EXISTENCE WITHOUT REACHABILITY: LOCAL EDITING BARRIERS FOR CONSTRAINED-TRAJECTORY WATERMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured tasks often admit many valid trajectories, allowing a provider to embed a watermark by choosing which solution to release. Whether an editor can remove the mark while keeping the task working is decided not by how many unmarked solutions exist, but by which of them it can reach. We make this precise through one object, the basin of the released solution: the trajectories reachable by local edits within a violation budget. For detectors that repair downhill, erasure depends only on which valid solutions the basin contains, and certified attribution on how far the basin extends. Local sensitivity and unique-neighbour expansion build energy walls that confine the basin. For parity tasks, the valid solutions in a basin form a coset of the group generated by closed syndrome walks, and, for moves shorter than its minimum distance, the task's code acquires a non-archimedean barrier norm whose balls are these groups. Our construction, BasinMark, is erased at the norm of the cheapest codeword its affine key can see; a random key sees the first jump of the norm except with probability at most $2^{-\ell}$. On 44 parity instances with 100–1600 decisions, certificates and replayed erasure paths locate this jump exactly on twenty. At $T = 400$ the exact threshold grows with the column weight from 0.01 to 0.04, and both bounds keep rising at higher weight. Heuristic attacks first erase at up to twice the slack at which optimized paths already succeed.

"A frog in a well cannot be talked with about the sea;—he is confined to the limits of his hole." — *Zhuangzi*, "Autumn Floods", trans. Legge (1891) (explained in Appendix K)

(a) Existence without reachability (b) Loops lift to codewords (c) The first jump rises with c

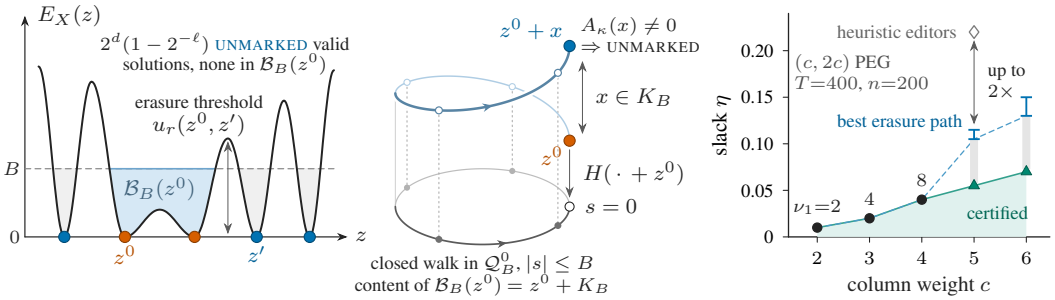

Figure 1: **Existence without reachability.** (a) Schematic: unmarked solutions abound (Theorem A.16), but under downhill repair a budget-B editor erases only if its basin holds one (Proposition 4.1). (b) Lifts of closed syndrome walks reach exactly $z^0 + K_B$ (Theorem 4.5). (c) The barrier norm's first jump rises with c; heuristic editors need up to twice the best path's slack (Table 1).

1 INTRODUCTION

Plans and tool-call sequences carry decisions that must work together: revising an early step can require coordinated changes later. When several trajectories solve the same task, a provider can

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.

encode a watermark in which one it returns while preserving task utility (Kahng et al., 2001; Qu & Potkonjak, 1998; Megerian et al., 2002). This raises a robustness question: if unmarked valid solutions are plentiful, can an editor reach one from the released output and keep the task working?

Consider routes passing on opposite sides of an obstacle. Each is valid, yet changing one into the other a few decisions at a time may break the route in between. Solution diversity supplies room for a watermark; the geometry of intermediate trajectories determines which alternatives an editor can reach. We study this distinction between *existence* and *reachability*.

The basin. Everything a utility-bounded local editor can do happens inside one set: the trajectories reachable from the released solution without exceeding the violation budget. Erasure is decided by which valid solutions this basin contains; attribution is certified by how far it extends.

The first question is what the basin contains. For detectors that repair downhill, an unmarked output is reachable exactly when an unmarked valid solution is. Erasure therefore depends on content alone, and its threshold is a barrier distance: the least peak energy of a walk to an unmarked solution. The second question is how far the basin extends. Positive attribution holds while the basin stays where the detector is certified. An energy wall around the released solution confines the basin, and local sensitivity with unique-neighbour expansion builds such walls from the task's constraints.

For parity tasks the content can be computed exactly. Energy depends only on the syndrome, so the trajectory graph covers a smaller syndrome graph, and walks around closed syndrome loops accumulate codewords. A basin reaches exactly a coset of the group these loops generate. For moves shorter than the minimum distance, the barrier distance becomes a non-archimedean norm on the task's code whose balls are these groups, and the watermark's affine key reads off the result: the mark is erased at the norm of the cheapest codeword the key can see. A random key sees the first jump of the norm except with probability at most $2^{-\ell}$, so the threshold belongs to the task rather than the key.

Contributions.

1. **Basins.** Erasure is decided by the content of a basin and certified attribution by its radius. One wall argument yields confinement, output fidelity, and finite certificates.
2. **The barrier norm.** For parity tasks, the content of a basin is a coset of a monodromy group. The code's barrier norm gives the exact erasure threshold of every affine key, shows that a random key cannot buy robustness beyond its first jump, and determines how many keyed checks can survive a given budget.
3. **Exact finite thresholds.** On 44 instances, complete component searches and replayed erasure paths locate the first jump exactly on twenty and bracket it on the rest. It rises with the column weight, and heuristic attacks overstate it.

The guarantees concern edits to a released artefact; discarding it and regenerating a solution, which is straightforward for parity tasks, lies outside this model.

2 Related Work

Constraint-based and trajectory watermarking. Classical constraint-based watermarking encodes information by biasing which valid solution a solver returns (Kahng et al., 2001). Its sequential-decision analogue uses the choice among legal moves as a carrier (Hernández-Castro et al., 2006). Recent work marks agent behavior through keyed hook actions, selection and tally channels, equivalent tool paths, or planning-level identifiers (Meng et al., 2026; Gao et al., 2026; Wang et al., 2026; Huang et al., 2026). We retain the established idea of selecting among valid solutions and study a different source of robustness: the geometry induced by the task's own consistency requirements.

Robustness and erasure attacks. Zhang et al. (2023) give erasure attacks when quality-preserving perturbations induce an efficiently mixing walk on high-quality outputs. Harel-Canada et al. (2025) examine this assumption empirically, observe slow mixing in text, and discuss possible high-quality islands. We turn this geometric intuition into an instance-level barrier: local task structure certifies when bounded final distortion or utility-preserving local edits cannot leave the neighborhood of the released solution. Abdalla & Vershynin (2025) instead obtain unremovability from a statistical-to-

computational gap. Our guarantees allow computationally unbounded, key-aware editors and restrict the outputs or edit paths they may use.

Error-correcting codes. Robustness can also be obtained by adding redundancy to a watermark payload, including error-correcting codes (Chao et al., 2024) and pseudorandom error-correcting codes with cryptographic robustness (Christ & Gunn, 2024). BasinMark also uses decoding, but its parity constraints are part of the task itself: the keyed affine check selects among their solutions, and the task's constraint geometry decides whether an editor can leave the decoder's certified neighborhood within its utility and editing budget.

Constraint geometry. Locked constraints make a single changed decision locally visible, while expander codes ensure that many constraints see exactly one change (Sipser & Spielman, 1996; Zémor, 2001; Zdeborová & Mézard, 2008b;a). Lemma 4.3 combines these roles to lower-bound the utility cost of leaving the detector's certified ball. The watermark-specific step is the interface of this barrier with three-valued detection, explicit editor classes, and constructive escape conditions. For parity tasks, our exact criterion specializes the classical connectivity theorem for voltage graphs (Gross & Tucker, 1987): closed-walk labels describe reachable valid differences, and the keyed affine map decides which of those differences erase the mark. The first jump of the resulting barrier norm is the classical counterpart of a code's energy barrier, studied for self-correcting quantum memories (Zhao et al., 2025). Appendix J expands these comparisons, including overlap-gap arguments and thermal stability.

3 Problem Setup and BasinMark

3.1 Tasks as Constraint Systems

Validity and utility. A trajectory $z \in \{0,1\}^T$ specifies T binary high-level decisions. Task X has n constraints, each acting on an ordered tuple C_j of Δ distinct decisions and accepting assignments in $R_j \subseteq \{0,1\}^\Delta$. We measure how many task rules a trajectory breaks:

$$E_X(z) = \big|\{j : z|_{C_j} \notin R_j\}\big|, \qquad U_X(z) = 1 - E_X(z)/n, \qquad \mathrm{Sol}(X) = \{z : E_X(z) = 0\}.$$

Thus E_X is an edit energy and U_X is its normalized utility. Utility slack $\eta = 1 - q$ allows at most ηn violated constraints. Each decision has degree at most c, with $3 \le c < \Delta$; under exact c-regularity, $cT = \Delta n$. Distance dist is Hamming distance in the specified decision representation.

Local geometry. A relation R_j is locally sensitive if flipping any one coordinate of any $y \in R_j$ leaves R_j. Parity and 1-in-K constraints are examples; linearity is not required. In the bipartite decision–constraint graph G_X, let $\Gamma(S)$ be the constraints adjacent to $S \subseteq [T]$. We say that small decision sets expand when

$$|\Gamma(S)| \ge (1-\epsilon)c|S| \quad \text{for all } |S| \le \gamma T.$$

Full definitions and a pairwise-overlap certificate appear in Appendix A.

3.2 Editing and Detection

Starting from a marked valid trajectory z^0, the attacker seeks a utility-feasible output that the detector labels UNMARKED. We study two ways of editing.

Final-output editing. An offline editor may pass through arbitrary intermediate candidates; only the final distortion $\mathrm{dist}(z, z^0)$ is restricted.

Pathwise editing. An attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ produces $z^0, \dots, z^N$ with

$$\mathrm{dist}(z^{t+1}, z^t) \le r \qquad \text{and} \qquad U_X(z^t) \ge q$$

at every step. Neither editor is computationally bounded, and both may know the key, detector, and an explicit evasive target.

3.3 THE BASINMARK CONSTRUCTION

BasinMark in three steps. Gaussian elimination exposes free coordinates for the valid solution space. The key selects one affine slice of those coordinates, from which the provider samples a valid marked trajectory. At detection time, a decoder first recovers a valid trajectory; an affine check then decides whether that trajectory belongs to the marked slice.

Information coordinates and marked generation. For parity constraints, write $\mathrm{Sol}(X) = \{z : Hz = b\}$ over $\mathbb{F}_2$, assuming consistency. Gaussian elimination identifies an information set I such that $z \mapsto u = z|_I$ is a bijection from $\mathrm{Sol}(X)$ to $\mathbb{F}_2^d$, where $d = T - \mathrm{rank}(H) \geq (1 - c/\Delta)T$. A key specifies a full-row-rank matrix $M_\kappa \in \mathbb{F}_2^{\ell \times d}$ and an offset $s_\kappa \in \mathbb{F}_2^\ell$, with $1 \leq \ell < d$. BasinMark samples

$$u \sim \mathrm{Unif}\{u : M_\kappa u = s_\kappa\}$$

and returns the unique valid trajectory with information coordinates u. The full procedure, including rank resampling, is Algorithm 1 in Appendix A.4.

Three-way detection. Detection first repairs a candidate z with the sequential decoder, which uses the syndrome $Hz \oplus b$ and flips a decision only when doing so strictly decreases E_X. If the result $\hat{c}$ is still invalid, the detector returns ABSTAIN, a decoding failure counted separately from erasure. Otherwise it returns MARKED exactly when $M_\kappa(\hat{c}|_I) = s_\kappa$, and UNMARKED otherwise. A certified integer radius a is therefore a ball on which decoding returns the original codeword and attribution is constant. With expansion through γT and $\epsilon < 1/4$, Proposition A.14 gives $a = \lceil (1 - 2\epsilon)\lfloor \gamma T \rfloor \rceil - 1$; at finite scale we can instead certify a by exhaustive decoding checks.

Utility and distributional properties. Every generated trajectory has $U_X = 1$. Under the uniform-key assumptions of Proposition A.15, one output, averaged over the key, is uniform on $\mathrm{Sol}(X)$; for any key-independent artefact, the false-positive probability is at most $2^{-\ell}$. The mark consumes ℓ bits of strategy entropy. For a fixed key, marked samples occupy one affine coset; the resulting multi-sample boundary is discussed in Section 6 and Appendix A.4.

4 BASINS, WALLS, AND THE BARRIER NORM

This section computes the two invariants of the basin, its content and its radius. Proofs are in Appendices B, C and D.

4.1 BASINS

Fix a move radius r and let G_r join trajectories at Hamming distance between 1 and r. For a budget $B \geq 0$, the sublevel set $X_B = \{z : E_X(z) \leq B\}$ inherits the edges of G_r. The *basin* $\mathcal{B}_B(z^0)$ is the connected component of X_B that contains z^0. An attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ is exactly a walk in $\mathcal{B}_B(z^0)$ with $B = \lfloor \eta n \rfloor$, whatever the attacker knows or computes. For valid z, z', the *barrier distance* $u_r(z, z') = \min\{B : z' \in \mathcal{B}_B(z)\}$ is the least peak energy of a walk between them. Concatenating walks gives the ultrametric inequality $u_r(x, z) \leq \max\{u_r(x, y), u_r(y, z)\}$. A basin has two invariants: its *content*, the valid solutions it contains, and its *radius*, $\max_{z \in \mathcal{B}_B(z^0)} \mathrm{dist}(z, z^0)$.

Proposition 4.1 (Erasure is decided by content). *Suppose the detector labels the valid solution it decodes to, returns every valid input unchanged, and repairs through energy-nonincreasing moves of G_r. An editor with budget B can then produce an* UNMARKED *output from a marked z^0 if and only if $\mathcal{B}_B(z^0)$ contains an unmarked valid solution. The erasure threshold is the barrier distance from z^0 to the unmarked solutions.*

Appending the detector's repair to an editor's walk never leaves the basin. Sequential energy-decreasing bit flipping satisfies the conditions for every $r \geq 1$, and linearity is not used.

Proposition 4.2 (Walls confine basins). *If every z with $a < \mathrm{dist}(z, z^0) \leq a + r$ has $E_X(z) > B$, then $\mathcal{B}_B(z^0)$ lies in the Hamming ball of radius a about z^0. If the detector returns* MARKED *throughout that ball, every walk in $\mathcal{B}_B(z^0)$ stays* MARKED.

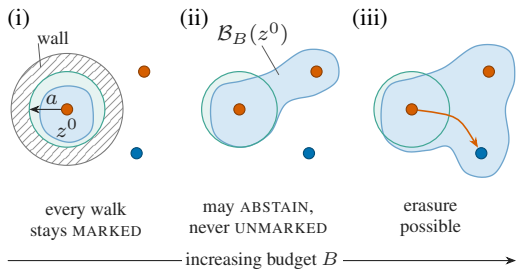

Figure 2: **Radius certifies, content decides.** Schematic basins $\mathcal{B}_B(z^0)$ for rising B, around the radius-a ball where the detector returns MARKED (teal); orange and blue dots are marked and unmarked valid solutions. (i) A hatched wall keeps the basin in the ball, so every walk stays MARKED (Proposition 4.2). (ii) The basin leaves the ball but holds no unmarked solution: outputs may ABSTAIN, never UNMARKED. (iii) Once it holds one, erasure is possible (Proposition 4.1).

A walk leaving the ball must first land within distance $a + r$. Certified attribution therefore concerns the basin's radius, and erasure its content (Figure 2). A basin may extend beyond the certified ball without an unmarked solution; its outputs can then cause abstention, never erasure.

4.2 Walls from Local Sensitivity and Expansion

Lemma 4.3 (Barrier). *Suppose every R_j is locally sensitive, every decision has degree at most c, and G_X is a $(\gamma T, (1-\epsilon)c)$-expander with $\epsilon < 1/2$. Let $z^* \in \mathrm{Sol}(X)$ and let $w = \mathrm{dist}(z, z^*)$ satisfy $0 < w \le \gamma T$. Then*

$$E_X(z) \;\ge\; (1-2\epsilon)\, c\, w.$$

Let S be the changed decisions. At most $c|S|$ edges leave S, while expansion forces at least $(1-2\epsilon)c|S|$ constraints to meet S exactly once. Each such constraint is violated by local sensitivity. Applied to the annulus just outside ζT, the lemma builds the wall of Proposition 4.2.

Theorem 4.4 (Confinement). *Assume the hypotheses of Lemma 4.3 with exact c-regularity, let $z^0 \in \mathrm{Sol}(X)$, and let $\eta = 1 - q$ be the utility slack. Put*

$$\zeta \;:=\; \frac{\eta}{(1-2\epsilon)\Delta}\,.$$

If $\zeta < \gamma$ and $r \le (\gamma - \zeta)T$, then every attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ started at z^0 satisfies $\mathrm{dist}(z^t, z^0) \le \zeta T$ for all t, on every realisation.

Read at a single point, the same wall gives output fidelity without any per-step condition: an output within distance γT whose utility is at least $1-\eta$, with $\eta < (1-2\epsilon)\Delta a/T$, is MARKED (Theorem D.1). When expansion is difficult to certify, a wall computed directly on a finite instance, the minimum energy $L_{a,b}$ on an annulus, plays the same role (Proposition D.2); for parity, binary linear feasibility checks certify it (Appendix F.4). Walls also separate valid solutions. Every walk between distinct valid solutions crosses the sphere of radius $\lfloor \gamma T \rfloor$ about its start, since distinct solutions are at least $\lfloor \gamma T \rfloor + 1$ apart (Corollary A.6). Hence $u_r(z, z') \ge (1-2\epsilon)c(\lfloor \gamma T \rfloor - r + 1)$ for distinct valid z, z' (Proposition D.3). With Proposition 4.1, no smaller budget erases the mark under a detector that repairs downhill, whatever radius its decoder is certified for.

4.3 Parity Basins and the Barrier Norm

For parity tasks the content of a basin can be computed exactly. If the current error is $e = z \oplus z^0$, its syndrome is $s = He$ and its energy is $|s|$. Fix a marked valid z^0, a budget B, and a translation-invariant set of allowed moves $\mathcal{R}$ (for radius one, the unit vectors). Form the labelled syndrome graph on $s \in \mathrm{im}\, H$ with $|s| \le B$: each move $v \in \mathcal{R}$ gives an edge $s \to s + Hv$ labelled v whenever both endpoints are allowed. Keep parallel edges and loops. In the component Q_B^0 of syndrome zero, choose a spanning tree and let p_s be the sum of move labels on its path from zero to s. Define

$$K_B = \mathrm{span}\{p_s + v + p_t : (s, t; v) \text{ is an edge of } Q_B^0\} \subseteq \ker H.$$

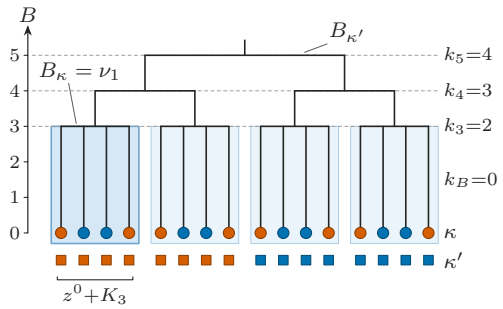

Figure 3: **The barrier norm is a tree.** Leaves are the valid solutions $z^0 \oplus x$ ($T = 14$, $n = 15$, column weight 3, $d = 4$, $r = 1$). Two merge at height $\nu(x \oplus x')$, so a budget B cuts the tree into cosets of K_B; dark water is the basin $\mathcal{B}_3(z^0)$ (Proposition D.4). A key colors leaves MARKED (orange) or UNMARKED (blue) and is erased at the lowest merge mixing them: κ at ν_1, as a random key is with probability $\geq 1 - 2^{-\ell}$; κ', designed to vanish on K_4, at 5 (Corollary 4.6).

The trajectory graph covers the syndrome graph, and K_B is the group of codewords accumulated around closed syndrome walks, its monodromy in the sense of graph covering theory (Gross & Tucker, 1987).

Theorem 4.5 (Reachable space and erasure)**.** *The errors reachable from zero under budget B are exactly $\bigcup_{s \in Q_B^0}(p_s + K_B)$. In particular, the reachable valid trajectories are $z^0 + K_B$. Put $k_B = \dim K_B$ and $t_B = \operatorname{rank}(v \mapsto M_\kappa(v|_I)$ on $K_B)$. Of these 2^{k_B} valid trajectories, exactly $2^{k_B} - 2^{k_B - t_B}$ are unmarked. If the detector fixes valid inputs and every successful repair uses allowed, energy-nonincreasing moves, an admissible erasure path exists if and only if $t_B > 0$.*

The content of every basin is thus a coset of K_B. Because energy is invariant under translation by codewords, the barrier distance between valid solutions depends only on their difference: $u_r(z, z \oplus x) = \nu(x)$ for $x \in \ker H$. The function ν satisfies $\nu(x \oplus x') \leq \max\{\nu(x), \nu(x')\}$, is positive on every nonzero codeword when $r < d_{\min}$, and has the subspaces K_B as its balls $\{\nu \leq B\}$. We call it the *barrier norm* of the task. With $A_\kappa(x) = M_\kappa(x|_I)$, Proposition 4.1 gives the erasure threshold of BasinMark exactly,

$$B_\kappa = \min\{\nu(x) : A_\kappa(x) \neq 0\} :$$

a mark is erased at the norm of the cheapest codeword its key can see (Proposition D.4).

Corollary 4.6 (What a key can buy)**.** *Let $\nu_1 = \min_{x \neq 0} \nu(x)$ be the first jump of the barrier norm. For a uniformly random full-row-rank key, $\Pr[B_\kappa > B] = \begin{bmatrix} d - k_B \\ \ell \end{bmatrix}_2 / \begin{bmatrix} d \\ \ell \end{bmatrix}_2$; hence $B_\kappa = \nu_1$ with probability at least $1 - 2^{-\ell}$. Some key with $\ell < d$ checks has $B_\kappa > B$ if and only if $k_B \leq d - \ell$.*

For a random key, the threshold is a property of the task: it is the first jump (Figure 3). A designed key survives a budget B only by vanishing on K_B, and at most $d - k_B$ checks, the codimension of that ball, can do so. The earlier bounds bracket the same invariant. Under expansion, $\nu_1 \geq (1 - 2\epsilon)c(\lfloor \gamma T \rfloor - r + 1)$ (Proposition D.3). Under exact c-regularity, flipping a minimum-weight codeword one bit at a time gives $\nu_1 \leq c\lfloor d_{\min}/2 \rfloor$ (Theorem A.19); Appendix A.6 adds a sufficient Metropolis escape-time bound. A complete traversal of Q_B^0 certifies the answer; a partial one supplies explicit paths to other valid trajectories but cannot prove that none exist. When $K_B = \{0\}$ and $\max_s |p_s| \leq a$, the basin lies inside radius a, so a decoder certified through radius a returns MARKED throughout (Proposition 4.2).

5 Parity Experiments

5.1 Experimental Design

Instances. The scale study uses $(c, \Delta) = (5, 10)$ parity systems with $n = T/2$, generated by progressive edge growth (Hu et al., 2005): four instances at each $T \in \{100, 200, 400, 800, 1600\}$ and two phase-sweep instances at each of $T \in \{400, 800\}$. A geometry study at $T = 400$ varies the column weight over $(2, 4)$, $(3, 6)$, $(4, 8)$ and $(6, 12)$ PEG systems and adds a $(5, 10)$ Gallager control with many four-cycles, four instances each; every family has rate $1/2$. Each instance carries a BasinMark key with $\ell = 16$.

Editors and outcomes. Besides the exact component search, an optimized editor finds low-weight codewords by randomized information-set search and follows beam-searched flip orders of their

Task geometry	T	Inst.	No erasure $\eta <$	Erasure walk η	Exact	Optimized attack η	Heuristic first erasure	Always MARKED $\eta <$
Scale								
(5,10) PEG	100	4	0.12	0.12	4/4	0.16–0.18	–	–
(5,10) PEG	200	4	0.08	0.08	4/4	0.12–0.14	–	0.06
(5,10) PEG	400	6	0.055	–	0/6	0.105–0.115	0.22	0.04
(5,10) PEG	800	6	0.0375	–	0/6	0.1175–0.1225	0.22	0.02
(5,10) PEG	1600	4	0.0175	–	0/4	0.125–0.135	–	0.01
Geometry at $T = 400$								
(2,4) PEG	400	4	0.01	0.01	4/4	0.01	–	–
(3,6) PEG	400	4	0.02	0.02	4/4	0.025–0.03	–	–
(4,8) PEG	400	4	0.04	0.04	4/4	0.08	–	–
(5,10) PEG†	400	6	0.055	–	0/6	0.105–0.115	0.22	0.04
(6,12) PEG	400	4	0.07	–	0/4	0.13–0.15	–	–
(5,10) Gallager	400	4	0.055	–	0/4	0.11–0.115	–	–

Table 1: Radius-one erasure thresholds of the original bit-flip detector. *No erasure*: a complete syndrome component with $K_B = \{0\}$ (all keys, all path lengths). *Erasure walk*: an explicit path from the same search, replayed with the detector; the threshold is exact when the two meet. *Optimized attack*: a beam-searched flip order of a low-weight codeword, replayed the same way. *Heuristic attacks*: first radius-one erasure by the editors of Figure 4 on the two phase-sweep instances of each size. *Always* MARKED: the basin stays within radius three, where the decoder of Section 5.3 is certified. A dash means no erasure walk within the search limit, a heuristic sweep not run, or radius-three decoding not certified, respectively. All families have rate $1/2$ ($n = T/2$); ranges are over instances. †Same instances as in the scale block.

supports. The heuristic sweep uses a blind local walk, an informed greedy editor, and a straight-line editor with random flip orders; the last two know the key and a target found by low-weight search. It covers the phase-sweep instances with ten slacks, $r \in \{1, 4, 16\}$, six targets per cell and 3×10^4 moves per attack. An attack succeeds only when its output satisfies the utility floor and the detector returns UNMARKED; decoder failures remain ABSTAIN.

5.2 Exact Erasure Thresholds

A complete component with $K_B = \{0\}$ certifies $\nu_1 > B$ (Section 4.3): no radius-one path, of any length and under any key, reaches an UNMARKED output when $\lfloor \eta n \rfloor \le B$. Conversely, each nonzero cycle label found by the search comes with an explicit editing path to another valid trajectory. We replay every such path with the task's energy and the three-valued detector and count it as an erasure only when the detector returns UNMARKED. When certificate and path meet at consecutive budgets, the threshold is exact and equals ν_1. Table 1 reports the calculation on 44 instances.

Exact thresholds. Certificate and erasure walk meet on all eight $(5, 10)$ instances with $T \le 200$ and on every instance of the three lower-degree families, so twenty thresholds are exact. At $T = 200$ the erasing walks change 56–70 of the 200 decisions while keeping the utility floor throughout.

Dependence on geometry. At fixed $T = 400$ and rate $1/2$, the exact threshold rises with column weight, from 0.01 for cycle codes ($c = 2$) through 0.02 to 0.04 ($c = 4$). For $c = 5$ and 6, the certificates (0.055 and 0.07) and upper bounds (0.105–0.115 and 0.13–0.15) keep rising. The Gallager control (59–78 four-cycles) matches PEG at both ends, so short cycles weaken neither bound.

Dependence on scale. In absolute terms, the barrier grows with the horizon. The exact erasure budget rises from six violated constraints at $T = 100$ to eight at $T = 200$, and certificates exceed ten and fourteen violations at $T = 400$ and 800. From $T = 400$ to 1600 the best constructive paths found need a stable fraction of violated constraints, 0.105–0.135; at $T = 400$, certificate and attack bracket the threshold within about a factor of two. Beyond $T = 200$ the certificate is limited by component size rather than by any erasure path: components of up to 5.8×10^7 syndromes are exhausted, and every larger budget exceeds the search limit.

5.3 Certified Radius and Decoder Coverage

The component calculation also certifies decoding. With column weight five, every nonzero codeword of weight at most six has a single-bit path of peak energy at most nine (Proposition C.6). Hence $K_9 = \{0\}$ gives minimum distance at least eight on every archived instance with $T \geq 400$, and exact feasibility checks close the four at $T = 200$. A bounded-distance decoder is therefore unambiguous through radius three. On 12,000 matched random corruptions, it recovers every error of weight at most three and abstains above; bit flipping still recovers 1,733–1,996 of 2,000 at those weights (Appendix F.1). When a complete component also stays inside radius three, this decoder returns MARKED along every path, which gives the last column of Table 1.

5.4 Computed Walls

Walls computed directly certify smaller regions. Searched minimum energies grow approximately linearly over a range that lengthens with T (Figure 12), but the original bit-flip decoder has exact radii of only one or two at $T \leq 400$ (Appendix F.6). On one $T = 400$ instance, exact minimum energies $5, 6, 7$ at weights one to three and feasibility checks at weights four and five give $L_{2,5}(z^0) = 7$. Proposition D.2 therefore certifies every output within distance five for $\eta < 0.035$ (Appendix F.4). The same profile gives $L_{0,5} = 5$, which forbids even a first move of radius at most five for $\eta < 0.025$; the component certificates of Table 1 instead cover nontrivial paths.

5.5 The Erasure Phase Diagram

Figure 4 places the original sweep against these bounds. The blind walk never erases. The informed editors show three regimes: neither abstention nor erasure at small slack, abstention at intermediate slack, and erasure from $\eta = 0.22$, or from 0.15 on one $T = 400$ instance with $r \geq 4$. The separation between abstention and erasure is visible only with the three-valued detector. The onset of erasure, however, reflects the editors' targets and flip orders, on which the component calculation does not depend: optimized paths already erase these instances at 0.115–0.1225.

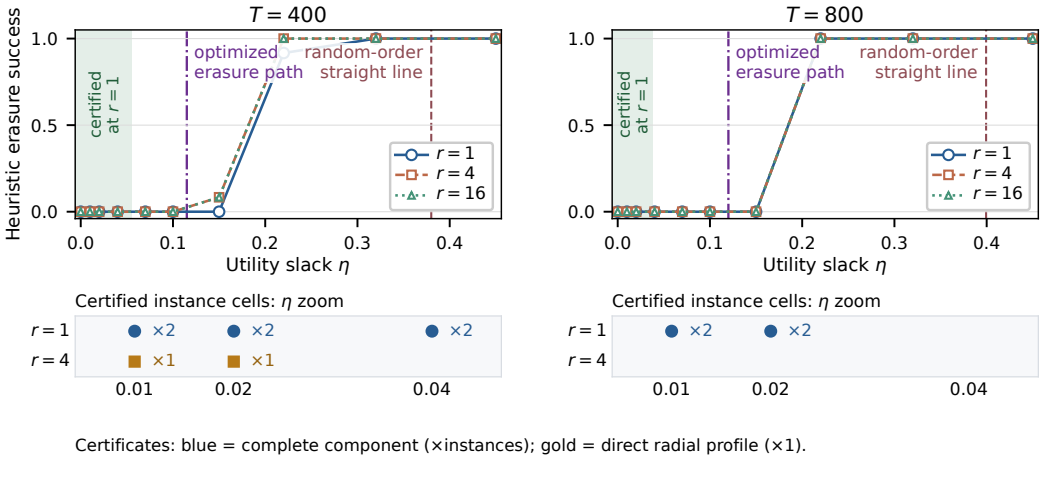

Figure 4: Heuristic erasure success across utility slack and move radius (informed greedy editor, mean of two instances). Shaded: certified radius-one no-erasure region from complete components. Dash-dot: optimized radius-one erasure path (mean of the two instances). Dashed: random-order straight-line peak. Lower strips: certified phase-grid cells with multiplicities; ten come from complete components and two radius-four cells from the direct profile. Abstention rates appear in Appendix F.17 and per-instance brackets in Table 2.

Committed execution. We also realize the pathwise threat model through irreversible execution, freezing each decision after a validated radius-one edit. For the monotone target-following editor, neither arm erases at $\eta \leq 0.15$; at $\eta = 0.22$ the committed and no-commit arms reach an exact-valid unmarked trajectory on $11/24$ and $24/24$ targets (Appendix F.15).

5.6 Editing Geometry Beyond Parity

Complete planning-space enumeration tests whether edit separation occurs outside the parity construction. Corridor mazes have three homotopy-aligned components (purity 1.00) below the critical radius $r_c = s+1$, where s is gate separation. Blocksworld remains connected at radius one. These are direct geometric observations, not checks of the expansion assumptions (Appendix F.2, Figure 11).

6 Discussion

6.1 Scope and Limitations

Regeneration and repeated releases. The threat model is artefact-relative. A key-aware re-planner can solve for another affine class, and a key-independent uniform valid sample is unmarked with probability $1 - 2^{-\ell}$, both at utility one. Fixed-key samples also reveal the affine carrier: three marked samples permit an affine forgery, while $d - \ell + 1$ affinely independent samples identify the accepted coset. The present construction therefore addresses constrained editing of a released trajectory, rather than regeneration or multi-sample unforgeability; a registry variant with a fresh private offset per release is analysed in Proposition C.7. The translation invariance that makes parity basins computable is also what makes fixed-key marks forgeable: both reflect that the solution set is a coset of the code.

Finite certified range. Thresholds are exact where components can be exhausted up to the first erasure budget. The two exact $(5, 10)$ values, 0.12 and 0.08, come from the smallest sizes; beyond them the certificate is limited by component size, so its relative slack shrinks with n while constructive attacks stay near 0.11–0.13. The expansion constants are far from practical ($\gamma \approx 10^{-14}$ for the $(5, 10)$ family; Remark I.3), and the finite thermal runs do not enter the predicted exponential regime.

Evidence beyond parity. The committed-prefix experiment measures one deterministic monotone editor, not the full class $\mathcal{A}_{\mathrm{loc}}(q, 1)$. Planning enumeration establishes edit-space separation as a phenomenon, without the constraint encoding required by Lemma 4.3. The ALFWORLD snapshot covers 35 short trajectories and has insufficient power for an end-to-end detector.

6.2 Outlook

First, scalable lower bounds on the path barrier, such as cut certificates, would close the remaining bracket at large T, and stronger decoders would extend positive attribution. Second, nonlinear locked families offer a setting in which the same local barrier may coexist with difficult regeneration; their basins and walls are unchanged, but without a symmetry group their content has no norm and must be explored directly.

7 Conclusion

We reduce the robustness of a task-embedded watermark to the basins of the task's energy landscape. Walls confine basins and certify attribution for any decoder certified inside them. For parity tasks, the content of a basin is a coset of a monodromy group, and the code's barrier norm fixes the erasure threshold of every affine key. For a random key this threshold is the task's first jump ν_1 except with probability at most $2^{-\ell}$, and it is computed exactly wherever the components can be exhausted.

AI Use Statement

The research question, the framing of watermark robustness as reachability in the task's edit geometry, the choice of experiments and the decisions taken at every stage of this work are the authors'. We used LLM-based assistants throughout as tools: for literature search, for coding assistance in implementing and running the experiment and analysis scripts and the figures, for drafting and editing text to improve readability, for formatting references, for suggesting titles and keywords, and for assistance in developing and checking the theoretical arguments, including drafting proof steps that the authors then verified.

We have reviewed all AI-assisted work: every theoretical claim was checked by hand and, where it admits a numerical test, against the finite-instance computations in Section 5 and Appendix F; every

reported experimental number is regenerated from archived run records by the accompanying analysis scripts; and the literature discussion was checked against the cited sources. We take responsibility for the final content of this work, including text, claims, code and artifacts produced with the aid of generative AI.

Reproducibility Statement

Full theoretical statements and proofs are provided in Appendices A–D. Appendix E specifies instance generation, decoding, and the agent pipeline; Appendix F gives supplementary measurements, certification ranges, censoring, and sample sizes. Code and data for the reported experiments and construction checks will be released upon publication.

References

Emmanuel Abbe, Shuangping Li, and Allan Sly. Binary perceptron: Efficient algorithms can find solutions in a rare well-connected cluster. *arXiv preprint arXiv:2111.03084*, 2021.

Pedro Abdalla and Roman Vershynin. LLM watermarking using mixtures and statistical-to-computational gaps. *arXiv preprint arXiv:2505.01484*, 2025.

Hyeseon An, Shinwoo Park, Dongsu Kim, and Yo-Sub Han. Sequential behavioral watermarking for LLM agents. *arXiv preprint arXiv:2605.11036*, 2026.

Ben Barber. A note on balanced independent sets in the cube. *Australasian Journal of Combinatorics*, 52:205–207, 2012. arXiv:1210.4029.

Marco Benedetti, Andrej Bogdanov, Enrico M. Malatesta, Marc Mézard, Gianmarco Perrupato, Alon Rosen, Nikolaj I. Schwartzbach, and Riccardo Zecchina. Are neural networks collision resistant? *arXiv preprint arXiv:2509.20262*, 2025.

Andrej Bogdanov, Alon Rosen, Neekon Vafa, and Vinod Vaikuntanathan. Adaptive robustness of hypergrid Johnson–Lindenstrauss. *arXiv preprint arXiv:2504.09331*, 2025.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

Patrick Chao, Yan Sun, Edgar Dobriban, and Hamed Hassani. Watermarking language models with error correcting codes. *arXiv preprint arXiv:2406.10281*, 2024.

Miranda Christ and Sam Gunn. Pseudorandom error-correcting codes. *arXiv preprint arXiv:2402.09370*, 2024.

Danilo Francati, Yevin Nikhel Goonatilake, Shubham Pawar, Daniele Venturi, and Giuseppe Ateniese. The coding limits of robust watermarking for generative models. *arXiv preprint arXiv:2509.10577*, 2025.

David Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41):e2108492118, 2021. doi: 10.1073/pnas.2108492118.

David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. *Annals of Probability*, 45(4):2353–2376, 2017.

Zheng Gao, Xiaoyu Li, Xiaoyan Feng, Jiaojiao Jiang, Yang Song, Yulei Sui, Zhenchang Xing, and Liming Zhu. TRACE: A two-channel robust attribution watermark via complementary embeddings for LLM-agent trajectories. *arXiv preprint arXiv:2607.08400*, 2026.

Sanjam Garg, Somesh Jha, Saeed Mahloujifar, and Mohammad Mahmoody. Adversarially robust learning could leverage computational hardness. *arXiv preprint arXiv:1905.11564*, 2019.

Jonathan L. Gross and Thomas W. Tucker. *Topological Graph Theory*. John Wiley & Sons, New York, 1987.

Fabrice Y. Harel-Canada, Boran Erol, Connor Choi, Jason Liu, Gary Jiarui Song, Nanyun Peng, and Amit Sahai. Sandcastles in the storm: Revisiting the (im)possibility of strong watermarking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29698–29735. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1436. arXiv:2505.06827.

Julio C. Hernández-Castro, Ignacio Blasco-López, Juan M. Estevez-Tapiador, and Arturo Ribagorda-Garnacho. Steganography in games: A general methodology and its application to the game of Go. *Computers & Security*, 25(1):64–71, 2006. doi: 10.1016/j.cose.2005.12.001.

Xiao-Yu Hu, Evangelos Eleftheriou, and Dieter M. Arnold. Regular and irregular progressive edge-growth Tanner graphs. *IEEE Transactions on Information Theory*, 51(1):386–398, 2005. doi: 10.1109/TIT.2004.839541.

Brice Huang and Mark Sellke. Tight Lipschitz hardness for optimizing mean field spin glasses. *arXiv preprint arXiv:2110.07847*, 2021.

Hanbo Huang, Yiran Zhang, Hao Zheng, Xuan Gong, Yihan Li, Lin Liu, Zhuotao Liu, and Shiyu Liang. RLCracker: Evaluating the worst-case vulnerability of LLM watermarks with adaptive RL attacks. *arXiv preprint arXiv:2509.20924*, 2025.

Kaibo Huang, Jin Tan, Yukun Wei, Wanling Li, Zipei Zhang, Hui Tian, Zhongliang Yang, and Linna Zhou. AgentMark: Utility-preserving behavioral watermarking for agents. *arXiv preprint arXiv:2601.03294*, 2026.

Morteza Ibrahimi, Yash Kanoria, Matt Kraning, and Andrea Montanari. The set of solutions of random XORSAT formulae. *Annals of Applied Probability*, 25(5):2743–2808, 2015.

Andrew B. Kahng, John Lach, William H. Mangione-Smith, Stefanus Mantik, Igor L. Markov, Miodrag Potkonjak, Paul Tucker, Huijuan Wang, and Gregory Wolfe. Constraint-based watermarking techniques for design IP protection. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 20(10):1236–1252, 2001. doi: 10.1109/43.952740.

James Legge. *The Texts of Taoism, Part I: The Tâo Teh King; The Writings of Kwang-dze, Books I–XVII*, volume 39 of *Sacred Books of the East*. Clarendon Press, Oxford, 1891.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2nd edition, 2017.

Seapahn Megerian, Milenko Drinić, and Miodrag Potkonjak. Watermarking integer linear programming solutions. In *Proceedings of the 39th Design Automation Conference (DAC)*, 2002.

Wenlong Meng, Chen Gong, Terry Yue Zhuo, Fan Zhang, Kecen Li, Zheng Liu, Zhou Yang, Chengkun Wei, and Wenzhi Chen. Watermarking LLM agent trajectories. *arXiv preprint arXiv:2602.18700*, 2026.

Carlos Munuera and Morgan Barbier. Wet paper codes and the dual distance in steganography. *arXiv preprint arXiv:1104.1970*, 2011.

Huy Phan, Kieu Dang, Ojaswi Dulal, Aiham AL Shukairi, Abby Shine, Chase Garner, and Phung Lai. LLM watermarking as big data provenance: A deployment-oriented systematization. *arXiv preprint arXiv:2607.10103*, 2026.

Gang Qu and Miodrag Potkonjak. Analysis of watermarking techniques for graph coloring problem. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, 1998.

Michael Sipser and Daniel A. Spielman. Expander codes. *IEEE Transactions on Information Theory*, 42(6):1710–1722, 1996. doi: 10.1109/18.556667.

Liwen Wang, Zongjie Li, Yuchong Xie, Shuai Wang, Dongdong She, Wei Wang, and Juergen Rahmel. On protecting agentic systems' intellectual property via watermarking. *arXiv preprint arXiv:2602.08401*, 2026.

Lenka Zdeborová and Marc Mézard. Constraint satisfaction problems with isolated solutions are hard. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(12):P12004, 2008a. arXiv:0810.1499.

Lenka Zdeborová and Marc Mézard. Locked constraint satisfaction problems. *Physical Review Letters*, 101(7):078702, 2008b. arXiv:0803.2955.

Gilles Zémor. On expander codes. *IEEE Transactions on Information Theory*, 47(2):835–837, 2001. doi: 10.1109/18.910593.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023. ICML 2024.

Guangqi Zhao, Andrew C. Doherty, and Isaac H. Kim. Energy barrier of hypergraph product codes. *Physical Review Letters*, 134(18):180601, 2025. arXiv:2407.20526.

Appendix

Table of Contents

A Extended Theoretical Development **14**
A.1 Preliminaries and Technical Background . . . 14
A.2 Expansion Forces a Linear Barrier . . . 15
A.3 Confinement: the Attacker Cannot Leave the Ball . . . 17
A.4 BasinMark: Marking the Free Coordinates . . . 18
A.5 Upper Bounds: Where the Barrier Stops . . . 20
A.6 Thermal Attackers: Paying for Violations . . . 21

B Theoretical Results **22**
B.1 Proof of Lemma A.4 . . . 22
B.2 Proof of Lemma A.5 . . . 22
B.3 Proof of Lemma 4.3 . . . 23
B.4 Proof of Corollary A.6 . . . 23
B.5 Proof of Lemma A.7 . . . 23
B.6 Proof of Theorem 4.4 . . . 23
B.7 Proof of Corollary A.9 . . . 24
B.8 Proof of Corollary A.10 . . . 24
B.9 Proof of Theorem D.1 . . . 24
B.10 Proof of Proposition D.2 . . . 24
B.11 Proof of Theorem A.12 . . . 25
B.12 Proof of Proposition A.14 . . . 25
B.13 Proof of Proposition A.15 . . . 25
B.14 Proof of Theorem A.16 . . . 26
B.15 Proof of Theorem A.18 . . . 26
B.16 Proof of Theorem A.19 . . . 26
B.17 Proof of Lemma A.21 . . . 26
B.18 Proof of Theorem A.22 . . . 27

C Exact Reachability in Parity Systems **27**
C.1 Lifting Syndrome Paths . . . 27
C.2 From Reachability to Erasure . . . 29
C.3 The Effect of a Random Affine Test . . . 29
C.4 Computing the Certificate . . . 30
C.5 A Decoding Certificate from the Component . . . 30
C.6 Independent Private Offsets Across Releases . . . 31

D Proofs for the Basin Framework **32**
D.1 Output fidelity and the finite-instance wall . . . 32
D.2 Proof of Proposition 4.1 . . . 32
D.3 Proof of Proposition 4.2 . . . 32
D.4 Barrier distance under expansion . . . 33
D.5 The barrier norm . . . 33
D.6 Proof of Corollary 4.6 . . . 34

E Experimental Protocol **34**
E.1 Constraint instances . . . 34
E.2 Decoder and detector . . . 34
E.3 Detector-radius certification . . . 34
E.4 ALFWORLD trajectories . . . 35

F Supplemental Experimental Results **35**
F.1 Exact Thresholds and Decoder Comparison . . . 35
F.2 Connectivity of Fully Enumerated Plan Spaces . . . 36
F.3 Barrier profiles . . . 36

F.4 A certified finite profile . . . 36
F.5 Girth, sampling, and local expansion . . . 38
F.6 Detector radius beyond weight two . . . 39
F.7 Detector coverage under corruption . . . 39
F.8 ALFWORLD geometry, per task type . . . 40
F.9 Interpreting finite measurements . . . 40
F.10 Certified and sampled detector radii . . . 40
F.11 Metropolis escape . . . 41
F.12 Agent-editing attacks . . . 41
F.13 Exact null and test power . . . 41
F.14 The theorem-covered phase cells . . . 41
F.15 Committed-prefix editing . . . 42
F.16 Barrier and reach profiles . . . 42
F.17 Three empirical regimes . . . 43
F.18 Exact planning-space geometry . . . 43
F.19 ALFWORLD census . . . 43
F.20 Supporting Figures and Tables . . . 44

G Scope of the Guarantees **45**

H Additional Analysis **45**
H.1 Certified radius . . . 45
H.2 Two barrier-slope summaries . . . 46
H.3 Abstention and erasure . . . 46
H.4 Instance-level checks . . . 46
H.5 Density is not the whole story . . . 46
H.6 A nonlinear candidate family . . . 47
H.7 A certified parameter point . . . 47
H.8 A barrier-profile interface . . . 47
H.9 Operational enforcement of condition (b) . . . 48
H.10 Why the straight-line attack is cheap . . . 48
H.11 Division of labour between task and key . . . 48
H.12 Open Problems . . . 48

I Technical Remarks **49**

J Additional Related Work **50**

K On the Epigraph **51**

A Extended Theoretical Development

This appendix develops the full model, auxiliary results, construction, attack bounds, and thermal extension. Lemma 4.3, Theorem 4.4 and Theorem D.1 are restated for continuity; their proofs follow in Appendix B.

A.1 Preliminaries and Technical Background

Definition A.1 (Task instance)**.** A task instance is $X = \big([T], \{(C_j, R_j)\}_{j=1}^{n}\big)$, where each C_j is an ordered tuple of Δ distinct elements of $[T]$ and $R_j \subseteq \{0,1\}^{\Delta}$. A *trajectory* is $z \in \{0,1\}^T$: a vector of T binary high-level decisions. Trajectory z *satisfies* constraint j if $z|_{C_j} \in R_j$, where $z|_{C_j}$ is read off in the order fixed by C_j. Write

$$E_X(z) = \big|\{j : z|_{C_j} \notin R_j\}\big|, \qquad U_X(z) = 1 - E_X(z)/n, \qquad V(q) = \{z : E_X(z) \le (1-q)n\},$$

and $\mathrm{Sol}(X) = V(1)$. Every decision lies in at most c constraints; where we additionally need exact c-regularity (so that $cT = \Delta n$) we say so. We assume throughout that $3 \le c < \Delta$.

The reading: decision i is one binary strategic choice the planner makes — which of two tool families to use at a stage, which of two decompositions, which side of an obstacle to pass — and constraint j says the Δ decisions it touches must be jointly consistent for the plan to execute. E_X counts how many consistency requirements a candidate breaks, and $\eta := 1 - q$ is the *utility slack* the attacker is permitted.

Definition A.2 (Locally sensitive constraint)**.** R_j is *locally sensitive* if for every $y \in R_j$ and every coordinate $p \in \{1, \dots, \Delta\}$, the vector obtained from y by flipping coordinate p is *not* in R_j; equivalently, any two distinct elements of R_j are at Hamming distance at least 2 (vacuously true when $|R_j| \leq 1$).

Local sensitivity says no single decision can be revised in isolation: every constraint is *tight*. For symmetric relations it is exactly the *locked* condition of Zdeborová & Mézard (2008b;a), whose second condition — every variable appears in at least two distinct constraints — is the leaf-free requirement our expansion hypothesis strengthens quantitatively. Their canonical locked examples are precisely parity, $R_j = \{y : \bigoplus_p y_p = b_j\}$, and 1-in-$K$; we take both from them. 1-in-3, $R_j = \{100, 010, 001\}$, is not affine, and finding a single solution is NP-hard (Remark I.1).

For the task model, these ingredients yield a deterministic energy formulation. Zdeborová and Mézard argue that locked solutions are isolated: at distance of order $\log N$ from the shortest loop through a typical variable in a locally tree-like factor graph, and extensively by a cavity-method prediction, both statements about *distance* between solutions in a random ensemble. Unique-neighbour energy bounds are classical for parity checks in expander codes; Lemma 4.3 packages the same counting mechanism for any locally sensitive relation on a given instance. It lower-bounds the number of broken constraints linearly in displacement, which is the interface the watermark argument consumes. A constraint that merely forbids one bad combination is *not* locally sensitive, so a barrier for such tasks requires a different argument.

Definition A.3 (Factor graph, expansion)**.** G_X is the bipartite graph on decisions $[T]$ and constraints $[n]$ with $i \sim j$ iff $i \in C_j$; $\Gamma(S)$ denotes the constraint neighbourhood of $S \subseteq [T]$. G_X is a $(\gamma T, (1-\epsilon)c)$-*expander* if $|\Gamma(S)| \geq (1-\epsilon)c|S|$ for every S with $|S| \leq \gamma T$.

Expansion says small sets of decisions touch nearly as many distinct constraints as their degree allows: constraints are spread out rather than piled on the same few decisions.

At small w, expansion follows from a pairwise overlap count that can be checked exhaustively on a given instance.

Lemma A.4 (Overlap expansion)**.** *Suppose every decision lies in* exactly c *constraints and every pair of decisions lies in at most μ common constraints. Then $|\Gamma(S)| \geq c|S| - \mu\binom{|S|}{2}$ for every $S \subseteq [T]$, so G_X is a $(w, (1-\epsilon)c)$-expander for every $w \leq 2\epsilon c/\mu + 1$.*

Proof in Appendix B.1.

Exact degree c is needed, not merely degree at most c: Definition A.3 compares against $c|S|$, so a decision of degree below c would have to be paid for by the same bound. Our instances are 5-regular by construction. The quantity μ is measurable; for $\mu = 2$ it is counted by the four-cycles of G_X. Section 5 therefore reports four-cycle counts directly.

Distances are measured in Hamming distance rather than the normalised overlap $1 - \mathrm{dist}(z, z')/T$ used in the landscape literature (Remark I.5).

A.2 Expansion Forces a Linear Barrier

Lemma A.5 (Unique neighbour)**.** *Let G be bipartite with left vertex set $[T]$, every left vertex of degree at most c, and let $S \subseteq [T]$ satisfy $|\Gamma(S)| \geq (1-\epsilon)c|S|$. Then the number u of right vertices with exactly one neighbour in S satisfies*

$$u \;\geq\; 2|\Gamma(S)| - c|S| \;\geq\; (1 - 2\epsilon)\, c\, |S|.$$

This is the unique-neighbour counting of the expander-code literature (Sipser & Spielman, 1996). The proof below records the code-free form used by Lemma 4.3.

Proof in Appendix B.2.

Restated (Lemma 4.3: Barrier). *Suppose every R_j is locally sensitive, every decision has degree at most c, and G_X is a $(\gamma T, (1-\epsilon)c)$-expander with $\epsilon < 1/2$. Let $z^* \in \mathrm{Sol}(X)$ and let $w = \mathrm{dist}(z, z^*)$ satisfy $0 < w \le \gamma T$. Then*

$$E_X(z) \;\ge\; (1-2\epsilon)\, c\, w.$$

Proof in Appendix B.3.

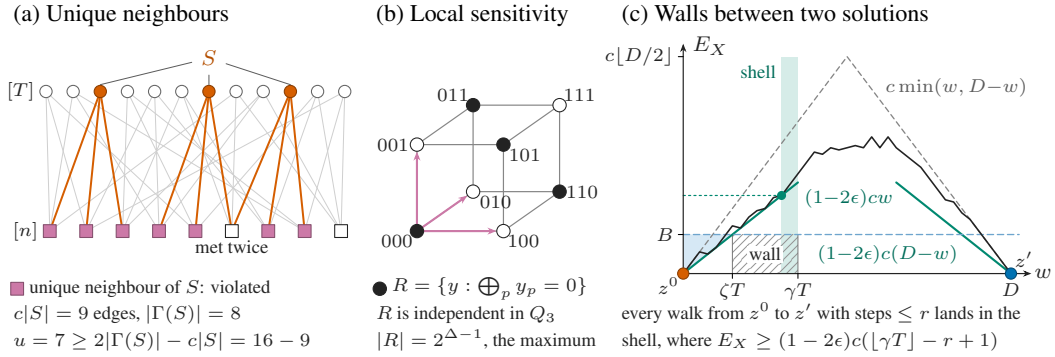

Figure 5: **Walls from expansion and local sensitivity.** (a) $T = 12$, $n = 9$, $c = 3$, $\Delta = 4$: S has $u = 7$ unique neighbours among $|\Gamma(S)| = 8$, so Lemma A.5 is tight ($\epsilon = 1/9$); all seven are violated (Lemma 4.3). (b) A locally sensitive R is independent in Q_Δ; the parity classes attain $|R| = 2^{\Delta-1}$ (Lemma A.7). (c) Schematic energy on the straight line between valid z^0 and z': the teal bounds hold only within γT of an endpoint, the dashed tent is Theorem A.19. The wall confines budget-B walks if $r \le (\gamma - \zeta)T$ (Theorem 4.4); walks with steps $\le r$ land in the shell (Proposition D.3).

Corollary A.6 (Linear minimum distance). *Under the hypotheses of Lemma 4.3, any two distinct $z^*, z' \in \mathrm{Sol}(X)$ satisfy $\mathrm{dist}(z^*, z') \ge \lfloor \gamma T \rfloor + 1$.*

Proof in Appendix B.4.

Why this is true. Expansion manufactures *witnesses*. Change w decisions and you touch at most cw constraints; expansion forces a constant fraction of those to be touched exactly once, and local sensitivity converts "touched exactly once" into "certainly broken." Neither step inspects the global structure of the task, which is why the bound is linear in w rather than merely positive.

A.2.1 How much room is there? Parity is extremal

Local sensitivity sharply restricts the admissible relation. Within this class, parity is the unique density-extremal relation.

Lemma A.7 (Extremal classification; classical). *Let $\Delta \ge 2$ and let $R \subseteq \{0,1\}^\Delta$ be locally sensitive. Then $|R| \le 2^{\Delta-1}$, and equality holds if and only if R is one of the two parity classes $\{y : \bigoplus_p y_p = b\}$, $b \in \{0,1\}$.*

Local sensitivity makes R an independent set in the hypercube Q_Δ. The cardinality bound is the classical coding fact $A(\Delta, 2) = 2^{\Delta-1}$, and the classification of maximum independent sets as the two parity classes appears in Barber (2012). A perfect matching already gives the bound; the subcube-averaging proof below also yields the equality case.

Proof in Appendix B.5.

Density alone does not decide abundance, and the non-linear route — families such as 1-in-Δ, where finding any solution is NP-hard — stays open; the first-moment criterion that settles the small-arity side and leaves the large-arity side open is in Appendix H.5 and Appendix H.6.

The expansion argument also identifies two quantitative controls: an absolute ceiling on the protected slack (Remark I.2) and the finite-size reach of the first-moment expansion certificate (Remark I.3). Section 5 complements the latter by measuring the barrier profile directly.

A.3 Confinement: the Attacker Cannot Leave the Ball

Definition A.8 (Local attack class)**.** Given a starting trajectory z^0, an attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ is a procedure producing a finite sequence $z^0, z^1, \ldots, z^N$ such that, for every realisation of its randomness,

(a) $\mathrm{dist}(z^{t+1}, z^t) \le r$ for all t (r-locality), and

(b) $U_X(z^t) \ge q$ for all t (utility preservation),

and outputting z^N. It *succeeds* if $D_\kappa(z^N) = \textsc{unmarked}$ and $U_X(z^N) \ge q$.

Conditions (a) and (b) are the class's only restrictions. Members may be randomized, may use unbounded computation, and may know the task, secret key, detector, and complete list of evasive trajectories. The resulting guarantee isolates geometric hardness from cryptographic hardness.

Restated (Theorem 4.4: Confinement)**.** *Assume the hypotheses of Lemma 4.3 with exact c-regularity, let $z^0 \in \mathrm{Sol}(X)$, and let $\eta = 1 - q$ be the utility slack. Put*

$$\zeta \;:=\; \frac{\eta}{(1-2\epsilon)\Delta}\,.$$

If $\zeta < \gamma$ and $r \le (\gamma - \zeta)T$, then every attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ started at z^0 satisfies $\mathrm{dist}(z^t, z^0) \le \zeta T$ for all t, on every realisation.

The proof uses exact c-regularity only through $cT = \Delta n$; under the weaker standing assumption that every decision has degree at most c one has $\Delta n \le cT$ and the same argument goes through unchanged.

Proof in Appendix B.6.

Confinement mechanism. The attacker's own quality requirement is what traps it. Every decision it changes costs $(1-2\epsilon)c$ violated constraints; with a budget of ηn violations it can afford $\eta n/((1-2\epsilon)c) = \zeta T$ changes, and the locality bound $r \le (\gamma - \zeta)T$ prevents it from stepping over the region where that accounting is valid. Nothing in the argument depends on what the attacker knows or how long it computes.

Corollary A.9 (Local erasure is impossible)**.** *In addition, let D_κ be a detector that is constant on the ball of radius a about z^0, with $\zeta T < a$, and let $D_\kappa(z^0) = \textsc{marked}$. Then under the hypotheses of Theorem 4.4 every attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ fails, on every realisation; the success probability is exactly 0. Equivalently, erasure is impossible whenever*

$$\eta \;<\; (1-2\epsilon)\,\Delta\,\frac{a}{T} \qquad \textit{and} \qquad r \;\le\; \Big(\gamma - \frac{\eta}{(1-2\epsilon)\Delta}\Big)T. \tag{1}$$

Proof in Appendix B.7.

A.3.1 Dropping the per-step condition entirely

Condition (b) models online editing in which every intermediate trajectory must remain usable. For offline editing, deleting (b) yields an output guarantee conditioned jointly on final utility and final Hamming distortion.

Corollary A.10 (Perfect-utility outputs)**.** *Assume the hypotheses of Lemma 4.3 with $z^* = z^0 \in \mathrm{Sol}(X)$. If an attacker outputs any $z \in \mathrm{Sol}(X)$ with $\mathrm{dist}(z, z^0) \le \gamma T$, then $z = z^0$. Erasure therefore requires changing more than γT decisions, whatever the attacker does in between and however long it computes.*

Proof in Appendix B.8.

Restated (Theorem D.1: Output-fidelity bound; no per-step condition)**.** *Assume the hypotheses of Lemma 4.3 with $z^* = z^0 \in \mathrm{Sol}(X)$, let $D_\kappa(z^0) = \textsc{marked}$, and let* Dec *be certified at radius a. Let z be* any *trajectory with*

$$\mathrm{dist}(z, z^0) \;\le\; \gamma T \qquad \textit{and} \qquad U_X(z) \;\ge\; 1-\eta, \qquad \eta < (1-2\epsilon)\,\Delta\,\frac{a}{T}.$$

Then $D_\kappa(z) = \text{MARKED}$. In particular no attacker whatsoever — offline, randomised, unbounded in computation, and possibly holding the key, detector, and list of evasive trajectories — can produce an unmarked output satisfying both the stated final-distortion and utility conditions.

Proof in Appendix B.9.

Comparing the two guarantees. Theorem D.1 is the output-only statement: it covers offline editors without constraining their intermediate states, conditional on the final artefact satisfying both stated bounds. Its price is that the final distortion must be at most γT. Theorem 4.4 says something different and sharper: an attacker that additionally has to remain deployable at every step never even reaches distortion γT — it is pinned at $\zeta T = \eta T/((1-2\epsilon)\Delta)$, which is smaller by the factor $\eta/((1-2\epsilon)\Delta\gamma)$. Thus the per-step condition yields an additional, quantitative confinement conclusion, while Theorem D.1 supplies the corresponding output-only lower bound.

Definition A.11 (Erasure bottleneck)**.** Let $a < b$ be integers. (X, D_κ, z^0, q) satisfies the (a, b)-*erasure bottleneck* if

(i) **abundance:** $|\{z \in \mathrm{Sol}(X) : D_\kappa(z) = \text{UNMARKED}\}| \geq 2^{\Omega(T)}$;

(ii) **basin fidelity:** $\mathrm{dist}(z, z^0) \leq a \implies D_\kappa(z) = D_\kappa(z^0) = \text{MARKED}$;

(iii) **gap:** $\{z : a < \mathrm{dist}(z, z^0) \leq b\} \cap V(q) = \varnothing$.

Theorem A.12 (Transfer)**.** *If (X, D_κ, z^0, q) satisfies the (a, b)-erasure bottleneck and $r \leq b - a$, then no realisation of any attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ started at z^0 succeeds; the success probability is exactly* 0.

Proof in Appendix B.11.

Only (ii) and (iii) are used; (i) is the non-triviality condition, without which (ii) and (iii) hold vacuously for the constant detector $D_\kappa \equiv \text{MARKED}$ and the statement says nothing. Theorem A.12 applied with a the detector radius and $b = \lfloor \gamma T \rfloor$ recovers Corollary A.9 with the slightly weaker radius bound $r \leq \lfloor \gamma T \rfloor - a$; the abstract version also applies to schemes that establish a bottleneck without a monotone barrier.

Appendices H.9, H.7, and H.8 develop enforcement models, finite-scale constants, and the abstract barrier interface, respectively.

The same ceiling applies to the output-fidelity bound (Remark I.4).

A.4 BasinMark: Marking the Free Coordinates

We now instantiate with parity, $R_j = \{y : \bigoplus_p y_p = b_j\}$, so $\mathrm{Sol}(X) = \{z : Hz = b\}$ for the Δ-sparse incidence matrix $H \in \mathbb{F}_2^{n \times T}$. Assuming consistency, $\mathrm{Sol}(X) = z_0 + \ker H$ is an affine code of dimension $d = T - \mathrm{rank}(H) \geq (1 - c/\Delta)T > 0$.

Information set. Let J be the set of pivot columns of the row-reduced echelon form of H, so H_J has full column rank and $\mathrm{col}(H_J) = \mathrm{col}(H)$, and set $I = [T] \setminus J$, $|I| = d$. Then $z \mapsto z|_I$ is a bijection $\mathrm{Sol}(X) \to \mathbb{F}_2^d$, computable and invertible once per task in $O(T^3)$ by Gaussian elimination. We call $u = z|_I$ the *strategy word*: the genuinely free part of the plan. *I is derived from the task instance alone, before any watermarking.* This is what makes the marked classes task-grounded rather than clusters fitted after the fact.

Definition A.13 (BasinMark)**.** A PRF on key κ, indexed by a counter, yields $M_\kappa \in \mathbb{F}_2^{\ell \times d}$ and $s_\kappa \in \mathbb{F}_2^\ell$ with $\ell < d$; the counter is advanced until M_κ has full row rank, so full rank holds by construction. *Generation:* sample u uniformly from $S_\kappa = \{u : M_\kappa u = s_\kappa\}$ and output the unique $z \in \mathrm{Sol}(X)$ with $z|_I = u$. *Detection:* run the decoder of Proposition A.14 on the syndrome $\sigma = Hz \oplus b$ to obtain an error pattern $\hat{e}$, and set $\hat{c} = z \oplus \hat{e}$. If the decoder does not terminate with $H\hat{c} = b$, output ABSTAIN; otherwise output MARKED if $M_\kappa(\hat{c}|_I) = s_\kappa$ and UNMARKED if not.

Detection is three-valued by design. An abstention is neither a provenance claim nor an erasure, so an attacker that manufactures an uncorrectable syndrome is credited with neither — which is the only thing a fail-closed convention would have bought. Collapsing abstention into MARKED would be worse than cosmetic: outside the decoder's convergence basin the resulting detector *is* the constant

detector $D_\kappa \equiv$ MARKED, which is exactly the degenerate object condition (i) of Definition A.11 exists to exclude.

The complete generation and detection procedure is Algorithm 1.

Algorithm 1: BasinMark: mark the information set of the task's own solution space

Input: task instance X with parity check H and syndrome b; information set I with $|I| = d$; key κ; payload width $\ell < d$.

Key setup `// once per key`
Draw $(M_\kappa, s_\kappa) \in \mathbb{F}_2^{\ell \times d} \times \mathbb{F}_2^\ell$ from the PRF on κ, advancing the counter until $\mathrm{rank}(M_\kappa) = \ell$
Generation `// cost: one Gaussian elimination`
Sample u uniformly from $S_\kappa = \{u : M_\kappa u = s_\kappa\}$
return the unique $z \in \mathrm{Sol}(X)$ with $z|_I = u$ `//` $U_X(z) = 1$ `exactly`
Detection(z) `// three-valued`
$\sigma \leftarrow Hz \oplus b$; run the flipping decoder of Proposition A.14 on σ to get $\hat{e}$; $\hat{c} \leftarrow z \oplus \hat{e}$
if *the decoder does not terminate with* $H\hat{c} = b$ **then return** ABSTAIN
if $M_\kappa(\hat{c}|_I) = s_\kappa$ **then return** MARKED
else return UNMARKED

Proposition A.14 (Sequential flipping decoder)**.** *Let every R_j be locally sensitive and let G_X be a $(\gamma T, (1-\epsilon)c)$-expander with $\epsilon < 1/4$. The sequential bit-flipping decoder — repeatedly flip any single decision whose flip* strictly decreases *the number of violated constraints — corrects every error pattern of weight strictly less than $a_{\mathrm{SS}} := (1-2\epsilon)\lfloor \gamma T \rfloor$, in $O(T)$ flips and $O(cT + c^2 \Delta\, a_{\mathrm{SS}})$ total work with incremental syndrome updates (Sipser & Spielman, 1996; Zémor, 2001).*

Sipser and Spielman prove the linear-time guarantee for expansion $3c/4$, correcting an $\alpha/2$ fraction of errors when sets of up to an α fraction expand (Sipser & Spielman, 1996, Thm. 10); the proof in Appendix B.12 follows their argument and gives the radius $(1-2\epsilon)\lfloor \gamma T \rfloor$ for every $\epsilon < 1/4$, and Zémor (2001) analyses a variant of their decoder for bipartite-graph codes.

For parity the rule coincides with the familiar one, "flip any decision lying in more violated than satisfied constraints": flipping i toggles all c_i of its constraints, so the violation count changes by exactly $c_i - 2v$ when i lies in v violated constraints, and $v > c_i/2$ is equivalent to a strict decrease. For a general locally sensitive R_j only one direction survives. A satisfied constraint at i must become violated, but a violated one need not become satisfied, so the change lies in $[c_i - 2v,\ c_i - v]$ and both endpoints occur — with 1-in-3 and $z|_{C_j} = 111$, every single flip leaves the constraint violated. Stating the rule in terms of the violation count rather than the violated-constraint degree is what keeps the proposition true outside parity.

Proof in Appendix B.12.

Write a for the largest integer radius at which the detector is certified, so that $\mathrm{dist}(z, z^0) \le a$ implies correct decoding; $a = \lceil a_{\mathrm{SS}} \rceil - 1$ for Proposition A.14 (whose guarantee is strict), and $a = \lceil d_{\min}/2 \rceil - 1$ for minimum-distance decoding, which is not efficient in general. Since a decoder correct at radius a forces $d_{\min} \ge 2a + 1$, the output within the radius is the strict unique nearest codeword. Section 5 measures a directly on the finite instances.

Proposition A.15 (Utility, distribution, false positives)**.** *Assume throughout that s_κ is uniform and independent of M_κ — automatic under Definition A.13, since the counter is advanced on the rank of M_κ alone — and that $S_\kappa \neq \varnothing$. Then: every BasinMark output lies in* $\mathrm{Sol}(X)$*, so $U_X = 1$ exactly; for a* single *output the law of u, marginalised over the key, is exactly uniform on $\mathbb{F}_2^d$; and for any z produced independently of κ, $\Pr_\kappa[D_\kappa(z) = \text{MARKED}] \le 2^{-\ell}$, with equality exactly when the decoder converges on z.*

Proof in Appendix B.13.

The scheme costs exactly ℓ bits of strategy entropy out of $d = \Theta(T)$ and is undetectable from one sample. Its key-independent and fixed-key properties separate as follows.

Key-independent false positives. With three-valued detection, $\Pr_\kappa[D_\kappa(z) = \text{MARKED}] \le 2^{-\ell}$ for *every* artefact produced independently of the key, not only for valid ones — $\approx 9 \times 10^{-13}$ at $\ell = 40$. This is the benefit of keeping abstention distinct from acceptance.

Fixed-key multi-sample behavior. Proposition A.15 gives *exact* uniformity for a single output, and for a single output only. Under a fixed key all outputs lie in one coset $S_\kappa = \{u : M_\kappa u = s_\kappa\}$ of $\ker M_\kappa$, an affine subspace of *dimension* $d - \ell$, and the three adversarial goals then separate sharply.

(i) Forgery costs three samples. Over $\mathbb{F}_2$, $M_\kappa(u_1 \oplus u_2 \oplus u_3) = s_\kappa \oplus s_\kappa \oplus s_\kappa = s_\kappa$, and $\mathrm{Sol}(X)$ is affine, so for any three outputs $z^* := z_1 \oplus z_2 \oplus z_3$ lies in $\mathrm{Sol}(X)$, has $U_X = 1$, and is accepted — with no knowledge of κ and no linear algebra at all. It is new unless two of the three coincide, of probability $O(2^{-(d-\ell)})$. Two samples give nothing, the affine hull of two points over $\mathbb{F}_2$ being those two points; m affinely independent samples give the 2^{m-1} points of their affine hull, a $2^{m-1-(d-\ell)}$ fraction of S_κ. The resulting affine forgery remains marked and therefore does not constitute the erasure considered in Corollaries A.9 and A.17.

(ii) Recovering the detector costs $d - \ell + 1$ affinely independent samples, since they must affinely *span* S_κ: after samples of affine rank k the number of consistent detectors is the Gaussian binomial $\begin{bmatrix} d-k \\ \ell \end{bmatrix}_2$, equal to 1 precisely when $k = d - \ell$. At $d = 1000$, $\ell = 40$, taking $m = \ell + 1$ leaves about 2^{36802} candidates and $m = d - \ell + 1$ leaves one. What is recovered is S_κ, never the pair: all pairs sharing the row space of $[M_\kappa \mid s_\kappa]$ induce the same detector.

(iii) Distinguishing costs $d - \ell + \Theta(1)$ samples. For uniform (M_κ, s_κ) the likelihood ratio of m marked outputs against m uniform ones depends on the tuple only through its affine rank k, equalling $2^{\ell(m-1)} \prod_{i<\ell}(2^{d-k} - 2^i)/(2^d - 2^i)$. The induced total variation distance is approximately $2^{-(d-\ell+1-m)}$ while $m \ll d - \ell$ — 9×10^{-10} at $d = 50$, $\ell = 10$, $m = \ell + 1$, and 4×10^{-6} at $d = 30$, $\ell = 6$ — rising to 0.42, 0.71 and ≈ 1 at $m = d - \ell$, $d - \ell + 1$ and $d - \ell + 2$.

The deployment bound is therefore $\Theta(d - \ell)$ outputs per key against a distinguisher, but only three against a forger. Both follow from reusing this affine carrier.

Distribution preservation is relative to the uniform law on $\mathrm{Sol}(X)$, rather than to an arbitrary agent policy over $\mathrm{Sol}(X)$.

Theorem A.16 (Abundance: the mark is maximally removable)**.** *With $c < \Delta$ and $\ell < d$,*

$$\big|\{z \in \mathrm{Sol}(X) : D_\kappa(z) = \textsc{unmarked}\}\big| \;=\; 2^d - 2^{d-\ell} \;=\; 2^d\big(1 - 2^{-\ell}\big), \qquad d \geq (1 - c/\Delta)T,$$

and every one of these trajectories has utility exactly 1.

Proof in Appendix B.14.

Why abundance matters. A $1 - 2^{-\ell}$ fraction of *all* perfectly valid trajectories evade the detector. Protection therefore comes from the attack and distortion constraints, rather than from an absence of evasive outputs. The guarantees allow unbounded computation within those constraints. This makes Corollary A.9 a separation rather than a vacuous one: combining the two — Corollary A.17 below — there are $2^d(1 - 2^{-\ell})$ places the attacker would be happy to reach and it can reach none of them.

Corollary A.17 (The separation)**.** *Let G_X be a $(\gamma T, (1-\epsilon)c)$-expander with $\epsilon < 1/2$, exact c-regularity, $c < \Delta$, $1 \leq \ell < d$, and let z^0 be a BasinMark output detected at radius a. If $\eta < (1 - 2\epsilon)\Delta a/T$, $\zeta := \eta/((1-2\epsilon)\Delta) < \gamma$ and $r \leq (\gamma - \zeta)T$, then $2^d(1 - 2^{-\ell})$ trajectories of utility exactly 1 evade the detector, and every attack in $\mathcal{A}_{\mathrm{loc}}(1-\eta, r)$ started at z^0 reaches none of them on any realisation — its success probability is exactly 0 — even given unbounded computation, the key, the detector, and the explicit list of all of them.*

Only Proposition A.14 needs $\epsilon < 1/4$; everything else here needs $\epsilon < 1/2$ together with a certified radius (Remark I.6).

A.5 Upper Bounds: Where the Barrier Stops

Throughout this section we retain the parity specialisation and locate the complementary attack regimes. Let $D^* = D^*(X, \kappa) = \min\{\mathrm{dist}(z^0, z) : z \in \mathrm{Sol}(X), D_\kappa(z) = \textsc{unmarked}\}$; by Corollary A.6, $D^* > \gamma T$. Define $r_c(q) = \inf\{r :$ some attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ started at z^0 succeeds with positive probability$\}$, a genuine threshold because $\mathcal{A}_{\mathrm{loc}}(q, r)$ is monotone in r.

Theorem A.18 (Locality threshold)**.** *Assume $1 \leq \ell < d$, so the evasive set is non-empty and D^* is defined. If $r \geq D^*$ then a single move erases the watermark while keeping utility exactly*

1; *hence* $r_c(q) \le D^*$ *unconditionally. If in addition the hypotheses of Corollary A.9 hold, then* $(\gamma - \zeta)T < r_c(q) \le D^*$.

Proof in Appendix B.15.

This is an existence threshold: finding a *nearest* evasive codeword is a separate minimum-weight problem for $\ker H$. The unrestricted problem is NP-hard, and no polynomial-time algorithm is known for the structured case. Theorem A.18 concerns the computationally unbounded class $\mathcal{A}_{\mathrm{loc}}$.

Theorem A.19 (Straight-line attack)**.** *Assume exact c-regularity. Let z' be any evasive valid trajectory, $e = z^0 \oplus z'$, $D = |e|$. Flip the coordinates of* $\mathrm{supp}(e)$ *one at a time, in any order. Every intermediate point z, having flipped the subset $S \subseteq \mathrm{supp}(e)$, satisfies*

$$E_X(z) \;\le\; c \cdot \min\big(|S|,\, D - |S|\big) \;\le\; c\lfloor D/2 \rfloor.$$

Consequently, if $\eta \ge \Delta D/(2T)$, this $r = 1$ attack lies in $\mathcal{A}_{\mathrm{loc}}(1-\eta, 1)$ and succeeds. Given z', it performs D single-bit flips in $O(Dc\Delta)$ time with incremental syndrome updates.

Proof in Appendix B.16.

Corollary A.20 (Two-sided phase diagram)**.** *Assume the hypotheses of Theorem A.19 and of Corollary A.9. With $\eta = 1 - q$ and $\zeta = \eta/((1-2\epsilon)\Delta)$:*

- ***(I)*** *if $\eta < (1-2\epsilon)\Delta a/T$ and $r \le (\gamma - \zeta)T$, the success probability is exactly 0;*
- ***(II)*** *if $\eta \ge \Delta D^*/(2T)$, an explicit $r = 1$ attack succeeds, in linear time once a minimum-weight evasive codeword is supplied.*

The two windows never overlap. Their ratio is

$$\frac{\eta_{\mathrm{att}}}{\eta_{\mathrm{sec}}} = \frac{\Delta D^*/(2T)}{(1-2\epsilon)\Delta a/T} = \frac{D^*}{2(1-2\epsilon)\,a} \;\ge\; \frac{d_{\min}}{(1-2\epsilon)(d_{\min}-1)} \;>\; \frac{1}{1-2\epsilon} \;\ge\; 1,$$

using $D^* \ge d_{\min}$ and, since a decoder certified at radius a forces $d_{\min} \ge 2a+1$, $a \le (d_{\min}-1)/2$.

Condition (II) is informative only when $D^* \le 2T/\Delta$. At $\Delta = 10$, the cutoff is $0.2T$. Table 6's median searched-target distances range from $0.185T$ to $0.201T$; they describe the target sampler, not the true nearest evasive distance D^*.

Appendices H.10 and H.11 further separate the straight-line attack cost from the respective roles of the key and task geometry.

The gap between the two windows, and what closes it, is Remark I.7.

A.6 Thermal Attackers: Paying for Violations

The hard utility constraint $V(q)$ can instead be relaxed into an energy penalty: the attacker may break constraints but pays $e^{-\beta}$ per violation. Let P_β be the single-flip Metropolis chain, reversible with respect to $\pi_\beta(z) \propto e^{-\beta E_X(z)}$.

Lemma A.21 (Escape bound)**.** *Let P have stationary distribution π with $\pi(x_0) > 0$, let $A \subseteq \Omega$ with $x_0 \in A$, and let $T_{A^c} = \inf\{s \ge 0 : X_s \notin A\}$. Then for every $t \ge 0$,*

$$\Pr_{x_0}\big[T_{A^c} \le t\big] \;\le\; \frac{t\,Q(A, A^c)}{\pi(x_0)}, \qquad Q(A, A^c) = \sum_{x \in A,\, y \notin A} \pi(x) P(x, y).$$

Proof in Appendix B.17.

This is the stationarity argument behind the bottleneck-ratio bound (Levin & Peres, 2017, Thm. 7.4), written for hitting rather than mixing.

The quantity we bound is the *basin-escape* time T_{A^c}, which lower-bounds the erasure hitting time because A contains no undetected trajectory.

Theorem A.22 (Exponential basin escape)**.** *Assume every R_j is locally sensitive, G_X is a $(\gamma T, (1-\epsilon)c)$-expander with $\epsilon < 1/2$, and $z^0 \in \mathrm{Sol}(X)$ with $D_\kappa(z^0) = \textsc{marked}$. Let* Dec *be certified at radius a, so D_κ is constant on the ball of radius a about z^0. Fix any integer w with $1 \le w \le \min(\gamma T,\, a+1)$, let $A = \{z : \mathrm{dist}(z, z^0) < w\}$, and let $B = \{z : D_\kappa(z) = \textsc{unmarked}\}$ — this is* every *trajectory the detector actively rejects, not merely every rejected codeword, since in this model validity is priced rather than required, so it includes the whole decoding ball around each evasive codeword. The identical proof bounds the hitting time of $B' = \{z : D_\kappa(z) \neq \textsc{marked}\}$ as well, since $A \cap B' = \varnothing$ by the same line, so the result covers denial of attribution too. Then $A \cap B = \varnothing$ and, with H the natural-logarithm binary entropy,*

$$\Pr_{z^0}\big[T_B \le t\big] \;\le\; t\exp\Big(T\,\mathsf{H}(w/T) - \beta(1-2\epsilon)c\,w\Big).$$

Hence for every fixed $\beta > \beta_c := \dfrac{T\,\mathsf{H}(w/T)}{(1-2\epsilon)c\,w}$, setting $\delta = \beta(1-2\epsilon)cw/T - \mathsf{H}(w/T) > 0$ and $t = e^{\delta T/2}$ gives $T_B > e^{\delta T/2}$ with probability at least $1 - e^{-\delta T/2}$. Taking $w = \lceil \theta T \rceil$ for a fixed $\theta < \min(\gamma,\, a/T)$ — available with θ up to $(1-2\epsilon)\gamma$ whenever $\epsilon < 1/4$, since then $a = \lceil a_{\mathrm{SS}} \rceil - 1$ is itself $\Theta(T)$ — makes $\beta_c \to \mathsf{H}(\theta)/((1-2\epsilon)c\theta)$, a constant independent of T. This is an asymptotic consequence under a linear certified radius; it does not certify the larger crossing radius used in our finite experiments.

Proof in Appendix B.18.

Finite-scale diagnostic. The same argument gives a sharper sufficient threshold in terms of the minimum energy on the crossing sphere:

$$\beta_c \;=\; \frac{T\,\mathsf{H}(w/T)}{E_{\min}(w)}, \qquad w = \lfloor \gamma T/2 \rfloor + 1,$$

where $E_{\min}(w) = \min_{|e|=w} E_X(z^0 \oplus e)$. In the experiment we substitute the smallest value found by search and obtain plug-in values between 2.6 and 4.0 (median 3.1) across $T = 80$ to 320, with no visible trend in T. Because the runs use $w = 0.05T > a + 1 = 3$, they diagnose scaling outside the theorem's certified crossing radius. Moreover, the searched energy upper-bounds the unknown minimum, so the resulting β_c is a *lower* plug-in estimate of the corresponding threshold. In the accessible regime, the escape time follows power laws whose exponent climbs from 0.8 to 2.7 as β approaches β_c; the exponentially slow regime lies beyond the sampling budget. The theorem gives a sufficient condition for a *fixed-temperature* single-flip chain. Time-inhomogeneous schedules such as simulated annealing, and converses below β_c, remain open.

B Theoretical Results

This appendix collects every proof deferred from the main text, in the order the statements appear there. Longer statements are reproduced for continuity; shorter propositions are referenced directly from the main text.

B.1 Proof of Lemma A.4

Restated (Lemma A.4: Overlap expansion)**.** *Suppose every decision lies in* exactly c *constraints and every pair of decisions lies in at most μ common constraints. Then $|\Gamma(S)| \ge c|S| - \mu\binom{|S|}{2}$ for every $S \subseteq [T]$, so G_X is a $(w, (1-\epsilon)c)$-expander for every $w \le 2\epsilon c/\mu + 1$.*

Proof. Bonferroni: $|\Gamma(S)| = |\bigcup_{i\in S}\Gamma(\{i\})| \ge \sum_{i\in S}|\Gamma(\{i\})| - \sum_{i<i'\in S}|\Gamma(\{i\}) \cap \Gamma(\{i'\})| \ge c|S| - \mu\binom{|S|}{2}$. This is $\ge (1-\epsilon)c|S|$ iff $\mu(|S|-1)/2 \le \epsilon c$. □

B.2 Proof of Lemma A.5

Restated (Lemma A.5: Unique neighbour)**.** *Let G be bipartite with left vertex set $[T]$, every left vertex of degree at most c, and let $S \subseteq [T]$ satisfy $|\Gamma(S)| \ge (1-\epsilon)c|S|$. Then the number u of right vertices with exactly one neighbour in S satisfies*

$$u \;\ge\; 2|\Gamma(S)| - c|S| \;\ge\; (1-2\epsilon)\,c\,|S|.$$

Proof. At most $c|S|$ edges leave S. Every neighbour that is not unique absorbs at least two of them, so $c|S| \geq u + 2(|\Gamma(S)| - u) = 2|\Gamma(S)| - u$, giving $u \geq 2|\Gamma(S)| - c|S| \geq 2(1-\epsilon)c|S| - c|S|$. □

B.3 Proof of Lemma 4.3

Restated (Lemma 4.3: Barrier)**.** *Suppose every R_j is locally sensitive, every decision has degree at most c, and G_X is a $(\gamma T, (1-\epsilon)c)$-expander with $\epsilon < 1/2$. Let $z^* \in \mathrm{Sol}(X)$ and let $w = \mathrm{dist}(z, z^*)$ satisfy $0 < w \leq \gamma T$. Then*

$$E_X(z) \;\geq\; (1 - 2\epsilon)\, c\, w.$$

Proof. Let $S = \{i : z_i \neq z_i^*\}$, so $|S| = w \leq \gamma T$ and expansion applies. Let j be a constraint with exactly one neighbour $i_0 \in S$; Lemma A.5 supplies at least $(1-2\epsilon)cw$ of them. Then $z|_{C_j}$ differs from $z^*|_{C_j}$ in the single coordinate of C_j equal to i_0. Since $z^*|_{C_j} \in R_j$ and R_j is locally sensitive, $z|_{C_j} \notin R_j$: constraint j is violated. Distinct such j are distinct violated constraints. □

B.4 Proof of Corollary A.6

Restated (Corollary A.6: Linear minimum distance)**.** *Under the hypotheses of Lemma 4.3, any two distinct $z^*, z' \in \mathrm{Sol}(X)$ satisfy $\mathrm{dist}(z^*, z') \geq \lfloor \gamma T \rfloor + 1$.*

Proof. Suppose $1 \leq w := \mathrm{dist}(z^*, z') \leq \gamma T$. Lemma 4.3 with $z = z'$ gives $E_X(z') \geq (1-2\epsilon)cw \geq (1-2\epsilon)c > 0$ using $w \geq 1$ and $\epsilon < 1/2$; since E_X is integer valued, $E_X(z') \geq 1$, contradicting $z' \in \mathrm{Sol}(X)$. □

B.5 Proof of Lemma A.7

Restated (Lemma A.7: Extremal classification; classical)**.** *Let $\Delta \geq 2$ and let $R \subseteq \{0,1\}^\Delta$ be locally sensitive. Then $|R| \leq 2^{\Delta-1}$, and equality holds if and only if R is one of the two parity classes $\{y : \bigoplus_p y_p = b\}$, $b \in \{0,1\}$.*

Proof. Q_Δ is Δ-regular and bipartite, hence has a perfect matching, and an independent set meets each matching edge at most once, giving $|R| \leq 2^{\Delta-1}$. Conversely each parity class has minimum distance 2 and cardinality $2^{\Delta-1}$, so both attain the bound.

For the equality case, let $|R| = 2^{\Delta-1}$. Every 2-dimensional subcube is a 4-cycle, so it meets R in at most two vertices, and if in exactly two then in a diagonal pair — two vertices of the same parity. Count incidences: each vertex lies in $\binom{\Delta}{2}$ subcubes and there are $\binom{\Delta}{2}2^{\Delta-2}$ subcubes, so

$$\sum_{Q \text{ 2-subcube}} |R \cap Q| \;=\; |R|\binom{\Delta}{2} \;=\; 2^{\Delta-1}\binom{\Delta}{2} \;=\; 2 \cdot \binom{\Delta}{2}2^{\Delta-2},$$

i.e. the average of $|R \cap Q|$ over subcubes is exactly 2. Since no term exceeds 2, every term equals 2: every 2-subcube meets R in a same-parity diagonal pair. Now fix $x \in R$. For any $i \neq j$, the subcube spanned by e_i, e_j at x meets R in two vertices, one of which is x; the other must have the same parity as x, so it is $x \oplus e_i \oplus e_j$. Hence R is closed under adding $e_i \oplus e_j$. These vectors generate the even-weight subspace, so R contains the whole parity class of x, which already has $2^{\Delta-1}$ elements. Therefore R is that parity class. □

B.6 Proof of Theorem 4.4

Restated (Theorem 4.4: Confinement)**.** *Assume the hypotheses of Lemma 4.3 with exact c-regularity, let $z^0 \in \mathrm{Sol}(X)$, and let $\eta = 1 - q$ be the utility slack. Put*

$$\zeta \;:=\; \frac{\eta}{(1-2\epsilon)\Delta}\,.$$

If $\zeta < \gamma$ and $r \leq (\gamma - \zeta)T$, then every attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ started at z^0 satisfies $\mathrm{dist}(z^t, z^0) \leq \zeta T$ for all t, on every realisation.

Proof. Induction on t. For $t = 0$, $\mathrm{dist}(z^0, z^0) = 0 \le \zeta T$. Assume $\mathrm{dist}(z^t, z^0) \le \zeta T$. By r-locality, $\mathrm{dist}(z^{t+1}, z^0) \le \zeta T + r \le \gamma T$. Suppose $\mathrm{dist}(z^{t+1}, z^0) > \zeta T$. Since $0 < \mathrm{dist}(z^{t+1}, z^0) \le \gamma T$, Lemma 4.3 applies with $z^* = z^0$ and gives

$$E_X(z^{t+1}) \;>\; (1-2\epsilon)\, c\, \zeta T \;=\; (1-2\epsilon)\, c\, T \cdot \frac{\eta}{(1-2\epsilon)\Delta} \;=\; \eta\, \frac{cT}{\Delta} \;=\; \eta\, n,$$

contradicting (b), which requires $E_X(z^{t+1}) \le \eta n$. Hence $\mathrm{dist}(z^{t+1}, z^0) \le \zeta T$. □

B.7 Proof of Corollary A.9

Restated (Corollary A.9: Local erasure is impossible)**.** *In addition, let D_κ be a detector that is constant on the ball of radius a about z^0, with $\zeta T < a$, and let $D_\kappa(z^0) =$ MARKED. Then under the hypotheses of Theorem 4.4 every attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ fails, on every realisation; the success probability is exactly 0. Equivalently, erasure is impossible whenever*

$$\eta \;<\; (1-2\epsilon)\,\Delta\,\frac{a}{T} \qquad \textit{and} \qquad r \;\le\; \Big(\gamma - \frac{\eta}{(1-2\epsilon)\Delta}\Big)T.$$

Proof. By Theorem 4.4 every z^t lies in the ball of radius $\zeta T < a$, on which D_κ is constant and equal to $D_\kappa(z^0) =$ MARKED. The first display in (1) is $\zeta T < a$ rearranged. □

B.8 Proof of Corollary A.10

Restated (Corollary A.10: Perfect-utility outputs)**.** *Assume the hypotheses of Lemma 4.3 with $z^* = z^0 \in \mathrm{Sol}(X)$. If an attacker outputs any $z \in \mathrm{Sol}(X)$ with $\mathrm{dist}(z, z^0) \le \gamma T$, then $z = z^0$. Erasure therefore requires changing more than γT decisions, whatever the attacker does in between and however long it computes.*

Proof. Immediate from Corollary A.6. □

B.9 Proof of Theorem D.1

Restated (Theorem D.1: Output-fidelity bound; no per-step condition)**.** *Assume the hypotheses of Lemma 4.3 with $z^* = z^0 \in \mathrm{Sol}(X)$, let $D_\kappa(z^0) =$ MARKED, and let* Dec *be certified at radius a. Let z be* any *trajectory with*

$$\mathrm{dist}(z, z^0) \;\le\; \gamma T \qquad \textit{and} \qquad U_X(z) \;\ge\; 1-\eta, \qquad \eta < (1-2\epsilon)\,\Delta\,\frac{a}{T}.$$

Then $D_\kappa(z) =$ MARKED. In particular no attacker whatsoever — offline, randomised, unbounded in computation, and possibly holding the key, detector, and list of evasive trajectories — can produce an unmarked output satisfying both the stated final-distortion and utility conditions.

Proof. If $\mathrm{dist}(z, z^0) \le a$ then $\mathrm{Dec}(z) = z^0$ and $D_\kappa(z) =$ MARKED. Otherwise $a < \mathrm{dist}(z, z^0) \le \gamma T$, so Lemma 4.3 gives $E_X(z) \ge (1-2\epsilon) c\, \mathrm{dist}(z, z^0) > (1-2\epsilon) c\, a$, hence $E_X(z)/n > (1-2\epsilon)\Delta a/T > \eta$, contradicting $U_X(z) \ge 1-\eta$. □

B.10 Proof of Proposition D.2

Proof. The utility condition gives $E_X(z) \le \eta n < L_{a,b}(z^0)$. Hence an output with $\mathrm{dist}(z, z^0) \le b$ cannot lie in the annulus $a < \mathrm{dist}(z, z^0) \le b$ and must lie in the marked radius-a ball. For the path statement, suppose inductively that z^t lies in that ball. The next point cannot lie in the annulus by the same utility argument, and reaching distance greater than b from distance at most a would require a step larger than $b - a$. Thus every point remains within radius a and is marked. □

B.11 Proof of Theorem A.12

Restated (Theorem A.12: Transfer)**.** *If (X, D_κ, z^0, q) satisfies the (a, b)-erasure bottleneck and $r \le b - a$, then no realisation of any attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ started at z^0 succeeds; the success probability is exactly* 0.

Proof. By (ii) applied to z^0, $D_\kappa(z^0) = \textsc{marked}$, so the $N = 0$ attack fails. Fix a realisation with $D_\kappa(z^N) = \textsc{unmarked}$; by (ii), $\mathrm{dist}(z^N, z^0) > a$, so $t^* = \min\{t : \mathrm{dist}(z^t, z^0) > a\}$ is well defined and $t^* \ge 1$, with $\mathrm{dist}(z^{t^*-1}, z^0) \le a$. By r-locality, $\mathrm{dist}(z^{t^*}, z^0) \le a + r \le b$. Thus $a < \mathrm{dist}(z^{t^*}, z^0) \le b$ while $U_X(z^{t^*}) \ge q$ by (b), contradicting (iii). □

B.12 Proof of Proposition A.14

Restated (Proposition A.14: Sequential flipping decoder)**.** *Let every R_j be locally sensitive and let G_X be a $(\gamma T, (1-\epsilon)c)$-expander with $\epsilon < 1/4$. The sequential bit-flipping decoder — repeatedly flip any single decision whose flip* strictly decreases *the number of violated constraints — corrects every error pattern of weight strictly less than $a_{\mathrm{SS}} := (1 - 2\epsilon)\lfloor \gamma T \rfloor$, in $O(T)$ flips and $O(cT + c^2 \Delta\, a_{\mathrm{SS}})$ total work with incremental syndrome updates (Sipser & Spielman, 1996; Zémor, 2001).*

Proof. Each flip strictly decreases the number of violated constraints, by the flip rule; hence the count never increases and the process terminates. Let S_t be the error support after t flips, with $|S_0| < a_{\mathrm{SS}}$. Initially at most $c|S_0|$ constraints are violated, and this count never increases, so it always stays below $c\, a_{\mathrm{SS}} = (1 - 2\epsilon)c\lfloor \gamma T \rfloor$.

Invariant $|S_t| \le \lfloor \gamma T \rfloor$. Since $|S_t|$ changes by one per flip, if it ever exceeded $\lfloor \gamma T \rfloor$ it would first equal $\lfloor \gamma T \rfloor$; but then Lemma 4.3 would force at least $(1 - 2\epsilon)c\lfloor \gamma T \rfloor$ violated constraints, contradicting the previous paragraph.

Progress. Suppose $S_t \ne \varnothing$. By the invariant, Lemma A.5 applies. Write u_i for the number of constraints whose unique element of S_t is i; each such constraint is violated, by local sensitivity, and $\sum_{i \in S_t} u_i \ge (1 - 2\epsilon)c|S_t|$, so some $i \in S_t$ has $u_i \ge (1 - 2\epsilon)c$. Flipping that i makes $z|_{C_j} = z^*|_{C_j} \in R_j$ for every one of those u_i constraints, repairing all of them, while each of the at most $c_i - u_i$ remaining constraints at i can at worst become violated. Hence the violation count changes by at most $c_i - 2u_i \le c - 2u_i \le (4\epsilon - 1)c < 0$, using $c_i \le c$ and $\epsilon < 1/4$. So a strictly improving flip exists and the decoder has not halted. Hence at termination $S_t = \varnothing$ and the decoder outputs the transmitted codeword. The decoder maintains, for each decision, the change in violation count its flip would cause; flipping i dirties only the c_i constraints containing i, so all $c_i\Delta$ affected entries are refreshed in $O(c\Delta)$, and a bucket of strictly-improving decisions supplies the next flip in $O(1)$. Each flip costs $O(c\Delta)$ with incremental syndrome updates and an $O(1)$ membership oracle for each R_j (immediate for parity, and obtained in general by tabulating each relation in $O(2^\Delta)$ preprocessing, Δ constant); without the oracle both terms gain a factor Δ. The number of flips is at most the initial violation count $c|S_0| < c\, a_{\mathrm{SS}}$, so the total is $O(cT + c^2\Delta\, a_{\mathrm{SS}})$ including the initial syndrome pass. □

B.13 Proof of Proposition A.15

Restated (Proposition A.15: Utility, distribution, false positives)**.** *Assume throughout that s_κ is uniform and independent of M_κ — automatic under Definition A.13, since the counter is advanced on the rank of M_κ alone — and that $S_\kappa \ne \varnothing$. Then: every BasinMark output lies in* $\mathrm{Sol}(X)$*, so $U_X = 1$ exactly; for a* single *output the law of u, marginalised over the key, is exactly uniform on $\mathbb{F}_2^d$; and for any z produced independently of κ, $\Pr_\kappa[D_\kappa(z) = \textsc{marked}] \le 2^{-\ell}$, with equality exactly when the decoder converges on z.*

Proof. The first claim is immediate from Definition A.13. For the second, conditioned on M_κ the vector s_κ is uniform, so S_κ is a uniformly random coset of $\ker M_\kappa$; u is uniform on that coset, and averaging over a uniform coset gives the uniform law on $\mathbb{F}_2^d$. For the third, the decoder is a function of (H, b, z) alone, so "the decoder converges on z" is a κ-independent event — which is what prevents an adversary from steering the abstention decision using knowledge of the key. On that event $\hat{c}$ is

a fixed codeword and s_κ is uniform and independent of $\hat{c}|_I$, so the ℓ keyed parities all match with probability exactly $2^{-\ell}$; off it the output is ABSTAIN, of probability 0. □

B.14 Proof of Theorem A.16

Restated (Theorem A.16: Abundance: the mark is maximally removable). *With $c < \Delta$ and $\ell < d$,*

$$\big|\{z \in \mathrm{Sol}(X) : D_\kappa(z) = \textsc{unmarked}\}\big| \;=\; 2^d - 2^{d-\ell} \;=\; 2^d\big(1 - 2^{-\ell}\big), \qquad d \geq (1 - c/\Delta)T,$$

and every one of these trajectories has utility exactly 1.

Proof. For $z \in \mathrm{Sol}(X)$ the decoder returns z itself, so $D_\kappa(z) = \textsc{marked}$ iff $z|_I \in S_\kappa$. The bijection $z \mapsto z|_I$ and $|S_\kappa| = 2^{d-\ell}$ give the count; $c < \Delta$ makes the exponent linear in T. □

B.15 Proof of Theorem A.18

Restated (Theorem A.18: Locality threshold). *Assume $1 \leq \ell < d$, so the evasive set is non-empty and D^* is defined. If $r \geq D^*$ then a single move erases the watermark while keeping utility exactly* 1*; hence $r_c(q) \leq D^*$ unconditionally. If in addition the hypotheses of Corollary A.9 hold, then $(\gamma - \zeta)T < r_c(q) \leq D^*$.*

Proof. For the upper bound, move in one step from z^0 to a nearest evasive valid trajectory: the step has length $D^* \leq r$, and both endpoints have utility 1, so (a) and (b) hold and the attack succeeds. For the lower bound, Corollary A.9 says every attack with $r \leq (\gamma - \zeta)T$ fails, so $r_c(q) > (\gamma - \zeta)T$. □

B.16 Proof of Theorem A.19

Restated (Theorem A.19: Straight-line attack). *Assume exact c-regularity. Let z' be any evasive valid trajectory, $e = z^0 \oplus z'$, $D = |e|$. Flip the coordinates of* $\mathrm{supp}(e)$ *one at a time, in any order. Every intermediate point z, having flipped the subset $S \subseteq \mathrm{supp}(e)$, satisfies*

$$E_X(z) \;\leq\; c \cdot \min\big(|S|,\, D - |S|\big) \;\leq\; c\lfloor D/2 \rfloor.$$

Consequently, if $\eta \geq \Delta D/(2T)$, this $r = 1$ attack lies in $\mathcal{A}_{\mathrm{loc}}(1 - \eta, 1)$ and succeeds. Given z', it performs D single-bit flips in $O(Dc\Delta)$ time with incremental syndrome updates.

Proof. Both z^0 and z' are codewords, so $He = 0$ and $|C_j \cap \mathrm{supp}(e)|$ is even for every j. If j is violated at the intermediate point then $|C_j \cap S|$ is odd, hence $|C_j \cap (\mathrm{supp}(e) \setminus S)| = |C_j \cap \mathrm{supp}(e)| - |C_j \cap S|$ is odd and in particular nonzero. So every violated constraint meets S, giving $E_X(z) \leq \#\{j : C_j \cap S \neq \varnothing\} \leq c|S|$, and also meets $\mathrm{supp}(e) \setminus S$, giving $E_X(z) \leq c(D - |S|)$. Hence $E_X(z) \leq c \min(|S|, D - |S|) \leq c\lfloor D/2 \rfloor$ for every ordering, since $|S|$ ranges over $\{0, \ldots, D\}$. Divide by n and use $c/n = \Delta/T$. □

B.17 Proof of Lemma A.21

Restated (Lemma A.21: Escape bound). *Let P have stationary distribution π with $\pi(x_0) > 0$, let $A \subseteq \Omega$ with $x_0 \in A$, and let $T_{A^c} = \inf\{s \geq 0 : X_s \notin A\}$. Then for every $t \geq 0$,*

$$\Pr_{x_0}\big[T_{A^c} \leq t\big] \;\leq\; \frac{t\, Q(A, A^c)}{\pi(x_0)}, \qquad Q(A, A^c) = \sum_{x \in A,\, y \notin A} \pi(x) P(x, y).$$

Proof. Run the chain from stationarity. A union bound over steps gives $\Pr_\pi[\exists\, s < t : X_s \in A,\, X_{s+1} \notin A] \leq t\, Q(A, A^c)$. That event contains $\{X_0 = x_0\} \cap \{T_{A^c} \leq t\}$, of probability $\pi(x_0) \Pr_{x_0}[T_{A^c} \leq t]$. □

B.18 Proof of Theorem A.22

Restated (Theorem A.22: Exponential basin escape)**.** *Assume every R_j is locally sensitive, G_X is a $(\gamma T, (1-\epsilon)c)$-expander with $\epsilon < 1/2$, and $z^0 \in \mathrm{Sol}(X)$ with $D_\kappa(z^0) = \textsc{marked}$. Let* Dec *be certified at radius a, so D_κ is constant on the ball of radius a about z^0. Fix any integer w with $1 \le w \le \min(\gamma T,\, a+1)$, let $A = \{z : \mathrm{dist}(z, z^0) < w\}$, and let $B = \{z : D_\kappa(z) = \textsc{unmarked}\}$ — this is* every *trajectory the detector actively rejects, not merely every rejected codeword, since in this model validity is priced rather than required, so it includes the whole decoding ball around each evasive codeword. The identical proof bounds the hitting time of $B' = \{z : D_\kappa(z) \neq \textsc{marked}\}$ as well, since $A \cap B' = \varnothing$ by the same line, so the result covers denial of attribution too. Then $A \cap B = \varnothing$ and, with H the natural-logarithm binary entropy,*

$$\Pr_{z^0}\big[T_B \le t\big] \;\le\; t\exp\Big(T\,\mathsf{H}(w/T) - \beta(1-2\epsilon)c\,w\Big).$$

Hence for every fixed $\beta > \beta_c := \dfrac{T\,\mathsf{H}(w/T)}{(1-2\epsilon)c\,w}$, setting $\delta = \beta(1-2\epsilon)cw/T - \mathsf{H}(w/T) > 0$ and $t = e^{\delta T/2}$ gives $T_B > e^{\delta T/2}$ with probability at least $1 - e^{-\delta T/2}$. Taking $w = \lceil \theta T \rceil$ for a fixed $\theta < \min(\gamma,\, a/T)$ — available with θ up to $(1-2\epsilon)\gamma$ whenever $\epsilon < 1/4$, since then $a = \lceil a_{\mathrm{SS}} \rceil - 1$ is itself $\Theta(T)$ — makes $\beta_c \to \mathsf{H}(\theta)/((1-2\epsilon)c\theta)$, a constant independent of T. This is an asymptotic consequence under a linear certified radius; it does not certify the larger crossing radius used in our finite experiments.

Proof. $A \cap B = \varnothing$: if $\mathrm{dist}(z, z^0) < w \le \gamma T$ then in particular $\mathrm{dist}(z, z^0) \le a$ whenever $w \le a+1$, which we assume, so $\mathrm{Dec}(z) = z^0$ and $D_\kappa(z) = \textsc{marked}$, i.e. $z \notin B$. Hence $T_B \ge T_{A^c}$. Single-flip moves leaving A land on the sphere $\mathcal{S} = \{z : \mathrm{dist}(z, z^0) = w\}$, and $w \le \gamma T$ by hypothesis, so Lemma 4.3 gives $E_X(y) \ge (1-2\epsilon)cw$ for $y \in \mathcal{S}$. By stationarity $Q(A, A^c) = Q(A^c, A) = \sum_{y \notin A} \pi_\beta(y) P(y, A) \le \sum_{y \in \mathcal{S}} \pi_\beta(y)$, since only $\mathcal{S}$ has $P(y, A) > 0$. As $E_X(z^0) = 0$ we have $\pi_\beta(z^0) = 1/Z$, so

$$\frac{Q(A, A^c)}{\pi_\beta(z^0)} \;\le\; \sum_{y \in \mathcal{S}} e^{-\beta E_X(y)} \;\le\; |\mathcal{S}|\, e^{-\beta(1-2\epsilon)cw}, \qquad |\mathcal{S}| = \binom{T}{w} \le e^{T\mathsf{H}(w/T)}.$$

Apply Lemma A.21. □

C Exact Reachability in Parity Systems

The radial certificates exclude an entire annulus around the released trajectory. For parity tasks, the content of a basin (Section 4.3) is computed exactly by tracing the syndromes along a path. The connectivity argument is a binary linear specialization of the classical voltage-graph lifting framework (Gross & Tucker, 1987). Its application here separates the task's reachable subspace from the affine test that labels that subspace.

C.1 Lifting Syndrome Paths

Fix a consistent parity task $Hz = b$, a valid starting trajectory z^0, and an integer violation budget $B \ge 0$. Write $C = \ker H$ and $d = \dim C$. The error $e = z + z^0$ has syndrome He and energy $|He|$, where all additions are over $\mathbb{F}_2$ and $|\cdot|$ denotes Hamming weight. Let $\mathcal{R} \subseteq \mathbb{F}_2^T \setminus \{0\}$ be a fixed set of permitted moves. For radius-one editing, $\mathcal{R} = \{e_1, \ldots, e_T\}$; for radius r, take all nonzero vectors of weight at most r. The same moves are available at every state. A budget-B path is a sequence of such moves whose visited trajectories all have energy at most B.

Construct the undirected labeled multigraph $\mathcal{Q}_B$ with vertices $\{\sigma \in \mathrm{im}\, H : |\sigma| \le B\}$. For every $v \in \mathcal{R}$, join σ to $\tau = \sigma + Hv$ when both are vertices, and label the edge by v. Retain loops when $Hv = 0$ and parallel edges arising from distinct moves. Let $\mathcal{Q}_B^0$ denote its component containing 0. Choose a spanning tree of this component and let p_σ be the sum of the move labels on the tree path from 0 to σ. Then $Hp_\sigma = \sigma$ and $p_0 = 0$. For each edge $(\sigma, \tau; v)$ in the component, define

$$g_{\sigma,\tau;v} = p_\sigma + v + p_\tau \in C, \qquad K_B = \mathrm{span}\{g_{\sigma,\tau;v} : (\sigma, \tau; v) \in E(\mathcal{Q}_B^0)\}. \tag{2}$$

Tree edges contribute zero. Each remaining generator records the net change made by a closed syndrome walk: follow the tree to σ, traverse the edge, and return along the tree from τ.

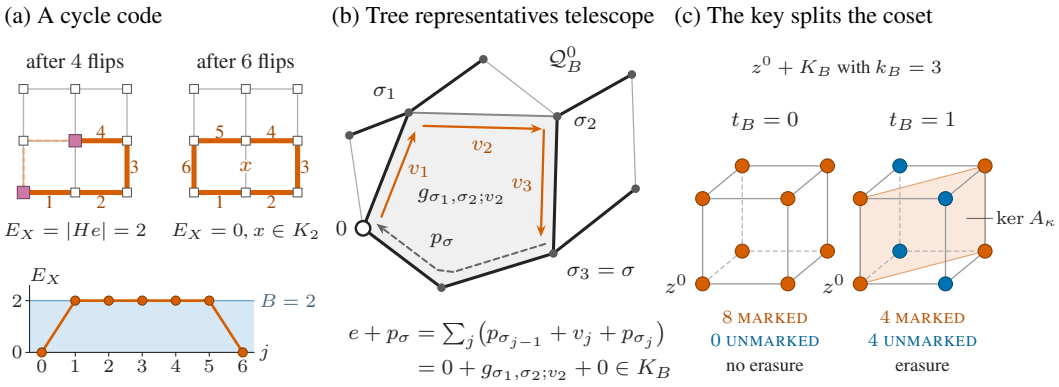

Figure 6: **Lifting syndrome walks.** (a) The cycle code of the 3×3 grid ($T = 12$ edges, $n = 9$ vertices, $d = 4$) at $r = 1$. Flipping the six-cycle x edge by edge keeps $E_X \le 2$, so $x \in K_2$. In any cycle code a codeword is an even-degree edge set, hence a sum of cycles, each walked this way, so $K_2 = C$; a first flip violates two constraints, so $K_1 = \{0\}$ (Theorem C.1). Thus $B_\kappa = 2$ in (3) for every nonzero M_κ, matching the $(2, 4)$ threshold $\eta = 0.01$ at $n = 200$ in Table 1. (b) In a schematic $\mathcal{Q}_B^0$ with a bold spanning tree, $e + p_\sigma$ telescopes to the non-tree generator (Theorem C.1). (c) For $k_B = 3$, $2^{k_B} - 2^{k_B - t_B}$ points of $z^0 + K_B$ are UNMARKED (Corollary C.3).

Theorem C.1 (Reachable subspace)**.** *The errors reachable from* 0 *at budget* B *are exactly*

$$\bigcup_{\sigma \in V(\mathcal{Q}_B^0)} (p_\sigma + K_B).$$

In particular, the reachable valid trajectories are $z^0 + K_B$*. If* $k_B = \dim K_B$*, there are* 2^{k_B} *such trajectories. Every connected component of the budget-*B *trajectory graph that meets* $\mathrm{Sol}(X)$ *contains* 2^{k_B} *valid trajectories and* $|V(\mathcal{Q}_B^0)|2^{k_B}$ *trajectories in total. There are* 2^{d-k_B} *components meeting* $\mathrm{Sol}(X)$*. The subspace* K_B *is independent of the spanning tree and is increasing with* B*.*

Proof. Consider a feasible path with syndrome sequence $\sigma_0 = 0, \ldots, \sigma_m = \sigma$ and move labels $v_1, \ldots, v_m$. Its final error $e = \sum_{j=1}^m v_j$ satisfies

$$e + p_\sigma = \sum_{j=1}^{m} (p_{\sigma_{j-1}} + v_j + p_{\sigma_j}) \in K_B,$$

because the intermediate tree representatives cancel. Thus every reachable error belongs to the displayed union.

Conversely, the closed tree–edge–tree walk associated with $g_{\sigma,\tau;v}$ is feasible and ends with precisely that error. Concatenating such walks realizes any sum of generators: starting a closed syndrome walk at a codeword translates its trajectory without changing its syndromes or energies. A final tree path realizes $p_\sigma + k$ for any $k \in K_B$. This proves equality.

The syndrome-zero slice is K_B, and translation by any codeword is an automorphism of the trajectory graph. Consequently its components meeting $\mathrm{Sol}(X)$ correspond to the cosets of K_B in C. The displayed union has disjoint syndrome slices, each of size 2^{k_B}, giving the claimed counts. Since K_B is the set of reachable valid differences, it is independent of the chosen tree. A budget-B path remains feasible at every larger budget, giving monotonicity. □

Corollary C.2 (Positive attribution from an isolated lift)**.** *Fix a marked valid* z^0*. Suppose* $K_B = \{0\}$ *and the decoder returns* z^0 *on every input within distance* a*. If* $\max_{s \in V(\mathcal{Q}_B^0)} |p_s| \le a$*, every trajectory reachable from* z^0 *under budget* B *is labeled* MARKED*. No energy-monotonicity condition on the decoder is required.*

Proof. Theorem C.1 identifies the reachable errors as exactly $\{p_s : s \in V(\mathcal{Q}_B^0)\}$. Each has weight at most a, so decoding returns the marked starting trajectory. □

In this isolated case, $\max_s |p_s|$ is the exact largest distortion of a reachable trajectory and is independent of the spanning tree. For a nonzero K_B, tree representatives alone do not determine that largest distortion.

C.2 From Reachability to Erasure

Let I be an information set, let z^0 be marked, and define the linear map $A_\kappa : C \to \mathbb{F}_2^\ell$ by $A_\kappa(k) = M_\kappa(k|_I)$. Write $t_B = \operatorname{rank}(A_\kappa|_{K_B})$. This rank measures how many independent mark equations can change along feasible paths.

Corollary C.3 (Exact erasure criterion)**.** *The marked valid trajectories reachable from z^0 are exactly $z^0 + (K_B \cap \ker A_\kappa)$. The number of reachable unmarked valid trajectories is*

$$2^{k_B} - 2^{k_B - t_B}.$$

Thus an exact-valid erasure exists if and only if $t_B > 0$.

Proof. By Theorem C.1, every reachable valid trajectory is $z^0 + k$ with $k \in K_B$. Since $M_\kappa(z^0|_I) = s_\kappa$, its mark equation holds precisely when $A_\kappa(k) = 0$. Rank–nullity gives $|K_B \cap \ker A_\kappa| = 2^{k_B - t_B}$. □

For the paper's decoder, this valid-endpoint criterion also settles erasure at arbitrary outputs. The reason is operational: an editor can append the detector's repair steps to its own path.

Proposition C.4 (Energy-monotone decoding)**.** *Suppose the detector returns each valid input unchanged before applying its affine mark test. Whenever it decodes an input to a valid trajectory, suppose that its repair uses moves in $\mathcal{R}$ and never increases energy. It may abstain on other inputs. Then a budget-B path from a marked z^0 reaches an output labeled* UNMARKED *if and only if $t_B > 0$.*

Proof. If an editor reaches an unmarked output, append the decoder's repair path. All appended states have energy at most the output's energy, hence at most B. The endpoint is valid and unmarked, so Corollary C.3 implies $t_B > 0$. Conversely, when $t_B > 0$ that corollary provides a feasible path to an unmarked valid trajectory, which the detector returns unchanged. □

Theorem 4.5 combines Theorem C.1, Corollary C.3, and Proposition C.4. Sequential strict energy-decreasing bit flipping satisfies the proposition for every $\mathcal{R}$ containing the unit vectors. Among decoders satisfying its assumptions, the existential erasure threshold is therefore

$$B_\kappa = \min\{B : t_B > 0\}, \tag{3}$$

with the minimum of the empty set taken as infinity. It depends on H, the permitted moves, and M_κ, and is unchanged by the offset or the choice of marked starting trajectory. Different decoders can still change the success and abstention rates of a particular editor. Repairs that require a higher-energy intermediate state fall outside the proposition.

The conclusion is resistance to UNMARKED outputs. Reachable states may produce ABSTAIN even when $K_B = \{0\}$, so this criterion does not give the positive attribution guarantee of Theorem D.1. It applies to fixed translation-invariant moves; committed-prefix restrictions depend on execution time and are not represented by $\mathcal{Q}_B$.

C.3 The Effect of a Random Affine Test

The next probability is over an ideal uniformly sampled full-row-rank matrix. The geometry, including K_B, is fixed. This explicitly specifies the random-key model used in the count.

Proposition C.5 (Probability that the reachable subspace remains marked)**.** *Let $M \in \mathbb{F}_2^{\ell \times d}$ be uniform among full-row-rank matrices, with $1 \le \ell < d$, and set $k = \dim K_B$. Then*

$$\Pr_M[A(K_B) = \{0\}] = \frac{\begin{bmatrix} d-k \\ \ell \end{bmatrix}_2}{\begin{bmatrix} d \\ \ell \end{bmatrix}_2} = \prod_{j=0}^{\ell-1} \frac{2^{d-k} - 2^j}{2^d - 2^j}, \tag{4}$$

where the probability is zero if $d - k < \ell$. For $k = 1$, it equals $(2^{d-\ell} - 1)/(2^d - 1)$; in general it is at most $2^{-\ell k}$.

Proof. Restriction to I is an isomorphism from C to $\mathbb{F}_2^d$, so $W = \{k|_I : k \in K_B\}$ has dimension k. The row space of a uniform full-rank matrix is a uniform ℓ-dimensional subspace. The event $MW = \{0\}$ says that this row space lies in the $(d-k)$-dimensional annihilator $W^\perp$. Counting subspaces gives the Gaussian-binomial ratio, and expanding it gives the product. For $k = 1$, equivalently count whether one fixed nonzero vector lies in the uniformly distributed $(d-\ell)$-dimensional kernel, giving the stated expression. When $d - k \geq \ell$, each factor in the product is at most 2^{-k}; the remaining case has probability zero. □

Thus as soon as the task admits one nonzero reachable valid difference, a random affine test labels it unmarked with probability $1 - (2^{d-\ell} - 1)/(2^d - 1)$. The subspace first changes because of the task's geometry; the key determines which of its directions change the mark. The formula also holds conditional on a released z^0 in the ideal generation model with an independent uniform offset: averaging over that offset makes z^0 uniform on $\mathrm{Sol}(X)$ and independent of M.

C.4 Computing the Certificate

Breadth-first search from syndrome zero computes $\mathcal{Q}_B^0$ without listing all 2^T trajectories. Store each discovered syndrome as a bit vector and its tree representative as an error vector. For each popped σ and each move v, form $\tau = \sigma + Hv$. Discard it if $|\tau| > B$. If it is new, set $p_\tau = p_\sigma + v$ and enqueue it. Otherwise, insert $p_\sigma + v + p_\tau$ into a binary linear basis for K_B. This preserves loops and parallel edges. The required dimension is the rank of their error labels, rather than the number of cycles in the syndrome graph.

After the queue is exhausted, the basis gives k_B, and applying A_κ to its vectors gives t_B. A generator with nonzero image immediately yields an erasure witness through its tree–edge–tree walk. A no-erasure certificate requires exhausting the search; terminating at a resource limit provides only the explored paths and witnessed subspace. For radius one, the search examines $T|V(\mathcal{Q}_B^0)|$ candidate edges, together with the cost of binary linear algebra. The syndrome component can itself be exponentially large. For small instances, repeating the search as B increases gives the complete sequence of reachable subspaces and the exact threshold (3).

C.5 A Decoding Certificate from the Component

Proposition C.6 (Minimum distance from low-energy connectivity)**.** *Let every column of H have weight five, and let K_B use single-coordinate moves. If $K_9 = \{0\}$, every nonzero vector in* $\ker H$ *has weight at least eight. Consequently, complete bounded-distance decoding through radius three is unambiguous.*

Proof. Odd column weight forces every codeword to have even weight. A weight-two codeword has a two-flip path with energies $0, 5, 0$. For a weight-four codeword, some pair of its columns overlaps; otherwise their sum could not vanish. Flip that pair first. The path energies are at most $0, 5, 8, 5, 0$, using that the last intermediate syndrome equals a single column.

For a weight-six codeword with support S, choose a uniform three-element subset A. A row meets S in $k \in \{0, 2, 4, 6\}$ columns, and its probability of contributing to $H1_A$ is respectively $0, 3/5, 2/5, 1$, each at most $3k/10$. The six columns have thirty incidences, so $\mathbb{E}_A|H1_A| \leq 9$. Choose a triple with energy at most nine. Its complement in S has the same syndrome. Each triple contains an overlapping pair, since three pairwise disjoint columns have syndrome weight fifteen. Order each triple with such a pair first; traverse the first triple in that order and the complement in reverse order. The resulting path energies are bounded by

$$0, \quad 5, \quad 8, \quad 9, \quad 8, \quad 5, \quad 0.$$

Every codeword of weight at most six therefore belongs to K_9 and is excluded by the hypothesis. Any two distinct codewords are then at distance at least eight, making radius-three decoding unique. □

The budget nine is sharp under the column-weight-five assumption. For the edge–vertex incidence matrix of the complete graph on six vertices, the only nonzero codeword is the all-ones vector, and the energy of a weight-w error is $w(6-w)$. Every unit-flip path between the codewords crosses weight three and energy nine, so $K_8 = \{0\}$ despite minimum distance six. This example has row weight two; it does not assert sharpness within the $(5, 10)$ family.

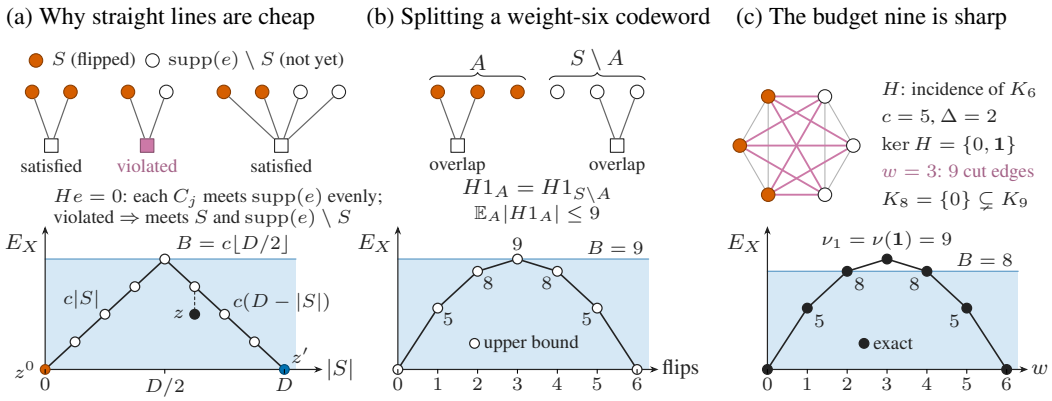

Figure 7: **Explicit flip paths.** (a) Theorem A.19 (schematic, $D = 8$): a violated constraint meets both S and $\mathrm{supp}(e) \setminus S$, so $E_X(z) \le c\min(|S|, D - |S|)$ and the path from z^0 to z' stays in the water $\{E_X \le B\}$ for $B = c\lfloor D/2 \rfloor$. (b) Proposition C.6: A and $S \setminus A$ have equal syndromes, and the split path has energies at most $0, 5, 8, 9, 8, 5, 0$, against 15 from Theorem A.19 at $c = 5$, $D = 6$. (c) For the incidence matrix of K_6 a weight-w error has energy $w(6 - w)$, so $K_8 = \{0\}$ and $\nu(\mathbf{1}) = 9$ although $d_{\min} = 6$. This example has row weight two, so it lies outside the $(5, 10)$ family.

C.6 Independent Private Offsets Across Releases

The fixed-offset construction has an affine-closure forgery. A release registry gives a precise way to change that multi-sample behavior while retaining affine generation. The following is a separate variant: sample one uniform full-rank M; for each release j, sample an independent uniform secret offset $s_j \in \mathbb{F}_2^\ell$ and publish exactly one strategy word u_j, sampled uniformly from $\{u : Mu = s_j\}$. Store the offsets privately. They are sampled independently with replacement; the construction does not condition on their being distinct. An authenticated release identifier selects one offset for detection, or the detector searches all N registered offsets. Decoding remains key-independent.

Proposition C.7 (Release-specific and registry-wide error probabilities)**.** *In this ideal model, the public transcript $(u_1, \ldots, u_N)$ is independent of M and consists of independent uniform strategy words. An adversary observes this transcript, has no access to the offsets or a detection oracle, and outputs a valid strategy u^*. If it chooses a target release j and $u^* \neq u_j$, then*

$$\Pr[Mu^* = s_j] = p_{d,\ell} := \frac{2^{d-\ell} - 1}{2^d - 1}.$$

The target and output may depend on the transcript. If instead the registry accepts a match to any offset and $u^* \notin \{u_1, \ldots, u_N\}$, its acceptance probability is at most $Np_{d,\ell}$. For an artifact produced independently of $(M, s_1, \ldots, s_N)$, the registry false-positive probability is at most

$$1 - (1 - 2^{-\ell})^N,$$

with equality whenever the key-independent decoder always converges on that artifact. Testing one independently selected release has false-positive probability at most $2^{-\ell}$.

Proof. For fixed M and u, the probability of generating u in release j is $2^{-\ell}2^{-(d-\ell)} = 2^{-d}$. Offsets and sampling randomness are independent across releases, proving that the transcript is uniform and independent of M. Conditional on that transcript and the adversary's independent randomness, $s_j = Mu_j$ and $v = u^* + u_j$ is a fixed nonzero vector. The kernel of M is a uniform $(d - \ell)$-dimensional subspace, so $\Pr[Mv = 0] = p_{d,\ell}$. Summing these probabilities over the registered releases gives the fresh-output bound.

For an independent artifact on which decoding converges, condition on M and its decoded strategy word. Each independent uniform offset matches its image with probability $2^{-\ell}$, independently across releases. The probability of any match is therefore $1 - (1 - 2^{-\ell})^N$. A decoder abstention contributes no acceptance. The single-release bound follows from the same argument with one offset. □

For registry-wide false-positive target α, it suffices to choose $\ell \geq \lceil \log_2(N/\alpha) \rceil$, subject to $\ell < d$. With one authenticated release identifier there is no multiplicity factor. These bounds exclude replay, which reproduces an accepted release. They also rely on private offsets and one generated sample per release: public pairs (u_j, s_j) expose linear equations about M, while repeated sampling under a fixed offset restores affine closure within that release. The proposition quantifies this variant's transcript-only forgery behavior; it does not supply security against a detection oracle or establish a new empirical result.

D Proofs for the Basin Framework

Throughout, G_r, X_B, $\mathcal{B}_B(z^0)$ and u_r are as in Section 4.1. An attack in $\mathcal{A}_{\mathrm{loc}}(q, r)$ has consecutive distances at most r and $U_X \geq q$ at every point. Since E_X is integer valued, $U_X(z) \geq 1 - \eta$ is equivalent to $E_X(z) \leq \lfloor \eta n \rfloor$, and steps of length zero repeat a point, so the reachable set is exactly $\mathcal{B}_{\lfloor \eta n \rfloor}(z^0)$. The barrier distance is finite because G_r connects the whole hypercube, whose energies are at most n. It is symmetric because walks reverse, and concatenating a walk from x to y with one from y to z gives $u_r(x, z) \leq \max\{u_r(x, y), u_r(y, z)\}$. It is a pseudo-ultrametric: $u_r(z, z') = 0$ for distinct valid z, z' exactly when single moves between valid solutions connect them.

D.1 Output fidelity and the finite-instance wall

Section 4.2 uses the following two statements; their direct proofs are in Appendix B.

Theorem D.1 (Output-fidelity bound; no per-step condition)**.** *Assume the hypotheses of Lemma 4.3 with $z^* = z^0 \in \mathrm{Sol}(X)$, let $D_\kappa(z^0) = \textsc{marked}$, and let* Dec *be certified at radius a. Let z be* any *trajectory with*

$$\mathrm{dist}(z, z^0) \;\leq\; \gamma T \qquad \textit{and} \qquad U_X(z) \;\geq\; 1 - \eta, \qquad \eta < (1 - 2\epsilon)\,\Delta\,\frac{a}{T}.$$

Then $D_\kappa(z) = \textsc{marked}$. In particular no attacker whatsoever — offline, randomised, unbounded in computation, and possibly holding the key, detector, and list of evasive trajectories — can produce an unmarked output satisfying both the stated final-distortion and utility conditions.

Proposition D.2 (Direct finite-instance certificate)**.** *Let D_κ return* MARKED *throughout the Hamming ball of integer radius a around z^0, and for $a < b \leq T$ define*

$$L_{a,b}(z^0) := \min_{a < \mathrm{dist}(z, z^0) \leq b} E_X(z).$$

If $\eta n < L_{a,b}(z^0)$, every output with $\mathrm{dist}(z, z^0) \leq b$ and $U_X(z) \geq 1 - \eta$ is marked. Moreover, any path started at z^0 that maintains this utility floor and has step radius $r \leq b - a$ remains in the marked radius-a ball.

Proof in Appendix B.10.

D.2 Proof of Proposition 4.1

If $\mathcal{B}_B(z^0)$ contains an unmarked valid z', an editor walks to it; the detector returns z' unchanged and labels it UNMARKED. Conversely, suppose a walk in the basin ends at an output z labelled UNMARKED. The detector decoded z to a valid unmarked $\hat{c}$ through moves of G_r along which the energy never exceeds $E_X(z) \leq B$. Appending these moves extends the walk inside X_B, so $\hat{c} \in \mathcal{B}_B(z^0)$. Since $\mathcal{B}_B(z^0)$ grows with B, the least budget at which it contains an unmarked solution is $\min\{u_r(z^0, z') : z' \in \mathrm{Sol}(X) \text{ unmarked}\}$. Sequential strict energy-decreasing bit flipping fixes valid inputs, since no flip decreases zero energy, and uses unit moves, which belong to G_r for every $r \geq 1$.

D.3 Proof of Proposition 4.2

Suppose a walk in $\mathcal{B}_B(z^0)$ leaves the ball of radius a, and let z^t be its first point with $\mathrm{dist}(z^t, z^0) > a$. Then $\mathrm{dist}(z^{t-1}, z^0) \leq a$ and the step has length at most r, so $a < \mathrm{dist}(z^t, z^0) \leq a + r$ and $E_X(z^t) > B$, contradicting $z^t \in X_B$. Hence the basin lies in the ball, and a detector that is MARKED on the ball is MARKED along every walk.

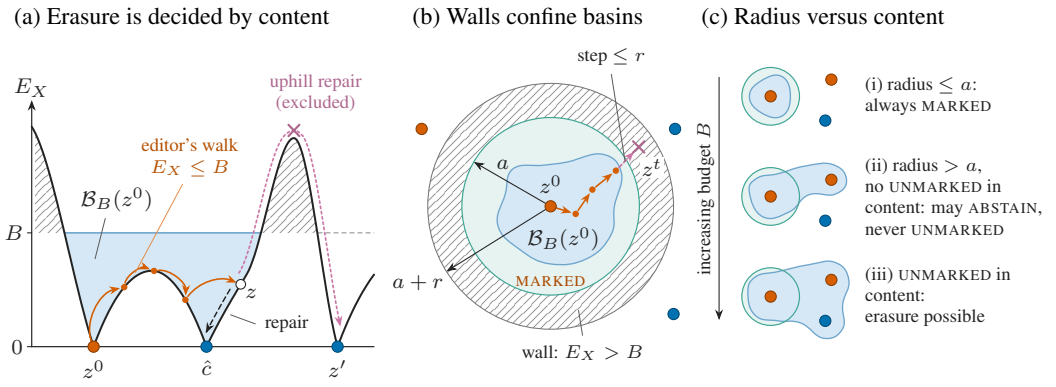

Figure 8: **The two invariants of a basin.** Schematic; hatching marks $E_X > B$. (a) The basin $\mathcal{B}_B(z^0)$ is the component of $\{E_X \le B\}$ containing z^0; the valley of z' is not in it. The energy-nonincreasing repair takes the editor's output z to $\hat{c}$, so the decoded target of any UNMARKED output lies in the basin (the 'only if' part of Proposition 4.1). An excluded uphill repair could decode to z' outside it. (b) With steps of length $\le r$, a walk's first point z^t beyond the certified ball of radius a would lie in the wall (Proposition 4.2). (c) Certified attribution depends on the radius, erasure on the content.

The guarantees of Section 4.2 are instances. For Theorem 4.4, take $a = \lfloor \zeta T \rfloor$. Every z with $\zeta T < \mathrm{dist}(z, z^0) \le \zeta T + r \le \gamma T$ has $E_X(z) \ge (1-2\epsilon)c\,\mathrm{dist}(z, z^0) > (1-2\epsilon)c\,\zeta T = \eta n \ge \lfloor \eta n \rfloor$ by Lemma 4.3 and $cT = \Delta n$, so these points form a wall of width r. For Proposition D.2, the annulus $a < \mathrm{dist}(z, z^0) \le b$ has energies at least $L_{a,b}(z^0) > \eta n$ and contains a wall of width $r \le b - a$; its output clause is the same annulus read at a single point. Theorem A.12 takes its gap condition as the wall. Theorem D.1 is the pointwise version: an output with $\mathrm{dist}(z, z^0) \le \gamma T$ and $E_X(z) \le \eta n$ cannot satisfy $a < \mathrm{dist}(z, z^0)$, because there $E_X(z) \ge (1 - 2\epsilon)c(a + 1) > (1 - 2\epsilon)\Delta a n/T > \eta n$, using $\Delta n \le cT$. Corollaries A.9 and C.2 are the attribution clause, the latter with Theorem C.1 supplying the basin radius $\max_s |p_s|$. The direct proofs in Appendices B and C remain valid.

D.4 Barrier distance under expansion

Proposition D.3 (Crossing bound)**.** *Under the hypotheses of Lemma 4.3 with $1 \le r \le \lfloor \gamma T \rfloor$, any two distinct valid z, z' satisfy $u_r(z, z') \ge (1 - 2\epsilon)c(\lfloor \gamma T \rfloor - r + 1)$.*

Proof. By Corollary A.6, $\mathrm{dist}(z, z') \ge \lfloor \gamma T \rfloor + 1$. Along any walk from z to z', the distance to z starts at zero, ends above $\lfloor \gamma T \rfloor$, and changes by at most r per step. Its first point at distance at least $\lfloor \gamma T \rfloor - r + 1$ therefore lies at a distance w with $\lfloor \gamma T \rfloor - r + 1 \le w \le \lfloor \gamma T \rfloor \le \gamma T$, where Lemma 4.3 gives energy at least $(1 - 2\epsilon)cw$. □

D.5 The barrier norm

Proposition D.4 (Barrier norm)**.** *Let X be a consistent parity task, $C = \ker H$, and let the permitted moves be those of G_r. For $x \in C$, the barrier distance $u_r(z, z \oplus x)$ is the same for every valid z; call it $\nu(x)$. Then (i) $\nu(x \oplus x') \le \max\{\nu(x), \nu(x')\}$; (ii) $\nu(x) = 0$ exactly when x lies in the span of the codewords of weight at most r, so ν is positive on $C \setminus \{0\}$ when $r < d_{\min}$; (iii) $\{x : \nu(x) \le B\} = K_B$ for every B; (iv) for an affine mark with linear part A_κ and a detector as in Proposition 4.1, the erasure threshold from every marked valid z^0 is $B_\kappa = \min\{\nu(x) : A_\kappa(x) \ne 0\}$.*

Proof. Translation by $x' \in C$ preserves E_X and the edges of G_r, so it maps $\mathcal{B}_B(z)$ onto $\mathcal{B}_B(z \oplus x')$. Hence $u_r(z, z \oplus x) = u_r(z \oplus x', z \oplus x' \oplus x)$, and ν is well defined because $\mathrm{Sol}(X) = z^0 \oplus C$. (i) The ultrametric inequality gives $u_r(z, z \oplus x \oplus x') \le \max\{u_r(z, z \oplus x), u_r(z \oplus x, z \oplus x \oplus x')\} = \max\{\nu(x), \nu(x')\}$. (ii) $\nu(x) = 0$ means that $z \oplus x$ is reached by a walk inside $X_0 = \mathrm{Sol}(X)$, each of whose steps is a codeword of weight between 1 and r; conversely, any sum of such codewords is such a walk. (iii) By Theorem C.1, the valid trajectories in $\mathcal{B}_B(z^0)$ are $z^0 \oplus K_B$. (iv) By Proposition 4.1,

the threshold is the least $\nu(x)$ over x with $z^0 \oplus x$ unmarked. Since z^0 is marked, $z^0 \oplus x$ is unmarked exactly when $A_\kappa(x) \neq 0$. □

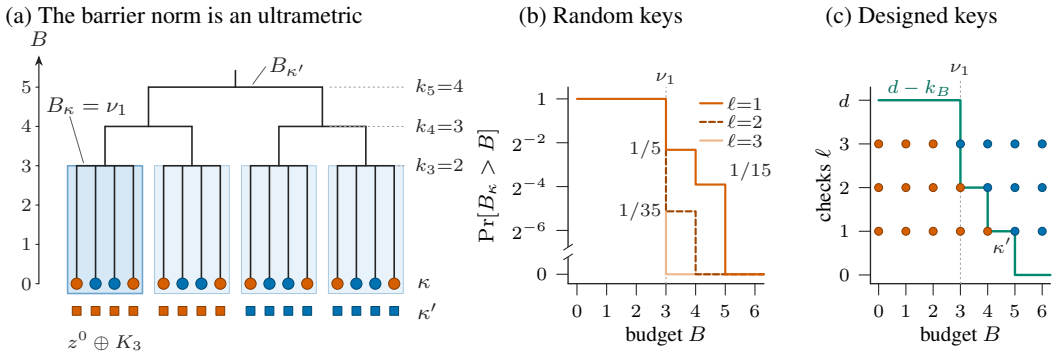

Figure 9: **The barrier norm of an exactly computed task** ($T = 14$, $n = 15$, column weight 3, $d = d_{\min} = 4$, $r = 1$). (a) Leaves $z^0 \oplus x$ and $z^0 \oplus x'$ merge at $\nu(x \oplus x') \in \{3, 4, 5\}$, so the water at $B = 3$ cuts the tree into the cosets of K_3 (Proposition D.4). Orange is marked and blue unmarked under a key κ nonzero on K_3 (circles) and a designed κ' vanishing on K_4 (squares); each is erased at the lowest merge mixing its colors. (b) For random keys, $\Pr[B_\kappa > B]$ falls at ν_1 to $1/5$, $1/35$, 0, below $2^{-\ell}$. (c) Some key with ℓ checks survives B (orange) exactly when $\ell \leq d - k_B$ (Corollary 4.6).

D.6 Proof of Corollary 4.6

By Proposition D.4(iii,iv), $B_\kappa > B$ exactly when A_κ vanishes on K_B, and Proposition C.5 gives the probability of this event for a uniformly random full-row-rank key. Always $B_\kappa \geq \nu_1$, and K_{ν_1} contains a nonzero codeword, so $\Pr[B_\kappa > \nu_1] \leq 2^{-\ell k_{\nu_1}} \leq 2^{-\ell}$. For the last claim, restriction to the information set is an isomorphism from C onto $\mathbb{F}_2^d$, so $W_B = \{x|_I : x \in K_B\}$ has dimension k_B. The map A_κ vanishes on K_B exactly when the row space of M_κ lies in the annihilator of W_B, which has dimension $d - k_B$. A full-row-rank M_κ with ℓ rows and this property exists if and only if $\ell \leq d - k_B$. Such a key keeps the false-positive bound $2^{-\ell}$ of Proposition A.15 whenever its offset is uniform and independent.

E Experimental Protocol

E.1 Constraint instances

Each instance is a $(c, \Delta) = (5, 10)$ biregular factor graph on T decisions and $n = cT/\Delta$ parity constraints. The main arm uses progressive edge growth (PEG); a Gallager ensemble supplies the low-girth control. We evaluate four independently generated instances at each size. Because neither construction is uniformly free of four-cycles at the smaller sizes, we report the four-cycle count for every instance. The certified detector radius is 2 for both construction arms at $T \geq 400$, while the Gallager control reaches zero four-cycles by $T = 800$ (§5.4).

E.2 Decoder and detector

The detector runs the sequential flipping decoder of Proposition A.14 and returns MARKED, UNMARKED, or ABSTAIN when decoding does not terminate at a codeword. This third outcome is essential: all 48 outputs that a fail-open rule would have counted as erasures were decoder stalls, not valid unmarked trajectories. Every experiment below uses the three-valued rule.

E.3 Detector-radius certification

We enumerate every error pattern through weight w_{exh} and return the largest weight at which all patterns decode correctly. A value equal to w_{exh} is right-censored. Weight three requires $\binom{T}{3}$

decodings—1.3×10^6 at $T = 200$ and 6.8×10^8 at $T = 1600$—so we complete it where feasible and mark all partial results as lower bounds.

E.4 ALFWORLD TRAJECTORIES

The policy ranks admissible commands by length-normalised log-likelihood and executes the highest ranked command. To avoid a fixed point after a no-op, commands already attempted in the same state are excluded. Counterfactual edits are replayed from the initial state, so success and failure are always judged by the environment rather than by a learned surrogate.

F SUPPLEMENTAL EXPERIMENTAL RESULTS

F.1 EXACT THRESHOLDS AND DECODER COMPARISON

Instances. The scale study reuses the twenty archived $(5, 10)$ instances: four at each of $T \in \{200, 400, 800, 1600\}$ and the two phase-sweep instances at each of $T \in \{400, 800\}$. It adds four instances at $T = 100$. The geometry study uses four instances of each of the $(2, 4)$, $(3, 6)$, $(4, 8)$ and $(6, 12)$ PEG families and of a $(5, 10)$ Gallager family at $T = 400$. All have rate $1/2$ and $n = T/2$. Each instance has a BasinMark key with $\ell = 16$ and a marked starting trajectory.

Certificates and erasure walks. For single-coordinate moves, we search the syndrome component at integer budgets starting at $B = c - 1$, below the energy of a single flip. A completed component with no nonzero label certifies $K_B = \{0\}$. When the search finds a nonzero label, we replay its tree–edge–tree walk from the marked trajectory. Every visited energy is at most B and the endpoint is valid; the walk counts as an erasure only if the three-valued detector returns UNMARKED. Each search stops at a size limit of 1.5×10^8 discovered syndromes at $T = 400$ and 800 and 8×10^7 at $T = 1600$; the limits used at $T \leq 200$ were never reached. An instance's budget sequence ends at its first erasure walk or after three consecutive incomplete budgets. Every incomplete search reached its limit without finding a nonzero label.

Optimized attack. Randomized information-set search collects the lightest codewords it finds, and beam search orders the flips of each of the lightest supports to minimize the peak energy. We keep the best order whose endpoint changes the mark and replay it with the detector. The search effort per instance decreases with T, so this slack is an upper bound whose tightness varies with T.

Results. Table 2 lists all 44 instances. Except on the cycle codes, every erasure walk is longer than its final distance, because it flips and later restores coordinates outside the codeword it reaches. This explains the gap between the two upper bounds. At $T = 200$ the codewords used by the optimized attack have weight 20–26, but the best orders found for flipping only their supports peak at energy 12–14. Walks through the component instead reach codewords of weight 56–70 at peak energy eight. An earlier search stopped at 5×10^4 syndromes and closed $B = 7$ at $T = 200$ and $B = 9$ above; the present certificates supersede it.

Complete radius-three decoding. We compare sequential bit flipping with a separate bounded-distance decoder. The latter stores all zero-, single-, and double-error syndromes in a table. A query first checks this table, then adds each single-column syndrome and checks again. It therefore finds a correction whenever one exists at distance at most three and abstains otherwise.

Uniqueness is certified through minimum distance. The pair tables contain no collisions between distinct errors of weight at most two, excluding nonzero codewords through weight four; odd column weight excludes odd-weight codewords. We then test weight-six feasibility with a 180-second solver limit. Ten calls prove infeasibility and ten reach the time limit. Independently, the complete $K_9 = \{0\}$ components certify minimum distance at least eight on all 16 archived instances with $T \geq 400$, by Proposition C.6. Six instances have both proofs, so their union certifies radius three on all 20. The ten solver timeouts remain unresolved solver decisions.

The bounded decoder need not repair through an energy-monotone path. Its pathwise attribution guarantee instead uses the completed components directly: when $K_B = \{0\}$ and $\max_s |p_s| \leq 3$,

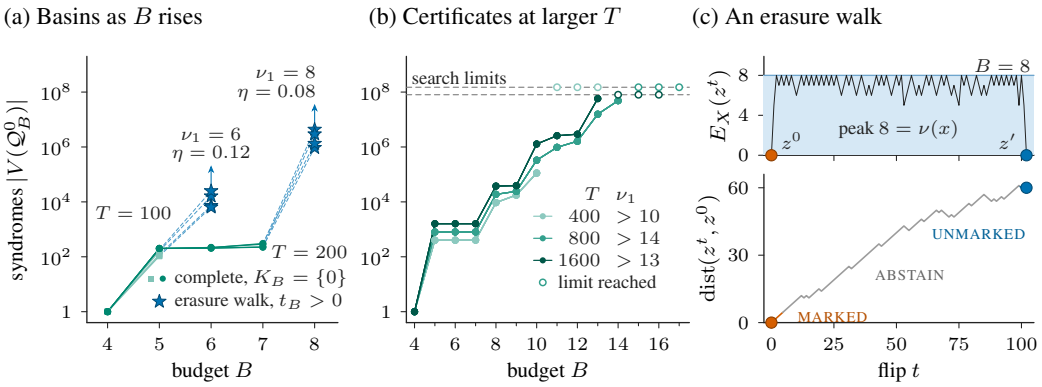

Figure 10: **Syndrome components of the** $(5, 10)$ **PEG instances.** (a) $|V(\mathcal{Q}_B^0)|$ at $T = 100$ and 200, four instances each. At the stars the search stopped after four nonzero labels, so these sizes are lower bounds; certificate and walk meet, so $\nu_1 = 6$ and 8 exactly (Table 1). (b) At $T \geq 400$ complete components certify $\nu_1 > B$ until the search limit (open circles). (c) One such walk ($T = 200$, $B = 8$): 102 flips with $E_X \leq 8$ reach the valid $z' = z^0 \oplus x$, $|x| = 60$; the output is MARKED for $t \leq 4$ and UNMARKED from $t = 100$. Since $K_7 = \{0\}$, the peak $\nu(x) = 8$ is exact.

every reachable trajectory is in its certified decoding ball. Table 2 gives the corresponding budgets B_M.

Matched random-error comparison. For each archived instance and each weight $w \in \{1, 2, 3, 4, 6, 10\}$, both decoders receive the same 100 uniformly sampled error supports, giving 12,000 matched trials. Correct recovery means returning the original valid trajectory; abstention means returning no valid trajectory; wrong recovery means returning a different valid trajectory. Both decoders recover every sampled weight-one, -two, and -three error. At the larger tested weights, bounded decoding abstains on every sample, while bit flipping often still recovers the original trajectory (Table 3). Neither decoder returns a wrong trajectory in these trials. The comparison exposes the coverage cost of the exact radius-three rule; the worst-case radius comes from the distance certificates above.

F.2 Connectivity of Fully Enumerated Plan Spaces

A corridor maze has two gates per wall separated by s; blocksworld with interleavable subgoals provides a loose-coupling control. We enumerate every valid plan and every integer edit radius under *state-sequence* distance. For $s \in \{2, \ldots, 8, 10\}$, the critical radius is $r_c = s + 1$ exactly. Below it, the edit graph has three components, each aligned perfectly with an environment-defined homotopy class (purity 1.00). Blocksworld remains connected at radius one, including at 2520 plans; the largest maze contains 3968 plans.

This enumeration establishes edit-space separation directly. Connecting it to Lemma 4.3 would additionally require an encoding-specific proof of local sensitivity and expansion for the planning constraints.

F.3 Barrier Profiles

Table 7 reports means over four instances per size. Figure 12(a) plots the smallest violation count found at each weight by adversarial search with restarts. Search can only upper-bound the true minimum, so these curves can falsify a proposed lower barrier but cannot certify one.

F.4 A Certified Finite Profile

On one $T = 400$ phase-sweep instance ($n = 200$), exhaustive enumeration gives $E_1 = 5$, $E_2 = 6$, and $E_3 = 7$. For each later weight we posed a binary linear feasibility problem that asks whether an

Geometry	T	Seed	4-cyc.	B_{cert}	Syndromes at B_{cert}	B_{wit}	Walk steps/dist.	Attack weight/peak	R3	B_M
(5,10) PEG	100	1000	88	5	159	6	32/20	12/8	–	–
(5,10) PEG	100	1001	87	5	177	6	30/18	12/9	–	–
(5,10) PEG	100	1002	92	5	120	6	26/18	12/9	–	–
(5,10) PEG	100	1003	89	5	106	6	26/24	12/9	–	–
(5,10) PEG	200	1000	8	7	287	8	102/60	24/12	S	5
(5,10) PEG	200	1001	8	7	228	8	104/64	26/13	S	7
(5,10) PEG	200	1002	8	7	305	8	92/56	24/13	S	5
(5,10) PEG	200	1003	4	7	224	8	114/70	20/14	S	6
(5,10) PEG	400	500	2	10	113,129	–	–	62/23	S+C	7
(5,10) PEG	400	501	1	10	110,496	–	–	62/23	S+C	7
(5,10) PEG	400	1000	3	10	114,206	–	–	64/22	S+C	7
(5,10) PEG	400	1001	1	10	112,601	–	–	62/21	S+C	7
(5,10) PEG	400	1002	1	10	110,351	–	–	54/23	S+C	7
(5,10) PEG	400	1003	1	10	112,389	–	–	64/23	S+C	7
(5,10) PEG	800	500	1	14	48,662,188	–	–	146/49	C	7
(5,10) PEG	800	501	0	14	48,509,828	–	–	146/47	C	8
(5,10) PEG	800	1000	0	14	48,566,755	–	–	144/49	C	8
(5,10) PEG	800	1001	3	14	48,860,030	–	–	146/48	C	8
(5,10) PEG	800	1002	0	14	48,438,330	–	–	144/47	C	8
(5,10) PEG	800	1003	3	14	48,711,345	–	–	142/48	C	7
(5,10) PEG	1600	1000	1	13	57,839,151	–	–	318/103	C	7
(5,10) PEG	1600	1001	0	13	57,808,062	–	–	322/100	C	9
(5,10) PEG	1600	1002	0	13	57,823,072	–	–	314/108	C	9
(5,10) PEG	1600	1003	1	13	57,826,810	–	–	314/103	C	9
(2,4) PEG	400	1000	0	1	1	2	6/6	6/2	–	–
(2,4) PEG	400	1001	0	1	1	2	6/6	6/2	–	–
(2,4) PEG	400	1002	0	1	1	2	6/6	5/2	–	–
(2,4) PEG	400	1003	0	1	1	2	6/6	6/2	–	–
(3,6) PEG	400	1000	0	3	401	4	60/36	26/6	–	–
(3,6) PEG	400	1001	0	3	401	4	58/38	16/6	–	–
(3,6) PEG	400	1002	0	3	401	4	82/68	16/6	–	–
(3,6) PEG	400	1003	1	3	412	4	90/70	26/5	–	–
(4,8) PEG	400	1000	1	7	9,203	8	48/46	51/16	–	–
(4,8) PEG	400	1001	1	7	11,360	8	45/41	53/16	–	–
(4,8) PEG	400	1002	1	7	9,393	8	48/42	51/16	–	–
(4,8) PEG	400	1003	1	7	10,406	8	44/36	49/16	–	–
(6,12) PEG	400	1000	3	13	113,986	–	–	65/30	–	–
(6,12) PEG	400	1001	9	13	114,295	–	–	65/30	–	–
(6,12) PEG	400	1002	3	13	115,215	–	–	63/30	–	–
(6,12) PEG	400	1003	5	13	113,492	–	–	66/26	–	–
(5,10) Gal.	400	1000	59	10	137,779	–	–	60/23	–	–
(5,10) Gal.	400	1001	68	10	146,991	–	–	62/22	–	–
(5,10) Gal.	400	1002	78	10	149,994	–	–	58/23	–	–
(5,10) Gal.	400	1003	66	10	139,836	–	–	50/23	–	–

Table 2: All 44 instances. B_{cert} is the largest budget whose complete component has $K_B = \{0\}$, so no radius-one path erases for $\eta < (B_{\mathrm{cert}} + 1)/n$. B_{wit} is the budget of the replayed erasure walk, given with its length and final Hamming distance; a dash means that the next three budgets were incomplete. *Attack* gives the codeword weight and peak energy of the optimized flip order. *R3* records the minimum-distance proof for radius-three decoding: S is a closed solver decision and C is $K_9 = \{0\}$ with Proposition C.6. B_M is the largest completed budget whose component stays within radius three, so radius-three decoding returns MARKED along every path for $\eta < (B_M + 1)/n$.

error of that weight can have energy below seven. Solver-closed decisions exclude all such errors at weights four and five (parity of the energy sharpens the tested caps), so

$$L_{2,5}(z^0) = \min_{3 \le w \le 5} E_w = 7.$$

Together with the independently verified radius $a = 2$, Proposition D.2 therefore guarantees detection through distance five whenever $\eta < 7/200 = 0.035$. The profile also identifies the boundary of this certificate: a weight-six perturbation with energy six gives $L_{2,b} \le 6$ for every $b \ge 6$ (without determining the exact value of E_6); the below-seven problem is infeasible at weight seven; and weights eight through ten remain unresolved. Thus the lower certificate is exact through weight five on this instance, while the weight-six witness is an upper bound beyond it. The closed feasibility checks took approximately 7.9, 12.0, and 31.6 minutes at weights four, five, and seven. The weight-six witness was found after 34.1 minutes; weights eight through ten each reached a one-hour limit without closing the question. With the original $a = 2$, increasing the outer radius alone cannot raise the utility threshold above 0.035: the weight-three minimum already gives $L_{2,b} \le 7$ for every $b \ge 3$.

Study	T	w	Trials	Bit flipping	Bounded radius three
Barrier	200	3	400	400/0/0	400/0/0
Barrier	200	4	400	396/4/0	0/400/0
Barrier	200	6	400	385/15/0	0/400/0
Barrier	200	10	400	138/262/0	0/400/0
Barrier	400	3	400	400/0/0	400/0/0
Barrier	400	4	400	400/0/0	0/400/0
Barrier	400	6	400	400/0/0	0/400/0
Barrier	400	10	400	397/3/0	0/400/0
Barrier	800	3	400	400/0/0	400/0/0
Barrier	800	4	400	400/0/0	0/400/0
Barrier	800	6	400	400/0/0	0/400/0
Barrier	800	10	400	400/0/0	0/400/0
Barrier	1600	3	400	400/0/0	400/0/0
Barrier	1600	4	400	400/0/0	0/400/0
Barrier	1600	6	400	400/0/0	0/400/0
Barrier	1600	10	400	400/0/0	0/400/0
Phase	400	3	200	200/0/0	200/0/0
Phase	400	4	200	200/0/0	0/200/0
Phase	400	6	200	200/0/0	0/200/0
Phase	400	10	200	198/2/0	0/200/0
Phase	800	3	200	200/0/0	200/0/0
Phase	800	4	200	200/0/0	0/200/0
Phase	800	6	200	200/0/0	0/200/0
Phase	800	10	200	200/0/0	0/200/0

Table 3: Matched decoder outcomes by study, size, and error weight. Entries are correct/abstain/wrong counts. Weights one and two, omitted for space, have perfect recovery for both decoders in every group. Each barrier row aggregates four instances and each phase row two, with 100 errors per instance and weight.

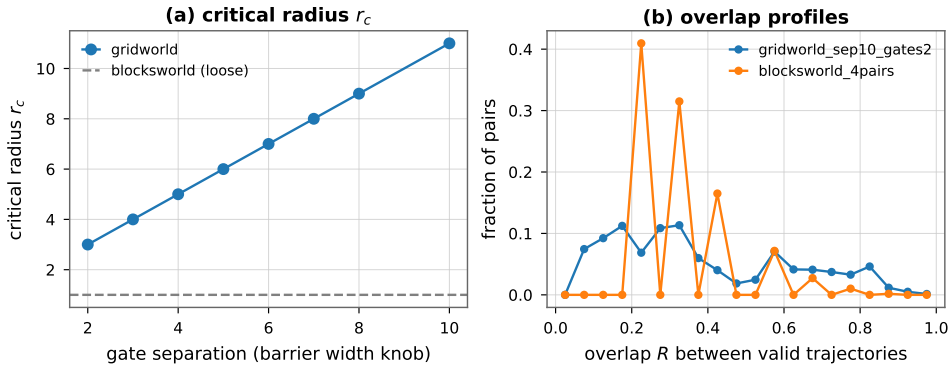

Figure 11: Connectivity of fully enumerated planning spaces. Corridor mazes satisfy $r_c = s + 1$, and their three subcritical components coincide with homotopy classes; the blocksworld control is connected at radius one. Distances use the state-sequence representation and the stated horizons.

F.5 Girth, sampling, and local expansion

The Sipser–Spielman radius these instances certify is $a_{\mathrm{SS}} = 1.2$: at $\epsilon = 0.2$ and $w \le 2$, $(1 - 2\epsilon)\lfloor \gamma T \rfloor = 1.2$, hence a certified radius $a = \lceil a_{\mathrm{SS}} \rceil - 1 = 1$.

With 30 samples per weight, the largest sampled decodable weight is $4.5, 11.5, 21.8, 55.0$ as T grows; at $T = 1600$ it falls from 55 to 41 when the sample count rises from 30 to 3000. The sampled quantity is therefore not a correction-radius certificate. The girth control points to a different bottleneck. PEG

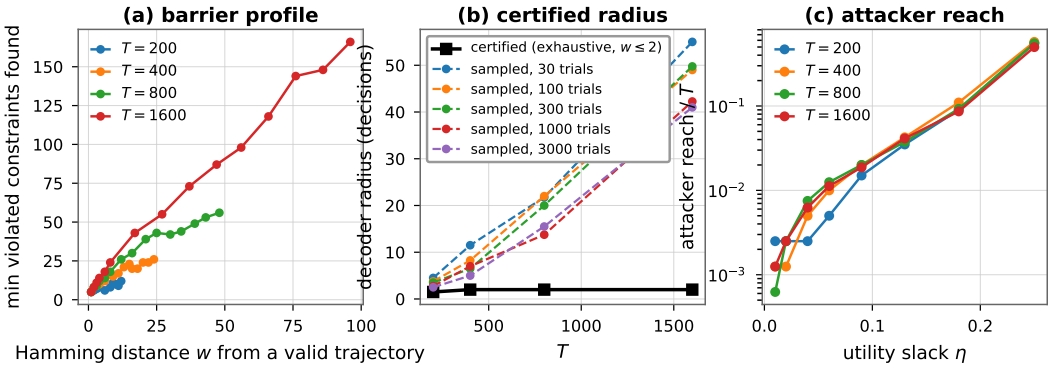

Figure 12: Original barrier, detector, and attack measurements. Left: smallest energy found at distance w, an upper bound on the true minimum. Middle: exhaustive bit-flip correction radii and sampled no-failure thresholds; arrows denote right-censoring. Right: distance reached by the utility-preserving greedy attack.

reduces the four-cycle count from 252 to 7 at $T = 200$, and the Gallager control reaches zero by $T = 800$, yet both arms have radius 2 at $T \geq 400$ (Table 7).

Exhaustive enumeration over $|S| \leq 3$ on all twenty instances gives $\min |N(S)| \in \{8, 9\}$ at $|S| = 2$ and $\{10, 11, 12\}$ at $|S| = 3$. Hence every instance certifies $\epsilon = 0.2$ through weight two. At weight three, six instances certify $\epsilon = 0.2$, twelve certify $\epsilon \leq 0.267$, and two certify $\epsilon \leq 1/3$. Proposition A.14 would require $\epsilon < 1/4$ out to $|S| = \gamma T$ to obtain $a = \Theta(T)$; only the constant-weight range is certified here. Lemma A.4 connects that range to pairwise overlap, which is why the tables retain the four-cycle counts.

F.6 Detector radius beyond weight two

Table 4: Original bit-flip decoder study, across four instances per size. The exhaustive column is exact through $T = 400$ and right-censored above it; the last column reports sampled no-failure thresholds after 30 and 3000 trials.

T	Exhaustive radius a	Status	Sampled: 30 / 3000 trials
200	$1, 2, 1, 2$	exact	4.5 / 2.5
400	$2, 2, 2, 2$	exact	11.5 / 5.0
800	≥ 2 in each instance	censored	21.8 / 15.5
1600	≥ 2 in each instance	censored	55.0 / 41.0

Verification to weight 3 gives, uncensored, $a = 1, 2, 1, 2$ on the four $T = 200$ instances and $a = 2, 2, 2, 2$ on the four at $T = 400$. Larger sizes remain right-censored. At $T = 800$ we verified 1.01×10^7 of the 8.50×10^7 weight-3 patterns on each of two instances (11.8% and 12.0%) with zero failures, and at $T = 1600$, 6.27×10^6 of 6.81×10^8 (0.92%) with zero failures. Sampling weight-3 errors at random, the fraction that fail to decode is 7.0×10^{-4} at $T = 200$, 5.0×10^{-5} at $T = 400$ and 0 in 2×10^4 trials at each of $T = 800$ and $T = 1600$. These partial and sampled checks do not certify radius three; the original theorem-linked calculations therefore retain the certified lower bound $a = 2$ at the larger sizes.

F.7 Detector coverage under corruption

Outside the decoder's convergence basin the scheme reports ABSTAIN. We measure this coverage directly ($T = 200$, $\ell = 3$, three instances, 300 random corruptions per weight): the MARKED rate is 1.000 out to edit weight 2, then 0.999, 0.994, 0.958, 0.788, 0.386, 0.081 at weights $3, 4, 6, 8, 10, 12$, and 0.000 from weight 16 on, where abstention is total. The complement is entirely abstention: the UNMARKED rate is 0.000 at every weight and on every instance. The detector therefore loses

coverage rather than changing attribution. Certification applies only through $w \leq a$, with $a = 1$ on two of these instances and $a = 2$ on the third.

F.8 ALFWORLD GEOMETRY, PER TASK TYPE

Table 5 breaks the position census down by ALFRED task type. The radius-2 search restricts later repairs to a same-verb candidate set, making every coupled count a lower bound. On three trajectories checked exhaustively, the restricted search recovered one of four nonzero coupled positions across the two trajectories that contained any; the third contained none. This confirms that the reported coupled counts can miss valid repairs, without providing a population recall estimate.

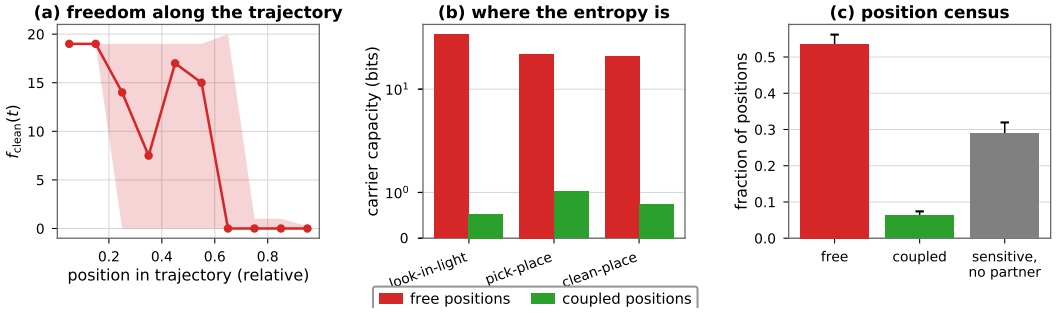

Figure 13: Geometry of 35 expert trajectories in ALFWORLD. Left: per-position freedom. Middle: measured capacity at free and coupled positions by task type (log scale). Right: the position census. Free and sensitive counts are exhaustive; coupled counts come from the incomplete radius-two search.

Table 5: Trajectory geometry on ALFWORLD `valid_unseen`. All columns after n are trajectory means. The free/sensitive split exhausts the admissible actions at each position; the restricted radius-two search makes the coupled column a lower bound.

task type	n	T	free frac.	sens. frac.	coupled/traj	bits free / coupled
look-in-light	18	17.4	0.64	0.24	0.44	46.0 / 0.53
pick-place	13	12.8	0.41	0.48	0.62	26.6 / 1.03
clean-place	4	15.5	0.47	0.41	0.75	24.9 / 0.75
all	35	15.5	0.54	0.35	0.54	36.4 / 0.74

F.9 INTERPRETING FINITE MEASUREMENTS

There are three distinct evidence directions. Adversarial search for $\min_{|e|=w} E_X(z^* \oplus e)$ returns an *upper* bound on the true minimum and can therefore falsify, but not certify, a barrier. Solver-closed infeasibility decisions in Appendix F.4 give *lower* bounds at their specified weights. A greedy utility-preserving walk gives a *lower* bound on the reach of one editor, whereas Theorem 4.4 upper-bounds every editor in its class when the certified hypotheses hold.

Table 7 reports $s_{\text{eff}}^{\text{probe}} = \min_{w \leq 0.06T} \widehat{E}_w / w$ for the full searched range. Table 6 instead uses $s_{\text{eff}}^{\leq \gamma T}$, minimised over the theorem's range $a < w \leq \gamma T$; with $a = 2$ this contains only $w = 3$. Both are computed from searched upper bounds, so the resulting $\widehat{\eta}_{\text{plug}}$ is indicative. The values are 0.82–1.79 and 3.00–3.67, respectively. Using the full-range slope would give $\widehat{\eta}_{\text{plug}} = 0.011$ at $T = 400$ and 0.006 at $T = 800$, compared with the small-radius values reported in Table 6.

F.10 CERTIFIED AND SAMPLED DETECTOR RADII

Exhaustive verification gives exact radii $1, 2, 1, 2$ at $T = 200$ and $2, 2, 2, 2$ at $T = 400$. The larger instances are right-censored at $a = 2$ (Table 4). A sampled threshold cannot replace this quantity: at $T = 1600$ it decreases from 55 to 41 when the number of trials increases from 30 to 3000. We therefore use only the per-instance certified radius in theorem-linked calculations. Proposition A.14's

asymptotic $a = \Theta(T)$ statement additionally requires expansion with $\epsilon < 1/4$ over a linear-sized range, which these finite instances do not certify.

F.11 METROPOLIS ESCAPE

Even a barrier-free chain needs $\Theta((\theta T)^2)$ steps to diffuse to radius θT. For each β we therefore fit both $\log(\text{escape}) = \lambda T + b$ and $\log(\text{escape}) = p \log T + b$ and report which has the smaller residual. The searched profile gives $\widehat{\beta}_c = T\mathsf{H}(w/T)/\widehat{E}_{\min}(w) \approx 3$, with no visible trend in T. Below that value, fitted polynomial exponents remain near 1 up to $\beta \approx 1.3$ and then rise sharply; no measured point favours the exponential fit. These runs cross at $w = 0.05T$, whereas Theorem A.22 requires $w \le \min(\gamma T, a + 1)$ and the certified value is $a + 1 = 3$. The experiment therefore describes escape from the larger geometric ball; it does not instantiate the theorem's exponential regime.

F.12 AGENT-EDITING ATTACKS

The attacker holds the key, the detector and the engine, and is limited only by its move radius and successful task completion. One attack changes green actions only at scored positions; on BASINMARK, the carrier definition gives $f_{\text{won}} = 0$ at each such position, so this attack has no legal move. A second attack searches every position. Because the scored set is recomputed after an edit, this attack can remove a carrier from the statistic without changing the carrier action itself. We report these DESCORED events separately, along with an edit-invariant variant that scores every position. The filtered statistic measures usable capacity; the all-position statistic isolates edit invariance.

F.13 EXACT NULL AND TEST POWER

The green count is a sum of fair coins, one per *distinct* action string among the scored positions, weighted by multiplicity — ALFWORLD trajectories repeat "go to …" often, so treating positions as independent would understate the variance. We compute this law exactly. If the number of distinct scored actions G is small, the 1% threshold can equal the largest attainable score; such a row has no power. We report G, the attainable maximum, the threshold, the realised false-positive rate, and the powered-row count, and compute detection rates only over powered rows.

F.14 THE THEOREM-COVERED PHASE CELLS

Take $\epsilon = 0.267$, certified through $w = 3$ on these instances (§5.4), so $(1 - 2\epsilon)\Delta = 4.67$ and $\gamma T = 3$. At $\eta = 0.01$ this gives $\zeta T = \eta T/((1 - 2\epsilon)\Delta) = 1.71$ at $T = 800$ and 0.86 at $T = 400$, both below the certified radius $a = 2$ and below γT. The allowed step sizes $r \le (\gamma - \zeta)T$ are 1.29 and 2.14, covering the $r = 1$ column. At $\eta = 0.02$, $\zeta T = 1.71$ at $T = 400$ but 3.43 at $T = 800$. Corollary A.9 therefore covers the $r = 1$ column at $\eta \le 0.02$ for $T = 400$ and at $\eta \le 0.01$ for $T = 800$: six instance–slack cells, with observed erasure $0/6$ for each attack.

Theorem D.1 gives slack cutoffs 0.0233 at $T = 400$ and 0.0117 at $T = 800$ for outputs within Hamming distance three. Those cutoffs contain the same six grid points, but the distortion condition is an output condition rather than a statement about an entire attack cell. On the directly optimized $T = 400$ instance, $L_{0,5} = 5$ also certifies the $r = 4$ cells at $\eta = 0.01$ and 0.02: no allowed nonzero first move fits the violation budget. The syndrome-component certificates in Section 4.3 provide a separate route to certifying radius-one paths after the first move becomes feasible.

With verified $a = 2$ and the searched slope, Table 6 gives the indicative $\widehat{\eta}_{\text{plug}} = 0.015$–$0.030$. The smallest slack at which a heuristic sweep attack erased is 0.15 at $T = 400$ and 0.22 at $T = 800$. The explicit radius-one straight-line path has median peak cost 0.38–0.40 and first succeeds on the grid at $\eta = 0.45$. Because the plug-in slope is a searched upper bound while the attacks give constructive upper bounds on erasure cost, these 5–$27\times$ ratios compare distinct evidence directions rather than bracket a phase boundary. Section 5.2 replaces them with a certified bracket: complete components exclude radius-one erasure for $\eta < 0.055$ at $T = 400$ and $\eta < 0.0375$ at $T = 800$, while optimized radius-one flip orders erase the same instances at 0.115–0.1225.

A key-aware replanner bypasses local editing by solving directly for a different affine class; a key-independent uniform draw from $\mathrm{Sol}(X)$ is unmarked with probability $1 - 2^{-\ell}$. This baseline is analytically characterised and is not included in the editing phase boundary.

T	a	a/T	$s_{\text{eff}}^{\leq \gamma T}$	$\widetilde{D}/T$	$\widehat{\eta}_{\text{plug}}$	$\widetilde{\eta}_{\text{path}}$	ratio
400	2	0.0050	3.00	0.190	0.0300	0.3825	12.75
400	2	0.0050	3.00	0.185	0.0300	0.3775	12.58
800	2	0.0025	3.67	0.196	0.0183	0.3987	21.75
800	2	0.0025	3.00	0.201	0.0150	0.4000	26.67

Table 6: Indicative plug-in comparisons, per instance. $\widehat{\eta}_{\text{plug}}$ inserts the verified radius a and the searched slope $s_{\text{eff}}^{\leq \gamma T}$ into the formula of Corollary A.9; it is not a certified security threshold because the slope is an upper bound on the unknown minimum. Here $a < w \leq \gamma T$ contains only $w = 3$, so each slope is a single-weight estimate rather than a robust profile. Its aggregation differs from Table 7, as Appendix H.2 explains. $\widetilde{D}$ is the median distance to the six searched evasive targets. For each target, we retain the smallest peak energy among four sampled bit orders; $\widetilde{\eta}_{\text{path}}$ is the median of those six peaks divided by n. The last column is $\widetilde{\eta}_{\text{path}}/\widehat{\eta}_{\text{plug}}$.

F.15 Committed-prefix editing

We operationalise the enforcement story of Appendix H.9 on four parity tasks ($T \in \{400, 800\}$, two instances each). Before tick k, an editor may make one radius-one flip in the uncommitted suffix; the environment accepts it only if the resulting trajectory meets the utility floor, then executes and freezes decision k. A paired control has the same marked start, evasive target, validator, horizon, and edit budget but no frozen prefix. We use six targets per task and thirteen slack values, for 312 paired protocols. All 24 no-validation sentinels reach their exact-valid targets, with no prefix mutation, step-radius violation, utility-floor violation, or sentinel failure.

	Committed prefix		No commit	
η	q-erasure	exact erasure	q-erasure	exact erasure
≤ 0.15	$0/24$	$0/24$	$0/24$	$0/24$
0.22	$23/24$	$11/24$	$24/24$	$24/24$
≥ 0.32	$24/24$	$24/24$	$24/24$	$24/24$

At $\eta = 0.22$, the exact-erasure counts split as $5/12$ versus $12/12$ at $T = 400$ and $6/12$ versus $12/12$ at $T = 800$. The corresponding formal q-erasure counts are $11/12$ versus $12/12$ and $12/12$ versus $12/12$: commitment changes exact reach much more than whether the editor ever touches an approximately valid unmarked state. These are ever-hit events—an attacker may stop at first success—so an earlier abstention and a later erasure are not mutually exclusive.

The editor follows one supplied target monotonically, so the comparison concerns this editor, not all of $\mathcal{A}_{\text{loc}}(q, 1)$. Within that scope, commitment reduces exact erasure at $\eta = 0.22$ from $24/24$ to $11/24$, while formal q-erasure changes only from $24/24$ to $23/24$. The difference isolates how irreversibility and deadlines obstruct exact target reach even when an approximately valid unmarked state is almost always touched.

F.16 Barrier and reach profiles

At $T = 1600$, taking the minimum over four instances, the smallest violation count found at distance w rises through $5, 8, 11, 14, 18, 24$ at $w = 1, 2, 3, 4, 6, 8$; the ratio $E_{\min}(w)/w$ then decreases from 2.53 at $w = 17$ and 2.04 at $w = 27$ into a plateau, sitting between 1.72 and 1.97 for $37 \leq w \leq 96 = 0.06T$. The smallest ratio over the probed range is $0.82, 1.08, 1.17, 1.72$ at $T = 200, 400, 800, 1600$. As searched upper bounds, these values show that the tested search did not find a sublinear profile; they do not certify linear growth.

Greedy reach shows a parallel T-scaled pattern. Writing $s = \eta n/\text{reach}(\eta)$ for the barrier slope implied by how far the greedy attacker actually got, the same η gives essentially the same s at every size: at $\eta = 0.09$, $s = 3.00, 2.25, 2.25, 2.40$ across $T = 200, 400, 800, 1600$; at $\eta = 0.13$, $s = 1.86, 1.53, 1.73, 1.58$. The reach itself is a fixed fraction of T — 1.5–2.0% at $\eta = 0.09$ and 3.5–4.3% at $\eta = 0.13$ across the four sizes — before breaking to $\approx 0.5T$ at $\eta = 0.25$. This is a lower bound on the reach of one greedy editor. Detector-radius certification and censoring are reported separately in Appendices F.6–F.10.

F.17 THREE EMPIRICAL REGIMES

At $T = 800$, the phase sweep separates into three bands. For $\eta \leq 0.04$, attacks neither erase nor cause abstention. At $\eta = 0.07$–0.15, erasure remains zero while the informed radius-one attacker's abstention rate rises $0.08 \to 0.58 \to 1.00$ (averaged over $r \in \{1, 4, 16\}$: $0.03 \to 0.22 \to 0.94$). This is denial of attribution, reported separately from erasure. The blind walk never erases but reaches total abstention from $\eta = 0.07$ at $r = 1$ and from $\eta = 0.15$ at every radius. At $\eta \geq 0.22$, the informed attack erases at every tested radius with success rate 1.00.

At $T = 400$ the middle band ends one grid point earlier and less sharply: one of the two instances already erases at $\eta = 0.15$, in 1 of 6 trials, at $r = 4$ and $r = 16$ but not at $r = 1$; and at $\eta = 0.22$ the success rate is 0.83–1.00. Thus the boundary is radius-independent in the $T = 800$ grid but not in the smaller instance. Across both sizes the blind walk never erases, whereas the informed attacker succeeds once slack permits it: target information, not slack alone, converts local motion into erasure. These bands describe the heuristic editors. Optimized flip orders erase the same instances at 0.115 ($T = 400$) and 0.1175–0.1225 ($T = 800$), inside the intermediate band (Section 5.2).

F.18 EXACT PLANNING-SPACE GEOMETRY

The corridor-maze and blocksworld studies enumerate the valid plan sets and their local-edit graphs exactly. Corridor-maze components coincide with route homotopy classes, while the matched blocksworld control remains connected. These experiments establish that edit-separated valid sets occur outside parity. They do not encode the domains in Definition A.1 or verify local sensitivity and expansion, so their connection to Lemma 4.3 is at the level of the geometric phenomenon rather than the sufficient mechanism.

F.19 ALFWORLD CENSUS

The census contains 35 expert trajectories from three task types, with mean length 15.5. A 0.536 fraction of positions are FREE and a 0.351 fraction are SENSITIVE; 19 of the 35 have at least one COUPLED position and the mean count is 0.54 per trajectory. Converting each into carrier capacity — $\sum_t \log_2(1 + f(t))$ over the relevant positions — gives 36.4 bits at free positions against 0.74 bits at coupled ones. Among the alternatives found by this search, most measured capacity lies where a single utility-preserving edit can reach it. The coupled capacity is a lower bound from an incomplete radius-two search, so this comparison does not upper-bound all coupled capacity in these trajectories. Table 5 breaks this down by task type and Figure 13 shows the per-position profile. The split varies by task type: `pick_and_place` ($n = 13$), where the choice of *which* object propagates to a later step, has the lowest free fraction (0.407, against 0.644 for `look_at_obj_in_light`, $n = 18$) and the highest measured coupled capacity (1.03 bits against 0.53).

The carrier sets are small enough that the 1% detection test is frequently powerless: with one or two distinct scored actions the 1% quantile of the exact null coincides with the largest attainable score. Such rows are counted and excluded from detection rates. Within the six evaluated trajectory–key rows per scheme, one utility-preserving radius-1 move reaches an ACTIONMARK carrier on every row and a BASINMARK carrier on none, while both schemes preserve the task exactly. This is the carrier behavior built into the constructions. The rows contain only 1.67 and 0.67 certified carriers, respectively, leaving all six rows per scheme powerless. The census therefore measures where candidate carrier capacity lies; it does not supply a powered scheme comparison.

F.20 Supporting Figures and Tables

T	n	d	4-cyc	a cert.	sampled (few)	sampled (many)	$s_{\text{eff}}^{\text{probe}}$
200	100	100	7.0	1.5	4.5	2.5	0.82
400	200	200	1.5	2.0	11.5	5.0	1.08
800	400	400	1.5	2.0	21.8	15.5	1.17
1600	800	800	0.5	2.0	55.0	41.0	1.72

Table 7: Barrier and detector measurements. Columns are means over four instances per size, except $s_{\text{eff}}^{\text{probe}}$, which is the minimum over the probed range. The certified radius exhausts every error pattern through weight two; the sampled columns report the largest no-failure weight after 30 and 3000 random patterns.

β	exp. rate λ	poly. exponent p	better fit	censored cells
0.40	0.0056	1.01	poly	0
0.70	0.0053	0.95	poly	0
1.00	0.0045	0.79	poly	0
1.30	0.0054	0.96	poly	0
1.60	0.0070	1.24	poly	0
1.90	0.0153	2.73	poly	0
2.20	–	–	–	5

Table 8: Growth rate of $\log$(basin-escape time) in T, by inverse temperature. Cells dominated by censored runs are excluded from the fit and counted.

domain	#valid	exhaustive	L	env. classes	r_c	comps. at $r < r_c$
gridworld_sep2_gates2	256	yes	16	3	3	3
gridworld_sep3_gates2	440	yes	17	3	4	3
gridworld_sep4_gates2	688	yes	18	3	5	3
gridworld_sep5_gates2	1008	yes	19	3	6	3
gridworld_sep6_gates2	1408	yes	20	3	7	3
gridworld_sep7_gates2	1896	yes	21	3	8	3
gridworld_sep8_gates2	2480	yes	22	3	9	3
gridworld_sep10_gates2	3968	yes	24	3	11	3
blocksworld_2pairs	6	yes	4	–	1	–
blocksworld_3pairs	90	yes	6	–	1	–
blocksworld_4pairs	2520	yes	8	–	1	–

Table 9: Connectivity of fully enumerated valid-plan sets. The critical radius r_c is the smallest radius at which the local-edit graph becomes connected. In every gridworld, the three subcritical components coincide with environment-defined homotopy classes (purity 1.00).

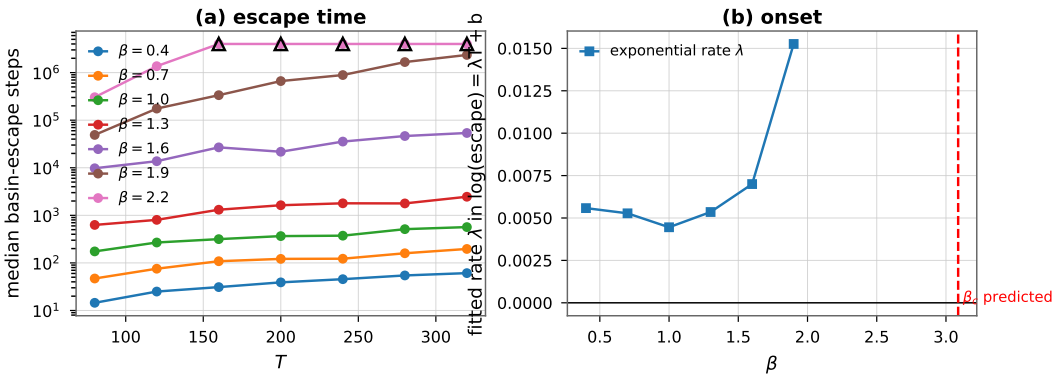

Figure 14: (a) Median basin-escape time against T, by inverse temperature; triangles mark censored cells. (b) The fitted exponential rate λ in $\log(\text{escape}) = \lambda T + b$, against β. It is positive but small at every accessible β and rises sharply as β grows; the predicted β_c (red dashed) lies beyond the measured range. Circled points would be those where the exponential fit beats the polynomial one — there are none, which is the finding.

G Scope of the Guarantees

The guarantees concern editors that remain close to the supplied trajectory or move through utility-preserving local steps. Inside this class they are unconditional, allow unbounded computation, and reveal the key and detector. Corollary A.9 applies to $\mathcal{A}_{\text{loc}}(q, r)$; Theorems A.18 and A.19 identify how larger steps or larger utility budgets leave that regime.

For sequential editing, condition (b) of Definition A.8 is enforced when the trajectory is executed while it is revised: committed decisions are irreversible and the environment validates each next step. Theorem D.1 covers a complementary setting: it constrains the final output within distortion γT and slack $(1 - 2\epsilon)\Delta a/T$, without assumptions on intermediate states. Section A.6 instead prices invalid intermediate states.

Re-planning is a different operation. In the parity instantiation, Gaussian elimination can solve directly for a different affine class, and a key-independent uniform valid draw is unmarked with probability $1 - 2^{-\ell}$ by Theorem A.16. The barrier is therefore useful when preserving and editing the supplied trajectory is materially cheaper than solving the task again. This distinction also separates our geometric obstruction from inference-hardness watermarks (Abdalla & Vershynin, 2025) and from unconstrained regenerators such as learned paraphrasers (Huang et al., 2025).

Lemma 4.3 itself extends to hard-to-solve locally sensitive families. The 1-in-Δ calculation in Appendix H.6 identifies large-arity parameters that survive the abundance first moment, including $(c, \Delta) = (5, 100)$ with exponent $+0.009$. An efficient generator and bounded-distance detector for such a family remain the missing ingredients for a construction that is also hard to re-plan.

H Additional Analysis

H.1 Certified radius

The radial output-fidelity threshold depends on the radius a at which the detector is certified constant. For the original bit-flip decoder, the sampled radius grows when the sample misses rare failures and falls from 55.0 to 41.0 at $T = 1600$ when trials increase a hundredfold. We therefore use certified a throughout. Table 6 combines this value with a searched, upper-bounding slope, so its plug-in threshold remains indicative rather than certified.

H.2 Two barrier-slope summaries

$s_{\text{eff}}^{\text{probe}}$ minimises over the whole probed range and summarises the search for departures from linear growth; $s_{\text{eff}}^{\le \gamma T}$ minimises over $a < w \le \gamma T$, the range used by Theorem D.1. Both derive from searched upper bounds. Their factor-of-three difference reflects the steeper profile at small w; Appendix F.9 reports the corresponding sensitivity calculation.

H.3 Abstention and erasure

Three-valued detection makes a third outcome visible: an attacker can drive the detector to abstain without producing a valid unmarked artefact. A blind radius-one walk reaches total abstention from slack 0.07 while never erasing. All tables therefore report abstention and erasure separately. In the agent study, the analogous event is an edit that drops a carrier from the scored set without changing its action.

H.4 Instance-level checks

An instance can be checked for local sensitivity, pairwise overlap μ, which certifies expansion for $w \le 2\epsilon c/\mu + 1$ through Lemma A.4 — here $w \le 3$; and the detector radius a. The available first-moment argument does not certify expansion to $w = \gamma T$ at simulated sizes (Remark I.3); finite instances consequently use constant-radius certificates unless a stronger instance-specific barrier certificate is supplied.

H.5 Density is not the whole story

Lemma A.7 might suggest that abundance forces near-parity density, but density alone is insufficient. The failure of this inference identifies the relevant nonlinear families. Write $P_R(\alpha)$ for the probability that a Δ-window with i.i.d. Bernoulli(α) entries lands in R. Over the random (c, Δ)-biregular ensemble, the first-moment (annealed) count obeys

$$\frac{1}{T}\log_2 \mathbb{E}_X|\mathrm{Sol}(X)| \;\le\; \max_{\alpha\in\mathcal{A}}\Big[\, H_2(\alpha) + \tfrac{c}{\Delta}\log_2 P_R(\alpha)\,\Big] + o(1), \tag{5}$$

where H_2 is the base-2 binary entropy — distinct from the natural-logarithm H used in §A.6; the counting arguments are natural in bits and the Gibbs arguments in nats — and $\mathcal{A} \subseteq [0,1]$ is the set of densities a solution can actually have. That last restriction is the whole point. Equation (5) is the (c, Δ)-biregular, bit-normalised specialisation of the annealed entropy of Zdeborová & Mézard (2008a). Over the configuration model the first moment is exactly

$$\mathbb{E}_X|\mathrm{Sol}(X)| = \sum_w \binom{T}{w}\binom{Tc}{wc}^{-1} [x^{wc}]\, f_R(x)^{\,n}, \qquad f_R(x) = \sum_k r_k x^k, \quad n = cT/\Delta,$$

with $r_k = |\{y \in R : |y| = k\}|$: left-degree regularity means only this weight enumerator enters. The bracket in (5) is that coefficient evaluated at the tilt $x = \alpha/(1-\alpha)$ rather than at the saddle, whence the inequality; it is an equality at a given α exactly when the tilt *is* the saddle, i.e. when $\mathbb{E}[\,|y| : y \in R\,] = \alpha\Delta$ under Bernoulli$(\alpha)^{\Delta}$. Note also that $\mathcal{A}$ is not an extra hypothesis but a consequence: $[x^{wc}]f_R^{\,n}$ vanishes off $\mathcal{A}$. For 1-in-Δ, $f_R(x) = \Delta x$ is a monomial, so $\mathcal{A} = \{1/\Delta\}$ is forced by the coefficient extraction alone. *When the degree count pins the solution weight, $\mathcal{A}$ is a single point and the maximum is not free to roam.* Being a first moment, (5) is an upper bound on the exponent, so a negative value rules abundance out while a positive one does not by itself establish it.

Two sanity checks. The naive product $2^T \prod_j (|R_j|/2^{\Delta})$ is the bracket evaluated at $\alpha = \frac{1}{2}$; since the bracket is now an upper bound, that alone no longer makes it a lower bound on the exponent, so it needs an independent reason, and it has one. Keeping only the term $w = \mu^* T$, where $\mu^* = f_R'(1)/(\Delta f_R(1))$ is the mean density of R, gives

$$\frac{1}{T}\log_2 \mathbb{E}_X|\mathrm{Sol}(X)| \;\ge\; \Big[1 + \tfrac{c}{\Delta}\log_2 \tfrac{|R|}{2^{\Delta}}\Big] + (c-1)\big(1 - H_2(\mu^*)\big) + o(1),$$

the bracketed term being exactly the naive product and the slack vanishing only when R is weight-balanced. This also quantifies how loose the naive product is when solutions concentrate away from

density $\frac{1}{2}$. And for parity $P_R(\alpha) = \left[1 \pm (1-2\alpha)^{\Delta}\right]/2$, so $P_R(\frac{1}{2}) = \frac{1}{2}$; nothing pins the weight, the maximiser is $\alpha = \frac{1}{2}$, and (5) returns $1 - c/\Delta$. For the even class this merely reproduces a deterministic count: $\mathrm{Sol}(X)$ is then the $\mathbb{F}_2$-kernel of the biadjacency matrix, so $|\mathrm{Sol}(X)| = 2^{T-\mathrm{rank}} \geq 2^{(1-c/\Delta)T}$ for *every* instance, and (5) is an exact sandwich rather than a one-sided bound. (We scope this to the even class deliberately: for the odd class the system is affine rather than a kernel and can be inconsistent, e.g. $c = \Delta = 2$ with T odd, where summing all $n = T$ equations gives $0 = 1$.)

H.6 A NONLINEAR CANDIDATE FAMILY

The natural non-linear locally sensitive family is 1-in-Δ, $R_j = \{y : \text{exactly one } y_p = 1\}$, for which finding a solution is NP-hard — exactly the property one wants if the barrier is to be informative against a re-planner (Appendix G). Its density is only $\Delta 2^{-\Delta}$, so the naive count is hopelessly negative. But the degree count pins the weight: summing $|C_j \cap \mathrm{supp}(z)|$ two ways gives $c|z| = n = cT/\Delta$, so *every* 1-in-Δ solution has weight exactly T/Δ. Thus $\mathcal{A} = \{1/\Delta\}$, and (5) must be evaluated there rather than maximised: with $P_R(\alpha) = \Delta\alpha(1-\alpha)^{\Delta-1}$ the exponent is $H_2(1/\Delta) + \frac{c}{\Delta}(\Delta - 1)\log_2(1-1/\Delta)$, which is $+0.059$ at $(c, \Delta) = (3, 10)$, $+0.005$ at $(4, 20)$, -0.065 at $(5, 20)$, $+0.009$ at $(5, 100)$, -0.005 at $(6, 100)$ and $+0.002$ at $(6, 200)$. Asymptotically the exponent is positive iff $\Delta \gtrsim e\, c^{-1}$; scanning exactly, the least admissible arity is $\Delta = 7$ at $c = 3$, 19 at $c = 4$, 53 at $c = 5$, 146 at $c = 6$ and 401 at $c = 7$. The small-arity values agree with Zdeborová & Mézard (2008a), whose Table I gives $L_s = 3$ for 1-in-3 through 1-in-6 in the regular ensemble — consistent with our "least admissible arity $\Delta = 7$ at $c = 3$" (their table stops at $K = 6$), by a sharper computation (exact replica-symmetric entropy, with a second moment in the balanced cases) than a first moment. The large-arity calculation extends beyond their $K \leq 6$ range. The family survives this abundance test only at *large* arity, and among examples with $c \geq 5$ — the range in which the expansion certificate of Remark I.3 can be non-empty at all — $(5, 100)$ qualifies and $(5, 20)$ and $(6, 100)$ do not.

This makes large-arity 1-in-Δ a concrete candidate for combining a locally sensitive barrier with NP-hard re-planning. Completing a watermark construction requires a generator for its solutions and a bounded-distance detector (Appendix H.12). Lemma A.7 says any non-parity locally sensitive relation has $|R| < 2^{\Delta-1}$ and so pays $\log_2(2^{\Delta-1}/|R|)$ bits per constraint relative to parity; (5) says that loss must be repaid by the solutions concentrating at a density where P_R is unusually large. 1-in-Δ repays it exactly this way, and only once Δ is large enough.

H.7 A CERTIFIED PARAMETER POINT

Take $(c, \Delta) = (12, 24)$ (rate $1/2$) and $\epsilon = 0.245$, inside the band $\epsilon \in (1/c, 1/4)$ of Remark I.3. The first-moment certificate there gives $\gamma = 7.0 \times 10^{-5}$, and Proposition A.14 certifies the detector at radius $a = \lceil (1-2\epsilon)\lfloor \gamma T \rfloor \rceil - 1$. At $T = 10^6$ decisions this yields $n = 5 \times 10^5$ constraints, $\gamma T = 70$, and $a = 35$. Taking $\eta = 4.2 \times 10^{-4}$, which satisfies (1), the attacker may leave at most 210 of the 500,000 constraints violated; the confinement radius is $\zeta T = 34.3 < 35$ decisions and the per-move radius may be as large as $r = 35$.

Thus, for *a trajectory of a million decisions, an attacker holding the key and the detector and unbounded computation, permitted to rewrite up to* 35 *decisions per move and to leave* 210 *constraints broken — and it provably cannot erase the mark, because it can never move more than* 35 *decisions from where it started.*

This example makes the asymptotic statement concrete. Its slack 4.2×10^{-4} and horizon 10^6 also display the cost of the first-moment constants. For comparison, the $(5, 10)$ certificate has no positive finite-radius consequence at $T \leq 1600$, although search finds a roughly linear profile to $0.06T$. The direct certificate in Appendix F.4 is the finite-instance route around this gap.

H.8 A BARRIER-PROFILE INTERFACE

Theorems 4.4 and D.1 both exploit a specific feature of expander instances: the certified lower bound grows linearly with distance to γT. Their proofs need only a lower-certified radial profile over the distances they traverse. Proposition D.2 states this weaker, instance-level interface and separates barrier certification from the particular expansion argument used to obtain it.

H.9 Operational enforcement of condition (b)

Condition (b) is enforced when the artefact is *in use while it is being edited*. In the committed-prefix model, the trajectory is executed left to right, decisions $z_1, \ldots, z_k$ have already been carried out and are irreversible, an editor may touch only the uncommitted suffix, and the environment validates the current trajectory before executing the next decision. This models issued tool calls, staged design checks, and plans revised during execution. Offline editing instead uses Theorem D.1 or the priced-validity model of Section A.6. Appendix F.15 implements the sequential model directly.

H.10 Why the straight-line attack is cheap

The mechanism is worth stating outside the proof, because it is the mirror image of Lemma 4.3. Both endpoints are valid, so the difference e is a codeword, so *every* constraint meets $\mathrm{supp}(e)$ an even number of times. A constraint can therefore only be broken while the walk has flipped some but not all of the difference it sees — which means every violated constraint is one the walk is halfway through, and there are at most $c\min(|S|, D - |S|)$ of those. Leaving a valid trajectory is expensive because expansion isolates the change; travelling between two valid trajectories is cheap because parity pairs the changes up.

H.11 Division of labour between task and key

Inside the confinement ball the attacker cannot reach *any* valid trajectory other than z^0 — Corollary A.10 — so it cannot change which plan it holds, let alone which mark. The key (M_κ, s_κ) does not enter that geometric argument. More generally, the task determines the reachable subspace K_B; the affine test determines which reachable directions change the mark. Independently, the affine test gives a false-positive probability of at most $2^{-\ell}$ for key-independent artefacts. Both the radial and syndrome certificates allow key-aware editors.

H.12 Open Problems

1. **Close the bracket at large T.** Complete components give exact thresholds through $T = 200$ and certified lower bounds beyond, and optimized paths bound the threshold from above within about a factor of two at $T = 400$ (Section 5.2). Closing the bracket for larger T needs lower bounds on the minimal path barrier $\min_{\text{paths}} \max_t E_X(z^t)$ that do not enumerate the component, for example cut certificates.
2. **Provable constants.** Remark I.3 shows that first-moment expansion certificates say nothing at all at simulable sizes, while heuristic searches only probe the barrier there. Our direct computation certifies $L_{2,5} = 7$ on one instance but finds a weight-six energy-six counterexample and leaves weights eight to ten unresolved. Extending exact profiles across instances and beyond this immediate boundary—or finding a certificate that verifies the *conclusion* of Lemma 4.3 directly on a given instance, rather than the expansion hypothesis—would extend the finite radial certificates; the direct finite-instance profile in Proposition D.2 states exactly what must be computed.
3. **A converse below β_c.** Theorem A.22 gives a sufficient condition for exponentially slow escape. Is erasure by Metropolis actually *fast* below the threshold, and is there a sharp transition? Our data are consistent with one but do not establish it. Time-inhomogeneous schedules (simulated annealing) are entirely open.
4. **Larger certified decoding basins.** The output-fidelity bound in Theorem D.1 improves with the certified radius a. Proposition A.14 gives $a = \Theta(T)$ under suitable expansion; at finite scale, ordered-statistics decoding or exhaustive repair may offer alternatives, but their radii still need certification. The syndrome erasure criterion takes a different route and does not depend on a. Combining larger certified basins with the reachable-component calculation would extend positive attribution as well as resistance to erasure.
5. **Beyond binary decisions and parity.** Lemma 4.3 extends verbatim to alphabets of size m with locally sensitive relations. Which non-linear locally sensitive families admit an efficient generator and an efficient bounded-distance detector? Answering this is the route to the hard-to-re-solve schemes discussed in Appendix G.

6. **Agent trajectories beyond one benchmark.** Appendix F.8 measures the geometry on ALFWORLD, where "locally sensitive constraint" is estimated by the engine rather than declared. Whether the same census holds on tool-calling and code-editing agents — and whether any agent task puts a usable amount of entropy behind a barrier — is open.

I TECHNICAL REMARKS

This section collects technical qualifications and comparisons referenced from the main text.

Remark I.1 (Beyond linear constraints)**.** Lemma 4.3 uses only Definitions A.2 and A.3. Over a linear family the conclusion is the classical confinement property of expander codes (Sipser & Spielman, 1996), and its qualitative non-linear counterpart — that leaving a solution requires flipping a closed loop — is known for locked CSPs (Zdeborová & Mézard, 2008b;a). Our pairwise sensitivity condition makes the edge count apply verbatim to nonlinear relations and retain a bound linear in w. It therefore covers families such as 1-in-3, where finding a solution is NP-hard, even though the efficient generator and detector in Section 3 use parity.

Remark I.2 (A ceiling that expansion imposes on itself)**.** Assume exact c-regularity, so $n = cT/\Delta$. Since $\Gamma(S) \subseteq [n]$, applying Definition A.3 at $|S| = \lfloor \gamma T \rfloor$ forces $(1-\epsilon)c\lfloor \gamma T \rfloor \le cT/\Delta$, hence

$$\Delta\gamma \;\le\; \frac{1}{1-\epsilon}\big(1+o(1)\big), \qquad \text{and so} \qquad (1-2\epsilon)\,\Delta\,\gamma \;\le\; \frac{1-2\epsilon}{1-\epsilon}\big(1+o(1)\big), \qquad \frac{1-2\epsilon}{1-\epsilon} \;\le\; 1. \tag{6}$$

Every utility-slack threshold we derive from Lemma 4.3 is therefore bounded by an absolute constant *no matter how large* Δ *is*: the bound $(1-2\epsilon)/(1-\epsilon)$ equals 1 at $\epsilon = 0$ and $2/3$ at $\epsilon = 1/4$, so in the regime $0 \le \epsilon < 1/4$ that Proposition A.14 needs it lies in $(2/3, 1]$ and is strictly below 1 when $\epsilon > 0$. Raising the constraint arity cannot evade this ceiling, because expansion forces γ down like $1/\Delta$. Thus the thresholds are $\Theta(\gamma\Delta) = O(1)$.

Remark I.3 (Provable versus measured constants)**.** Two conditions must hold simultaneously for the decoder-based certificate. First $\epsilon < 1/4$, so that Proposition A.14's decoder is available. Second, ϵ cannot be too small: taking $S = C_j$ shows that any $(\gamma T, (1-\epsilon)c)$-expander with $\gamma T \ge \Delta$ obeys $|\Gamma(S)| \le c\Delta - (\Delta - 1)$ and hence $\epsilon \ge (1 - 1/\Delta)/c$, a hard combinatorial necessity. Separately, the first-moment certificate below is empty unless $c\epsilon > 1$. Together these force $\epsilon \in (1/c, 1/4)$ for the certificate to say anything, a band that is *empty for* $c \le 4$: at least $c = 5$ is required.

Within that band, the first-moment certificate — a random (c, Δ)-biregular graph is a $(\gamma T, (1-\epsilon)c)$-expander whp for every γ below the first positive root of $H_2(\gamma) + (c/\Delta)H_2((1-\epsilon)\Delta\gamma) + c\gamma \log_2((1-\epsilon)\Delta\gamma)$ — gives very small constants:

(c, Δ)	ϵ	γ_{FM}	η_{sec}	$T_{\min} = 2/\gamma$	
$(5, 10)$	0.245	1.0×10^{-14}	2.6×10^{-14}	2.0×10^{14}	*experimental family*
$(8, 16)$	0.245	4.0×10^{-6}	1.7×10^{-5}	5.0×10^{5}	
$(10, 20)$	0.245	2.8×10^{-5}	1.5×10^{-4}	7.1×10^{4}	
$(12, 24)$	0.245	7.0×10^{-5}	4.3×10^{-4}	2.9×10^{4}	

For the experimental $(5, 10)$ family, this argument gives $\gamma \approx 10^{-14}$ and requires $T \approx 2 \times 10^{14}$ before the protected annulus is nonempty. The asymptotic example in Appendix H.7 instead uses $(12, 24)$ and $T = 10^6$. At simulated sizes, an instance-level profile can replace the first-moment expansion route because the theorems need the *conclusion* of Lemma 4.3. Appendix F.4 certifies this profile through distance five on one $T = 400$ instance; its weight-six witness sets the immediate boundary. The adversarial search at $T = 1600$ reaches ratio about 1.7 through $w = 96 = 0.06T$, but remains an upper bound on the unknown minimum.

Remark I.4 (The ceiling, concretely)**.** A detector certified at radius a forces $d_{\min} \ge 2a + 1$, so $a \le (d_{\min} - 1)/2$; and by Remark I.2 the protected-slack coefficient $(1-2\epsilon)\Delta\gamma$ is asymptotically at most 1 (strictly below 1 when $\epsilon > 0$). Whenever $a \le \gamma T$ the first condition in (1) therefore never protects a slack η beyond $(1-2\epsilon)/(1-\epsilon) \le 1$. This abstract ceiling only rules out protecting the entire unit slack range (strictly so for $\epsilon > 0$); for $0 \le \epsilon < 1/4$ it lies in $(2/3, 1]$ and therefore does *not*, by itself, imply a near-perfect-utility regime. The near-perfect thresholds in our concrete certificates come instead from the very small certified γ and a/T. Increasing Δ cannot remove the

abstract ceiling, but the present constants — not this abstract upper bound — are what make our proved numerical regime narrow.

Remark I.5 (Relation to overlap gaps)**.** Condition (iii) is universally quantified over an *instance*: no trajectory whatsoever with utility above the floor lies in the band. Overlap-gap results in random optimization instead describe where *typical* near-optima sit in a random ensemble, and they rule out *stable* algorithms rather than all efficient ones (Gamarnik, 2021; Gamarnik & Sudan, 2017; Huang & Sellke, 2021). A barrier established only for typical solutions can be circumvented by a rare, well-connected, subdominant cluster, which is exactly what happens for the symmetric binary perceptron (Abbe et al., 2021). Because (iii) quantifies over everything, no such cluster can exist. Section 4.2 establishes this stronger, instance-wide statement from checkable structure.

Remark I.6 (The role of $\epsilon < 1/4$)**.** Only Proposition A.14 uses $\epsilon < 1/4$. Lemma 4.3, Theorem 4.4, Corollary A.9 and Theorem D.1 need $\epsilon < 1/2$ together with a radius a at which the detector is certified — by whatever means. Exhaustive certification of a therefore lets the finite-profile result apply at $w = 3$ even on instances whose weight-three expansion gives only $\epsilon \leq 0.267$.

Remark I.7 (Scale of the finite-instance gap)**.** The ratios in Table 6 are constant-order in T but are not two certified sides of a bracket. The plug-in value $\widehat{\eta}_{\mathrm{plug}} = 0.015$–$0.030$ uses a certified radius and a searched, upper-bounding slope; the constructed paths have median peak cost 0.38–0.40; and the smallest slack at which a heuristic sweep attack erased is 0.15–0.22. These figures differ by factors of 5–27, but their evidence directions do not locate the true boundary between them. Section 5.2 locates it within about a factor of two: certified at 0.055 and constructively exceeded at 0.105–0.115 for $T = 400$. In particular, the data do not justify assigning "almost all" of the difference to $a = 2$. Raising a is one actionable decoder question (Appendix H.12); separately, the geometric quantity is the minimal path barrier between codewords, for which the straight-line construction is only an upper bound.

J Additional Related Work

Limits of watermark robustness. Zhang et al. (2023) show that no watermark can be strongly robust once a quality-preserving perturbation oracle induces an efficiently mixing walk on high-quality outputs; the result is an attack, i.e. an upper bound, and assumes irreducibility and fast mixing rather than deriving them. Harel-Canada et al. (2025) test that assumption empirically, find that real text mixes slowly, and raise the possibility of disconnected "islands" in the high-quality subgraph. Our structural condition makes this intermediate geometry checkable on a task instance and turns it into a lower bound. Francati et al. (2025) give sharp unconditional corruption-rate thresholds (matched, below the threshold, by information-theoretic constructions) through a coding abstraction; those are information-theoretic limits at the symbol level, with no computational or locality parameter, and are complementary to ours.

Constraint-based watermarking. Kahng et al. (2001) watermark the solution of a design task by biasing which valid solution the solver returns, with antecedents and companions in graph colouring (Qu & Potkonjak, 1998) and integer programming (Megerian et al., 2002); the sequential-decision analogue, hiding information in the free choice among legal moves, is due to Hernández-Castro et al. (2006). That literature measures a mark's credibility — the probability that the keyed constraints are satisfied by chance. We take the carrier and the credibility metric as given and analyse resiliency to editing.

Codes the watermarker builds. A large body of work obtains robustness by *adding* redundancy: error-correcting codes over the payload (Chao et al., 2024), pseudorandom error-correcting codes with cryptographic robustness (Christ & Gunn, 2024), and, in steganography, wet paper codes whose free and locked coordinates superficially resemble our information set (Munuera & Barbier, 2011). The difference is whose code it is. There the watermarker constructs the code and pays for its rate; here the code is the task's own consistency structure, the watermarker adds nothing, and the constraint geometry that supplies the editing barrier is inherited from the task.

Watermarking agent trajectories. Marking agent behaviour is an active area: keyed hook actions in trajectory datasets (Meng et al., 2026), two-channel attribution with selection and tally channels (Gao et al., 2026), biased choice among functionally equivalent tool-execution paths (Wang et al.,

2026), planning-level multi-bit identifiers (Huang et al., 2026), and history-conditioned transition patterns (An et al., 2026); see Phan et al. (2026) for a survey. These works develop constructions and empirical robustness studies; our complementary question is when task structure itself certifies an obstruction to erasure.

Landscape barriers, and existence versus discoverability. Near-optimal solutions of random problems can cluster and obstruct local search, as studied by the overlap gap property (Gamarnik & Sudan, 2017; Gamarnik, 2021) and its branching refinement (Huang & Sellke, 2021); the same geometry has recently been used as a security criterion for collision resistance (Benedetti et al., 2025) and robust hashing (Bogdanov et al., 2025). The broader idea that robust classifiers or attacks may exist while being hard to find is older still in adversarial robustness (Bubeck et al., 2018; Garg et al., 2019), and in watermarking specifically Abdalla & Vershynin (2025) obtain unremovability from a statistical-to-computational gap. Two distinctions locate our result within this line. First, the hardness there is *inference* hardness — the adversary cannot estimate the key-bearing distribution — whereas ours is *search* hardness over the output space, and neither guarantee implies the other. Second, our condition is instance-wise and universally quantified rather than a statement about typical solutions in an ensemble, which rules out the rare well-connected cluster of Abbe et al. (2021); see Remark I.5. Finally, the clustering of random XORSAT solution sets (Ibrahimi et al., 2015) concerns the geometry of solutions rather than the cost of moving between them.

Locked CSPs and expander codes. The closest structural precedent is not XORSAT but the *locked* constraint satisfaction problems of Zdeborová & Mézard (2008b;a), which cover *both* of our constraint families: their locking condition is exactly local sensitivity for symmetric relations, parity and 1-in-K are their canonical examples, and they already state the qualitative consequence — solution clusters are point-like, so moving between solutions requires flipping a closed loop of variables. Their $\log N$ separation comes from the length of the shortest loop through a typical variable in a locally tree-like factor graph (Zdeborová & Mézard, 2008a, Sec. II.B), and their extensive separation is a cavity-method prediction. Lemma 4.3 differs in three respects that the watermarking application needs: it is deterministic and instance-wise, it is an *energy* statement — a count of violated constraints, not merely non-existence — and it is linear in w rather than a threshold at one scale. On the linear side the same statement is the confinement/robustness property of expander codes (Sipser & Spielman, 1996). The increment is the combination and the quantitative form, not either ingredient.

Two further connections clarify the role of our upper bounds. The quantity §4.3 bounds — $\min_{\text{paths}} \max_t E_X(z^t)$, the least peak violation count on a path between two codewords — is the *energy barrier* of a code; Zhao et al. (2025) give tight results for the quantum analogue on hypergraph product codes, and Theorem A.19 is the standard direct-path upper bound on it. And pairing linear confinement with a sphere-entropy count to get a size-independent critical inverse temperature, which is what Theorem A.22 does, is the classical thermal-stability argument for code Hamiltonians; what is ours there is its use to price an *attacker's* budget rather than a thermal environment's.

Relation to the overlap gap property. Our condition creates a forbidden band of overlaps, and the overlap-gap literature motivates this geometric viewpoint (Gamarnik, 2021; Gamarnik & Sudan, 2017; Huang & Sellke, 2021). It is not an OGP result in the technical sense; see Remark I.5. Using landscape geometry as a security criterion is itself not new (Benedetti et al., 2025; Bogdanov et al., 2025); using it for watermark erasure is.

K On the Epigraph

The epigraph comes from the "Autumn Floods" chapter of the *Zhuangzi*, one of the Outer Chapters traditionally attributed to Zhuang Zhou and usually dated to the third century BCE; we quote the translation of Legge (1891). In the chapter, the River Earl follows his current to the sea and sees it for the first time. The Lord of the Northern Sea tells him that only now, having left his banks, can he be spoken to about great things: the frog is limited by its well, the summer insect by its season. The sea exists all along; what limits the frog is its place.

This is the distinction the paper makes precise. The sea is the set of unmarked valid solutions, which is exponentially large (Theorem A.16). The well is the basin $\mathcal{B}_B(z^0)$, the trajectories an editor reaches from the released solution without exceeding its violation budget, and its sides are the energy walls

that the task's own constraints build (Lemma 4.3, Proposition 4.2). As in the parable, the limitation is one of place. Our editors may be computationally unbounded and may know the key, the detector and a target (Section 3); an editor whose basin contains no unmarked solution still cannot produce an unmarked output (Proposition 4.1). What decides erasure is not what the editor knows, but what it can reach.

The parable also has two limits, and both appear in the paper. First, the well has a finite depth. Once the budget exceeds the barrier distance, the basin contains an unmarked solution and erasure becomes possible; for a random key on a parity task this happens at the first jump ν_1 of the barrier norm except with probability at most $2^{-\ell}$ (Corollary 4.6), and Table 1 locates it exactly on twenty instances. Second, an editor that discards the released trajectory and regenerates a solution is not in the well at all; regeneration lies outside the local-editing model (Section 6).