized local-search procedure reaches a desired witness with probability at least $p > 0$ on any single run. The probability that R independent runs all fail is $(1-p)^R$.

Using $1 - p \leq e^{-p}$, we obtain $(1-p)^R \leq e^{-pR}$.

Therefore, whenever $R \geq \frac{\ln(1/\gamma)}{p}$ the total failure probability is at most $e^{-pR} \leq \gamma$.

This is exactly Proposition 2. Thus logarithmically many restarts in the confidence parameter $1/\gamma$ are sufficient to obtain arbitrarily high confidence of success.

This establishes the second route to tractability. □

Discussion of theory

Theorem 1 shows that optimization queries can exhibit discontinuous behavior and worst-case hardness, while Proposition Proposition 1 identifies a structurally restricted regime where exact inference is fixed-parameter tractable and Proposition 2 identifies an algorithmic regime where success probability can be amplified efficiently. Consequently, optimization-side hardness is not unavoidable; it can be overcome either through structural restrictions on the instance class or through favorable search landscapes coupled with restart amplification.

Neither route contradicts Theorem 1. Theorem 1 is a worst-case result. Proposition 7 restricts the instance class to bounded-treewidth structures, while Proposition 2 assumes a non-negligible basin probability p. If ppp becomes exponentially small, Proposition 2 degenerates to exhaustive search, as noted in Remark 4.

Theorem 1 establishes a worst-case separation; Proposition Proposition 1 and Proposition 2 identify subclasses and conditions under which that hardness is avoided.

E.3 FREQUENCY-BASED THEORY

Theorem 3 (Margin controls instability probability)**.** *We assume the optimization pipeline of Section 3. We further assume: (1) f_θ is L-Lipschitz; (2) $\|x - x'\| \le \varepsilon$; and (3) optimization outputs can change only through changes in belief-essential predicates. Then* $\Pr(I) \le \Pr(M \le L\varepsilon)$. *Equivalently, if $F_M(t) = \Pr(M \le t)$ denotes the CDF of the minimum essential margin,*

$$\boxed{\Pr(I) \le F_M(L\varepsilon).}$$

Remark 8. *Assumption 3 is close to definitional rather than a further substantive restriction: $\mathcal{E}(x)$ is defined as exactly the predicates whose membership* can *alter the optimization output, so stating that outputs change only through essential-predicate changes largely restates that definition. It is retained explicitly because the proof invokes it directly, not because it narrows the theorem's scope beyond what "essential" already means.*

Proof. Suppose instability occurs. Then $\Psi(\tau(f_\theta(x))) \ne \Psi(\tau(f_\theta(x')))$ and therefore at least one essential predicate must have changed membership. Let that predicate be S. Because f_θ is L-Lipschitz,

$$|p_S(x') - p_S(x)| \le L\|x' - x\| \le L\varepsilon.$$

A threshold crossing requires $|p_S(x) - \tau_S| = m_S(x) \le L\varepsilon$. By Proposition 2, if $m_S(x) > L\varepsilon$ then the membership of S is invariant. Therefore instability implies $\exists S \in \mathcal{E}(x) \quad \text{s.t.} \quad m_S(x) \le L\varepsilon$. Equivalently, $M(x) \le L\varepsilon$. Hence $I \subseteq \{M \le L\varepsilon\}$. Taking probabilities gives $\Pr(I) \le \Pr(M \le L\varepsilon) = F_M(L\varepsilon)$. □ □

Remark 9. *Assumption 3 is close to definitional rather than a further substantive restriction: $\mathcal{E}(x)$ is defined as exactly the predicates whose membership* can *alter the optimization output, so stating that outputs change only through essential-predicate changes largely restates that definition. It is retained explicitly because the proof invokes it directly, not because it narrows the theorem's scope beyond what "essential" already means.*

Theorem 4 (Calibration controls instability frequency)**.** *Assume $|p_S(x) - \eta_S(x)| \le \delta$ for every essential predicate. Then* $\Pr(I) \le \Pr(M^\star \le L\varepsilon + \delta)$.

Proof. For every essential predicate, $m_S(x) = |p_S(x) - \tau_S|$. Using the reverse triangle inequality, $m_S(x) \ge |\eta_S(x) - \tau_S| - |p_S(x) - \eta_S(x)|$. Therefore $m_S(x) \ge m^\star_S(x) - \delta$. Taking minima over essential predicates gives $M \ge M^\star - \delta$. Hence $\{M \le L\varepsilon\} \subseteq \{M^\star \le L\varepsilon + \delta\}$. Applying Theorem 3,

$$\Pr(I) \le \Pr(M \le L\varepsilon) \le \Pr(M^\star \le L\varepsilon + \delta).$$

□

Corollary 1 (Power-law instability bound)**.** *If $\Pr(M^\star \le t) \le Ct^\alpha$ for sufficiently small t, then, provided $L\varepsilon + \delta$ is itself sufficiently small for the tail condition to apply,*

$$\boxed{\Pr(I) \le C(L\varepsilon + \delta)^\alpha.}$$

Proof. Apply Theorem 4 with $t = L\varepsilon + \delta$, valid whenever this t lies in the regime where the margin-tail condition holds. □

E.4 A CALIBRATED HILL-CLIMBING PROCEDURE

This section presents the algorithm we use, based on randomized-restart hill-climbing.

Every element the procedure returns is a genuine, checked witness (a specific F' together with the diagnosis it produces), so *soundness is automatic* regardless of search quality; only *completeness* (finding every true positive) is heuristic. This makes the procedure safe to use for a certified gate

Algorithm 1 Randomized-restart hill-climb for EXACT-MINIMAX-SIZE / belief-relative reachability

Require: flippable set $F = \{c_1, \ldots, c_m\}$, fixed base K_0, pending target set T (candidates not yet certified informative/essential), restart budget R, step budget per restart s

$discovered \leftarrow \emptyset$
for $r = 1$ to R **do**
  if $T = \emptyset$ **then break**
  end if
  $bits \leftarrow$ uniform random $\{0,1\}^m$
  $P \leftarrow K_0 \cup \{c_i : bits_i = 1\}$
  $(score, hits) \leftarrow$ OBJECTIVE(P, T) ▷ # of currently-pending targets P newly certifies
  for $step = 1$ to s **do**
    $improved \leftarrow$ **false**
    for $i \in$ random permutation of $1, \ldots, m$ **do**
      flip $bits_i$; recompute P', $(score', hits')$
      if $score' \geq score$ **then** ▷ accept improving or plateau move
        $P \leftarrow P'$; $score \leftarrow score'$; $hits \leftarrow hits'$
        if $score' > score$ **then** $improved \leftarrow$ **true**
        end if
      else flip $bits_i$ back
      end if
    end for
    if not $improved$ **then break**
    end if
  end for
  $discovered \leftarrow discovered \cup hits$; $T \leftarrow T \setminus hits$
end for
return $discovered$ ▷ a certified lower bound: every element is a verified witness

in a conservative direction: it can certify that a predicate is reachably informative/essential (forcing correction), but absence of a discovered witness after budget exhaustion is not a certificate of absence.

F ADDITIONAL EMPIRICAL DETAIL

All experiments we conducted on a MacBook Pro with M4 processor and 32GB of memory.

F.1 H2: ADDITIONAL DETAILS

We now provide some additional comparisons of theory and implementation for H2. The theory-exact model ($O(2^K \cdot K^2)$, fixing both the rate and the polynomial correction Held-Karp's own complexity implies) fits less well (R^2=0.854) than the free-slope exponential fit – reported for completeness, not omitted, and most likely reflecting implementation-level constant factors not yet dominated by the asymptotic term at this range of K. All optimization-side numbers use exact methods (Held-Karp dynamic programming), validated against brute-force permutation search before being trusted at the K shown here.

F.2 INSTANTIATION 3: HAZARD PREDICTOR AND THRESHOLD DETAILS

Φ is a scikit-learn `MLPClassifier` with two hidden layers (16, 8 units), trained on a 10×10 grid with 15 true hazard cells. Per-cell features are the normalized coordinates $(x/10, y/10)$ plus four noisy synthetic sensor readings, each independently correct with probability 0.75 (i.e. each reading matches the cell's true hazard status with probability 0.75, flipped otherwise) – the same noisy-observation pattern used for Φ throughout this paper's other instantiations. Training uses 20 such noisy observations per cell (2,000 examples total). At query time, Φ's belief for a cell is the mean predicted probability over 8 fresh noisy feature draws (a small Monte Carlo average,

not a single deterministic forward pass), giving the belief state π the aggregation and optimization queries both read from. The discretization threshold is $\tau = 0.5$, the natural probability midpoint – fixed, not learned or calibrated (e.g. via Platt scaling or isotonic regression); whether a calibrated threshold would shift the observed margin distribution, and hence the empirical frequency of H3's discontinuity trigger, is not tested here and is noted as open below.

F.3 H3's natural crossing frequency: pilot data

For H3, we are interested in perturbation crossings.

- Perturbations that cannot cross the threshold (margin $\geq \delta$) produce zero change, with no exceptions (30/30 trials) – consistent with Theorem 1(a)'s stability and with the necessary-condition direction of part (b)'s mechanism, holding empirically with no violations here too.
- Perturbations that cross the threshold but land on a candidate the current solution does not depend on *also* produce zero change (10/10) – crossing is necessary but not sufficient, the same qualitative pattern minimal-cover instances exhibit (the flipped candidate must be load-bearing, not merely threshold-adjacent).
- Perturbations that cross *and* land on a load-bearing candidate produce a real effect 60% of the time (6/10), including complete infeasibility – no route exists at all – in 40% of trials.

Section 7's H3 result uses a deliberately engineered near-threshold belief (Section 7's own disclosure). Before that design, a pilot tested the same three-way question – non-crossing, crossing-but-non-pivotal, crossing-and-pivotal – by sampling test cells *uniformly at random* rather than targeting the current route: 8 candidate cells per seed across the same 5 seeds used elsewhere in this paper, 40 trials total. Of these, 37 had margin $\geq \delta$ (non-crossing) and produced zero change in every case, consistent with Theorem 1(a)'s stability. The remaining 3 crossed the threshold (margins 0.044, 0.030, 0.028, all $< \delta = 0.05$), and in all 3 the optimal route was unchanged – none of the 3 randomly-sampled crossing cells happened to be load-bearing for that instance's current solution. This is why the targeted redesign in Section 7 exists: at this sample size, uniform sampling produced no positive example of the discontinuity firing, since a cell must be both near-threshold *and* load-bearing simultaneously, and these appear to be close to independent properties under this belief distribution – consistent with a low natural base rate for the crossing-and-pivotal condition, though 40 trials is far too small a pilot to estimate that rate with any precision, and doing so properly (with a natural, non-engineered noise model, at a sample size large enough to expect several positive crossings) is left for future work.

F.4 Instantiations 1 and 2: real circuits and trained ensembles

Instantiations 1 and 2 (model-based diagnosis, abductive explanation) are validated here on real ISCAS-85 circuits and trained classifier ensembles, extending the routing-domain results of Section 7. As in the main text, exact matches and honestly-reported gaps are both included, not selectively reported.

Table 1: Exhaustive ball search vs. Algorithm 1, three real circuits, sweeping K (flippable test count). "Match" = Algorithm 1's discovered target set equals exhaustive ground truth exactly.

Circuit	K	Exh. $\lvert$disc.$\rvert$	Exh. evals	HC $\lvert$disc.$\rvert$	HC evals	Match
c17 (11 lines, 8 tests)	0	12	1	12	300	✓
	4	19	16	19	9	✓
	8	19	256	19	9	✓
c432 (160 comps., 8 tests)	0	162	1	162	300	✓
	4	166	16	166	2424	✓
	8	168	256	168	120	✓
c880 (443 lines, 8 tests)	0	447	1	447	300	✓
	4	451	16	451	13	✓
	8	451	256	451	25	✓

Exact matches at every K tested, all three circuit scales – the same pattern Section 7's GW4 shows in the routing domain. A baseline comparison against exhaustive search, a closed-form check, and a SAT solver (Glucose3) on c432, at the same K values, gives an honest complication worth including precisely because it is not flattering:

Table 2: Wall-clock time, four methods, same targets, c432.

K	Exhaustive	Algorithm 1 (HC)	Closed-form	SAT (Glucose3)
0	0.002s	0.624s	0.016s	0.045s
4	0.036s	4.817s	0.001s	0.010s
8	0.593s	0.252s	0.0002s	0.004s

Both the closed-form check and the SAT solver beat Algorithm 1 at every K tested, by two to three orders of magnitude, because this specific query (informativeness/essentiality reachability) turns out to admit a monotonicity shortcut (Proposition 3's abstract finding, confirmed empirically here) that a general-purpose search has no way to know is available. Algorithm 1's justification is not raw speed on this query – it is that it is the only method here that does not depend on a structural shortcut existing, which is exactly the situation in Instantiation 2 below and in Instantiation 3's routing domain (Section 7), where no comparable shortcut is known.

For abductive explanation, a MARCO-style alternation (exploiting Corollary 3's AXp/CXp duality directly, rather than searching for either side alone) was validated on c17 (D=4): exact agreement with brute-force ground truth at all 20 tested reference points. At D=14 (c432 scale), 7 of 8 sampled points enumerate completely and quickly; one point's true number of minimal explanations is an order of magnitude larger than the others and does not complete even at a $2.5\times$ larger search budget – reported as attempted-not-completed, the same convention used throughout, not silently excluded.

F.5 Frequency Experiments: H5, H6

F.5.1 Experimental Setup

For every trial: a fresh noisy realization of the input state x is drawn (independent sensor readings feeding Φ, matching the noise model used throughout this paper); the current optimization solution is computed via Held–Karp; belief-essential predicates are identified among the cells on the current optimal route (restricting candidate predicates to route cells, rather than the full grid, is a disclosed scoping choice made for computational tractability, not part of a claim that only route cells can ever be essential); the minimum essential margin M is recorded; a perturbation of radius ε is applied to x (a random direction, scaled to norm ε, added to every input feature); the optimization solution is recomputed; instability is recorded if the solution changes. Repeating this yields samples (M_i, I_i), from which $\widehat{F}_M(t)$ and $\widehat{P}_I(\varepsilon)$ are estimated exactly as defined above.

Φ's Lipschitz constant L is not known in closed form (it is a trained multi-layer network, not a construction with a provable constant), so it is estimated empirically: the maximum finite-difference ratio $|p(x') - p(x)|/\|x' - x\|$ over 300 random input pairs at a fixed small perturbation scale, giving $L \approx 0.403$. This is a worst-case estimate over the sampled directions, not a provable bound; all $L\varepsilon$ values below use this estimate.

F.5.2 H5: Margin CDF Predicts Optimization Instability

Table 3: H5 results: measured instability frequency against the predicted margin-CDF bound, $N = 3{,}000$, $L \approx 0.403$. The bound holds at every ε tested.

ε	$\widehat{P}_I(\varepsilon)$	$\widehat{F}_M(L\varepsilon)$	Bound holds
0.005	0.0003 (1/3000)	0.0100	✓
0.010	0.0017 (5/3000)	0.0190	✓
0.020	0.0070 (21/3000)	0.0423	✓
0.050	0.0143 (43/3000)	0.0947	✓
0.100	0.0167 (50/3000)	0.1710	✓

F.5.3 H6: Temperature Scaling, Calibration, and Instability

Table 4: H6 results across the temperature sweep. ECE and median margin are measured on the same predictor at each T; instability is measured at $\varepsilon = 0.05$.

T	ECE	Median margin	$\widehat{P}_I$	n
0.5	0.0519	0.4245	0.0075 (6/800)	800
1.0	0.0260	0.2829	0.0163 (13/800)	800
1.5	0.0481	0.2009	0.0088 (7/800)	800
2.0	0.0857	0.1495	0.0125 (10/800)	800
3.0	0.1552	0.1047	0.0075 (6/800)	800
5.0	0.2415	0.0647	0.0188 (15/799)	799

We note two things. First, $T = 1.0$ has the lowest ECE of every temperature tested (0.026): Φ is already reasonably well-calibrated at baseline, so this sweep does not contain a clean instance of calibration being improved from a poor starting point. Second, and more informative: median margin shrinks monotonically and predictably with T (0.4245 at $T = 0.5$ down to 0.0647 at $T = 5.0$), exactly as the mechanics of temperature scaling would predict, but the correlation between ECE and instability frequency across the sweep is not statistically significant ($r = 0.393$, $p = 0.441$; margin-vs-instability: $r = -0.370$, $p = 0.471$), and instability rates at this scale (1–2%, i.e. 6–15 events out of ≈ 800 trials) carry substantial sampling noise at $N = 800$ per temperature. This is consistent with Remark 5, not evidence against Theorem 4. The theorem's δ is a worst-case bound against the true conditional probability η_S, which is not directly observed here and is a different quantity from ECE, an average-case diagnostic. The clean, mechanical margin-vs-temperature relationship supports the margin-based mechanism Theorem 3 identifies; the ECE-vs-instability question remains open at the sample size tested here, and resolving it would need either substantially more trials per temperature or a redesigned protocol that controls δ directly rather than using ECE as a proxy for it.

F.6 Summary

The existing H3 experiment (Section 7) demonstrates that optimization discontinuities exist and may be severe when activated. H5 tests a stronger, quantitative claim – that the empirical frequency of such discontinuities is upper-bounded by the margin CDF exactly as Theorem 3 predicts – and the bound holds at every ε tested, though loosely, for reasons intrinsic to the gap between worst-case and random-direction perturbations. H6 confirms the mechanical margin-temperature relationship the theory is built on, while leaving the calibration-diagnostic (ECE) connection to instability frequency as an open empirical question rather than a confirmed one.

Together with Theorem 1, these results separate two questions: how large an optimization failure can be, and how often should one expect to observe it. The first is answered by the aggregation–optimization stability gap; the second, now with empirical support rather than only a proposed protocol, by the margin-distribution theory developed in this section.