MAXIMIN-SUFFICIENT LEARNING UNDER STRATEGIC FEEDBACK IN ZERO-SUM GAMES WITH AN INFORMED OPPONENT

Anonymous authors
Paper under double-blind review

ABSTRACT

A defender can test a policy in resettable diagnostic episodes before deployment, but an informed opponent may choose, among its best responses, those that hide the unknown game model. Identifying the model may then be impossible, and it is stronger than needed: the defender only needs one policy that is ε-maximin for every model the data cannot rule out. We ask when this weaker target can be certified. Because best responses can tie, each diagnostic query induces a set of transcript laws, from which the opponent selects adaptively. For a learner that sees only the recorded transcripts, we show that certification is impossible whenever a group of models that no single policy serves within tolerance shares a common transcript law at every accessible query. For finite models, queries and transcripts, with deployment and inference sharing one tolerance and small enough δ, this joint obstruction is the only one: a confidence sequence that tests empirical transcript frequencies against each model's set of laws returns an ε-maximin policy with probability at least $1 - \delta$ exactly when no such group exists. An acquisition score from minimax-regret duality (RDI) targets the deployment decision and affects only speed. On matrix, patrolling, search and poker designs, identifying the model takes 1 to 19 times as many episodes as reaching the policy certificate on the same runs. On the patrolling and search designs, RDI query selection is about three times faster than uniform querying, mostly through an allocation that could be fixed before any data; the exact RDI score is at most 1.31 times faster than a cheaper pairwise score there, and six times faster on an interdiction design built so that pairwise evidence vanishes. The certificate is conditional on the candidate library: in a misspecification study whose library misses the true game, the method certified every run, and 92% of the runs had a gap above the tolerance.

1 INTRODUCTION

A defender calibrating a policy against an unknown adversarial environment faces two uncertainties: which game is played, and which best response an informed attacker picks. The second can hide the first. A diagnostic policy may reveal the model under one tie-break and nothing under another, equally optimal one, so fitting a single likelihood cannot certify deployment against the whole response class.

Identifying the model is sufficient for deployment but not necessary. Deployment needs only that every group of models that the opponent can keep indistinguishable admits a common ε-maximin policy. Indistinguishable models are harmless when they share such a policy and fatal when they do not.

We model calibration as resettable diagnostic episodes. A query specifies a reset law, an announced defender policy and a recorded transcript; the attacker knows the model, best-responds to that episode's policy and may condition on history, so each model and query induce a set $\mathcal{P}_\theta(q)$ of transcript laws. For a plausible set C the target is $\Gamma(C) = \min_{x \in X} \sup_{\theta \in C} \mathrm{Gap}_\theta(x)$, with X the defender realization-plan polytope and $\mathrm{Gap}_\theta(x)$ the loss relative to the model's maximin value, so deployment can be certified while several models remain plausible.

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.

For a learner that keeps only the transcripts, certification is impossible whenever a group of models that no single policy serves within tolerance has a common transcript law at every accessible query (Theorem 4.3). In the finite exact branch (finite models, queries and transcripts; one tolerance for deployment and inference; small δ) the absence of such a group is also sufficient, attained by a frequency confidence sequence (Theorem 4.4). Acquisition affects only speed. RDI, a score of decision loss from minimax-regret duality, has one guarantee: a lower bound on the information every admissible selector must release (Proposition 5.1). Empirically, identification takes 1 to 19 times the episodes of certification; RDI beats uniform querying mostly through a static allocation, and on two-sided designs exact RDI is at most 1.31 times faster than a pairwise score except where pairwise evidence vanishes.

Our contributions are three. (i) A maximin-sufficient formulation of calibration under strategic set-valued feedback, with a deployment certificate valid for every acquisition rule (Theorem 4.1), and a characterization of when certification is possible: an obstruction by a jointly common transcript law (Theorem 4.3), which is the only obstruction in the finite exact branch (Theorem 4.4). (ii) An anytime-valid confidence sequence for set-valued response channels that tests empirical transcript frequencies against each model's channel polytope, needs neither common support nor pairwise separation, and attains that characterization (Theorem 4.4). (iii) An acquisition score from minimax-regret duality that targets the deployment decision (Theorem 4.2, Proposition 5.1), and experiments that measure what stopping at sufficiency saves relative to identification and how much acquisition adds (Section 6).

2 Related Work

With a known model, one-sided partially observable stochastic games are solved by value iteration and heuristic search (Horák et al., 2017; 2023); our problem precedes planning: the model is uncertain and the feedback is strategic. Our target is sufficient for the decision, as in fixed-confidence and task-optimal exploration (Garivier & Kaufmann, 2016; Kaufmann et al., 2016; Degenne & Koolen, 2019; Wagenmaker et al., 2021; Wagenmaker & Jamieson, 2022), decision-focused learning (Elmachtoub & Grigas, 2022; Donti et al., 2017; Wilder et al., 2019), reward-free exploration (Jin et al., 2020) and best-policy identification (Ménard et al., 2021; Al Marjani & Proutière, 2021). $\Gamma(C)$ is the minimax-regret criterion used by Braziunas & Boutilier (2007) with Helly-type bounds (Campi & Garatti, 2008); RDI is narrower than Bayesian experimental design and Blackwell comparison (Pukelsheim, 2006; Chaloner & Verdinelli, 1995; Foster et al., 2019; Blackwell, 1953). The four closest lines are equilibrium learning from payoff queries (Zhou et al., 2017; Maiti et al., 2024; Poiani et al., 2026), commitment learning from follower responses (Letchford et al., 2009; Blum et al., 2014; Peng et al., 2019; Bacchiocchi et al., 2025), fitting games to observed play (Waugh et al., 2011; Ling et al., 2018) and security games and nonmyopic followers (Kiekintveld et al., 2009; Amin et al., 2013; Haghtalab et al., 2022). Haghtalab et al. (2022) already let the agent play any element of an approximate best-response set and break ties favorably only in the regret benchmark, and Poiani et al. (2026) already handle infinitely many correct answers, including approximate Nash equilibria from noisy payoff queries. Our setting differs in the feedback: a query returns one symbol of a partial log, drawn from a set $\mathcal{P}_\theta(q)$ of laws, and an informed opponent selects from this set adaptively. Theorem 4.3 gives a condition under which this feedback admits no uniformly correct learner, and our sufficiency result allows adaptive selection whenever no joint common law exists. Informed or deceptive followers can block, slow or mislead an uninformed leader in general-sum play (Ananthakrishnan et al., 2024; Zhao et al., 2023; Birmpas et al., 2021); those results concern regret, equilibrium payoffs or induced commitments and issue no deployment certificate. Our confidence sequences use time-uniform inference, safe and convex composite testing, and log-optimal numeraire e-variables (Abbasi-Yadkori et al., 2011; Howard et al., 2021; Ramdas et al., 2023; Grünwald et al., 2024; Goldenshluger et al., 2015; Koolen & Grünwald, 2022; Larsson et al., 2025; Wasserman et al., 2020) on set-identified (Chernozhukov et al., 2007) channels that depend on the selector, so validity must hold conditionally. Robust MDPs and double oracles (Iyengar, 2005; Nilim & El Ghaoui, 2005; McMahan et al., 2003; Lanctot et al., 2017) optimize against a given set; we certify over the current confidence set. Table 2 and Appendices A.7 and A.1 give details.

3 PROBLEM SETTING

Game family and objective. Let Θ be a finite or compact model class, each θ specifying a two-player zero-sum game $G(\theta)$ with horizon H in which the defender maximizes total reward and the attacker minimizes it. The attacker observes the model and, in the one-sided formulation, the physical state; the defender acts from its own information history. For $\mu \in \Pi_D$, $\nu \in \Pi_A$ write $V_\theta^{\mu,\nu} = \mathbb{E}_\theta^{\mu,\nu}[\sum_{h=1}^{H} r_{\theta,h}]$, $g_\theta(\mu) = \inf_\nu V_\theta^{\mu,\nu}$, $v_\theta^\star = \sup_\mu g_\theta(\mu)$ and $\mathrm{Gap}_\theta(\mu) = v_\theta^\star - g_\theta(\mu)$; the last is the deployment loss. Finite extensive-form statements use perfect recall and common realization-plan polytopes X, Y; all statements about convexity, Helly bounds, minimax and the robust radius use realization coordinates $x \in X$, with chance probabilities absorbed into A_θ. Appendix F.6 gives the catalog conditions: a common tree, negative payoffs with a normalizing shift common to all models (Appendix B), and chance probabilities that are zero in some models only.

For a nonempty plausible set $C \subseteq \Theta$, $\Gamma(C) = \min_{x \in X} \sup_{\theta \in C} \mathrm{Gap}_\theta(x)$ is the smallest deployment gap attainable by one policy that must work for every plausible model; the guarantee is one-sided maximin and certifies no Nash pair. $\Gamma(C)$ can be small while the parameter diameter is large, so a raw diameter certifies nothing about deployment without a calibrated sensitivity bound (Appendix A.2). Stopping therefore uses a certificate: output a policy only when a certified upper bound U_n on $\Gamma(C_n)$ satisfies $U_n + 2\varepsilon_{\mathrm{span}} \le \varepsilon$, with $\varepsilon_{\mathrm{span}}$ bounding the value distance between the deployed game $G^\dagger$ and the reference $G(\theta^\star)$. The worst case over C needs no prior: it suits a fixed unknown model that the attacker knows.

Diagnostic queries and strategic feedback. Before deployment the defender may run resettable diagnostic episodes. A query $q \in \mathcal{Q}_{\mathrm{acc}}$ specifies a reset law, an announced calibration policy μ^q sampled during the episode in place of the current robust deployment candidate, and a transcript map producing $Z \in \mathcal{Z}_q$, possibly using audit information revealed at the end of the episode (realized actions, chance events, rewards) that need not be visible at deployment. Tier I assumes a fixed finite pool; Tier II adds priced query columns. The attacker may choose any best response $\mathrm{BR}_\theta(q) = \arg\min_\nu V_\theta^{q,\nu}$, so θ induces the channel set $\mathcal{P}_\theta(q) = \{\mathcal{L}_\theta^{q,\nu}(Z) : \nu \in \mathrm{BR}_\theta(q)\}$, randomized best responses included. An admissible selector may depend on the full history and pick a law in this set after seeing the announced query. All confidence and impossibility claims quantify over these selectors, since no algorithm can distinguish models that share a law at every accessible query.

The behavioral assumption is myopia at the episode level: the attacker may choose the least informative current best response, but may not sacrifice current diagnostic payoff to manipulate future learning. Myopia is an assumption about a controller re-optimized in each episode; it does not follow from reset access. Announcement reveals the defender's mixed policy but not its realized seed. An adversarial evaluation run scenario by scenario satisfies reset, announcement and myopia jointly, and none of the three implies the others (Appendices A.2 and A.6).

Running example: a censored sentinel. Two models are 2×2 matrix games in which the defender plays row one with probability x: $A_0 = \mathrm{diag}(1,1)$ and $A_1 = \mathrm{diag}(3,1)$, with maximin points $x = 1/2$ and $x = 1/4$. The best common policy $x = 3/8$ has gap $1/8$ in both, so $\Gamma(\{0,1\}) = 1/8$. A query q_x announces x and records the realized actions and reward. At $q_{1/2}$ model 0's attacker is indifferent and may play column two, model 1's unique best response, so both models produce the same transcript law. At $q_{3/8}$ the unique best responses differ and one episode identifies the model. Whether the censoring matters depends on the tolerance. For $\varepsilon \ge 1/8$ the policy $x = 3/8$ is certified with no data, although $q_{1/2}$ can never identify the model; for $\varepsilon < 1/8$ no single policy serves both models, and certification needs a separating query such as $q_{3/8}$ (Proposition B.29, Appendix A.2).

What the learner may condition on, and normal stopping. The obstruction results quantify over all the learner's information, so we fix the filtration. It is generated by the query draws and the recorded transcripts $Z_1, \dots, Z_n$ and nothing else. The defender thus keeps neither its realized trajectory nor an execution seed that reconstructs it; even a θ-independent seed carries information, because the response couples it with Z_t. A learner that keeps either has a finer filtration. Under it the reported times remain upper bounds, but Theorem 4.3 must be re-checked (Appendix A.2). After n episodes the learner holds $D_n = \{(q_t, Z_t)\}_{t=1}^n$ and a confidence set C_n. Validity means $\mathbb{P}_{\theta^\star}\{\theta^\star \in C_n \ \forall n \ge 0\} \ge 1 - \delta$ under every admissible selector. An empty C_n is a failure state and certifies nothing; normal stopping outputs a policy only after the robust certificate passes.

4 CERTIFICATION AND LEARNABILITY

Oracle programs are in Appendix B, proofs in Appendices C, D and E, qualifications in Appendix A.5.

4.1 DEPLOYMENT SAFETY AND ITS OPERATIONAL BUDGET

The reference model $\theta^\star$ must describe the actual diagnostic laws. With d_V the supremum value difference over both players' policies, the three conditions are a span bound on the deployed game, uniform coverage of $\theta^\star$, and a certified robust solve at every nonempty C_n:

$$d_V(G^\dagger, G(\theta^\star)) \le \varepsilon_{\mathrm{span}}, \tag{4.1}$$

$$\mathbb{P}\{\theta^\star \in C_n \ \forall n \ge 0\} \ge 1-\delta, \tag{4.2}$$

$$\sup_{\theta\in C_n} \mathrm{Gap}_\theta(\mu_n) \le U_n \le \Gamma(C_n) + \eta_{\mathrm{comp}}, \quad \eta_{\mathrm{comp}} \ge 0. \tag{4.3}$$

Theorem 4.1 (Deployment safety from coverage and a robust bound)**.** *Assume* (4.1)*,* (4.2)*, and the first inequality in* (4.3)*. On one event of probability at least* $1-\delta$*, simultaneously at every round with* $C_n \neq \varnothing$*,* $\mathrm{Gap}_{G^\dagger}(\mu_n) \le U_n + 2\varepsilon_{\mathrm{span}}$*. Let* τ *be the normal stopping time at which a nonempty-set certificate first satisfies* $U_\tau + 2\varepsilon_{\mathrm{span}} \le \varepsilon$*, and* $\tau = \infty$ *otherwise. Then*

$$\mathbb{P}\{\tau < \infty,\ \mathrm{Gap}_{G^\dagger}(\mu_\tau) > \varepsilon\} \le \delta. \tag{4.4}$$

This holds for every adaptive acquisition rule. If $\tau < \infty$ *almost surely, the returned policy is* ε*-maximin with probability at least* $1-\delta$*.*

Given the span budget, the safety of a returned policy rests on two components only, valid coverage and a certified robust bound; the acquisition rule affects only when the procedure stops. The span budget must come from an external audit that bounds the value distance uniformly over policies, and if $2\varepsilon_{\mathrm{span}} + \eta_{\mathrm{comp}} \ge \varepsilon$ the theorem has no positive accuracy budget. In our experiments the span budget is zero, because calibration and deployment share one specified game (Appendix A.5). Coverage (4.2) presumes $\theta^\star \in \Theta$, so the certificate is conditional on the candidate library: it has no finite-sample power against a library that misses the true game (Section 6).

4.2 GEOMETRY OF THE ROBUST RADIUS

In realization coordinates $g_\theta(x) = \min_{y\in Y} x^\top A_\theta y$ and $p = \dim(\mathrm{aff}\, X)$; assume X, Y compact, the payoff matrix continuous on compact C, and put $u(\pi) = \max_x \mathbb{E}_\pi g_\theta(x)$ for a prior π on C.

Theorem 4.2 (Minimax-regret duality and a finite witness)**.** *Under the common sequence-form and continuity assumptions,*

$$\Gamma(C) = \max_{\pi\in\Delta(C)} \{\mathbb{E}_\pi[v_\theta^\star] - u(\pi)\}. \tag{4.5}$$

A maximizing least-favorable prior exists, and there are $W \subseteq C$ *with* $1 \le |W| \le p+1$ *and a prior* π_W *on* W *such that* $\Gamma(W) = \Gamma(C) = \min_{x\in X} \sum_{\theta\in W} \pi_W(\theta)\, \mathrm{Gap}_\theta(x)$*. For finite* C *the support of any least-favorable prior has robust radius* $\Gamma(C)$ *and admits such a reduction.*

This specializes classical minimax decision duality and Helly's theorem; the bound $p+1$ improves to $r+1$ when every loss factors through an r-dimensional affine image of X (Corollary D.5). With the true model added as an anchor, the same bound keeps the conflict groups of Section 4.4 small, at most $p+2$ models, and the witness is the decision problem on which Section 5 scores queries.

4.3 IMPOSSIBILITY UNDER STRATEGIC CENSORING

Write $\mathcal{U}_a(\theta) = \{x \in X : \mathrm{Gap}_\theta(x) \le a\}$. A finite family Ξ is jointly strategically equivalent when $\bigcap_{\theta\in\Xi} \mathcal{P}_\theta(q) \neq \varnothing$ for every accessible query (with a measurable selection of the common law for infinite query classes). This joint condition is stronger than pairwise intersection.

Theorem 4.3 (Joint worst-selector obstruction)**.** *Suppose that* Ξ *is jointly strategically equivalent and that* $\bigcap_{\theta\in\Xi} \mathcal{U}_a(\theta)$ *is empty. Then there are admissible selectors, one under each model of* Ξ*, inducing a single common law for the entire adaptive calibration experiment, including termination and any returned policy. Write* $k := |\Xi|$ *and* $m := \max_{x\in X} \sum_{\theta\in\Xi} \mathbf{1}\{x \in \mathcal{U}_a(\theta)\} \le k-1$*.*

Every algorithm has some $\theta \in \Xi$ for which $\mathbb{P}_\theta\{\widehat{\mu} \in \mathcal{U}_a(\theta)\} \le m/k$ for any finite policy output $\widehat{\mu}$. In particular, a conflicting pair prevents success probability strictly above $1/2$ in both models. If a confidence sequence has coverage $1-\delta$ under each model and selector, and a valid robust certificate permits normal output only when $U_n \le a$, then under these common-law selectors

$$\mathbb{P}(\tau < \infty) \le k\delta, \tag{4.6}$$

informative only when $k\delta < 1$. For the span-adjusted rule one may take $a = \varepsilon - 2\varepsilon_{\mathrm{span}} \ge 0$.

If the opponent can make a group of models that no single policy serves look identical at every query the defender may run, then no amount of transcript data and no acquisition rule yields a valid certificate except with probability at most $k\delta$; the defender must then enlarge the query pool, record more of each episode, or relax the tolerance. In the running example with pool $\{q_{1/2}\}$ and $\varepsilon < 1/8$, $\Xi = \{0, 1\}$ is jointly strategically equivalent and conflicting, so no learner succeeds with probability above $1/2$ in both models (Corollary B.30); adding $q_{3/8}$ removes the obstruction, because the two models force different columns there.

4.4 Confidence sets and a matched characterization

Assume finite models, queries and alphabets with compact convex channels of common positive support. For each ordered pair take a minimizer of $\mathrm{KL}(P\|K)$ over $P \in \mathcal{P}_{\theta'}(q)$, $K \in \mathcal{P}_\theta(q)$, set $e_q^{\theta,\theta'} = P^*/K^*$ with the first superscript the tested null, and mix them uniformly over alternatives. This is the confidence sequence C_n^e of (B.19)–(B.21). Its factors are conditionally e-valid, its worst-case log-growth against a null θ is $J_q^{\mathrm{wc}}(\theta^\star, \theta) = \inf_{P \in \mathcal{P}_{\theta^\star}(q)} \mathbb{E}_P[\log e_q^{\theta,\theta^\star}]$, and it covers $\theta^\star$ at all times with probability $1-\delta$ under every adaptive query rule and admissible selector (Theorem B.16).

Pairwise validity does not imply completeness: three models can have channels that intersect pairwise and no joint common law, and the example of Remark B.17 has radius $2/3$ while every pairwise factor equals one. Composite event tests resolve that example (Proposition D.4) and support elimination bounds stopping without common support (Proposition D.3), but the bets are chosen by hand for each instance. Appendix E.2 replaces them with one test that is valid and complete for arbitrary finite channel polytopes. Let $\widehat{P}_{q,N}$ be the empirical transcript frequency over the N rounds that use q, and take the sup-norm Azuma boundary (E.3) $\beta_N = \big(\frac{1}{2N}\log[2N(N+1)|\mathcal{Q}_{\mathrm{acc}}|SM/\delta]\big)^{1/2}$, $S = \max_q |\mathcal{Z}_q|$, $M = |\Theta|$. The test eliminates θ as soon as $\mathrm{dist}_\infty(\widehat{P}_{q,N}, \mathcal{P}_\theta(q)) > \beta_N$ for some q; each check is one linear program over the channel's vertices. With $\varepsilon_0 = \varepsilon - 2\varepsilon_{\mathrm{span}} - \eta_{\mathrm{comp}} > 0$ the budget left after the span and solver charges, call Ξ an *anchored conflict* when $\theta^\star \in \Xi$, $2 \le |\Xi| \le p+2$ (Helly) and $\Gamma(\Xi) > \varepsilon_0$, so that no single policy serves Ξ, and let $\mathfrak{C}_{\varepsilon_0}(\theta^\star)$ be the finite collection of these (Appendix B.4). Conflict separation and its worst case are

$$\varsigma(\Xi; \theta^\star) = \max_q \min_{P \in \mathcal{P}_{\theta^\star}(q)} \max_{\theta \in \Xi \setminus \{\theta^\star\}} \mathrm{dist}_\infty\big(P, \mathcal{P}_\theta(q)\big), \quad \varsigma_0(\theta^\star) = \min_{\Xi \in \mathfrak{C}_{\varepsilon_0}(\theta^\star)} \varsigma(\Xi; \theta^\star).$$

Forced exploration draws each query from $(1-\gamma)\lambda_n + \gamma\lambda_0$, with λ_n any acquisition allocation and λ_0 of full support (Algorithm 1, Section 5). Here (ε, δ)-*correctness* is $\mathbb{P}_\theta\{\tau < \infty, \mathrm{Gap}_\theta(\widehat{x}) \le \varepsilon\} \ge 1-\delta$ under every model and admissible selector (B.17), for τ normal stopping with output $\widehat{x}$.

Theorem 4.4 (A matched condition in the exact branch)**.** *Assume finite models, queries and alphabets with channel polytopes given by an explicit finite representation; assume neither common support nor pairwise separation. The resulting confidence sets cover $\theta^\star$ at all times with probability at least $1-\delta$ under every adaptive query rule and admissible selector. Now restrict to the exact branch $\varepsilon_{\mathrm{span}} = \eta_{\mathrm{comp}} = 0$, so that the deployment and inference tolerances are the same ε, and fix $\delta < 1/(p+2)$. Then $\min_{\theta \in \Theta} \varsigma_0(\theta) > 0$ if and only if (ε, δ)-correctness on Θ is achievable, the converse being stated for the transcript-only filtration of Section 3; and when the left side holds, Algorithm 1 with forced exploration and $C_n = C_n^{\mathrm{freq}}$ is such an algorithm, stopping, under Assumption E.14, with probability at least $1-\delta-\delta'$ by $n^\dagger = O\big((N^\star(\varsigma_0) + \log(|\mathcal{Q}_{\mathrm{acc}}|/\delta'))/(\gamma\lambda_0^{\min})\big)$, $N^\star(\varsigma) = \min\{N : 2\beta_N < \varsigma\}$, with constants in* (E.5) *and δ' a second budget for the forced-exploration query counts. The forward direction is proved model by model; the converse is Theorem 4.3 applied to the offending anchored conflict set.*

In the finite exact setting the joint obstruction of Theorem 4.3 is therefore the only obstruction to (ε, δ)-correctness: whether a library and query pool admit a correct learner is decided by a condition

on small groups of conflicting models, and when that condition holds, the frequency test with forced exploration attains it without assuming common support or pairwise separation.

The equivalence has three limits. First, it holds only in the exact branch; with approximation error charged separately, $\varsigma_0 > 0$ remains sufficient but the two directions no longer combine, and our certified oracle is outside this branch by the certified η_{comp} reported in Section 6. Second, it characterizes correctness only, and almost sure termination can fail: Remark E.9 gives a positively separated instance with two models on which the confidence set empties by chance, so the single budget δ must cover both termination and accuracy. Third, the converse is proved for the transcript filtration; sufficiency holds for any finer filtration, and Proposition A.1 is a 2×2 instance on which the converse fails for a learner that retains its executed action (Appendix A.2 checks our designs). The constants are loose: the deadline $n^\dagger$ proves finite termination and exceeds the medians of Section 6 by factors of 6×10^2 to 4×10^5, and the condition $\delta < 1/(p+2)$, a worst case fixed a priori, weakens to $\delta < 1/\min\{M, p+1\}$ for the converse (Remark E.8).

5 Algorithm and Acquisition

RDI-EASI (robust decision-relevant information with equilibrium-aware safe inference) has three modules: a confidence sequence, a certified robust deployment solver and an acquisition rule. Only the first two are needed for the safety of a returned policy (Theorem 4.1); termination in Theorem 4.4 also uses the floor of forced exploration, and otherwise the acquisition rule is a heuristic that affects only speed. "Equilibrium-aware" refers to the observation model: the transcript law is selected from a best-response face.

At round n the robust solver returns a policy μ_n with a certified bound $U_n \geq \sup_{\theta\in C_n} \mathrm{Gap}_\theta(\mu_n)$, and the algorithm stops if $U_n + 2\varepsilon_{\mathrm{span}} \leq \varepsilon$. Otherwise a witness oracle finds a finite least-favorable $W_n \subseteq C_n$ and prior π_n exposing the current robust radius (Theorem 4.2). For such a witness (π, W) of C the parent impurity $\mathcal{R}(\pi; W) = \min_{\mu\in\Pi_D} \sum_{\theta\in W} \pi(\theta)\,\mathrm{Gap}_\theta(\mu)$ equals $\Gamma(C)$. The child impurity $\mathcal{R}^{\mathrm{wc}}_q(\pi; W) = \inf_{\psi:\mathcal{Z}_q\to X} \sup_{K_i\in\mathcal{P}_i(q)} \sum_i \pi_i \mathbb{E}_{K_i}[\mathrm{Gap}_i(\psi(Z))]$ lets the policy depend on the transcript of q but takes each model's worst admissible selector; min and max interchange on finite transcript polytopes, and the exact child LP is (B.7). RDI is the decrease $I(q) = \mathrm{RDI}_q(\pi; W) = \mathcal{R}(\pi; W) - \mathcal{R}^{\mathrm{wc}}_q(\pi; W)$, written $I_n(q)$ at (π_n, W_n). It evaluates a frozen least-favorable decision experiment; its Bayesian reading for singleton channels, and a counterexample in which $I(q) = 0.4$ leaves the radius at 0.5, are in Appendix D.4. In the running example, with witness $\{0, 1\}$, $I(q_{1/2}) = 0$ because the worst-case selector returns a common law, while $I(q_{3/8}) = \Gamma(\{0, 1\}) = 1/8$ because the transcript identifies the model, so RDI selects the separating query.

The stopping guarantee of forced exploration (Theorem D.1) holds for any predictable exploitation rule. The guarantee specific to RDI is the following.

Proposition 5.1 (RDI lower-bounds the information a query must release)**.** *Let $\mathcal{Z}_q$ be finite, $\pi \in \Delta(W)$, and $G_W = \max_{\theta\in W} \sup_x \mathrm{Gap}_\theta(x)$. Let $I^{\mathrm{wc}}_\pi(q)$ be the smallest mutual information between a π-distributed model index and the transcript, minimized over admissible channel selections. Then $\mathrm{RDI}_q(\pi; W) \leq G_W\sqrt{I^{\mathrm{wc}}_\pi(q)/2}$, so $I^{\mathrm{wc}}_\pi(q) \geq 2\,\mathrm{RDI}_q(\pi; W)^2/G^2_W$. If moreover $\theta^\star \in W$, $G_W > 0$, $\pi(W \setminus \{\theta^\star\}) > 0$ and the selector realizes $K_{\theta^\star}$, then some $\theta \in W \setminus \{\theta^\star\}$ has $\mathrm{dist}_{\mathrm{TV}}(K_{\theta^\star}, \mathcal{P}_\theta(q)) \geq \mathrm{RDI}_q(\pi; W)/G_W$.*

By the second statement, a query with a high RDI score must accumulate evidence against some witness alternative however ties are broken. The converse fails, and the proposition does not cover a true model outside the witness (Appendix E.1).

The witness may contain decisive models with low prior weight, so RDI-EASI protects conflicting pairs $b = (\theta, \theta')$ by a floor on their squared Hellinger channel separation $D_q(b)$: the allocation over the pool maximizes average RDI subject to a κ_{sep} fraction of the pool's best max-min separation ((B.9) and (B.10) in Appendix B). PS-EASI uses pairwise safe log-growth acquisition with the same confidence sequence and solver and evaluates no child LP. The exact child LP runs only on the small designs, where it ties with PS-EASI on S4 and is faster elsewhere, on S5 by a factor of 1.20 (Section 6); we use it as an information benchmark and PS-EASI as the scalable rule. The witness oracle, tie handling and numerical tolerances are part of the specification (Appendix E.5).

Algorithm 1 Tier I RDI-EASI over a fixed finite query pool. Inputs: Θ, pool $\mathcal{Q}_{\mathrm{acc}}$, target ε, level δ, span bound $\varepsilon_{\mathrm{span}}$, separation fraction κ_{sep}, exploration probability $\gamma \in (0, 1]$, λ_0 with full support.

1: $D_0 \leftarrow \varnothing$, $C_0 \leftarrow \Theta$
2: **for** $n = 0, 1, 2, \ldots$ **do**
3: Build C_n from D_n by the confidence sequence; **return** statistical failure if $C_n = \varnothing$ *(every experiment instead runs Algorithm 1′, replacing an empty C_n by the model with the least evidence against it; this removes the failure exit for any module and restores a.s. stopping only for frequency-constrained ones; Prop. E.10)*
4: $(\mu_n, U_n) \leftarrow \textsc{CertifiedRobustSolve}(C_n)$; **return** $(\mu_n, U_n + 2\varepsilon_{\mathrm{span}})$ if $U_n + 2\varepsilon_{\mathrm{span}} \le \varepsilon$
5: Canonical witness (W_n, π_n) by the dual-face rule of Appendix E.5; protected pairs $\mathcal{B}_n \subseteq W_n \times W_n$ (the pairs (θ, θ') with $\mathcal{U}_a(\theta) \cap \mathcal{U}_a(\theta') = \varnothing$, $a = \varepsilon - 2\varepsilon_{\mathrm{span}}$)
6: *(optional; disabled in all experiments, and by Assumption E.14(i))* **return** obstruction certificate if a full-universe common-law test certifies one
7: Evaluate $I_n(q)$ and $D_q(b)$ on $\mathcal{Q}_{\mathrm{acc}}$; solve the RDI allocation over the pool for λ_n
8: Draw $q_{n+1} \sim (1-\gamma)\lambda_n + \gamma\lambda_0$, execute it, observe Z_{n+1}, set $D_{n+1} \leftarrow D_n \cup \{(q_{n+1}, Z_{n+1})\}$
9: **end for**

Tier II (Algorithm 2, Appendix A.4) replaces the fixed pool by column generation, separation pricing followed by RDI pricing, with certified residuals and finite termination (Propositions B.5 and B.6). Only separation pricing is certified, under full logging (Proposition B.8); RDI pricing needs a global optimization oracle, since the child impurity is not concave in the diagnostic plan, and Tier II results use the separation family. Like an inexact acquisition LP, inexact query generation can slow the procedure and leaves the certificate valid (Theorem 4.1), so only the statistical and deployment programs are certified numerically (Remark E.13).

Finite stopping under forced exploration. Forced exploration is part of the algorithm for finite channels and acts also when all one-step scores vanish. Under the finite conditions with common support, the span and solver bounds and the pairwise construction, it makes every anchored conflict at ε_0 lose an alternative by an explicit deadline $n^\star$ from the worst-case drift $\gamma \sum_q \lambda_0(q) J_q^{\mathrm{wc}}(\theta^\star, \theta)$ and an Azuma boundary, so $\mathbb{P}\{\tau \le n^\star, \mathrm{Gap}_{G^\dagger}(\mu_\tau) \le \varepsilon\} \ge 1 - \delta - \delta'$ under every admissible selector (Theorem D.1). Its $\delta \downarrow 0$ scaling, normalized factors, singular channels and the floor's growth are in Remark E.19, Corollary D.2 and Appendix A.5.

Scope and limitations. Only the branch with finite models, queries and transcripts has a complete set of assumptions, and RDI guarantees neither contraction of the confidence radius nor instance optimality (Appendix E.4). The lower bound on the expected sample size (Theorem B.12) is loose: it needs $\delta < 1/(p+2)$, loses a factor $p+1$, and even at the exit time differs from the upper bounds in selector class, calibration and cost functional (Remark E.19). Appendix A.3 lists each main result with its hypotheses and the scope restriction that is easiest to overlook.

6 Experiments

We study four questions. **(Q1)** Does stopping at a maximin-sufficient set rather than at an identified model save episodes? **(Q2)** Does acquisition driven by the witness help, and does the exact RDI score help more than a cheap pairwise score, an allocation frozen at the start, or the best fixed query? **(Q3)** Which inference module is needed, and what does each miss? **(Q4)** Where does the usable information come from? The results are in Figure 1 here and in Figures 3, 4, 5 and 6 of Appendix F, which gives every number, ablation and convention.

Designs and protocol. S1–S3 are exact instances with three or four models whose deployment payoff has identical attacker columns, so only calibration is strategic. S4 (patrolling), S5 (search, unknown detector degradation) and S6 (interdiction) are two-sided at deployment and small enough for the exact child LP (B.7) at every witness; S1–S3 also run full RDI. S5 differs from S4 by design: S4's two maximal compatible sets (sets with $\Gamma \le \varepsilon$) are disjoint while S5's overlap, and eight of its nineteen queries are structurally blind. S6, built for Q3, is pairwise-blind at every query (Appendix F.3). D5 (sequence-form patrolling; its model 3 has $\mathrm{Gap}_3 \equiv 0$ and so constrains no

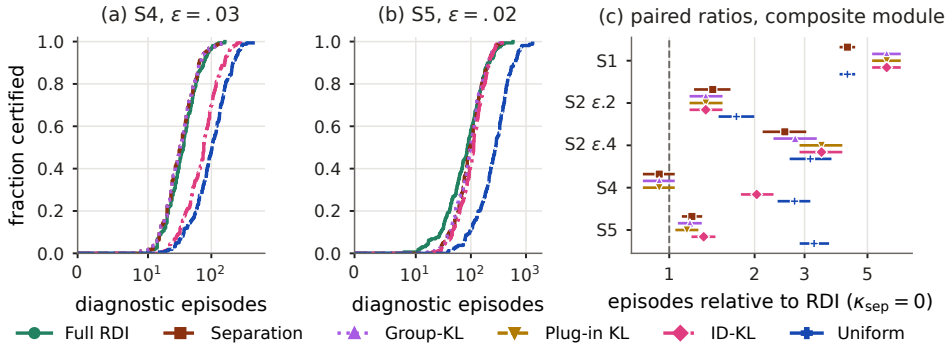

Figure 1: Q2. (a,b) CDFs of the normal certification time on S4 and S5, composite module, 160 runs per curve. (c) Paired ratios against full RDI with the floor released; right of the dashed line means RDI is faster.

deployment) and D6 (poker with a biased deck) are too large for the exact child LP. Every method on finite channels uses $\delta = .05$, forced exploration with $\gamma = .1$ and the same certified oracle, with $\eta_{\text{comp}} \leq 1.1{\times}10^{-14}$ on S1–S6 and $1.1{\times}10^{-13}$ with D5/D6. The runs use Algorithm $1'$, whose confidence set never empties; Table 27 reports Algorithm 1 itself. All times are under the greedy evidence-minimizing selector, a one-step heuristic that need not attain the worst case of the theory; it reads the learner's evidence, so it must be held fixed when modules are compared. We compare uniform acquisition, ID-KL (aimed at identification), PS-EASI ("Separation") and full RDI with static baselines; all comparators are our own, because published fixed-confidence algorithms need a single likelihood per query (Appendix G.5). A *median* inverts the certification CDF over all runs of a cell; a *paired ratio* is a log-mean over matched runs. All comparisons are exploratory and uncorrected for multiplicity; the S5 comparison at $\varepsilon = .02$ with 160 runs, used as the primary one, was chosen after the null result on S4, without pre-registration. Seed counts, background mixing rates η and fallback counts are in Appendix F.

Q1: savings from maximin sufficiency. Each run records both times on the same trajectory and continues past the sufficient certificate, so the comparison isolates the stopping target. Across designs and rules the ratio ranges from 1.00 to 19.2 (Figure 3). On S4, pairwise separation certifies at a median of 32 episodes and needs 624 to isolate one model, a paired ratio of 19.2 $[16.8, 21.8]$. The pair it stops separating shares an ε-maximin policy, so 19.2 is the cost of running a rule that ignores this pair until it identifies the model. Full RDI on S4 gives 2.87, ID-KL 1.35, and on D5 every rule has ratio one.

Q2: the value of acquisition. Full RDI is 2.77 $[2.44, 3.11]$ times faster than uniform querying on S4 and 3.24 $[2.92, 3.62]$ times on S5 (Figure 1), and an allocation fixed before any data achieves most of this gain. Table 26 decomposes it (in a separate study with 20 seeds) against static comparators that use the same library and no knowledge of the truth. On S5 the median falls as $285 \to 154 \to 92 \to 78$: $1.85\times$ from dropping the eight structurally blind queries, $1.67\times$ from freezing the initial RDI allocation, and $1.18\times$ from adaptivity. These are ratios of medians without a paired test; the only paired ratio, uniform querying against full RDI, is 3.34 $[2.93, 3.79]$. On S4, which has no blind query, the frozen allocation equals the adaptive one (34) and the single query chosen a priori matches it (34), so adaptivity adds nothing measurable there. ID-KL costs twice as much as full RDI on S4 and a third more on S5. Against the pairwise score, the exact child LP is indistinguishable on S4, 0.92 $[0.82, 1.03]$, although it costs orders of magnitude more per distinct confidence set. It is faster on S5 (1.20 $[1.12, 1.29]$ at the primary endpoint; 1.31 and 1.19 on two disjoint fresh seed roots, all excluding one; covering 1 at $\varepsilon = .03$ and .04) and on S6 (6.00 $[5.64, 6.39]$, at a fixed composite module). On S1 and S2, where only calibration is strategic, it is faster by 1.42 to 4.25 (Table 21). On S5 and S6 the single query chosen a priori is the query that the ex post search returns. Counting compute per episode, exact RDI pays off on S5 only when an interaction costs at least 6 ms (119 ms with a cold cache), and never on S4 (Appendix G.7).

Q3: which inference module is needed. No module dominates (Figure 4). Adding composite event bets to the pairwise e-process saves two orders of magnitude on S1 (18 against 1,405) but costs 29% more on S4 and 16% more on S5. Since the selector reads each module's evidence, Table 16 also holds the response vertex constant, on S1's pool restricted to $\{q_0\}$. There the composite module and its intersection with the frequency test reach the same median, because in the intersection the frequency test eliminates nothing (its boundary exceeds 1 for the first four counts). They do not agree run by run: running each component at $\delta/2$ avoids the 2 and 1 truth eliminations that the composite module suffers. On that restricted pool, and on all seven queries of S6, every anchored conflict is *pairwise-blind* (the best certified pairwise log-growth is zero to machine precision) while the joint separation is positive. The pairwise e-process certifies on none of those paths, and the joint modules on all of them. S6 shows this with a deployment attacker that mixes at every model's maximin point, which S1 lacks. The frequency module contributes generality and a matched stopping condition; with full RDI its median is 0.83 to 1.44 times that of the composite module.

Limitation: a misspecified candidate library. No inference module detects it. We generate S4 and S5 data from laws outside every candidate polytope, at sup-norm distances .053 to .296 (perturbing the game and recomputing the attacker's best response). All 720 runs per cell certify under every module, and the frequency test fires in 2 of the 4,320 runs whose module includes it. The realized gap exceeds ε on every S4 run and in three of S5's four perturbed worlds, that is, on 7,920 of the 8,640 runs (91.7%; Figure 6(b)). The certificate guarantees deployment performance only when the candidate library contains the true game, and the span budget of Section 4.1 does not cover a library that misses it; adequacy must come from an external model audit.

Q4, validity and cost. A field ablation alone cannot separate chance from strategic information, since suppressing a field changes the observation law; Figure 5 therefore also varies the generating mechanism. In the study that restricts the transcript (80 pairs; Q2 uses 160), S4's paired ratio of uniform querying to the design's rule is 2.57 $[2.12, 3.14]$ with the full record and 6.14 $[5.21, 7.14]$ with only the two alarm bits, and on D5 keeping only the alarms raises the median from 16 to 2,030 episodes. On both designs, only the view with alarms alone has an interval disjoint from that of the full record. The D6 controls do not isolate a strategic channel: the benefit survives a fixed deal law and, smaller, a frozen responder, whose bound $(1-\eta)\max \mathrm{TV}(D_\theta, D_{\theta'}) = 3/35$ is attained exactly. Coverage is per model, so Figure 6(a) reports every cell separately: every rate lies in $[.000, .038]$, and 9 of the 42 intervals reach above .05. Appendix G.7 reports computational cost.

7 Conclusion

Strategic ambiguity matters only when it crosses a decision boundary: models that the opponent can keep indistinguishable at every available query are harmless if they share an ε-maximin policy and fatal if they do not. For finite models, queries and transcripts, with deployment and inference sharing one tolerance and small δ, this obstruction by a joint common law is the only barrier to (ε, δ)-correct certification, and a confidence sequence on transcript frequencies attains the resulting characterization. The characterization concerns correctness and does not imply almost sure termination, and its converse is stated for a learner that keeps only the recorded transcripts (Proposition A.1). Empirically, stopping at the policy certificate rather than at identification saves up to a factor of 19 on the same runs, and nothing on D5. On the patrolling and search designs, RDI query selection beats uniform querying about threefold, mostly through a static allocation, and on the two-sided designs exact RDI is at most 1.31 times faster than a cheap pairwise score except where pairwise evidence vanishes. We therefore view RDI as a criterion of decision loss with a certified lower bound on information, and the pairwise score as its scalable surrogate. The certificate is conditional on the candidate library: when the library misses the true game, every module still certified in our study. Four problems remain open: sharp complexity bounds for multiple correct answers; an instance on which RDI provably beats pairwise separation; a test of library adequacy with finite-sample power; and scalable pricing of the child impurity.

AI use statement

In this work, we used generative AI tools to implement methods (writing and refactoring the experimental code to our specification), to provide feedback on the research methodology and experiments

(including simulated reviews that led us to add baselines, ablations and consistency checks), and to help interpret results (drafting and revising the text that reports them). We also used them to edit the paper for readability and to produce figures from our records. We have not used generative AI tools to develop the theoretical model or the conceptual framework, to formulate the mathematical claims, or to propose or refine hypotheses. We used them in one proof: during final review, an AI-assisted critique found that the proof of Theorem E.6(2) invoked a concentration bound where only confidence-set coverage had been established, and the replacement step, a nearest-point argument, was drafted with AI assistance; the statement of the theorem is unchanged. No other proof or proof ingredient came from AI. Generating synthetic data sets, translation, cleaning or reformatting datasets, and qualitative or thematic data analysis are not applicable to this work: all data are simulated by our code from the specifications in Appendix F. We have reviewed all AI-assisted work. The authors wrote every other proof and checked every proof. The supplementary code regenerates, from the run records, every reported figure and every reported table except Table 15; a check suite tests numerical properties of the code and the records. We take responsibility for the final content of this work, including all text, claims, code and results.

Ethics statement

This work involves no human subjects, no personal data and no released dataset; all games, transcripts and opponents are synthetic and specified in Appendix F. The application framing is defensive: the object we certify is an upper bound on one defender's deployment gap against an informed adversary, and the two larger security designs (patrolling and search) are standard benchmarks and model no deployed system. The work has two dual-use aspects. First, the obstruction results (Theorem 4.3) also tell an adversary how to prevent certification: respond so that models that no single policy serves keep a common transcript law. We think publication is justified, because a defender who does not know which query families are obstructed cannot avoid them, while an adversary able to exploit the obstruction already knows the model by assumption. Second, and more serious, the misspecification study in Section 6 shows that the method returns a certificate on a candidate library that does not contain the true game, with a realized gap above the requested tolerance and no warning from any module. A certificate from this method therefore depends on an external audit of library adequacy and of the span budget, and should not be presented as an unconditional safety guarantee for any deployed system.

Reproducibility statement

The theoretical claims are stated with their assumption sets in Sections 4 and 5, with the exact oracle programs in Appendix B, complete proofs in Appendices C, D and E, and a scope table in Appendix E.4 that records, for each branch, which assumptions each result needs. Definitions and conventions that the main text states briefly are given in full in Appendix A. For the experiments, Appendix F specifies S1–S6 and D5's information structure down to their payoff matrices, channel polytopes, query pools and attacker response sets, with the full D5 and D6 definitions in Appendix G.3 and the D1/D2 bandit designs in Appendix G.4; Appendix F.4 defines each method name; Appendix F.8 lists the software environment, hardware and seed counts; Appendix F gives every number behind every figure, together with the reporting conventions used in the main text (medians that are observed episode counts, and paired bootstrap ratios); and Appendix F.7 states how numerical error is charged to the certificate. Source code and the run records behind every reported table and figure are provided in the anonymous supplementary material, together with a manifest of their SHA-256 hashes.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011. URL https://proceedings.neurips.cc/paper/2011/hash/e1d5be1c7f2f456670de3d53c7b54f4a-Abstract.html.

Aymen Al Marjani and Alexandre Proutière. Adaptive sampling for best policy identification in Markov decision processes. In *Proceedings of the 38th International Conference on Machine*

Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 7459–7468. PMLR, 2021. URL https://proceedings.mlr.press/v139/marjani21a.html.

Steve Alpern and Shmuel Gal. *The Theory of Search Games and Rendezvous*, volume 55 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, 2003. doi: 10.1007/b100809.

Kareem Amin, Afshin Rostamizadeh, and Umar Syed. Learning prices for repeated auctions with strategic buyers. In *Advances in Neural Information Processing Systems*, volume 26, pp. 1169–1177, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/996a7fa078cc36c46d02f9af3bef918b-Abstract.html.

Nivasini Ananthakrishnan, Nika Haghtalab, Chara Podimata, and Kunhe Yang. Is knowledge power? On the (im)possibility of learning from strategic interactions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 23852–23880, 2024. doi: 10.52202/079017-0752.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi: 10.1137/S0097539701398375.

Francesco Bacchiocchi, Matteo Bollini, Matteo Castiglioni, Alberto Marchesi, and Nicola Gatti. The sample complexity of Stackelberg games. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2053–2061. PMLR, 2025. URL https://proceedings.mlr.press/v258/bacchiocchi25a.html.

Arie Beresteanu, Ilya Molchanov, and Francesca Molinari. Sharp identification regions in models with convex moment predictions. *Econometrica*, 79(6):1785–1821, 2011. doi: 10.3982/ECTA8680.

Georgios Birmpas, Jiarui Gan, Alexandros Hollender, Francisco J. Marmolejo-Cossío, Ninad Rajgopal, and Alexandros A. Voudouris. Optimally deceiving a learning leader in Stackelberg games. *Journal of Artificial Intelligence Research*, 72:507–531, 2021. doi: 10.1613/jair.1.12542.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Learning optimal commitment to overcome insecurity. In *Advances in Neural Information Processing Systems*, volume 27, pp. 1826–1834, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/d3eea90e8e9cff58fc84a4c40d22d95a-Abstract.html.

Darius Braziunas and Craig Boutilier. Minimax regret based elicitation of generalized additive utilities. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 25–32, 2007.

Marco C. Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008. doi: 10.1137/07069821X.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/ss/1177009939.

Victor Chernozhukov, Han Hong, and Elie Tamer. Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284, 2007. doi: 10.1111/j.1468-0262.2007.00794.x.

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14564–14573, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/60cb558c40e4f18479664069d9642d5a-Abstract.html.

Priya L. Donti, Brandon Amos, and J. Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5484–5494, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3fc2c60b5782f641f76bcefc39fb2392-Abstract.html.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68 (1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

Adam Foster, Martin Jankowiak, Elias Bingham, Paul Horsfall, Yee Whye Teh, Thomas Rainforth, and Noah Goodman. Variational Bayesian optimal experimental design. In *Advances in Neural Information Processing Systems*, volume 32, pp. 14036–14047, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/d55cbf210f175f4a37916eafe6c04f0d-Abstract.html.

Alfred Galichon and Marc Henry. Set identification in models with multiple equilibria. *The Review of Economic Studies*, 78(4):1264–1298, 2011. doi: 10.1093/restud/rdr008.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027. PMLR, 2016. URL https://proceedings.mlr.press/v49/garivier16a.html.

Alexander Goldenshluger, Anatoli Juditsky, and Arkadi Nemirovski. Hypothesis testing by convex optimization. *Electronic Journal of Statistics*, 9(2):1645–1712, 2015. doi: 10.1214/15-EJS1054.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

Nika Haghtalab, Thodoris Lykouris, Sloan Nietert, and Alexander Wei. Learning in Stackelberg games with non-myopic agents. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 917–918, 2022. doi: 10.1145/3490486.3538308. Extended abstract; full version available as arXiv:2208.09407.

Karel Horák, Branislav Bošanský, and Michal Pěchouček. Heuristic search value iteration for one-sided partially observable stochastic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 558–564, 2017. doi: 10.1609/aaai.v31i1.10597.

Karel Horák, Branislav Bošanský, Vojtěch Kovařík, and Christopher Kiekintveld. Solving zero-sum one-sided partially observable stochastic games. *Artificial Intelligence*, 316:103838, 2023. doi: 10.1016/j.artint.2022.103838.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Garud N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2): 257–280, 2005. doi: 10.1287/moor.1040.0129.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4870–4879. PMLR, 2020. URL https://proceedings.mlr.press/v119/jin20d.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Christopher Kiekintveld, Manish Jain, Jason Tsai, James Pita, Fernando Ordóñez, and Milind Tambe. Computing optimal randomized resource allocations for massive security games. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems – Volume 1*, pp. 689–696, 2009. URL https://www.cs.utep.edu/kiekintveld/papers/2009/kjtpot-massive-security-games.htm.

Wouter M. Koolen and Peter Grünwald. Log-optimal anytime-valid E-values. *International Journal of Approximate Reasoning*, 141:69–82, 2022. doi: 10.1016/j.ijar.2021.09.010.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Perolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4190–4203, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3323fe11e9595c09af38fe67567a9394-Abstract.html.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *The Annals of Statistics*, 53(3):1015–1043, 2025. doi: 10.1214/24-AOS2487.

Joshua Letchford, Vincent Conitzer, and Kamesh Munagala. Learning and approximating the optimal strategy to commit to. In *Algorithmic Game Theory*, volume 5814 of *Lecture Notes in Computer Science*, pp. 250–262. Springer, 2009. doi: 10.1007/978-3-642-04645-2_23.

Chun Kai Ling, Fei Fang, and J. Zico Kolter. What game are we playing? End-to-end learning in normal and extensive form games. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 396–402, 2018. doi: 10.24963/ijcai.2018/55.

Arnab Maiti, Ross Boczar, Kevin Jamieson, and Lillian J. Ratliff. Near-optimal pure exploration in matrix games: A generalization of stochastic bandits & dueling bandits. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2602–2610. PMLR, 2024. URL https://proceedings.mlr.press/v238/maiti24a.html.

H. Brendan McMahan, Geoffrey J. Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 536–543, 2003. URL https://cdn.aaai.org/ICML/2003/ICML03-071.pdf.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7599–7608. PMLR, 2021. URL https://proceedings.mlr.press/v139/menard21a.html.

Arnab Nilim and Laurent El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005. doi: 10.1287/opre.1050.0216.

Binghui Peng, Weiran Shen, Pingzhong Tang, and Song Zuo. Learning optimal strategies to commit to. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2149–2156, 2019. doi: 10.1609/aaai.v33i01.33012149.

Riccardo Poiani, Martino Bernasconi, and Andrea Celli. Pure exploration with infinite answers. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2989–2997. PMLR, 2026. URL https://proceedings.mlr.press/v300/poiani26a.html.

Friedrich Pukelsheim. *Optimal Design of Experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2006. ISBN 978-0-89871-604-7. doi: 10.1137/1.9780898719109.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Andrew Wagenmaker and Kevin Jamieson. Instance-dependent near-optimal policy identification in linear MDPs via online experiment design. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5968–5981, 2022. doi: 10.52202/068431-0432.

Andrew J. Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-optimal exploration in linear dynamical systems. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10641–10652. PMLR, 2021. URL https://proceedings.mlr.press/v139/wagenmaker21a.html.

Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. doi: 10.1073/pnas.1922664117.

Kevin Waugh, Brian D. Ziebart, and J. Andrew Bagnell. Computational rationalization: The inverse equilibrium problem. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1169–1176, 2011. URL https://icml.cc/Conferences/2011/papers/599_icmlpaper.pdf.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Geng Zhao, Banghua Zhu, Jiantao Jiao, and Michael Jordan. Online learning in Stackelberg games with an omniscient follower. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42304–42316. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhao23o.html.

Yichi Zhou, Jialian Li, and Jun Zhu. Identify the Nash equilibrium in static games with random payoffs. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 4160–4169. PMLR, 2017. URL https://proceedings.mlr.press/v70/zhou17b.html.

A Deferred Material from the Main Text

This appendix gives the full versions of passages that the main text summarizes: a comparison with the lines of prior work closest to each of our components, supporting arguments for the problem setting, an index of results, the Tier II procedure for query generation, deferred qualifications and protocol details, and extended related work. Figure 2 shows where the three modules of Section 5 sit in the calibration loop, and which of them are needed for validity.

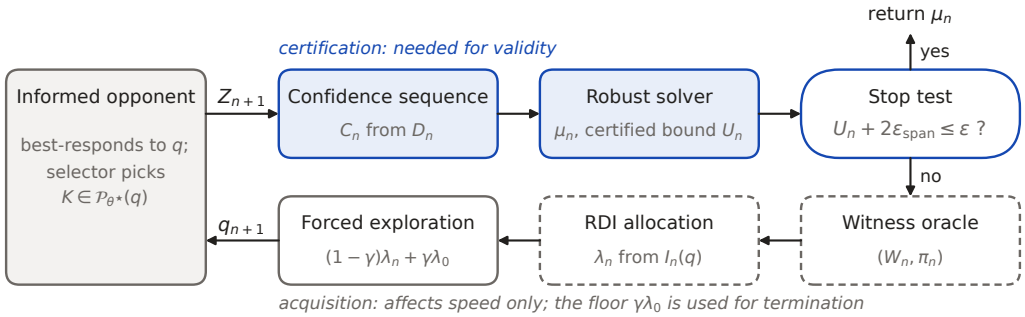

Figure 2: The certified calibration loop of Section 5 (Algorithm 1), read clockwise from the opponent. The certificate uses only the confidence sequence and the robust solver (top row), so a returned policy is safe under every acquisition rule (Theorem 4.1). Query acquisition (bottom row) scores queries by a decision experiment on the witness; its dashed steps affect only speed, and termination also uses the forced-exploration floor (Theorem 4.4). Forced exploration and the composite/support evidence of the confidence sequence address different causes of statistical stalling.

A.1 What is inherited and what is new

We compare each of our components with the closest line of prior work. *Multiple-answer pure exploration* (Degenne & Koolen, 2019) already optimizes over a set of correct answers and obtains matching asymptotics. We inherit the idea that the answer set defines the target. Their complexity result does not carry over: their experiment has a single law per arm, whereas each query here induces a set of laws chosen by an adaptive adversary, and our lower bound still loses a $p+1$ pigeonhole factor (Remark E.19). *Safe testing of convex composite nulls* (Grünwald et al., 2024) supplies the pairwise factor: for a convex pair of channel sets the likelihood ratio of the joint (reverse) information projection is the known growth-optimal e-variable, and Theorem B.16 is its direct application. Goldenshluger et al. (2015) construct the analogous fixed-sample test for convex hypotheses from the pair closest in Hellinger distance; our factor uses the pair closest in KL divergence. Beyond that application, pairwise validity is incomplete for these families (Remark B.17), and the completeness condition for the induced decision problem is the joint conflict separation of Theorem 4.4, which is necessary as well as sufficient in the finite exact branch with $\delta < 1/(p+2)$. *Log-optimal and numeraire e-variables* (Koolen & Grünwald, 2022; Larsson et al., 2025) characterize the optimal bet against a composite null. Our exact factor attains that optimum for one ordered pair and one step. The fixed mixture over pairs is not shown to be optimal for eliminating a union of models or for resolving a deployment decision, and Section 6 shows it is not uniformly best. *Learning commitment strategies from follower responses* (Letchford et al., 2009; Blum et al., 2014; Peng et al., 2019) shares the announcement protocol. Letchford et al. (2009) and Blum et al. (2014) assume a fixed tie-break that favors the leader, and Peng et al. (2019) let the learner choose among tied best responses; this statement concerns those three papers only. Bacchiocchi et al. (2025) drop Peng et al.'s tie-control and non-degeneracy assumptions but keep a tie-break that favors the leader, and manage the trade-off between sample complexity and termination probability that arises when leader strategies have finite precision. Haghtalab et al. (2022) let the agent play an arbitrary element of an ε-approximate best-response set and break ties in favor of the principal only in the regret benchmark, which is separate from the feedback model. None of these lines has a counterpart of the obstruction in Theorem 4.3, which is a condition on whether a set of laws admits a law common to a family of conflicting models. Theorem 4.3 does not imply that learning from strategic responses requires

a fixed tie-break: our sufficiency result permits fully adaptive selection whenever no common law exists. Our new contributions are therefore the joint common-law obstruction with its bound m/k on overlapping answers, the coupling of minimax-regret duality to selector-adversarial channels, the tight characterization by conflict separation, and the stopping rule that uses only the certificate and never reads the acquisition score. The remaining components are specializations of known tools.

What is left when every channel is a singleton. Only the impossibility and completeness results need set-valued channels to be non-trivial; the information bound is non-trivial also for singleton channels, where it is weaker. Suppose every $\mathcal{P}_\theta(q)$ is a single law $K_\theta(q)$. Then joint strategic equivalence of Ξ becomes "$K_\theta(q)$ is the same for all $\theta \in \Xi$ at every accessible query", which coincides with its pairwise version, so the distinction between joint and pairwise conditions disappears and Theorem 4.3 reduces to the classical change-of-measure impossibility for identically distributed experiments: the common law needs no construction, and the bound m/k follows by averaging over the k models. Likewise $\varsigma(\Xi) > 0$ becomes "some accessible query separates $\theta^\star$ in sup norm from at least one alternative of Ξ", the compatibility test becomes a sup-norm concentration test against one law (for which neither common support nor pairwise separation is an obstacle), and Theorem 4.4 reduces to the standard statement that (ε, δ)-correct identification of a sufficient answer is achievable exactly when every anchored conflict is distinguishable at some accessible query. Proposition 5.1 is the only result that does not reduce to a classical statement: $I^{\mathrm{wc}}_\pi(q)$ becomes the ordinary mutual information $I(\theta; Z)$ and the bound $\mathrm{RDI}_q \le G_W \sqrt{I(\theta; Z)/2}$ remains a Pinsker-type relation between the one-step reduction of decision loss and released information. It loses its robustness to the selector: the minimization over admissible selections and the "however the attacker breaks ties" clause of the second statement are both empty.

Published algorithms as baselines. Every comparator in Section 6 is our own construction; the observation model forces this. Published fixed-confidence and pure-exploration algorithms, including the multiple-answer algorithm of Degenne & Koolen (2019), the track-and-stop family and best-policy identification, compute their stopping statistic from a single likelihood per arm or query. Here a (model, query) pair carries a set of transcript laws chosen by an adaptive selector, so the likelihood those algorithms need does not exist, and any fixed choice of member changes the guarantee from "valid under every admissible selector" to "valid under that selector". Adapting one of them means replacing its stopping statistic, and the result is a new algorithm. We therefore compare against the interfaces that are defined in our setting: full identification, fixed queries, uniform querying and the pairwise separation score.

Simple static comparators, all implementable, are reported in Appendix G.5, and Section 6 reports that decomposition alongside the ratio against uniform querying. Three of the adaptive comparators adapt published rules: Plug-in KL (Proposition B.23) is track-and-stop's plug-in allocation (Garivier & Kaufmann, 2016) with the single likelihood of each arm replaced by the directed worst-case projection J^{wc}_q over the best-response face; Group-KL is the same rule whose alternative set is the models failed by the robust policy of the plug-in model's maximal compatible set, which is a multiple-answer target in the sense of Degenne & Koolen (2019) carried over to a set-valued channel; and ID-KL is the same construction taking all other active models as alternatives, so its acquisition aims at identification while its stopping target remains the sufficiency certificate (Table 4). We inherit none of their guarantees: the worst-case projection changes the stopping statistic, so the asymptotic optimality proved for those algorithms does not transfer. The one published algorithm we did run, EXP3 (Auer et al., 2002), optimizes cumulative regret, has no stopping rule, and is therefore reported as non-comparable.

A.2 Supporting arguments for the problem setting

The running example: a censored sentinel game. Take $A_0 = \mathrm{diag}(1, 1)$ and $A_1 = \mathrm{diag}(3, 1)$, whose defender maximin row-one probabilities are $1/2$ and $1/4$; one policy for both has minimum worst-case gap $1/8$, attained at $x = 3/8$, so $\Gamma(\{0, 1\}) = 1/8$. A diagnostic announces x and logs realized actions and reward. At $x = 1/2$ both types can choose column two, which removes the model information; this is the censoring shown by the query $q_{1/2}$ of Section 4.3. At $x = 3/8$ their unique best-response columns differ and one episode separates them, which is why adding $q_{3/8}$ removes the obstruction.

Small diagnostic KL does not bound decision distance. For losses $(x, 1-x)$ and diagnostic laws $\mathrm{Ber}(1/2+t)$ and $\mathrm{Ber}(1/2-t)$, the directed KL is $2t\log((1/2+t)/(1/2-t)) \to 0$, while the answer sets are disjoint at every $\varepsilon < 1/2$. A diagnostic reset state can supply these laws separately from the deployment state. A relation between diagnostic sensitivity and deployment loss is needed to convert statistical proximity to a gap bound.

A system in which reset, announcement and myopia hold jointly. The three assumptions hold jointly in an adversarial evaluation protocol run scenario by scenario, of the kind used for scheduled assessment of a controller for patrol, interdiction or intrusion response. Each scenario restores a fixed environment snapshot, which is the reset. The controller under evaluation is published to the adversarial team or solver before the scenario runs, because the assessment audits the worst case; that is the announcement. The adversarial side is re-optimized for each scenario against the published controller, is scored only on that scenario, and carries no state across scenarios; that is myopia. None of the three follows from the others: an audit could publish the controller without resetting, or reset without publishing. The combination ensures that the diagnostic channel at a query is exactly the best-response face, so the opponent's learning dynamics do not enter the calibration data. The assumptions fail whenever the same adversarial team is scored on cumulative damage across the campaign.

A learner that retains its own executed actions. A learner that retained its own executed actions would have a finer, possibly strictly finer, filtration; its e-factors and frequency tests remain valid, because they are evaluated against the conditional law given the announced query, so all reported certification times are upper bounds for that richer learner. What changes is the obstruction statement: Theorem 4.3 asks for a common law of the entire experiment, so it must be checked on the joint law of the retained information, and a common law for the recorded transcript alone does not suffice when the executed action is remembered. We state the obstruction for the discarding interface.

The two halves of Theorem 4.4 are therefore quantified over different learners: sufficiency holds for any filtration at least as fine as the transcript filtration, because the tests are evaluated against the conditional law given the announced query, whereas the converse is proved for the transcript filtration alone. Two facts describe this gap.

First, the general implication is false, as a two-action instance shows; we verify the instance in exact rational arithmetic.

Proposition A.1 (Transcript equivalence does not imply action-retaining equivalence)**.** *Let the defender be the row player and take*

$$A_0 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \qquad A_1 = \begin{pmatrix} 0 & 3 \\ 1 & 0 \end{pmatrix},$$

with a single accessible diagnostic that plays each row with probability $\frac{1}{2}$ and records only the realized reward, $Z = R$. Then

1. *the first column is the unique best response under both models, so the response face is a singleton and no tie-breaking freedom is used;*
2. *$Z \sim \mathrm{Ber}(\frac{1}{2})$ under both models, so the transcript channels are equal, not merely intersecting;*
3. *the joint laws of (executed row, Z) are supported on $\{(a_1, 1), (a_2, 0)\}$ and $\{(a_1, 0), (a_2, 1)\}$ respectively, which are disjoint, so a single observation identifies the model;*
4. *$v_0^\star = v_1^\star = \frac{3}{4}$ and $\Gamma(\{0,1\}) = \frac{1}{4}$, attained at the uniform defence, so $\{0,1\}$ is a decision conflict at every $\varepsilon < \frac{1}{4}$; and*
5. *both deployment games have fully mixed equilibria, with the attacker playing $(\frac{3}{4}, \frac{1}{4})$ in each, so the conflict does not rely on an inert deployment opponent.*

Hence joint strategic equivalence of the transcript channels does not imply joint strategic equivalence of the action-retaining experiment, and no general argument extends the converse half of

Theorem 4.4 to the action-retaining filtration. To extend it, one must re-verify the common-law condition on the joint channel of the retained information, which is a separate hypothesis. Proposition A.1 does not affect Theorems 4.3 and 4.4 themselves: both are stated for the transcript filtration, and the counterexample shows that this restriction is necessary.

The phrase "retaining the executed action" has two possible readings, so we fix one. It means that the learner appends the realized pure action (the tour, plan or probe it played) to the record. The announced mixed policy is public in this protocol and already in the transcript, so remembering it changes nothing.

Remark A.2 (A model-independent execution seed carries information). A randomization seed drawn independently of θ can still carry information. A learner that keeps its own randomization device but discards the realized action does not stay in the transcript filtration, as the instance of Proposition A.1 shows. Let $B \sim \mathrm{Ber}(\frac{1}{2})$, independent of θ, be the bit that selects the executed row of the uniform probe, and let R be the realized reward against the unique diagnostic best response, column 1. Then $R = 1 - B$ under $\theta = 0$ and $R = B$ under $\theta = 1$, so

$$\mathrm{supp}(B, R) = \{(0, 1), (1, 0)\} \text{ or } \{(0, 0), (1, 1)\},$$

disjoint across the two models, while B and R are each $\mathrm{Ber}(\frac{1}{2})$ under both. Under a uniform prior, $I(\theta; B) = I(\theta; R) = 0$ but $I(\theta; B, R) = 1$ bit: a learner that keeps the seed identifies the model in one episode, without storing the executed action explicitly. What matters is that the attacker's response couples the seed with the transcript; marginal independence of the seed from θ does not prevent this coupling. We check this in exact rational arithmetic.

Consequently the transcript-only converse covers exactly those learners that discard the executed action and every execution seed or random tape from which it could be reconstructed. Selection randomness used to draw the query is unrestricted, because it is generated before the transcript and is conditionally independent of θ (Appendix B). For a learner that retains execution randomness, the common-law hypothesis must be re-checked on the joint channel that includes it.

Second, on the instances reported here the gap is narrow but not void. S4 records the realized tour index, S5 the realized plan index, and every D5 logging view retains the realized tour, so for those the transcript determines the executed action and the two filtrations coincide. There are two exceptions. The ablation of Table 19 that keeps only the alarms has a record of four symbols without the tour index. S6 logs one of four symbols and does not record the executed probe: its five pure probes are determined by the query, but its two probe mixtures with equal weights are not, and since every symbol has positive probability under both component resolutions the record does not reveal which probe ran. For those two rows the transcript filtration is strictly coarser, so no conclusion from the full log transfers to them.

For S6 we prove that the structure survives the finer filtration. Every probe row of S6 has the same payoff row, so all rows in a mixture's support share the response face $\{\text{strike } j : j \neq \theta\}$. Write V_c for the joint law of (executed row, symbol) when the attacker plays the pure strike column c. The action-retaining polytope is $\mathrm{conv}\{V_c : c \neq \theta\}$, so the three models' polytopes are again the three edges of a triangle, here in $|\mathrm{supp}| \times 4$ dimensions. Any two edges therefore meet at the vertex of the remaining index, so S6 stays pairwise-blind at every query even when the executed probe is logged. The triple intersection is empty exactly when V_0, V_1, V_2 are affinely independent, which holds at every query of positive resolution; we verified this at the six such queries (all but the structurally blind probe at $w = 0$). The mixture queries are strictly more informative under the finer filtration than the probe with averaged resolution that the transcript reports, so the argument must use the structure of S6, since the two filtrations differ there.

A.3 Index of results: hypotheses, scope and where each is used

Table 1 lists, for each main result, its hypotheses and its scope. The "Scope" column names the restriction that is easiest to overlook; each restriction is also stated in the result itself.

A.4 Tier II query generation

Tier II keeps the outer loop of Algorithm 1 and replaces the fixed pool by column generation, separation pricing followed by RDI pricing at a fixed threshold, as described in Section 5. Certified

Table 1: The main results, their hypotheses, and the scope restriction that is easiest to overlook. "Exact branch" means $\varepsilon_{\text{span}} = \eta_{\text{comp}} = 0$; our certified deployment oracle is in the approximate branch by at most 1.1×10^{-13} (Remark E.12).

Result	Hypotheses	Scope: what it does not say
Thm 4.1	Span bound (4.1); anytime-valid confidence sequence; certified deployment oracle.	Bounds the deployment gap of one policy; certifies no Nash pair, and is conditional on $\theta^\star \in \Theta$.
Thm 4.3	Joint common law at every accessible query; decision conflict; transcript filtration.	Excludes learners that retain their own executed action; Prop. A.1 shows the restriction is necessary. Does not imply that fixed tie-breaking is needed.
Thm 4.2	Convexity, compactness, minimax exchange, Helly.	A geometric identity for Γ; no statistical content.
Prop. 5.1	Bounded loss range G_W; one step.	Lower-bounds the information released in one step; the confidence radius need not contract.
Thm 4.4	Finite polytopes; exact branch; same ε for deployment and inference; $\delta < 1/(p+2)$ (class-level: $\delta < 1/\min\{M, p+1\}$, Remark E.8).	Characterizes (ε, δ)-correctness; almost sure termination is not implied. Converse holds for the transcript filtration only.
Thm E.7	As above, plus Assumption E.14.	The deadline $n^\dagger$ proves finite termination from the exploration floor, exceeds realized medians by factors of 6×10^2 to 4×10^5, and does not depend on the acquisition rule.
Prop. E.10	Never-empty confidence set; for part 2 also Assumption E.3, the exact branch, forced exploration, Assumption E.14, and a confidence set containing the frequency constraint (E.7).	Part 2 excludes modules that use only e-processes; Remark E.11 is a counterexample. A never-empty set only removes the exit for statistical failure and guarantees no policy.
Thm B.16	Convex channel pairs; reverse information projection.	Valid but not complete: Remark B.17.
Thm B.12	Finite models, queries and alphabets with convex channels (Assumption B.9); $\delta < 1/(p+2)$; stationary selectors at $\theta^\star$.	A loose lower bound that loses a factor $p+1$; no instance optimality.
Thm D.1	Finite common-support channels; span and solver bounds; pairwise confidence construction; forced exploration with full-support λ_0.	High-probability deadline only; the same for every acquisition rule.

residuals for the two pricing problems are Proposition B.5; finite termination of accepted columns is Proposition B.6, which requires $\kappa_{\text{sep}} < 1$ and strictly positive acceptance tolerances. Only separation pricing has a certified oracle (Proposition B.8); line 19 calls a global optimization oracle over $\mathcal{Q}_{\text{acc}}$ and the procedure returns UNKNOWN when none is supplied.

A.5 Deferred qualifications of the main results

The span budget. The span budget of Theorem 4.1 must come from an external model audit that is uniform over policies, such as confidence bounds on the finite probabilities of initial states and signals, on reward means, and on transition and signal kernels over every accessible history. The failure probability of their union bound must be added to δ, and the coupling bound in Corollary B.1 converts those bounds to $\varepsilon_{\text{span}}$. Held-out average performance alone does not bound the supremum over policies. Numerical error is obtained from feasible primal/dual bounds and verified best responses; a solver's success flag is never used. If $2\varepsilon_{\text{span}} + \eta_{\text{comp}} \geq \varepsilon$, the sufficient stopping theorem has no positive accuracy budget: improve the model or solver, relax the requested tolerance, or return UNKNOWN. In our experiments the span budget is zero, $\varepsilon_{\text{span}} = 0$, because calibration and deployment share one specified game (Appendix F).

Nestedness of the confidence sets. The sets C_n^e are nested (Theorem B.16); the optional moment ellipsoids of Theorem B.26 need not be nested, and intersecting their history restores nestedness and preserves coverage.

The run conditions behind the frequency deadline. The deadline of Theorem 4.4 assumes Assumption E.14: the optional obstruction exit disabled, every oracle returning, no administrative cap, and the test schedule of Lemma E.15, which costs an explicit factor $5/4$. Validity holds without any of these; only the deadline needs them. Outside the exact branch the condition is sufficient only, for the reason given in Corollary E.16.

Normalized factors and the composite module. In Theorem D.1, for numerically normalized factors replace exact KL by verified growth lower bounds g_q. If $g_q \geq -\xi$ rather than zero, the valid

Algorithm 2 Tier II query generation

Require: Witness W, prior π, protected pairs $\mathcal{B}$, seed pool Q_0, accessible universe $\mathcal{Q}_{\mathrm{acc}}$, separation fraction $\kappa_{\mathrm{sep}} \in [0, 1)$, tolerances $\tau_{\mathrm{sep}}, \tau_{\mathrm{gen}} > 0$
Ensure: Expanded finite pool Q and allocation $\lambda \in \Delta(Q)$, or the status UNKNOWN

$Q \leftarrow Q_0$
if $\mathcal{B} = \varnothing$ **then**
  Skip the separation master and set the threshold $b_0 \leftarrow 0$
else
  repeat
    Solve the separation master (B.9) on Q; let (w, t) be its optimal dual pair
    Call separation pricing over $\mathcal{Q}_{\mathrm{acc}}$ with weights w
    if the call returns no certified global upper bound **then**
      return UNKNOWN
    end if
    if the priced query's *exact* score exceeds $t + \tau_{\mathrm{sep}}$ **then**
      Verify its realized best-response face after the perturbation into the relative interior of Appendix D.6; add it to Q
    end if
  until no separation column is accepted
  $b_0 \leftarrow \kappa_{\mathrm{sep}} s^\star(Q)$
end if
repeat
  Solve the RDI allocation master (B.10) on Q at threshold b_0; if $b_0 = 0$ use the convention $(\alpha, \eta) = (0, \max_{q \in Q} I(q))$
  Call RDI pricing over $\mathcal{Q}_{\mathrm{acc}}$ with the dual weights
  if the call returns no certified global upper bound **then**
    return UNKNOWN
  end if
  if the priced query's exact score exceeds $\eta + \tau_{\mathrm{gen}}$ **then**
    Add it to Q and re-solve the RDI master at the *unchanged* threshold b_0 (Proposition B.5)
  end if
until no RDI column is accepted
return the final pool Q and finite-pool allocation λ

forced drift is $\gamma \sum_q \lambda_0(q) g_q - (1 - \gamma)\xi$; use this value in the deadline and require it to be positive. The combined confidence module gives the same argument with $h_\delta = \log(2(M - 1)/\delta)$ for its pairwise component of weight one half, and the support/evidence version covers singular channels.

The gap between the lower and upper bounds. Even at the realized exit time of Appendix E.3, the lower bound of Theorem B.12 and the upper bounds differ in the selector class, the probability calibration and the cost functional (Remark E.19). Whether the pigeonhole factor $p + 1$ can be removed is open: the multiple-answer argument of Degenne & Koolen (2019) is the natural comparison, but its bandit experiment with a single law per arm does not establish a sharp bound for a continuum of policy answers and history-dependent set-valued selectors. For experiments with a single law per arm the continuum of answers has since been treated by Poiani et al. (2026), whose asymptotically matching bounds include ε-Nash identification in two-player zero-sum matrix games from payoff queries, so the open part of the question is the selector-dependent channel set. Resets, announcement and myopia are assumptions of the experiment model.

A.6 Deferred details of the protocol and the algorithm

Announcement, off-path variation, and the response polytope. Announcement reveals the defender's mixed calibration policy to the attacker; its realized private seed stays hidden. Hiding the query would require a specified attacker belief over calibration policies and a different response correspondence; it can prevent this form of censoring, which depends on the query, but we assert no general ordering of informativeness between the two protocols. The best-response face is ex-ante optimal, and its off-path variation has zero realization probability under the query and cannot change the transcript law; the image channel polytope can therefore have lower dimension than the face itself, and such variation adds no informative threat. The κ-response extension and the corresponding sweep quantify bounded payoff sacrifice through the full response polytope (Proposition D.6).

Witness oracle, ties, and numerical tolerances. The witness oracle, tie handling and numerical tolerances belong to Algorithm 1 itself; their exact rules (the dual-face witness, the lexicographic tie-break and the tolerances) are collected in Appendix E.5. A singleton witness has zero RDI and should be handled by the deployment solver. If $\kappa_{\mathrm{sep}} = 0$ or $s_n^\star(Q) = 0$, the separation constraint protects no pair; this is a fallback state and is no evidence of a full-universe obstruction. Positive pricing tolerances mean that a strictly positive priced score is not enough to add a column, so an implementation must either use zero acceptance tolerances in an exact theorem or verify that the certified improvement exceeds τ_{sep} or τ_{gen}.

A.7 Extended related work

Table 2 summarizes the positioning of Section 2 by line of work against six properties that this paper combines: (A) the responder knows the game and the learner does not; (B) each experiment induces a set of laws selected adaptively by the responder, with no tie-break or response model assumed; (C) the learner stops at one sufficient decision without identifying the model; (D) the certificate is time-uniform and valid under adaptive querying; (E) the possibility condition is matched by an impossibility result, which for this paper holds in the exact branch with finite polytopes; and (F) the deployment game is sequential, which here means the finite sequence-form setting. The marks describe each line's standard setting, and the paragraphs below give the detail behind each row.

Table 2: Positioning against the closest lines of work. A: informed responder; B: set-valued feedback; C: decision target; D: anytime certificate; E: matched impossibility; F: sequential games; ∘ marks a partial match.

Line of work	A	B	C	D	E	F
One-sided POSG solvers with a known model (Horák et al., 2017; 2023)	∘	×	✓	×	∘	✓
Pure exploration, multiple or infinite answers (Degenne & Koolen, 2019; Poiani et al., 2026)	×	×	✓	✓	✓	×
Equilibrium identification from payoff queries (Zhou et al., 2017; Maiti et al., 2024)	×	×	✓	✓	∘	×
Commitment learning from best responses (Letchford et al., 2009; Peng et al., 2019; Bacchiocchi et al., 2025)	✓	×	✓	×	∘	×
Informed, deceptive or nonmyopic responders (Birmpas et al., 2021; Haghtalab et al., 2022; Zhao et al., 2023; Ananthakrishnan et al., 2024)	✓	∘	×	×	∘	×
Inverse games and utility fitting (Waugh et al., 2011; Ling et al., 2018)	∘	×	×	×	×	✓
Safe testing of composite nulls (Grünwald et al., 2024; Larsson et al., 2025)	×	✓	×	✓	∘	×
Set identification under multiple equilibria (Chernozhukov et al., 2007; Beresteanu et al., 2011; Galichon & Henry, 2011)	∘	✓	×	×	✓	×
Minimax-regret and robust decisions (Braziunas & Boutilier, 2007; Iyengar, 2005; Nilim & El Ghaoui, 2005)	×	×	✓	×	×	∘
This paper	✓	✓	✓	✓	✓	✓

One-sided partially observable stochastic games are solved by value iteration and heuristic search once the model is known (Horák et al., 2017; 2023); our problem is upstream of planning, with an uncertain model and feedback selected by a strategic best responder. Fixed-confidence pure exploration asks how many adaptive samples identify a correct answer (Garivier & Kaufmann, 2016; Kaufmann et al., 2016; Degenne & Koolen, 2019) and task-optimal exploration asks only what a downstream control task needs (Wagenmaker et al., 2021; Wagenmaker & Jamieson, 2022); RDI-EASI likewise targets only what the decision needs, but the answer is a robust maximin policy and each experiment carries a selector-adversarial channel set. Decision-focused learning shares the principle that parameter error need not measure decision error (Elmachtoub & Grigas, 2022; Donti et al., 2017; Wilder et al., 2019) but trains a predictor instead of selecting experiments under an anytime certificate, and reward-free exploration (Jin et al., 2020) and best-policy identification (Ménard et al., 2021; Al Marjani & Proutière, 2021) bracket our target from either side.

With $u_\theta(\mu) = g_\theta(\mu)$ the radius of a common policy is a minimax-regret criterion,

$$\Gamma(C) = \inf_{\mu} \sup_{\theta \in C} \{ \sup_{\mu'} u_\theta(\mu') - u_\theta(\mu) \},$$

already used for stopping and query selection in preference elicitation by minimax regret (Braziunas & Boutilier, 2007), and Helly-type bounds on support constraints are equally classical and underlie the scenario bounds of Campi & Garatti (2008); what is new is the coupling to zero-sum maximin deployment, set-valued best-response channels and confidence sequences that hold under every admissible selector. Learning equilibria from payoff queries (Zhou et al., 2017; Maiti et al., 2024) and learning commitment strategies from observed follower responses (Letchford et al., 2009; Blum et al., 2014; Peng et al., 2019) assume samples of individual payoff entries, or a best-response oracle whose ties are broken in the leader's favor or chosen by the learner. Poiani et al. (2026) already treat a continuum of correct answers, including learning an approximate Nash equilibrium of a two-player zero-sum matrix game from noisy payoff queries, so a continuous answer space does not distinguish our work. The difference is that their feedback is a noisy payoff sample from a fixed law, whereas ours is one symbol of a partial log, drawn from a set of laws that an informed opponent selects from adaptively. Theorem 4.3 gives a sufficient condition for the set-valued channel to admit no uniformly correct learner, and Theorem 4.4 shows it is the only obstruction in the finite exact branch with $\delta < 1/(p+2)$; neither implies that the assumptions of the papers above are necessary for the problems they study. Inverse equilibrium and differentiable inverse game methods fit utilities through a postulated response model (Waugh et al., 2011; Ling et al., 2018), whereas we retain a set of exact or boundedly suboptimal responses and never select one; commitment in security games (Kiekintveld et al., 2009) supplies the patrolling application, which we study under a maximin objective; and learning from nonmyopic buyers and Stackelberg agents (Amin et al., 2013; Haghtalab et al., 2022) models an incentive our myopic or fixed-κ class does not cover.

Optimal and Bayesian experimental design evaluate experiments by their effect on estimation or downstream decisions (Pukelsheim, 2006; Chaloner & Verdinelli, 1995; Foster et al., 2019), and Blackwell's comparison ranks them for all decision problems (Blackwell, 1953); RDI is narrower, a one-step decrease of the current least-favorable deployment impurity robustified against admissible selectors. Because the certificate must survive adaptive querying and optional stopping we use ideas on confidence sequences and e-processes from time-uniform inference and safe testing (Abbasi-Yadkori et al., 2011; Howard et al., 2021; Ramdas et al., 2023; Grünwald et al., 2024), to keep a running confidence set covering the true model under all admissible selectors, with no posterior interpretation. Set-identified econometric models distinguish a unique parameter from a consistent set (Chernozhukov et al., 2007); a strategic channel set plays a related role, but data laws change predictably with the selector and query, so our anytime supermartingale construction needs conditional validity that fixed-sample theory does not supply. Robust MDPs optimize against a prescribed uncertainty set (Iyengar, 2005; Nilim & El Ghaoui, 2005) and PSRO and double-oracle methods approximate large games (McMahan et al., 2003; Lanctot et al., 2017); our robust solver instead certifies an upper bound on the deployment gap over the current confidence set, so approximate query generation can slow the procedure but leaves the certificate valid.

B Technical Statements and Oracle Formulations

This appendix gives oracle formulations and supporting statements for the main results (Theorem 4.4 and Proposition 5.1 are detailed in Appendix E), query pricing, the non-sharp information lower bound, optional plug-in acquisition, linear moments and the illustrative bandit separation.

Notation and information structure. Throughout, $\delta, \delta' \in (0,1)$, logarithms are natural and error tolerances are nonnegative unless stated otherwise. For a defender policy μ, let $g_\theta(\mu) := \inf_{\nu \in \Pi_A} V_\theta^{\mu,\nu}$, $v_\theta^\star := \sup_{\mu \in \Pi_D} g_\theta(\mu)$ and $\mathrm{Gap}_\theta(\mu) := v_\theta^\star - g_\theta(\mu)$. For $\varnothing \neq C \subseteq \Theta$ and $a \geq 0$, let

$$\Gamma(C) := \min_{x \in X} \sup_{\theta \in C} \mathrm{Gap}_\theta(x), \qquad \mathcal{U}_a(\theta) := \{x \in X : \mathrm{Gap}_\theta(x) \leq a\}. \tag{B.1}$$

The finite extensive-form games have perfect recall and common nonempty compact realization-plan polytopes X, Y with $g_\theta(x) = \min_{y \in Y} x^\top A_\theta y$, $v_\theta^\star = \max_{x \in X} g_\theta(x)$ and $p := \dim(\mathrm{aff}\, X)$. The information structure and polytopes are fixed across models; a compact model class, whenever used, is a compact metric space with $\theta \mapsto A_\theta$ continuous. We use behavioral policies and realization-equivalent plans interchangeably. Rewards have a known bound $r_{\theta,h} \in [0, R_{\max}]$; $R_{\mathrm{tot}} := H R_{\max}$.

Negative rewards and the shift convention. Five designs have negative payoffs (S4, S5 and D5 throughout, S6 and D6 in part), so we make the normalization $r_{\theta,h} \geq 0$ explicit: replace $r_{\theta,h}$ by $r_{\theta,h} + c$ for a constant $c \geq \max_{\theta,h}(-r_{\theta,h})$ that does *not* depend on θ. Three quantities are invariant under this substitution and one is not. (i) Every value and every policy's worst-case payoff shift by Hc ($v^\star_\theta \mapsto v^\star_\theta + Hc$), so $\mathrm{Gap}_\theta(x)$, hence $\Gamma(C)$ and the decision problem, are unchanged. (ii) Best-response faces are unchanged, since adding a constant to every entry of a column-indexed payoff vector does not move its $\arg\min$; hence so are $\mathcal{P}_\theta(q)$, the channel polytopes and every statistical statement in this paper. (iii) The tolerance ε, being a gap, needs no rescaling. (iv) Scale constants change: the shifted span $R'_{\max} = \max_{\theta,h} r_{\theta,h} - \min_{\theta,h} r_{\theta,h}$ enters $R_{\mathrm{tot}} = HR'_{\max}$, the loss range G_W of Proposition 5.1, and the $(HR'_{\max}/2)$-sub-Gaussian constant of the reward concentration arguments. We always take the smallest such c, so $R'_{\max}$ is the realized span, measured on the deployment matrices of S4, S5, S6 and D6 and the terminal payoff vectors of the sequence-form D5: .562, .752, 4, .733 and 2.476 on S4, S5, S6, D5 and D6. No computation in this paper uses $R'_{\max}$. The constant entering Proposition 5.1 is the loss range $G_W = \max_{\theta\in W} \sup_x \mathrm{Gap}_\theta(x)$ at the canonical witness; our numerical check of Corollary E.2 uses its upper bound $\bar{G}_W := \max_{i\in W}(v_i^{\mathrm{upper}} - \min A_i) \geq G_W$, recomputed from the matrices (0.258, 0.324 and 2.000 on S4, S5 and S6), where $v_i^{\mathrm{upper}} \geq v_i^\star$ is the value of a feasible attacker mixture and $\min A_i$ the smallest payoff entry; a larger constant only weakens the checked inequality. We use one shift for all models throughout. A shift c_θ that depends on the model adds, over a fixed horizon, Hc_θ to $g_\theta(x)$ and to $v^\star_\theta$, so Gap_θ, $\Gamma(C)$ and every deployment claim are invariant under it too, though value differences change. When the recorded transcript contains reward information, however, it can change the diagnostic channel: two models sharing a record law before the shift need not share one after, and the common-law hypothesis would need rechecking. A single c for all models is thus a convention that keeps the channel fixed; the deployment problem does not require it.

A diagnostic query $q \in \mathcal{Q}_{\mathrm{acc}}$ specifies a reset law, an announced calibration policy μ^q and a recorded transcript Z. Let $\mathrm{BR}_\theta(q) := \arg\min_\nu V_\theta^{q,\nu}$ and $\mathcal{P}_\theta(q) := \{\mathcal{L}_\theta^{q,\nu}(Z) : \nu \in \mathrm{BR}_\theta(q)\}$; randomized best responses are included. With $\mathcal{F}_{t-1}$ the completed calibration history, put $\mathcal{G}_t := \sigma(\mathcal{F}_{t-1}, q_t)$, so $\mathcal{F}_{t-1} \subseteq \mathcal{G}_t \subseteq \mathcal{F}_t$. Admissibility means $\mathcal{L}_\theta(Z_t \mid \mathcal{G}_t) \in \mathcal{P}_\theta(q_t)$ almost surely. Selectors may depend on the entire history. Only one of two kinds of learner randomness can be added to the filtration at no cost. *Selection* randomness, the draw of q_t from a history-measurable distribution, is generated before Z_t and independent of θ given the history, so adding it changes no conditional law and leaves the transcript-only converse unaffected; it is already inside $\mathcal{G}_t$. *Execution* randomness, the bits selecting which pure action the announced mixed policy plays inside an episode, cannot be added in this way: it is independent of θ marginally but coupled with Z_t through the attacker's response, so the pair can carry information that neither component carries alone (Remark A.2: a 2×2 instance on which each component is uninformative and the pair identifies θ in one episode). The transcript filtration of Section 3 therefore contains only the query draws and $Z_1, \ldots, Z_n$. A learner covered by the converse must discard the executed action and any execution seed or random tape from which it could be reconstructed; retaining either gives a finer, possibly strictly finer, filtration, and the common-law hypothesis must be checked again on the joint channel that includes it. At execution time the defender samples $a_h \sim \mu_h^{q_t}(\cdot \mid I_h^D)$ in place of the robust deployment candidate. All kernels and decision rules are measurable; transcript spaces are standard Borel.

B.1 Deployment certificates and strategic obstructions

For two games with matched policy spaces define $d_V(G, G') := \sup_{\mu,\nu} |V_G^{\mu,\nu} - V_{G'}^{\mu,\nu}|$. The main statement is Theorem 4.1. The three conditions used throughout, stated in Section 4.1, are the span bound (4.1) between the deployment game $G^\dagger$ and the reference model $\theta^\star$, the coverage requirement (4.2), a separate assumption about the actual calibration experiment, and the certified solver bracket (4.3) returned at every nonempty C_n, whose η_{comp} includes certified optimization and the equilibrium error of each model and cannot substitute for statistical coverage. An empty confidence set is a statistical failure status, never treated as a vacuously valid deployment certificate.

Corollary B.1 (Primitive sufficient condition for span error)**.** *Suppose the two games have the same horizon and information structure; their joint initial state-and-signal laws differ in total variation by at most ϵ_{init}; at every common feasible history, their conditional expected stage rewards differ by at most ϵ_r and lie in $[0, R_{\max}]$; and their joint kernels for the next physical state and all signals*

entering either player's subsequent information history differ in total variation by at most ϵ_K. Then

$$d_V(G^\dagger, G(\theta^\star)) \le H\epsilon_r + HR_{\max}\epsilon_{\text{init}} + \frac{H(H-1)}{2}R_{\max}\epsilon_K. \tag{B.2}$$

Observed rewards must be included among these signals whenever they affect subsequent actions; a small difference in reward means alone does not control what reward observations reveal.

Definition B.2 (Joint strategic equivalence)**.** A nonempty finite set $\Xi \subseteq \Theta$ is jointly strategically observationally equivalent relative to $\mathcal{Q}$ if $\bigcap_{\theta\in\Xi}\mathcal{P}_\theta(q) \neq \varnothing$ for all $q \in \mathcal{Q}$. For infinite $\mathcal{Q}$, a common law must admit a measurable query-indexed selection realizable by admissible response selectors. For $|\Xi| = 2$ it is pairwise strategic observational equivalence; unqualified equivalence refers to $\mathcal{Q}_{\text{acc}}$.

The main statement is Theorem 4.3. Pairwise channel intersections need not supply a joint common law. The obstruction is defined over the full accessible query universe, whatever the current Tier II pool. Its probability bound concerns a fixed Ξ and the complete experiment, with no conditional probability claim for every data-dependent confidence set.

B.2 Duality for hidden information and finite witnesses

For a Borel prior π supported on a nonempty compact C, let $u(\pi) := \max_{x\in X}\mathbb{E}_{\theta\sim\pi}[g_\theta(x)]$; each informed attacker type best responds separately in this value. The main statement is Theorem 4.2. The dual prior measures the decision cost of hidden model information; it does not generate data, need not contain the true model in its support and never replaces the statistical model class. A sparse witness need not be easy to find: certifiably maximizing $\text{Gap}_\theta(x)$ over continuous C can be nonconvex. Here p is the affine dimension of the defender realization-plan polytope; that of the unknown parameter plays no role. In extensive-form games p can be large, so the theorem asserts that an exact finite witness exists and gives no computational compression in general. Affine primitive kernels need not make A_θ affine, as chance factors that depend on the parameter multiply along trajectories.

B.3 RDI acquisition and certified query generation

For $W = \{\theta_1, \dots, \theta_K\}$ and $\pi \in \Delta(W)$ define, with the infimum over measurable post-query rules,

$$\mathcal{R}(\pi; W) := \min_{x\in X}\sum_i \pi_i \,\text{Gap}_{\theta_i}(x), \tag{B.3}$$

$$\mathcal{R}^{\text{wc}}_q(\pi; W) := \inf_{\psi:\mathcal{Z}_q\to X}\sup_{K_i\in\mathcal{P}_{\theta_i}(q)}\sum_i \pi_i \int \text{Gap}_{\theta_i}(\psi(z))K_i(dz), \tag{B.4}$$

$$\text{RDI}_q(\pi; W) := \mathcal{R}(\pi; W) - \mathcal{R}^{\text{wc}}_q(\pi; W). \tag{B.5}$$

Proposition B.3 (Decision-impurity reduction)**.** *For every W, π, q, $0 \le \text{RDI}_q(\pi; W) \le \mathcal{R}(\pi; W)$. If the supported models admit a common law under q, their RDI is zero. If each channel is a singleton with density p^q_i relative to ρ, put $\bar{p}^q(z) := \sum_i \pi_i p^q_i(z)$ and $\pi^z_i := \pi_i p^q_i(z)/\bar{p}^q(z)$ when $\bar{p}^q(z) > 0$. Then $\text{RDI}_q(\pi; W) = \mathcal{R}(\pi; W) - \int \bar{p}^q(z)\mathcal{R}(\pi^z; W)\rho(dz)$.*

Proposition B.4 (Finite linear programs)**.** *Suppose $\mathcal{Z}_q$ is finite and $\mathcal{P}_{\theta_i}(q)$ is a polytope with extreme laws $k_{i\ell}$, $\ell \in L_i$. Write $A_i = A_{\theta_i}$ and $v^\star_i = v^\star_{\theta_i}$. The parent impurity is the value of*

$$\min_{x,g}\sum_i \pi_i g_i \quad \textit{s.t.} \quad x \in X,\ g_i \ge 0,\ g_i \ge v^\star_i - x^\top A_i y \quad (i \in [K],\ y \in \text{ext}\, Y). \tag{B.6}$$

The worst-selector child impurity is the value of

$$\begin{aligned}\min_{\{x_z\},\{g_{iz}\},\{t_i\}}\sum_i \pi_i t_i \quad \textit{s.t.} \quad & x_z \in X,\ g_{iz} \ge 0,\ t_i \ge \sum_z k_{i\ell}(z) g_{iz} \quad (i \in [K],\ \ell \in L_i),\\ & g_{iz} \ge v^\star_i - x_z^\top A_i y \quad (i \in [K],\ z \in \mathcal{Z}_q,\ y \in \text{ext}\, Y).\end{aligned} \tag{B.7}$$

The robust radius is

$$\Gamma(W) = \min_{x\in X, t}\{t : t \ge v^\star_i - x^\top A_i y \textit{ for all } i,\ y \in \text{ext}\, Y\}. \tag{B.8}$$

For a fixed finite-game query, the best-response realization plans form a face of a polytope and the transcript-law map is linear; therefore the channel set is a polytope.

Appendix D.6 sizes the child LP (one block of defender flow constraints per symbol), gives an executable scheme by constraint generation and separates acquisition residuals from the error of the deployment solver.

Use the convention $H^2(P,K) = 1 - \int \sqrt{dP\, dK}$ and let $d_q(\theta,\theta') := \inf_{P\in\mathcal{P}_\theta(q),\, K\in\mathcal{P}_{\theta'}(q)} H(P,K)$. At a nonterminal round, take an exact witness (W_n, π_n), let $a = \varepsilon - 2\varepsilon_{\mathrm{span}} \geq 0$, and define $\mathcal{A}_n := \{(\theta,\theta') \in W_n^2 : \mathcal{U}_a(\theta) \cap \mathcal{U}_a(\theta') = \varnothing\}$. Set $\mathcal{B}_n = \mathcal{A}_n$. If it is empty, impose no pairwise floor; forced exploration and composite inference handle remaining joint conflicts. Write $D_q(b) = d_q^2(\theta,\theta')$ for $b = (\theta,\theta')$ and $I_n(q) = \mathrm{RDI}_q(\pi_n; W_n)$. For a nonempty finite pool Q and nonempty $\mathcal{B}_n$, let

$$s_n^\star(Q) := \max_{\lambda\in\Delta(Q)} \min_{b\in\mathcal{B}_n} \sum_{q\in Q} \lambda(q) D_q(b). \tag{B.9}$$

For an infinite accessible universe, replace sums by integrals over Borel probability measures and maxima by suprema. For $\kappa_{\mathrm{sep}} \in [0,1]$, the RDI allocation is

$$\max_{\lambda\in\Delta(Q)} \sum_{q\in Q} \lambda(q) I_n(q) \quad \text{s.t.} \quad \sum_{q\in Q} \lambda(q) D_q(b) \geq \kappa_{\mathrm{sep}} s_n^\star(Q) \quad (b \in \mathcal{B}_n). \tag{B.10}$$

For empty $\mathcal{B}_n$, omit the separation constraints. Scores are assumed Borel measurable. Tier I uses a finite $Q = \mathcal{Q}_{\mathrm{acc}}$; Tier II grows a finite initial pool via the two dual pricing problems below.

Proposition B.5 (Tier II pricing certificates)**.** *Fix a finite nonempty $\mathcal{B}_n$ and a finite nonempty pool Q, and suppose the finite masters are solved exactly. The separation dual is $\min_{w\in\Delta(\mathcal{B}_n),\, t}\{t : \sum_b w(b) D_q(b) \leq t \ (q \in Q)\}$. Let (w,t) be optimal and $s = s_n^\star(Q) = t$. If a certified full-universe bound gives $\sup_{q\in\mathcal{Q}_{\mathrm{acc}}} \sum_b w(b) D_q(b) \leq t + \tau_{\mathrm{sep}}$, then $s \leq s_n^\star(\mathcal{Q}_{\mathrm{acc}}) \leq s + \tau_{\mathrm{sep}}$, so every feasible RDI allocation has separation at least $\kappa_{\mathrm{sep}} s_n^\star(\mathcal{Q}_{\mathrm{acc}}) - \kappa_{\mathrm{sep}}\tau_{\mathrm{sep}}$. For the threshold $b_0 = \kappa_{\mathrm{sep}} s$, which stays fixed for the whole RDI loop so that adding an RDI column does not move it, the RDI dual is $\min_{\alpha\geq 0,\eta}\{\eta - b_0 \sum_b \alpha(b) : I_n(q) + \sum_b \alpha(b) D_q(b) \leq \eta \ (q \in Q)\}$, with the convention $(\alpha,\eta) = (0, \max_{q\in Q} I_n(q))$ when $b_0 = 0$, where α carries a zero objective coefficient, so any pair b with $D_q(b) = 0$ for every $q \in Q$ leaves $\alpha(b)$ undetermined at the optimum. If an optimal dual solution satisfies $\sup_{q\in\mathcal{Q}_{\mathrm{acc}}}\{I_n(q) + \sum_b \alpha(b) D_q(b)\} \leq \eta + \tau_{\mathrm{gen}}$, the current objective is within τ_{gen} of the best accessible-universe RDI objective at the same threshold b_0.*

The last comparison does not use the stronger threshold $\kappa_{\mathrm{sep}} s_n^\star(\mathcal{Q}_{\mathrm{acc}})$. An added RDI column requires re-solving the RDI master at unchanged b_0; the separation master is not re-solved in the RDI loop, since refreshing b_0 from the enlarged pool would break the fixed-threshold hypothesis of Proposition B.6's finite-termination argument. Certifying pricing termination needs a global upper bound on the pricing supremum; a locally found query only bounds it below. Heuristic pricing, approximate RDI, lookahead or fallback exploration preserve Theorem 4.1 if the statistical and deployment certificates stay valid. A zero one-step RDI score is not by itself an obstruction certificate.

Proposition B.6 (Finite column addition in Tier II)**.** *Fix a round with finite W, prior π, nonempty finite $\mathcal{B}$, finite nonempty pool Q, and tolerances $\tau_{\mathrm{sep}}, \tau_{\mathrm{gen}} > 0$. Suppose every finite master is solved exactly, each pricing call returns a candidate in finite time, and every acceptance test uses the exact score of the returned column. When the separation threshold is zero, take the RDI dual convention $(\alpha,\eta) = (0, \max_{q\in Q} I(q))$. Then the separation-pricing loop can accept only finitely many columns in the round, and if $\kappa_{\mathrm{sep}} \in [0,1)$ so can the RDI pricing loop. Consequently the implemented Tier II subroutine returns after finitely many pricing calls. Under heuristic pricing this certifies termination of the implemented finite-candidate search only, and carries no full-universe optimality claim.*

The two pricing branches differ in what can be executed, for a structural reason independent of implementation effort. Separation pricing has a certified oracle: by Proposition B.8, under full logging it reduces to a maximum, over finitely many polyhedral cells, of a convex minimization solvable as a second-order-cone program, so its supremum over $\mathcal{Q}_{\mathrm{acc}}$ has a certified upper bound. RDI pricing has no such oracle: the worst-selector child impurity is not concave in the diagnostic plan, so $\sup_q\{I_n(q) + \sum_b \alpha(b) D_q(b)\}$ is a global non-convex program, and without an external global optimizer Algorithm 2 must answer UNKNOWN. The Tier II experiments D3 and D4 of Appendix G.4 therefore price on separation alone (D3 over a grid of diagnostic plans, D4 over a finite candidate list). They test the separation score and the finite termination of Proposition B.6; the full-universe certificate of Proposition B.8 is not implemented.

Assumption B.7 (Fully logged diagnostics)**.** The accessible universe is $\mathcal{Q}_{\mathrm{acc}} = \{q_x : x \in X_{\mathrm{acc}}\}$ for a polytope $X_{\mathrm{acc}} \subseteq X$ of diagnostic defender realization plans, all sharing one reset law. The diagnostic transcript records the full terminal history, including both players' realized actions, chance events, and realized rewards. Stage rewards are deterministic known functions $r_\theta(h)$ of the model and terminal history h.

Proposition B.8 (Certified separation pricing under full logging)**.** *Assume finite sequence-form games, Assumption B.7, a finite witness W, and a nonempty finite pair set $\mathcal{B}$. Write $d_q^2(b) = d_q^2(\theta, \theta')$ for $b = (\theta, \theta')$. (1) For each $\theta \in W$ and each face F of Y, the set of diagnostic plans x for which F is the attacker's optimal face is polyhedral. The common refinement over all $\theta \in W$ gives finitely many relatively open polyhedral cells on which every best-response face is constant. (2) Fix such a cell T and its frozen faces F_θ. For $b = (\theta, \theta')$, let $\mathcal{H}_b$ be the terminal histories whose recorded rewards agree under the two models, $\sigma_D(h)$ and $\sigma_A(h)$ the defender and attacker sequences at h, and $\kappa_b(h)$ the geometric mean of the two chance probabilities along h. On $\operatorname{cl} T$ define $\mathsf{Aff}_b(x) := \max_{y \in F_\theta,\, y' \in F_{\theta'}} \sum_{h \in \mathcal{H}_b} x_{\sigma_D(h)} \kappa_b(h) \sqrt{y_{\sigma_A(h)} y'_{\sigma_A(h)}}$. This affinity is continuous, convex in x and computable by a second-order-cone program. For x in the relative interior of T, $d_{q_x}^2(b) = 1 - \mathsf{Aff}_b(x)$. (3) For every $w \in \Delta(\mathcal{B})$, $\sup_{x \in X_{\mathrm{acc}}} \sum_{b \in \mathcal{B}} w(b) d_{q_x}^2(b) = \max_T [1 - \min_{x \in \operatorname{cl} T} \sum_{b \in \mathcal{B}} w(b) \mathsf{Aff}_b(x)]$. The supremum need not be attained by an executable query; reserve half of τ_{sep} for the verified interior perturbation of Appendix D.6. Each inner problem is convex over a polytope, and boundary points cannot exceed adjacent-cell values because enlarging a best-response face can only increase affinity. Hence separation pricing has a certified implementation at any positive tolerance on this fully logged query family. (4) Let y and y' be two distinct extreme realization plans in an attacker optimal face at x. If $x' \mapsto x'^\top A_\theta (y - y')$ is not identically zero on $\operatorname{aff} X_{\mathrm{acc}}$, then x lies on the proper equalization hyperplane $x^\top A_\theta y = x^\top A_\theta y'$ defining the cell decomposition. This concerns distinct realization-plan vertices or a positive-dimensional optimal face; two nonzero sequence coordinates of one pure plan do not suffice. The proposition certifies separation pricing only; it does not make the RDI pricing objective globally convex.*

If a closed-cell optimizer lies on a boundary, its value upper-bounds what can be executed inside any adjacent best-response cell. A Tier II implementation must either perturb to a query in the relative interior and verify the realized best-response face, or charge the possible score loss to the pricing tolerance; the boundary certificate alone is not an executable query.

B.4 Strategic information times

Assumption B.9 (Finite strategic channels)**.** The model class Θ and accessible class $\mathcal{Q}_{\mathrm{acc}}$ are finite, $|\Theta| \geq 2$, and $|\mathcal{Q}_{\mathrm{acc}}| \geq 1$. Each transcript alphabet $\mathcal{Z}_q$ is finite and each $\mathcal{P}_\theta(q)$ is nonempty, compact, and convex. The common sequence-form representation above holds. Unless otherwise stated, this subsection uses a realizable deployment model, $\varepsilon_{\mathrm{span}} = 0$.

We write the finite-alphabet divergence as $\mathrm{KL}(P\|K) := \sum_z P(z) \log(P(z)/K(z))$ and the binary divergence as $\mathrm{kl}(u, v) := u \log(u/v) + (1-u) \log((1-u)/(1-v))$, with the usual extended-real conventions: a zero numerator contributes zero, and a positive numerator over zero contributes $+\infty$. A singleton model class requires only solving that model. For a nonempty finite Ξ, compactness gives $\Gamma(\Xi) \leq \varepsilon$ if and only if $\bigcap_{\theta \in \Xi} \mathcal{U}_\varepsilon(\theta) \neq \varnothing$. Call Ξ incompatible when $\Gamma(\Xi) > \varepsilon$, and define the anchored conflict collection $\mathfrak{C}_\varepsilon(\theta^\star) := \{\Xi \subseteq \Theta : \theta^\star \in \Xi,\ 2 \leq |\Xi| \leq p + 2,\ \Gamma(\Xi) > \varepsilon\}$.

Definition B.10 (Selector-adversarial and selector-uniform times)**.** A stationary selector at $\theta^\star$ is $s = (K_q^s)_{q \in \mathcal{Q}_{\mathrm{acc}}}$ with $K_q^s \in \mathcal{P}_{\theta^\star}(q)$; its space is $\mathfrak{S}(\theta^\star) = \prod_q \mathcal{P}_{\theta^\star}(q)$. The directed divergences are

$$J_q(s, \theta) := \inf_{K \in \mathcal{P}_\theta(q)} \mathrm{KL}(K_q^s \| K), \tag{B.11}$$

$$J_q^{\mathrm{wc}}(\theta^\star, \theta) := \inf_{P \in \mathcal{P}_{\theta^\star}(q),\, K \in \mathcal{P}_\theta(q)} \mathrm{KL}(P \| K). \tag{B.12}$$

Both infima are attained as extended-real minima. The data-generating law is the first KL argument. For $\mathrm{Alt}_\varepsilon(x) := \{\theta : \mathrm{Gap}_\theta(x) > \varepsilon\}$, let

$$D_\varepsilon(s, \lambda) := \sup_{x \in \mathcal{U}_\varepsilon(\theta^\star)} \inf_{\theta \in \mathrm{Alt}_\varepsilon(x)} \sum_q \lambda(q) J_q(s, \theta) \tag{B.13}$$

and $D_\varepsilon(s) := \sup_{\lambda\in\Delta(\mathcal{Q}_{\mathrm{acc}})} D_\varepsilon(s,\lambda)$. Replacing $J_q(s,\theta)$ by $J_q^{\mathrm{wc}}(\theta^\star,\theta)$ defines $D_\varepsilon^{\mathrm{wc}}(\theta^\star,\lambda)$ and $D_\varepsilon^{\mathrm{wc}}(\theta^\star)$. Set

$$T_\varepsilon^{\mathrm{sel}}(\theta^\star) := \sup_{s\in\mathfrak{S}(\theta^\star)} \frac{1}{D_\varepsilon(s)}, \qquad T_\varepsilon^{\mathrm{wc}}(\theta^\star) := \frac{1}{D_\varepsilon^{\mathrm{wc}}(\theta^\star)}. \tag{B.14}$$

The conventions are $\inf\varnothing = +\infty$, $0\cdot(+\infty) = 0$, $1/0 = +\infty$, and $1/(+\infty) = 0$. Pointwise comparison implies $T_\varepsilon^{\mathrm{sel}}(\theta^\star) \le T_\varepsilon^{\mathrm{wc}}(\theta^\star)$.

Proposition B.11 (Anchored Helly representation)**.** *Under Assumption B.9 and with the costs $c_\theta := \sum_q \lambda(q) J_q(s,\theta)$,*

$$D_\varepsilon(s,\lambda) = \min_{\Xi\in\mathfrak{C}_\varepsilon(\theta^\star)} \max_{\theta\in\Xi\setminus\{\theta^\star\}} c_\theta. \tag{B.15}$$

The minimum over an empty collection is $+\infty$. The same identity holds with the two-sided divergences. Moreover, $D_\varepsilon(s) = 0$ if and only if some $\Xi \in \mathfrak{C}_\varepsilon(\theta^\star)$ satisfies

$$K_q^s \in \mathcal{P}_\theta(q) \quad \text{for all } \theta\in\Xi,\ q\in\mathcal{Q}_{\mathrm{acc}}. \tag{B.16}$$

The bound $p+2$ includes the prescribed true model and at most $p+1$ alternatives; it is distinct from the unanchored $p+1$ witness bound. Eliminating at least one alternative from each incompatible conflict set suffices for decision resolution.

Theorem B.12 (Expected-sample lower bound)**.** *Let Assumption B.9 hold, $\varepsilon > 0$, and $0 < \delta < 1/(p+2)$. Call an algorithm (ε,δ)-correct if, under every model and admissible selector,*

$$\mathbb{P}_\theta\{\tau<\infty,\ \mathrm{Gap}_\theta(\widehat{x}) \le \varepsilon\} \ge 1-\delta, \tag{B.17}$$

where τ is normal stopping and $\widehat{x}$ is its output. An obstruction or failure status does not count as a successful policy output. Let $N_q(n) := \sum_{t=1}^n \mathbf{1}\{q_t = q\}$, with $N_q(\tau)$ the corresponding count up to normal stopping. Every such algorithm satisfies, for every stationary $s\in\mathfrak{S}(\theta^\star)$,

$$\mathbb{E}_{\theta^\star,s}\tau \ge \frac{\mathrm{kl}((1-\delta)/(p+1),\,\delta)}{D_\varepsilon(s)}, \tag{B.18}$$

and $\sup_s \mathbb{E}_{\theta^\star,s}\tau \ge \mathrm{kl}((1-\delta)/(p+1),\,\delta)\, T_\varepsilon^{\mathrm{sel}}(\theta^\star)$. More precisely, if $0 < \mathbb{E}_{\theta^\star,s}\tau < \infty$ and $\bar\lambda(q) = \mathbb{E}_{\theta^\star,s}N_q(\tau)/\mathbb{E}_{\theta^\star,s}\tau$, then every $\Xi\in\mathfrak{C}_\varepsilon(\theta^\star)$ gives $\mathbb{E}_{\theta^\star,s}\tau\, \max_{\theta\in\Xi\setminus\{\theta^\star\}} \sum_q \bar\lambda(q) J_q(s,\theta) \ge \mathrm{kl}((1-\delta)/(|\Xi|-1),\delta)$.

For fixed p, the numerator in (B.18) is $(p+1)^{-1}\log(1/\delta) + O(1)$ as $\delta\downarrow 0$, so the bound keeps a pigeonhole factor and does not reach instance optimality. First, the uniform condition $\delta < 1/(p+2)$ is a worst-case geometric bound; the pigeonhole step of the proof needs only $\delta < 1/|\Xi|$ for the offending conflict set. The bound $|\Xi| \le p+2$ is Helly's bound in $\mathrm{aff}\, X$; when every gap factors through an affine map T, the reduction proving Corollary D.5 runs in $\mathrm{aff}\, T(X)$ instead. The refined bound above already has $\mathrm{kl}((1-\delta)/(|\Xi|-1),\delta)$, with no $p+1$ penalty; Remark E.8 gives details and lists $|\Xi|$ for the designs of Section 6. Second, the convention $\tau=\infty$ on non-output paths helps correctness but can make the expectation bound vacuous for procedures that return failure with positive probability, so the bound is informative only for procedures that never abort. Theorem E.18 restates it at the algorithm's actual exit time σ, the relevant form for procedures that may report failure or abstain. The empirical reports list normal certification, statistical failure, obstruction and administrative censoring separately.

Assumption B.13 (Common support)**.** For every q, every law in $\bigcup_{\theta\in\Theta}\mathcal{P}_\theta(q)$ is strictly positive on the same finite alphabet $\mathcal{Z}_q$.

Corollary B.14 (Infinite information time and structural diagnosis)**.** *Under Assumption B.9, the existence of a jointly strategically equivalent $\Xi\in\mathfrak{C}_\varepsilon(\theta^\star)$ implies $T_\varepsilon^{\mathrm{sel}}(\theta^\star) = T_\varepsilon^{\mathrm{wc}}(\theta^\star) = +\infty$, and no algorithm is (ε,δ)-correct on a class containing Ξ when $\delta < 1/|\Xi|$. Under Assumption B.13, the converse holds for $T_\varepsilon^{\mathrm{sel}}$: its being infinite implies the existence of such a joint conflict through $\theta^\star$. For a nonempty finite $C\subseteq\Theta$, every model in C belongs to a joint obstruction at tolerance $a\ge 0$ precisely when, for every $\theta\in C$, there is $S_\theta\subseteq C\setminus\{\theta\}$, $1\le|S_\theta|\le p+1$, for which $\Xi_\theta = \{\theta\}\cup S_\theta$ satisfies $\Gamma(\Xi_\theta) > a$ and $\bigcap_{\vartheta\in\Xi_\theta}\mathcal{P}_\vartheta(q)\neq\varnothing$ for all $q\in\mathcal{Q}_{\mathrm{acc}}$. If this test fails for a particular $\theta^\star$, then $D_a(s) > 0$ for every stationary selector there, with the information definitions restricted to C.*

A certified obstruction requires a lower bound proving $\Gamma(\Xi_\theta) > a$ and full-universe common-law feasibility; a deployment upper bound cannot prove this strict inequality. For explicitly polyhedral channels and finite query classes the feasibility checks are LPs; a failed search gives the status unknown. Reporting an obstruction can be optional, as the actual selector can be more informative than the worst one.

Remark B.15 (Comparison of the two times). If one stationary selector $\bar{s}$ satisfies $J_q(\bar{s}, \theta) = J_q^{\mathrm{wc}}(\theta^\star, \theta)$ for all q, θ, the two times coincide; this holds for two models or when every true-model channel is a singleton. In general, true-side minimizers against different alternatives need not coincide, so infinite $T_\varepsilon^{\mathrm{wc}}$ alone does not prove a joint obstruction. The lower bound ranges over stationary selectors; it does not characterize optimal complexity against all history-dependent selectors.

B.5 Confidence sequences robust to the selector

Assume B.9 and B.13. For an ordered pair $\theta \neq \theta'$ and query q, fix a joint information projection

$$(P_q^\star, K_q^\star) \in \arg\min_{P \in \mathcal{P}_{\theta'}(q),\, K \in \mathcal{P}_\theta(q)} \mathrm{KL}(P \| K), \qquad e_q^{\theta,\theta'}(z) := \frac{P_q^\star(z)}{K_q^\star(z)}. \tag{B.19}$$

The argument order is $\boxed{e^{\theta,\theta'} = e^{\mathrm{null}:\theta|\mathrm{alt}:\theta'},\ \mathrm{KL}(P_{\mathrm{alt}} \| K_{\mathrm{null}})}$. This factor is the reverse information projection for the least-favorable pair. Related work on composite nulls includes Larsson et al. (2025), Wasserman et al. (2020), and Grünwald et al. (2024). Our pairwise mixture is not shown to minimize the time to certify a decision. For $M = |\Theta|$, define

$$E_n(\theta; \theta') := \prod_{t=1}^{n} e_{q_t}^{\theta,\theta'}(Z_t), \qquad E_0 = 1, \tag{B.20}$$

$$\bar{E}_n(\theta) := \frac{1}{M-1} \sum_{\theta' \neq \theta} E_n(\theta; \theta'), \qquad C_n^e := \bigcap_{m=0}^{n} \{\theta \in \Theta : \bar{E}_m(\theta) < 1/\delta\}. \tag{B.21}$$

Thus $C_0^e = \Theta$. The mixture always uses the fixed initial class and weights; neither model elimination nor witness changes renormalize it.

Theorem B.16 (Conditional e-validity and worst-case log-growth)**.** *Under Assumptions B.9 and B.13, the factors of* (B.19) *satisfy*

$$\sup_{K \in \mathcal{P}_\theta(q)} \mathbb{E}_K[e_q^{\theta,\theta'}(Z)] \leq 1, \tag{B.22}$$

$$\inf_{P \in \mathcal{P}_{\theta'}(q)} \mathbb{E}_P[\log e_q^{\theta,\theta'}(Z)] = J_q^{\mathrm{wc}}(\theta', \theta), \tag{B.23}$$

and for every true model, adaptive query rule, and admissible selector,

$$\mathbb{P}_{\theta^\star}\{\theta^\star \in C_n^e \text{ for all } n \geq 0\} \geq 1 - \delta. \tag{B.24}$$

The running intersection gives $C_{n+1}^e \subseteq C_n^e$, hence $\Gamma(C_{n+1}^e) \leq \Gamma(C_n^e)$ while the sets are nonempty.

Remark B.17 (Pairwise e-processes can miss joint constraints). The e-process above is valid but not complete for all finite convex channel families. Let $\Theta = \{1, 2, 3\}$, $X = \Delta_3$, and $g_i(x) = x_i$, so $v_i^\star = 1$ and $\mathrm{Gap}_i(x) = 1 - x_i$. Then $\Gamma(\{1, 2, 3\}) = 2/3$ while every subset of two models has radius $1/2$; for $\varepsilon = 0.6$, eliminating any one false model suffices. Give the algorithm one query with alphabet $\{1, 2, 3\}$ and channel sets $\mathcal{P}_1 = \mathrm{conv}\{(.1, .8, .1), (.1, .1, .8)\}$, $\mathcal{P}_2 = \mathrm{conv}\{(.8, .1, .1), (.1, .1, .8)\}$ and $\mathcal{P}_3 = \mathrm{conv}\{(.8, .1, .1), (.1, .8, .1)\}$. All probabilities are strictly positive, every pair of channel sets intersects, and the three sets have no common intersection. Therefore every pairwise KL projection has value zero and the factors (B.19) are identically one, so C_n^e never shrinks. There is, however, no joint strategic obstruction here: under model i, every admissible law has $\mathbb{P}(Z = i \mid \mathcal{F}_{t-1}) = .1$, and the three empirical frequency constraints cannot all hold because the frequencies sum to one. A confidence procedure using these joint constraints can eliminate at least one model while retaining the truth. With the uniform least-favorable prior, the query has $\mathcal{R} = 2/3$, $\mathcal{R}_q^{\mathrm{wc}} = 0.55$, and $\mathrm{RDI} = 7/60 > 0$. Pairwise e-process confidence sets are therefore a conservative statistical module, needing either additional tests of joint constraints or assumptions under which pairwise separation is sufficient.

For the stopping results define $B_e := \max_{\theta\neq\theta',q,z} |\log e_q^{\theta,\theta'}(z)| < \infty$, $h_\delta := \log((M-1)/\delta)$ and

$$b_n^e(\delta') := 2B_e\sqrt{2n\log\frac{Mn(n+1)}{\delta'}}, \qquad n \geq 1. \tag{B.25}$$

Consider an exact-solver run that checks $\Gamma(C_n^e) \leq \varepsilon$ starting at $n = 0$, has no imposed budget, and does not terminate early on an obstruction report; an empty set produces failure and no policy. The process can be mathematically continued with admissible kernels after termination for concentration arguments; this requires no interaction after termination.

Theorem B.18 (Fixed-allocation stopping bound)**.** *Under the preceding assumptions and run conventions, suppose $q_t \mid \mathcal{F}_{t-1} \sim \lambda$ for a fixed allocation, using fresh draws, including in the mathematical continuation. Let $c_\lambda(\theta) := \sum_q \lambda(q) J_q^{\mathrm{wc}}(\theta^\star,\theta)$ and $n_\theta := \min\{n \geq 1 : nc_\lambda(\theta) - b_n^e(\delta') \geq h_\delta\}$, where $\min\varnothing = \infty$. If the conflict collection is nonempty, put*

$$n^\star(\delta,\delta') := \max_{\Xi\in\mathfrak{C}_\varepsilon(\theta^\star)} \min_{\theta\in\Xi\setminus\{\theta^\star\}} n_\theta. \tag{B.26}$$

When $n^\star < \infty$, every admissible selector satisfies $\mathbb{P}_{\theta^\star}\{\tau \leq n^\star,\ \mathrm{Gap}_{\theta^\star}(x_\tau) \leq \varepsilon\} \geq 1-\delta-\delta'$. If the conflict collection is empty, normal stopping occurs at $n = 0$. If it is nonempty and $D_\varepsilon^{\mathrm{wc}}(\theta^\star,\lambda) > 0$, then, for fixed $\delta' \in (0,1)$ and $\delta \downarrow 0$,

$$n^\star(\delta,\delta') = (1+o(1))\frac{\log(1/\delta)}{D_\varepsilon^{\mathrm{wc}}(\theta^\star,\lambda)}. \tag{B.27}$$

A maximizing allocation exists and gives coefficient $T_\varepsilon^{\mathrm{wc}}(\theta^\star)$, but may depend on the unknown true model. If λ has full support and $T_\varepsilon^{\mathrm{wc}}(\theta) < \infty$ for every θ, the run is (ε,δ)-correct.

The asymptotic statement fixes δ'; taking δ' of order δ changes the concentration term. The theorem bounds normal stopping with high probability; it says nothing about expected stopping time.

Corollary B.19 (Adaptive accumulated information)**.** *Allow any adaptive randomized queries within the fixed finite $\mathcal{Q}_{\mathrm{acc}}$. Define $A_n(\theta) := \sum_{t=1}^n J_{q_t}^{\mathrm{wc}}(\theta^\star,\theta)$. With probability at least $1-\delta'$, simultaneously over $n \geq 1$ and $\theta \neq \theta^\star$,*

$$\log E_n(\theta;\theta^\star) \geq A_n(\theta) - b_n^e(\delta'). \tag{B.28}$$

For the exact-solver run above, with probability at least $1-\delta-\delta'$, simultaneously for all $n \geq 1$,

$$\min_{\Xi\in\mathfrak{C}_\varepsilon(\theta^\star)} \max_{\theta\in\Xi\setminus\{\theta^\star\}} A_n(\theta) \geq h_\delta + b_n^e(\delta') \quad\Longrightarrow\quad \tau \leq n. \tag{B.29}$$

An empty conflict collection again permits stopping at zero. This condition is sufficient; whether a particular adaptive rule reaches the threshold is left open.

Proposition B.20 (Numerically certified e-factors)**.** *Suppose that, before Z_t is observed, a positive approximate factor $\widetilde{e}_t^{\theta,\theta'}$ and a finite bound $c_t^{\theta,\theta'}$ are $\mathcal{G}_t$-measurable and satisfy $c_t^{\theta,\theta'} \geq \sup_{K\in\mathcal{P}_\theta(q_t)} \mathbb{E}_K[\widetilde{e}_t^{\theta,\theta'}(Z)]$. The normalized factor $e_{t,\mathrm{safe}}^{\theta,\theta'}(z) := \widetilde{e}_t^{\theta,\theta'}(z)/\max\{1, c_t^{\theta,\theta'}\}$ retains conditional e-validity, and its products with fixed-weight mixtures retain* (B.24)*. Growth bounds must use certified lower conditional log-growth bounds and a valid bound on the safe log increments, rather than automatically using the exact J_q^{wc} and B_e.*

Corollary B.21 (Deployment and solver tolerances)**.** *Suppose the actual diagnostic laws still validate the confidence construction for $\theta^\star$,* (4.1) *holds, the solver satisfies both inequalities in* (4.3)*, and $\varepsilon_0 := \varepsilon - 2\varepsilon_{\mathrm{span}} - \eta_{\mathrm{comp}} > 0$. Then the fixed-allocation, forced-exploration, and adaptive sufficient-stopping bounds all apply with ε_0 in place of ε in the answer sets and conflict collections, and the resulting normal output is certified at deployment tolerance ε. The safety statement remains* (4.4)*. This corollary relaxes the convention $\varepsilon_{\mathrm{span}} = 0$ of Assumption B.9.*

Remark B.22 (Supports and the scope of the rate theorem)*.* Common support excludes noiseless transcripts that reveal the model. An infinite divergence and zero characteristic time can still leave an $O(1)$ diagnostic cost; observations are needed even then. For finite alphabets, the rule that eliminates θ only when $\sup_{K\in\mathcal{P}_\theta(q_t)} K(Z_t) = 0$ has zero coverage error. Alternatively, an actual observation randomization $P \mapsto (1-\eta)P + \eta U_q$, with $0 < \eta < 1$ and U_q strictly positive, creates common support; merely smoothing the likelihood of untransformed observations does not justify the theorem. The rates for finite classes need not extend to an infinite Tier II query universe.

B.6 Optional plug-in KL acquisition

This variant shares the confidence and deployment modules of the default RDI rule. For nonempty C_n^e, choose $\check{\theta}_n \in \arg\min_{\theta\in C_n^e} \bar{E}_n(\theta)$ and $x_n^{\mathrm{cand}} \in \arg\min_{x\in X} \mathrm{Gap}_{\check{\theta}_n}(x)$. Choose $a > 0$, for example $a = \varepsilon_0$ from Corollary B.21, and set $\mathcal{A}_n^{\mathrm{ct}} := \{\theta \in C_n^e : \mathrm{Gap}_\theta(x_n^{\mathrm{cand}}) > a\}$. An approximate candidate is allowed only with a certificate that its gap at $\check{\theta}_n$ is at most a. For nonempty $\mathcal{A}_n^{\mathrm{ct}}$ and finite active pool Q, use

$$\lambda_n^{\mathrm{ct}} \in \arg\max_{\lambda\in\Delta(Q)} \min_{\theta\in\mathcal{A}_n^{\mathrm{ct}}} \sum_{q\in Q} \lambda(q) J_q^{\mathrm{wc}}(\check{\theta}_n, \theta). \tag{B.30}$$

Proposition B.23 (KL acquisition and its guarantees)**.** *Under Assumptions B.9 and B.13,* (B.30) *is a finite LP for a fixed candidate. Its dual pricing objective for weights $w \in \Delta(\mathcal{A}_n^{\mathrm{ct}})$ is $\sup_{q\in\mathcal{Q}_{\mathrm{acc}}} \sum_{\theta\in\mathcal{A}_n^{\mathrm{ct}}} w(\theta) J_q^{\mathrm{wc}}(\check{\theta}_n, \theta)$. A certified pricing residual gives the same additive max–min certificate as the separation part of Proposition B.5. Furthermore, for either ordering of the arguments on the right, $d_q^2(\theta,\theta') \le \frac{1}{2} J_q^{\mathrm{wc}}(\theta,\theta')$ and $d_q^2(\theta,\theta') \le \frac{1}{2} J_q^{\mathrm{wc}}(\theta',\theta)$, the left side being symmetric while J_q^{wc}, by Definition B.10, puts the data-generating model first; in* (B.30) *the first argument is the plug-in model $\check{\theta}_n$. If $\mathcal{A}_n^{\mathrm{ct}} = \varnothing$ is certified over all of C_n^e, the candidate may be deployed only after obtaining a certified $U_n^{\mathrm{cand}} \ge \sup_{\theta\in C_n^e} \mathrm{Gap}_\theta(x_n^{\mathrm{cand}})$ with $U_n^{\mathrm{cand}} + 2\varepsilon_{\mathrm{span}} \le \varepsilon$.*

The plug-in model is excluded from $\mathcal{A}_n^{\mathrm{ct}}$; using the robust deployment candidate instead can include the plug-in model as an alternative and force a zero objective. Optimizing one plug-in optimal answer need not attain the full objective of the characteristic time, which jointly optimizes the answer and allocation. The RDI protected pairs and KL alternatives need not coincide. These rules inherit statistical safety, but no optimal adaptive sample complexity follows from the fixed-allocation theorem.

B.7 Observable linear moments

Assumption B.24 (Conditional linear moments with bounded bias)**.** For known $a_q \in \mathbb{R}^m$ and $M_q \in \mathbb{R}^{m\times d}$, the recorded statistic satisfies $Y_t = a_{q_t} + M_{q_t}\theta^\star + b_t + \xi_t$, where b_t is $\mathcal{G}_t$-measurable and $\|b_t\|_2 \le b$. Uniformly over admissible selectors, $\mathbb{E}[\exp(u^\top \xi_t) \mid \mathcal{G}_t] \le \exp(\sigma^2\|u\|_2^2/2)$ for $u \in \mathbb{R}^m$, with $\sigma > 0$. Also $\|\theta^\star\|_2 \le B_\theta$ and $\|M_q\|_{\mathrm{op}} \le L_M$. In the unbiased branch $b = 0$, all selectors have the same conditional mean $a_q + M_q\theta^\star$. In the biased branch, b_t may depend on the actual selector; only its uniform bound is required.

Proposition B.25 (When observed payoffs give linear moments)**.** *Suppose the diagnostic total payoff $R = \sum_{h=1}^H r_{\theta,h}$ is observed, and put $m_q(\theta) := \inf_\nu V_\theta^{q,\nu}$. Every exact best response has $\mathbb{E}_\theta^{q,\nu} R = m_q(\theta)$. For $\mathrm{BR}_\theta^\kappa(q) := \{\nu : V_\theta^{q,\nu} \le m_q(\theta) + \kappa\}$, the mean lies in $[m_q(\theta), m_q(\theta) + \kappa]$. Under every such selector the conditionally centered payoff is $(HR_{\max}/2)$-sub-Gaussian: our convention is $\mathbb{E}e^{u\xi} \le e^{u^2\sigma^2/2}$, so the variance proxy is $\sigma^2 = H^2 R_{\max}^2/4$. If $m_q(\theta) = a_q + M_q\theta$ on the model class, exact responses give Assumption B.24 with $b = 0$, and κ-best responses give it with a selector-dependent $0 \le b_t \le \kappa$. If the query's sequence-form payoff matrix A_θ^q is affine on a convex parameter domain, m_q is concave and piecewise affine there. A sufficient condition for affineness is a single attacker realization plan minimizing $x_q^\top A_\theta^q y$ for every model.*

Mean invariance does not imply distributional invariance or affineness in θ. A known mean and a sub-Gaussian envelope do not specify a likelihood or a Bayesian posterior. RDI evaluation requires channel laws in addition to Assumption B.24. In the unbiased Gaussian model, $Y - a_q \mid (\theta, q) \sim N(M_q\theta, \sigma^2 I_m)$ and $H^2(P_\theta^q, P_{\theta'}^q) = 1 - \exp(-\|M_q(\theta-\theta')\|_2^2/(8\sigma^2))$. Thus squared mean separation serves only as a local proxy for likelihood-based information gain. With bias, separation must account for the admissible conditional-mean sets. For $\lambda_{\mathrm{reg}} > 0$ define

$$V_n := \lambda_{\mathrm{reg}} I_d + \sum_{t=1}^n M_{q_t}^\top M_{q_t}, \qquad \widehat{\theta}_n := V_n^{-1} \sum_{t=1}^n M_{q_t}^\top (Y_t - a_{q_t}), \tag{B.31}$$

$$\beta_n(\delta) := \sigma\sqrt{2\log\frac{\det(V_n)^{1/2}}{\det(\lambda_{\mathrm{reg}} I_d)^{1/2}\delta}} + \sqrt{\lambda_{\mathrm{reg}}} B_\theta, \quad r_n(\delta) := \beta_n(\delta) + b\sqrt{n}. \tag{B.32}$$

Theorem B.26 (Anytime moment confidence and deployment fallback). *Under Assumption B.24, the sets* $C_n^{\mathrm{lin}} := \{\theta \in \Theta : \|\theta - \widehat{\theta}_n\|_{V_n} \le r_n(\delta)\}$ *contain* $\theta^\star$ *simultaneously with probability at least* $1-\delta$. *On that event,* $\mathrm{diam}_2(C_n^{\mathrm{lin}}) \le 2r_n(\delta)/\sqrt{\lambda_{\min}(V_n)}$. *If* $\mathrm{Gap}_\theta(x)$ *is* $2L_G$*-Lipschitz in* θ *uniformly over* x, *and* (4.1)–(4.3) *hold with these confidence sets, then*

$$\mathrm{Gap}_{G^\dagger}(\mu_n) \le 2\varepsilon_{\mathrm{span}} + \eta_{\mathrm{comp}} + \frac{4L_G r_n(\delta)}{\sqrt{\lambda_{\min}(V_n)}}. \tag{B.33}$$

If $\lambda_{\min}(\sum_{t=1}^n M_{q_t}^\top M_{q_t}) \ge n\alpha^2$ *with* $\alpha > 0$, *the denominator is at least* $\sqrt{\lambda_{\mathrm{reg}} + n\alpha^2}$, *and the contribution of the bias term is at most* $4L_G b/\alpha$.

This is a conservative fallback based on the full diameter; the robust radius can certify earlier. The quantity $4L_G b/\alpha$ appears in an upper bound; it does not lower-bound the achievable deployment error. Still, bounded adversarial bias alone can prevent arbitrarily accurate identification and deployment.

Lemma B.27 (Optional full-rank exploration). *Suppose that* $\alpha_0 > 0$ *and a finite* $\mathcal{Q}_0 \subseteq \mathcal{Q}_{\mathrm{acc}}$ *satisfy* $|\mathcal{Q}_0|^{-1}\sum_{q\in\mathcal{Q}_0} M_q^\top M_q \succeq \alpha_0^2 I_d$. *At each round, independently with probability* $\gamma \in (0,1]$, *replace the active query by a uniform draw from* $\mathcal{Q}_0$. *There is a universal* $c > 0$ *such that with probability at least* $1-\delta'$, *for every* $n \ge 1$, $\lambda_{\min}(\sum_{t=1}^n M_{q_t}^\top M_{q_t}) \ge \gamma\alpha_0^2 n - cL_M^2(\sqrt{n\ell_n} + \ell_n)$, *where* $\ell_n := \log(2d(n+1)^2/\delta')$. *Consequently* $\alpha^2 = \gamma\alpha_0^2/2$ *is valid once the deviation term is at most* $\gamma\alpha_0^2 n/2$. *A sufficient threshold is* $n \ge C\, L_M^4 \gamma^{-2}\alpha_0^{-4}\log(edL_M^4/(\delta'\gamma^2\alpha_0^4))$ *for a sufficiently large universal* C.

For κ-response channels (Proposition D.6), the certificate and obstruction arguments carry over only when coverage is valid for the enlarged class, the e-process rate theorem also needs its assumptions of finite channels and common support, and an unbounded sacrifice of diagnostic payoff for future strategic gain is not covered by a fixed κ model.

B.8 Decision resolution versus full identification

Consider the family of hidden rooms with $d \ge 2$, $\Theta = [0,1]^d$, $H = 1$, an inert attacker, and two defender actions. Deployment starts in room s_1, with rewards $r_\theta(s_1, a_1) = \theta_1$ and $r_\theta(s_1, a_2) = 1 - \theta_1$. In room s_j, $j \ge 2$, both actions have reward θ_j. Query q_j resets to room s_j and exposes only $Y = \theta_j + \xi$, $\xi \sim N(0, \sigma^2)$, with independent diagnostic noise; the noiseless reward is not observed during calibration. Let $\mathcal{Q}_{\mathrm{acc}} = \{q_1, \dots, q_d\}$.

Proposition B.28 (Illustrative bandit separation). *Fix* $\varepsilon \in (0, 1/4)$, $\delta \in (0, 1/3)$, *and* $\sigma > 0$. *Run exact RDI-EASI without forced exploration. Maintain* $C_n = I_{1,n} \times [0,1]^{d-1}$, *where* $I_{1,n}$ *is the Gaussian confidence interval specified in the proof, and use the endpoint witness whose nuisance coordinates are all* $1/2$. *Before normal stopping,* q_1 *is the unique query with positive RDI, and Tier I chooses only* q_1. *Exact Tier II pricing over the same accessible class generates* q_1 *if absent and then chooses only it when the acceptance tolerances are zero, or when the certified improvement margin at the current round exceeds the positive tolerances. There is a universal* C *such that* $\mathbb{P}_\theta\{\tau \le N,\ \mathrm{Gap}_\theta(x_\tau) \le \varepsilon\} \ge 1-\delta$ *with* $N \le C[1 + (\sigma^2/\varepsilon^2)\{\log(1/\delta) + \log(e + \sigma/\varepsilon)\}]$. *The bound is independent of* d. *In contrast, let* $\theta^0 = (1/2, \dots, 1/2)$ *and* $0 < r \le 1/8$. *Every procedure using these diagnostic observations, stopping almost surely under every model, and satisfying* $\mathbb{P}_\theta\{\theta \in S_\tau,\ \mathrm{diam}_2(S_\tau) \le r\} \ge 1-\delta$ *for all* $\theta \in \Theta$ *obeys*

$$\mathbb{E}_{\theta^0}\tau \ge \frac{d\sigma^2}{2r^2}\,\mathrm{kl}(1-\delta, \delta). \tag{B.34}$$

The diameter criterion is a strictly stronger accuracy objective: $d_V(G(\theta), G(\theta')) \le |\theta_1 - \theta_1'| \le \|\theta - \theta'\|_2$, *so solving any representative of a covered set of diameter* r *incurs gap at most* $2r$. *Taking* $r = \varepsilon/2$ *is sufficient for the requested deployment accuracy, but not equivalent to that objective.*

Canonicalizing witnesses matters for acquisition: models with the same deployment payoff may have different diagnostic laws. Consolidating duplicate gap functions does not justify replacing a channel by the union of all channels in that equivalence class.

Proposition B.29 (A censored matrix-game example). *Let* $\Theta = \{0, 1\}$ *and* $A_\theta = \mathrm{diag}(1 + 2\theta, 1)$, *with the defender maximizing. Query* q_x *announces row-one probability* x *and records* $Z =$

(A, B, R). *The robust radius of the two models is* $1/8$. *For* $0 \le a < 1/8$, *the sets* $\mathcal{U}_a(0) = [1/2 - a, 1/2 + a]$ *and* $\mathcal{U}_a(1) = [1/4 - a/3, 1/4 + a]$ *are disjoint. The restricted class* $\{q_{1/2}\}$ *admits a common transcript law. Every* q_x *with* $1/4 < x < 1/2$ *forces different unique best-response columns in the two models. The resulting transcript laws are mutually singular, so* $d_{q_x}(0, 1) = 1$ *and the model is identified in one episode. In particular, at* $q_s = q_{3/8}$ *the payoff statistic has* $\mathbb{E}[R \mid \theta, q_s] = \frac{3}{8} + \frac{1}{4}\theta$ *for* $\theta \in \{0, 1\}$*; this affine identity concerns the two-point class only.*

Corollary B.30 (Failure of restricted-query sufficiency)**.** *In the instance of Proposition B.29, let* $\mathcal{Q}_{\mathrm{restr}} = \{q_{1/2}\}$. *For every accuracy* $a < 1/8$, *no adaptive algorithm whose normal queries remain in* $\mathcal{Q}_{\mathrm{restr}}$ *can guarantee an* a*-optimal finite normal policy output with success probability strictly larger than* $1/2$ *uniformly over the two models, under admissible least-informative best-response selectors. Equivalently, for any target confidence* $1 - \delta > 1/2$, *repeated sampling from the current restricted pool alone is insufficient for a valid normal maximin-sufficiency certificate at tolerance* a. *This is only a restricted-pool obstruction: in the same instance, every* q_x *with* $1/4 < x < 1/2$ *separates the two models, and* $q_{3/8}$ *identifies the model in one episode.*

Proof. At $q_{1/2}$, model 0 has both columns as best responses, while model 1 has b_2 as its unique best response. The least-informative selector for model 0 chooses b_2, so both models induce the same full-transcript law at the only restricted query. Because the restricted pool is a singleton, every restricted-pool adaptive experiment has the same law under the two models. For $a < 1/8$, Proposition B.29 gives disjoint answer sets $\mathcal{U}_a(0)$ and $\mathcal{U}_a(1)$. Let μ be the common subprobability law of the algorithm's finite normal output, assigning no mass when the algorithm does not stop normally. Then $\mu(\mathcal{U}_a(0)) + \mu(\mathcal{U}_a(1)) \le 1$, so at least one of the two models has success probability at most $1/2$. The separating-query statement is the full-query claim of Proposition B.29. □

The full query class has reward bound $R_{\max} = 3$. For exact responses at $q_{1/2}$ and $q_{3/8}$ the realized rewards are in $\{0, 1\}$, so the centered payoff there is $1/2$-sub-Gaussian. The separating full-transcript channels have disjoint supports and are outside Assumption B.13; zero-error elimination applies instead. Noising only B does not restore common support when the disjoint raw (A, R) observations remain available.

C Proofs of the Main Results

Theorem 4.4 and Proposition 5.1 are proved in Appendix E, through Theorems E.6 and E.7 and Proposition E.1, respectively.

C.1 Certificates, span transfer, and strategic obstructions

Proof of Theorem 4.1. Let $\mathcal{E} = \{\theta^\star \in C_n \text{ for every } n \ge 0\}$. For any fixed defender policy, the definition of d_V gives

$$|g_{G^\dagger}(\mu) - g_{\theta^\star}(\mu)| \le \varepsilon_{\mathrm{span}}.$$

Taking suprema over policies also gives $|v^\star_{G^\dagger} - v^\star_{\theta^\star}| \le \varepsilon_{\mathrm{span}}$. Therefore, for every μ,

$$\mathrm{Gap}_{G^\dagger}(\mu) \le \mathrm{Gap}_{\theta^\star}(\mu) + 2\varepsilon_{\mathrm{span}}.$$

On $\mathcal{E}$, the certified upper bound implies, simultaneously for every nonempty-set round,

$$\mathrm{Gap}_{G^\dagger}(\mu_n) \le \sup_{\theta \in C_n} \mathrm{Gap}_\theta(\mu_n) + 2\varepsilon_{\mathrm{span}} \le U_n + 2\varepsilon_{\mathrm{span}}.$$

Consequently $\{\tau < \infty, \mathrm{Gap}_{G^\dagger}(\mu_\tau) > \varepsilon\} \subseteq \mathcal{E}^c$. Coverage proves the claim. No step uses the acquisition objective or the second solver inequality. □

Proof of Corollary B.1. Fix matched behavioral-policy kernels in the two games. Couple the initial state-and-signal draws maximally. As long as all relevant histories match, use common action randomization and maximally couple the next joint state-and-signal draws. If D_h denotes divergence before stage h, then $\mathbb{P}(D_h) \le \epsilon_{\mathrm{init}} + (h - 1)\epsilon_K$. On matching histories the conditional expected

rewards differ by at most ϵ_r; otherwise their difference is at most $R_{\max}$. The tower property and the triangle inequality give

$$|V_{G^\dagger}^{\mu,\nu} - V_{\theta^\star}^{\mu,\nu}| \le \sum_{h=1}^{H} \{\epsilon_r + R_{\max}[\epsilon_{\text{init}} + (h-1)\epsilon_K]\}.$$

Evaluating the sum and taking the supremum over matched policy pairs proves (B.2). The coupling requires equality of the complete action-relevant histories. This is why observed rewards are included among the signals controlled by ϵ_K. □

Proof of Theorem 4.3. For each accessible query choose $K_q \in \bigcap_{\theta\in\Xi} \mathcal{P}_\theta(q)$, using the assumed measurable selection when necessary. Under every model use an admissible selector inducing K_q whenever query q is announced. The algorithm has the same initial randomization law in every model. Conditional on any common past, its next query has the same law, and the conditional transcript law is K_q in every model. Induction yields one common law $\mathbb{L}$ for all finite histories. The same stochastic-kernel construction yields a common infinite history law and hence common stopping and output laws.

Let $A_\theta = \{\text{a finite policy output belongs to } \mathcal{U}_a(\theta)\}$. Their intersection is empty, including on non-termination outcomes. By the overlap multiplicity definition, $\sum_{\theta\in\Xi} \mathbf{1}_{A_\theta} \le m$ pointwise, so some model has $\mathbb{L}(A_\theta) \le m/k$.

For the certificate claim let $\mathcal{E}_\theta = \{\theta \in C_n \text{ for all } n\}$. Coverage under each model is a statement under the same law $\mathbb{L}$, so $\mathbb{L}(\bigcap_{\theta\in\Xi} \mathcal{E}_\theta) \ge 1 - k\delta$. On this intersection, normal stopping with $U_\tau \le a$ would produce a policy in every $\mathcal{U}_a(\theta)$, a contradiction. This proves (4.6). □

C.2 Duality, sparse witnesses, and RDI

Proof of Theorem 4.2. Write $f_\theta(x) = v_\theta^\star - g_\theta(x)$. The minimum of linear functions defining g_θ is concave in x, so f_θ is convex and continuous. Continuity of A_θ and compactness of X, Y imply joint continuity of $f_\theta(x)$. For $F(x,\pi) = \int_C f_\theta(x)\pi(d\theta)$, F is convex and continuous in x, and affine and weakly continuous in π. The spaces X and $\Delta(C)$ are compact and convex. The convex minimax theorem gives

$$\min_{x\in X} \sup_{\pi\in\Delta(C)} F(x,\pi) = \sup_{\pi\in\Delta(C)} \min_{x\in X} F(x,\pi).$$

Dirac measures show that the left side is $\Gamma(C)$. The inner minimum on the right equals $\mathbb{E}_\pi v_\theta^\star - \max_x \mathbb{E}_\pi g_\theta(x)$. It is upper semicontinuous in π, being an infimum of continuous functions, so its supremum is attained.

We give a compactness-and-Helly proof of the exact witness bound. If $\Gamma(C) = 0$, any singleton has the same radius. Otherwise take nonnegative $a_j \uparrow \Gamma(C)$ with $a_j < \Gamma(C)$. The closed convex sets

$$F_{\theta,j} := \{x \in X : f_\theta(x) \le a_j\}$$

have empty total intersection. Because they are closed subsets of compact X, some finite subfamily has empty intersection. Helly's theorem in $\operatorname{aff} X$, of dimension p, reduces it to at most $p+1$ models. Thus there is W_j with $|W_j| \le p+1$ and $\Gamma(W_j) > a_j$. Pad each ordered list with repetitions to length $p+1$. Compactness of C^{p+1} supplies a convergent subsequence, with limit list defining a nonempty set $W \subseteq C$.

Joint continuity on compact $C \times X$ implies uniform convergence of the corresponding finite maxima of f_θ. Taking minima over X preserves this convergence. Hence $\Gamma(W) = \lim_j \Gamma(W_j) = \Gamma(C)$. Applying the proved dual identity to this finite W supplies π_W. Finally, if C is finite and $\pi^\star$ is least favorable,

$$\Gamma(\operatorname{supp}\pi^\star) \ge \min_x \sum_\theta \pi^\star(\theta) f_\theta(x) = \Gamma(C).$$

Inclusion gives the reverse inequality. Applying the witness argument within that support completes its reduction. □

Proof of Proposition B.3. All gaps are nonnegative, so $\mathcal{R}_q^{\mathrm{wc}} \geq 0$. A constant post-query rule can ignore the transcript and attain $\mathcal{R}$, giving $\mathcal{R}_q^{\mathrm{wc}} \leq \mathcal{R}$. These inequalities prove both endpoints. If a common law K exists for the supported models, then every post-query rule satisfies

$$\sum_i \pi_i \int f_{\theta_i}(\psi(z))K(dz) = \int \sum_i \pi_i f_{\theta_i}(\psi(z))K(dz) \geq \mathcal{R}.$$

The adversary may choose these common laws, so equality holds and RDI is zero.

For singleton channels, any measurable rule has objective

$$\int \bar{p}^q(z) \sum_i \pi_i^z f_{\theta_i}(\psi(z))\rho(dz).$$

It is bounded below by $\int \bar{p}^q(z)\mathcal{R}(\pi^z; W)\rho(dz)$. To justify the reverse inequality without assuming a posterior selection rule, take a countable dense sequence (x_j) in compact X. For any $\zeta > 0$, choose the first j for which the continuous objective at x_j is within ζ of its minimum. This index is measurable because each objective and their countable infimum are measurable. The resulting rule has objective within ζ of the displayed integral. Let $\zeta \downarrow 0$. On $\{\bar{p}^q = 0\}$ define the posterior arbitrarily. □

Proof of Proposition B.4. For every x,

$$f_{\theta_i}(x) = \max_{y \in \operatorname{ext} Y} \{v_i^\star - x^\top A_i y\}.$$

Thus g_i in (B.6) is precisely an epigraph variable for the model gap, and minimizing its weighted sum gives the parent impurity. This remains true when some weights vanish.

For finite transcripts, a post-query rule consists of one plan x_z per transcript. The variables g_{iz} bound its model gaps. For fixed $\{x_z\}$, model i's worst transcript law maximizes a linear functional over $\mathcal{P}_{\theta_i}(q)$ and attains its maximum at some $k_{i\ell}$. The constraints on t_i describe this second epigraph. The law for each model can be chosen separately, so their weighted sum gives (B.7). Maximizing the original epigraph over both the model and attacker vertex gives (B.8). Finally, minimizing a linear payoff over the attacker's realization-plan polytope exposes an optimal face. With the defender query fixed, terminal-history probabilities, and therefore transcript probabilities, are linear in the attacker's realization plan. The linear image of that face is a polytope. □

Proof of Proposition B.5. Write $s_Q = s_n^\star(Q)$ and $s_{\mathrm{acc}} = s_n^\star(\mathcal{Q}_{\mathrm{acc}})$. Support inclusion implies $s_Q \leq s_{\mathrm{acc}}$. Finite LP duality gives $t = s_Q$. For any accessible allocation Λ,

$$\min_b \int D_q(b)\Lambda(dq) \leq \sum_b w(b) \int D_q(b)\Lambda(dq) \leq \sup_q \sum_b w(b)D_q(b) \leq s_Q + \tau_{\mathrm{sep}}.$$

Taking the supremum over Λ proves the separation interval. Feasibility gives separation at least $\kappa_{\mathrm{sep}} s_Q$, which is at least the stated full-universe lower bound.

For the RDI master at threshold b_0, denote its finite-pool optimum by J_Q. Strong duality gives $J_Q = \eta - b_0 \sum_b \alpha(b)$. The pricing certificate gives the pointwise inequality $I_n(q) \leq \eta + \tau_{\mathrm{gen}} - \sum_b \alpha(b) D_q(b)$. Integrate against any accessible allocation satisfying the same threshold. Since $\alpha(b) \geq 0$, its objective is at most $\eta + \tau_{\mathrm{gen}} - b_0 \sum_b \alpha(b) = J_Q + \tau_{\mathrm{gen}}$. The finite-pool optimum is also feasible for the full problem. These two comparisons prove the claim. □

Proof of Proposition B.6. For separation pricing, each accepted column is $c = (D_q(b))_{b \in \mathcal{B}} \in [0,1]^{\mathcal{B}}$. Let (w^j, t^j) be an optimal separation dual when the jth accepted column c^j is priced. Acceptance means $\langle w^j, c^j \rangle > t^j + \tau_{\mathrm{sep}}$. Once c^j is added to the pool, every later dual (w^k, t^k) satisfies $\langle w^k, c^j \rangle \leq t^k$ for $k > j$. If infinitely many columns were accepted, compactness of $[0,1]^{\mathcal{B}} \times \Delta(\mathcal{B}) \times [0,1]$ gives a convergent subsequence $(c^{j_i}, w^{j_i}, t^{j_i}) \to (\bar{c}, \bar{w}, \bar{t})$. Passing to the limit in the later feasibility inequalities gives $\langle \bar{w}, \bar{c} \rangle \leq \bar{t}$, whereas the acceptance inequalities give $\langle \bar{w}, \bar{c} \rangle \geq \bar{t} + \tau_{\mathrm{sep}}$, a contradiction.

For RDI pricing, each accepted column is $(I(q), (D_q(b))_{b \in \mathcal{B}}) \in [0, \bar{H}] \times [0,1]^{\mathcal{B}}$ with $\bar{H} := \mathcal{R}(\pi; W)$ by Proposition B.3. It remains only to keep the optimal RDI dual iterates in a compact set. If the

threshold base s is zero, the stated convention gives $\alpha = 0$ and $\eta = \max_q I(q) \le \bar{H}$. If $s > 0$, let λ^{sep} be an optimal separation allocation for the pool on which s was computed, extended by zero to later columns. It is feasible for the RDI master with slack at least $\omega := (1 - \kappa_{\mathrm{sep}})s > 0$ in every separation constraint. For an optimal RDI dual, strong duality and evaluation at λ^{sep} give

$$V = \max_{\lambda \in \Delta(Q)} \left[\sum_q \lambda(q) I(q) + \sum_b \alpha(b) \left(\sum_q \lambda(q) D_q(b) - \kappa_{\mathrm{sep}} s \right) \right] \ge \omega \sum_b \alpha(b),$$

where $V \le \bar{H}$. Hence $\|\alpha\|_1 \le \bar{H}/\omega$. Choosing the smallest feasible η in an optimal dual gives $\eta \le \bar{H} + \|\alpha\|_1$. Because the threshold $b_0 = \kappa_{\mathrm{sep}} s$ stays fixed for the whole RDI loop and λ^{sep} stays feasible as RDI columns are added, these bounds hold uniformly for all RDI dual solutions of the round. The same subsequence argument as above, now applied to the RDI acceptance inequalities $I^j + \langle \alpha^j, D^j \rangle > \eta^j + \tau_{\mathrm{gen}}$ and the later feasibility inequalities $I^j + \langle \alpha^k, D^j \rangle \le \eta^k$, rules out infinitely many accepted RDI columns. □

Proof of Proposition B.8. For part 1, a face F of Y is optimal for the attacker under model θ and diagnostic plan x exactly when $A_\theta^\top x$ belongs to the corresponding cone in the normal fan of Y. The normal fan is finite and polyhedral, and the map $x \mapsto A_\theta^\top x$ is linear. Linear preimages of those cones therefore give a finite polyhedral partition. Refining over the finite witness set keeps the partition finite.

For part 2, full logging makes a transcript value identify the terminal history and reward label. For $y \in F_\theta$ and $y' \in F_{\theta'}$, the common-support contribution to the Hellinger affinity is

$$x_{\sigma_D(h)} \sqrt{\pi_\theta^{\mathrm{ch}}(h) \pi_{\theta'}^{\mathrm{ch}}(h)} \sqrt{y_{\sigma_A(h)} y'_{\sigma_A(h)}},$$

and histories with different reward labels contribute zero because their recorded transcripts are disjoint. This gives the displayed formula for Aff_b. Introducing variables t_h with rotated cone constraints $t_h^2 \le y_{\sigma_A(h)} y'_{\sigma_A(h)}$ yields a second-order-cone representation; at an optimum, positive objective coefficients make $t_h = \sqrt{y_{\sigma_A(h)} y'_{\sigma_A(h)}}$ whenever the term matters. For fixed feasible (y, y') the objective is linear in x, and the feasible product is compact, so Aff_b is the support function of a compact set. It is therefore convex and continuous. Since $d^2 = 1 - \sup(\text{Hellinger affinity})$ under the paper's normalization, the identity $d^2_{q_x}(b) = 1 - \mathsf{Aff}_b(x)$ holds inside the cell.

For part 3, within a relative cell interior the frozen and true best-response faces coincide. Continuity extends the cell value to its closure. At a boundary point, the true best-response face contains each adjacent frozen face, which can only enlarge the admissible channel sets, increase the supremal affinity, and decrease d^2. Therefore the global separation-pricing supremum is the maximum of the displayed closed-cell convex minimization values. Standard conic dual certificates provide the claimed tolerance certificate.

For part 4, if two distinct extreme realization plans y, y' are both optimal at x, then

$$x^\top A_\theta y = x^\top A_\theta y'.$$

When the linear functional $x' \mapsto x'^\top A_\theta (y - y')$ is not identically zero on the accessible affine hull, this equality defines a proper hyperplane of $\mathrm{aff}\, X_{\mathrm{acc}}$. It is one of the equalization hyperplanes in the normal-fan refinement. If the functional is identically zero, or if the observation is merely that a single realization plan has several nonzero sequence coordinates, no boundary conclusion follows; these degenerate cases are excluded by the statement. □

C.3 Information times and the lower bound

Proof of Proposition B.11. Each answer set is a nonempty compact convex polytope: f_θ is a finite maximum of affine functions and is zero at a maximin plan. For any nonnegative extended-real costs c_θ, write

$$L = \sup_{x \in \mathcal{U}_\varepsilon(\theta^\star)} \inf_{\theta \in \mathrm{Alt}_\varepsilon(x)} c_\theta, \quad R = \min_{\Xi \in \mathfrak{C}_\varepsilon(\theta^\star)} \max_{\theta \in \Xi \setminus \{\theta^\star\}} c_\theta.$$

Every x feasible for the true model violates at least one other model in any conflict Ξ. Therefore its inner infimum is at most $\max_{\Xi \setminus \{\theta^\star\}} c_\theta$, proving $L \le R$.

For the reverse inequality let $t < R$ be real and put $B_t = \{\theta \neq \theta^\star : c_\theta \leq t\}$. If $\{\theta^\star\} \cup B_t$ were incompatible, the convex sets $\mathcal{U}_\varepsilon(\theta^\star) \cap \mathcal{U}_\varepsilon(\theta)$, $\theta \in B_t$, would have empty intersection. Helly's theorem would supply at most $p+1$ such sets. Together with $\theta^\star$, their models would form a member of $\mathfrak{C}_\varepsilon(\theta^\star)$ with maximum alternative cost at most t, a contradiction. Hence some true-model answer satisfies every model in B_t. All its alternatives have cost strictly greater than t. Finiteness of Θ makes their infimum greater than t, with value $+\infty$ if there are no alternatives. Letting $t \uparrow R$ proves $L \geq R$, also for $R = +\infty$. The two choices of divergences give the stated identities.

If $D_\varepsilon(s) = 0$, apply the identity to a full-support uniform allocation. Some finite conflict has all its nonnegative query divergences zero. The KL infima are attained by compactness and lower semicontinuity, and KL vanishes only between equal laws. Thus (B.16) holds. Conversely that condition makes one conflict's maximum cost zero under every allocation, so $D_\varepsilon(s) = 0$. □

Proof of Theorem B.12. Fix $(\theta^\star, s)$ and denote its law by $\mathbb{P}$. If $\mathbb{E}\tau = \infty$ or $D_\varepsilon(s) = \infty$, the claimed bound is immediate. Otherwise the conflict collection is nonempty and τ is finite $\mathbb{P}$-almost surely. Fix Ξ in that collection and set $k = |\Xi| - 1 \leq p + 1$. Define

$$A = \{\tau < \infty, \widehat{x} \in \mathcal{U}_\varepsilon(\theta^\star)\}, \quad B_\theta = A \cap \{\widehat{x} \notin \mathcal{U}_\varepsilon(\theta)\}.$$

Incompatibility gives $A \subseteq \bigcup_{\theta \in \Xi \setminus \{\theta^\star\}} B_\theta$. Since $\mathbb{P}(A) \geq 1 - \delta$, some θ_0 has $\mathbb{P}(B_{\theta_0}) \geq (1-\delta)/k$. For each query choose $K_q \in \mathcal{P}_{\theta_0}(q)$ attaining $J_q(s, \theta_0)$. These laws form an admissible stationary selector. Let $\mathbb{Q}$ be its calibration law at θ_0. Correctness gives $\mathbb{Q}(B_{\theta_0}) \leq \delta$.

For an integer T, set $B_{\theta_0,T} = B_{\theta_0} \cap \{\tau \leq T\}$. This event is measurable at $\tau \wedge T$. The stopped KL chain rule and data processing give

$$\sum_q \mathbb{E}_{\mathbb{P}} N_q(\tau \wedge T)\, \mathrm{KL}(K_q^s \| K_q) \geq \mathrm{kl}\big(\mathbb{P}(B_{\theta_0,T}), \mathbb{Q}(B_{\theta_0,T})\big).$$

The algorithm's query and randomization kernels are identical under both models, so only transcript laws contribute. This finite-time inequality allows infinite divergences; zero expected query counts contribute zero. Monotone convergence of counts and lower semicontinuity of binary KL, as $T \to \infty$, yield

$$\sum_q \mathbb{E} N_q(\tau) J_q(s, \theta_0) \geq \mathrm{kl}\big(\mathbb{P}(B_{\theta_0}), \mathbb{Q}(B_{\theta_0})\big) \geq \mathrm{kl}((1-\delta)/k, \delta).$$

The last inequality uses $(1-\delta)/k > \delta$ and monotonicity of binary KL on these ranges. Its positive right side rules out $\mathbb{E}\tau = 0$. Normalize expected counts to form $\bar{\lambda}$ and maximize over the alternatives in Ξ. This proves the more precise inequality. Also $\mathrm{kl}((1-\delta)/k, \delta) \geq \mathrm{kl}((1-\delta)/(p+1), \delta)$. Minimizing over conflicts and using (B.15) gives

$$\mathrm{kl}((1-\delta)/(p+1), \delta) \leq \mathbb{E}\tau\, D_\varepsilon(s, \bar{\lambda}) \leq \mathbb{E}\tau\, D_\varepsilon(s).$$

If the latter information value is zero this contradicts finite expected stopping; otherwise divide by it. Taking the supremum over s proves the second bound. □

Proof of Corollary B.14. A joint common law on a conflict defines a stationary selector satisfying (B.16), so $D_\varepsilon(s) = 0$. The comparison of the two information times makes both infinite. The success obstruction follows from Theorem 4.3.

For the converse, assume common support. Compactness and finiteness imply a positive uniform lower bound on every channel probability. KL is continuous on these compact domains. Minimization over a compact alternative-channel set shows that $J_q(s, \theta)$ is continuous in s. The Helly formula is a finite minimum of finite maxima of continuous functions of (s, λ). Maximization over the compact allocation simplex preserves continuity. Thus $D_\varepsilon(s)$ attains its minimum on compact $\mathfrak{S}(\theta^\star)$. An empty conflict collection gives time zero, so if the time is infinite the attained minimum must be zero. Proposition B.11 then supplies a joint conflict.

For the structural test, any displayed Ξ_θ is itself a joint obstruction. Conversely, if a larger joint obstruction contains θ, intersect its answer sets with $\mathcal{U}_a(\theta)$ and apply the anchored Helly argument. It reduces the alternatives to at most $p+1$ while preserving incompatibility; the common-law property is preserved under taking subsets. The equivalence with zero $D_a(s)$ follows from Proposition B.11, without requiring common support. These are pointwise structural facts about C.

For the comparison in Remark B.15, a simultaneous minimizer makes the two divergence arrays identical. With two models choose a joint KL minimizer separately at each query and take its true-side law; there is only one nontrivial alternative. Singleton true channels also fix that law automatically. □

C.4 E-PROCESSES AND STOPPING BOUNDS

Proof of Theorem B.16. Fix (θ, θ', q) and a minimizer $(P^\star, K^\star)$. Common support makes KL differentiable on all feasible line segments. For $K \in \mathcal{P}_\theta(q)$, optimality in the second argument gives

$$0 \le \frac{d}{du}\,\mathrm{KL}(P^\star \| (1-u)K^\star + uK)\Big|_{u=0+} = 1 - \sum_z K(z)\frac{P^\star(z)}{K^\star(z)}.$$

This proves (B.22). For $P \in \mathcal{P}_{\theta'}(q)$, optimality in the first argument similarly gives

$$\begin{aligned} 0 &\le \frac{d}{du}\,\mathrm{KL}((1-u)P^\star + uP \| K^\star)\Big|_{u=0+} \\ &= \sum_z (P(z) - P^\star(z))\left\{\log\frac{P^\star(z)}{K^\star(z)} + 1\right\} \\ &= \mathbb{E}_P \log e_q^{\theta,\theta'}(Z) - \mathrm{KL}(P^\star \| K^\star). \end{aligned}$$

Equality holds at $P = P^\star$, proving (B.23).

At the true tested model, admissibility and conditional e-validity give

$$\mathbb{E}_{\theta^\star}[e_{q_t}^{\theta^\star,\theta'}(Z_t) \mid \mathcal{G}_t] \le 1.$$

By the tower property, each $E_n(\theta^\star;\theta')$ is a nonnegative supermartingale with initial value one relative to $(\mathcal{F}_n)$. The fixed-weight mixture $\bar{E}_n(\theta^\star)$ is also such a supermartingale. The maximal inequality for nonnegative supermartingales yields

$$\mathbb{P}_{\theta^\star}\{\exists n : \bar{E}_n(\theta^\star) \ge 1/\delta\} \le \delta.$$

This is exactly (B.24); no union bound over the unknown true model is needed. □

Proof of Theorem B.18. For $\theta \ne \theta^\star$, set $\ell_t(\theta) = \log e_{q_t}^{\theta,\theta^\star}(Z_t)$. The fixed conditional allocation and (B.23) imply

$$\mathbb{E}[\ell_t(\theta) \mid \mathcal{F}_{t-1}] = \mathbb{E}\{\mathbb{E}[\ell_t(\theta) \mid \mathcal{G}_t] \mid \mathcal{F}_{t-1}\} \ge c_\lambda(\theta).$$

The centered increments have absolute value at most $2B_e$. Azuma–Hoeffding, followed by a union bound using $\sum_{n\ge1}[n(n+1)]^{-1} = 1$, gives an event $\mathcal{E}_2$ of probability at least $1-\delta'$ on which

$$\log E_n(\theta;\theta^\star) \ge n c_\lambda(\theta) - b_n^e(\delta') \quad \text{for every } \theta \ne \theta^\star,\ n \ge 1.$$

The individual failure probability is at most $\delta'/[Mn(n+1)]$. If $B_e = 0$ the assertion is deterministic. Intersect with the coverage event $\mathcal{E}_1 = \{\theta^\star \in C_n^e \text{ for all } n\}$. At any finite n_θ, $E_{n_\theta}(\theta;\theta^\star) \ge (M-1)/\delta$, so $\bar{E}_{n_\theta}(\theta) \ge 1/\delta$. The running intersection permanently removes θ.

On $\mathcal{E}_1 \cap \mathcal{E}_2$, by time $n^\star$ at least one alternative of every anchored conflict has been removed. If normal stopping had not already occurred and $C_{n^\star}^e$ were incompatible, it would contain $\theta^\star$, and the anchored Helly argument would give a conflict wholly inside it. This contradicts those removals. Thus the set is nonempty and compatible, and the exact solver stops. On coverage, its output is correct. If the conflict collection is empty, the same Helly argument shows that Θ is compatible at time zero.

For fixed δ', $b_n^e(\delta') = O(\sqrt{n\log n}) = o(n)$. When $c_\lambda(\theta) > 0$, comparison with $h_\delta/c_\lambda(\theta)$ and with $(1+\zeta)h_\delta/c_\lambda(\theta)$ for each fixed $\zeta > 0$ gives

$$n_\theta = (1+o(1))\frac{\log(1/\delta)}{c_\lambda(\theta)}.$$

When $c_\lambda(\theta) = 0$, $n_\theta = \infty$. Positive $D_\varepsilon^{\mathrm{wc}}(\theta^\star,\lambda)$ means that each conflict has an alternative of positive cost. Finiteness and the Helly formula therefore imply

$$\max_{\Xi} \min_{\theta\in\Xi\setminus\{\theta^\star\}} \frac{1}{c_\lambda(\theta)} = \frac{1}{\min_\Xi \max_{\theta\in\Xi\setminus\{\theta^\star\}} c_\lambda(\theta)} = \frac{1}{D_\varepsilon^{\mathrm{wc}}(\theta^\star,\lambda)}.$$

This proves the asymptotic bound. With a nonempty conflict collection the information value is a continuous function of λ on its compact simplex, hence has a maximizer.

Finally, if its supremum is positive, every conflict has some alternative with positive divergence on some query. A full-support allocation consequently gives positive cost to at least one alternative in each conflict. Under the hypothesis for every true model, $n^\star(\delta, \delta')$ is finite for every $\delta' > 0$. For each $\delta' > 0$, successful finite stopping has probability at least $1 - \delta - \delta'$. That event does not depend on δ'; letting $\delta' \downarrow 0$ proves (B.17). □

Proof of Corollary B.19. Keep $\ell_t(\theta)$ as above, but center after the query is selected:

$$m_t(\theta) = \mathbb{E}[\ell_t(\theta) \mid \mathcal{G}_t], \qquad D_t(\theta) = \ell_t(\theta) - m_t(\theta).$$

The tower property gives $\mathbb{E}[D_t(\theta) \mid \mathcal{F}_{t-1}] = 0$, and $|D_t(\theta)| \le 2B_e$. The same Azuma–Hoeffding union bound gives $\sum_{t=1}^{n} D_t(\theta) \ge -b_n^e(\delta')$ simultaneously. Conditional log-growth satisfies $m_t(\theta) \ge J_{q_t}^{\mathrm{wc}}(\theta^\star, \theta)$, which proves (B.28). This centering is valid even when q_t is not $\mathcal{F}_{t-1}$-measurable.

On its intersection with coverage, the premise of (B.29) supplies, in every conflict, an alternative whose pairwise log process is at least h_δ. Its mixture crosses $1/\delta$ and removes that alternative. The nonempty remaining set is compatible by the anchored Helly argument, so exact normal stopping has occurred. □

Proof of Proposition B.20. For every admissible null law, the expectation of the normalized factor is at most $c_t^{\theta,\theta'} / \max\{1, c_t^{\theta,\theta'}\} \le 1$. Conditioning first on $\mathcal{G}_t$ and then on $\mathcal{F}_{t-1}$ proves the supermartingale property of the product. Fixed nonnegative weights summing to one preserve it, so the same maximal inequality proves coverage. For growth, normalization subtracts $\log \max\{1, c_t^{\theta,\theta'}\}$ from every log increment. Without a certified bound on this loss and on the approximate factor's alternative-side log expectation, the exact KL rate cannot be substituted. □

Proof of Corollary B.21. Use the conflict argument at accuracy ε_0. When $\Gamma(C_n) \le \varepsilon_0$,

$$U_n + 2\varepsilon_{\mathrm{span}} \le \Gamma(C_n) + \eta_{\mathrm{comp}} + 2\varepsilon_{\mathrm{span}} \le \varepsilon.$$

Theorem 4.1 proves deployment safety. The statistical proof still requires the actual diagnostic experiment to satisfy its coverage assumptions; the value-distance inequality was used only for deployment transfer. □

The zero-error rule in Remark B.22 also has a direct proof. Conditional on any history and selected query, an admissible true law assigns zero probability to the set of symbols to which every true-model law assigns probability zero. Thus the probability of eliminating the true model at that round is zero. A countable union over rounds proves simultaneous coverage with probability one.

Proof of Proposition B.23. For a fixed candidate, introduce a scalar ρ and maximize it subject to

$$\lambda \in \Delta(Q),$$
$$\sum_q \lambda(q) J_q^{\mathrm{wc}}(\check{\theta}_n, \theta) \ge \rho \qquad (\theta \in \mathcal{A}_n^{\mathrm{ct}}).$$

Common support makes all coefficients finite. The dual minimizes t over $w \in \Delta(\mathcal{A}_n^{\mathrm{ct}})$ subject to

$$\sum_\theta w(\theta) J_q^{\mathrm{wc}}(\check{\theta}_n, \theta) \le t \qquad (q \in Q).$$

Hence the pricing objective is exactly the one stated. The separation proof of Proposition B.5 applies verbatim with these coefficients.

For any two laws, Jensen's inequality gives, with extended-real conventions,

$$\mathrm{KL}(P\|K) \ge -2 \log \int \sqrt{dP\, dK} = -2 \log(1 - H^2(P, K)) \ge 2H^2(P, K).$$

Taking infima over the same two channel sets proves the comparison. If the alternative set is certified empty, the candidate has gap at most a throughout the confidence set. The asserted deployment test then follows from the same coverage-and-transfer argument as Theorem 4.1. If certification adds a residual of at most η_{comp}, the sufficient margin is $a + \eta_{\mathrm{comp}} + 2\varepsilon_{\mathrm{span}} \le \varepsilon$. □

C.5 Linear moments and exploration

Proof of Proposition B.25. By definition, every exact best response attains $m_q(\theta)$; every κ-best response has expected payoff between this minimum and this minimum plus κ. Randomized choices of responses preserve the same equality or interval. Since $R \in [0, HR_{\max}]$, conditional Hoeffding's lemma gives

$$\mathbb{E}\Big[e^{u(R-\mathbb{E}[R|\mathcal{G}_t])} \mid \mathcal{G}_t\Big] \le \exp(u^2 H^2 R_{\max}^2/8).$$

If the minimum payoff has the asserted affine representation, take $b_t = \mathbb{E}[R \mid \mathcal{G}_t] - a_{q_t} - M_{q_t}\theta^\star$ and center the remaining noise. This bias is zero for exact responses and lies in $[0, \kappa]$ otherwise.

For an affine sequence-form payoff matrix,

$$m_q(\theta) = \min_{y \in \operatorname{ext} Y} x_q^\top A_\theta^q y$$

is a finite minimum of affine functions, so it is concave and piecewise affine. If one plan minimizes for all models, $m_q(\theta) = x_q^\top A_\theta^q \bar{y}_q$ on the model class and is affine there. □

Proof of Theorem B.26. Write $M_t = M_{q_t}$, $\lambda = \lambda_{\mathrm{reg}}$, $S_n = \sum_{t=1}^n M_t^\top \xi_t$, and $A_n = \sum_{t=1}^n M_t^\top M_t = V_n - \lambda I_d$. For each fixed $v \in \mathbb{R}^d$, the conditional moment bound, first given $\mathcal{G}_t$ and then by the tower property, shows that

$$L_n(v) = \exp\left(\frac{v^\top S_n}{\sigma} - \frac{1}{2} v^\top A_n v\right)$$

is a nonnegative supermartingale with initial value one. Mixing with $v \sim N(0, \lambda^{-1} I_d)$ and evaluating the Gaussian integral gives another such supermartingale,

$$L_n = \frac{\det(\lambda I_d)^{1/2}}{\det(V_n)^{1/2}} \exp\left(\frac{\|S_n\|_{V_n^{-1}}^2}{2\sigma^2}\right).$$

The maximal inequality implies that, with probability at least $1 - \delta$, simultaneously for every n,

$$\|S_n\|_{V_n^{-1}} \le \sigma \sqrt{2 \log \frac{\det(V_n)^{1/2}}{\det(\lambda I_d)^{1/2} \delta}}.$$

This derivation permits vector observations and freshly randomized adaptive queries without imposing independence within a vector.

Let $B_n = \sum_{t=1}^n M_t^\top b_t$. Stack the M_t into a matrix $\mathbf{M}_n$ and the b_t into a vector $\mathbf{b}_n$. Since every eigenvalue of $\mathbf{M}_n(\lambda I + \mathbf{M}_n^\top \mathbf{M}_n)^{-1}\mathbf{M}_n^\top$ lies in $[0, 1]$,

$$\|B_n\|_{V_n^{-1}}^2 \le \|\mathbf{b}_n\|_2^2 \le n b^2.$$

This bound is pathwise and allows selector-dependent biases. The ridge equation is $V_n(\widehat{\theta}_n - \theta^\star) = S_n + B_n - \lambda\theta^\star$. Also $\|\lambda\theta^\star\|_{V_n^{-1}} \le \sqrt{\lambda} B_\theta$. Combining these facts proves simultaneous containment with radius $r_n(\delta)$.

For two points in the ellipsoid, their V_n-distance is at most $2r_n(\delta)$, and their Euclidean distance is therefore at most $2r_n(\delta)/\sqrt{\lambda_{\min}(V_n)}$. Let $x^\star$ be optimal at $\theta^\star$. On coverage,

$$\Gamma(C_n^{\mathrm{lin}}) \le \sup_{\theta \in C_n^{\mathrm{lin}}} \mathrm{Gap}_\theta(x^\star) \le 2L_G \operatorname{diam}_2(C_n^{\mathrm{lin}}).$$

The solver upper residual and span transfer yield (B.33). Under excitation, $\lambda_{\min}(V_n) \ge \lambda + n\alpha^2$ and $b\sqrt{n}/\sqrt{\lambda + n\alpha^2} \le b/\alpha$.

To verify the limitation of bias-only assumptions, consider the one-room deployment game with rewards $(\theta, 1-\theta)$, $\theta \in \{1/2 - b, 1/2 + b\}$, $0 < b \le 1/2$, and $M_q = 1$. Choose diagnostic biases $+b$ and $-b$, respectively. Both models then produce the same $N(1/2, \sigma^2)$ observations. Their deployment gaps are $2bx$ and $2b(1-x)$, whose robust radius is b. No method can uniformly resolve every tolerance below b with success greater than $1/2$ in both models. This example shows one possible obstruction under the generic bounded-bias model. It does not give a universal lower bound with constant $4L_G/\alpha$. □

Proof of Lemma B.27. Let $X_t = M_{q_t}^\top M_{q_t}$ and $\bar{X}_t = \mathbb{E}[X_t \mid \mathcal{F}_{t-1}]$. The independent forced component and positivity of the remaining component give $\bar{X}_t \succeq \gamma\alpha_0^2 I_d$. Also $0 \preceq X_t, \bar{X}_t \preceq L_M^2 I_d$. Thus $Z_t = X_t - \bar{X}_t$ is a self-adjoint martingale difference, $\|Z_t\|_{\mathrm{op}} \le L_M^2$, and its accumulated conditional matrix variance is bounded by $nL_M^4 I_d$. The matrix Freedman inequality yields a universal c such that

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_{t=1}^{n} Z_t\right) < -cL_M^2(\sqrt{nx} + x)\right\} \le de^{-x}.$$

Use $x = \ell_n$ and union bound over $n \ge 1$. The total failure probability is at most $\sum_{n\ge1} \delta'/[2(n+1)^2] < \delta'$. Weyl's inequality and $\sum_t \bar{X}_t \succeq n\gamma\alpha_0^2 I_d$ prove the stated bound. For the sufficient threshold put $k = L_M^4/(\gamma^2\alpha_0^4) \ge 1$. Each of the two deviation terms is at most $n\gamma\alpha_0^2/4$, and hence their sum is at most $n\gamma\alpha_0^2/2$, for $n \ge Ck\log(edk/\delta')$ with sufficiently large universal C. This follows from the elementary bound $\log(n+1) = o(n/k)$ and, quantitatively, by substituting $n = Ck\log(edk/\delta')$; thereafter $n/\log(n+1)$ increases. □

C.6 Dimension separation and strategic censoring

Proof of Proposition B.28. Write $x = \mu(a_1 \mid s_1)$ and $\Delta = \theta_1 - 1/2$. Direct evaluation gives

$$\mathrm{Gap}_\theta(x) = \begin{cases} 2\Delta(1-x), & \Delta \ge 0, \\ 2|\Delta|x, & \Delta < 0. \end{cases}$$

Thus no nuisance coordinate affects deployment. After $n \ge 1$ observations of q_1, define

$$\bar{Y}_n = \frac{1}{n}\sum_{t=1}^{n} Y_t, \qquad w_n = \sigma\sqrt{\frac{2}{n}\log\frac{\pi^2(n+1)^2}{3\delta}}, \qquad I_{1,n} = [\bar{Y}_n - w_n, \bar{Y}_n + w_n] \cap [0,1].$$

Use $I_{1,0} = [0,1]$. A Gaussian tail bound gives

$$\mathbb{P}_\theta(|\bar{Y}_n - \theta_1| > w_n) \le \frac{6\delta}{\pi^2(n+1)^2}.$$

Summing over $n \ge 1$ proves simultaneous coverage at least $1-\delta$. For analysis the iid q_1 sequence may be generated in advance. Empty intervals off the coverage event count as failure statuses and carry no policy certificate.

Let a nonempty interval be $[\ell, u]$. If it is on one side of $1/2$, one pure action is optimal throughout it and its robust radius is zero. Otherwise write $w_+ = u - 1/2 > 0$ and $w_- = 1/2 - \ell > 0$. The two endpoints attain the worst gap of every policy, and

$$\Gamma(C_n) = \min_{x\in[0,1]} \max\{2w_+(1-x),\, 2w_- x\} = \frac{2w_+w_-}{w_+ + w_-} \le \frac{w_+ + w_-}{2} \le w_n.$$

The canonical endpoint models set all nuisance coordinates to $1/2$. A least-favorable prior gives upper-endpoint weight $\pi_+ = w_-/(w_+ + w_-)$ and lower-endpoint weight $\pi_- = w_+/(w_+ + w_-)$. Put $c = \pi_+ w_+ = \pi_- w_- > 0$ and let P_+, P_- be their q_1 Gaussian laws. The parent impurity is $2c$. Pointwise minimization of the post-query expected gap gives

$$\mathcal{R}_{q_1}^{\mathrm{wc}}(\pi; W) = 2c\int \min\{dP_+, dP_-\} = 2c[1 - \mathrm{TV}(P_+, P_-)],$$

so $\mathrm{RDI}_{q_1} = 2c\,\mathrm{TV}(P_+, P_-) > 0$. Every q_j, $j \ge 2$, has identical laws at the canonical endpoints and hence zero RDI and zero pair separation. Only q_1 has positive pair separation. Before stopping, the endpoint pair is conflicting at tolerance ε. Exact Tier I allocation therefore chooses only q_1 for every $\kappa_{\mathrm{sep}} \in [0,1]$. If a Tier II pool lacks q_1, its separation value is zero, while the full pricing supremum is strictly positive and uniquely attained by q_1. With zero acceptance tolerance, pricing adds it without using a diagnostic episode; thereafter the same allocation argument applies. With positive acceptance tolerances, the same conclusion requires the certified separation or RDI improvement margin at that round to exceed the relevant tolerance. We make no claim for arbitrary fixed positive tolerances.

On the coverage event, the procedure stops by the deterministic first $n \geq 1$ such that $w_n \leq \varepsilon$. This is the first integer satisfying

$$n \geq \frac{2\sigma^2}{\varepsilon^2} \log \frac{\pi^2 (n+1)^2}{3\delta}.$$

For $a > 0$ and $c_0 > 1$, the first positive integer satisfying $n \geq a \log(c_0(n+1)^2)$ is bounded by $C[1 + a\{\log c_0 + \log(e+a)\}]$. To see this, set the right side as a candidate: its logarithm is bounded by a constant multiple of $\log c_0 + \log(e+a)$, and increasing the universal C absorbs that multiple and integer rounding. Substitution of $a = 2\sigma^2/\varepsilon^2$ gives the stated bound. The additive one covers the small-noise regime. Coverage and the robust certificate prove the success assertion.

For full identification define $\theta^{(j)} = \theta^0 + 2re_j$, $j \in [d]$, which belongs to $[0,1]^d$. Let $A = \{\theta^0 \in S_\tau, \mathrm{diam}_2(S_\tau) \leq r\}$. Uniform validity gives $\mathbb{P}_{\theta^0}(A) \geq 1 - \delta$. Since the two parameters are $2r > r$ apart, event A is disjoint from successful coverage of $\theta^{(j)}$, so $\mathbb{P}_{\theta^{(j)}}(A) \leq \delta$. Only query q_j distinguishes these models, and its KL divergence is

$$\mathrm{KL}\{N(1/2, \sigma^2) \| N(1/2 + 2r, \sigma^2)\} = 2r^2/\sigma^2.$$

The stopped change-of-measure argument used in Theorem B.12 therefore gives $\mathbb{E}_{\theta^0} N_j(\tau)\, 2r^2/\sigma^2 \geq \mathrm{kl}(1-\delta, \delta)$. If the expected stopping time is infinite the conclusion is automatic; otherwise sum these inequalities and use $\sum_j N_j(\tau) = \tau$. This proves (B.34). □

Proof of Proposition B.29. The worst-case policy values and game values are

$$g_0(x) = \min\{x, 1-x\}, \quad v_0^\star = 1/2, \qquad g_1(x) = \min\{3x, 1-x\}, \quad v_1^\star = 3/4.$$

Solving $g_i(x) \geq v_i^\star - a$ gives the displayed intervals. They are disjoint exactly when $1/4 + a < 1/2 - a$, that is, $a < 1/8$. At $x = 3/8$ both gaps are $1/8$; no smaller robust radius is possible because the answer sets for every smaller tolerance are disjoint. Hence $\Gamma(\{0,1\}) = 1/8$.

At $q_{1/2}$ both columns are best responses in model zero, whereas column two is the unique best response in model one. Choosing column two in both produces the common law

$$\mathbb{P}\{(A,B,R) = (a_1, b_2, 0)\} = 1/2, \qquad \mathbb{P}\{(A,B,R) = (a_2, b_2, 1)\} = 1/2.$$

For $1/4 < x < 1/2$, column one uniquely minimizes in model zero and column two uniquely minimizes in model one. The observed column therefore reveals the model, and every admissible law under model zero is singular with every admissible law under model one. Hence $d_{q_x}(0,1) = 1$. At $x = 3/8$, the expected payoffs are $3/8$ and $5/8$, giving the stated two-point affine representation.

At this separating query, model zero's (A,R) support is $\{(a_1,1),(a_2,0)\}$, whereas model one's support is $\{(a_1,0),(a_2,1)\}$. Both pairs have positive probability and the supports are disjoint. Randomizing only the recorded column cannot remove this distinction. The realized rewards at the two specified exact-response queries are zero or one; over the full continuous query class reward three can occur, so its uniform reward bound must be three. □

D Additional Guarantees and Operational Details

D.1 Finite stopping under forced exploration

Fix a finite query class and a full-support distribution λ_0. The actual sampling distribution in Algorithm 1 is $\widetilde{\lambda}_n = (1-\gamma)\lambda_n + \gamma\lambda_0$, with fixed $0 < \gamma \leq 1$. The following proof applies to RDI, PS-EASI, and the plug-in rule, because their exploitation distributions are predictable.

Set $\varepsilon_0 = \varepsilon - 2\varepsilon_{\mathrm{span}} - \eta_{\mathrm{comp}} > 0$. An anchored conflict is an incompatible subset containing $\theta^\star$, at most $p+1$ additional models being needed; write $\mathfrak{C}_{\varepsilon_0}(\theta^\star)$ for this finite collection (Appendix B.4, Proposition B.11).

Theorem D.1 (Safety and finite stopping with forced exploration)**.** *Assume the finite common-support channel conditions, the span/solver bounds, and the pairwise confidence construction. Run Algorithm 1 with $q_{n+1} \sim (1-\gamma)\lambda_n + \gamma\lambda_0$, where λ_0 has full support and $0 < \gamma \leq 1$. Assume each round's oracles finish, no administrative cap is imposed, and no optional obstruction report terminates the run early. Let $c_0(\theta) = \sum_q \lambda_0(q) J_q^{\mathrm{wc}}(\theta^\star, \theta)$ and $n_\theta = \min\{n \geq 1 :$*

$n\gamma c_0(\theta) - b_n^e(\delta') \geq h_\delta\}$, *where* $h_\delta = \log((M-1)/\delta)$, $b_n^e(\delta') = 2B_e\sqrt{2n\log(Mn(n+1)/\delta')}$, *and* $B_e = \max|\log e_q^{\theta,\theta'}(z)|$. *Set* $n^\star = \max_{\Xi\in\mathfrak{C}_{\varepsilon_0}(\theta^\star)}\min_{\theta\in\Xi\setminus\{\theta^\star\}} n_\theta$. *If* $n^\star < \infty$, *then under every admissible history-dependent selector,* $\mathbb{P}\{\tau \leq n^\star, \mathrm{Gap}_{G^\dagger}(\mu_\tau) \leq \varepsilon\} \geq 1-\delta-\delta'$. *An empty conflict collection permits stopping at zero. If* $D^{\mathrm{wc}}_{\varepsilon_0}(\theta,\lambda_0) > 0$ *for every model, successful finite output has probability at least* $1-\delta$ *under every model, and for fixed* δ' *and* $\delta \downarrow 0$ *the bound's leading coefficient is at most* $1/[\gamma D^{\mathrm{wc}}_{\varepsilon_0}(\theta^\star,\lambda_0)]$.

Proof of Theorem D.1. For each false null θ, write $L_t = \log e_{q_t}^{\theta,\theta^\star}(Z_t)$. Conditioning before the randomized query is drawn gives

$$\mathbb{E}[L_t \mid \mathcal{F}_{t-1}] \geq \sum_q \widetilde{\lambda}_{t-1}(q) J_q^{\mathrm{wc}}(\theta^\star,\theta) \geq \gamma c_0(\theta),$$

because every exact worst-case KL growth is nonnegative. The centered increments have absolute value at most $2B_e$. The same union bound as in Theorem B.18 therefore yields $\log E_n(\theta;\theta^\star) \geq n\gamma c_0(\theta) - b_n^e(\delta')$ for every n,θ, with probability at least $1-\delta'$. At n_θ the corresponding mixture crosses $1/\delta$, and nestedness makes this elimination permanent. At $n^\star$ every anchored conflict has lost an alternative. On coverage the remaining set contains the truth and has radius at most ε_0. The solver then gives $U_n + 2\varepsilon_{\mathrm{span}} \leq \varepsilon$, so $\tau \leq n^\star$, and Theorem 4.1 gives $\mathrm{Gap}_{G^\dagger}(\mu_\tau) \leq U_\tau + 2\varepsilon_{\mathrm{span}} \leq \varepsilon$ on the coverage event. Coverage and the concentration event fail with probabilities at most δ and δ', so both hold with probability at least $1-\delta-\delta'$. If the conflict collection is empty, anchored Helly gives $\Gamma(\Theta) \leq \varepsilon_0$, so the certificate already passes at $n=0$. If $D^{\mathrm{wc}}_{\varepsilon_0}(\theta^\star,\lambda_0) > 0$, all necessary elimination times are finite. Applying the bound for every $\delta' > 0$ and letting $\delta' \downarrow 0$ proves successful finite output with probability at least $1-\delta$. For fixed δ', the concentration term is $o(n)$; the asymptotic coefficient is at most $1/[\gamma D^{\mathrm{wc}}_{\varepsilon_0}(\theta^\star,\lambda_0)]$. This argument uses the actual randomized sampling law; no fixed allocation is substituted for the adaptive algorithm. □

Corollary D.2 (What a separation floor guarantees)**.** *Consider a block of rounds on which a pair* $(\theta^\star,\theta)$ *is protected and the exploitation allocation satisfies* $\sum_q \lambda_n(q) d_q^2(\theta^\star,\theta) \geq \kappa_{\mathrm{sep}} s_0 > 0$. *Under the finite common-support assumptions, its conditional pairwise log-growth is at least*

$$2(1-\gamma)\kappa_{\mathrm{sep}} s_0 + \gamma \sum_q \lambda_0(q) J_q^{\mathrm{wc}}(\theta^\star,\theta).$$

Starting from zero log evidence, a block of length m *eliminates the false null with probability at least* $1-\delta'$ *whenever* m *times this lower bound exceeds* $h_\delta + b_m^e(\delta')$. *For a later block subtract its initial log evidence from the required threshold. The block may be started predictably; simultaneous guarantees over several blocks require an error allocation across blocks.*

Proof. The inequality $d_q^2 \leq J_q^{\mathrm{wc}}/2$ follows from $\mathrm{KL}(P\|K) \geq -2\log\int\sqrt{dP\,dK} \geq 2H^2(P,K)$, followed by infimization. Integrate against the exploitation and exploration allocations and use conditional e-growth and the martingale deviation bound. Only pairs involving the data-generating model admit this argument; two false models need not accumulate evidence against each other. Changing witnesses also prevents applying a single-round floor to an entire unprotected future block. □

D.2 Supports, safe factors, and a hybrid bound

Common support suffices for the differentiable KL formula; a valid confidence sequence does not require it. For every null define $S_\theta(q) = \{z : \sup_{K\in\mathcal{P}_\theta(q)} K(z) > 0\}$. Intersect the running confidence set with $\{\theta : Z_t \in S_\theta(q_t)\}$. Also maintain positive, bounded, conditionally valid factors $e_q^{\theta,j}$ with any fixed initial mixture weights (the constant one always qualifies; clipped proposals are normalized as in Proposition B.20). This changes only the betting function.

Proposition D.3 (Hybrid support and evidence stopping)**.** *For each false null* θ, *choose a mixture component with weight* $w_\theta > 0$ *and a bounded log factor,* $|\log e| \leq B$. *Suppose its certified minimum conditional growth* $g_q(\theta) \geq 0$ *under every true-model law is known. Put*

$$\alpha_q(\theta) = \inf_{P\in\mathcal{P}_{\theta^\star}(q)} P(S_\theta(q)^c), \quad \alpha_0(\theta) = \sum_q \lambda_0(q)\alpha_q(\theta), \quad g_0(\theta) = \sum_q \lambda_0(q) g_q(\theta).$$

Let n_θ be the smaller of

$$\left\lceil \frac{\log(2(M-1)/\delta')}{\gamma\alpha_0(\theta)} \right\rceil \quad \textit{and} \quad \min\left\{ n : n\gamma g_0(\theta) - 2B\sqrt{2n\log\frac{2Mn(n+1)}{\delta'}} \geq \log\frac{1}{w_\theta\delta} \right\}.$$

A zero denominator or empty minimum gives infinity. Replace the elimination times in Theorem D.1 by these n_θ; its $1-\delta-\delta'$ stopping and deployment guarantee holds without common support whenever the resulting $n^\star$ is finite.

Proof. Support elimination cannot remove the truth. Conditional on any surviving history, the false null's removal hazard is at least $\gamma\alpha_0(\theta)$. Thus its support-survival probability at time n is at most $(1-\gamma\alpha_0)^n \leq e^{-n\gamma\alpha_0}$. Union bound the chosen support deadlines with budget $\delta'/2$. Apply the bounded-increment martingale argument to all chosen evidence components with the remaining $\delta'/2$. A component exceeding $1/(w_\theta\delta)$ forces the fixed mixture above $1/\delta$. For each null use whichever deadline is smaller. The same anchored-conflict argument and coverage event finish the proof. □

In particular, a singular separating query can give a support certificate after one episode even when a global log-ratio bound is arbitrarily poor.

D.3 Composite inference when pairwise factors are uninformative

For a transcript event A, a null channel polytope supplies the interval $[l,u] = [\inf_K K(A), \sup_K K(A)]$. Choose a fixed $v \in (0,1)$ and let $p_0 = \mathrm{proj}_{[l,u]}\, v$. For $0 < p_0 < 1$, define

$$e_{A,v}(z) = \begin{cases} v/p_0, & z \in A, \\ (1-v)/(1-p_0), & z \notin A. \end{cases}$$

When $v > u$ the null expectation is largest at u; when $v < l$ it is largest at l; otherwise the factor is one. Thus it is valid for the whole null polytope. Boundary cases use support elimination and the bounded normalization above. Products for fixed query/event/bet components, set to one on other queries, and a fixed mixture over components give a composite null e-process $E_n^{\mathrm{comp}}(\theta)$. We use

$$\bar{E}_n^{\mathrm{joint}}(\theta) = \tfrac{1}{2}\bar{E}_n^{\mathrm{pair}}(\theta) + \tfrac{1}{2}E_n^{\mathrm{comp}}(\theta), \qquad C_n = \bigcap_{t\leq n}\{\theta : \bar{E}_t^{\mathrm{joint}}(\theta) < 1/\delta\}$$

together with support elimination. The weights are fixed before any data are seen.

Proposition D.4 (Explicit resolution of the three-model example)**.** *For the channel sets in Remark B.17, the composite construction can eliminate a model after a deterministic finite number of uses of that query, even though all pairwise factors are one. If each null's mixture assigns weight at least $w > 0$ to its event $A_i = \{i\}$ and bet $v = 1/3$, a model is eliminated by*

$$n_0 = \left\lceil \frac{3\log(1/(w\delta))}{\log(4000/2187)} \right\rceil$$

uses of the query. At tolerance 0.6, on coverage this gives a correct deployment certificate while two models may remain.

Proof. Under null i, $P(Z=i) = .1$ for every admissible selector, so the two factor values are $10/3$ and $20/27$. For every observed symbol, the product of the three null factors is $(10/3)(20/27)^2 = 4000/2187 > 1$. After n observations, one of the three products is at least $(4000/2187)^{n/3}$. Its mixture reaches $1/\delta$ by n_0. The true null is retained with probability at least $1-\delta$, and any two-model subset has radius $1/2 < 0.6$. With a $1 + 10^{-12}$ numerical normalization margin, subtract $3\log(1+10^{-12})$ in the denominator; the resulting bound changes by less than one episode at the settings used here. □

D.4 RDI, MINIMAX INTERCHANGE, AND THE NEXT CONFIDENCE RADIUS

For finite $\mathcal{Z}_q$, the post-query rule space is the compact convex product $X^{|\mathcal{Z}_q|}$. The objective is continuous and convex in this rule and affine in the compact convex product of channel polytopes. The minimax theorem therefore gives

$$\min_{\psi} \max_{K_i} \sum_i \pi_i \mathbb{E}_{K_i} \operatorname{Gap}_i(\psi(Z)) = \max_{K_i} \min_{\psi} \sum_i \pi_i \mathbb{E}_{K_i} \operatorname{Gap}_i(\psi(Z)),$$

with a saddle point. For a general Borel alphabet the definition keeps its stated order; the evaluated finite LPs have equality.

In singleton channels, exact Bayesian updating gives

$$\mathbb{E}\mathcal{R}(\pi^Z) = \Gamma(C) - \mathrm{RDI}_q(\pi; C) \le \mathbb{E}\Gamma(\operatorname{supp} \pi^Z)$$

when π is least favorable on C. If every nonzero posterior weight is at least $\alpha > 0$, also

$$\mathbb{E}\Gamma(\operatorname{supp} \pi^Z) \le \frac{\Gamma(C) - \mathrm{RDI}_q(\pi; C)}{\alpha}.$$

This holds because a weighted-risk minimizer has each supported loss at most its weighted risk divided by α. These are Bayesian support statements; no contraction theorem for an anytime confidence set follows.

There is no universal positive function f with $\mathbb{E}\Gamma(C_{n+1}) \le \Gamma(C_n) - f(\mathrm{RDI})$. Take losses $(x, 1-x)$, prior $(1/2, 1/2)$, and a binary signal with model probabilities 0.9 and 0.1. The parent radius is $1/2$, the posterior Bayes risk is 0.1, and RDI is 0.4, but posterior support remains both models and its radius remains $1/2$. A confidence set can likewise remain unchanged. The witness is frozen within this one-step experiment, and the post-query rule is used to score a query; deployment still uses the robust solver after the actual confidence update.

For the three-model example the optimal post-query policies are $\psi(1) = (0, 1/2, 1/2)$, $\psi(2) = (1/2, 0, 1/2)$, $\psi(3) = (1/2, 1/2, 0)$. They give worst-case child loss $11/20$. Conversely choose the channel with probability $.1$ at its model index and $.45$ at each other index. Its Bayes success probability is $.45$ for any optimal post-query decision, which proves the matching lower bound. This gives $2/3 - 11/20 = 7/60$ without a numerical LP solution.

D.5 EFFECTIVE DIMENSION

The separation floor does not by itself target the decision: if the witness endpoints differ in a nuisance coordinate (in hidden rooms, endpoints other than the canonical ones with equal nuisance), both RDI and pair separation can rank a nuisance query highly, so the canonical convention is a condition of Proposition B.28, and we do not prove robustness to it.

Corollary D.5 (A smaller bound when loss has a lower-dimensional image)**.** *Suppose all deployment gaps factor as* $\operatorname{Gap}_\theta(x) = f_\theta(Tx)$ *through the same affine map* T*, and let* $r = \dim(\operatorname{aff} T(X))$*. Then the exact witness size bound is* $r + 1$ *in place of* $p + 1$*.*

Proof. Apply the compactness and Helly proof of Theorem 4.2 in the convex compact set $T(X)$. Continuity and convexity descend by the stated factorization. □

Counting attacker facets alone does not give this reduction: even one halfspace per model can require a Helly witness of $p + 1$ models.

D.6 COMPUTATION AND NON-ATTAINED PRICING VALUES

For s transcript symbols, K witness models, n_x defender coordinates, n_y attacker vertices, and L_i channel vertices for model i, the child LP has $sn_x + sK + K$ scalar variables, $sKn_y + \sum_i L_i$ gap/channel inequalities, and one copy of the defender flow constraints for each symbol. A raw alphabet of tens of thousands of symbols therefore gives as many defender blocks. On such instances we do not use the literal formula for acquisition.

An implementation by constraint generation starts with a small set of attacker and response constraints, solves the restricted child LP, and for each (i, z) minimizes $x_z^\top A_i y$ over Y. Violated gap inequalities are added. For each i, maximize $\sum_z K(z) g_{iz}$ over its response polytope and add a violating law. Repeat until both maximum residuals are at most ρ_g, ρ_K. Raising each g_{iz} by ρ_g and each t_i by $\rho_g + \rho_K$ gives a feasible child upper bound; the restricted LP gives a lower bound. The resulting RDI interval has width at most $\rho_g + \rho_K$. A deployment solve similarly reports verified upper and lower gap bounds; their difference belongs in η_{comp}. An error in the RDI score affects only acquisition.

In Proposition B.8, the left-hand side is a supremum and need not be attained by an executable query. On a frozen cell, $\mathsf{Aff}_b(x)$ is Lipschitz in ℓ_1 with constant $L_b \le \sum_{h \in \mathcal{H}_b} \kappa_b(h)$, since all realization coordinates lie in $[0, 1]$. For a closed-cell optimizer x and known interior x_0, use $x_\zeta = (1 - \zeta)x + \zeta x_0$. Reserve $\tau_{\mathrm{sep}}/2$ for optimization error and $\tau_{\mathrm{sep}}/2$ for the perturbation, choosing $\zeta \sum_b w(b) L_b \|x_0 - x\|_1 \le \tau_{\mathrm{sep}}/2$ and verifying the best-response face at the executed query.

D.7 Approximate responses and operational status

Proposition D.6 (Monotone degradation with diagnostic payoff sacrifice)**.** *With fixed deployment losses and query universe, let $\mathcal{P}_\theta^\kappa(q)$ be the image of the full polytope $\{y \in Y : x_q^\top A_\theta y \le m_q(\theta) + \kappa\}$. For $\kappa' \ge \kappa$, channels are nested, both directed information minima and Hellinger separations cannot increase, and $T_\varepsilon^{\mathrm{wc}}$ and $T_\varepsilon^{\mathrm{sel}}$ cannot decrease. A joint common-law obstruction persists after enlargement.*

Proof. The defining inequality weakens with κ. Infima of divergence over larger feasible sets cannot increase. For a fixed true selector, enlarging alternative sets lowers its information value; enlarging the true selector set also enlarges the outer supremum of reciprocal information times. The answer sets remain fixed, so all comparisons pass through their minima and maxima. Existing common laws remain feasible. □

Status convention. The output status is one of: *certified*, carrying a policy and verified upper bound; *obstruction*, carrying an incompatible subset and a common law for every accessible query; *statistical failure*, when the confidence set is empty; *budget censored*, when an administrative cap is reached; or *unknown*, when an oracle cannot certify a required bound. Only the first is a deployment guarantee.

E Completeness, Acquisition Value, and Cost Accounting

This appendix gives three further results: a link between the RDI score and the statistical evidence a query must release, a general inference procedure with a matching finite-stopping characterization for arbitrary finite channel polytopes, and a lower bound stated at the algorithm's actual exit time.

E.1 RDI certifies the information a query must release

Theorem D.1 holds for any predictable exploitation rule. For RDI, a positive score certifies that every admissible selector must release a quantified amount of evidence at that query.

Throughout write $\|P - K\|_{\mathrm{TV}} = \frac{1}{2} \sum_z |P(z) - K(z)|$, $\mathrm{dist}_{\mathrm{TV}}(P, \mathcal{C}) = \min_{K \in \mathcal{C}} \|P - K\|_{\mathrm{TV}}$, and, for a finite witness $W = \{\theta_1, \ldots, \theta_K\}$,

$$G_W := \max_{i \le K} \sup_{x \in X} \mathrm{Gap}_{\theta_i}(x) \;\le\; R_{\mathrm{tot}}.$$

Proposition E.1 (RDI lower-bounds the separation any selector must leave)**.** *Let $\mathcal{Z}_q$ be finite, let $\mathcal{P}_{\theta_i}(q)$ be nonempty, compact and convex, and let $\pi \in \Delta(W)$. For an admissible selection $\mathbf{K} = (K_i)_{i \le K}$ with $K_i \in \mathcal{P}_{\theta_i}(q)$ put $\bar{K}_\pi = \sum_i \pi_i K_i$ and*

$$I_\pi(q; \mathbf{K}) := \sum_i \pi_i \, \mathrm{KL}(K_i \| \bar{K}_\pi), \qquad I_\pi^{\mathrm{wc}}(q) := \min_{\mathbf{K}} I_\pi(q; \mathbf{K}).$$

Then:

1. *for every $\rho \in \Delta(\mathcal{Z}_q)$ and every admissible $\mathbf{K}$,*
$$\mathrm{RDI}_q(\pi; W) \;\le\; G_W \sum_i \pi_i \|K_i - \rho\|_{\mathrm{TV}}; \tag{E.1}$$
2. *consequently $\mathrm{RDI}_q(\pi; W) \le G_W \sqrt{I^{\mathrm{wc}}_\pi(q)/2}$, i.e. $I^{\mathrm{wc}}_\pi(q) \ge 2\,\mathrm{RDI}_q(\pi; W)^2/G_W^2$;*
3. *suppose $\theta^\star \in W$, $G_W > 0$ and $\pi_\star := \pi(W \setminus \{\theta^\star\}) > 0$. Then for* every *law $P \in \mathcal{P}_{\theta^\star}(q)$ (in particular, for every law an admissible selector may realize at q),*
$$\max_{\theta \in W \setminus \{\theta^\star\}} \mathrm{dist}_{\mathrm{TV}}\big(P, \mathcal{P}_\theta(q)\big) \;\ge\; \frac{\mathrm{RDI}_q(\pi; W)}{G_W\,\pi_\star} \;\ge\; \frac{\mathrm{RDI}_q(\pi; W)}{G_W}. \tag{E.2}$$

Proof. (1) Fix ρ and an admissible $\mathbf{K}$. Because the supremum over channel selections dominates the value at $\mathbf{K}$,
$$\mathcal{R}^{\mathrm{wc}}_q(\pi; W) \;\ge\; \inf_\psi \sum_i \pi_i \mathbb{E}_{K_i}\,\mathrm{Gap}_{\theta_i}(\psi(Z)).$$
Each $\mathrm{Gap}_{\theta_i}(\psi(\cdot))$ takes values in $[0, G_W]$, so $|\mathbb{E}_{K_i} f - \mathbb{E}_\rho f| \le G_W \|K_i - \rho\|_{\mathrm{TV}}$ and
$$\sum_i \pi_i \mathbb{E}_{K_i}\,\mathrm{Gap}_{\theta_i}(\psi(Z)) \;\ge\; \sum_i \pi_i \mathbb{E}_\rho\,\mathrm{Gap}_{\theta_i}(\psi(Z)) - G_W \sum_i \pi_i \|K_i - \rho\|_{\mathrm{TV}}.$$
Under the single law ρ the transcript is uninformative: $\inf_\psi \sum_i \pi_i \mathbb{E}_\rho\,\mathrm{Gap}_{\theta_i}(\psi(Z)) = \mathbb{E}_\rho \min_x \sum_i \pi_i\,\mathrm{Gap}_{\theta_i}(x) = \mathcal{R}(\pi; W)$. Taking the infimum over ψ gives (E.1).

(2) Apply (E.1) at a minimizing selection $\mathbf{K}^\star$ and $\rho = \bar{K}^\star_\pi$. Pinsker's inequality and concavity of the square root give $\sum_i \pi_i \|K^\star_i - \bar{K}^\star\|_{\mathrm{TV}} \le \sqrt{I^{\mathrm{wc}}_\pi(q)/2}$. The minimum is attained because the channel polytopes are compact and the objective is continuous on the finite alphabet.

(3) Fix $P \in \mathcal{P}_{\theta^\star}(q)$. *Before* applying (E.1), construct the selection that realizes the distances: for each i with $\theta_i \ne \theta^\star$ choose
$$K^\dagger_i \in \arg\min_{K \in \mathcal{P}_{\theta_i}(q)} \|K - P\|_{\mathrm{TV}},$$
attained because $\mathcal{P}_{\theta_i}(q)$ is compact and the norm is continuous, and set $K^\dagger_i = P$ for the components with $\theta_i = \theta^\star$. Every $K^\dagger_i$ lies in its own channel set, so $\mathbf{K}^\dagger$ is an admissible selection and (E.1) applies to it with $\rho = P$:
$$\mathrm{RDI}_q(\pi; W) \;\le\; G_W \sum_i \pi_i \|K^\dagger_i - P\|_{\mathrm{TV}} \;=\; G_W \sum_{i:\,\theta_i \ne \theta^\star} \pi_i\,\mathrm{dist}_{\mathrm{TV}}\big(P, \mathcal{P}_{\theta_i}(q)\big),$$
because the $\theta^\star$-components contribute zero. Bounding each distance by the maximum and dividing by $G_W \pi_\star > 0$ gives (E.2). □

Parts 2 and 3 prove Proposition 5.1; its separation statement is part 3 with $P = K_{\theta^\star}$.

The non-degeneracy conditions in (3) are needed: if $G_W = 0$ or $\pi_\star = 0$, then $\mathrm{RDI}_q \equiv 0$ and the claim is vacuous.

First, the result is one-sided. Large information does not imply large RDI: two models with the same payoff matrix, $A_\theta = A_{\theta'}$, have $\mathrm{RDI}_q \equiv 0$ for every query, while their channels may be mutually singular. RDI measures only information that reduces the decision loss.

Second, because (E.2) holds for every admissible P, we may minimize over P. This gives a statement about the fixed quantity that controls Theorem E.7.

Corollary E.2 (RDI lower-bounds the conflict separation of its own witness)**.** *Under the hypotheses of Proposition E.1(3), let $\varsigma^{\mathrm{TV}}_q(W; \theta^\star) := \inf_{P \in \mathcal{P}_{\theta^\star}(q)} \max_{\theta \in W \setminus \{\theta^\star\}} \mathrm{dist}_{\mathrm{TV}}(P, \mathcal{P}_\theta(q))$. Then*
$$\varsigma^{\mathrm{TV}}_q(W; \theta^\star) \;\ge\; \frac{\mathrm{RDI}_q(\pi; W)}{G_W}, \qquad \textit{and hence} \qquad \varsigma_q(W; \theta^\star) \;\ge\; \frac{2\,\mathrm{RDI}_q(\pi; W)}{|\mathcal{Z}_q|\,G_W},$$
where $\varsigma_q(\cdot\,; \theta^\star)$ is the sup-norm conflict separation of Definition E.4, anchored at $\theta^\star$, and the second inequality uses $\|\cdot\|_{\mathrm{TV}} \le \frac{1}{2}|\mathcal{Z}_q|\,\|\cdot\|_\infty$. Consequently, if the current canonical witness W is itself an anchored conflict set and $\mathrm{RDI}_q(\pi_n; W) \ge s > 0$, then on the validity event the frequency test of Appendix E.2 removes some member of $W \setminus \{\theta^\star\}$ after at most $N^\star\big(2s/(|\mathcal{Z}_q| G_W)\big)$ tested *uses of q.*

Here $N^\star(\varsigma) = \min\{N \geq 1 : 2\beta_N < \varsigma\}$ counts *tested* uses of q; converting it into episodes costs the schedule factor $5/4$ of Lemma E.15 and the exploration factor, both of which the deadline (E.5) includes.

Third, the corollary bounds the time to break one conflict set, namely the one carried by the current witness. The stopping deadline of Theorem E.7 depends on $\varsigma_0(\theta^\star) = \min_\Xi \varsigma(\Xi; \theta^\star)$ over all anchored conflict sets, and a large RDI score at the current witness says nothing about the other sets. The $|\mathcal{Z}_q|^{-1}$ factor comes only from the norm conversion $\|\cdot\|_{\mathrm{TV}} \leq \frac{1}{2}|\mathcal{Z}_q|\|\cdot\|_\infty$ and enters the corollary's budget quadratically, but not the deadline (E.5) or Table 15, which use the tabulated ς_0 and see the alphabet only through the $\log S$ term of (E.3); whether it can be removed is open. The result therefore does not establish a global sample-complexity advantage for RDI; the experiments of Section 6 measure the realized effect, which differs across designs. The corollary also quantifies over all admissible P before taking the maximum; a pigeonhole argument round by round would not suffice, because averaging realized laws can move the average back inside an alternative's polytope.

E.2 A COMPLETE INFERENCE PROCEDURE FOR GENERAL CHANNEL POLYTOPES

Theorem D.1 needs $J_q^{\mathrm{wc}}(\theta^\star, \theta) > 0$ for a pairwise divergence, and Remark B.17 gives a three-model family where every pairwise factor is identically one. The composite bets of Appendix D.3 resolve that example with events chosen by hand. This subsection replaces both with a single procedure that is valid for arbitrary finite channel polytopes. Its stopping condition is necessary as well as sufficient.

Assumption E.3 (Finite polytope channels)**.** Θ and $\mathcal{Q}_{\mathrm{acc}}$ are finite, every $\mathcal{Z}_q$ is finite, and every $\mathcal{P}_\theta(q)$ is a nonempty polytope given by an explicit finite representation (a finite vertex list or a finite facet system). No common support and no pairwise separation are assumed.

The finite representation makes the membership test of (E.4) a linear program. For the best-response channels of Section 3, it is the vertex list of the attacker's optimal face, which our implementation enumerates. If one assumes only nonempty compact convex sets, every statement below still holds with "linear program" replaced by "call to an exact projection oracle $P \mapsto \mathrm{dist}_\infty(P, \mathcal{P}_\theta(q))$", and the computational claims should be read relative to that oracle.

In this subsection, "algorithm" means an algorithm measurable with respect to the transcript filtration of Section 3. The sufficiency half holds for any finer filtration; the converse is proved only for the transcript filtration, and Proposition A.1 shows that this restriction is necessary.

Let $N_q(n) = \sum_{t \leq n} \mathbf{1}\{q_t = q\}$ and let $\widehat{P}_{q,n} \in \Delta(\mathcal{Z}_q)$ be the empirical transcript frequency over the rounds that used q. With $S = \max_q |\mathcal{Z}_q|$ and $M = |\Theta|$ set

$$\beta_N := \sqrt{\frac{1}{2N} \log \frac{2N(N+1)\,|\mathcal{Q}_{\mathrm{acc}}|\,S\,M}{\delta}}, \qquad N \geq 1, \tag{E.3}$$

and define the running confidence set

$$C_n^{\mathrm{freq}} := \bigcap_{m \leq n} \Big\{\theta : \ \mathrm{dist}_\infty\big(\widehat{P}_{q,N_q(m)}, \mathcal{P}_\theta(q)\big) \leq \beta_{N_q(m)} \ \text{ for every } q \text{ with } N_q(m) \geq 1\Big\}. \tag{E.4}$$

Each membership test is one linear program over the channel's vertex representation. The intersection over m makes eliminations permanent. When the tests are evaluated on the schedule $\mathcal{T}$ of Lemma E.15, the condition for q is imposed only at the rounds m with $N_q(m) \in \mathcal{T}$; this scheduled set is the one run in Theorem E.7 and in the experiments.

Definition E.4 (Conflict separation)**.** For $\Xi \subseteq \Theta$ with $|\Xi| \geq 2$ and a distinguished *anchor* $\vartheta \in \Xi$ set

$$\varsigma_q(\Xi; \vartheta) := \min_{P \in \mathcal{P}_\vartheta(q)} \max_{\theta \in \Xi \setminus \{\vartheta\}} \mathrm{dist}_\infty\big(P, \mathcal{P}_\theta(q)\big), \qquad \varsigma(\Xi; \vartheta) := \max_{q \in \mathcal{Q}_{\mathrm{acc}}} \varsigma_q(\Xi; \vartheta).$$

The minimum is attained by compactness. The anchor is part of the definition. We write $\varsigma(\Xi)$ only as an abbreviation for $\varsigma(\Xi; \theta^\star)$ in passages where $\theta^\star$ is the data-generating model. Every statement that uses a different anchor writes it out.

Lemma E.5 (Conflict separation characterizes joint equivalence)**.** *Under Assumption E.3, for every anchor $\vartheta \in \Xi$, $\varsigma(\Xi; \vartheta) = 0$ if and only if Ξ is jointly strategically observationally equivalent in the sense of Definition B.2.*

Proof. $\varsigma_q(\Xi;\vartheta) = 0$ holds exactly when some $P \in \mathcal{P}_\vartheta(q)$ has zero distance to every $\mathcal{P}_\theta(q)$, $\theta \in \Xi$; the sets are closed, so this means $P \in \bigcap_{\theta\in\Xi} \mathcal{P}_\theta(q)$. Requiring this at every accessible query is Definition B.2. □

Positivity therefore does not depend on the anchor, but the value does, and it sets the deadline (E.5) through $N^\star(\varsigma_0)$; every reported value states its anchor, and $\varsigma_0(\theta)$ below is anchored at the same θ that indexes it.

Theorem E.6 (Anytime validity and completeness)**.** *Under Assumption E.3:*

1. (Validity.) *For every adaptive query rule and every admissible selector,* $\mathbb{P}_{\theta^\star}\{\theta^\star \in C_n^{\mathrm{freq}}$ *for all* $n \geq 0\} \geq 1-\delta$.
2. (Completeness.) *Let* $\theta^\star \in \Xi$ *and* $\varsigma(\Xi;\theta^\star) > 0$*, and let* $q^\sharp$ *attain* $\varsigma(\Xi;\theta^\star)$*. Put* $N^\star(\varsigma) := \min\{N \geq 1 : 2\beta_N < \varsigma\}$*, which is finite. On the validity event, as soon as* $N_{q^\sharp}(n) \geq N^\star(\varsigma(\Xi;\theta^\star))$ *the set* C_n^{freq} *omits at least one element of* $\Xi \setminus \{\theta^\star\}$.

Proof. (1) Fix q and $z \in \mathcal{Z}_q$. Let $t_1 < t_2 < \cdots$ enumerate the rounds with $q_t = q$; these are stopping times of the calibration filtration because the query is $\mathcal{G}_t$-measurable. Under $\theta^\star$ the conditional law $\mu_t := \mathcal{L}(Z_t \mid \mathcal{G}_t)$ lies in $\mathcal{P}_{\theta^\star}(q)$ by admissibility, and $D_j := \mathbf{1}\{Z_{t_j} = z\} - \mu_{t_j}(z)$ is a martingale difference sequence which, conditionally on $\mathcal{G}_{t_j}$, takes values in the predictable interval $[-\mu_{t_j}(z),\, 1-\mu_{t_j}(z)]$ of length exactly 1. The bounded-range form of the Azuma–Hoeffding inequality therefore applies with range 1 and gives the constant below. The cruder form that uses only $|D_j| \leq 1$ would double the boundary. With the union bound $\sum_{N\geq 1} \delta_1/(N(N+1)) = \delta_1$ this gives $\mathbb{P}\{\exists N : |\sum_{j\leq N} D_j| > N\beta_N^{(1)}\} \leq \delta_1$ with $\beta_N^{(1)} = \sqrt{\log(2N(N+1)/\delta_1)/(2N)}$. Taking $\delta_1 = \delta/(|\mathcal{Q}_{\mathrm{acc}}|SM)$ and a union bound over q, z and the M null indices yields, simultaneously for all n, q, $\|\widehat{P}_{q,N} - \bar{\mu}_{q,N}\|_\infty \leq \beta_N$, where $\bar{\mu}_{q,N} = N^{-1}\sum_{j\leq N} \mu_{t_j}$. Convexity of $\mathcal{P}_{\theta^\star}(q)$ gives $\bar{\mu}_{q,N} \in \mathcal{P}_{\theta^\star}(q)$, so $\mathrm{dist}_\infty(\widehat{P}_{q,N}, \mathcal{P}_{\theta^\star}(q)) \leq \beta_N$ and $\theta^\star$ is never eliminated.

(2) Write $N = N_{q^\sharp}(n)$. On the validity event $\theta^\star$ is not eliminated at round n, so $\mathrm{dist}_\infty(\widehat{P}_{q^\sharp,N}, \mathcal{P}_{\theta^\star}(q^\sharp)) \leq \beta_N$, and since $\mathcal{P}_{\theta^\star}(q^\sharp)$ is compact some $P' \in \mathcal{P}_{\theta^\star}(q^\sharp)$ has $\|\widehat{P}_{q^\sharp,N} - P'\|_\infty \leq \beta_N$. Because Definition E.4 minimizes over every element of $\mathcal{P}_{\theta^\star}(q^\sharp)$, there is $\theta \in \Xi\setminus\{\theta^\star\}$ with $\mathrm{dist}_\infty(P', \mathcal{P}_\theta(q^\sharp)) \geq \varsigma(\Xi;\theta^\star)$. The sup-norm distance to a fixed set is 1-Lipschitz, hence

$$\mathrm{dist}_\infty\big(\widehat{P}_{q^\sharp,N}, \mathcal{P}_\theta(q^\sharp)\big) \;\geq\; \varsigma(\Xi;\theta^\star) - \beta_N \;>\; \beta_N$$

whenever $2\beta_N < \varsigma(\Xi;\theta^\star)$, and that θ is eliminated. Finiteness of $N^\star$ follows from $\beta_N \to 0$. □

Theorem E.7 (Exact branch: a matched characterization)**.** *Work under Assumption E.3 in the* exact *branch* $\varepsilon_{\mathrm{span}} = \eta_{\mathrm{comp}} = 0$*, so that the deployment tolerance and the inference tolerance are the same number* $\varepsilon > 0$*. Fix* $\delta \in (0, 1/(p+2))$ *and* $\delta' \in (0,1)$*; part 2 in fact only needs* $\delta < 1/|\Xi|$ *for the offending conflict set* Ξ*, and* $|\Xi| \leq p+2$ *by Helly, so* $\delta < 1/(p+2)$ *is the a-priori worst case (Remark E.8). For* $\theta \in \Theta$ *put*

$$\varsigma_0(\theta) := \min_{\Xi\in\mathfrak{C}_\varepsilon(\theta)} \varsigma(\Xi;\theta), \qquad \min_{\varnothing} := +\infty.$$

Run Algorithm 1 with $C_n = C_n^{\mathrm{freq}}$*, tested on the schedule of Lemma E.15 (or its intersection with any other valid sequence, splitting* δ*; then* β_N*,* $N^\star$ *and* $n^\dagger$ *are evaluated at the frequency component's share of* δ*), and forced exploration* $q_{n+1} \sim (1-\gamma)\lambda_n + \gamma\lambda_0$ *for a full-support* λ_0*,* $\lambda_0^{\min} = \min_q \lambda_0(q) > 0$*, under the* run conditions *of Assumption E.14. Then:*

1. (Sufficiency, fixed true model.) *If* $\varsigma_0(\theta^\star) > 0$ *then with* $\mathbb{P}_{\theta^\star}$*-probability at least* $1-\delta-\delta'$ *the run stops normally by*

$$n^\dagger(\theta^\star) := \left\lceil \frac{5}{4} \max\left\{ \frac{2N^\star(\varsigma_0(\theta^\star))}{\gamma\lambda_0^{\min}},\; \frac{8}{\gamma\lambda_0^{\min}} \log\frac{|\mathcal{Q}_{\mathrm{acc}}|}{\delta'} \right\} \right\rceil, \tag{E.5}$$

where $N^\star(\varsigma) = \min\{N \geq 1 : 2\beta_N < \varsigma\}$ *and the factor* $5/4$ *pays for the test schedule of Lemma E.15; and its output satisfies* $\mathrm{Gap}_{\theta^\star}(x_\tau) \leq \varepsilon$.

2. (Necessity, fixed true model.) *If* $\varsigma_0(\theta^\star) = 0$*, some* $\Xi \in \mathfrak{C}_\varepsilon(\theta^\star)$ *is jointly strategically equivalent, and by Theorem 4.3 every algorithm, including ours, has some* $\theta \in \Xi$ *with* $\mathbb{P}_\theta\{$*finite policy output in* $\mathcal{U}_\varepsilon(\theta)\} \le (|\Xi|-1)/|\Xi| < 1-\delta$*, the last inequality being exactly* $\delta < 1/|\Xi|$*. Hence no algorithm is* (ε,δ)*-correct on* Ξ*, and a fortiori none is on* Θ*.*

3. (Class-level characterization.) *Therefore*

$$\min_{\theta\in\Theta} \varsigma_0(\theta) > 0 \iff \textit{some}\ \text{transcript-filtration}\ \textit{algorithm is}\ (\varepsilon,\delta)\textit{-correct on}\ \Theta,$$

in the sense of (B.17)*, where "transcript-filtration" means measurable with respect to the transcript filtration of Section 3; when the left side holds the displayed algorithm is such an algorithm. The tolerance, the confidence, the correctness definition and the quantifier over* Θ *match across the two directions; the filtration does not, which is why the right-hand side names it: sufficiency holds for any finer filtration, the converse only for the transcript filtration (Appendix A.2, which also covers the designs on which the two coincide).*

Remark E.8 (How sharp the confidence hypothesis is). For the class-level impossibility the bound can be sharpened. The $p+2$ above comes only from requiring a designated anchor. Suppose $\Xi \subseteq \Theta$ is jointly strategically equivalent with $\Gamma(\Xi) > \varepsilon$, i.e. $\bigcap_{\theta\in\Xi} \mathcal{U}_\varepsilon(\theta) = \varnothing$. Each $\mathcal{U}_\varepsilon(\theta)$ is a compact convex subset of $\operatorname{aff} X$, of dimension p, so by Helly's theorem, applied without a required member, some $\Upsilon \subseteq \Xi$ with $|\Upsilon| \le p+1$ already has $\bigcap_{\theta\in\Upsilon} \mathcal{U}_\varepsilon(\theta) = \varnothing$, hence $\Gamma(\Upsilon) > \varepsilon$. Joint strategic equivalence passes to subsets, since a law common to Ξ at a query is common to Υ there. So Υ is itself jointly equivalent and incompatible, and running the obstruction on Υ needs only

$$\delta \;<\; \frac{1}{\min\{M,\, p+1\}}, \qquad M = |\Theta|. \tag{E.6}$$

The term M enters because $|\Upsilon| \le |\Xi| \le M$ regardless of the dimension. The anchored form keeps $p+2$: there $\theta^\star$ must survive the Helly reduction, which costs one set. The sufficiency half and the deadline are stated in terms of anchored conflict sets, and nothing above changes them. Equation (E.6) only shows that the impossibility direction is less restrictive than the uniform $1/(p+2)$ suggests: D5 and D6 have $M = 6$, so it requires only $\delta < 1/6$, which $\delta = .05$ satisfies although their sequence-form p makes $1/(p+2)$ much smaller. Since the offending Ξ is not known in advance, the class hypothesis is the uniform $\delta < 1/\max_\Xi |\Xi|$, bounded by $1/(p+2)$ via Helly (only $\delta < 1/2$ if every conflict set is a pair); the sharper $\delta < 1/|\Xi|$ can be checked once the anchored conflict sets are enumerated, as Table 14 does for S1, S2 and S4–S6. We do not claim the characterization, or its other hypotheses, for D5 or D6: their child LPs are too expensive to solve, they run pairwise separation acquisition, and their anchored conflict sets are not enumerated. On S1–S5, $p = \dim(\operatorname{aff} X)$ is 2 (S1–S3), 6 (S4) and 15 (S5), so $1/(p+2)$ is .25, .125 and .0588, and $\delta = .05$ satisfies $\delta < 1/(p+2)$ on all of them, on S5 with a margin of about 15%; Table 14 shows $|\Xi| \le 4$ on S4 and S5 and $|\Xi| \le 3$ on S1, S2 and S6, so the sharper form requires only $\delta < .25$ there and also covers S6.

Remark E.9 (The normal stopping time is not almost surely finite). Part 3 asserts (ε,δ)-correctness only. For the algorithm displayed here its right-hand side cannot be strengthened to "(ε,δ)-correct with an almost surely finite normal stopping time", and the cause lies in the algorithm itself.

Take $\Theta = \{0,1\}$, $X = [0,1]$, $\mathrm{Gap}_0(x) = x$, $\mathrm{Gap}_1(x) = 1-x$, $\varepsilon = 0.1$, $\delta = 0.05$, a single query q, and the singleton channels $\mathcal{P}_0(q) = \{\mathrm{Ber}(0.5)\}$, $\mathcal{P}_1(q) = \{\mathrm{Ber}(0.5001)\}$. Then $\Gamma(\Theta) = 0.5 > \varepsilon$, the unique anchored conflict set is Θ itself and $\varsigma_0(0) = \varsigma_0(1) = 10^{-4} > 0$, so the hypothesis of part 1 holds under both models; every oracle is exact and no run condition of Assumption E.14 is violated. Consider the path whose first 24 observations are all 0. The boundary (E.3) gives $\beta_{20} = 0.5271486632$ and $\beta_{24} = 0.4888784671$, and $N = 20$ and $N = 24$ are consecutive test times of Lemma E.15; the counts 21–23 are not tested. (Testing at every count would empty the set already at $N = 23$, where $\beta_{23} \approx 0.4976$.) At $N = 20$ the empirical frequency is at sup-distance 0.5 and 0.5001 from the two channels and neither exceeds β_{20}, so both models survive; at $N = 24$ the same two distances both exceed β_{24}, both models are eliminated, $C_{24} = \varnothing$ and the run returns STATISTICAL FAILURE. Under model 0 that path has probability $2^{-24} \approx 5.96 \times 10^{-8} > 0$, so $\mathbb{P}_0\{\tau = \infty\} > 0$ even though the separation condition holds strictly on the whole class.

The difference lies in the event on which each statement holds. Part 1 bounds τ on an event of probability at least $1 - \delta - \delta'$. Letting $\delta' \downarrow 0$ keeps only the intersection with the coverage event, which gives (B.17). Nothing in the argument controls the complement, of probability up to δ, on

which the confidence set may become empty by chance. Assumption E.14 removes the obstruction exit, the UNKNOWN oracle reply and the administrative cap. It cannot remove this exit, because an erroneously emptied confidence set is a failure of the confidence event itself. A finite exit time is not a substitute: the run above exits at $N = 24$, but without a policy. An almost surely finite normal stopping time requires a different algorithm and a new correctness proof. Proposition E.10 gives one.

We distinguish three properties below.

(S1) No STATISTICAL FAILURE. The run never takes the STATISTICAL FAILURE exit, which is triggered by $C_n = \varnothing$. The never-empty replacement gives this for any statistical module, by construction. It does not guarantee that the run outputs a policy.

(S2) Almost surely finite normal stopping, with no cap. The certificate passes at a finite round with probability one. This concerns the shrinking of the confidence set and does not follow from (S1).

(S3) A policy within a finite budget. The certificate passes before a pre-declared deadline. This is (E.5), and it requires a quantitative bound on the separation.

The proposition below gives (S1) unconditionally and (S2) only for confidence sets that contain the frequency-compatibility constraint. Remark E.11 shows that (S2) can fail without this constraint.

Proposition E.10 (A variant with an almost surely finite normal stopping time)**.** *Let* Algorithm 1′ *be Algorithm 1 with the confidence set C_n replaced by*

$$C'_n := \begin{cases} C_n, & C_n \neq \varnothing, \\ \{\widehat{\theta}_n\}, & C_n = \varnothing, \end{cases} \qquad \widehat{\theta}_n := \min\Big(\arg\min_{\theta\in\Theta} M_n(\theta)\Big),$$

where $M_n(\theta)$ is the running incompatibility margin of θ (the largest value so far of $\mathrm{dist}_\infty(\widehat{P}_{q,N}, \mathcal{P}_\theta(q)) - \beta_N$ for the frequency module, of the log-evidence $\log \bar{E}_n(\theta)$ for an e-process module, and of their sum for an intersection), and the outer $\min$ is the model-label tie-break. The two parts have different hypotheses. Part 1 assumes only the definition of the replacement. It is pathwise and, in particular, does not use Assumption E.14, whose clause (iv) presupposes a frequency test. Part 2 assumes Assumption E.3, the exact branch, forced exploration $q_{n+1} \sim (1-\gamma)\lambda_n + \gamma\lambda_0$ with λ_0 of full support, and the run conditions of Assumption E.14. Then

1. (Inheritance, and (S1); any module.) *Algorithm 1′ is (ε, δ)-correct whenever Algorithm 1 is: on every path on which the original algorithm returns a policy the two agree, and the replacement only converts failure exits into outputs that are then subject to the same robust certificate. Property* (S1) *holds for every module.*

   This is only an inheritance statement: it does not make Algorithm 1′ correct whenever $\min_\theta \varsigma_0(\theta) > 0$, and Remark E.11 gives a counterexample. Positive joint separation gives correctness only through a module that can act on it, which is the case of part 2.

2. (Almost surely finite normal stopping; frequency-constrained modules only.) *Suppose in addition that the confidence set satisfies, at every round and every tested count,*

$$C_n \;\subseteq\; \big\{\theta :\; \mathrm{dist}_\infty(\widehat{P}_{q,N_q(n)}, \mathcal{P}_\theta(q)) \le \beta_{N_q(n)} \text{ for every } q\big\}, \tag{E.7}$$

   that is, membership in C_n implies frequency compatibility at every query. If $\min_{\theta\in\Theta} \varsigma_0(\theta) > 0$ then $\mathbb{P}_\theta\{\tau' < \infty\} = 1$ for every $\theta \in \Theta$, where τ' is the normal stopping time of Algorithm 1′; that is, (S2) *holds.*

The frequency module and the intersection module (frequency and comp.+freq.) satisfy (E.7) *by definition, with β_N evaluated at the frequency component's share of δ. The pairwise and composite e-process modules do not, and for them part 2 is not claimed.*

Proof. Part 1. The argument is a pathwise coupling and uses neither the coverage event nor any property of the statistical module. An argument through the coverage event A and Theorem E.6(1) would not suffice, since $\mathbb{P}(S) \ge 1-\delta$ and $\mathbb{P}(A) \ge 1-\delta$ do not give $\mathbb{P}(S \cap A) \ge 1-\delta$.

Fix $\theta^\star$ and an admissible selector. Run the two algorithms on a common probability space, coupling the query draws, the transcripts and any solver randomness, so that they agree until the first round at which $C_n = \varnothing$. This is the first round at which their definitions can differ. Up to that round the two confidence sets are equal by construction, and so are the acquisition decisions, the certificate tests and the outputs.

Let $S_{\mathrm{Alg}\,1} := \{\tau < \infty,\ \mathrm{Gap}_{\theta^\star}(\widehat{x}_\tau) \le \varepsilon\}$ and let $S_{\mathrm{Alg}\,1'}$ be the same event for Algorithm $1'$. On a path in $S_{\mathrm{Alg}\,1}$ the original algorithm returns a policy at the finite time τ, so it did not take the STATISTICAL FAILURE exit at any round $\le \tau$, so $C_n \neq \varnothing$ for all $n \le \tau$. On such a path the replacement is never invoked: $C'_n = C_n$ throughout, and Algorithm $1'$ has the same history, the same stopping time and the same output policy. Hence that path lies in $S_{\mathrm{Alg}\,1'}$, so

$$S_{\mathrm{Alg}\,1} \ \subseteq\ S_{\mathrm{Alg}\,1'}, \qquad \text{and so} \qquad \mathbb{P}_{\theta^\star}(S_{\mathrm{Alg}\,1'}) \ \ge\ \mathbb{P}_{\theta^\star}(S_{\mathrm{Alg}\,1}) \ \ge\ 1-\delta$$

whenever Algorithm 1 is (ε,δ)-correct. As the selector and $\theta^\star$ were arbitrary, Algorithm $1'$ is (ε,δ)-correct. The inclusion can be strict: Algorithm $1'$ may succeed on a path where the original failed, as it does on one run of Table 27. This is why only one direction is claimed.

Property (S1) is immediate from the definition of C'_n: the replacement removes the exit that the test $C_n = \varnothing$ triggers.

Part 2. Forced exploration draws $q_{n+1} = q$ with conditional probability at least $\gamma\lambda_0^{\min} > 0$ given the past, so by Lévy's extension of the Borel–Cantelli lemma every query is played infinitely often almost surely and $N_q(n) \to \infty$ for every q. Since $\beta_N \to 0$ and the schedule of Lemma E.15 tests infinitely many counts, the tested boundaries tend to zero. Moreover $\|\widehat{P}_{q,N} - \bar{\mu}_{q,N}\|_\infty \to 0$ almost surely, by the same sub-Gaussian bound that gives (E.3) together with $\sum_N e^{-2N\beta_N^2} < \infty$. Work on the intersection of these probability-one events.

The sets C_n are nonincreasing, so $S_\infty := \bigcap_n C_n$ is well defined. If $S_\infty = \varnothing$ then $C_n = \varnothing$ at some finite n, whence C'_n is a singleton; and if $|S_\infty| = 1$ then C'_n is a singleton from some finite round on. In either case $\Gamma(C'_n) = 0$, because the maximin policy of a single model attains that model's own value; the exact branch sets $\eta_{\mathrm{comp}} = 0$, so the certified solver returns $\widehat{U} \le \Gamma(C'_n) + \eta_{\mathrm{comp}} = 0 \le \varepsilon$ and $\tau' < \infty$.

Suppose then $|S_\infty| \ge 2$, and suppose for contradiction that $\Gamma(S_\infty) > \varepsilon$. This is the only place where hypothesis (E.7) is used, and it is essential. The argument below needs every permanently retained model to satisfy the frequency constraint at every tested count. This is exactly what (E.7) says, and the bare property "C_n never empties" does not give it. Fix $\theta_0 \in S_\infty$. By Helly's theorem in $\mathrm{aff}\,X$ applied to the sets $\mathcal{U}_\varepsilon(\theta)$, $\theta \in S_\infty$, there is $\Xi \subseteq S_\infty$ with $\theta_0 \in \Xi$, $2 \le |\Xi| \le p+2$ and $\Gamma(\Xi) > \varepsilon$; that is, $\Xi \in \mathfrak{C}_\varepsilon(\theta_0)$.

Fix a query q. Whatever the confidence event does, the data are generated by $\theta^\star$, so every conditional law the selector plays at q lies in $\mathcal{P}_{\theta^\star}(q)$, and by convexity their running average $\bar{\mu}_{q,N}$ lies in $\mathcal{P}_{\theta^\star}(q)$ as well. Each $\theta \in S_\infty$ is never eliminated, so $\mathrm{dist}_\infty(\widehat{P}_{q,N}, \mathcal{P}_\theta(q)) \le \beta_N$ at every tested N; with $\|\widehat{P}_{q,N} - \bar{\mu}_{q,N}\|_\infty \to 0$ and $\beta_N \to 0$ this gives $\mathrm{dist}_\infty(\bar{\mu}_{q,N}, \mathcal{P}_\theta(q)) \to 0$ along the tested counts. The set $\mathcal{P}_{\theta^\star}(q)$ is compact, so passing to a subsequence (simultaneously for the finitely many queries), there is $\bar{P}_q \in \mathcal{P}_{\theta^\star}(q)$ with $\bar{\mu}_{q,N} \to \bar{P}_q$, and $\mathrm{dist}_\infty(\bar{P}_q, \mathcal{P}_\theta(q)) = 0$ for every $\theta \in S_\infty$. Each $\mathcal{P}_\theta(q)$ is closed, so $\bar{P}_q \in \bigcap_{\theta\in S_\infty} \mathcal{P}_\theta(q)$, and in particular $\bar{P}_q \in \mathcal{P}_{\theta_0}(q)$.

Therefore, for every q,

$$\inf_{P\in\mathcal{P}_{\theta_0}(q)} \max_{\theta\in\Xi\setminus\{\theta_0\}} \mathrm{dist}_\infty\big(P, \mathcal{P}_\theta(q)\big) \ \le\ \max_{\theta\in\Xi\setminus\{\theta_0\}} \mathrm{dist}_\infty\big(\bar{P}_q, \mathcal{P}_\theta(q)\big) = 0,$$

so $\varsigma(\Xi;\theta_0) = 0$. Since $\Xi \in \mathfrak{C}_\varepsilon(\theta_0)$ this forces $\varsigma_0(\theta_0) = 0$, contradicting $\min_\theta \varsigma_0(\theta) > 0$. Hence $\Gamma(S_\infty) \le \varepsilon$, the stopping test passes at the first round at which $C'_n = S_\infty$, and $\tau' < \infty$. □

Remark E.11 (Never-empty alone does not give almost sure stopping). Hypothesis (E.7) cannot be dropped; Remark B.17 gives a counterexample. There three channel polytopes meet pairwise but have empty triple intersection, every pairwise e-factor is identically one, and the initial radius is $\Gamma(\Theta) = 2/3 > \varepsilon$. Run the pairwise module on that design with the single available query. Then $C_n = \Theta$ for every n: no model is ever eliminated, so the confidence set is never empty, the never-empty fallback is never invoked, and $\Gamma(C_n) = \Gamma(\Theta) > \varepsilon$ forever. The run neither fails nor certifies,

$\tau' = \infty$ with probability one, and (B.17) fails for every $\delta < 1$. The pairwise row of Table 16 measures this situation: none of the 120 runs certifies, and none ends in statistical failure.

So the property "the confidence set is never empty" gives (S1) and nothing else. An almost sure stopping claim for an e-process-only module needs its own positive-growth argument, that is, a guarantee that the log-evidence against every wrong model diverges. A positive joint separation $\varsigma_0 > 0$ does not give such a guarantee, because joint separation can come entirely from constraints that the pairwise module never evaluates.

Proposition E.10 does not change Algorithm 1 (Remark E.9 still applies to it), leaves the deadline unaffected because the two algorithms coincide on the coverage event, and does not improve accuracy: off that event Algorithm 1′ certifies the maximin policy of a model the data have contradicted. The proposition is a device for proving almost sure stopping and does not recommend Algorithm 1′ for practice; Appendix G.6 reports both algorithms. Sufficiency is proved model by model; we do not claim that $\varsigma_0(\theta^\star) > 0$ for a single $\theta^\star$ is necessary for anything, since an algorithm told the true model needs no data.

Remark E.12 (The reported runs are in the approximate branch, and by how much). Part 2 above and the characterization of Theorem E.7 are exact-branch statements about an idealized solver with $\eta_{\mathrm{comp}} = 0$. The certified deployment oracle used in our runs returns a bracket of positive width. Every run in Section 6 is therefore in the approximate branch of Corollary E.16, where the separation condition is only sufficient and no converse is claimed.

Equation (4.3) asks for $U(C) \le \Gamma(C) + \eta_{\mathrm{comp}}$, and Table 6 also certifies the lower side. For every design and every subset C of its candidate library, we compute a certified lower bound $L_\Gamma(C) \le \Gamma(C)$ from a Lagrangian dual pair, namely a prior w on C and a mixed attacker y_θ for each model. The evaluation is conservative. The model values are taken from below. The inner maximum is taken from above, by an explicit maximum over rows in the matrix form and by the certified backward induction of Appendix F.7 in the sequence form. Every dot product is charged its forward-error bound. Any dual pair gives a valid certificate, so a poor dual solution loses tightness but never validity. Then

$$\eta_{\mathrm{comp}} \;\le\; \max_{C \subseteq \Theta} \big(U(C) - L_\Gamma(C)\big),$$

and this maximum covers all contributions at once, since it bounds the whole difference, whereas the v_{upper} allowance is only one summand. The measured values are in Table 6: at most 1.1×10^{-14} on S1–S6, where Theorem 4.4 is used, and at most 1.1×10^{-13} over every design, including D5 and D6; each exceeds the corresponding v_{upper} allowance, as it must, and all are at least ten orders of magnitude below the smallest tolerance used ($\varepsilon = .002$).

Remark E.13 (Which linear programs carry correctness). Our implementation uses three families of linear programs, and they play different roles for correctness. We certify them separately, and the first two are certified in the direction that matters for correctness.

The statistical elimination LP. Eliminating θ at query q requires $\mathrm{dist}_\infty(\widehat{P}_{q,N}, \mathcal{P}_\theta(q)) > \beta_N$. This distance is the value of a minimization. A feasible primal point, which is what a solver returns and what the reported objective is, bounds it from above and so cannot justify an elimination. Overestimating the distance near the boundary would remove the true model and break the coverage event on which the guarantee rests. Our implementation therefore never eliminates on the reported objective. It computes a certified interval $[L_{\theta,q}, U_{\theta,q}]$ from the dual characterization of the distance to a compact convex set,

$$\mathrm{dist}_\infty(P, \mathcal{P}_\theta(q)) = \max_{\|u\|_1 \le 1} \Big(\langle u, P \rangle - \max_{V \in \mathrm{ext}\, \mathcal{P}_\theta(q)} \langle u, V \rangle \Big),$$

under which every u in the ℓ_1 ball gives a certified lower bound. We evaluate the solver's dual multipliers, rescaled onto the ball so that feasibility holds by construction, and the $2S$ coordinate functionals $\pm e_i$. We then shrink the best value by the forward-error bound of the two inner products. A model is eliminated if and only if $L_{\theta,q} > \beta_N^{\mathrm{impl}} := \beta_N + \rho_{\mathrm{num}}$, with ρ_{num} the sampler-to-storage discrepancy of Appendix F.7; every theoretical statement in this appendix ((E.5), $N^\star(\varsigma)$ and part 2 of Proposition E.10) concerns the ideal $\beta_N \to 0$ in the exact branch. Since $\beta_N^{\mathrm{impl}} \to \rho_{\mathrm{num}} > 0$, the almost sure stopping argument of part 2 does not transfer directly to an implementation with fixed precision, which needs a separation margin above the error floor ($\varsigma_0 > 2\rho_{\mathrm{num}}$) or a precision that improves with N. The first holds here by thirteen orders of magnitude: $\rho_{\mathrm{num}} \le 3.7 \times 10^{-15}$ on the

designs that run the frequency test ($\leq 4.1 \times 10^{-13}$ over every row of Table 6, the largest on D5, which runs the pairwise module), against $\varsigma_0 \geq .047$ in Table 15. Widening a radius can only make elimination harder, so this affects only power; the widest certified interval over the grid of Table 6 is below 10^{-9} on every design.

The deployment LP, which decides when the algorithm stops and what it guarantees, is certified on both sides (Remark E.12). *The acquisition LPs* (child impurity, Hellinger floor, maximin allocation) affect only efficiency, because the stopping rule never uses an acquisition score; we report their residuals but do not certify them.

Assumption E.14 (Run conditions)**.** For the deadline in Theorem E.7 the run must satisfy: (i) the optional full-universe obstruction test of Algorithm 1 is disabled, so the loop cannot exit with an obstruction certificate instead of a policy; (ii) every oracle call returns in finite time and none reports UNKNOWN, so no round is skipped and no early exit occurs; (iii) no administrative cap is imposed; (iv) the frequency test is evaluated on the schedule of Lemma E.15; (v) the confidence budget δ is shared across all components of C_n. Violating (i)–(iii) does not affect validity, since the coverage statement of Theorem E.6 holds regardless. Each of them can, however, produce a terminal status other than a policy, and then the deadline does not apply. The experiments report those statuses separately.

Lemma E.15 (The implemented test schedule costs a factor $5/4$)**.** *Let the compatibility test for query q be evaluated when its own count N lies in $\mathcal{T} = \{N \leq 8\} \cup \{N > 8 : N \equiv 0 \bmod 2^{\lfloor \log_2 N \rfloor - 2}\}$. Then for every $N \geq 1$ there is $N' \in \mathcal{T}$ with $N \leq N' \leq 5N/4$. Consequently every deadline stated for per-round testing holds for the implemented schedule after multiplication by $5/4$.*

Proof. For $N \leq 8$ take $N' = N$. For $N > 8$ the step is $2^{\lfloor \log_2 N \rfloor - 2} \leq N/4$, and the smallest multiple of the step at or above N is less than $N + N/4$. □

Evaluating the membership test on any predictable subset of rounds preserves anytime validity, because the union bound in Theorem E.6 already covers every N and the boundary is evaluated at the actual count. It does not preserve the per-round deadline, which is why (E.5) includes the factor $5/4$. Our implementation uses this to keep the number of linear programs logarithmic in the horizon, and any valid time-uniform envelope for the same martingale could replace (E.3) without changing the proofs.

Proof of Theorem E.7. Part 2 is Lemma E.5 together with Theorem 4.3: for a jointly equivalent Ξ the obstruction gives a common law for the entire adaptive experiment, so the returned answer has one law and $\sum_{\theta \in \Xi} \mathbf{1}\{\widehat{x} \in \mathcal{U}_\varepsilon(\theta)\} \leq |\Xi| - 1$ forces some model to succeed with probability at most $(|\Xi| - 1)/|\Xi|$. Since $|\Xi| \leq p + 2$ and $\delta < 1/(p + 2)$, this is below $1 - \delta$.

For part 1, forced exploration makes $N_q(n)$ stochastically dominate a $\mathrm{Bin}(n, \gamma\lambda_0(q))$ variable, because at each round the executed query is drawn from λ_0 with probability at least γ independently of the past. The multiplicative Chernoff bound $\mathbb{P}\{\mathrm{Bin}(n, p) < np/2\} \leq e^{-np/8}$ and a union bound over queries give $N_q(n) \geq \gamma\lambda_0(q)n/2$ for every q with probability at least $1 - \delta'$ once $n \geq 8 \log(|\mathcal{Q}_{\mathrm{acc}}|/\delta')/(\gamma\lambda_0^{\min})$. At $n = n^\dagger$ every query therefore has a *tested* count at least $N^\star(\varsigma_0(\theta^\star))$, by Lemma E.15. By Theorem E.6(2), on the validity event every $\Xi \in \mathfrak{C}_\varepsilon(\theta^\star)$ has lost at least one alternative, while $\theta^\star$ is retained.

Suppose $\Gamma(C_{n^\dagger}) > \varepsilon$. Then $\bigcap_{\theta \in C_{n^\dagger}} \mathcal{U}_\varepsilon(\theta) = \varnothing$, and Helly's theorem in $\mathrm{aff}\, X$ supplies at most $p+1$ members of $C_{n^\dagger}$ with empty intersection; adding $\theta^\star$ gives $\Xi \subseteq C_{n^\dagger}$ with $\theta^\star \in \Xi$, $|\Xi| \leq p + 2$ and $\Gamma(\Xi) > \varepsilon$, i.e. a member of $\mathfrak{C}_\varepsilon(\theta^\star)$ entirely contained in $C_{n^\dagger}$, a contradiction. This argument uses only anchored sets. If two models other than $\theta^\star$ are jointly indistinguishable from each other, both may survive forever without affecting the conclusion, because a conflict set that certifies $\Gamma(C_n) > \varepsilon$ can always be taken to contain $\theta^\star$. An obstruction among false models therefore does not invalidate part 1.

Hence $\Gamma(C_{n^\dagger}) \leq \varepsilon$. In the exact branch the certified solver returns $U_n \leq \Gamma(C_n)$ and stopping tests $U_n \leq \varepsilon$; deployment accuracy is Theorem 4.1 on the same coverage event.

For part 3, if $\min_\theta \varsigma_0(\theta) > 0$ then part 1 applies under every model; applying it for every $\delta' > 0$ and letting $\delta' \downarrow 0$ gives $\mathbb{P}_\theta\{\tau < \infty,\ \mathrm{Gap}_\theta(x_\tau) \le \varepsilon\} \ge 1 - \delta$ for every $\theta \in \Theta$, which is (ε, δ)-correctness. Conversely if $\varsigma_0(\theta) = 0$ for some θ, part 2 rules out every algorithm. □

Theorem E.6(1) and Theorem E.7 together prove Theorem 4.4: the former is its coverage claim, part 3 of the latter is its characterization, and part 1 is its deadline.

Corollary E.16 (Approximate branch: sufficiency only)**.** *Drop $\varepsilon_{\mathrm{span}} = \eta_{\mathrm{comp}} = 0$ and assume only $\varepsilon_0 := \varepsilon - 2\varepsilon_{\mathrm{span}} - \eta_{\mathrm{comp}} > 0$. If $\varsigma_0^{(\varepsilon_0)}(\theta^\star) := \min_{\Xi \in \mathfrak{C}_{\varepsilon_0}(\theta^\star)} \varsigma(\Xi; \theta^\star) > 0$ then part 1 of Theorem E.7 holds verbatim with ε_0 in the deadline and the deployment guarantee at $\mathrm{Gap}_{G^\dagger}(x_\tau) \le \varepsilon$. This is only a sufficient condition. The corresponding necessity statement obtained from Theorem 4.3 at tolerance ε_0 concerns (ε_0, δ)-correctness, a strictly stronger accuracy requirement than the deployment target ε; the two cannot be combined into one characterization, and we do not claim a matched condition outside the exact branch.*

E.3 A LOWER BOUND AT THE ALGORITHM'S ACTUAL EXIT TIME

Theorem B.12 sets $\tau = \infty$ on every path that does not return a policy. Any procedure that reports statistical failure with positive probability then has $\mathbb{E}\tau = \infty$, and the bound constrains nothing. The following version counts the executed episodes.

Definition E.17 (Exit time and soundness)**.** An algorithm's *exit time* σ is the index of the last diagnostic episode it runs, on every path, whatever status it reports. It is (ε, δ)-sound if $\sigma < \infty$ almost surely with $\mathbb{E}_{\theta,s}\sigma < \infty$ under every model and admissible selector, its output is either a policy $\widehat{x}$ or an abstention symbol, and

$$\mathbb{P}_\theta\{\text{a policy is returned and } \mathrm{Gap}_\theta(\widehat{x}) \le \varepsilon\} \ \ge\ 1 - \delta. \tag{E.8}$$

A single budget δ covers every non-conforming outcome: an incorrect policy, statistical failure, an obstruction report, a budget-censored exit and an UNKNOWN oracle result are all counted in the same δ, and all of them are subsumed by the abstention symbol for the purposes of (E.8).

Theorem E.18 (Exit-time lower bound)**.** *Let Assumption B.9 hold, $\varepsilon > 0$ and $0 < \delta < 1/(p+2)$. Every (ε, δ)-sound algorithm satisfies, for every stationary selector $s \in \mathfrak{S}(\theta^\star)$,*

$$\mathbb{E}_{\theta^\star,s}\sigma \ \ge\ \frac{\mathrm{kl}\big((1-\delta)/(p+1),\ \delta\big)}{D_\varepsilon(s)}, \tag{E.9}$$

with the conventions $1/0 = +\infty$ and $0 \cdot (+\infty) = 0$. More precisely, with $\bar\lambda(q) = \mathbb{E}_{\theta^\star,s}N_q(\sigma)/\mathbb{E}_{\theta^\star,s}\sigma$, every $\Xi \in \mathfrak{C}_\varepsilon(\theta^\star)$ gives

$$\mathbb{E}_{\theta^\star,s}\sigma \max_{\theta \in \Xi \setminus \{\theta^\star\}} \sum_q \bar\lambda(q) J_q(s,\theta) \ \ge\ \mathrm{kl}\Big(\frac{1-\delta}{|\Xi|-1}, \delta\Big).$$

Proof. Pigeonhole. Fix $\Xi \in \mathfrak{C}_\varepsilon(\theta^\star)$, so $\bigcap_{\theta\in\Xi} \mathcal{U}_\varepsilon(\theta) = \varnothing$ and $|\Xi| \le p+2$. For $x \in \mathcal{U}_\varepsilon(\theta^\star)$ choose $\iota(x) \in \Xi \setminus \{\theta^\star\}$ with $x \notin \mathcal{U}_\varepsilon(\iota(x))$, possible exactly because the intersection is empty. The sets $B_\theta = \{x \in \mathcal{U}_\varepsilon(\theta^\star) : \iota(x) = \theta\}$ partition $\mathcal{U}_\varepsilon(\theta^\star)$ into at most $|\Xi| - 1 \le p+1$ pieces. Let $A_\theta \in \mathcal{F}_\sigma$ be the event that a policy is returned and lies in B_θ. By (E.8) under $\theta^\star$, $\sum_\theta \mathbb{P}_{\theta^\star}(A_\theta) \ge 1 - \delta$, so some θ has $\mathbb{P}_{\theta^\star}(A_\theta) \ge (1-\delta)/(|\Xi|-1) \ge (1-\delta)/(p+1)$; under that θ, A_θ implies a returned policy outside $\mathcal{U}_\varepsilon(\theta)$, so $\mathbb{P}_\theta(A_\theta) \le \delta$.

The alternative experiment. Write $K_q := K_q^s \in \mathcal{P}_{\theta^\star}(q)$ for the stationary true-model law. If $J_q(s,\theta) = +\infty$ for some q with $\mathbb{E}_{\theta^\star,s}N_q(\sigma) > 0$, the left-hand side of the displayed inequality in the statement is $+\infty$ and there is nothing to prove; queries with $\mathbb{E}N_q(\sigma) = 0$ contribute nothing under the convention $0 \cdot \infty = 0$. So assume $J_q(s,\theta) < \infty$ for every query used with positive expected count, and choose for each such q a minimizer

$$K_q^\theta \in \arg\min_{K \in \mathcal{P}_\theta(q)} \mathrm{KL}(K_q \| K),$$

attained because $\mathcal{P}_\theta(q)$ is compact and KL is lower semicontinuous on a finite alphabet. Finiteness of $\mathrm{KL}(K_q \| K_q^\theta)$ forces $\operatorname{supp} K_q \subseteq \operatorname{supp} K_q^\theta$, so on the support of K_q the log-ratio

$\log(K_q(z)/K_q^\theta(z))$ is finite; the alphabet is finite, hence $|\log(K_q(Z)/K_q^\theta(Z))| \le B_q < \infty$ almost surely under $\theta^\star$. This boundedness comes from finiteness of J_q alone. Admissibility of the selector would not give it: admissibility permits an alternative law that vanishes where K_q does not, and then the ratio is unbounded and the divergence infinite. Let $\mathbb{P}_\theta$ denote the law of the whole experiment under model θ with the stationary selector $(K_q^\theta)_q$, which is admissible by construction.

Change of measure at a truncated exit time. For $T \in \mathbb{N}$ the time $\sigma \wedge T$ is a bounded stopping time and

$$L_{\sigma\wedge T} = \sum_{t\le\sigma\wedge T} \log \frac{K_{q_t}(Z_t)}{K^\theta_{q_t}(Z_t)}$$

has bounded increments with conditional means $J_{q_t}(s,\theta)$, so Wald's identity gives $\mathbb{E}_{\theta^\star,s} L_{\sigma\wedge T} = \sum_q \mathbb{E}_{\theta^\star,s}[N_q(\sigma\wedge T)]\, J_q(s,\theta)$. The data-processing inequality for kl applied to the event $A_\theta \cap \{\sigma \le T\} \in \mathcal{F}_{\sigma\wedge T}$ yields

$$\begin{aligned}\mathrm{kl}\big(\mathbb{P}_{\theta^\star}(A_\theta \cap \{\sigma\le T\}),\, \mathbb{P}_\theta(A_\theta\cap\{\sigma\le T\})\big)& \\ \le \sum_q \mathbb{E}_{\theta^\star,s}[N_q(\sigma\wedge T)]\, J_q(s,\theta) \;\le\; \sum_q \mathbb{E}_{\theta^\star,s}[N_q(\sigma)]\, J_q(s,\theta),&\end{aligned}$$

the last step because $J_q \ge 0$ and $N_q(\sigma\wedge T) \le N_q(\sigma)$.

Limit. Since $\sigma < \infty$ almost surely under both models, the events $A_\theta \cap \{\sigma \le T\}$ increase to A_θ and both probabilities converge. The function kl is lower semicontinuous on $[0,1]^2$, so the displayed bound passes to the limit:

$$\mathrm{kl}\big(\mathbb{P}_{\theta^\star}(A_\theta), \mathbb{P}_\theta(A_\theta)\big) \;\le\; \sum_q \mathbb{E}_{\theta^\star,s}[N_q(\sigma)]\, J_q(s,\theta) \;=\; \mathbb{E}_{\theta^\star,s}\sigma \sum_q \bar\lambda(q) J_q(s,\theta),$$

the last equality using $\mathbb{E}_{\theta^\star,s}\sigma < \infty$. Monotonicity of kl over the relevant range and the anchored Helly representation (B.15), minimized over $\mathfrak{C}_\varepsilon(\theta^\star)$, give (E.9). □

Remark E.19 (Where the upper and lower bounds fail to meet). Both bounds measure cost in executed diagnostic episodes, but three gaps remain.

1. *Selector class.* The lower bound ranges over stationary selectors and uses $D_\varepsilon(s)$; the stopping bounds use the selector-uniform $D_\varepsilon^{\mathrm{wc}}(\theta^\star,\cdot)$, which can be strictly smaller (Remark B.15). The ratio $T_\varepsilon^{\mathrm{wc}}/T_\varepsilon^{\mathrm{sel}}$ can be infinite.

2. *Probability calibration.* Theorem D.1 gives $1-\delta-\delta'$, with δ the confidence-set budget and δ' the forced-exploration drift budget, so fixing δ' keeps the success probability bounded away from one as $\delta \downarrow 0$, and $\delta' = \delta$ gives $1-2\delta$ rather than $1-\delta$ (a common nominal $1-\delta_{\mathrm{total}}$ needs $\delta = \delta' = \delta_{\mathrm{total}}/2$). Either way the leading coefficient changes, because each budget enters a logarithm. With $\delta' = \delta$, writing $L = \log(1/\delta)$, $c = \gamma D_{\varepsilon_0}^{\mathrm{wc}}(\theta^\star,\lambda_0)$ and $B_e = \max|\log e_q^{\theta,\theta'}|$, the deadline condition $nc - 2B_e\sqrt{2n\log(Mn(n+1)/\delta)} \ge \log((M-1)/\delta)$ with $n = \alpha L$ becomes $\alpha c - 2\sqrt{2}\, B_e\sqrt{\alpha} \ge 1$, so

$$n^\star = (1+o(1))\,\alpha^\star \log\frac{1}{\delta}, \qquad \alpha^\star = \left(\frac{\sqrt{2}\,B_e + \sqrt{2B_e^2 + c}}{c}\right)^2 \;\ge\; \frac{1}{c}, \tag{E.10}$$

   with equality only in the degenerate case $B_e = 0$. The $1/[\gamma D^{\mathrm{wc}}]$ coefficient quoted after Theorem D.1 therefore holds at fixed δ'. It does not hold once the two confidence budgets are matched.

3. *Expectation versus high probability.* (E.9) bounds an expectation; the stopping results bound a high-probability deadline. Converting the latter to an expectation requires a tail bound beyond $n^\star$.

In the worst case over these three gaps the comparison of coefficients is $(p+1)\,\alpha^\star\, D_\varepsilon(s)$, between a $1-\delta$ statement and a $1-2\delta$ statement. Nothing in this paper establishes a sharp characteristic time for maximin-sufficient certification.

E.4 Scope of the theoretical branches

For each branch of the theory (finite or compact model classes, finite or continuous query universes, finite alphabets or Gaussian observations, exact or numerically normalized factors), Table 3 states whether it gives a computable acquisition rule, an anytime-valid confidence sequence with explicit constants, a finite-stopping guarantee and an executable oracle; a dash marks a gap in what is proved.

Table 3: Scope of the theoretical branches. "Acquisition" means a rule that can be evaluated exactly at each round; "confidence" means an anytime-valid sequence with explicit constants; "stopping" means a finite high-probability deadline for the implemented rule, never an almost surely finite normal stopping time; "matched" in the stopping column means that the same separation condition characterizes (ε, δ)-correctness, with termination left open; "oracle" means every required optimization admits a certified implementation. Witness existence for compact models (Theorem 4.2) is listed separately; the paper gives no learning pipeline for compact models.

Branch	Acquisition	Confidence	Stopping	Oracle
Finite models, finite pool, finite alphabet, exact factors	yes	yes	sufficient, Thm D.1	yes
Polytope compatibility, exact branch ($\varepsilon_{\mathrm{span}} = \eta_{\mathrm{comp}} = 0$)	yes	yes	matched, Thm E.7	yes
Polytope compatibility, approximate branch	yes	yes	sufficient, Cor E.16	yes
Finite models, finite pool, singular channels	yes	yes	yes, Prop D.3	yes
Finite models, finite pool, normalized factors	yes	yes	yes, with drift correction	yes
Finite models, Tier II continuous pool, full logging	priced	yes	—	separation only, Prop B.8
Finite models, Tier II continuous pool, general RDI pricing	heuristic	yes	—	—
Compact models, witness existence	—	—	—	—
Compact models, linear moments with bounded bias	optional	yes, Thm B.26	diameter fallback only	yes
Gaussian bandit illustration (hidden rooms)	closed form	yes	sufficient, Prop B.28	closed form

E.5 Collected operational definitions

Witness and prior. The least-favorable prior on a nonempty finite C is read off the dual of the linear program for the robust radius. Write the deployment polytope as $X = \{x \geq 0 : Ex = b\}$; for a simplex $E = \mathbf{1}^\top$, $b = 1$. With variables $\lambda_{ij} \geq 0$ indexed by model $i \in C$ and attacker column j, and a free multiplier μ of the same dimension as b, solve

$$\max \sum_{i,j} v_i^\star \lambda_{ij} + b^\top \mu \quad \text{s.t.} \quad \sum_{i,j} (A_i)_{kj} \lambda_{ij} + (E^\top \mu)_k \leq 0 \ \ (\forall k), \qquad \sum_{i,j} \lambda_{ij} \leq 1, \tag{E.11}$$

whose value is $\Gamma(C)$ and whose prior is $\pi_i = \sum_j \lambda_{ij}$. The canonical witness maximizes π lexicographically in the model label order on the optimal face of (E.11), and W is the support of the resulting prior. Permuting the label order defines the alternative witnesses used in the ablation. Robust policies and optimal allocations are chosen by the analogous lexicographic rule on their own optimal faces; these finite LP tie-breaks give unique choices and a measurable algorithm.

This dual-face rule defines the canonical witness throughout the paper, and our implementation computes it. It differs from the rule that first chooses the subset W with $\Gamma(W) = \Gamma(C)$ of smallest cardinality, first in lexicographic order, and only then solves for a prior on W. The two rules can disagree. Take $X = [0, 1]$ and three models with $\mathrm{Gap}_0(x) = 2x$, $\mathrm{Gap}_1(x) = 2(1 - x)$ and $\mathrm{Gap}_2(x) = \max\{0, 3 - 4x\}$, all realizable as matrix games with $v_0^\star = v_1^\star = 2$ and $v_2^\star = 3$. Then

$\Gamma(\{0,1,2\}) = \Gamma(\{0,1\}) = 1$, so the rule that picks the subset first returns $W = \{0,1\}$ with prior $(1/2, 1/2)$; every optimal dual solution has $\pi = \big((1+t)/2, (1-3t)/2, t\big)$ for $t \in [0, 1/3]$, so lexicographic maximization of π_0 forces $t = 1/3$ and the rule used here returns $\pi = (2/3, 0, 1/3)$ and $W = \{0,2\}$. We recompute this instance as a consistency check on both the prior and its support.

Two conventions complete the definition. First, the rule has an explicit branch at $\Gamma(C) = 0$, applied before any lexicographic optimization: the prior is uniform on C and the witness is all of C. The dual does not force this choice (exact lexicographic maximization would return a point mass on the model with the lowest index, making the reported $|W_n|$ misleading), and nothing downstream depends on it, since every RDI score of a parent with zero radius is zero under any prior. Second, the witness returned by this rule is not shown to have minimum cardinality: the bound $|W| \le p+1$ of Theorem 4.2 is an existence statement about some witness, so we record the realized $|W_n|$ at each round. Our implementation covers the simplex case $E = \mathbf{1}^\top$, $b = 1$, which is the deployment space of S1–S6 (S5 is the 15-dimensional simplex over its 16 pure search plans). The sequence-form designs D5 and D6 do not call a witness oracle, since PS-EASI and the plug-in rules never need one.

PS-EASI. The pairwise baseline builds the set of active pairs $\{(i,j) \subseteq C_n : \Gamma(\{i,j\}) > \varepsilon\}$, scores each pair at each query by the symmetrized certified log-growth $\min\{g_q(i,j), g_q(j,i)\}$ where g_q is the audited lower bound on the conditional log-growth of the safe factor, and takes the maximin allocation over queries. If no active pair exists, it plays λ_0. It never calls a witness oracle or a child impurity LP; "directed-growth aggregation" means exactly the displayed symmetrized minimum.

Negative certified growth. A numerically normalized factor whose audited growth g_q is below -10^{-9} on any alternative vertex is replaced by the constant factor one. Growth in $[-10^{-9}, 0]$ is treated as numerical zero, and the drift correction $\gamma \sum_q \lambda_0(q) g_q - (1-\gamma)\xi$ of Appendix A.5 is used with $\xi = 10^{-9}$.

Numerical witness tolerance. A witness is accepted when its certified radius interval lies within 10^{-9} of the parent radius interval; the realized loss is recorded per round.

Weights in Proposition D.4. The weight w is the mixture weight of the event component *inside* the composite e-process, after the outer one-half split between the pairwise and composite halves. With the uniform mixture over $|\mathcal{Q}_{\mathrm{acc}}|$ queries, $|\mathcal{Z}_q|$ single-symbol events and eight bets, $w = \frac{1}{2} \cdot (8|\mathcal{Q}_{\mathrm{acc}}||\mathcal{Z}_q|)^{-1}$; the outer one-half is therefore already included.

F Design Specifications, Reproduction and Full Tables

With Appendix G, which has the S1–S3 deployment payoffs, S3's model 3 and the full D5 and D6 games, this appendix gives every problem definition needed to reimplement the designs, and it states the D5 information structure exactly as used. Section 6 reports each answer once, with a figure; from Appendix F.8 on we give, organized by the same four questions, the numbers behind those figures, the ablations they omit and the statistical conventions. Most tables are further views of the runs of Section 6; the rest are separate studies with their own seeds (the sweep, view, selector, baseline and generating-mechanism studies with 20 seeds, the coverage runs with 250 and 400 seeds, the misspecification runs with 60 seeds and the replication of S5 on fresh seed roots).

F.1 S4: a small two-sided deployment game with an exact child LP

S1–S3 have identical attacker columns in their deployment payoffs, so the deployment opponent is inert and only calibration is strategic. S4 removes this limitation, and its child LP (B.7) is still solved exactly at every witness.

Game. On the path graph $0-1-2$ the defender starts at node 1 and occupies a node at times $1, 2, 3$, moving to an adjacent node or staying: seven tours $(1,0,0)$, $(1,0,1)$, $(1,1,0)$, $(1,1,1)$, $(1,1,2)$, $(1,2,1)$, $(1,2,2)$. The attacker picks a target/time pair $(j,t) \in \{0,1,2\} \times \{2,3\}$ (six pure actions); the defender's reward is $-w_\theta(j)\{1 - 0.9(1-e_\theta)\}$ if it occupies j at time t and $-w_\theta(j)$ otherwise.

Models. Four types, with $(w_\theta(0{:}2), e_\theta, \text{base}_\theta)$ equal to $((0.62, 0.55, 0.40), 0.05, 0.05)$, $((0.40, 0.55, 0.60), 0.10, 0.06)$, $((0.60, 0.56, 0.42), 0.22, 0.07)$ and $((0.42, 0.56, 0.58), 0.30, 0.08)$ for $\theta = 0, 1, 2, 3$. All four deployment games have a strictly mixed equilibrium on both sides, with the maximin defender strategy on three tours and the attacker strategy on three target/time pairs; maximin strategies are not unique, since only the occupancy marginals of a tour mixture at each step enter the payoff. The subset radii are $\Gamma(\{0, 2\}) = .00303$, $\Gamma(\{1, 3\}) = .00443$, $\Gamma(\{2, 3\}) = .05316$, $\Gamma(\{0, 3\}) = .05639$, $\Gamma(\{1, 2\}) = .05752$ and $\Gamma(\{0, 1\}) = \Gamma(\Theta) = .06175$, so for any $\varepsilon \in [.005, .053)$ there are exactly two maximal compatible groups, $\{0, 2\}$ and $\{1, 3\}$: the learner must decide which group contains the truth but never which member. At $\varepsilon < .003$ full identification is required; the main runs use $\varepsilon = .03$, and the sweep crosses both regimes. The canonical witness of the full set, and the initial witness of every run, is $\{0, 1\}$ with unique least-favorable prior $(.3048, .6952, 0, 0)$. So in half the configurations, exactly when $\theta^\star \in \{2, 3\}$, the true model is outside it and the separation floor protects no pair containing the true model at the first round: the situation of Corollary D.2, which the floor cannot fix and the floor ablation measures.

Calibration. Eighteen announced mixtures over the seven tours (Appendix F.6). The transcript records the realized tour index, two alarm bits and one detection report ($7 \times 4 \times 2 = 56$ symbols), but neither the attacker's action nor the exact reward. The bits are read, in this order, at the nodes the defender occupies at steps $s = 2$ and $s = 3$; none is recorded at the known start node $s = 1$, shared by every tour (D5 records one bit per step; Appendix F.5). The alarm probability at node i and step s is $\text{clip}\big(\text{base}_\theta + 0.10(1 - e_\theta) w_\theta(j) \rho(i, j) \tau(s, t),\ 0.02,\ 0.95\big)$, with $\rho = 1$ at the target, 0.4 at an adjacent node and 0.1 otherwise, and $\tau = 1$ for $s \le t$ and 0.3 afterwards. The detection report is an explicit sensor in place of an exact reward label, reporting a true interception with probability 0.55 and a non-interception with probability 0.20 (0.45 miss and 0.20 false-alarm rates). These rates belong to the environment, which is why S4 needs only $\eta = .05$ of uniform record randomization for common support. Distinct transcript laws per (query, model) cell: mean 2.11, maximum 6; eight of the 72 cells have a singleton channel. The child LP is instantiated at the current witness W: $56 \cdot 7 + 56|W| + |W|$ variables and $56 \cdot 6|W|$ gap inequalities plus channel rows. For every rule that calls the canonical witness oracle the witness is a conflicting pair, $|W_n| = 2$, on every recorded round (506 variables, 672 gap inequalities); only the full-set witness ablation is larger, $|W| \in \{3, 4\}$, with worst case 620 variables and 1344 inequalities at four models. The LP is solved exactly, without constraint generation, at every distinct witness.

F.2 S5: A SEARCH GAME WITH UNKNOWN DETECTOR DEGRADATION

The one-sided partially observable stochastic game solvers of Horák et al. (2017; 2023) are evaluated on three benchmark families, pursuit-evasion, patrolling and search, the last with a long history in operations research (Alpern & Gal, 2003). S4 and D5 are patrolling games and no other design is a search game; in search, the censoring studied here arises from the physics of the domain. S5, a search game small enough for the exact child LP (B.7) at every witness and the second design with full RDI and a two-sided deployment game, tests whether S4's answer to Q2 replicates.

Game. Four cells $\{0, 1, 2, 3\}$ carry asset values $v = (1.00, 0.86, 0.72, 0.62)$. A pure plan of the defender's search with two looks is an ordered pair (c_1, c_2) of cells, possibly equal (sixteen plans); the intruder strikes once, choosing where and when, $(j, t) \in \{0, 1, 2, 3\} \times \{1, 2\}$ (eight pure actions). The defender's reward $-v_j\big(1 - d_\theta(j)\, \mathbf{1}\{c_t = j\}\big)$ is a loss of v_j unless the look at period t is at the struck cell and that cell's detector fires. As the damage depends only on the cell, strikes on one cell at different periods pay the same whenever neither look covers it, which, like S4's (target, time) ties, creates best-response ties and keeps the response face set-valued where it would generically be a singleton.

Models. A model describes the sensor network: exactly one of the four detectors is degraded. With θ the degraded cell, $d_\theta(\theta) = 0.15$ and $d_\theta(c) = 0.60$ for $c \neq \theta$. All four deployment games have a strictly mixed equilibrium on both sides; only the look marginals of a plan mixture enter the payoff, so maximin plan mixtures are not unique, and ours use three plans for $\theta \in \{0, 1\}$ and five for $\theta \in \{2, 3\}$, against intruder mixtures over two and three cells, respectively. With $\mu_t(c)$ the probability that look t inspects cell c, the maximin schedules do not depend on t: $(\mu(0), \ldots, \mu(3))$ is

$(.985, .015, 0, 0)$, $(.369, .631, 0, 0)$, $(.501, .311, .189, 0)$ and $(.540, .357, .102, 0)$ for $\theta = 0, 1, 2, 3$, with maximin values -0.8523, -0.7786, -0.6996 and -0.6757. Cell 3 is never worth searching under any model, so its detector is never tested and model 3 is separated from the others only through what they say about cells 0–2.

An overlapping compatible cover. The subset radii are $\Gamma(\{2,3\}) = .00670$, $\Gamma(\{1,2\}) = .01714$, $\Gamma(\{1,3\}) = \Gamma(\{1,2,3\}) = .02334$, $\Gamma(\{0,2\}) = .03437$, $\Gamma(\{0,3\}) = .04071$, $\Gamma(\{0,1\}) = \Gamma(\{0,1,2\}) = .04272$ and $\Gamma(\Theta) = .04293$. At the tolerance $\varepsilon = .02$ of the main runs the maximal compatible sets, $\{1,2\}$, $\{2,3\}$ and $\{0\}$, overlap, so the cover is not a partition and "the compatible group" is not well defined. S4's two compatible sets are disjoint; D5's cover also overlaps, but D5 is too large for the exact child LP; the only other two-sided design with full RDI and an overlapping cover is S6, whose maximal compatible sets at $\varepsilon = .25$ are the three pairs $\{i, j\}$, which overlap pairwise, the triple being incompatible. At $\varepsilon = .02$ the learner therefore needs only a confidence set inside some member of the cover, which member depending on the realized data. The sweep of Table 9 adds four tolerances: at $\varepsilon = .005 < .0067$ the cover is the four singletons and full identification is required; at .01 it is $\{2,3\}, \{0\}, \{1\}$; at .03, $\{1,2,3\}, \{0\}$; at .04, $\{1,2,3\}, \{0,2\}$, which overlaps again. A tolerance of .05 exceeds $\Gamma(\Theta) = .04293$, so the whole class is compatible and the certificate passes at episode zero (Algorithm 1 evaluates the stopping test at $n = 0$, before any query); such tolerances are not swept. The canonical witness of the full model set is $\{0, 1, 3\}$ with least-favorable prior $(.3984, .4633, .1383)$; model 2, with zero prior mass, is outside it, so at the first acquisition round the true model is outside the witness in one of the four configurations: again the situation of Corollary D.2, with a witness of size three instead of two. Over the acquisition rounds before certification in the runs of full RDI under the composite module (witnesses recorded for five seeds per true model), the true model is inside the witness 93.7% of the time (82.3% to 93.7% across the four modules). The witness changes as the confidence set shrinks: the recorded witnesses are $\{0,1\}$, $\{0,1,3\}$, $\{0,3\}$ and, on a single round, $\{0,2\}$, with mean size 2.29 weighted by acquisition rounds of the composite module.

Calibration and transcript. Nineteen announced mixtures over the sixteen plans: the four holding one cell for both looks, the four sweeping from c to $c + 1 \bmod 4$, the uniform mixture, the maximin schedules of the four models and the six midpoints of pairs of maximin schedules. The transcript records the realized plan index and one binary sensor report per look ($16 \times 4 = 64$ symbols). A look at cell c reports one with probability $f = 0.18$ when the intruder is elsewhere (clutter) and $d_\theta(c) + (1 - d_\theta(c))f$ when it is in c at that period, independently across looks given the plan and the strike; the transcript thus records neither the struck cell, the strike period, whether a report was a true detection or clutter, nor the realized damage. As the clutter rate is positive and the same for both detector states, every symbol in the plan's support has positive probability under every model: common support follows from the sensor model itself, so S5 needs no record randomization and runs at $\eta = 0$ (S4 uses $\eta = .05$, D6 $.85$ and D5 $.95$).

Structurally uninformative queries. A detector's reliability is observable only at a look that coincides with the strike. Under a pure plan the intruder knows which two (cell, period) pairs are covered and can always strike an uncovered one (with these values, cell 1 under the plan that holds cell 0 and cell 0 under every other pure plan), so the transcript is pure clutter, with the same law under all four models. Comparing every model's best-response laws at every query confirms that exactly the eight pure plans are uninformative, in agreement with the certified pairwise log-growth there, zero to machine precision. The queries that cannot be censored are the equalizing ones: at a model's own maximin schedule every best response lies on a searched cell, so the intruder is exposed somewhere and can only choose where. As 42% of the queries carry no information under any response, the choice of query matters on this design, and the blind queries are a property of the game.

Sizes. Best-response columns per (query, model) cell: mean 2.11, as on S4, maximum 6. Distinct transcript laws per cell are fewer, $134/76 = 1.76$ on average, because of censoring: on S4 the two coincide, since no two best responses of one model induce the same law, while on S5 they do (at a blind query every response of every model induces the same law), and 34 of the 76 cells have a singleton channel, against 8 of 72 on S4. The child LP (B.7) has $64 \cdot 16 + 4 \cdot 64 + 4 = 1284$ variables and $64 \cdot 4 \cdot 8 = 2048$ gap inequalities plus channel rows at the full model set, and 1219 variables

and 1536 gap inequalities at the canonical witness; as on S4, it is solved exactly, without constraint generation, at every distinct witness.

F.3 S6: A TWO-SIDED DESIGN WHOSE ONLY CONFLICT IS JOINT

Three kinds of design matter for the completeness theorem. S1 with its pool restricted to $\{q_0\}$ has the phenomenon the theorem is about, a decision conflict that no pair of channel polytopes separates, so that every pairwise e-factor is identically one, but its identical attacker columns make its deployment game a one-player decision problem. S4 and S5 are two-sided at deployment, but none of the 24 anchored conflicts of either is pairwise-blind, so they never test joint completeness. S6, the third kind, is two-sided at deployment and pairwise-blind at every query in the pool; it is kept simple because it tests a structural property.

Game. A defender allocates hardening across three postures; an attacker who knows the allocation strikes along its worst direction. The hardening is tuned against exactly one of three attack vectors, and the model $\theta \in \{0,1,2\}$ says which one. Write $T = \{v \in \mathbb{R}^3 : \sum_j v_j = 0\}$ for the tangent plane of the simplex. Deployment uses $K = 3$ *directional* attacker columns with directions $u_0 = \frac{1}{\sqrt{6}}(1,1,-2)$, $u_1 = \frac{1}{\sqrt{6}}(-2,1,1)$ and $u_2 = \frac{1}{\sqrt{6}}(1,-2,1)$, that is, $u_k = -\sqrt{3/2}\left(e_{\pi(k)} - \frac{1}{3}\mathbf{1}\right)$ for a relabeling π. Posture i pays $-\langle u_k, e_i - x^\star_\theta\rangle$ against directional column k and $+2$ against every *strike* column, where $x^\star_\theta = \frac{1}{3}\mathbf{1} + \rho\,(e_\theta - \frac{1}{3}\mathbf{1})$ with $\rho = 0.4$, so $x^\star_0 = (.6,.2,.2)$ and its rotations. A posture mixture x thus pays $-\langle u_k, x - x^\star_\theta\rangle$ against column k and $+2$ against every strike column, so the attacker answers with a directional column, the lower envelope has value zero, attained exactly at $x^\star_\theta$, and $\mathrm{Gap}_\theta(x) = h(x - x^\star_\theta)$ with $h(v) := \max_k\langle u_k, v\rangle = \sqrt{3/2}\,\max_j(-v_j)$ for $v \in T$. The design relies on h not being a norm. It is convex, positively homogeneous and strictly positive on $T \setminus \{0\}$ (the directions positively span T) but not symmetric: $h(u_0) = 1$ while $h(-u_0) = \frac{1}{2}$. As the support function of the triangle $\mathrm{conv}\{u_0,u_1,u_2\}$, which is not centrally symmetric, h is an asymmetric support loss, and $\mathrm{Gap}_\theta(x)$ charges only the coordinates in which x falls short of $x^\star_\theta$. The radii below are therefore not Chebyshev radii of any norm; Proposition F.1 shows that no norm can produce them.

Proposition F.1 (S6 robust radii)**.** *For every nonempty $C \subseteq \{0,1,2\}$, with $m_j := \max_{\theta\in C} x^\star_{\theta,j}$, $\Gamma(C) = \sqrt{3/2}\,(\sum_j m_j - 1)/3 = \rho\,(|C|-1)/\sqrt{6}$, so $\Gamma(\{i\}) = 0$, $\Gamma(\{i,j\}) = \rho/\sqrt{6} = 0.1632993\ldots$ and $\Gamma(\{0,1,2\}) = 2\rho/\sqrt{6} = 0.3265986\ldots$ at $\rho = .4$. No norm on T can produce these three values.*

Proof. Probes carry no mass in any robust solution (Proposition F.2, whose proof does not use this one), so it suffices to minimize over x in the posture simplex. There $\max_{\theta\in C}\mathrm{Gap}_\theta(x) = \sqrt{3/2}\,\max_j(m_j - x_j)$, since the maximum of a maximum may be taken in either order. Minimizing $\max_j(m_j - x_j)$ subject to $\sum_j x_j = 1$, $x \geq 0$: for any feasible x, summing $m_j - x_j \leq t$ over j gives $\sum_j m_j - 1 \leq 3t$, so $t \geq (\sum_j m_j - 1)/3$; and $x_j = m_j - (\sum_j m_j - 1)/3$ is feasible and attains it. It sums to one, and it is nonnegative: for $|C| \leq 2$ we have $\min_j m_j = \frac{1-\rho}{3}$ and $(\sum_j m_j - 1)/3 \leq \frac{\rho}{3}$ with $\frac{1-\rho}{3} \geq \frac{\rho}{3}$ for $\rho \leq \frac{1}{2}$, while for $|C| = 3$ we have $\min_j m_j = \frac{1+2\rho}{3} > \frac{2\rho}{3}$. Finally $m_j = \frac{1}{3} + \frac{2\rho}{3}$ for $j \in C$ and $m_j = \frac{1}{3} - \frac{\rho}{3}$ otherwise, so $\sum_j m_j = 1 + \rho(|C|-1)$. For the last sentence, suppose $\mathrm{Gap}_\theta(x) = N(x - x^\star_\theta)$ for a norm N on T. Write $a := \Gamma(\{i,j\})$. A norm ball is symmetric, so the midpoint of $x^\star_i$ and $x^\star_j$, which lies in the simplex, certifies $\Gamma(\{i,j\}) \leq \frac{1}{2}N(x^\star_i - x^\star_j)$, while for any x, $N(x^\star_i - x^\star_j) \leq N(x^\star_i - x) + N(x - x^\star_j) \leq 2\max_{\theta\in\{i,j\}} N(x - x^\star_\theta)$, so minimizing over x gives $N(x^\star_i - x^\star_j) \leq 2\Gamma(\{i,j\})$. Hence $N(x^\star_i - x^\star_j) = 2a$ for every pair. The barycenter $\bar{x} = \frac{1}{3}(x^\star_0 + x^\star_1 + x^\star_2)$ lies in the simplex and satisfies $\bar{x} - x^\star_i = \frac{1}{3}\left[(x^\star_j - x^\star_i) + (x^\star_k - x^\star_i)\right]$, whence $N(\bar{x} - x^\star_i) \leq \frac{1}{3}(2a + 2a) = \frac{4}{3}a$ for each i, so $\Gamma(\{0,1,2\}) \leq \frac{4}{3}a < 2a$. The reported triple radius is exactly $2a$, so no such N exists. □

The doubling is a consequence of the asymmetry: $\Gamma(C)$ is linear in $|C| - 1$ because only shortfalls are charged, so the binding envelope m gains one ρ of excess mass per extra model. On the payoff matrices the closed form agrees with the certified oracle to 1.2×10^{-16} on all seven subsets, with probe mass 0 in each. Hence $\varepsilon = .25$ asks the learner to remove one model and allows it to

stop at two. Each model's maximin allocation is interior, and there the attacker mixes uniformly over the three directional columns: the equilibrium attacker policy is $(0, 0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ under every model (the three strike columns first, then u_0, u_1, u_2), so the deployment attacker is not inert, as the supplementary code checks.

Diagnostics. The defender may instead announce a *probe* posture, which pays a flat -2 against every *strike* column except the one the hardening blunts, which pays -1.5, and -1.75 against every directional column. Under model θ the attacker's best responses to any probe are therefore exactly the two strike columns $j \neq \theta$, never a directional column.

Proposition F.2 (probes are strictly dominated)**.** *Let x put total mass $p \in (0, 1)$ on probe rows, and let x_D be x restricted to the postures and renormalized. Then for every θ, $\mathrm{Gap}_\theta(x) \geq \mathrm{Gap}_\theta(x_D) + p\,(1.5 - \mathrm{Gap}_\theta(x_D)) \geq \mathrm{Gap}_\theta(x_D) + 0.765\,p$. Hence for every nonempty C the robust objective $\max_{\theta \in C} \mathrm{Gap}_\theta$ is strictly decreased by removing probe mass. At the remaining endpoint $p = 1$, where x_D is undefined, every probe row is the same row, so the worst-case payoff is -2 under every model and $\mathrm{Gap}_\theta(x) = 2$; the uniform posture deployment gives $\max_\theta \mathrm{Gap}_\theta = 2\rho/\sqrt{6} = 0.3266 < 2$, so a pure probe mixture is not robust-optimal either. No robust solution therefore places any probe mass.*

Proof. Every entry of every probe row is at most -1.5. Write $x = (1-p)x_D + px_P$ with x_P supported on probes, and let $c^\star$ attain $\min_c (x_D^\top A_\theta)_c$. Then $g_\theta(x) \leq (1-p)(x_D^\top A_\theta)_{c^\star} + p\,(x_P^\top A_\theta)_{c^\star} \leq (1-p) g_\theta(x_D) - 1.5\,p$. Since $v_\theta^\star = 0$ and $\mathrm{Gap}_\theta = -g_\theta$, this is the first inequality. For the second, $\mathrm{Gap}_\theta(x_D) = h(x_D - x_\theta^\star) = \sqrt{3/2}\max_j (x_{\theta,j}^\star - x_{D,j}) \leq \sqrt{3/2}\max_j x_{\theta,j}^\star = \sqrt{3/2}\,(.6) = 0.7349$, so $1.5 - \mathrm{Gap}_\theta(x_D) \geq 0.765$. □

A probe entry below the model value does not by itself rule a row out of a robust mixture; the margin $1.5 - \max_\theta \sup_{x_D} \mathrm{Gap}_\theta(x_D) = 0.765$ does. The robust solution on the full library is $(0, 0, 0, 0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ (the five probe rows first, in the order of the pool below, then postures $0, 1, 2$). With L_j the record law of strike vector j (the same under every model), the channel polytopes are the three edges of a triangle, $\mathcal{P}_0(q) = \mathrm{conv}\{L_1, L_2\}$, $\mathcal{P}_1(q) = \mathrm{conv}\{L_0, L_2\}$ and $\mathcal{P}_2(q) = \mathrm{conv}\{L_0, L_1\}$. Any two meet, at the vertex with the remaining index, so no pairwise test can separate any pair at any query (the largest certified pairwise log-growth over the pool is -1.0×10^{-12}, zero to machine precision), while all three have empty intersection whenever L_0, L_1, L_2 are affinely independent, so the frequency test separates them. S1's channel table is specified by hand; this triangle is induced by the game: the model enters the transcript only through which vectors the attacker is willing to play, the mechanism this paper studies.

Records and pool. A record is one symbol of $\{\text{vector } 0, \text{vector } 1, \text{vector } 2, \text{inconclusive}\}$, with $L_j = w\,e_j + (1-w)\big(.4 \cdot \mathrm{unif}_3 + .6\,e_{\mathrm{inc}}\big)$ at a probe of resolution w; every symbol has positive probability under every law for $w < 1$, so $\eta = 0$ and support elimination never fires. The pool is five pure probes at $w \in \{.80, .55, .35, .15, 0\}$ plus the even mixtures of the probes at $w = .80$ and 0 and at $w = .55$ and $.15$; a mixture of probes has the same payoff row, hence the same response face, so it is again a probe, of the averaged resolution ($.40$ and $.35$). The probe at $w = 0$ collapses the triangle to a point and is *structurally blind*: no algorithm should spend an episode on it, so the choice of query matters here, as Table 26 measures.

Results. The pairwise module certifies on none of its runs, all censored at the cap (0 of 60 under each of the three acquisition rules run there, 0 of 180 in total, 3 models × 20 seeds per rule); it uses 20 seeds and three rules, instead of the 40 seeds and eleven rules of the other modules, only to save compute, since none of its runs can stop before the cap. The composite, frequency and intersection modules certify on all 120 runs per cell (Table 10), with channel polytopes induced by best responses, so the gap between joint and pairwise tests in Theorem 4.4 appears on a deployment game in which the attacker matters. S6 is a constructed instance: it shows that the phenomenon is compatible with two-sided deployment and says nothing about how often it arises in applications.

F.4 METHOD AND MODULE DICTIONARIES

Each method name in this paper denotes exactly one rule and each module name exactly one module: Table 4 gives the acquisition rules and Table 5 the inference modules. Other names are defined where reported: Static RDI, Static separation and Uniform, informative only in Appendix G.5; the witness variants of Full RDI (full-set, multi-witness, permuted labels) with Table 12; RDI with $\kappa_{\text{sep}} = 0$ or 1, that is, Full RDI with that floor, with Table 13; the hidden-rooms rules in Appendix G.4. All runs use one implementation and record their parameters; every table states the acquisition rule, inference module, selector, ε, κ_{sep} and seed count, or inherits them from the protocol of Table 7. Separation falls back to λ_0. ID-KL is an acquisition rule that differs from Plug-in KL only in its alternative set; it stops at the sufficient certificate.

Table 4: Algorithm dictionary. "Sufficient" stopping is the certificate test $\widehat{U} \leq \varepsilon$ of Theorem 4.1. Every run records both its sufficient time and its singleton time on the same trajectory, so the singleton target needs no separate method and the two targets are compared within a trajectory.

Name	Acquisition	Fallback	Stopping
Separation (PS-EASI)	maximin over queries of $\min\{g_q(i,j), g_q(j,i)\}$ over the *active conflicting* pairs $\{i,j\} \subseteq C_n$ with $\Gamma(\{i,j\}) > \varepsilon$ (as in Appendix E.5; with $\varepsilon_{\text{span}} = 0$ throughout this coincides with the span-adjusted $a = \varepsilon - 2\varepsilon_{\text{span}}$ at which Appendix B defines the RDI rule's *protected* pairs)	λ_0 when no active pair conflicts	sufficient
Full RDI	exact child LP (B.7) on the canonical witness, then (B.10) with the floor κ_{sep} over the conflicting pairs *inside the witness*	RDI argmax when the protected set is empty	sufficient
Plug-in KL	maximin certified growth of the plug-in model against the models on which its own optimal policy has gap above ε	λ_0 when that set is empty	sufficient
Group-KL	as Plug-in KL, but against the models on which the robust policy of the largest maximal compatible subset of C_n containing the plug-in model has gap above ε (ties to the lexicographically first subset)	λ_0 when that set is empty	sufficient
ID-KL	maximin certified growth of the plug-in model against *all* other active models	λ_0 when alone	sufficient
Uniform	λ_0	—	sufficient
Fixed q_k	point mass on query k	—	sufficient

Table 5: Dictionary of inference modules. A module is the map from the transcript to the confidence set C_n. It is independent of the acquisition rule, and every table cell names both. δ is the total confidence budget, and the last column shows how it is split. Support elimination runs inside every module. It costs no δ, because it never eliminates a model whose stored channel can produce the observed symbol, and it is exact in floating point (Appendix F.7).

Module	Components	Confidence allocation
pairwise	support elimination; pairwise e-process, each null a uniform mixture over its $M-1$ competitor components	all of δ to the e-process threshold $\log(1/\delta)$
composite	support elimination; the same pairwise e-process; event bets, uniform over queries, events and the bet grid $v \in \{.01, .05, .15, 1/3, .5, .7, .9, .99\}$	the two e-processes are averaged with weights $1/2, 1/2$ and the mixture is compared with $\log(1/\delta)$; the mixture of two anytime-valid e-processes is anytime valid at the same level
frequency	support elimination; the pairwise e-process at level $\delta/2$; *intersected with* the frequency test with boundary (E.3) at level $\delta/2$	$\delta/2$ to each; the intersection is valid at δ
intersection ("comp.+freq.")	as the frequency module, with the composite e-process in place of the pairwise one	$\delta/2$ to each

The frequency column of every table therefore reports the frequency test of Theorem 4.4 intersected with a pairwise e-process at half the budget. The test acts alone only on the restricted pool of Table 16 and on S6 (Appendix F.3), where every pairwise information projection is identically zero, so the pairwise component eliminates nothing and every elimination is due to the frequency test. Which component eliminated a model is logged at the round of elimination. The Separation fallback has two visible consequences. On S1, S3 and S6 no pair of models conflicts at the stated tolerance (every pair of models has radius at most $1/2$ on S1 and S3, against $\varepsilon = .6$, and .1633 on S6, against $\varepsilon = .25$), so the protected set is empty: Separation plays λ_0 and is identical to Uniform, hence the

coinciding columns there, while RDI falls back to its own argmax. For the same reason κ_{sep} has no effect on these designs: with no pair to protect, the floor constrains nothing.

F.5 THE D5 INFORMATION STRUCTURE

D5's eight calibration tours are not its deployment strategy space, so we state the structure used in the experiments. (1) *When the attacker commits.* Once, at the start of the episode, the attacker chooses one target/time pair $(j, t) \in \{0, \dots, 4\} \times \{1, 2, 3\}$ (fifteen pure actions); as an open-loop column player it best responds to the announced defender plan and cannot revise its choice after observing alarms or the defender's realized position. (2) *Order within a step.* The defender moves first and then observes the step's alarm bit a_s, generated at the node it has already entered at step s. (3) *Closed-loop deployment, open-loop calibration.* The deployment policy may condition on the alarm history and on its own realized path; the calibration queries do not, each executing one of the twelve announced open-loop mixtures over the eight tours. (4) *Deployment strategy space.* The full perfect-recall sequence-form polytope over information states $(\text{visited path}, \text{alarm history})$: 65 information sets, 228 sequence variables, 196 terminal edges; neither the eight tours nor the 49 open-loop walks. (5) *Sequence-form constraints.* One flow row per information set, $\sum_a x_{h,a} = 1$ at the root and $\sum_a x_{h,a} - x_{\mathrm{parent}(h)} = 0$ otherwise, a 65×229 equality system (the last column holds the auxiliary scalar). Alarm probabilities are folded into the terminal payoff coefficients, which is valid because they factor along the path. (6) *An asymmetry between the two trees.* The deployment tree stops expanding after the third move and so branches on two alarm bits, while the calibration transcript records all three; the third alarm is recorded but affects no decision. (7) *Alarm rates.* The alarm model has a baseline $0.04 + 0.01\theta$ that depends on the type, plus a signal scaled by proximity and timing; across types the realized detection probability at the target ranges from 0.43 to 0.60 and the alarm rate far from it from 0.04 to 0.13. The model has no separate miss or false-alarm parameter: any such rate quoted for D5 must be derived from that expression, written out in Appendix G.3 with the graph, the table of target weights and the tours; the S4 design of Appendix F.1 uses explicit rates instead.

These choices affect the reported quantities. On the sequence-form polytope the robust radius of the full set of six models is $\Gamma(\Theta) = 0.18197601975129515$, with solver residual $1.2 \cdot 10^{-15}$. Two comparisons with a smaller deployment space are possible, and they can differ in sign. The first restricts the defender's strategy set and keeps the benchmark fixed, minimizing $\max_\theta\{v^\star_\theta - g_\theta(x)\}$ over x in the smaller set with the same $v^\star_\theta$, computed on the full sequence-form polytope. Over the eight calibration tours this gives $\Gamma_{\mathrm{tours}}(\Theta) = 0.3426136363636367 > \Gamma(\Theta)$, as it must, since shrinking a minimization's feasible set cannot lower its value. Restricting instead to the 49 legal open-loop walks of length three gives 0.18197601975129518; the two solver outputs agree to sixteen digits, inside D5's certified tolerance $\eta_{\mathrm{comp}} = 1.1 \cdot 10^{-13}$ (Table 6), so in this instance deployment conditioned on alarms gives no improvement, within that tolerance, over the best distribution over open-loop walks. This is a measured property of this instance; in general an improvement is not excluded, as the alarm law depends on the attacker's committed action. Seven of the fifteen D5 pairs have radius exactly zero. Enumerating all 63 nonempty subsets on the full sequence-form polytope, the maximal compatible sets at $\varepsilon = .01$ are $\{0, 3, 5\}$, $\{2, 3, 5\}$, $\{1, 3\}$ and $\{3, 4\}$, the last two maximal because they lie in no compatible triple. This has two consequences for how D5 can be described. First, the maximal sets overlap: model 3 belongs to all four, so they cover Θ but do not partition it, and "the compatible group" is not a well-defined label. Second, model 3 is in every maximal set for a structural reason: its target weights peak at node 3 with $w_3(3) = 0.9$, at graph distance two from the patrol start 0, whose neighbors are $\{1, 2, 4\}$, so the attacker can strike node 3 at time 1, when the defender cannot be there, and obtain -0.9, the most negative payoff in that model. Hence $\min_j p_{3,j}^\top x = -0.9 = v^\star_3$ for every realization plan and $\mathrm{Gap}_3(x) \equiv 0$: model 3 imposes no deployment constraint, every defender policy being optimal for it. Its channels differ from the others, so it contributes to the diagnostic problem but not to the decision problem, which explains why D5's compatible sets are large.

F.6 FULL QUERY CATALOGS

A query pool is part of the problem definition, so the catalogs are stated in full.

S1–S3 (five queries). q_0 is the strategic reset with $\mathcal{P}_i(q_0) = \mathrm{conv}\{(.1, .1, .1) + .7e_j : j \neq i\}$; q_1 is the weak chance signal $(.3, .3, .3) + .1e_i$, identical on the whole best-response face; q_{k+2}, $k = 0, 1, 2$, are the signals $(p, (1-p)/2, (1-p)/2)$ with $p = 1/3 + a_k$ when $i = k$ and $p = 1/3 - a_k/2$ otherwise, and $(a_0, a_1, a_2) = (.09, .07, .05)$. S3's model 3 copies model 0's $q_1, \ldots, q_4$ channels and has the singleton q_0 law $(.8, .1, .1)$.

S4 (eighteen queries). The seven pure tours, ordered as in Appendix F.1; the uniform tour mixture; the maximin mixtures $\mu^\star_\theta$ of the four models, $\theta = 0, 1, 2, 3$; the six midpoints $\frac{1}{2}(\mu^\star_i + \mu^\star_j)$, $i < j$.

D5 (twelve queries). The eight tours are 0123, 0432, 0234, 0104, 0201, 0402, 0120, 0210, written as node sequences from the start node 0. The catalog is, in order: the three pure specialist tours 0432, 0402, 0210 (tour indices $1, 5, 7$); the uniform mixture over all eight; the four even mixtures of adjacent tours $(0, 1), (2, 3), (4, 5), (6, 7)$; the uniform mixtures over the even and over the odd indices, $(0, 2, 4, 6)$ and $(1, 3, 5, 7)$; and over the first and the second half, $(0, 1, 2, 3)$ and $(4, 5, 6, 7)$.

D6 model class. The six vectors of card counts over the ranks (J, Q, K), in the index order of every D6 table, are $\theta_0 = (4, 2, 1)$, $\theta_1 = (2, 4, 1)$, $\theta_2 = (1, 2, 4)$, $\theta_3 = (4, 1, 2)$, $\theta_4 = (2, 1, 4)$ and $\theta_5 = (1, 4, 2)$: every deck has seven cards, and the six models are the six permutations of the multiset $\{4, 2, 1\}$. Three pairs, $\{0, 3\}$, $\{1, 5\}$ and $\{2, 4\}$, admit a common maximin defender policy and have deployment radius exactly zero; every other pair has radius at least $4/105 \approx .0381$, so at $\varepsilon = .03$ the compatible/conflicting partition is unambiguous. Cards are dealt without replacement, so equal ranks can occur whenever a rank has multiplicity at least two, and a showdown between equal ranks pays zero. Each deck has eight support deals: the six ordered pairs of distinct ranks and the two pairs of equal ranks that the counts 4 and 2 make possible. The action ordering, showdown stakes, terminal payoffs and information sets of the extensive form are in Appendix G.3; with these definitions the instance can be rebuilt from this description alone.

D6 controls for the information source. Question Q4 of Section 6 uses two derived families, each controlling exactly one thing (Appendix F.12 reports what each isolates). *Shared deal distribution*: all six models deal from the uniform $(2, 2, 2)$ deck and differ only in the rank order, the six permutations of (J, Q, K) in lexicographic order $(0, 1, 2)$, $(0, 2, 1)$, $(1, 0, 2)$, $(1, 2, 0)$, $(2, 0, 1)$, $(2, 1, 0)$ (each entry the strength rank of the corresponding card), so the deal law is identical across models by construction. *Chance only*: the six original biased decks, with one common rank order and the responder frozen at the single pure strategy $(1, 1, 0, 0, 1, 1)$, in the six-bit encoding of player two's strategies of Appendix G.3.

D6 (forty-six queries). Defender mixtures over the 64 pure strategies of player one, in order: the uniform mixture; the maximin policies of the six decks, computed by linear programming; the fifteen midpoints of pairs of maximin policies; six pure heuristics, given as the six information-set bits (bet/check by rank, then call/fold by rank), $(0, 0, 1, 0, 0, 1)$, $(0, 1, 1, 0, 1, 1)$, $(1, 0, 1, 0, 0, 1)$, $(1, 1, 1, 1, 1, 1)$, $(0, 0, 0, 0, 0, 0)$ and $(0, 0, 0, 1, 1, 1)$; and eighteen $\mathrm{Dirichlet}(0.25 \cdot \mathbf{1}_{64})$ draws, taken in sequence from a pseudorandom generator with seed 606606, so the catalog is deterministic given that seed.

Tie-breaking and randomization. Attacker response vertices are enumerated as the vertices of $\{y \in \Delta : (\mu^q)^\top A_\theta y \leq m_q(\theta) + \kappa\}$, including every intersection of a simplex edge with the payoff halfspace (keeping only the κ-near-optimal pure actions would give too small a response set), in the index order of that enumeration. The lexicographic selector takes vertex 0. The greedy evidence-minimizing selector is a myopic heuristic and need not attain the worst case over admissible selectors: at each round it takes the vertex of the true response face that minimizes the evidence the current module would gain, that is, it minimizes over $P \in \mathcal{P}_{\theta^\star}(q)$ the one-step log-gain $\sum_{i \in C_n \setminus \{\theta^\star\}} w_i \, \mathbb{E}_P[\log(\text{one-step factor against } i)]$, with weights $w \propto \exp(\log \bar{E}_n(\cdot))$ that depend on the history, plus, with composite inference, the current log gains of the event components weighted by their wealth, with ties broken by the smallest index. The lookahead selector optimizes the same objective over two rounds, enumerating the next round's response vertices and the transcript distribution they induce. Observation streams and query randomization use two separate pseudorandom generators, each seeded by a fixed constant together with the true model index and

the run seed, so within each configuration and true model the nth use of a query consumes the same uniform under every method.

The observation stream does not depend on the episode cap. Each query has its own random stream, derived from the (true model, seed) stream, so what a query reads is a function of (θ, seed, q) alone and raising the cap extends every path without changing it. Two studies of one design with different caps, such as the main tables with 40 seeds at cap $= 6000$ and the sweeps with 20 seeds at 20,000, therefore share prefixes at shared seeds, and a study with 20 seeds is a sub-sample of the corresponding study with 40.

F.7 How numerical error enters the certificate

Four numerical error sources are kept separate: the gap between the stored channel table and the law the sampler realizes, the e-validity margin of a normalized factor, the deployment certificate, and acquisition residuals. Only the third enters η_{comp}; none is inferred from a solver residual's size.

The experimental model. The mathematical objects of Sections 3–4 are channel polytopes $\mathcal{P}_\theta(q)$ defined by exact best-response faces. The experiments instead store a finite table of double-precision transcript laws, one set of rows per (q, θ): the experimental model, to which every statement below refers. Rows agreeing to twelve decimals are merged and the sampler draws from the merged rows, so a row that is not stored is never a data-generating law, and the maximum defining the safe normalizer ranges over exactly the laws that can occur; by linearity of expectation the bound extends to every mixture in the convex hull, which is what an admissible selector may play. Response vertices are collected with a 10^{-10} optimality tolerance, so the stored face might omit a true best response or add a near-optimal column; exact arithmetic settles both.

Lemma F.3 (The stored face contains the exact face)**.** *For every design, query q and model θ used in this paper, the stored response face contains the exact best-response face of the stored game at the stored query.*

Proof. Doubles are dyadic rationals, so $\lambda_q^\top A_\theta$ can be evaluated exactly in $\mathbb{Q}$, and its $\arg\min$ over attacker columns (the exact best-response face) is obtained by exact comparisons. Carrying out this evaluation on all 802 (design, query, model) cases used in the experiments gives the inclusion in every case. □

This is the direction validity needs, and the safe one wherever the face is used: enlarging $\mathcal{P}_\theta(q)$ enlarges the set over which the safe normalizer maximizes, the polytope the frequency test must exclude before eliminating, and the attacker's admissible set in the deployment certificate, so a stored face that is too large costs power and keeps validity. Equality holds on 689 of the 802 cases; on the remaining 113 the tolerance admits a near-optimal column that is not optimal, so the certified statements below transfer to the described game conservatively.

Source 1: the sampler does not realize the stored row. Every draw (transcript symbols, and the actions of a deployed behavioral policy) selects a category from a stored vector p of doubles with a uniform u on the grid $\{j2^{-53} : 0 \le j < 2^{53}\}$. With the floating-point cumulative sum $c_0 = p_0$, $c_k = \mathrm{fl}(c_{k-1} + p_k)$, the standard inverse-CDF rule

$$z = \min\big\{\min\{k : c_k \succeq u\},\, s - 1\big\}, \tag{F.1}$$

where $\succeq$ is $\ge$ under the right-closed endpoint convention (category k receives $(c_{k-1}, c_k]$) and $>$ under the left-closed one (it receives $[c_{k-1}, c_k)$), and $\min \varnothing := s$, can return a category of stored probability zero in two separate ways: under the right-closed convention $u = 0$ selects the first category even when its stored mass is zero, and when the stored doubles sum to one as rationals but their floating-point cumulative sum stops short of one, every u above that sum falls through to the final category, whatever its stored mass. Both cases occur for explicit dyadic vectors, so (F.1) does not preserve support. Our sampler instead restricts to the positive entries $I = \{k : p_k > 0\}$, in increasing order $I[0], I[1], \ldots$, with C_j the floating-point cumulative sum over I alone:

$$z = I\big[\min\{\min\{j : C_j > u\},\, |I| - 1\}\big], \qquad \min \varnothing := |I|. \tag{F.2}$$

A category with $p_k = 0$ is not in I and cannot be returned, whatever u is and wherever the cumulative sum ends; the remaining tail mass goes to the last positive category. The law $\widetilde{p}$ realized by (F.2) can be computed exactly in rational arithmetic. We set $\rho_{\mathrm{num}} := \max_{q,\theta,\mathrm{row}} \|\widetilde{p} - p\|_1$, rounded up to the next double (for each design in Table 6); the subscript separates it from the unrelated uses of ρ: the proximity factor $\rho(i, j)$ of the S4 and D5 alarm models (Appendices F.1 and G.3), the spread of S6's maximin allocations (Appendix F.3), a reference law and a dominating measure in Propositions E.1 and B.3, respectively, and an auxiliary scalar in the proof of Proposition B.23.

This ρ_{num} is charged wherever the stored table is used as if it were the sampling law. Since $\mathbb{E}_{\widetilde{p}}[r] \le \mathbb{E}_p[r] + \rho_{\mathrm{num}}\|r\|_\infty$ for a nonnegative factor r, the safe normalizer becomes $\max\{1, \widehat{v}\}(1 + \zeta_s) + \rho_{\mathrm{num}}\|r\|_\infty$, a factor's certified drift is reduced by a further $\rho_{\mathrm{num}}\|\log r\|_\infty$, and our implementation uses the radius $\beta_N^{\mathrm{impl}} := \beta_N + \rho_{\mathrm{num}}$ instead of the ideal β_N of (E.3). Support elimination is then exactly valid and uses no confidence budget: an observed symbol z was returned by (F.2) under the true model, so $z \in I$ for the vertex that generated it and the true model's stored law gives z positive probability. The reverse collapse, two positive entries so close that the cumulative sum does not separate them, costs power but cannot cause a false elimination. No random run reaches these boundary cases, so we check them on constructed inputs instead.

Source 2: e-validity of the normalized factor. A proposed factor is divided by $\max\{1, \widehat{v}\}(1 + \zeta_s) + \rho_{\mathrm{num}}\|r\|_\infty$, with $\widehat{v}$ the largest computed expectation over the vertices of the null. As $\widehat{v}$ is a dot product of s nonnegative doubles, the classical forward error bound gives $v \le \widehat{v}(1 + 2\gamma_s)$, where $\gamma_k := k\mathbf{u}/(1 - k\mathbf{u})$ is the standard constant for a length-k dot product, $\mathbf{u} = 2^{-53}$ the unit roundoff (not the uniform u of Source 1) and $s = |\mathcal{Z}_q|$ the transcript alphabet size (not the time step of Appendix F.1). Hence $\sup_{P \in \mathcal{P}_\theta(q)} \mathbb{E}_P[e] \le 1$ holds for the stored table and for the realized sampler laws, with no audit tolerance, as soon as $\zeta_s \ge 2\gamma_s$. We take $\zeta_s = \max\{10^{-12}, 4\gamma_s\}$, twice the certified requirement, and build no channel table whose alphabet would violate it. The floor binds for the alphabets of the main designs and not for D5's raw alphabet, so the lossless aggregation of identical symbols of the common background is part of what makes the margin certifiable. The cost is a certified log-growth reduction of $\log(1 + \zeta_s) \le 10^{-12}$ per episode, at most $2 \cdot 10^{-8}$ nats over the longest cap for finite channels, 20,000 episodes, one contributor to the tolerance ξ of the drift correction under forced exploration in Appendix A.5; the implementation uses $\xi = 10^{-9}$ (Appendix E.5).

Source 3: deployment feasibility and the certificate. Clipping a solver's raw vector to the nonnegative orthant and charging the constraint violation to η_{comp} works on a simplex but not on the sequence form, where a vector violating the flow constraints is the realization plan of no behavioral policy, and η_{comp} bounds optimization error only for a feasible policy, saying nothing about an infeasible one. We therefore distinguish three objects. The executed behavioral policy σ draws an action at each information set by (F.2) applied to the stored share vector; this is what a user runs, and the certificate is about it. Its exact realization plan $r^\star$ satisfies $Er^\star = b$, $r^\star \ge 0$ with zero residual, because the realized action probabilities are exact rationals summing to one, but $r^\star$ is not stored. The certificate computation has only the stored float vector x, obtained by multiplying rounded parent masses by rounded shares, so it must carry a bound on $\|r^\star - x\|_1$, which we propagate through the reconstruction itself rather than read off the flow residual, because the constant relating a flow violation to an objective error depends on the tree.

Lemma F.4 (Forward error of the behavioral reconstruction)**.** *Write $p(I)$ for the parent sequence of information set I, s_I for its stored share vector, q_I for the law* (F.2) *realizes from s_I, and $\mathbf{u} = 2^{-53}$. Then for every child sequence c of I, $|r^\star_c - x_c| \le |r^\star_{p(I)} - x_{p(I)}|\, q_{I,c} + x_{p(I)}\,|q_{I,c} - s_{I,c}| + \mathbf{u}\, x_{p(I)} s_{I,c}$, and evaluating this in breadth-first order (each parent is bounded before its children) and summing over all sequences gives a computable bound on $\|r^\star - x\|_1$.*

Proof. $x_c = \mathrm{fl}(x_{p(I)} s_{I,c})$, so $|x_c - x_{p(I)} s_{I,c}| \le \mathbf{u}\, x_{p(I)} s_{I,c}$, while $r^\star_c = r^\star_{p(I)} q_{I,c}$ by definition of σ. Subtracting and inserting $x_{p(I)} q_{I,c}$ gives $r^\star_c - x_{p(I)} s_{I,c} = (r^\star_{p(I)} - x_{p(I)}) q_{I,c} + x_{p(I)}(q_{I,c} - s_{I,c})$. The deviations $|q_{I,c} - s_{I,c}|$ are the exact rationals of Source 1, so every term is computable, and the recursion terminates because the tree is finite and the root has $r^\star = x = 1$. □

On a simplex the same statement holds with one layer, the executed policy being a single draw from the stored vector, so $\|r^\star - x\|_1 = \|\tilde{x} - x\|_1$ exactly. The certificate is assembled as follows.

The plan is repaired by renormalizing the raw shares within each information set in breadth-first order and setting $x(h,a) = x(\mathrm{parent}(h))\,\beta(h,a)$; Lemma F.4, evaluated on the computed numbers, bounds its distance to the executed plan, and the flow residual is reported alongside as a diagnostic only. Each model's value is bracketed by certified bounds. The lower bound is the worst-case payoff of an exactly feasible plan, reduced by its own dot-product error and by the transport term above. The upper bound is evaluated by backward induction over doubles, with its forward error accumulated alongside, and is applied to the exact probability vector that (F.2) realizes from the stored attacker vector, whose exact rational defect $|1-\sum_j \mathrm{fl}(y_j)|$ is charged to the allowance, so the inequality $v^\star \le \max_{x\in X} x^\top A y$ is applied to a simplex point that exists. The reported bound is $\widehat{U} = \max_{\theta\in C_n}\{v_{\mathrm{upper},\theta} - g_{\mathrm{lower},\theta}(x)\} + \text{allowance}$, with allowance $= \max_{\theta\in C_n}\{\gamma_k \max_j \sum_i |p_{\theta ji}x_i| + \alpha_\theta\} + \|r^\star - x\|_1 \max_{\theta,j,i}|p_{\theta ji}|$, where γ_k is the constant of Source 2 with k the number of entries of x, $g_{\mathrm{lower},\theta}$ is the minimum over attacker columns of the double-precision dot product at the stored plan, and α_θ is the certified error of $v_{\mathrm{upper},\theta}$. The forward error of $g_{\mathrm{lower},\theta}$ is the allowance's first term, so the reported bound never treats the floating-point value as exact. Two further conditions are enforced: every allowance, and every step of both error recursions, is rounded up to the next double, so an allowance can be loose but never too small; and since the relative bound $\gamma_k|x|$ is the standard one for normal doubles, an absolute term of one smallest normal double replaces it whenever a magnitude entering a product is subnormal. Maximized over all $2^{|\Theta|}-1$ confidence sets, the realized allowance is $8.9\cdot 10^{-16}$ on S4 and $3.5\cdot 10^{-14}$ on D5, against the blanket 10^{-9} that a construction based on residuals would need. With it, $\widehat{U} \ge \sup_\theta \mathrm{Gap}_\theta(\sigma)$ for the executed policy, which is σ and not the stored vector x. These allowances certify $\widehat{U}$ as an upper bound on the executed policy's worst gap. The quantity η_{comp} is the whole gap $\widehat{U} - \Gamma(C)$ in (4.3), which also includes the suboptimality of the robust optimization and the lower value error, and neither $8.9\cdot 10^{-16}$ nor $3.5\cdot 10^{-14}$ bounds it without a further argument. Remark E.12 and Table 6 give that argument, certifying the other side, $L_\Gamma(C) \le \Gamma(C)$, and reporting $\max_C\{\widehat{U}(C) - L_\Gamma(C)\}$ for each design. The two sets of numbers are consistent and bound different quantities: on D5 the certified η_{comp} is $1.1\cdot 10^{-13}$, three times the $3.5\cdot 10^{-14}$ allowance above, because it bounds strictly more error terms. By Lemma F.3 this transfers to the described game, exactly where the stored face is the exact face and conservatively on the 113 cases where it is a strict superset.

Table 6: Certified numerical quantities, each a maximum over every deployment solve of the design. $[L_\Gamma, U](\Theta)$ is the certified bracket on the robust radius of the full library. $\eta_{\mathrm{comp}} = \max_C\{U(C) - L_\Gamma(C)\}$ over *every* subset C, so it dominates the suboptimality of the robust optimization, both value errors and the policy repair at once; the v_{upper} allowance beside it is one summand. "Plan L_1" is the bound of Lemma F.4 on the distance from the stored vector to the plan the executed behavioral policy realizes, ρ_{num} is the sampler discrepancy, and ζ_s is the e-validity margin, at its floor 10^{-12} on every design, against the required $2\gamma_s$. "Stat. LP" is the widest certified interval on an elimination distance over a grid containing every channel vertex and a Dirichlet sample of the simplex. The certified value brackets $[v_{\mathrm{lower}}, v_{\mathrm{upper}}]$ have width at most $3.3\cdot 10^{-14}$. The flow residual of the stored vector and the repair shift, reported as diagnostics, are at most $2.2\cdot 10^{-16}$ and $5.0\cdot 10^{-13}$.

Design	$[L_\Gamma, U](\Theta)$	η_{comp}	v_{upper} allow.	Plan L_1	ρ_{num}	ζ_s	$2\gamma_s$	Stat. LP
S1	[0.666667, 0.666667]	2.4e-15	3.3e-16	1.1e-16	2.8e-16	1.0e-12	6.7e-16	3.9e-16
S2	[0.491620, 0.491620]	2.6e-15	3.3e-16	2.2e-16	2.8e-16	1.0e-12	6.7e-16	3.9e-16
S3	[0.666667, 0.666667]	2.9e-15	3.3e-16	2.2e-16	2.8e-16	1.0e-12	6.7e-16	3.9e-16
S4	[0.061754, 0.061754]	3.4e-15	3.1e-16	4.4e-16	2.2e-15	1.0e-12	1.2e-14	5.5e-15
S5	[0.042935, 0.042935]	8.5e-15	8.6e-16	4.4e-16	3.7e-15	1.0e-12	1.4e-14	5.1e-15
S6	[0.326599, 0.326599]	1.1e-14	1.3e-15	2.2e-16	3.6e-16	1.0e-12	8.9e-16	4.4e-16
D5	[0.181976, 0.181976]	1.1e-13	9.0e-15	4.7e-15	4.1e-13	1.0e-12	2.3e-13	1.1e-13
D6	[0.131217, 0.131217]	6.8e-14	4.6e-15	5.0e-16	1.2e-14	1.0e-12	6.7e-14	6.7e-15

Source 4: acquisition error. Residuals of the child LP and of the Hellinger projections affect only which query is played. They are recorded for each witness and never enter $\widehat{U}$.

Table 7: Design catalog. Full RDI (the child LP (B.7), solved exactly at every witness) runs on S1–S6. It is too large to run on D5 and D6, where acquisition uses the pairwise separation score. The last two columns are setup costs, paid once and shared across methods and seeds. D5's raw alphabet has 40,320 symbols; lossless aggregation of identical symbols of the common background leaves 1017.

Design	Models	Queries	Symbols	Seeds/type	ε	η	Channels (s)	Subset cache (s)
S1	3	5	3	40	0.6	0.0	0.01	0.01
S2, ε=.2	3	5	3	40	0.2	0.0	0.01	0.01
S2, ε=.4	3	5	3	40	0.4	0.0	0.02	0.02
S3	4	5	3	40	0.6	0.0	0.04	0.04
S4	4	18	56	40	0.03	0.05	0.68	0.03
S5	4	19	64	40	0.02	0.0	0.37	0.04
S6	3	7	4	40	0.25	0.0	0.06	0.02
D5	6	12	1017	30	0.01	0.95	2.48	0.65
D6	6	46	300	30	0.03	0.85	3.91	0.46

F.8 Reproduction, protocol and reporting conventions

Software and hardware. All experiments run in Python with NumPy and SciPy, using HiGHS for every linear program and SLSQP for the Hellinger and information projections, on a Linux workstation with 20 logical cores and 31 GiB of memory; every stage is single-threaded apart from the solver's internal linear algebra calls. Versions: Python 3.10.0, NumPy 2.2.6, SciPy 1.15.3, pandas 2.3.3, Matplotlib 3.10.7.

Protocol and comparison units. Table 7 gives the main settings. Every method with finite channels uses $\delta = .05$, forced exploration $\gamma = .1$ toward a uniform query distribution, one of the pairwise, composite, frequency or intersection confidence modules, and the same certified deployment oracle, which returns an exactly feasible policy with a bound built from feasible primal and dual objects instead of solver residuals (HiGHS feasibility tolerances are 10^{-9}; Appendix F.7 explains how feasibility repair, payoff residual and optimization error enter the reported bound). The main cap is $N = 3000$ for S1/S2, 6000 for S4 and S5 and 10000 for D5/D6. The RDI separation parameter is $\kappa_{\text{sep}} = .25$; finite pools require no pricing tolerances. All response vertices are included, among them the mixed vertices of a κ-best-response polytope. Unless stated to be lexicographic or lookahead, the selector is the greedy evidence-minimizing one, and methods share random streams (both in Appendix F.6). Calibration and deployment use the specified synthetic model, so $\varepsilon_{\text{span}} = 0$. We report the first time the normal certificate passes; runs that fail, return an obstruction or reach the administrative cap keep their own status and contribute no event to the certification CDF. For means and paired comparisons, $T^N = \min(T, N)$ assigns N to every path without normal certification; a statistical failure is not treated as independent right censoring. On D5 and D6 every path certifies normally, so capped and conditional means coincide at every cap (23.7 episodes for Separation against 28.4 for Uniform on D5), and we draw no efficiency conclusion from D5's means.

Seed counts. The default is 40 independent seeds per true model on S1–S6 and 30 in the main D5/D6 runs. The coverage runs use 400 seeds per true model on S2 and 250 on S4–S6 (Appendix F.13); the S6 pairwise cells, the sweeps, transcript views, selector study and baseline study use 20. Every table states its own count.

Reporting conventions. A *median* is always $\inf\{v : F_n(v) \geq 1/2\}$, with F_n the empirical distribution of the event time over all n runs of the cell and a run without the event counted as $+\infty$: the $\lceil n/2 \rceil$-th smallest event time, hence an observed integer and never the average of two order statistics, and $+\infty$ exactly when fewer than half the runs produced the event (a *censored* median, reported with its censoring rate). A *capped mean* replaces each censored time by the cap; a *conditional* median is taken over the certified runs only; both are labeled as such wherever they appear. A *paired ratio* is $\exp\{\text{mean} \log(T_B^N / T_A^N)\}$ over runs matched on true model and seed; a ratio of medians is a different quantity and is labeled as such. Intervals use 2000 bootstrap resamples of independent seeds within each fixed model/configuration stratum, preserving method pairs.

Supplementary material. The supplementary material contains the code; the records of all runs behind every table and figure (one record per true model, seed and method, with the terminal status, both stopping times, the certificate, the realized gap, the eliminations by component, the count of each query and the wall-clock times), with checksums; and tests that re-check the numerical claims of this appendix on the designs, including the decisions of the acquisition rules in constructed states. Apart from wall-clock times, every reported quantity is a deterministic function of the fixed seed streams.

F.9 Q1 TABLES: THE STOPPING TARGET

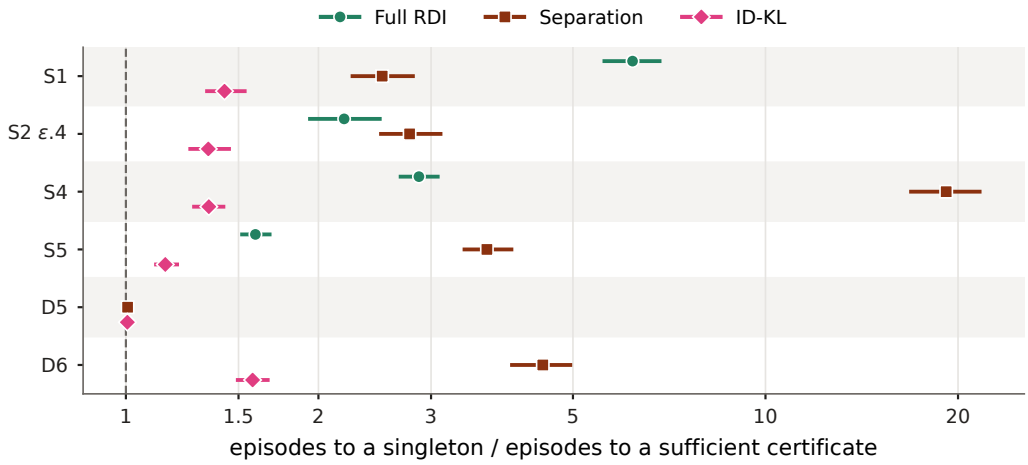

Figure 3: Q1. Paired ratio of the singleton time to the sufficient time, both read off the same trajectory, with 95% stratified bootstrap intervals. Composite module, greedy evidence-minimizing selector (Appendix F.13). Full RDI is not run on D5/D6.

Since the target is a set of models, we first measure the cost of requiring a single model, all else unchanged, recording both times on the same trajectory: after its sufficient certificate the algorithm continues until the confidence set is a singleton, with the same acquisition rule, weight of forced exploration, canonical witness and random streams (when every RDI score is zero the allocation is a point mass on the maximizer with the lowest index, so the continuation is a deterministic function of the trajectory). Each tracked trajectory ends as *singleton*, *budget censored* or *emptied after certificate*; only trajectories that reach both targets enter the paired ratio. No run on S1–S6, D5 or D6 ends with UNKNOWN or an obstruction certificate (every oracle these designs call is a finite linear program, and the optional full-universe obstruction test is disabled throughout), nor in STATISTICAL FAILURE: the runs execute Algorithm $1'$ of Proposition E.10, whose confidence set never empties. Its fallback fired on 73 of the 94,450 recorded runs (one run per seed, true model and method), mostly on S4: 12 of 2,000 in its baseline study, 6 of 3,000 in its coverage cells, 5 of 240 on the $\delta = .2$ arm of its confidence sweep, and one to three each on its main cells and its other sweep, selector and view arms; the rest are 4 of 3,600 on S2's coverage cells and single firings on six further designs. Every firing is a coverage failure, so these runs fall within the δ budget that Table 23 measures. We claim no accuracy for their output; under Algorithm 1 they would have exited with STATISTICAL FAILURE and no policy.

Table 8 reports this contrast for each acquisition rule, and the result depends on the rule: the paired ratio ranges from 1.00 to 19.17 (on S4, 2.87 for full RDI, 1.35 for ID-KL and 19.17 for the separation rule; on D5, 1.01 for both rules, because the transcripts identify the model as soon as they certify). The ratio compares two stopping targets under a fixed acquisition and continuation rule, on paired trajectories, and so measures the extra cost of requiring a single model while keeping the rule one already uses. No rule here is optimal for either target, and a rule designed to reach a singleton would choose different queries after the certificate, so the ratio does not compare the optimal sample complexities of the decision and identification problems. A rule aimed at the decision certifies early and then makes slow progress toward identification. On S4, Separation certifies at a median of 32

Table 8: Contrast of the stopping targets alone, composite module throughout, greedy evidence-minimizing selector (a heuristic that need not attain the worst case; Appendix F.13), $\kappa_{\mathrm{sep}} = .25$. "Sufficient"/"Singleton" are the median sufficient/singleton times on the same trajectory. Every tracked run reached both targets and every continuation ended as a singleton, so "Runs" is also the number of certified, singleton and paired runs. "Later": fraction of runs whose singleton time strictly exceeds the sufficient one.

Design	Acquisition	Runs	Sufficient	Singleton	Paired ratio [95% CI]	Later
S1	Full RDI	120	18	111	6.20 [5.56, 6.88]	1.00
	Separation	120	81	197	2.52 [2.25, 2.83]	1.00
	ID-KL	120	112	149	1.43 [1.33, 1.54]	0.95
S2, ε=.2	Full RDI	120	98	98	1.00 [1.00, 1.00]	0.00
	Separation	120	168	168	1.00 [1.00, 1.00]	0.00
	ID-KL	120	149	149	1.00 [1.00, 1.00]	0.00
S2, ε=.4	Full RDI	120	35	97	2.19 [1.93, 2.51]	0.61
	Separation	120	85	275	2.78 [2.49, 3.13]	0.80
	ID-KL	120	115	149	1.35 [1.25, 1.46]	0.70
S4	Full RDI	160	36	98	2.87 [2.67, 3.09]	1.00
	Separation	160	32	624	19.17 [16.77, 21.77]	0.99
	ID-KL	160	76	98	1.35 [1.27, 1.43]	0.59
S5	Full RDI	160	87	126	1.59 [1.51, 1.69]	0.52
	Separation	160	95	348	3.67 [3.36, 4.03]	0.56
	ID-KL	160	112	123	1.15 [1.11, 1.21]	0.48
D5	Separation	180	21	21	1.01 [1.00, 1.01]	0.02
	ID-KL	180	20	20	1.01 [1.00, 1.02]	0.01
D6	Separation	180	51	217	4.49 [3.99, 4.99]	0.96
	ID-KL	180	57	79	1.58 [1.49, 1.68]	0.64

but needs 624 to isolate a model: once the surviving set is $\{0, 2\}$ (radius .00303) or $\{1, 3\}$ (.00443), no active pair conflicts at $\varepsilon = .03$, and the fallback of Table 4 makes Separation query λ_0 for the rest of the trajectory, so 19.17 is the cost of running a rule aimed at the decision, with its uniform fallback, until it reaches a singleton. On S5 Separation certifies at 95 and needs 348; on 80 of its 160 runs (every run whose true model is 2 or 3) the certified set is exactly $\{2, 3\}$, of radius .0067, the detectors of these two models differing observably only at cell 2. For ID-KL, a rule aimed at identification, the paired ratio lies between 1.00 and 1.58 on every design of Table 8. A comparison across methods compares whole algorithms and says nothing about the stopping target.

On S4 we can also vary the task instead of the rule, since its subset radii cover a range of tolerances: Table 9 sweeps ε across the threshold at which the compatible groups $\{0, 2\}$ and $\{1, 3\}$ disappear. Below .003 the learner must identify the model, at about three times the cost; between .005 and .05 the median is flat, because the compatible structure does not change again until .053. Varying the tolerance changes the witness and the active pairs as well as the stopping target, so this sweep is about the task; the ablation of the stopping target alone is Table 8.

The same table covers the other sensitivity questions. Forced exploration costs under fifteen per cent at the operating point and about a factor of three in the limit: the median rises from 30 to 34 between $\gamma = 0$ and $\gamma = .1$, and to 88 at $\gamma = 1$, where the rule is Uniform by construction. Tightening δ from .2 to .001 multiplies the median by 3.8 (19 to 72) while $\log(1/\delta)$ grows by a factor of 4.3: the empirical scaling follows the δ-dependence of the deadline, although the deadline's absolute value is a worst-case sufficient condition. Enlarging the response polytope through κ reduces structural information monotonically (Proposition D.6); here the empirical times are also monotone, from 34 at $\kappa = 0$ to 145 at $\kappa = .04$, although the theory does not require this under a greedy selector.

F.10 Q2 TABLES: ACQUISITION

We draw three conclusions from Tables 10, 21 and 13. First, on S4 and S5 every acquisition rule aimed at the decision (Full RDI, Separation, Group-KL and Plug-in KL) needs about a third of the episodes of uniform querying (for Full RDI 2.77 $[2.44, 3.11]$ and 3.24 $[2.92, 3.62]$), under every inference module and every selector we tried (Table 22). Second, ID-KL, which aims at identification, costs more than full RDI on the same design (paired ratios 2.03 $[1.81, 2.31]$ on S4 and 1.32

Table 9: Parameter sweeps on S4 and S5, median episodes to normal certification, 20 seeds per type, composite module, episode cap 20,000. The fourth column is the maximal compatible cover at that tolerance, enumerated over all nonempty subsets. On S5 the cover overlaps at $\varepsilon = .02$ and .04, so it is not a partition; $\Gamma(\Theta) = .0429$ on S5, so any larger tolerance would certify at episode zero and is not swept. The observation stream of a query does not depend on the cap (Appendix F.6), so a cell with 20 seeds here is a sub-sample of the corresponding cell with 40 seeds in the main tables, and any difference between them is due to sample size.

Design	Parameter	Value	Maximal compatible sets	Full RDI	Separation	Group-KL	ID-KL	Uniform
S5	ε	0.005	{0}; {1}; {2}; {3}	122	143	119	119	572
S5	ε	0.01	{2,3}; {0}; {1}	78	102	89	110	285
S5	ε	0.02	{1,2}; {2,3}; {0}	78	103	98	107	285
S5	ε	0.03	{1,2,3}; {0}	78	85	75	94	234
S5	ε	0.04	{1,2,3}; {0,2}	78	84	75	94	234
S4	ε	0.002	{0}; {1}; {2}; {3}	99	108	90	90	604
S4	ε	0.005	{0,2}; {1,3}	34	33	33	74	88
S4	ε	0.011	{0,2}; {1,3}	34	33	33	74	88
S4	ε	0.03	{0,2}; {1,3}	34	33	33	74	88
S4	ε	0.05	{0,2}; {1,3}	34	33	33	74	88
S4	γ	0	–	30	30	–	–	88
S4	γ	0.05	–	31	31	–	–	88
S4	γ	0.1	–	34	33	–	–	88
S4	γ	0.3	–	38	34	–	–	88
S4	γ	1	–	88	88	–	–	88
S4	δ	0.001	–	72	62	–	–	200
S4	δ	0.01	–	50	43	–	–	133
S4	δ	0.05	–	34	33	–	–	88
S4	δ	0.2	–	19	19	–	–	47
S4	κ	0	–	34	33	–	–	88
S4	κ	0.01	–	50	41	–	–	144
S4	κ	0.04	–	145	145	–	–	582

$[1.21, 1.44]$ on S5), in agreement with Q1. Third, the comparison between the exact child LP and the cheap pairwise separation score depends on the separation fraction κ_{sep}, which Table 13 varies with the witness rule fixed at its canonical value: at $\kappa_{\text{sep}} = 0$ the rule is the unconstrained RDI argmax, and at $\kappa_{\text{sep}} = 1$, the tightest floor, it coincides with Separation on every instance with protected pairs except S5, where it certifies at 92 against 95. The two rules differ: they optimize different objectives (child impurity versus certified log-growth) over different constraint sets (conflicting pairs inside the witness versus active conflicting pairs of the whole confidence set), and they differ where the protected set is empty. The last column of Table 13 shows where the floor can act; on S1 and S3 it has no effect, and Separation coincides with Uniform (Appendix F.4). Against pairwise separation with the floor released (Table 21), the exact decision score does not beat the separation score on S4, although it costs orders of magnitude more per distinct confidence set (Table 29); it beats it on S1 and on the graded S2 instances, and on S5 by a smaller margin, reported next.

Replication on a second two-sided game. We designed S5, and chose its operating tolerance $\varepsilon = .02$, after S4 had given a null result for the comparison of full RDI with separation, without pre-registration: S5 is a targeted replication, a second two-sided instance built to differ from S4. Exactly one comparison is designated primary; every other paired ratio in the paper is exploratory and uncorrected. With the main protocol (40 seeds per model, cap 6,000, 160 paired runs), the comparison of full RDI with the pairwise separation score gives 1.199 $[1.119, 1.290]$ at $\varepsilon = .01$, 1.202 $[1.120, 1.292]$ at .02 (full RDI faster by 1.20 $[1.12, 1.29]$), 1.039 $[0.921, 1.164]$ at .03 and 1.037 $[0.919, 1.163]$ at .04; the median of full RDI is 87 at every tolerance and the separation medians are 94, 95, 79 and 79. The cell at .02 is the primary endpoint of Table 21; the cells at .01, .03 and .04 appear only here, with records in the supplementary material. At .02 the pairwise, composite, frequency and intersection modules give 1.16, 1.20, 1.21 and 1.15, every interval excluding one, although the modules' own medians range from 75 to 108; on S4 the same comparison gives 0.90 to 0.93, every interval covering one. The sweep of Table 9, with 20 seeds, gives 1.31 $[1.18, 1.46]$

Table 10: Median episodes to normal certification for every small design and every inference module, 120 runs per cell on S1, S2 and S6 (3 models $\times$ 40 seeds; the S6 pairwise cells use 20 seeds) and 160 on S4 and on S5 (4 models $\times$ 40 seeds). Greedy evidence-minimizing selector. ε is the value in the row label if there is one, and otherwise the design default of Table 7: .6 on S1, .03 on S4 and .02 on S5. $\kappa_{\text{sep}} = .25$ for Full RDI and 0 for the variant with the floor released. Full RDI solves the exact child LP at every witness; Separation is the pairwise score used by PS-EASI.

Design	Module	Full RDI	RDI, κ_{sep}=0	Separation	Plug-in KL	Group-KL	ID-KL	Uniform
S1	pairwise	1,405	1,405	144	84	84	84	144
S1	composite	18	18	81	112	112	112	81
S1	frequency	18	18	83	109	109	109	83
S1	comp.+freq.	20	20	92	126	126	126	92
S2, ε=.2	pairwise	331	1,479	131	123	123	123	245
S2, ε=.2	composite	98	111	168	149	149	149	197
S2, ε=.2	frequency	122	124	169	150	150	150	235
S2, ε=.2	comp.+freq.	115	135	196	184	184	184	217
S2, ε=.4	pairwise	212	1,418	71	93	69	93	152
S2, ε=.4	composite	35	21	85	115	92	115	94
S2, ε=.4	frequency	29	18	88	115	91	115	83
S2, ε=.4	comp.+freq.	41	24	108	139	109	139	111
S4	pairwise	28	28	26	26	26	61	77
S4	composite	36	36	32	32	32	76	102
S4	frequency	36	36	32	32	32	76	103
S4	comp.+freq.	41	41	40	40	40	89	124
S5	pairwise	75	75	73	79	80	87	212
S5	composite	87	87	95	95	99	112	281
S5	frequency	88	88	93	91	92	102	282
S5	comp.+freq.	108	108	115	110	106	130	336
S6	pairwise	∞	–	∞	–	–	–	∞
S6	composite	9	9	54	9	9	9	54
S6	frequency	13	13	73	13	13	13	73
S6	comp.+freq.	9	9	59	9	9	9	59

at $\varepsilon = .02$; as a sub-sample of the same paths, it differs from 1.20 through sample size. The effect therefore appears at $\varepsilon = .01$ and .02, its interval covers one at .03 and .04, and the design suggests why. A pair is *active* at tolerance ε when $\Gamma(\{i,j\}) > \varepsilon$. S5's pair radii (Appendix F.2), .00670 for $\{2,3\}$ and at least .03437 for every pair containing model 0, place the relevant thresholds at $\Gamma(\{1,2\}) = .01714$ and $\Gamma(\{1,3\}) = .02334$. Below .02334 at least one active conflicting pair does not contain model 0; at or above it every active pair contains model 0, so the only question left for either rule is whether the truth is model 0, and the deployment pricing of the child LP has nothing to exploit beyond what the equalization of the pairwise score already uses. The measured ratios follow this order: the times of the separation arm change on 159 of the 160 runs between $\varepsilon = .02$ and .03, where the threshold is crossed, but on only 5 between .03 and .04. We offer this as a mechanism for two instances; it was not the criterion by which $\varepsilon = .02$ was chosen. The comparisons against Uniform and ID-KL replicate without qualification: their ratios above hold on S5 as on S4.

In the witness ablation of Table 12, replacing the canonical witness by the full-set witness, a multi-witness mixture or a random label permutation does not change the S4 median, as expected: S4's least-favorable prior is unique (Appendix F.1), so a label permutation has no freedom to use. S5 shows the same more strongly: all three variants give trajectories identical to the canonical rule's, although the witness is not constant as the confidence set shrinks (Appendix F.2). The choice of witness matters exactly when the model class contains duplicates that are equivalent for deployment, which S3 is built to isolate: there the least-favorable prior is not unique (the canonical rule returns $\{0,1,2\}$ or $\{1,2,3\}$ depending on the label order), and the permuted and multi-witness variants move the median in opposite directions under the two modules, with paired ratios 0.75 and 0.56 under pairwise inference, where both variants beat the canonical rule, and 1.34 and 1.75 under composite inference, where the canonical rule beats both. Canonical lexicographic tie-breaking makes the algorithm well defined, and the results stay sensitive to it: no choice among these witnesses is best across inference modules. On S4 the true model is outside the initial witness $\{0,1\}$ whenever $\theta^\star \in \{2,3\}$, a situation the floor cannot fix (Appendix F.1). The full-set variant keeps every model

of the current confidence set in W with the same exact least-favorable prior, not restricted to its support, so models with zero weight stay in W and contribute protected pairs. Baselines with a fixed query limit what these comparisons can show.

Table 11: Baselines with a fixed query, 40 seeds per true model (120 paired runs per cell), greedy evidence-minimizing selector. Columns three to eight are medians of full RDI and of each fixed rule $q_0, \ldots, q_4$, which plays the named query at every round with the same forced exploration $\gamma = .1$. "Full RDI" is the default arm $\kappa_{\text{sep}} = .25$; the released arm $\kappa_{\text{sep}} = 0$ is the reference of Table 21. The last two columns are paired ratios (point estimates) of the best and of the worst fixed query to full RDI; a value below one means the fixed query is faster. Best and worst are selected *ex post* among the five queries, so neither is an implementable rule.

Design	Module	Full RDI	q_0	q_1	q_2	q_3	q_4	Best/RDI	Worst/RDI
S1	pairwise	1,405	1,405	78	70	101	200	0.05	1.00
S1	composite	18	18	102	85	129	252	1.00	12.46
S1	frequency	18	18	103	88	135	254	1.00	13.70
S2, ε=.2	pairwise	331	2,363	150	1,343	1,280	1,143	0.45	6.55
S2, ε=.2	composite	98	2,082	177	1,722	1,661	1,295	1.71	10.89
S2, ε=.2	frequency	122	1,798	180	845	919	650	1.46	4.84
S2, ε=.4	pairwise	212	1,608	100	71	101	1,143	0.32	6.74
S2, ε=.4	composite	35	21	130	85	132	1,295	1.53	18.27
S2, ε=.4	frequency	29	18	135	88	135	619	0.61	18.31

On S1 every pair of models has radius $1/2 < \varepsilon = .6$, so eliminating any one model suffices; until then the confidence set, witness and RDI score stay constant and RDI plays q_0 every round. Under every inference module fixed q_0 therefore matches full RDI exactly, and under the pairwise module the best fixed query beats full RDI (Table 11); S1 thus tests only which initial query is chosen. On S2 at $\varepsilon = .2$ all three pairwise radii exceed the tolerance and the witness changes after each elimination; full RDI with the composite module (the default arm of Table 11) beats every fixed query, with paired ratio 10.89 $[8.21, 14.37]$ against the worst query (q_0) and 1.71 $[1.49, 1.95]$ against the best (q_1); only the second is a claim about the value of acquisition. At $\varepsilon = .4$ fixed q_0 has the smaller median (21 against 35) but a paired log-mean ratio of 1.53 $[1.10, 2.16]$ against full RDI; the two statistics disagree in sign, which is why both are tabulated. The fixed query is faster on 70.8% of the 120 paired runs, with no ties, but loses on the mean because of a heavy upper tail: its quantiles at $(.25, .5, .75, .9, .95)$ are $(18, 21, 1{,}217, 2{,}368, 2{,}961)$, against RDI's $(28, 35, 82, 136, 199)$, and it reaches the 3,000 cap, while RDI's slowest run stops at 327. Under the composite and frequency modules the exact decision score beats the best fixed query chosen ex post on S2 at $\varepsilon = .2$ (paired ratios 1.71 and 1.46). At $\varepsilon = .4$ the composite comparison is mixed (the fixed query is faster in median, RDI in paired log-mean), and under the frequency module the best fixed query is faster (0.61); under the pairwise module the best fixed query is faster at both tolerances (0.45 and 0.32). On S1 the exact score is identical to the best fixed query under the composite and frequency modules and beaten by it under the pairwise module. The exact score therefore does not uniformly beat an oracle that picks the best query of the pool ex post. The spread across fixed queries is large (the paired ratio of the worst to the best query is 18.8 under the pairwise module and 12.5 under the composite one on S1, and 14.5 and 6.4 on S2 at $\varepsilon = .2$), so we compare against the best.

Table 12: Witness ablation. Entries are medians with the paired ratio to the canonical rule of full RDI in parentheses; a ratio below one means the variant is faster.

Design	Full RDI	RDI, full-set witness	RDI, multi-witness	RDI, permuted labels	Separation
S4, composite	36	36 (1.00)	36 (1.00)	36 (1.00)	32 (0.92)
S5, composite	87	87 (1.00)	87 (1.00)	87 (1.00)	95 (1.20)
S3, pairwise	118	118 (1.00)	84 (0.56)	94 (0.75)	128 (0.70)
S3, composite	71	71 (1.00)	100 (1.75)	90 (1.34)	128 (2.20)

Table 13: Ablation of the separation floor at a fixed witness rule, composite module, greedy evidence-minimizing selector. "Full RDI" is the default arm $\kappa_{\text{sep}} = .25$; "RDI, κ_{sep}=0" is the released arm, used as reference in Table 21. The last column is the mean number of protected pairs seen by RDI. Where it is zero, κ_{sep} has no effect.

Design	RDI, κ_{sep}=0	Full RDI	RDI, κ_{sep}=1	Separation	Uniform	Mean protected pairs
S1	18	18	18	81	81	0.00
S2, ε=.2	111	98	168	168	197	1.38
S2, ε=.4	21	35	85	85	94	1.00
S4	36	36	32	32	102	1.00
S5	87	87	92	95	281	1.59
S3	71	71	71	128	128	0.00

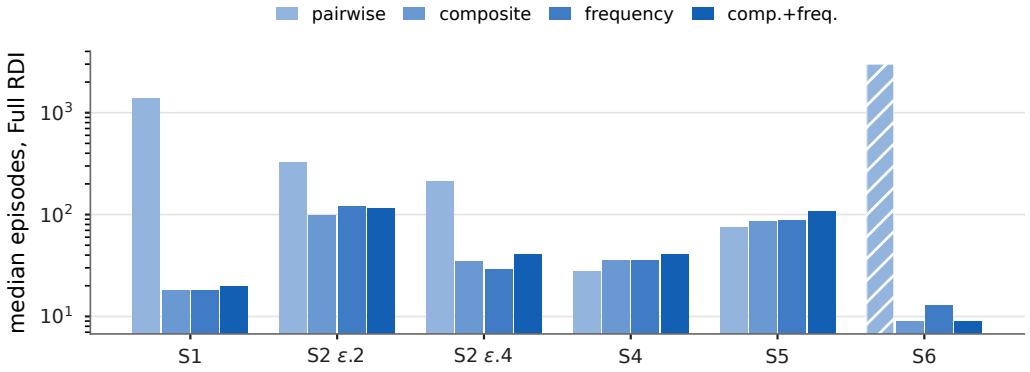

Figure 4: Q3. Median episodes to normal certification for full RDI under each inference module. No module is best everywhere. Adding the composite event bets helps by about two orders of magnitude on S1 and by factors of 3.4 and 6.1 on S2 (at $\varepsilon = .2$ and .4), and costs 16–29% on S4 and S5. The hatched bar (S6, pairwise) is a censored median: no run certifies.

F.11 Q3 TABLES: INFERENCE MODULES AND MISSPECIFICATION

Table 10 compares the pairwise, composite and frequency modules (the last given by Theorem 4.4) on every small design, and Table 11 adds the baselines with a fixed query under each. Adding the composite event bets to the pairwise e-process gains two orders of magnitude on S1 (18 episodes against 1,405) and a factor of 3.4 on S2 at $\varepsilon = .2$ (98 against 331), and costs 29% more on S4 (36 against 28) and 16% more on S5 (87 against 75), where the pairwise projections are already informative.

The S1 case has a mechanism that can be shown directly by removing every other query. Table 14 lists the anchored conflict sets of Definition B.10, the sets Ξ with $\theta^\star \in \Xi$, $2 \le |\Xi| \le p + 2$ and $\Gamma(\Xi) > \varepsilon$ (without this filter the counts would be too large): three on S1 and twenty-four each on S4 and S5. None of the conflicts of S4 or S5 is pairwise-blind, so on these designs the frequency module has no structural gap to exploit, and Table 10 gives no evidence about such a gap. Restricting S1's pool to $\{q_0\}$ makes all three conflicts *pairwise-blind*: the best certified pairwise log-growth is zero to machine precision, while the joint conflict separation is $\varsigma = 0.35$.

The deadline (E.5) **for each design.** The frequency module's stopping guarantee is an explicit episode count. With ς_0 given in Table 14 and $|\mathcal{Q}_{\text{acc}}|$, $S = \max_q |\mathcal{Z}_q|$ and $M = |\Theta|$ in Table 7, Table 15 evaluates the count exactly from printed constants, without new runs. With uniform λ_0, as in every run, $\lambda_0^{\min} = 1/|\mathcal{Q}_{\text{acc}}|$ and $\gamma = .1$, so $\gamma\lambda_0^{\min} = .1/|\mathcal{Q}_{\text{acc}}|$. A total of .05 is split evenly between the two budgets: the boundary (E.3) is evaluated at $\delta = .025$ and the budget for the query counts is $\delta' = .025$, for a nominal $1 - \delta - \delta' = .95$. The exploration term $8\log(|\mathcal{Q}_{\text{acc}}|/\delta')/(\gamma\lambda_0^{\min})$ never binds here (9.5×10^3 on S4, against 1.08×10^7 for the $N^\star$ term), so $n^\dagger = \lceil \frac{5}{4} \cdot 2N^\star(\varsigma_0)|\mathcal{Q}_{\text{acc}}|/\gamma \rceil = \lceil 25\, N^\star(\varsigma_0)\, |\mathcal{Q}_{\text{acc}}| \rceil$ in every row. The split does not explain the

Table 14: Anchored conflict sets, enumerated with the defining filter $\Gamma(\Xi) > \varepsilon$. "Pairwise-blind" counts the sets whose best certified pairwise log-growth is at most 10^{-9} at every query. A growth column entry such as $-1.00\mathrm{e}{-12}$ is a floating-point zero, not a negative divergence: Appendix E.5 treats growth in $[-10^{-9}, 0]$ as numerical zero.

Design	Anchored conflicts	$\lvert\Xi\rvert$	$\min \Gamma(\Xi)$	$\min_{\Xi} \varsigma(\Xi)$	min pairwise growth	Pairwise-blind
S1	3	3–3	0.6667	0.3500	3.90e-02	0
S1 pool $\{q_0\}$	3	3–3	0.6667	0.3500	-1.00e-12	3
S2, ε=.2	9	2–3	0.3259	0.1050	2.88e-02	0
S2, ε=.4	5	2–3	0.4444	0.1350	3.90e-02	0
S4	24	2–4	0.0532	0.0471	1.16e-01	0
S5	24	2–4	0.0233	0.0585	4.06e-02	0
S6	3	3–3	0.3266	0.4000	-1.00e-12	3

size: evaluating the boundary at the runs' own $\delta = .05$ and taking $\delta' = .05$ lowers every $n^\dagger$ by only 2.2% to 4.1%.

Table 15: Quantity (a): the deadline (E.5) evaluated for each design at $\gamma = .1$, uniform λ_0 and $\delta = \delta' = .025$, against the realized median of the frequency module; $N^\star(\varsigma) = \min\{N \geq 1 : 2\beta_N < \varsigma\}$ with β_N the boundary (E.3). The deadline exceeds the realized median by a factor between 6×10^2 and 4×10^5 on every design. The gap is largest on the two two-sided designs in the table, S4 and S5 (S6 is also two-sided but is not included).

Design	$\lvert\mathcal{Q}_{\mathrm{acc}}\rvert$	S	M	ς_0	$N^\star(\varsigma_0)$	$\gamma\lambda_0^{\min}$	$n^\dagger$	Realized median
S1	5	3	3	0.3500	323	0.0200	40,375	18
S1, pool $\{q_0\}$	1	3	3	0.3500	293	0.1000	7,325	12
S2, ε=.2	5	3	3	0.1050	4,541	0.0200	567,625	122
S2, ε=.4	5	3	3	0.1350	2,627	0.0200	328,375	29
S4	18	56	4	0.0471	30,025	0.0056	13,511,250	36
S5	19	64	4	0.0585	19,041	0.0053	9,044,475	88

The paper reports two such large episode counts, measuring different things. *Quantity (a)* is $n^\dagger(\theta^\star)$ of (E.5) in Table 15, the episode count by which the run is guaranteed to stop normally, driven by the separation ς_0 of the hardest anchored conflict set and ranging from 7.3×10^3 to 1.35×10^7. *Quantity (b)*, in the misspecification discussion below, is the episode count at which the sufficient condition $2\beta_N < D$ would allow eliminating a single law at the measured library distance D, roughly 1.2×10^4 to 3.7×10^6. The two use different separations and count different events, and only (a) includes the $5/4$ schedule factor; both show that the guarantee lies orders of magnitude beyond the point where certification happens.

Two response regimes. This design is the best place to compare the statistical modules: it has one query (so acquisition cannot differ) and every pairwise e-factor is identically one. The selector must, however, be held fixed, since under the greedy adversary each module faces a different response sequence; Table 16 reports both regimes. Under the greedy adversary the composite and intersection modules both certify at a median of 14 and the frequency module at 12; with the response vertex held fixed, so that all four modules see the same observations, the composite and intersection modules certify at 5 and the frequency module again at 12, because its own evidence component is the pairwise one, whose log-growth is identically zero here, so its selector objective is constant. The difference between 14 and 5 is the adversary's effect on the data. The pairwise module certifies on none of the 120 paths in either regime. Equal medians do not mean equal runs: the composite module eliminates the true model on 2 of its 120 runs under the greedy adversary and on 1 under the fixed vertex, the intersection module on 0 runs in both regimes, because its halved budget raises the composite component's thresholds enough to protect the truth on exactly those runs. The frequency component cannot eliminate anything in the first four observations: with $Q = 1$, $S = M = 3$ and its share $\delta/2 = .025$ in the intersection module, the boundary (E.3), which decreases in N, is $\beta_4 = \sqrt{\frac{1}{8}\log\frac{2\cdot4\cdot5\cdot1\cdot3\cdot3}{.025}} \approx 1.094$, while the ℓ_∞ distance between two probability vectors never

Table 16: S1 with the query pool restricted to $\{q_0\}$, 120 runs per module, uniform acquisition. Every anchored conflict is pairwise-blind (joint separation $\varsigma = 0.35$). Under "Greedy" a row describes the module together with the behavior it induces; under "fixed vertex" the four modules share one observation stream per (true model, seed), and the rows compare the tests alone. "Elim. comp."/"Elim. freq." count the models removed by the composite/frequency component.

Regime	Module	Runs	Certified	Median	Elim. comp.	Elim. freq.	Truth elim.
greedy	pairwise	120	0.000	∞	0	0	0
greedy	composite	120	1.000	14	120	0	2
greedy	frequency	120	1.000	12	0	120	0
greedy	comp.+freq.	120	1.000	14	120	0	0
fixed vertex	pairwise	120	0.000	∞	0	0	0
fixed vertex	composite	120	1.000	5	120	0	1
fixed vertex	frequency	120	1.000	12	0	120	0
fixed vertex	comp.+freq.	120	1.000	5	120	0	0

exceeds 1; over all 240 runs of the intersection module, all 240 eliminations are attributed to the composite component and none to the frequency test.

What the selector uses. The greedy objective (Appendix F.6), used unchanged for every module, has three inputs: the surviving set C_n, the running log-evidence $\log \bar{E}_n$ and, when the evidence component has composite bets, the current betting mixture. It does not use the confidence budget: neither $\log \bar{E}_n$ nor the betting weights depend on δ, which enters only the elimination threshold $\log(1/\delta)$, so splitting δ between two components changes the selector only through C_n, by keeping a model in the set slightly longer. It also ignores the frequency component, which has no one-step log-growth to minimize, so the adversary here is greedy against the e-process only; this limits the selector, and it is one reason the regime with a fixed vertex gives the cleaner comparison. The table therefore supports only the structural claim that the pairwise module cannot finish here and the joint modules can, a direct test that the frequency module goes beyond the pairwise separation condition without timings from a pool where other queries help the pairwise process. Both joint modules certify on all 120 paths, so the frequency test reaches the outcome of the composite bets without the bets having to be chosen; we have no instance on which the frequency module certifies and the composite module does not. The claim for the frequency module is generality: validity and completeness for arbitrary finite channel polytopes, with a stopping condition that has a matching impossibility result (Theorem 4.4). With the acquisition rule fixed at the default arm ($\kappa_{\mathrm{sep}} = .25$) and only the module varied, the frequency module ties the composite module on S1 and S4 (18 and 36 episodes each), is one episode slower on S5 (88 against 87), beats it on S2 at $\varepsilon = .4$ (29 against 35), and loses on S6 (13 against 9) and on S2 at $\varepsilon = .2$ (122 against 98); all values are in Table 10.

No inference module in this paper detects a misspecified candidate library at the sample sizes where certification occurs. We generated S4 and S5 data from laws outside every candidate channel polytope, at a measured sup-norm distance, recording both the exit status and the realized deployment gap in the true perturbed game.

Perturbation. We perturb the game itself. The four S4 types are specified by a node weight vector w_θ, an evasion probability and a false-alarm base (Appendix F.1); for each base index $b \in \{0, 1, 2, 3\}$ one data-generating world replaces type b's parameters by $w' = \big(\mathrm{clip}(w_0 - \sigma), \mathrm{clip}(w_1 + \sigma), \mathrm{clip}(w_2 - \sigma)\big)$ and evasion$' = \min\{0.9, \text{evasion} + \sigma\}$, with $\sigma \in \{0.12, 0.40\}$ and clip the truncation to $[0.05, 0.95]$. The payoff matrix, the attacker's best response to each announced query in the perturbed game and the transcript law (mixed with the uniform law at $\eta = .05$, as in the unperturbed design) are recomputed; the candidate library is left unchanged. On S5 one candidate's detection vector is perturbed coordinatewise by σ and the common clutter rate raised by $\sigma/2$. Each world is run with 60 seeds under Full RDI, Separation and Uniform. The distance to the library, $D = \max_q \min_\theta \mathrm{dist}_\infty(P_q, \mathcal{P}_\theta(q))$, is positive exactly when some single query rules out every candidate, so it is a sufficient but not a necessary indicator of misspecification: a truth outside the library can have $D = 0$ and be ruled out only by several queries together. A test of library adequacy would have to target the weaker quantity $\min_\theta \max_q \mathrm{dist}_\infty(P_q, \mathcal{P}_\theta(q))$.

Table 17: Diagnostic misspecification, 720 runs per row (4 perturbed worlds × 60 seeds × 3 acquisition rules), greedy evidence-minimizing selector, on S4 and S5. The data-generating law is outside the candidate library, so no coverage claim applies. $D = \max_q \min_\theta \mathrm{dist}_\infty$, shown as its range over the four perturbed worlds; $\widehat{U}$ is the returned certificate. Support elimination and the event bets eliminated nothing; the frequency test eliminated a model in 2 of the 4,320 runs whose module includes it (both on S4); every other elimination is pairwise. Only the (design, module) combinations listed were run.

Design	Module	σ	D range	Elim./run	1st elim.	Med. gap	$\Pr[\text{gap} > \varepsilon]$	$\Pr[\text{gap} > \widehat{U}]$
S4	pairwise	0.12	0.053–0.056	2.11	49	0.1120	1.000	1.000
S4	pairwise	0.40	0.273–0.279	2.12	52	0.2156	1.000	1.000
S4	composite	0.12	0.053–0.056	2.12	67	0.1120	1.000	1.000
S4	composite	0.40	0.273–0.279	2.12	67	0.2156	1.000	1.000
S4	frequency	0.12	0.053–0.056	2.13	67	0.1120	1.000	1.000
S4	frequency	0.40	0.273–0.279	2.12	68	0.2156	1.000	1.000
S4	comp.+freq.	0.12	0.053–0.056	2.12	90	0.1120	1.000	1.000
S4	comp.+freq.	0.40	0.273–0.279	2.11	92	0.2156	1.000	1.000
S5	composite	0.12	0.095–0.112	2.28	66	0.0298	0.750	0.996
S5	composite	0.40	0.288–0.296	2.23	63	0.0763	0.750	0.997
S5	frequency	0.12	0.095–0.112	2.28	66	0.0298	0.750	0.996
S5	frequency	0.40	0.288–0.296	2.23	64	0.0763	0.750	0.997

Outcome. Every run certifies, under all four modules and both perturbation sizes. On S4 the deployed policy's realized gap in the true game exceeds $\varepsilon = .03$ on every run, and the returned certificate understates it on every run. The modules eliminate models (the confidence set ends at size 2 on 88.1% of runs and at size 1 on the rest, never at 4), but the component that could detect the misspecification, the frequency test, eliminates a model in only 2 of the 2,880 S4 runs whose module includes it, both at $\sigma = 0.40$ (one under each such module), and neither elimination prevents an unsafe deployment on its run, while across all 5,760 S4 runs the pairwise e-process makes 12,202 eliminations. Support elimination and the event bets contribute none, the confidence set is never emptied, and STATISTICAL FAILURE is never returned. The median first elimination, at episode 67–68 under the composite and frequency modules, precedes the median certification at 89–92 (from the run records, not tabulated), so the pairwise process discards candidates on data that no candidate explains. Since the frequency rows include a pairwise e-process at level $\delta/2$ (Appendix F.4), given the attribution they match, to that precision, what the pairwise component alone would have produced. The rule the implementation evaluates, $\mathrm{dist}_\infty(\widehat{P}_{q,N}, \mathcal{P}_\theta(q)) > \beta_N$, can fire at any N; the inequality $2\beta_N < D$ used in Theorem E.6 is only a sufficient condition, and it holds only after 1.2×10^4 to 3.7×10^6 episodes (quantity (b) above). At $D = .28$ it needs about 660 observations of a single query, roughly 11,900 episodes under a uniform $1/18$ share and 119,000 under the floor of forced exploration, $\gamma/|\mathcal{Q}_{\mathrm{acc}}| = 1/180$; at $D = .056$ it needs about 370,000 and 3.7 million episodes. The median certification time is near 90.

The S5 study replicates the finding and adds one detail. Its four worlds are at $D = 0.095$–0.112 for $\sigma = 0.12$ and $D = 0.288$–0.296 for $\sigma = 0.40$. Every run certifies, the frequency test fires in none of the 1,440 S5 runs whose module includes it, and the certificate understates the realized gap on 99.6% and 99.7% of them. The unsafe fraction is 0.75, split exactly by world: on the three worlds perturbed from models 1, 2 and 3 every run deploys a policy whose realized gap exceeds $\varepsilon = .02$ (median gaps between .030 and .055 in these worlds at $\sigma = 0.12$), and on the world perturbed from model 0 no run does (median gap .0005). The failure therefore depends on the direction of the perturbation, which the learner cannot observe. No module checks the candidate library: the coverage statement (4.2) is conditional on $\theta^\star \in \Theta$, and library adequacy must be established by the same external model audit that Section 4.1 requires for the span budget. In our view, a test of library adequacy is the most important missing part of the framework.

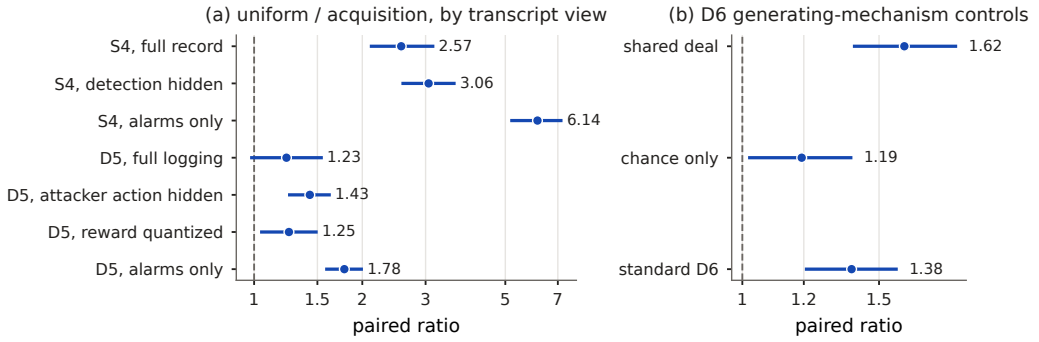

Figure 5: Q4. Paired ratio of uniform querying to the design's acquisition rule (full RDI on S4, pairwise separation on D5 and D6); a larger value means acquisition matters more. (a) Feedback restrictions that a logging system would impose. (b) D6 controls that hold one generating mechanism fixed. Neither isolates a purely strategic channel; the labels state what is held fixed.

F.12 Q4 TABLES: WHERE THE INFORMATION COMES FROM

Hiding a transcript field changes the observation law, so removing fields cannot by itself separate chance information from strategic information. Table 18 instead varies the generating mechanism with the two D6 families of Appendix F.6, each controlling exactly one component.

The family with a *shared deal distribution* controls the deal distribution and nothing else. It neither isolates response information nor freezes the responder: each model still best-responds to the announced query, so the response differs across models by construction, and two channels remain confounded with the deal, the recorded showdown reward, which the rank order determines, and the response itself. Over every query and model pair, the largest pairwise total variation of the stored laws is 0.080 at the first query, q_0, and 0.150 over the pool, attained at q_{27} between models 0 and 3. The stored law is $(1-\eta)$ times a physical mixture plus η times the uniform law, with $\eta = 0.85$, so the physical mixtures at q_{27} are at total variation $0.150/0.15 = 1$, that is, they have disjoint supports. The largest L_1 distance at q_0 before randomization is exactly $16/15 = 1.0667$, an L_1 distance, not a total variation. For these reasons we do not call the family "strategic only".

In the family "Chance only" the responder is fixed, so all channels are singletons, and an independent upper bound holds, which we check. The models differ in card counts and share one rank order, so the bound comes from data processing: with D_θ the law of the ordered deal of two cards under deck θ, the rank order common to all models, the responder frozen and the query policy announced, the whole transcript is the image of the deal under one Markov kernel common to all models, so $\mathrm{TV}(P^q_\theta, P^q_{\theta'}) \le \mathrm{TV}(D_\theta, D_{\theta'})$ for every query q. Enumerating the six decks exactly gives $\max_{\theta\neq\theta'} \mathrm{TV}(D_\theta, D_{\theta'}) = 4/7$, attained for instance between $(4,2,1)$ and $(1,2,4)$, so the stored laws satisfy $\mathrm{TV}(P^q_{\theta,\eta}, P^q_{\theta',\eta}) \le (1-\eta)\cdot\frac{4}{7} = 0.15\cdot\frac{4}{7} = \frac{3}{35} \approx 0.0857143$; the computed maximum, 0.0857, attains the bound. The bound fails for the family with a shared deal distribution, which does not freeze the responder. The frozen column is a best response in only 3 of the 276 (query, deck) cells, with mean excess payoff $.487$ over the best response, so this family changes the response mechanism itself; it lies outside the paper's behavioral assumption and serves as a control, showing what can still be learned when responses carry no information.

The advantage of query selection over uniform querying is largest with the shared deal distribution and smallest in the family "Chance only", so the benefit remains when the deal distribution is identical across models. The data leave open whether it comes from strategic response behavior, because with a shared deal the rank order is revealed in two ways, through the reward field and through the response itself, which that family does not hold fixed. We read Table 19 only as a statement about logging requirements. The second half of Q4 asks how much of D5's difficulty comes from record randomization. At $\eta = .95$, 95% of D5 records are drawn from the common background, and without randomization one full transcript identifies the type. Table 19 replaces randomization with restrictions that a logging system would impose.

Table 18: Controls of the generating mechanism on D6, $\varepsilon = .03$, 20 seeds per type. Each control row fixes one component, and neither isolates a purely strategic channel. "Shared deal distribution" holds the deal distribution identical across models; "Chance only" holds the responder fixed at a column outside the best-response face; "Both" is the standard D6, with nothing held fixed. "Mean face" is the mean number of columns in a best-response face.

Family	Mean face	Separation	Uniform	Uniform/Sep. ratio [95% CI]
Shared deal distribution	1.89	51	84	1.62 [1.39, 1.88]
Chance only (frozen responder)	1.00	65	85	1.19 [1.02, 1.38]
Both (standard D6)	3.55	45	62	1.38 [1.21, 1.58]

Table 19: Study that restricts the transcript. "Acquisition" is full RDI on S4 and pairwise separation on D5. 20 seeds per true model and an episode cap of 20,000. The S4 ratio with the full record, 2.57 $[2.12, 3.14]$, is therefore over 80 matched pairs and comes from a different study than the 2.77 $[2.44, 3.11]$ over 160 pairs of Table 21, which uses the 40 seeds and cap 6,000 of the main protocol. Hiding the attacker's realized action is a stronger restriction than record randomization, and it is the regime in which acquisition matters most.

Transcript view	Symbols	Acquisition	Median	Uniform	Uniform ratio [95% CI]
S4, full	56	Full RDI	34	88	2.57 [2.12, 3.14]
S4, detection hidden	28	Full RDI	334	1,092	3.06 [2.60, 3.60]
S4, alarms only	4	Full RDI	394	2,839	6.14 [5.21, 7.14]
D5, full logging	40320	Separation	16	21	1.23 [0.98, 1.54]
D5, attacker action hidden	2688	Separation	315	437	1.43 [1.26, 1.62]
D5, reward quantized	2880	Separation	37	45	1.25 [1.05, 1.49]
D5, alarms only	64	Separation	2,030	3,661	1.78 [1.59, 1.99]

The table supports only a comparison of the two ends of the range. On S4 the paired ratio is 2.57 $[2.12, 3.14]$ with the full record, 3.06 $[2.60, 3.60]$ with the detection report hidden and 6.14 $[5.21, 7.14]$ with only the two alarm bits; the first two intervals overlap, and the value with only the alarms is separated from both. On D5 full logging gives 1.23 $[0.98, 1.54]$, and the two intermediate restrictions give 1.43 $[1.26, 1.62]$ with the attacker's action hidden and 1.25 $[1.05, 1.49]$ with the reward quantized; full logging is the only view whose interval covers one, and these three intervals overlap. The view with only the alarms, at 1.78 $[1.59, 1.99]$, is separated from full logging and raises the median from 16 to 2,030 episodes; hiding the attacker's realized action alone raises D5's median by more than an order of magnitude, to 315. The intervals support two statements: on S4 and D5 the advantage of acquisition is largest under the poorest admissible view, and the intermediate views are not ordered. Two designs and one ordered endpoint do not support the general rule that acquisition matters most when feedback is scarce.

F.13 Larger benchmarks, selectors and coverage

Table 20 gives the D5 and D6 comparisons. Full RDI is not run on these two designs, whose child LPs have tens of thousands of variables per query; the exact score is evaluated on S4 and S5 instead. Composite tests increase the median for every D5/D6 method shown, so the composite module solves S1's case without being uniformly stronger. Changing the selector from lexicographic to greedy moves the medians by at most three episodes on either design (diagnostics below).

Selector ambiguity. The greedy evidence-minimizing selector ("greedy" in the selector columns; Appendix F.6) is a myopic heuristic and need not attain the worst case over admissible selectors. Coverage and the certificate are valid for every admissible selector, by Theorems B.16 and E.6, but every timing in this paper is under one named selector. Table 22 reports, in a study with 20 seeds, the structural quantities that determine whether a selector has any freedom (the largest number of vertices of a response face, the fraction of (query, model) cells whose channel is a singleton, and the number of distinct transcript laws), with timings under the lexicographic, greedy and two-step lookahead selectors. S1 has at most two vertices per face and a singleton channel in 80% of its cells, so all of its ambiguity is in the single query q_0; there the greedy adversary more than doubles

Table 20: Median episodes to normal certification, 180 runs per method, $\varepsilon = .01$ on D5 and $\varepsilon = .03$ on D6, under the algorithm definitions of Table 4. "Greedy" is the greedy evidence-minimizing selector of Appendix F.13.

Design	Selector	Module	Separation	Plug-in KL	Group-KL	ID-KL	Uniform
D5	lexicographic	pairwise	19	19	19	19	22
D5	greedy	pairwise	16	18	16	18	19
D5	greedy	composite	21	20	20	20	27
D6	lexicographic	pairwise	44	47	40	46	60
D6	greedy	pairwise	44	46	43	46	62
D6	greedy	composite	51	54	49	57	73

Table 21: Paired comparisons against full RDI with the separation floor released ($\kappa_{\text{sep}} = 0$, the rule with the floor released in Table 4), composite module, greedy evidence-minimizing selector, 120 runs per cell on S1 and S2 and 160 on S4 and S5. Ratios above one favor RDI. Intervals are stratified paired percentile bootstrap intervals over runs matched on true model and seed; the cap convention is the one stated in the protocol. This is the primary table of paired comparisons. Its row for S5, $\varepsilon = .02$, against Separation is the single primary endpoint of Section 6. This endpoint was designated after the S4 results were known (Appendix F.10). Tables 8, 19, 18, 11 and 12 also report paired ratios, with the same definition, and are exploratory.

Design	Comparator	Median	Paired ratio to RDI (κ_{sep}=0) [95% CI]	Pairs
S1	Separation	81	4.25 [4.03, 4.48]	120
S1	Group-KL	112	5.85 [5.25, 6.48]	120
S1	Plug-in KL	112	5.85 [5.25, 6.48]	120
S1	ID-KL	112	5.85 [5.25, 6.48]	120
S1	Uniform	81	4.25 [4.03, 4.48]	120
S2, ε=.2	Separation	168	1.42 [1.24, 1.63]	120
S2, ε=.2	Group-KL	149	1.35 [1.19, 1.52]	120
S2, ε=.2	Plug-in KL	149	1.35 [1.19, 1.52]	120
S2, ε=.2	ID-KL	149	1.35 [1.19, 1.52]	120
S2, ε=.2	Uniform	197	1.73 [1.51, 1.97]	120
S2, ε=.4	Separation	85	2.56 [2.16, 3.01]	120
S2, ε=.4	Group-KL	92	2.78 [2.36, 3.28]	120
S2, ε=.4	Plug-in KL	115	3.44 [2.91, 4.03]	120
S2, ε=.4	ID-KL	115	3.44 [2.91, 4.03]	120
S2, ε=.4	Uniform	94	3.14 [2.69, 3.69]	120
S4	Separation	32	0.92 [0.82, 1.03]	160
S4	Group-KL	32	0.92 [0.82, 1.03]	160
S4	Plug-in KL	32	0.92 [0.82, 1.03]	160
S4	ID-KL	76	2.03 [1.81, 2.31]	160
S4	Uniform	102	2.77 [2.44, 3.11]	160
S5	Separation	95	1.20 [1.12, 1.29]	160
S5	Group-KL	99	1.18 [1.08, 1.29]	160
S5	Plug-in KL	95	1.15 [1.06, 1.25]	160
S5	ID-KL	112	1.32 [1.21, 1.44]	160
S5	Uniform	281	3.24 [2.92, 3.62]	160

full RDI's median relative to lexicographic tie-breaking, from 7 to 18 (a factor of 2.6), by making that query uninformative. S4 has up to six vertices, a singleton channel in only 11% of its cells and 152 distinct laws, and the same adversary raises the median by three episodes, from 31 to 34. The two-step lookahead selector does not make the adversary stronger: it brings the algorithm back to 6 episodes on S1 and is within two episodes of the greedy rule on S4. Neither rule bounds the optimal nonmyopic selector.

Anytime coverage. The guarantee is per model: Theorem 4.1 states that for each $\theta^\star$ the simultaneous coverage event fails with probability at most δ. A failure rate pooled across models does not test this, so Table 23 reports the losses of every true model separately and Figure 6(a) plots every (design, method, true model) cell. The default of 40 seeds cannot test $\delta = .05$ (with zero

Table 22: Selector ambiguity diagnostics and timings. Composite module, $\varepsilon = .6$ on S1 and $\varepsilon = .03$ on S4, $\kappa_{\text{sep}} = .25$, 20 seeds per true model. This study is separate from the main tables with 40 seeds and uses the same settings. Where a row differs from Table 10 (S4 gives $(34, 33, 88)$ here against $(36, 32, 102)$ there, and S1 $(18, 80)$ against $(18, 81)$), the difference is due to the seed count, and the two should not be compared directly.

Design	Selector	Max face	Singleton fraction	Distinct laws	Full RDI	Separation	Uniform
S1	lexicographic	2	0.800	18	7	31	31
S1	greedy evidence-minimizing	2	0.800	18	18	80	80
S1	two-step lookahead	2	0.800	18	6	32	32
S4	lexicographic	6	0.111	152	31	31	83
S4	greedy evidence-minimizing	6	0.111	152	34	33	88
S4	two-step lookahead	6	0.111	152	32	34	88

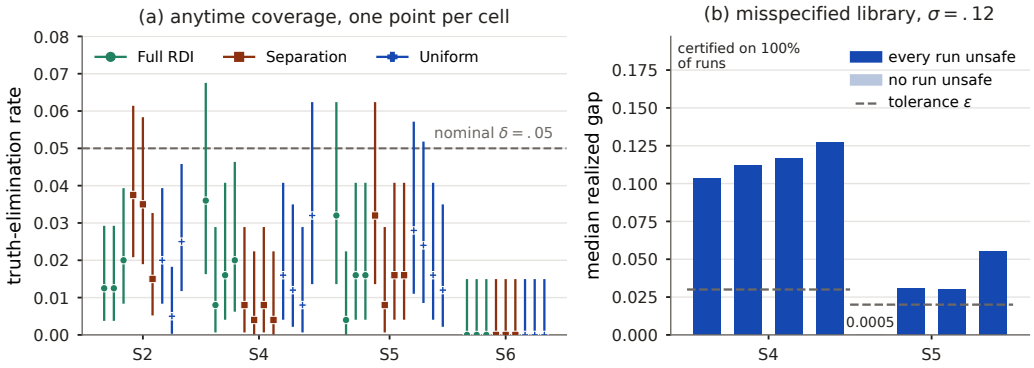

Figure 6: (a) Anytime coverage: one point per (design, method, true model) cell, with a Clopper–Pearson interval. Truth elimination is the complement of the simultaneous coverage event; intervals are per cell and uncorrected for multiplicity. (b) Misspecified candidate library at $\sigma = .12$, composite module: median realized gap of the deployed policy in each perturbed world, against that design's tolerance; every run certified.

failures, the one-sided 95% binomial upper bound is still 7.2%); hence the coverage seed counts of Appendix F.8. Within a cell the seeds are independent draws given the fixed model, method and configuration, so the binomial model applies; across cells the runs share seed streams by design, so the rows are not independent and are never pooled. The observed rate of every cell is below the nominal .05: nine of the 42 cells show no failure, and the rest lie between .0040 and .0375. Nine of the 42 intervals reach above .05, so the data are consistent with failure probabilities slightly above .05 in some cells, even though no point estimate is. With 42 cells the rates carry no simultaneous claim, and the greedy selector need not attain the worst case. Validity rests on the proof; the runs check the implementation and show how conservative it is here.

Truth elimination and bound violation. The flags in Table 23 are computed independently, from different objects. *Ever lost*: $\theta^\star$ was eliminated at some round, according to the raw elimination log. This is the complement of the simultaneous event $\{\theta^\star \in C_n \ \forall n\}$ that Theorem 4.1 controls. Membership at the stop can differ from it, because the never-empty replacement of Algorithm $1'$ can reinstate an eliminated true model as the singleton with the least evidence against it, which it does on 61 of the 94,450 recorded runs, 8 of them in these coverage cells. *Bound violation*: $\text{Gap}_{\theta^\star}(\widehat{x}_\tau) > U_\tau$, that is, the realized gap of the returned, executable policy in the true game exceeds the issued certificate; the gap is recomputed as $v_{\text{upper}}(\theta^\star) - \min_j(\widehat{x}_\tau^\top A_{\theta^\star})_j$ and never read from the elimination flag. $\text{Gap} > \varepsilon$: the same recomputed gap exceeds the tolerance, the quantity (B.17) constrains. The theory gives the one-sided inclusion $\{\text{Gap}_{\theta^\star}(\widehat{x}_\tau) > U_\tau\} \subseteq \{\theta^\star \text{ was removed from the confidence set}\}$, because while $\theta^\star \in C_n$ the certificate maximizes over a set that contains $\theta^\star$; the reverse inclusion does not follow. Over the 42 cells there are 179 truth eliminations and 171 bound violations: no run violates its certificate without losing the truth, as the inclusion requires, and 8 runs lose the truth without a violation. On all eight the raw set emptied, the

Table 23: Anytime coverage, one row per (design, method): greedy evidence-minimizing selector, composite module on S2, S4 and S5 and comp.+freq. on S6, 400 runs per true model on S2 and 250 on S4, S5 and S6. "Lost per $\theta^\star$" counts, for each true model in turn, the runs that eliminated the true model at some round, read off the raw elimination log (Appendix F.13). "Max rate" is the largest rate over true models, with the Clopper–Pearson interval of that cell, uncorrected for multiplicity. "Viol." counts runs whose recomputed true gap exceeds the issued certificate, "Lost only" losses without a violation, and the last column runs whose recomputed gap exceeds the tolerance itself.

Design	Method	Runs	Lost per $\theta^\star$	Max rate [95% CP]	Viol.	Lost only	Gap $> \varepsilon$
S2	Full RDI	400	5/5/8	0.0200 [0.0087, 0.0390]	18	0	18
S2	Separation	400	15/15/6	0.0375 [0.0211, 0.0611]	35	1	35
S2	Uniform	400	9/2/11	0.0275 [0.0138, 0.0487]	20	2	20
S4	Full RDI	250	9/2/4/6	0.0360 [0.0166, 0.0672]	20	1	19
S4	Separation	250	2/1/2/1	0.0080 [0.0010, 0.0286]	6	0	6
S4	Uniform	250	5/5/3/8	0.0320 [0.0139, 0.0621]	17	4	11
S5	Full RDI	250	8/1/4/4	0.0320 [0.0139, 0.0621]	17	0	17
S5	Separation	250	8/2/4/4	0.0320 [0.0139, 0.0621]	18	0	17
S5	Uniform	250	7/6/4/3	0.0280 [0.0113, 0.0568]	20	0	17
S6	Full RDI	250	0/0/0	0.0000 [0.0000, 0.0146]	0	0	0
S6	Separation	250	0/0/0	0.0000 [0.0000, 0.0146]	0	0	0
S6	Uniform	250	0/0/0	0.0000 [0.0000, 0.0146]	0	0	0

replacement fired, and the singleton it selected, the one with the least evidence against it, was the true model, so $C'_\tau = \{\theta^\star\}$, the issued certificate is $U_\tau = \Gamma(\{\theta^\star\}) = 0$ and the recomputed true gap is also 0: a coverage failure occurred earlier in the run, but the certificate holds exactly. A second mechanism can also make the inclusion one-way in general (another surviving model with the same deployment loss but a different diagnostic channel), but it did not occur in these runs. Because the confidence set never empties under Algorithm $1'$, a run that loses the truth returns a policy rather than exiting with STATISTICAL FAILURE, so the violations are visible in the data.

F.14 Computation

Online wall-clock times include acquisition and confidence updates; the deployment cache over all subsets and the channel construction count as setup. Table 24 separates the cost of a cold start from that of a warm cache, counts the deployment LPs, and reports the cache's size and hit rate. Precomputing all $2^M - 1$ nonempty subsets is cheap at these sizes (15 LPs on S4 and 63 on D6, both under 40 KiB). The acquisition cache matters more: cold, full RDI costs up to $13.5\times$ more per episode than warm (S5; $10.2\times$ on S4 and $2.2\times$ on S6), the largest in Table 24, while the pairwise rule and every other rule without RDI lose at most a factor of 1.2. Reported times per episode use a warm cache.

G Additional Experimental Results

G.1 Exact small strategic instances and canonicalization

S1 has three models $i = 0, 1, 2$, three deployment actions, and losses $1 - x_i$. The deployment payoff matrix has three identical columns with row i equal to one and the other rows zero. Calibration uses five reset states; the reset index announces the query. At every reset, action $j = i$ has diagnostic payoff one and $j \neq i$ has payoff zero, so the attacker's best-response face consists of mixtures of the latter two actions. The observation kernel with three symbols is the S1–S3 catalog of Appendix F.6; the reward and attacker action are not in this diagnostic signal. This construction embeds deployment and calibration in one finite game with initial states that depend on the reset and a common tree. The defender's only nontrivial deployment simplex has dimension two. On S1, full RDI therefore faces set-valued strategic diagnostic channels and an easy deployment game.

S2 replaces the deployment weights by $w = (1, .8, .55)$; all three coordinates affect loss. On an active set, the unique least-favorable prior is $\pi_i = (1/w_i)/\sum_j 1/w_j$, and the witness that preserves the radius is that active set. S3 adds model 3 with model 0's deployment loss and its $q_1, \ldots, q_4$

Table 24: Costs with a cold start and with a warm cache. "Cold" resets the acquisition cache at every seed; "warm" shares it, as the main runs do. Every baseline gets the same cache treatment. A ratio at or below 1.0 (for example S4 Plug-in KL at 0.9) is timer noise for a rule that costs a few milliseconds per episode. The setup times (channels, subset cache) are wall-clock measurements from separate runs and differ from those of Table 7.

Design	Method	Channels (s)	Subset cache (s)	LPs	Cache (KiB)	Cold ms/ep	Warm ms/ep	Ratio	Cold hit %	Warm hit %
S4	Full RDI	0.41	0.026	15	1.8	31.10	3.04	10.2	95.6	99.7
S4	Separation	0.41	0.026	15	1.8	0.69	0.79	0.9	94.9	99.6
S4	Plug-in KL	0.41	0.026	15	1.8	2.02	2.34	0.9	94.9	99.6
S4	Uniform	0.41	0.026	15	1.8	0.55	0.60	0.9	98.3	99.9
S5	Full RDI	0.19	0.026	15	2.8	24.46	1.81	13.5	97.8	99.8
S5	Separation	0.19	0.026	15	2.8	0.52	0.47	1.1	98.2	99.8
S5	Plug-in KL	0.19	0.026	15	2.8	2.19	1.92	1.1	98.1	99.8
S5	Uniform	0.19	0.026	15	2.8	0.47	0.57	0.8	99.2	99.9
S6	Full RDI	0.04	0.009	7	0.9	8.57	3.81	2.2	90.4	99.7
S6	Separation	0.04	0.009	7	0.9	3.57	2.94	1.2	98.5	100.0
S6	Uniform	0.04	0.009	7	0.9	3.13	2.83	1.1	98.5	100.0
D6	Separation	2.49	0.414	63	35.4	0.28	0.23	1.2	94.5	99.1
D6	Plug-in KL	2.49	0.414	63	35.4	1.70	2.01	0.8	94.5	99.2
D6	Uniform	2.49	0.414	63	35.4	0.20	0.23	0.9	95.5	99.5

channels, but with singleton q_0 law $(.8, .1, .1)$. When both duplicate deployment models remain, the forward witness keeps 0 and the reverse witness keeps 3. Both witnesses have the exact parent radius. S3 has 160 runs per method; its witness ablation is discussed in Appendix F.10.

G.2 Inference, selectors, and numerical tolerances

The pairwise and composite constructions are those of Table 5. A fixed component is updated only on its own query; the small designs use events of single symbols and the large designs each model's union of physical support symbols, with the probabilities of the common background accounted for exactly. The mixture weights remain fixed; current evidence is used only by the greedy selector and the acquisition rules. Support elimination intersects both constructions and never invents an unobserved symbol. Proposals for pair factors minimize KL over channel mixtures with SLSQP (tolerance 10^{-10}, at most 120 iterations); each ratio is clipped to $[10^{-8}, 10^{8}]$ and divided by the safe normalizer of Appendix F.7. Its margin ζ_s stays at the floor 10^{-12} on every reported design: it would rise only for alphabets of at least 2252 symbols, and the largest reported alphabet has 1017. A proposal whose certified growth is below -10^{-9} on some alternative vertex is replaced by the constant factor one, with the drift correction described after Theorem D.1. This fallback replaces 2 of the 228 pair factors on S5, 12 of 216 and 5 of 216 in the S4 sweep arms $\kappa = .04$ and $\kappa = .01$, 6 of 216 in the S4 view with only the alarms, 2 of 360 in the corresponding D5 view, and none elsewhere. For more than 80 mixture variables the implementation replaces the optimization by a normalized barycenter proposal, which keeps conditional e-validity and loses the worst-case log-growth optimality of Theorem B.16; no response face in the reported runs has more than 24 mixture variables, so this proposal is never used. No omitted support symbol or violation of null feasibility above 10^{-10} was found before numerical duplicate pruning.

G.3 Full D5 and D6 definitions

D5 uses the graph

$$0 : \{1, 2, 4\}, \quad 1 : \{0, 2\}, \quad 2 : \{0, 1, 3\}, \quad 3 : \{2, 4\}, \quad 4 : \{0, 3\}.$$

The defender starts at node 0, horizon is $H = 3$, and at each step it may stay or move to an adjacent node. The six attacker types are vectors of target weights with evasion parameters

θ	$w_\theta(0{:}4)$	e_θ
0	$(0.9, 0.2, 0.4, 0.7, 0.3)$	0.05
1	$(0.2, 0.85, 0.3, 0.4, 0.7)$	0.10
2	$(0.4, 0.3, 0.9, 0.2, 0.6)$	0.15
3	$(0.65, 0.4, 0.2, 0.9, 0.3)$	0.20
4	$(0.3, 0.75, 0.55, 0.25, 0.9)$	0.25
5	$(0.55, 0.3, 0.75, 0.6, 0.35)$	0.30.

An attacker action is a pair of target and time $(j, t) \in \{0, \dots, 4\} \times \{1, 2, 3\}$. If the defender is not at j at time t, reward is $-w_\theta(j)$; if it is at j, reward is $-w_\theta(j)\{1 - 0.85(1 - e_\theta)\}$. Alarm probability at node i and time s is clipped to $[0.02, 0.95]$ after adding false alarm $0.04 + 0.01\theta$ to

$$0.65(1 - e_\theta)w_\theta(j)\,\rho(i, j)\,\tau(s, t),$$

where $\rho = 1$ at the target, 0.45 at an adjacent node, and 0.12 otherwise; $\tau = 1$ for $s \le t$ and 0.25 afterward. The eight deterministic tours and the twelve calibration policies are listed in Appendix F.6. The D5 transcript records the tour index, attacker action, three alarm bits, and reward label; the alphabet size is 40,320. Smoothing to common support uses $P_\eta = (1 - \eta)P + \eta U$, with $\eta = 0.95$ in D5 and $\eta = 0.999$ at its sweep point with weak feedback.

D6 uses the six regimes of deck counts and the dealing rule of Appendix F.6. Pure strategies of player one are six bits: bet/check by private rank and call/fold after checking and facing a bet by private rank. Pure strategies of player two are six analogous bits, one per information set: call/fold after a first bet, and bet/check after a first check. If player one bets and player two folds, reward is $+1$ and player two's card is not revealed. If the bet is called, showdown uses stake 2. If both check, showdown uses stake 1. If player one checks, player two bets, and player one folds, reward is -1 and player two's card is not revealed; if called, showdown uses stake 2. The recorded transcript is (player-one card, terminal line, reward, revealed player-two card/hidden), giving $3 \cdot 5 \cdot 5 \cdot 4 = 300$ symbols. D6 uses the same smoothing with $\eta = 0.85$.

G.4 Hidden rooms and small verification instances

D1–D4 and the EXP3 baseline below use a separate, simpler simulator that does not use the modules of Tables 4 and 5. Their numbers are never pooled with the others. D1 uses $d \in \{2, 4, 8, 16, 32, 64, 128\}$, $\theta_1 \in \{.4, .45, .5, .55, .6\}$, and 30 seeds per configuration. D2 uses $d \in \{8, 32, 128\}$, $\theta_1 = .5$, noise ratios $\sigma_{\text{dec}}/\sigma_{\text{nuis}} \in \{1, 2, 4, 8\}$, and 30 seeds. Both have $\sigma_{\text{dec}} = .25$, $\varepsilon = \delta = .05$, and cap 200000. The Gaussian proxy is $\frac{1}{2}\log(1 + V_j/\sigma_j^2)$ with $V_j = (12 + n_j/\sigma_j^2)^{-1}$; the truncated proxy replaces V_j by the variance of the Gaussian likelihood under a uniform prior on $[0, 1]$. Neither is an exact posterior predictive entropy. Both use their sequential scores, with random tie-breaking. The truncated variance has a numerical fallback in the tail, and its use is logged for each run.

Gaussian observations can be simulated from the sufficient statistics of each query; Uniform's waiting times for the relevant query are exactly geometric. Certification is checked when that coordinate changes. The learner never uses a future observation or the true parameter to choose a policy. Only the simulator uses true parameters, to generate data and to check the returned gap. The "Fixed q_1" row plays q_1 at every round; it computes no RDI score and does not search for the informative coordinate. It is an oracle benchmark with a fixed query, because the single coordinate relevant to deployment is known here by construction. The Euclidean identification target of Proposition B.28 is a stronger target chosen for illustration and is not an equivalent deployment baseline. The design tests neither strategic acquisition nor graded decision relevance, so no claim in this paper rests on it.

Table 25: Simulation of hidden rooms. Entries are medians as defined in Appendix F (the $\lceil n/2 \rceil$-th smallest observed episode count, hence always an integer), over 150 runs per method and dimension in D1 and 120 in D2 (pooling the listed fixed configurations). No run in this table is censored. The "Fixed q_1" column plays that query at every round; it is an oracle benchmark with a fixed query. D1 omits the truncated proxy by design; the dash does not indicate censoring.

Design	d	Fixed q_1	Gaussian proxy	Trunc. variance	Uniform
D1	2	214	427	–	422
D1	32	219	6,993	–	6,773
D1	128	206	26,288	–	26,932
D2	8	857	6,856	6,743	6,864
D2	32	850	27,200	27,043	27,200
D2	128	847	108,416	107,052	106,722

D3 is a worked example for $A_\theta = \text{diag}(1 + 2\theta, 1)$ with $\theta = 0, 1$. The restricted pool $\{q_{1/2}\}$ has a common least-informative law and radius $1/8$. On this pool the output is an obstruction, with no

policy certificate. The supplied $q_{3/8}$, or the same query found by the Tier II grid search over the diagnostic plan (Appendix B.3), separates the models in one observation of the full log, followed by a deployment solve with zero radius. These outcomes hold for both types and are deterministic. Tie-breaking known to be favorable defines a different response model and is excluded from the comparisons, whose certificates hold under every selector.

D4 pairs two seeding rules by instance, seed and model count on the 100 conflict instances retained from 344 ensembles of random games with five actions. Anchored seeding uses zero pricing calls on 99 instances; naive seeding uses zero on 76. The mean saving is $.23$ calls, paired bootstrap interval $[.15, .32]$. Both rules find a separating query on every accepted instance. The statement is conditional on the $100/344$ conflict filter and counts calls to query generation; it measures neither certification episodes nor wall-clock cost. We report it only as a sanity check on our pricing implementation.

Finally, EXP3 (Auer et al., 2002) over the 64 pure D6 player-one strategies (horizon $T = 4000$, exploration $\min\{1, \sqrt{64 \log(64)/((e-1)T)}\}$, an exact opponent best response to the current mixed strategy at each deployment episode, the importance-weighted update applied to $(r+2)/4$, twenty seeds for each deck type) reaches a mean ex post gap of the averaged policy of $.196$ $[.195, .198]$ after 4000 deployment episodes, from $.305$ after 10. It returns no anytime certificate, and deployment episodes are a different cost axis from diagnostic episodes, so we draw no efficiency comparison from it.

G.5 Sources of the acquisition advantage

Section 6 reports that full RDI certifies 2.77 and 3.24 times faster than uniform querying on S4 and S5. These numbers compare against only one alternative, so they credit the decision-theoretic score with gains that have other causes. On S5 eight of the nineteen queries are structurally blind, and uniform querying spends $8/19$ of its episodes on them. Table 26 splits these factors into their parts; in this study the paired ratio of uniform querying to Full RDI on S5 is 3.34 $[2.93, 3.79]$ over 80 pairs. It adds four comparators. All of them use the same candidate library and no information about the true model, so all are implementable:

Uniform, informative only. λ_0 restricted to the queries that are not structurally blind. A query is blind when every model's every response induces the same transcript law, which is read off the channel table itself. This is the cheapest baseline that already knows the library.

Static RDI. The RDI allocation of the initial library, computed once and played for the whole run without recomputation. This is an experiment design fixed at the start.

Static RDI, full-set prior. The same, with the least-favorable prior on the whole library instead of on the canonical witness. It is identical run for run to Static RDI on S4, S5 and S6, so the table omits it.

Static separation. The maximin symmetrised separation allocation over the initial conflicting pairs, kept fixed.

With the rows for single queries, these separate four sources of advantage: avoiding queries that are provably useless, experiment design fixed at the start, adaptivity driven by the confidence set, and the exact child LP compared with the cheap separation score. The rows for single queries are split into the query chosen by a rule fixed a priori and the query that turned out best. The latter is labeled *ex post* because it is not implementable.

The query chosen a priori is the argmax of the RDI allocation on the initial library. This rule uses only the candidate library, so it is implementable. It selects q_{12} on S4, q_{17} on S5 and q_0 on S6. On S5 and S6 this is the same query that the ex post search returns, with the same median; on S4 it is one episode behind the ex post best. So on these designs the ex post advantage of the best fixed query over the one chosen a priori is at most one episode, and the rows for single queries can be read as implementable baselines.

Every method, including one with a fixed query, still uses the uniform forced-exploration weight $\gamma = .1$. A row for a structurally blind query therefore measures what forced exploration alone achieves. We report it for completeness only.

Table 26: Simple static baselines, 20 seeds per true model, cap 2000, composite module on S4 and S5 and composite plus frequency on S6. Read from the uniform rows upward, each block adds one capability: knowing which queries are blind, then a design fixed at the start, then adaptivity; the rows with a fixed query follow. The *ex post* row is not implementable and is labeled as such.

Design	Acquisition	Fraction certified	Median episodes	Mean (capped)
S4	Full RDI (adaptive)	1.000	34	43
S4	Separation (adaptive)	1.000	33	37
S4	Static RDI (frozen)	1.000	34	43
S4	Static separation (frozen)	1.000	33	37
S4	Uniform, informative only	1.000	88	116
S4	Uniform, all queries	1.000	88	116
S4	Fixed q_{12} (a priori)	1.000	34	43
S4	Fixed q_7 (*ex post* best)	1.000	33	37
S5	Full RDI (adaptive)	1.000	78	107
S5	Separation (adaptive)	1.000	103	122
S5	Static RDI (frozen)	1.000	92	135
S5	Static separation (frozen)	1.000	129	151
S5	Uniform, informative only	1.000	154	180
S5	Uniform, all queries	1.000	285	316
S5	Fixed q_{17} (a priori)	1.000	92	135
S5	Fixed q_{17} (*ex post* best)	1.000	92	135
S5	Fixed, structurally blind (8 queries)	0.341	∞	1749
S6	Full RDI (adaptive)	1.000	9	10
S6	Separation (adaptive)	1.000	57	62
S6	Static RDI (frozen)	1.000	9	10
S6	Static separation (frozen)	1.000	57	62
S6	Uniform, informative only	1.000	50	56
S6	Uniform, all queries	1.000	57	62
S6	Fixed q_0 (a priori)	1.000	9	10
S6	Fixed q_0 (*ex post* best)	1.000	9	10
S6	Fixed, structurally blind (1 query)	1.000	593	635

G.6 Algorithm 1 itself, reported separately

Every main table of this paper runs Algorithm $1'$, the never-empty variant that Proposition E.10 needs for almost sure stopping. Off the coverage event it certifies the maximin policy of a model that the data have already contradicted. It is therefore only a device for the proof. Table 27 reports Algorithm 1 (the raw confidence set, with a STATISTICAL FAILURE exit) next to it, on the same designs, seeds and streams. Following the (S1)–(S3) split before Proposition E.10, we report three quantities separately: the *failure* rate, the *correct-output* rate (certified, with the recomputed true gap of the returned policy within ε), and the median. The never-empty fallback turns a failure into an output that need not be correct, and the two columns show this. Wall-clock times are omitted because the two arms ran with differently filled acquisition caches.

Part 1 of the proposition gives the inclusion $D \subseteq A^c$: the set D of runs whose terminal status differs lies outside the coverage event. (On the coverage event A the true model is always in C_n, so C_n is never empty and the two algorithms coincide.) The reverse inclusion does not hold: the truth can be eliminated while some other model survives, and then both algorithms return the same policy and the fallback never fires. These runs show that the inclusion is strict. Over the $3{,}360$ tabulated runs the truth is lost on 28, the terminal status differs on 1, and 23 runs return a policy whose true gap exceeds ε, almost all with the same status in the two arms. The only cell where the two algorithms differ is S4, composite, Uniform. On one of its 160 runs the confidence set empties at episode 92. Algorithm 1 exits there with no policy, while Algorithm $1'$ substitutes the singleton with the least evidence against it, whose radius is zero, and certifies at the same episode. The correct-output rate goes from 0.969 to 0.975, so here the fallback happened to return an acceptable policy. This was chance on one run, and the proposition guarantees no such outcome.

Table 27: Algorithm 1 against Algorithm $1'$ on the same seeds and streams. "Failure" is the fraction of runs exiting with STATISTICAL FAILURE; "Correct" is the fraction that certified and whose recomputed true gap is within ε; "Differs" counts the runs reaching different terminal states. A failure rate of zero is only an observation: across the paper the never-empty fallback fired on 73 of $94{,}450$ runs, exactly the runs on which Algorithm 1 would have failed.

Design	Module	Method	Algorithm 1 (raw set)			Algorithm $1'$ (never-empty)			
			Failure	Correct	Median	Failure	Correct	Median	Differs
S1	composite	Full RDI	0.000	1.000	18	0.000	1.000	18	0/120
S1	composite	Separation	0.000	0.992	81	0.000	0.992	81	0/120
S1	composite	Uniform	0.000	0.992	81	0.000	0.992	81	0/120
S1	comp.+freq.	Full RDI	0.000	1.000	20	0.000	1.000	20	0/120
S1	comp.+freq.	Separation	0.000	1.000	92	0.000	1.000	92	0/120
S1	comp.+freq.	Uniform	0.000	1.000	92	0.000	1.000	92	0/120
S4	composite	Full RDI	0.000	0.981	36	0.000	0.981	36	0/160
S4	composite	Separation	0.000	0.994	32	0.000	0.994	32	0/160
S4	composite	Uniform	0.006	0.969	103	0.000	0.975	102	1/160
S4	comp.+freq.	Full RDI	0.000	0.994	41	0.000	0.994	41	0/160
S4	comp.+freq.	Separation	0.000	1.000	40	0.000	1.000	40	0/160
S4	comp.+freq.	Uniform	0.000	0.988	124	0.000	0.988	124	0/160
S5	composite	Full RDI	0.000	0.988	87	0.000	0.988	87	0/160
S5	composite	Separation	0.000	0.988	95	0.000	0.988	95	0/160
S5	composite	Uniform	0.000	0.981	281	0.000	0.981	281	0/160
S5	comp.+freq.	Full RDI	0.000	0.994	108	0.000	0.994	108	0/160
S5	comp.+freq.	Separation	0.000	1.000	115	0.000	1.000	115	0/160
S5	comp.+freq.	Uniform	0.000	0.988	336	0.000	0.988	336	0/160
S6	composite	Full RDI	0.000	1.000	9	0.000	1.000	9	0/120
S6	composite	Separation	0.000	1.000	54	0.000	1.000	54	0/120
S6	composite	Uniform	0.000	1.000	54	0.000	1.000	54	0/120
S6	comp.+freq.	Full RDI	0.000	1.000	9	0.000	1.000	9	0/120
S6	comp.+freq.	Separation	0.000	1.000	59	0.000	1.000	59	0/120
S6	comp.+freq.	Uniform	0.000	1.000	59	0.000	1.000	59	0/120

G.7 CERTIFIED TWO-SIDED BOUNDS, FRESH SEEDS, AND COST GROWTH

Certification. Table 6 (Appendix F.7) supports Remarks E.12 and E.13; its first column is also an independent check of the design geometry, bracketing $\Gamma(\Theta)$ to within 10^{-14} on every finite design.

Fresh seeds. Re-running the original seeds reproduces the reported runs exactly and so cannot test how stable the S5 effect is under resampling. Table 28 therefore repeats S5 with two disjoint observation seed roots. The observation stream is a function of (true model, seed, query) alone, so a fresh root re-randomizes every path. These are independent replications of the design. At the primary endpoint the paired ratio of Separation to RDI is 1.31 $[1.21, 1.41]$ and 1.19 $[1.10, 1.29]$ on the two fresh roots, against 1.20 $[1.12, 1.29]$ on the original root; every interval excludes one.

Table 28: S5 on independent observation seed roots at $\varepsilon = .02$, 40 seeds per true model: medians of five acquisition rules and the paired ratios [95% CI] of Uniform and of Separation to full RDI. The row "original" is the original seed root, used by every other S5 table; 5150220 and 7710431 are two fresh roots.

Seed root	Full RDI	Separation	Group-KL	ID-KL	Uniform	Uniform/RDI	Separation/RDI
original	87	95	99	112	281	3.24 [2.92, 3.62]	1.20 [1.12, 1.29]
5150220	82	100	98	118	265	3.33 [2.98, 3.71]	1.31 [1.21, 1.41]
7710431	89	93	98	114	271	3.27 [2.98, 3.59]	1.19 [1.10, 1.29]

Cost growth. Table 24 gives the cost per episode with the acquisition cache cold and warm on S4, S5 and S6. On S5, the design behind the claim on sample efficiency, cold full RDI costs 24.5 ms per episode against 0.52 ms for Separation, a ratio of 47. Its cold start also adds 0.19 s of channel construction and 0.026 s to fill the deployment cache over all subsets. With a warm cache the

same comparison is 1.8 ms against 0.47 ms. Table 29 gives the complementary quantity: the cold, uncached cost of one allocation as the library, the pool and the channel vertex count grow. It is measured on sub-libraries and sub-pools of S5, so everything else about the design is fixed. In this range (one design, $m \in \{2, 3, 4\}$, a single cold allocation), the cost of the exact child LP grows about linearly in the number of queries and is flat in the number of models, which is the purpose of the canonical witness. This does not identify the pool, rather than the library, as the main obstacle to scaling: witness search, the confidence update, channel construction and the deployment cache over $2^M - 1$ subsets all grow with $M = |\Theta|$, and four models are too few to separate these effects, so we leave library scaling open. At the full S5 pool, one cold allocation of full RDI costs two to three orders of magnitude more than one separation allocation. Let c_R and c_S be the compute per episode of the two rules, with comparable episode counts $N_S = rN_R$, $r > 1$. Equating total cost shows that full RDI is cheaper when one interaction costs at least

$$\frac{c_R - r\, c_S}{r - 1} \;=\; \frac{c_R - c_S}{r - 1} - c_S,$$

which is about 119 ms on S5 with the figures for a cold cache in Table 24 and $r \approx 1.20$, and about 6.2 ms with the figures for a warm cache, which the main runs use. Because r is a paired geometric mean of episode ratios, this threshold only approximates the interaction cost at which the two rules break even. On S4, where $r \approx 0.92 < 1$, full RDI is not cheaper for any interaction cost.

Table 29: Cold, uncached acquisition cost per distinct confidence set, on sub-libraries and sub-pools of S5. It differs from the cost per episode of Table 24, which spreads the cache over a run: it is the worst-case cost for a new confidence set, and it is the quantity that grows.

Models	Queries	Max vertices	Distinct laws	Full RDI (ms)	Separation (ms)	Uniform (ms)	RDI/Sep.
2	5	1	10	329.44	2.08	0.18	158.4
2	10	4	26	568.84	1.58	0.14	360.0
2	19	4	64	936.55	1.82	0.13	514.6
3	5	1	15	237.30	1.40	0.03	169.5
3	10	4	38	523.11	0.94	0.05	556.5
3	19	6	97	969.97	1.38	0.24	702.9
4	5	1	20	275.00	0.93	0.02	295.7
4	10	4	50	497.50	1.21	0.02	411.2
4	19	6	134	1125.37	1.48	0.03	760.4

G.8 Scope of the benchmark suite

The designs are not a sample from any population of applications. Four properties of the larger designs, each stated where it arises, limit what can be concluded from them. (i) Full RDI runs only on the libraries S1–S6, which have three or four models; D5 and D6 run PS-EASI. The cost curve of Table 29 is evidence about how the exact child LP fails to scale. (ii) Of the designs that run full RDI, only S4, S5 and S6 are two-sided at deployment, and they give different answers: the exact score has no advantage on S4 (0.92) and a factor 1.20 on S5, and S6 is built to isolate the choice of inference module. (iii) D5 and D6 use background mixing rates $\eta = .95$ and $.85$, so every D5 and D6 number describes the smoothed channel, and D5's full transcript identifies the model almost immediately. (iv) D5's model 3 has $\mathrm{Gap}_3 \equiv 0$ (Appendix F.5), so it belongs to every compatible set and cannot be part of a decision conflict. The theoretical claims do not depend on these properties: they are statements about finite channel polytopes, with hypotheses that can be checked for each instance (listed in Appendix A.3). The empirical claims do depend on them, so we state them per design and point out where the designs disagree.