BETTER AGENTS, WORSE TEAMS: WHEN LOW-ORDER TESTS CERTIFY INDEPENDENT UPGRADES

Anonymous authors
Paper under double-blind review

ABSTRACT

Independently improved agents can form a worse team. We study when tests of single and paired replacements certify every allowed mixed-version deployment. Bounds on policy changes define an exactly realizable activation class. We derive its sharp pairwise remainder and prove that, for m agents with equal exposure p, the anchored predictor is minimax exactly when $(1-p)^{m-2} \geq 1/2$. Verified causal dependencies give smaller remainders and polynomial target bounds. Synthetic experiments compare different statistical information: on a 16-agent workflow, pairwise testing certifies three of three pools at 64 million episodes with a residual-coupling known-moment bound, and at 128 million with empirical Bernstein and common random numbers. Independently fitted 12–16-agent teams include compatible and incompatible upgrades; pilot-allocated exhaustive testing narrows some cost gaps. Multi-step local tasks reveal failures missed by uncorrected pairwise fits. Episode gates certify only the gate-averaged canary mixture; small gains require low gate rates and large evaluation budgets. Exact feasible intervals and attained-remainder sampling separate uncertainty caused by missing information from conservative statistical bounds. The results identify when low-order tests are informative, optimal, and cheaper than direct evaluation.

1 INTRODUCTION

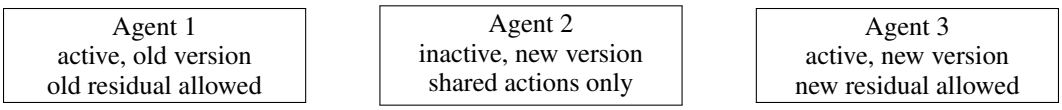
A team can improve each agent separately and still deploy a worse combination. This problem arises when components have separate release schedules, tenants enable different subsets, or operators roll back individual services. Testing one deployment shows whether that combination works. It does not certify a family of combinations that can arise independently.

The distinction already appears with small behavioral changes. There are three-agent cooperative processes with exposure $p = .1$ whose baseline, single-replacement, and pair-replacement returns are all 50.00%, yet whose fully replaced returns are respectively 50.40% and 49.60%. The two processes are indistinguishable from the low-order values. Additional observations at those same configurations cannot identify which process is deployed. The useful question is therefore: *what extra information controls the untested interactions, and when is collecting it cheaper than testing every target?*

We take fixed old and new policies as inputs. Our core result connects two questions: which value tables are possible under bounded policy changes, and how much of an untested value can be recovered from low-order means? The activation representation answers the first exactly. For the second, a matching construction gives a sharp threshold: with equal exposures, the anchored pairwise predictor is minimax if and only if the chance that all agents outside a specified pair stay inactive is at least one half. This gives an information limit for small-change extrapolation.

Finite data introduce another cost. A pairwise evaluation has fewer configurations, but its extrapolated contrast has a larger range than a direct difference. We compare empirical-variance confidence bounds under ordinary common random numbers (CRN), a stronger moment bound under an explicit coupling, and an exhaustive method that allocates holdout samples using a pilot. Independently fitted large teams supplement the controlled workflow: the crossover persists in an underused-resource regime, while crowded and larger-update regimes expose abstention and incompatibility.

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.



$$S = \{2, 3\}, \quad C = \{1, 3\} \quad \Longrightarrow \quad S \cap C = \{3\} \quad \Longrightarrow \quad u_{\{1,3\},\{3\}}$$

Figure 1: One coupled episode. Changing an inactive agent's version does not change the trajectory on this draw. The activation set is a proof device, not a required observation. Pairwise reconstruction is exact on draws with at most two active target agents.

Two forms of additional information complement the global analysis. Complete causal dependencies can prevent an agent's influence from reaching most rewards, even over many steps. Independent episode-level gates instead change the deployment itself, reducing exposure without inspecting an opaque policy. We quantify both their benefits and costs. The resulting decision is concrete: compare the available information, the worst-mixture margin, and the full evaluation cost before choosing extrapolation over direct tests.

2 Problem and the activation model

There are n agents with fixed old/new policies π_i^0, π_i^1, a common initial-state law, at most H decision rounds, and return $G \in [0, 1]$. Policies can use complete observation histories. The environment reads actions rather than version labels, and policy randomization factors across agents conditional on their histories. For $S \subseteq [n]$, write

$$F(S) = \mathbb{E}_{\pi_S^1, \pi_{-S}^0} G, \qquad D(S) = F(S) - F(\varnothing). \tag{1}$$

An allowed deployment family $\mathcal{R}$ is compatible at tolerance ε if $\min_{S \in \mathcal{R}} D(S) \geq -\varepsilon$. We use the full family in experiments. A known rolling order has only $n + 1$ targets and should normally be tested directly. Independent enable decisions, rollbacks, and availability constraints motivate much larger families.

One source of valid exposure probabilities is a uniform bound on total variation (TV) distance, denoted D_{TV}:

$$\sup_{t, h_i} D_{\mathrm{TV}}(\pi_{i,t}^0(\cdot \mid h_i), \pi_{i,t}^1(\cdot \mid h_i)) \leq d_i, \qquad p_i = 1 - (1 - d_i)^H. \tag{2}$$

Tabular policies permit exact computation; per-decision mixtures $(1 - \lambda_i)\pi_i^0 + \lambda_i \widetilde{\pi}_i$ supply $d_i \leq \lambda_i$. An order-k query family is $\mathcal{Q}_k = \{Q : |Q| \leq k\}$; its anchored predictor is denoted P_k, never $\mathcal{Q}_k$.

Let C include each agent independently with probability p_i, and set

$$w_C = \prod_{i \in C} p_i \prod_{i \notin C} (1 - p_i). \tag{3}$$

Theorem 1 (Realizable activation class)**.** *The set of value tables retaining only the exposure summaries is exactly*

$$\mathcal{A}(p) = \left\{ F : F(S) = \sum_C w_C u_{C, S \cap C}, \quad 0 \leq u_{C,T} \leq 1 \right\}. \tag{4}$$

Every process satisfying (2) *belongs to this class, and every table in the class is realizable by a finite cooperative process satisfying those bounds.*

At each history the two action kernels share a common part of mass $1 - d_i$. Independent residual flags select the remaining mass. Couple all version configurations with the same flags and environment randomness. An agent's version can matter only if one of its flags occurs. Conditioning on the set of such agents yields (4). A recording process with neutral, old-specific, and new-specific symbols realizes every conditional reward table. Appendix A gives the full proof. Here $\mathcal{A}(p)$ names the primitive activation class; (d, H) is one way to supply p.

3 WHAT PAIRWISE DATA DETERMINE

Define individual gains and pair interactions,

$$g_i = F(\{i\}) - F(\varnothing), \qquad h_{ij} = F(\{i,j\}) - F(\{i\}) - F(\{j\}) + F(\varnothing),$$
$$P_2(S) = F(\varnothing) + \sum_{i \in S} g_i + \sum_{\{i,j\} \subseteq S} h_{ij}. \tag{5}$$

A negative h_{ij} measures interference beyond the separately observed gains. Pairwise fitting uses $1 + n + \binom{n}{2}$ configurations; certification also needs a remainder bound.

Theorem 2 (Sharp anchored error)**.** *Let $M_S = |C \cap S|$. Over $\mathcal{A}(p)$, the exact worst-case errors of the anchored predictors are*

$$\sup_F |F(S) - P_1(S)| = B_1(p_S) = \mathbb{E}[M_S \mathbf{1}\{M_S \geq 2\}], \tag{6}$$

$$\sup_F |F(S) - P_2(S)| = B_2(p_S) = \mathbb{E}[(M_S - 1)^2 \mathbf{1}\{M_S \geq 3\}] \leq 4 \sum_{\{i,j,k\} \subseteq S} p_i p_j p_k. \tag{7}$$

Both signs are attainable for each specified target.

Conditioning on m active coordinates makes the omitted term a zero-sum linear functional on a unit box. Its positive coefficient mass is $(m-1)^2$ for $m \geq 3$ and zero otherwise. Choosing rewards by coefficient sign attains the bound, and Theorem 1 realizes those choices. The resulting radius is a sharp error for this estimator. Minimax optimality additionally requires two feasible processes that agree on the queried means and attain opposite errors.

When the simple predictor is optimal. Let $v_{C,T} = w_C u_{C,T}$, let $a_S v = \sum_C v_{C,S \cap C}$, and write $W_{C,T} = w_C$. Classical optimal recovery over symmetric convex classes (Micchelli & Rivlin, 1977; Foucart & Paouris, 2023) gives the following specialization.

Theorem 3 (Recovery in the activation class)**.** *With queried rows $A_{\mathcal{Q}}$ including the baseline, the worst-data minimax absolute-error radius is*

$$R_*(S) = \max_{|z_j| \leq W_j/2,\ A_{\mathcal{Q}} z = 0} a_S z = \min_{\lambda} \frac{1}{2} \sum_j W_j \left| a_S(j) - \sum_{Q \in \mathcal{Q}} \lambda_Q a_Q(j) \right|. \tag{8}$$

The worst queried data are all $1/2$. An optimal dual multiplier gives an optimal affine estimator; at particular data the midpoint of the feasible target interval may be more accurate.

The half-difference of two indistinguishable worlds lies in the centered box and query kernel. Conversely, either sign of any such vector is feasible at the central data. Box support and linear programming (LP) duality then give (8). The activation-specific attaining construction strengthens this general recovery principle:

Theorem 4 (Exact homogeneous threshold)**.** *For $m \geq 3$, common exposure $0 < p < 1$, and baseline, singleton, and pair means,*

$$R_* = B_2(p, \ldots, p) \quad \Longleftrightarrow \quad (1-p)^{m-2} \geq \tfrac{1}{2}. \tag{9}$$

The key obstruction is a cancellation budget. To attain the anchored remainder, high-activation strata force a negative second difference for every queried pair. Only the stratum where that pair alone is active can cancel it. Cancellation is possible exactly when this low stratum has at least half the pair's activation mass. Appendix O supplies a bounded, balanced construction proving sufficiency and a support argument proving necessity.

At $m = 5, p = .2$, the condition holds because $(4/5)^3 = 64/125 > 1/2$. At $m = 16, p = .02$, the anchored radius 1.60×10^{-2} is minimax; its threshold is $p \leq 1 - 2^{-1/14} \approx 4.83 \times 10^{-2}$. This explains why improving the recovery rule cannot remove the small-exposure uncertainty simply by reweighting the same queried means. Outside this region, compare the optimized radius with the *range-aware* benchmark $\min\{B_2, 1/2\}$: predicting $1/2$ requires no data. For $m = 4, p = .4$, both this benchmark and R_* are $1/2$; for $m = 6, p = .2$, they are approximately 0.49 and 0.42.

Appendix O reports both comparisons and rational checks. A heterogeneous sufficient condition remains in Corollary 12.

The same geometry gives a sharp *data-dependent* gain evaluation:

$$L_{\mathrm{LP}}(S) = \min_{v}(a_S - a_\varnothing)v \quad \text{s.t. } 0 \le v_j \le W_j, \quad \ell_Q \le a_Q v \le u_Q. \tag{10}$$

It is exact for nonempty feasible evidence in the summary class (Appendix E). It has 3^n variables; we use it to study information limits, not as the large-team implementation. If rewards G_t are paid earlier with weights β_t, their prefix exposures give $B_2^{\mathrm{time}}(S) = \sum_t \beta_t B_2(p(t)_S)$. The exact timed class is the Minkowski sum $\sum_t \beta_t \mathcal{A}(p(t))$, so timing also enters the LP (Appendices G and M).

4 Information that makes the bounds useful

4.1 Reachability and occupancy: what can be relaxed

The supremum in (2) can be restricted to a *verified* union of histories reachable under the queried and allowed configurations. Extending both kernels identically outside that union proves the same upper bound. Observed baseline histories alone do not establish this support condition.

Occupancy averages support a different guarantee. If μ_t^T is the reference configuration's history law and δ_t bounds the joint-action TV distance, first-disagreement coupling gives

$$|F(S) - F(T)| \le \min\{1, \sum_t \mathbb{E}_{\mu_t^T}\delta_t\}. \tag{11}$$

They do not generally give a cubic interaction remainder. A public rare context of probability q can make three deterministic policies differ together. Every agent's expected TV is q under *every* mixed configuration, yet an unobserved triple failure has probability q, not q^3. Appendix J proves both statements. The obstruction is shared activation, not merely loss of exactness of the summary class.

4.2 Local causal structure

Suppose the normalized return is a weighted sum of bounded components, $G = \sum_a \beta_a G_a$, with $\beta_a \ge 0$ and $\sum_a \beta_a \le 1$. A supplied causal cone V_a contains every agent whose version can affect component G_a, including effects through future states and observations.

Theorem 5 (Local interaction certificate)**.** *Under this cone condition,*

$$|F(S) - P_2(S)| \le B_{\mathrm{loc}}(S) := \sum_a \beta_a B_2(p_{S\cap V_a}). \tag{12}$$

If $|V_a| \le q$ and $p_i \le p_{\max}$, then $B_{\mathrm{loc}}(S) \le 4\binom{q}{3}p_{\max}^3$. Interactions h_{ij} vanish unless some cone contains both agents.

Apply Theorem 2 to each component and add its bound. The number of agents in a verified cone, rather than the entire team, now controls the remainder. Size-two cones make the predictor exact. Immediate locality or bounded graph degree alone is insufficient; repeated transitions can enlarge a reward's ancestors.

This structure also makes target search tractable at some cost in tightness. Write the coarse local penalty as $\sum_{i<j<k} c_{ijk}x_ix_jx_k$, $x_i = \mathbf{1}\{i \in S\}$. Since $x_ix_jx_k \le (x_i + x_j + x_k)/3$, it is at most $\sum_i b_i x_i$ for known $b_i \ge 0$. Evaluate targets of size at most two from direct values or verified zero interactions, with zero remainder. For larger targets minimize the quadratic predictor minus this unary penalty over the McCormick relaxation $0 \le x_i, z_{ij} \le 1$, $z_{ij} \le x_i, x_j$, $z_{ij} \ge x_i + x_j - 1$, adding $\sum_i x_i \ge 3$. Taking the smaller of this bound and the directly evaluated small-target bounds is valid and at least as tight as the unsplit relaxation (Proposition 15). Only $n + |E|$ variables are needed, where E contains the cone-pairs; only $1 + n + |E|$ configuration values are needed. Appendix K gives the statistical extension and checked dual calculation.

4.3 KNOWN EXPOSURES WITHOUT INSPECTING CANDIDATE POLICIES

A different option is to change the deployment contract. For each enabled agent, an independent gate with probability λ_i chooses its opaque candidate once at episode start; otherwise it keeps the old policy. Let $f(T)$ denote the *ungated* candidate team's return. Then

$$F_\lambda(S) = \sum_{T \subseteq S} w_S(T) f(T), \qquad w_S(T) = \prod_{i \in T} \lambda_i \prod_{i \in S \setminus T} (1 - \lambda_i).$$

The gate is fixed for the whole episode, so exposure does not compound with H.

Theorem 6 (Episode-gated evaluation)**.** *Given simultaneous intervals for the ungated values $f(Q)$, $Q \in \mathcal{Q}_2$, an exact lower endpoint for the gated gain is*

$$L_\lambda(S) = \sum_{1 \le |T| \le 2, T \subseteq S} w_S(T) \ell_T - (1 - w_S(\varnothing)) u_0. \tag{13}$$

At exact data the feasible interval has width $\Pr(|S \cap B| > 2)$, and half this width is minimax. For equal gates, a polynomial cardinality bound certifies all targets using the minimum queried gain at each cardinality (Appendix L).

The required capability is an outer request-pinned router and a harness that can force ungated singleton/pair candidate runs. No candidate probabilities or controllable simulator randomness are needed. The proof is binomial reweighting, assigning return zero to unmeasured subsets. The contract certifies expected return over the gates for each enable mask, rather than each realized candidate subset or unrestricted full rollout.

For a worked example, take 64 enabled components, baseline $f(\varnothing) = .5$, all queried nonempty means $21/40$, and gate rate 5.00×10^{-3}. Let P_j be the probability of exactly j active gates. The exact lower gain is

$$\frac{1}{40}(P_1 + P_2) - .5 \Pr(M > 2) \approx 4.68 \times 10^{-3}.$$

Only .32 candidates run on average per episode. At gate rate .01 the same lower bound is -2.06×10^{-3}; more samples cannot make it positive. This illustrates both how (13) is used and why a gated guarantee should not be described as approval of the ungated upgrade.

5 FINITE-SAMPLE EVALUATION AND COST

Fixed candidates are evaluated on fresh episodes. For a queried mean, simultaneous bounded-data intervals can be propagated through (10), (13), or signed predictor coefficients. Baseline reuse is accounted for before propagating errors (Appendix F). A simulator can also share randomness across query configurations. One batch contains one episode per query; batches are independent, while their columns may be correlated.

Let $Y_{S,r}$ be the anchored gain contrast in batch r. Its expectation is $P_k(S) - F(\varnothing)$. Its range length is $2A_k(|S|)$; for pairs, $A_2(1) = A_2(2) = 1$ and $A_2(m) = m(m-2)$ for $m \ge 3$.

Theorem 7 (Coupled fixed-budget certificate)**.** *For $N \ge 2$ independent and identically distributed (i.i.d.) batches with correct bounded marginals, unbiased variance $\widehat{V}_S$, $K = |\mathcal{R} \setminus \{\varnothing\}|$, and $\ell = \log(2K/\alpha)$, simultaneously with probability at least $1 - \alpha$,*

$$D(S) \ge \overline{Y}_S - \mathcal{B}(S) - \sqrt{2\widehat{V}_S \ell / N} - \frac{14 A_k(|S|)\ell}{3(N-1)}. \tag{14}$$

Here $\mathcal{B}$ is a justified anchored, reward-timed, or local remainder.

This is empirical Bernstein (EB) concentration on independent rows (Maurer & Pontil, 2009). Any correct-marginal coupling can use it. The specific residual-flag coupling additionally gives $|Y_S| \le A_k(M_S)$ and a known-moment alternative with $V_S = \mathbb{E}A_k(M_S)^2$ (Appendix H). We do not apply that moment bound to arbitrary common random numbers. The baseline itself has known gain zero.

The certificate approves a family only if every lower bound exceeds $-\varepsilon$; otherwise it abstains. Its costs include queried configurations, total configuration-episodes, and target search. A complete

Table 1: Workflow approvals out of three pools. Budgets are millions of total configuration-episodes. KM denotes known-moment bounds. Pair/exhaustive EB use equal allocation; pilot direct EB includes its pilot cost. A dash denotes an unrun cell. The .05 row is a compatible regime where pairwise loses, not an incompatible control.

n	p	Budget	Pair KM	Pair EB	Direct EB	Pilot EB
8	.02	16	3	3	3	3
12	.02	16	3	0	0	0
12	.02	64	3	3	0	3
16	.02	64	3	0	0	0
16	.02	128	3	3	0	0
16	.02	256	3	3	0	0
10	.05	64	0	0	3	3

N-batch design costs $N|\mathcal{Q}|$ episodes even when randomness is shared. The statistical range grows with target size and can offset configuration savings. We implement a stronger direct comparator: a charged pilot estimates difficulty by target size, chooses holdout counts, and is discarded from the final intervals. Conditional on the pilot, the holdout counts are fixed, so empirical Bernstein and a union bound remain valid (Appendix P). Its allocation is a heuristic, not an oracle or an optimal stopping rule. Exact-margin cost envelopes are reported separately in Appendix N. Each reported method/budget uses $\alpha = .05$ separately. Selecting a favorable method, budget, or stopping time afterward requires additional error control (Howard et al., 2021).

6 Experiments

Candidate training and evaluation use disjoint randomness. Each comparison matches total configuration-episodes, charging pilot and baseline runs; configurations within a coupled row share randomness. Fixed-budget statements use $\alpha = .05$ separately.

6.1 A crossover under two different statistical access models

The three-state workflow has horizon three. An upgraded agent intervenes with probability d_i each round, while an old agent does not. Given intervening set A, the workflow advances with probability $1 - \prod_{i \in A}(1 - e_i)$, capped at state two; terminal Bernoulli means are $(.10, .46, .83)$. Exposures and efficacies vary across agents. We compare known-moment and empirical-Bernstein bounds on 12 pools spanning four size/exposure settings, with three seeds each. Pilot allocation uses independently sampled observations. Candidates here are fixed, and all pools are compatible.

Because the old workflow policy is deterministic, inverse cumulative distribution function (CDF) sampling with common random numbers coincides pathwise with residual flags. EB needs only correct marginals and a known range; the known-moment (KM) interval additionally uses the flag-derived moment bound. Both are computed from the same rows. This separates the effect of statistical information from a change in the observed samples. On the learned task below, old policies are stochastic and no KM interval is used.

At $n = 16, p = .02$, pairwise EB certifies at 128 million episodes but not 64 million. The smaller-budget crossover therefore uses KM information. Equal direct EB still certifies none at 256 million, and the pilot method also certifies none there. At $n = 12$, however, pilot allocation closes the EB gap at 64 million. The crossover depends on allocation as well as exposure and team size. Increasing exposure can remove the pairwise advantage: at $n = 10, p = .05$, direct and pilot EB certify all three at 64 million, while pairwise certifies none.

6.2 Independently learned 12–16-agent replacements

The shared-resource task has three Markov states, horizon three, and two actions per agent. Clients choose whether to request a resource; the state-dependent capacity is one or two, and overload lowers immediate reward and makes a degraded state more likely. Each agent independently fits a three-state, two-action Markov decision process (MDP) from 8,192 sampled one-step outcomes

Table 2: Learned-team EB approvals out of six seeds. All methods use generic common random numbers. The large-update rows contain incompatible teams; the others are compatible. Pilot results at unrun budgets are omitted.

Regime	n	Budget (millions)	Pair	Direct	Pilot	False approvals
Underused	12	16	6	0	–	0
Underused	12	64	6	0	6	0
Underused	16	64	6	0	–	0
Underused	16	128	6	0	0	0
Near capacity	12	64	0	0	0	0
Near capacity	16	128	0	0	0	0
Larger update	12	64	0	0	0	0
Larger update	16	128	0	0	0	0

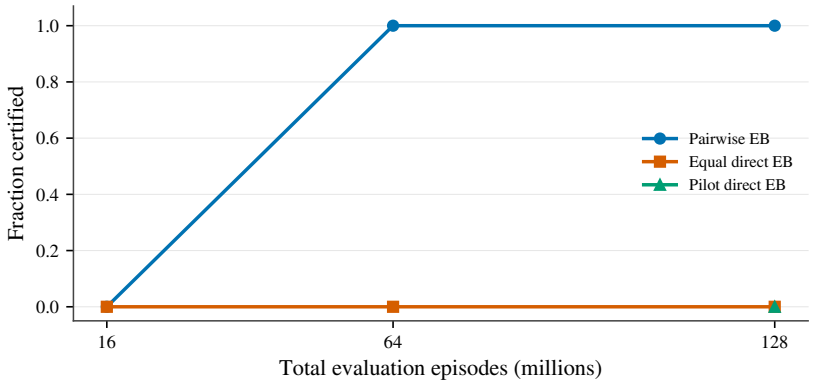


Figure 2: Measured learned-team certification under empirical Bernstein, restricted to the underused 16-agent regime. Each point aggregates six independently fitted pools. Pairwise testing benefits from more observations per queried configuration; the two exhaustive methods remain limited by weak direct target margins. All costs include baseline episodes.

per state/action, keeping all teammates at their old policies. It solves the fitted MDP and mixes its candidate with the old policy. Neither exact team values nor evaluation samples enter training.

We use six seeds for each of two team sizes and three regimes, without filtering candidates by evaluation success: an underused system with mixture 6.00×10^{-3}, a near-capacity system with the same mixture, and a near-capacity system with mixture .06. This gives 36 pools: 24 compatible and 12 incompatible. Every candidate improves individually in the executed pools, including those in incompatible teams. Exact all-mask values are computed by a product transform and dynamic programming in $O(Hn2^n)$ operations per fixed state count. Sampled evaluators receive returns and policy-change bounds, not this dynamics table.

The underused regime shows that the EB crossover also occurs with learned candidates. At 12 agents, pairwise EB certifies all six pools at 16 million episodes; direct EB certifies none even at 64 million, but pilot EB catches up there. At 16 agents, pairwise certifies at 64 million; both direct designs abstain at 128 million. Near capacity, the independently learned improvements are too small for any sampled method to certify. Larger changes create degrading mixtures and all methods abstain. The successful regime still has small changes and spare capacity. These results extend the finding beyond the workflow, but do not establish that it holds across many tasks.

6.3 Multi-step causal cones and hidden incompatible triples

A local reservation system has disjoint queues with three clients each. At every step an unreserved client may submit a request; a successful reservation persists. Terminal reward increases with reservations, with a penalty if all three clients reserve the same queue. There are $n = 12, 24, 48$ clients and horizons $H = 3, 6, 12$. Propagating the supplied factored transition graph verifies that each

queue's reward depends on its three clients throughout the episode. Cones stay bounded because queues have isolated state, not because the horizon is one.

We evaluate exact mean queries on 108 pools from six seeds and two regimes. In every incompatible pool, one queue has a hidden triple penalty large enough to outweigh its individually positive gains. The uncorrected pairwise predictor approves all 108 pools, falsely approving all 54 incompatible ones. The local certificate and its cardinality-split polynomial relaxation approve all 54 compatible pools and no incompatible pool. The global certificate approves none. Normalizing each reward component by its verified range makes the local bound informative; full-team exposure alone discards this information. The sparse design uses $1 + 2n$ configurations rather than every pair. This is a static program-level dependency specification, not a learned causal-discovery method. Cross-queue state or observations would enlarge the cones and invalidate the stated sparse design unless included.

Separate signed-quadratic tests quantify target-search cost using six instances per size, with unary gains in $[-5.00 \times 10^{-3}, 3.50 \times 10^{-2}]/n$, pair coefficients in $[-.04, .04]/n$ on first- and second-neighbor ring edges, and local penalty exposure $p = .05$. At $n = 16$, enumeration is faster than the relaxation and the maximum relaxation gap is 1.61×10^{-4}. At $n = 256$, the measured median solve-plus-rational-dual time is 0.15 seconds with 768 variables. These measurements describe this optimizer benchmark, not end-to-end learned-policy performance or a general runtime guarantee. The cardinality cut removes unnecessary remainder charges from directly tested small targets at one additional sparse constraint.

6.4 What the statistical and operational limits cost

Boundary ambiguity versus statistical slack. The balanced small-margin split has 54 compatible and 54 incompatible pools. Its exact pairwise bound is negative at the hidden triple even on the compatible half: this measures a *dead zone*, not tight calibration. At 32,768 rows per pair query, the local remainder plus statistical radius is a median 28.2 times the absolute triple margin (minimum 5.27). The exact data-dependent activation LP also abstains on all 72 three- and four-agent cases; its mean hidden-triple feasible width is 3.75×10^{-2}. This is an ambiguity result for the retained global activation information, not for every richer locality contract.

To isolate the statistical layer, we sample worlds attaining the negative anchored remainder. There $D(S) - L(S)$ equals confidence radius minus centered sampling error. With one target ($K = 1$), 10,000 repetitions, and 1,048,576 rows each, EB coverage-failure rates range from 3.00×10^{-4} to 1.80×10^{-3}, well below nominal $.05$; no exponential union bound is involved. These are exact multinomial sufficient-statistic draws, not explicitly advanced trajectories. Thus structural ambiguity and conservative concentration both contribute to abstention.

Realistic gates can be expensive. For 64 components with baseline $.5$ and low-order gains of 1–5% of baseline, gate rates $.01$ or larger give a negative exact full-mask lower bound on the tested grid. At a 5% gain and gate 5.00×10^{-3}, all six binomial evaluation repetitions certify at 524,288 samples per query: about 1.09 billion nominal Bernoulli observations over 2,081 configurations. At a 1% gain and gate 2.50×10^{-3}, all six certify only at the largest tested count, about 17.46 billion observations. These costs limit the operational benefit: low gates protect a canary mixture and do not resolve certification of full rollout.

7 Connections to prior work

Better components need not preserve system behavior: human–artificial intelligence (AI) update compatibility (Bansal et al., 2019), backward compatibility in prediction systems (Srivastava et al., 2020), and compatible visual representations (Shen et al., 2020) study related problems with different compatibility criteria. Ad hoc teamwork (Stone et al., 2010), joint-policy correlation (Lanctot et al., 2017), Other-Play (Hu et al., 2020), trajectory diversity (Lupu et al., 2021), and human-aware co-play (Carroll et al., 2019; Strouse et al., 2021) emphasize learning behavior that transfers to other partners. Here the candidates are fixed and the output is a post-training certificate.

Cooperative actor-critic and value-factorization methods (Lowe et al., 2017; Foerster et al., 2018; Sunehag et al., 2018; Rashid et al., 2018; Son et al., 2019; Yu et al., 2022) address training and credit assignment. Sequential trust-region updates (Kuba et al., 2022) and the conditional plug-and-

play guarantees of Sequential Agent Tuning (SAT) (Xie et al., 2026) control how updates are produced; independent successor-feature composition has its own failure modes (Zhao & Li, 2026). Our family-wide evaluation requires neither their training objectives nor their policy libraries. Safe policy improvement and off-policy evaluation provide complementary statistical guarantees (Thomas et al., 2015; Jiang & Li, 2016; Laroche et al., 2019; Simão & Spaan, 2019; Wienhöft et al., 2023; Bianchi et al., 2024). Performance-difference and TV bounds motivate our occupancy transport result (Kakade & Langford, 2002; Schulman et al., 2015; Achiam et al., 2017), while robust transition-set methods retain different uncertainty information (Iyengar, 2005; Nilim & El Ghaoui, 2005).

The mathematical ingredients are established: coupling (Thorisson, 2000), anchored/Möbius expansions (Grabisch et al., 2000; Kuo et al., 2010), higher-order policy expansions (Tang et al., 2020), random restrictions (O'Donnell, 2014), and concentration (Hoeffding, 1963; Maurer & Pontil, 2009). Common random numbers have conditional variance-reduction benefits (Glasserman & Yao, 1992). Optimal linear recovery from noiseless linear information over symmetric convex classes is classical (Micchelli & Rivlin, 1977; Foucart & Paouris, 2023); the box dual is an application of that principle. Our additions are the realizable policy-version class, its exact homogeneous anchored-optimality threshold and attaining constructions, and the resulting exposure-aware evaluations. Fractional factorial designs separate main effects and interactions (Box et al., 2005); t-way testing and covering arrays seek configuration-pattern coverage (Kuhn et al., 2004). Coverage alone does not identify expected returns under untested backgrounds. Appendix I gives the $1+n+\binom{n}{2}$ rank requirement for unrestricted quadratic identification. Our polynomial search uses existing convex and roof-duality methods (McCormick, 1976; Hammer et al., 1984; Boros et al., 2008).

8 Limitations and conclusion

The global certificate requires verified exposure bounds; arbitrary neural fine-tuning need not provide small ones. Horizon saturation remains a real limit. Reachability, causal cones, controllable coupling, and gate-bypass tests require additional capabilities. Gates change the deployed policy and average over their own random choices. Synthetic results with three workflow or six learned seeds per setting do not establish effectiveness for real language-model or robotic teams. Exact summary LPs remain exponential; polynomial relaxations trade tightness for tractability. Expected-return compatibility does not imply trajectory safety, robustness to deployment shift, or validity under adaptive evaluation reuse.

Choose the evaluation method based on the allowed deployments and available information. Fixed rolling orders usually favor direct tests. Pilot estimates may guide allocation, but final confidence calculations need fresh holdouts. If an evaluation abstains, more samples address statistical noise; richer queries or verified structure address intrinsic ambiguity. Reducing a gate rate instead changes the deployed policy. These remedies solve different problems.

AI USE STATEMENT

In this work, we used generative AI tools to draft sections of the paper and polish writing. Additionally, we used generative AI tools for editing code. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31, 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2429–2437, 2019. doi: 10.1609/aaai.v33i01.33012429. URL https://ojs.aaai.org/index.php/AAAI/article/view/4087.

Federico Bianchi, Edoardo Zorzi, Alberto Castellini, Thiago D. Simão, Matthijs T. J. Spaan, and Alessandro Farinelli. Scalable safe policy improvement for factored multi-agent MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3952–3973. PMLR, 2024. URL https://proceedings.mlr.press/v235/bianchi24b.html.

Endre Boros, Peter L. Hammer, Richard Sun, and Gabriel Tavares. A max-flow approach to improved lower bounds for quadratic unconstrained binary optimization (QUBO). *Discrete Optimization*, 5(2):501–529, 2008. doi: 10.1016/j.disopt.2007.02.001. URL https://www.sciencedirect.com/science/article/pii/S1572528607000400.

George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley, 2 edition, 2005. ISBN 9780471718130. URL https://bcs.wiley.com/he-bcs/Books?action=index&bcsId=9686&itemId=0471718130.

Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca D. Dragan. On the utility of learning about humans for human-AI coordination. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11794. URL https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Simon Foucart and Grigoris Paouris. Near-optimal estimation of linear functionals with log-concave observation errors. *Information and Inference: A Journal of the IMA*, 12(4):2546–2561, 2023. doi: 10.1093/imaiai/iaad038. URL https://arxiv.org/abs/2301.07228.

Paul Glasserman and David D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992. doi: 10.1287/mnsc.38.6.884. URL https://business.columbia.edu/faculty/research/some-guidelines-and-guarantees-common-random-numbers.

Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000. doi: 10.1287/moor.25.2.157.12225. URL https://orbilu.uni.lu/handle/10993/1945.

Peter L. Hammer, Pierre Hansen, and Bruno Simeone. Roof duality, complementation and persistency in quadratic 0–1 optimization. *Mathematical Programming*, 28:121–155, 1984. doi: 10.1007/BF02612354.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

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. URL https://arxiv.org/abs/1810.08240.

Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. "Other-Play" for zero-shot coordination. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4399–4410. PMLR, 2020. URL https://proceedings.mlr.press/v119/hu20a.html.

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

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274. Morgan Kaufmann, 2002. URL https://people.eecs.berkeley.edu/~pabbeel/cs287-fa09/readings/KakadeLangford-icml2002.pdf.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EcGGFkNTxdJ.

D. Richard Kuhn, Dolores R. Wallace, and Albert M. Gallo. Software fault interactions and implications for software testing. *IEEE Transactions on Software Engineering*, 30(6):418–421, 2004. doi: 10.1109/TSE.2004.24. URL https://csrc.nist.gov/pubs/journal/2004/06/software-fault-interactions-and-implications-for-s/final.

F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Mathematics of Computation*, 79(270):953–966, 2010. doi: 10.1090/S0025-5718-09-02319-9.

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, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3323fe11e9595c09af38fe67567a9394-Abstract.html.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3652–3661, 2019. URL https://proceedings.mlr.press/v97/laroche19a.html.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/68a9750337a418a86fe06c1991a1d64c-Abstract.html.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7204–7213, 2021. URL https://proceedings.mlr.press/v139/lupu21a.html.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Garth P. McCormick. Computability of global solutions to factorable nonconvex programs: Part I—convex underestimating problems. *Mathematical Programming*, 10:147–175, 1976. doi: 10.1007/BF01580665.

Charles A. Micchelli and Theodore J. Rivlin. A survey of optimal recovery. In *Optimal Estimation in Approximation Theory*, pp. 1–54. Springer, 1977. doi: 10.1007/978-1-4684-2388-4_1.

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. URL https://people.eecs.berkeley.edu/~elghaoui/pubs_rob_mdp.html.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782. URL https://arxiv.org/abs/2105.10386.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4295–4304, 2018. URL https://proceedings.mlr.press/v80/rashid18a.html.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

Yantao Shen, Yuanjun Xiong, Wei Xia, and Stefano Soatto. Towards backward-compatible representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Shen_Towards_Backward-Compatible_Representation_Learning_CVPR_2020_paper.html.

Thiago D. Simão and Matthijs T. J. Spaan. Safe policy improvement with baseline bootstrapping in factored environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4967–4974, 2019. doi: 10.1609/aaai.v33i01.33014967. URL https://ojs.aaai.org/index.php/AAAI/article/view/4427.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5887–5896, 2019. URL https://proceedings.mlr.press/v97/son19a.html.

Megha Srivastava, Besmira Nushi, Ece Kamar, Shital Shah, and Eric Horvitz. An empirical analysis of backward compatibility in machine learning systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3272–3280, 2020. doi: 10.1145/3394486.3403379. URL https://arxiv.org/abs/2008.04572.

Peter Stone, Gal A. Kaminka, Sarit Kraus, and Jeffrey S. Rosenschein. Ad hoc autonomous agent teams: Collaboration without pre-coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1504–1509, 2010. doi: 10.1609/aaai.v24i1.7529. URL https://ojs.aaai.org/index.php/AAAI/article/view/7529.

DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14502–14515, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/797134c3e42371bb4979a462eb2f042a-Abstract.html.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087, 2018. URL https://www.ifaamas.org/Proceedings/aamas2018/pdfs/p2085.pdf.

Yunhao Tang, Michal Valko, and Remi Munos. Taylor expansion policy optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9397–9406. PMLR, 2020. URL https://proceedings.mlr.press/v119/tang20d.html.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2380–2388, 2015. URL https://proceedings.mlr.press/v37/thomas15.html.

Hermann Thorisson. *Coupling, Stationarity, and Regeneration*. Springer, 2000. doi: 10.1007/978-1-4612-1236-2. URL https://hermann.hi.is/iid/csr/.

Patrick Wienhöft, Marnix Suilen, Thiago D. Simão, Clemens Dubslaff, Christel Baier, and Nils Jansen. More for less: Safe policy improvement with stronger performance guarantees. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4406–4415, 2023. doi: 10.24963/ijcai.2023/490. URL https://www.ijcai.org/proceedings/2023/490.

Yi Xie, Yangyang Xu, Yi Fan, and Bo Liu. SAT: Sequential agent tuning for coordinator free plug and play multi-LLM training with monotonic improvement guarantees. In *Proceedings of the 25th International Conference on Autonomous Agents and Multiagent Systems*, 2026. URL https://arxiv.org/abs/2605.05216.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24611–24624, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract.html.

Zijian Zhao and Sen Li. Is per-agent policy composition safe? rethinking successor-feature transfer in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2608.11658*, 2026. URL https://arxiv.org/abs/2608.11658.

A FULL PROOF OF THE ACTIVATION NORMAL FORM

We work with finite action sets and a finite horizon. Observation histories and states can also be taken finite; this is sufficient for the constructive sharpness claims and all experiments. The argument extends to kernels admitting a common measurable dominating construction, but no such extension is needed here. The episode reward is an arbitrary measurable functional in $[0, 1]$, not necessarily a sum of unit-bounded stage rewards. The experiments use the average of bounded stage rewards, a special case.

A.1 COMMON-KERNEL DECOMPOSITION

Fix an agent, time, and history, and denote its two action laws by K^0, K^1. Put $\delta = D_{\mathrm{TV}}(K^0, K^1) \leq d$. For $\delta < 1$, define the probability law

$$C(a) = \frac{\min\{K^0(a), K^1(a)\}}{1 - \delta}. \tag{15}$$

The numerator sums to $1 - \delta$. Since $d \geq \delta$, the factor $(1 - d)/(1 - \delta)$ is at most one, and hence $(1 - d)C(a) \leq K^v(a)$ for both versions. When $d > 0$, the residual kernels

$$R^v(a) = \frac{K^v(a) - (1 - d)C(a)}{d} \tag{16}$$

are nonnegative and each sums to one. If $d = 0$, the laws agree and one takes $C = K^0 = K^1$; the residual is unused. If $\delta = 1$, necessarily $d = 1$, the common kernel is unused, and the residual kernels are the original laws. When $d = 1$ but $\delta < 1$, the same conclusion follows from (16).

The parameter d is uniform over histories even when the actual distance δ varies. This uniformity permits residual flags independent of the visited state. Normalizing the overlap with the local δ, but selecting it with the fixed probability $1 - d$, is important: selecting residuals with a state-dependent probability would not produce independent episode-level activation events.

A.2 A JOINT COUPLING OF EVERY VERSION CONFIGURATION

For each agent i and each decision time $t \in \{1, \ldots, H\}$, generate an independent flag $B_{i,t} \sim \mathrm{Bernoulli}(d_i)$. Generate independent uniform random variables for sampling all action kernels, environmental transitions, and any reward randomness. Reuse these flags and uniforms across the 2^n counterfactual version configurations. At a visited history, use the common kernel when $B_{i,t} = 0$ and the selected version's residual kernel when $B_{i,t} = 1$. Each counterfactual trajectory has the correct marginal law, by the decomposition.

Let $C = \{i : \max_t B_{i,t} = 1\}$. Its membership events are independent over agents and have probabilities $p_i = 1 - (1 - d_i)^H$. Fix all randomness, including the flags. Compare two version assignments differing only in an agent outside C. They start from the same state, and at every step this agent uses the same common history-dependent kernel. Induction over time shows that every action, observation, transition, and reward agrees across these two trajectories: the induction hypothesis supplies equal histories, and the shared uniforms then supply equal actions and transitions. Consequently, the fixed-randomness episode return depends on the version assignment only through the coordinates in C.

Conditional on C, average over the residual-flag timings and all other randomness. Define $u_{C,T}$ to be this conditional expected return when the upgraded active agents are T. Versions outside C are irrelevant; $u_{C,T} \in [0, 1]$. The distribution of C does not depend on the version configuration. The law of total expectation yields (4). Zero-probability activation strata can be assigned arbitrary u values.

This proof does not require that trajectories for two configurations remain similar after an active residual choice. They may diverge arbitrarily. Nor does it assume a local reward decomposition. It only uses the fact that a version cannot matter before its agent has selected a version-dependent component under the constructed coupling.

A.3 Realizability of arbitrary conditional rewards

Fix any horizon H, distances d_i, and collection $u_{C,T} \in [0,1]$. Give agent i three actions: a neutral symbol $\perp$, an old-specific symbol o_i, and a new-specific symbol n_i. At every time and history its old policy chooses $\perp$ with probability $1-d_i$ and o_i with probability d_i; its new policy chooses $\perp$ with probability $1-d_i$ and n_i with probability d_i. Their total variation distance is exactly d_i.

The environment records, for each agent, whether a nonneutral symbol has appeared and which type. Under a fixed version an agent never emits conflicting old and new symbols. At the end of round H, let C be the agents that have emitted a nonneutral symbol and let T be those whose nonneutral symbol is new-specific. Return the deterministic terminal reward $u_{C,T}$. The recording state is finite, the environment has no direct access to policy version labels, and reward lies in $[0,1]$. Membership in C is independent across agents with the required probabilities, so the resulting value table is exactly (4).

The construction allows the environment and policy pair to vary while holding the retained summaries fixed. It therefore proves exactness of the summary class. It does not say that every u table can be realized by changing the reward of an arbitrary *particular* fixed policy pair. Detailed knowledge of that pair or its environment can restrict the class further.

A.4 Transport as a consequence

In the joint coupling, configurations S and T yield the same return whenever $C \cap (S \triangle T) = \varnothing$. The absolute difference is at most one otherwise. Taking expectations gives

$$|F(S) - F(T)| \le 1 - \prod_{i \in S \triangle T} (1 - p_i). \tag{17}$$

This remains valid with p_i obtained from upper bounds on the true distances; the proof may use residual kernels that activate more often than strictly necessary.

B Full proof of the sharp remainder

For completeness, define the order-k anchored operator and its norm for all $k \ge 0$:

$$\Delta_T F = \sum_{U \subseteq T} (-1)^{|T|-|U|} F(U), \qquad P_k(S) = \sum_{T \subseteq S,\, |T| \le k} \Delta_T F. \tag{18}$$

$$\gamma_k(m) = \begin{cases} 0, & m \le k, \\ \dfrac{1}{2}\left(1 + \displaystyle\sum_{j=0}^{k} \binom{m}{j}\binom{m-j-1}{k-j}\right), & m > k, \end{cases} \qquad B_k(p_S) = \mathbb{E}[\gamma_k(M_S)]. \tag{19}$$

The exact worst-case remainder is B_k for every order. This general statement specializes to Theorem 2; the following proof establishes it directly.

B.1 Anchored truncation coefficients

Let f be a function on subsets of a set of size m. By expanding the mixed differences, the coefficient of $f(U)$, $|U| = j \le k$, in $P_k([m])$ is

$$\sum_{r=j}^{k} (-1)^{r-j} \binom{m-j}{r-j} = (-1)^{k-j} \binom{m-j-1}{k-j}, \qquad m > k. \tag{20}$$

The alternating partial-sum identity follows by induction from Pascal's identity. If $m \le k$, the complete Möbius inversion gives $P_k([m]) = f([m])$. Applied to the constant function one, the truncation is always one, so its coefficients sum to one.

When $m > k$, define $\Lambda_{k,m}(f) = f([m]) - P_k([m])$. The coefficient of $f([m])$ is one, the other nonzero coefficients are the negatives of (20), and their total sum is zero. For any finite zero-sum

coefficient vector b,

$$\sup_{x\in[0,1]^q} b^\top x = \sum_{j:b_j>0} b_j = \frac{1}{2}\sum_j |b_j|, \qquad \inf_{x\in[0,1]^q} b^\top x = -\frac{1}{2}\sum_j |b_j|. \tag{21}$$

Both extrema are attained by setting coordinates to one on the appropriate sign and zero on the opposite sign. Multiplying the absolute coefficient for each cardinality by its multiplicity $\binom{m}{j}$ proves that the norm of $\Lambda_{k,m}$ is exactly $\gamma_k(m)$ in (19).

B.2 Conditioning on activation

For fixed C, the function $f_C(S) = u_{C,S\cap C}$ is constant along every coordinate outside C. Every mixed difference containing such a coordinate vanishes by pairwise cancellation. Therefore truncating f_C at order k on the target S is the same as truncating its restriction to the $m = |C \cap S|$ active target coordinates. The coordinates in $C \setminus S$ are held at their old versions.

Linearity, the conditional norm, and nonnegative weights imply

$$\begin{aligned} |F(S) - P_k(S)| &= \left| \sum_C w_C \big[f_C(S) - P_k^{f_C}(S) \big] \right| \\ &\le \sum_C w_C \gamma_k(|C \cap S|) = \mathbb{E}[\gamma_k(M_S)]. \end{aligned} \tag{22}$$

For a specified target S, choose in every activation stratum a sign-maximizing function from (21) on $C\cap S$. Values at arguments that upgrade agents outside S may be filled arbitrarily. When $|C\cap S| \le k$, choose any constant. Every conditional remainder is then positive and attains its norm, so equality holds in (22). The realizability part of Theorem 1 supplies a single admissible game attaining the sum. Complementing every reward supplies the negative extreme. This proves sharpness for each specified target; it does not assert that one game simultaneously attains every target's radius.

B.3 Special cases and comparison with the coarse bound

For $k = 0$, the remainder is a difference of two bounded values, giving $\gamma_0(m) = \mathbf{1}\{m \ge 1\}$. For $k = 1$ and $m \ge 2$, its positive coefficient mass is $1 + (m-1) = m$. For $k = 2$ and $m \ge 3$, write the remainder explicitly:

$$f([m]) - \binom{m-1}{2} f(\varnothing) + (m-2)\sum_i f(\{i\}) - \sum_{i<j} f(\{i,j\}). \tag{23}$$

Its positive mass is $1 + m(m-2) = (m-1)^2$; the negative mass is the same. For smaller m the second-order truncation is exact.

For general k and $m > k$, the identity

$$\binom{m}{j}\binom{m-j-1}{k-j} = \binom{m}{k+1}\binom{k+1}{j}\frac{k+1-j}{m-j} \le \binom{m}{k+1}\binom{k+1}{j} \tag{24}$$

shows that

$$\gamma_k(m) \le \frac{1 + \binom{m}{k+1}(2^{k+1}-1)}{2} \le 2^k \binom{m}{k+1}. \tag{25}$$

Independence gives $\mathbb{E}\binom{M_S}{k+1} = \sum_{|T|=k+1, T\subseteq S} \prod_{i\in T} p_i$. Since $(\sum_i p_i)^{k+1}$ includes $(k+1)!$ copies of every product with distinct indices,

$$B_k(p_S) \le 2^k \sum_{\substack{T\subseteq S \\ |T|=k+1}} \prod_{i\in T} p_i \le \frac{2^k}{(k+1)!}\rho_S^{k+1}. \tag{26}$$

For $|S| = k+1$, the only contributing activation event is that all target agents are active, and $\gamma_k(k+1) = 2^k$. Hence $B_k = 2^k \prod_{i\in S} p_i$ exactly.

B.4 Stable computation and endpoint behavior

The law of M_S is Poisson-binomial. Its probabilities can be computed by starting with $b_0 = 1$ and updating, for each p_i,

$$b_j^{\text{new}} = (1 - p_i) b_j^{\text{old}} + p_i b_{j-1}^{\text{old}}, \tag{27}$$

with missing indices zero. This costs $O(|S|^2)$ arithmetic operations and uses nonnegative quantities. Then sum $b_m \gamma_k(m)$. The implementation uses this recurrence rather than a cancellation-prone closed form.

For interpretation, let $e_1 = \sum_{i \in S} p_i$ and $e_2 = \sum_{i<j \in S} p_i p_j$. The special radii also satisfy

$$B_1 = e_1 - \mathbb{P}(M_S = 1), \qquad B_2 = 1 - e_1 + 2e_2 - \mathbb{P}(M_S = 0) - \mathbb{P}(M_S = 2). \tag{28}$$

These identities follow from the first two moments of M_S. Zero exposure makes that coordinate irrelevant. At exposure one, the coordinate is always active. If all agents are always active, the radius reduces to the deterministic operator norm, which can exceed one because the truncated polynomial itself need not lie in $[0, 1]$.

C Matching data-conditional indistinguishability

Theorem 8 (Three-agent information obstruction)**.** *For three agents with $p_i \le 1/5$, two admissible processes agree at value $1/2$ on every proper subset, while their full-target values are $1/2 \pm 4p_1p_2p_3$. No estimator of the full value from these low-order means can have worst-case absolute error smaller than $4p_1p_2p_3$.*

This is a special case of the following construction. For $n = k + 1$ and $0 \le p_i \le (1 + 2^k)^{-1}$, there are tables satisfying

$$F_+(S) = F_-(S) = \tfrac{1}{2} \quad (S \subsetneq [n]), \qquad F_\pm([n]) = \tfrac{1}{2} \pm 2^k \prod_i p_i. \tag{29}$$

We prove Theorem 8 for positive p_i first. Let $A = [n]$, $n = k + 1$, and $w = w_A = \prod_i p_i$. For $x_i = \mathbf{1}\{i \in S\}$, the odd-parity function satisfies

$$\mathrm{Odd}(x) - \tfrac{1}{2} = -\tfrac{1}{2} \prod_{i=1}^{n} (1 - 2x_i) = \sum_{C \subseteq A} a_C \prod_{i \in C} x_i, \qquad a_C = (-1)^{|C|+1} 2^{|C|-1}. \tag{30}$$

The empty-set product is one and $a_\varnothing = -1/2$.

Initialize all conditional rewards to $1/2$. In the fully active stratum set $u_{A,T} = \mathbf{1}\{|T| \text{ is odd}\}$. In every proper stratum $C \subsetneq A$ set

$$u_{C,T} = \tfrac{1}{2} - a_C \frac{w}{w_C} \mathbf{1}\{T = C\}. \tag{31}$$

These modifications cancel every proper monomial in the contribution of (30). Specifically, $\mathbf{1}\{S \cap C = C\} = \prod_{i \in C} x_i$, so the resulting value is

$$F(S) = \tfrac{1}{2} + a_A w \prod_{i=1}^{n} x_i. \tag{32}$$

Only the fully upgraded configuration differs from $1/2$.

It remains to check bounded rewards. For a proper C of size r, the magnitude of the correction is

$$|a_C| \frac{w}{w_C} = 2^{r-1} \prod_{i \notin C} \frac{p_i}{1 - p_i} \le 2^{r-1-k(n-r)} \le \tfrac{1}{2}, \tag{33}$$

because $p_i/(1-p_i) \le 2^{-k}$ and $r \le k = n-1$. The last inequality is equivalent to $r - k(n-r) \le 0$, whose maximum over proper r is zero at $r = k$. Thus every reward is in $[0, 1]$. The full-stratum rewards are binary already. If any $p_i = 0$, the claimed separation is zero and two constant-$1/2$ worlds suffice; the construction can also be interpreted by continuity.

The top coefficient a_A has magnitude $2^{n-1} = 2^k$. Choosing the table just constructed or its pointwise complement produces both signs in (29). Realizability follows from Theorem 1. The two worlds have identical full sets of proper-subset observations, not just identical individual or pair gains. Any estimator given these observations returns the same number in both worlds, and its maximum absolute error is at least half their separation, $2^k w$. The anchored estimator returns $1/2$ and attains that error. For $k = 2, n = 3$, Theorem 2 gives the same value $4w$.

In the positive world, all proper gains are zero and the full gain is nonnegative. In the negative world, the full gain is $-2^k w$. If $\varepsilon < 2^k w$, they lie on opposite sides of the compatibility decision. A deterministic sound evaluator cannot approve the shared evidence. A randomized evaluator whose false-approval probability is at most α in the negative world can approve that identical evidence with probability at most α in the positive world as well. This is an information obstruction, not a finite-sample artifact.

Independent return-only sampling. Replace the deterministic terminal payoff $u_{C,T}$ by a Bernoulli draw with success probability $u_{C,T}$, using independent episode randomness. This preserves the value table and makes the marginal terminal return on every proper query Bernoulli$(1/2)$ in both worlds. By induction on an adaptively selected sequence of proper queries, the observed query/return transcript has the same law in both worlds: conditional on the transcript, the next query is selected in the same way and its response has the same Bernoulli law. More independent return-only samples cannot resolve the ambiguity. This argument does not identify the full trajectory laws or coupled cross-version return laws; access to those is richer than the value-summary or independent return-only oracle.

D Unrestricted configuration-query lower bound

Proposition 9 (Hidden failure among unrestricted mixtures)**.** *Fix a known baseline value $1/2$, allow every nonbaseline configuration, and let $0 \le \varepsilon < 1/2$. An exact-value deterministic evaluator that always accepts compatible tables and always rejects incompatible tables requires $2^n - 1$ nonbaseline queries in the worst case. A randomized evaluator with error at most $\delta < 1/2$ on each of the following good/bad classes, and at most q queries on every run, satisfies*

$$q \ge (1 - 2\delta)(2^n - 1). \tag{34}$$

Proof. Put $N = 2^n - 1$. The good table is $F(S) = 1/2$ everywhere. For each nonempty T, the bad table agrees everywhere except $F(T) = 0$. Such tables are one-step cooperative games with deterministic binary-version policies, so no extra representational assumption is needed.

Run a deterministic correct evaluator on the good table. If a nonempty T remains unqueried, all observed answers are identical on its bad table, forcing the same output and contradicting correctness. For randomized evaluation, couple the random seed of runs on the good table and on a uniformly selected bad table. Before the hidden configuration is queried, transcripts coincide. On the good run at most q distinct nonbaseline queries are made, so the probability of hitting the uniform hidden location is at most q/N. The difference in acceptance probabilities between the good table and the uniform bad distribution is therefore at most q/N. Correctness requires the former acceptance probability to be at least $1 - \delta$ and the latter at most δ. Rearranging gives (34). □

This result addresses distinguishing or complete certification, not soundness alone. Allowing arbitrary abstention without any completeness requirement permits a zero-query rule that never certifies. Also, an evaluator asked to evaluate only the fully upgraded endpoint can simply query that endpoint; the exponential obstruction is a consequence of the deployment family.

E Sharp polytope evaluation and verified numerical duals

Theorem 10 (Exact conditional summary evaluation)**.** *For nonempty feasible evidence,* (10) *is the smallest gain among all activation-class processes consistent with the queried intervals. Robust compatibility over that class holds exactly when its minimum over allowed targets is at least $-\varepsilon$. Simultaneous input intervals covering all queried values with probability $1 - \alpha$ give false-certification probability at most α.*

The number of coordinates (C, T) with $T \subseteq C \subseteq [n]$ is $\sum_C 2^{|C|} = 3^n$. Define the box

$$\mathcal{V}(p) = \{v : 0 \le v_{C,T} \le w_C\}. \tag{35}$$

For each target S, $A_S v = \sum_C v_{C,S\cap C}$. By Theorem 1, the image of this box under all rows A_S is exactly $\mathcal{A}(p)$. Restricting the image to the queried intervals is therefore equivalent to adding the corresponding linear inequalities. A nonempty intersection is compact, and each linear objective has an attained minimum. Conversely, a minimizer determines $u_{C,T} = v_{C,T}/w_C$ whenever $w_C > 0$, and an arbitrary value when $w_C = 0$; the realizability construction produces an admissible process attaining it. This proves Theorem 10.

The statement of necessity and sufficiency is relative to *all* processes consistent with the retained intervals. Different targets may have different extremizing processes; this does not weaken the uniform criterion. Robust compatibility fails if there is at least one consistent process and one allowed target with a degrading gain, equivalently if the minimum over both process and target is below threshold. If the evidence constraints are infeasible, the implementation raises an error and abstains. It does not interpret inconsistent assumptions as a vacuous safety certificate.

E.1 A DUAL CERTIFICATE INDEPENDENT OF SOLVER FEASIBILITY TOLERANCES

Stack the upper and lower interval inequalities as $Av \le b$, where the lower rows are negated. For a fixed target put $c = A_S - A_\varnothing$, and let w_j be the upper bound for coordinate j. For any vector $y \le 0$, define

$$\mathcal{D}(y) = b^\top y + \sum_j w_j \min\{0, c_j - (A^\top y)_j\}. \tag{36}$$

For every feasible v, since $y^\top Av \ge y^\top b$,

$$\begin{aligned} c^\top v &= y^\top Av + (c - A^\top y)^\top v \\ &\ge b^\top y + \min_{0\le z\le w} (c - A^\top y)^\top z = \mathcal{D}(y). \end{aligned} \tag{37}$$

Thus (36) is a lower bound for *any* such y, not only an optimal numerical multiplier. Linear-programming duality also implies that maximizing it recovers the primal optimum for a feasible problem, but validity of an individual checked certificate does not require proving numerical optimality.

The implementation obtains inequality marginals from HiGHS, clips positive roundoff to zero, and reevaluates (36) with exact rational arithmetic. Every input floating-point number is treated as its exact binary rational value. In particular, w_C is recomputed as a rational product of the supplied p_i and $1-p_i$, rather than taken from a rounded product. The resulting rational lower bound is converted to floating point and rounded downward by one representable step. The empty target is set to zero exactly. Primal values, dual values, and constraint residuals remain available as diagnostics.

This procedure verifies the optimization lower bound for the supplied numerical summaries and intervals. It does *not* prove that estimated policy-distance bounds are valid, that confidence intervals cover true values, or that a floating-point simulator exactly represents an intended physical environment. These are separate assumptions. Our linear-programming experiments with exact dynamic programming (DP) pad value intervals by 2×10^{-12} and report comparisons at explicit numerical tolerances.

F STATISTICAL PROOFS AND SHARED-OBSERVATION ACCOUNTING

Set $q = |\mathcal{Q}_k|$. A simultaneous interval construction is

$$r_Q = \sqrt{\frac{\log(2q/\alpha)}{2N_Q}}, \qquad [\ell_Q, u_Q] = [\widehat{F}(Q) - r_Q, \widehat{F}(Q) + r_Q] \cap [0, 1]. \tag{38}$$

Using the combined gain coefficients $a_Q(S)$, its lower bound is

$$L_k(S) = \sum_{a_Q(S)\ge 0} a_Q(S)\ell_Q + \sum_{a_Q(S)<0} a_Q(S)u_Q - B_k(p_S). \tag{39}$$

With independent samples across queries, a direct contrast is

$$L_{\mathrm{ctr}}(S) = \sum_Q a_Q(S)\widehat{F}(Q) - B_k(p_S) - \sqrt{\frac{\log(|\mathcal{R}|/\alpha)}{2}\sum_Q \frac{a_Q(S)^2}{N_Q}}. \tag{40}$$

The empty target is always assigned the exact gain zero. Using $|\mathcal{R}|$ rather than excluding that target only makes this independent baseline conservative. Let $X_{Q,r} \in [0,1]$, $r = 1, \ldots, N_Q$, be independent within each queried configuration, each with mean $F(Q)$. Candidate policies and sample counts are chosen without using these observations. Hoeffding's inequality gives

$$\mathbb{P}\{|\widehat{F}(Q) - F(Q)| > r_Q\} \le 2\exp(-2N_Q r_Q^2). \tag{41}$$

Substituting (38) and applying a union bound gives an event E of probability at least $1-\alpha$ on which every interval contains its value. This interval construction does not require independence *between* configurations; the direct contrast construction below does.

On E, the minimum of the linear gain predictor over the interval box is exactly the signed endpoint expression in (39). Subtracting the deterministic remainder gives a lower bound on $D(S)$. Because E concerns the queried values once and for all, this conclusion holds simultaneously for every target in the deployment family, including a target selected after examining the intervals. The same event makes the true activation vector feasible in (10), proving LP statistical validity.

For transport, $D(S) \ge F(Q) - F(\varnothing) - \mathrm{dist}_p(S,Q)$, so $\ell_Q - u_\varnothing - \mathrm{dist}_p(S,Q)$ is valid on E. For $Q = \varnothing$, the baseline cancels and the sharper expression is simply $-\mathrm{dist}_p(S,\varnothing)$. Also $D(S) \ge -u_\varnothing$ because $F(S) \ge 0$. These are the interval implementation's optional improvements.

F.1 Direct independent contrasts

Now assume all $X_{Q,r}$ are independent across both indices. For a target S, the predictor error is

$$\sum_Q a_Q(S)(\widehat{F}(Q) - F(Q)) = \sum_{Q,r} \frac{a_Q(S)}{N_Q}(X_{Q,r} - F(Q)). \tag{42}$$

Each summand has range length $|a_Q(S)|/N_Q$. A one-sided Hoeffding bound therefore gives

$$\mathbb{P}\left\{\sum_Q a_Q(S)(\widehat{F}(Q) - F(Q)) > t\right\} \le \exp\left[-\frac{2t^2}{\sum_Q a_Q(S)^2/N_Q}\right]. \tag{43}$$

When every coefficient is zero the predictor is exact and no tail bound is needed. Otherwise, choose t as in (40) and union over the allowed targets. Combining this with the deterministic remainder proves that all direct lower bounds hold with probability at least $1-\alpha$.

Taking the maximum of the interval and contrast methods is valid only after accounting for both possible failure events. We allocate $\alpha/2$ to each and apply a union bound. Taking the maximum of multiple lower bounds derived from the *same* simultaneous interval event requires no additional split. The experiments follow these rules.

F.2 Explicit pairwise gain coefficients

For a target with $m \ge 1$ upgraded agents, the coefficients of $P_2(S) - F(\varnothing)$ are

$$a_\varnothing(S) = \frac{m(m-3)}{2}, \quad a_{\{i\}}(S) = 2 - m\ (i \in S), \quad a_{\{i,j\}}(S) = 1\ (i,j \in S), \tag{44}$$

with absent indices omitted and all other coefficients zero. Formula (44) also covers $m = 1, 2$ and reduces to the observed value difference. At $m = 3$, the gain predictor is $\sum_{i<j} F(\{i,j\}) - \sum_i F(\{i\})$, with no baseline term. The implementation constructs these combined coefficients before propagating uncertainty.

For a single fixed target and total evaluation budget $B = \sum_Q N_Q$, Cauchy–Schwarz implies

$$\sum_Q \frac{a_Q(S)^2}{N_Q} \ge \frac{(\sum_Q |a_Q(S)|)^2}{B}. \tag{45}$$

Ignoring integrality and zero coefficients, equality allocates N_Q proportional to $|a_Q(S)|$. This elementary observation explains why shared baseline leverage matters, but it is not a solution to optimal allocation over a whole deployment family. We use fixed designs and make no optimal sample-complexity claim.

G Reward-timed representation and sharp remainder

Theorem 11 (Sharp reward-timed anchored bound)**.** *If $G = \sum_t \beta_t R_t$ with $0 \le R_t \le 1$, $\beta_t \ge 0$, $\sum_t \beta_t \le 1$, and R_t paid by round t, then the anchored order-k remainder is at most $B_k^{\mathrm{time}}(S) = \sum_t \beta_t B_k(p(t)_S)$. This bound is attained in the reward-timing summary class for every fixed target and each sign.*

Write $F_t(S) = \mathbb{E}_S[R_t]$, where $R_t \in [0,1]$ is measurable with respect to the trajectory through the completion of decision round t. Assume $G = \sum_t \beta_t R_t$ with deterministic nonnegative weights whose sum is at most one. The condition is a property of the return definition, not a decomposition inferred by the evaluator from a few observed rewards. In particular, it cannot be assumed for an arbitrary terminal objective.

At round r, agent i's two kernels differ by at most the supplied $d_{i,r}$ at every specified history. Repeat the common-kernel construction of Appendix A with independent flags $B_{i,r} \sim \mathrm{Bernoulli}(d_{i,r})$. Let $C_t = \{i : B_{i,r} = 1 \text{ for some } r \le t\}$. For each fixed t, its membership events are independent across agents, with probabilities

$$p_{i,t} = 1 - \prod_{r=1}^{t}(1 - d_{i,r}), \qquad w_C^{(t)} = \prod_{i \in C} p_{i,t} \prod_{i \notin C}(1 - p_{i,t}). \tag{46}$$

The sets C_t are generally dependent across time; independence across time of these sets is neither assumed nor needed. Conditional on a full random tape, $R_t(S)$ depends only on the versions in C_t, by the same induction as before. Consequently,

$$F_t(S) = \sum_C w_C^{(t)} u_{t,C,S\cap C}, \quad 0 \le u_{t,C,T} \le 1, \qquad F(S) = \sum_t \beta_t F_t(S). \tag{47}$$

Anchored truncation is linear. The sharp prefix bound and nonnegative weights give

$$|F(S) - P_k F(S)| = \left| \sum_t \beta_t \big(F_t(S) - P_k F_t(S)\big) \right| \tag{48}$$

$$\le \sum_t \beta_t B_k(p_{S,t}). \tag{49}$$

Here $P_k F$ denotes the anchored operator applied to the value table F; the query family is $\mathcal{Q}_k$.

Attainment within a single process. Use the three-action recording process from Appendix A, now emitting a version-specific symbol with probability $d_{i,r}$ in round r. At each time t the state records (C_t, T_t), the activated agents and the observed new-specific versions. Assign the stage reward $R_t = u_{t,C_t,T_t}$. Any collection of such tables is realizable in one finite process, even though activation sets are nested. For the specified target S and order k, choose every $u_{t,C,\cdot}$ to maximize the positive remainder functional on $C \cap S$. Each prefix then attains $+B_k(p_{S,t})$. There is no conflict between these choices: rewards at different times may be different deterministic functions of the recording state and time. Their weighted sum attains the upper bound. Pointwise complements give the negative sign. This also explains why the claim is sharp over the reward-timing summary class but may not be sharp for a particular fixed reward function.

Dominance and limiting cases. The finite functional norm $\gamma_k(m)$ is nondecreasing in m: any table on m coordinates embeds into $m+1$ by ignoring the last coordinate, preserving the remainder. Couple prefix active counts so that $M_{S,t} \le M_{S,H}$. It follows that $B_k(p_{S,t}) \le B_k(p_{S,H})$, and summing weights at most one gives the stated dominance. With all weight at H, equality recovers the original remainder. With $H = 1$, the averaged-return and terminal-return versions coincide. For homogeneous small changes, the reward-timed leading triple sum is $4\sum_t \beta_t \sum_{i<j<l} p_{i,t}p_{j,t}p_{l,t}$; it can be materially smaller than evaluating every factor at H, but still increases as more interactions can precede rewards.

H COUPLED OBSERVATIONS: RANGE, MOMENTS, AND CONCENTRATION

A coupled batch is a random vector $X_r = (X_{Q,r} : Q \in \mathcal{Q}_k)$ with correct bounded marginal returns. Batches $r = 1, \dots, N$ are independent and identically distributed; columns within a batch may have arbitrary dependence. We form the full linear gain contrast $Y_{S,r} = \sum_Q a_Q(S) X_{Q,r}$ *before* applying concentration. Treating common-random-number columns as independent in the weighted Hoeffding formula would generally be invalid.

H.1 DETERMINISTIC RANGE AND ACTIVATION MOMENT

The gain coefficients sum to zero. Over a full $[0,1]$ cube, the maximum and minimum of their linear functional are therefore $\pm A_k(m)$, where

$$A_k(m) = \frac{1}{2} \sum_Q |a_Q([m])|, \qquad m = |S|. \tag{50}$$

The coefficients are those of the *gain*, including the baseline subtraction; they differ from the remainder coefficients defining γ_k. Thus $Y_{S,r}$ has range length $W_S = 2A_k(|S|)$ under any coupling. For $k = 2$, the explicit coefficients in (44) give $A_2(0) = 0$, $A_2(1) = A_2(2) = 1$, and $A_2(m) = m(m-2)$ for $m \geq 3$. For first order, $A_1(m) = m$. For exhaustive evaluation, the direct contrast $X_{S,r} - X_{\varnothing,r}$ has range length two, regardless of team size.

Under the residual-flag coupling specifically, a fixed random tape produces a bounded function of $S \cap C$. All mixed differences involving inactive coordinates cancel. Applying the gain operator then reduces to the operator on $M_S = |S \cap C|$ coordinates, giving the pointwise bound

$$|Y_{S,r}| \leq A_k(M_S), \qquad \mathrm{Var}(Y_S) \leq \mathbb{E}Y_S^2 \leq V_S^\star := \mathbb{E}[A_k(M_S)^2]. \tag{51}$$

For the exhaustive gain contrast the corresponding moment bound is $1 - \prod_{i \in S}(1 - p_i)$. These are upper bounds, not claims that activation coupling minimizes variance. The implemented sampler uses independent flags with the maximum time/history distance, together with shared action, transition, reward, and initial-state uniforms. It pays for every simulated configuration. The time-specific remainder can still be used with this coarser flag coupling: one bounds the deterministic expectation error and the other bounds the sampling variation.

H.2 EMPIRICAL AND KNOWN-MOMENT BERNSTEIN BOUNDS

Let $\widehat{V}_S = (N-1)^{-1} \sum_r (Y_{S,r} - \overline{Y}_S)^2$. Apply Theorem 4 of Maurer & Pontil (2009) to $Z_r = (A_k(m) - Y_{S,r})/W_S \in [0,1]$. After multiplying back by W_S, with failure probability at most δ,

$$\mathbb{E}Y_S \geq \overline{Y}_S - \sqrt{\frac{2\widehat{V}_S \log(2/\delta)}{N}} - \frac{7W_S \log(2/\delta)}{3(N-1)}. \tag{52}$$

If the functional is identically zero, no concentration is needed. Substitute $\delta = \alpha/K$, union over the nonempty allowed targets, and subtract the deterministic remainder to prove Theorem 7. Dependence among target contrasts is harmless to this union bound. Either the terminal or reward-timed deterministic remainder is valid. Sample variance is calculated on the entire contrast, retaining cross-column covariance.

Using the known upper variance bound (51), ordinary Bernstein concentration (Theorem 3 of Maurer & Pontil (2009)) gives the alternative

$$L_{\mathrm{act}}(S) = \overline{Y}_S - B_k(S) - \sqrt{\frac{2V_S^\star \log(K/\alpha)}{N}} - \frac{W_S \log(K/\alpha)}{3N}. \tag{53}$$

The relevant centered upper-tail variable is bounded above by W_S, and $V_S^\star$ bounds its variance. This form follows, for example, from the standard sub-gamma inversion of Bernstein's inequality. It should not be used with arbitrary common-random-number coupling unless the activation envelope has been established for that coupling. The empirical bound, by contrast, needs only independent bounded rows with correct marginals.

What is and is not uniform. For each predetermined budget and method the lower bounds hold simultaneously over configurations. They do not automatically hold over an adaptively selected budget, candidate pool, or statistical method. The plots reuse prefixes for efficiency but make fixed-budget comparisons; no optional stopping is performed. Selecting the maximum of the empirical and known-moment methods would require splitting the failure budget. Our tables report these methods separately at .05 and never present their unadjusted maximum as a valid combined certificate. In all algorithms the baseline comparison is set to zero exactly.

I GLOBAL MINIMAX RECOVERY FROM LOW-ORDER VALUES

Optimal recovery of a scalar linear functional from linear information over symmetric convex classes is classical (Micchelli & Rivlin, 1977; Foucart & Paouris, 2023). This section specializes its geometry to the activation representation and distinguishes the error of a prescribed predictor from the error of the best possible estimator. The loss is absolute error at a specified target, the observations are exact queried means, and the uncertainty set is the activation class $\mathcal{A}(p)$ in Theorem 1. These qualifications identify a complete estimation problem rather than an asymptotic approximation.

Write $j = (C, T)$, $W_j = w_C$, and $a_S(j) = \mathbf{1}\{S \cap C = T\}$. Center the box representation by $z_j = v_j - W_j/2$. Since $\sum_j a_S(j) W_j = 1$, every value has the form $F(S) = 1/2 + a_S z$. Set $\mathcal{Z} = \{z : |z_j| \le W_j/2\}$ and let $A_\mathcal{Q}$ contain the queried incidence rows. For observed centered values y, define the conditional radius

$$r(y) = \frac{1}{2}\left[\max_{z \in \mathcal{Z}: A_\mathcal{Q} z = y} a_S z - \min_{z \in \mathcal{Z}: A_\mathcal{Q} z = y} a_S z\right].$$

The midpoint of this interval minimizes worst-case absolute error conditional on y; no statistical assumption is involved.

Proof of Theorem 3. If z^+, z^- are feasible for the same y, their half-difference $z = (z^+ - z^-)/2$ lies in $\mathcal{Z}$ and satisfies $A_\mathcal{Q} z = 0$. Hence every conditional half-width is at most

$$R_* = \max_{z \in \mathcal{Z}: A_\mathcal{Q} z = 0} a_S z.$$

Conversely, a maximizing z exists by compactness. Both z and $-z$ are feasible at $y = 0$ and have target separation $2R_*$. The worst data therefore occur at the common query value $1/2$, and the displayed maximum is the global minimax radius. Because the baseline is queried, the same radius applies to estimating $D(S)$.

For an arbitrary multiplier λ, the equality constraints give

$$a_S z = (a_S - \lambda^\mathsf{T} A_\mathcal{Q}) z \le \frac{1}{2} \sum_j W_j \left| a_S(j) - \sum_{Q \in \mathcal{Q}} \lambda_Q a_Q(j) \right|.$$

Linear-program duality makes the infimum of this support bound equal to R_*. The dual is attained: it is a finite feasible linear program after introducing absolute-value slack variables. If λ_* minimizes it, the fixed affine estimator

$$\widehat{F}_*(S) = \tfrac{1}{2} + \sum_{Q \in \mathcal{Q}} \lambda_{*,Q} \big(F(Q) - \tfrac{1}{2}\big)$$

has worst-case error at most R_* over the entire class. The centered pair above proves the reverse inequality for every estimator, including nonlinear estimators. Thus this affine estimator is globally minimax, while conditional interval midpoints can be better at particular noncentral data. □

I.1 AN ANALYTIC MINIMAX REGIME FOR EVERY TARGET SIZE

Corollary 12 (A sufficient small-exposure regime)**.** *For a target S of size $m \ge 3$, if*

$$\prod_{k \in S \setminus \{i,j\}} (1 - p_k) \ge \frac{4}{5} \qquad \textit{for every } \{i, j\} \subseteq S, \tag{54}$$

then $R_*(S) = B_2(p_S)$ for pairwise data. Equal exposures satisfy this condition when $p \le 1 - (4/5)^{1/(m-2)}$. The condition is sufficient, not necessary. The exact homogeneous threshold is given by Theorem 4.

Proof. It is enough to construct a centered world attaining the anchored radius while all pairwise means equal $1/2$. Restrict to the target agents; any others can be ignored in the reward. Let $M = |C|$, $\tau = \Pr(M \ge 3)$, $a_i = \Pr(i \in C, M \ge 3)$, and $b_{ij} = \Pr(i, j \in C, M \ge 3)$. Write centered conditional rewards as $x_{C,T} = u_{C,T} - 1/2$. In strata $|C| \ge 3$, set $x_{C,T} = 1/2$ if $|T| = 1$ or $T = C$, set it to $-1/2$ if $|T| = 0$ or 2, and set all remaining coordinates to zero. These are precisely the signs of the nonzero anchored-remainder coefficients. In low-order strata set only

$$x_{\varnothing,\varnothing} = \frac{\tau}{2w_\varnothing}, \quad x_{\{i\},\{i\}} = -\frac{a_i}{w_{\{i\}}}, \quad x_{\{i,j\},\{i,j\}} = \frac{2b_{ij}}{w_{\{i,j\}}}$$

nonzero. A coordinate with zero mass can be set to zero.

For a query Q of size at most two, the high-order strata contribute $-\tau/2 + \sum_{i\in Q} a_i - 2\sum_{\{i,j\}\subseteq Q} b_{ij}$ to its centered value. The three displayed low-order assignments cancel these terms exactly. They do not change the anchored error, because the anchored operator is exact within strata of size at most two. It remains to verify that the assignments lie in the centered unit box.

Condition (54) implies $1 - p_i \ge .8$ for every target agent, by choosing a pair not containing it. It also gives $w_\varnothing \ge (4/5)^3 = 64/125$, so $\tau \le 1 - w_\varnothing \le w_\varnothing$. For a pair with nonzero mass, put $q_{-ij} = \prod_{k\ne i,j}(1 - p_k)$. Then $2b_{ij}/w_{\{i,j\}} = 2(1 - q_{-ij})/q_{-ij} \le 1/2$. For a singleton let $q_{-i} = \prod_{j\ne i}(1 - p_j) \ge .64$. Independence and $r_j = p_j/(1 - p_j)$ give

$$\frac{a_i}{w_{\{i\}}} = q_{-i}^{-1} - 1 - \sum_{j\ne i} r_j \le q_{-i}^{-1} - 1 + \log q_{-i} \le .64^{-1} - 1 + \log(.64) < \tfrac{1}{2}.$$

Here $\sum r_j \ge \sum \log(1 + r_j) = -\log q_{-i}$, and the final function decreases on $(0, 1]$. Thus every assignment is feasible. The positive world attains $+B_2$, and its complement attains $-B_2$ with the same central query values. The minimax lower bound follows. For equal exposures, (54) is $(1 - p)^{m-2} \ge 4/5$. □

I.2 Equal exposures and a small, verifiable numerical problem

For a target of m agents with common exposure p, query every subset of size at most k. Permuting the agents leaves the constraints, box, and full-target objective invariant. Averaging a feasible centered conditional-reward table over permutations therefore preserves its objective and feasibility. Its entries depend only on $c = |C|$ and $j = |T|$, and can be written $x_{c,j} \in [-1/2, 1/2]$. The query row for a set of size ℓ is

$$a_\ell(c, j) = \binom{\ell}{j}\binom{m - \ell}{c - j} p^c (1 - p)^{m-c}, \tag{55}$$

where binomial coefficients outside their ranges equal zero. The reduced primal maximizes $a_m x$ subject to $a_\ell x = 0$, $0 \le \ell \le k$. It has $(m + 1)(m + 2)/2$ box variables and $k + 1$ equality constraints, instead of 3^m variables and all low-order rows.

We solve this program for $m = 3, 4, 5, 6$ and eleven rational exposures from .01 to 1. The reported lower endpoint is an exactly feasible rational primal value: box-active coordinates are fixed at $\pm 1/2$, and the remaining equality system is solved in rational arithmetic. The upper endpoint is the rational support bound for the numerical dual multiplier. This gives certified brackets even if the floating-point optimizer stops short of optimality. The largest bracket width on the grid is below 9×10^{-9}; the entries in Table 3 agree to better than 10^{-10}. At $p = 1$, an unqueried full-target payoff is unconstrained in $[0, 1]$, so $R_* = 1/2$ for $m > 2$, while $B_2 = (m - 1)^2$. This simple endpoint already precludes a general minimax interpretation of B_2.

I.3 Random restrictions and identifiability of quadratic responses

An active coordinate is an alive coordinate in a random restriction. The activation class permits a separate function u_C for each restriction C; it is therefore larger than applying a fixed noise or

restriction operator to one Boolean function. This distinction explains both its pessimism and the smaller uncertainty of the episode-gated class in Appendix L. Bounds on mixed differences follow because a difference over coordinates J vanishes unless $J \subseteq C$. Independence supplies the factor $\prod_{i \in J} p_i$. Random restrictions and their interaction with Fourier structure are classical (O'Donnell, 2014); the present coordinates are anchored set differences, not an assertion that Fourier and Möbius coefficients coincide.

The scalar-value design problem also has a basic rank obstruction. Let $d_2 = 1 + n + \binom{n}{2}$, the dimension of unrestricted quadratic pseudo-Boolean functions. Any nonadaptive collection of fewer than d_2 evaluations has an evaluation matrix with a nonzero kernel. A sufficiently small perturbation along this kernel, around the constant function $1/2$, gives two distinct bounded quadratic responses with identical evaluations. For a deterministic adaptive design, run it along the constant-response transcript and apply the same kernel argument to the queried rows. Thus exact identification of every quadratic response needs at least d_2 scalar evaluations in the worst case. Fractional factorial designs may improve balance and statistical conditioning; they do not evade this dimension requirement (Box et al., 2005). This argument concerns full quadratic identification, not the weaker task of selectively certifying some teams.

J Reachability, occupancy, and dependent activations

J.1 Restricting the supremum to a verified reachable set

Let $\mathcal{H}_i$ include every history of agent i reachable under any allowed deployment or queried configuration. Suppose both version kernels are defined there and their total variation is bounded by d_i. Extend the two kernels identically outside $\mathcal{H}_i$. The modified kernels satisfy the uniform bound everywhere. Induction on decision time shows that each allowed or queried configuration has its original trajectory law: by definition it never leaves the supplied reachable set. Applying Theorem 1 to the extension proves the same upper bound for the retained configurations. A reachable set supplied by an exact support analysis is sufficient; a finite sample of observed histories is not a support certificate. Retaining a particular support graph can further shrink the class, so this argument supplies an upper bound rather than an exact characterization relative to that additional information.

J.2 An occupancy-weighted transport bound

Let S, T be two fixed configurations. At joint history h and time t, let $\delta_t(h)$ upper-bound the total variation between their joint action kernels. If policy randomization factors, one available choice is

$$\delta_t(h) = 1 - \prod_{i \in S \triangle T} \big(1 - D_{\mathrm{TV}}(\pi^0_{i,t}(\cdot \mid h_i), \pi^1_{i,t}(\cdot \mid h_i))\big).$$

Let μ_t^T be the unconditional history law of the reference configuration T. Couple the two joint action laws maximally while their histories agree, and use the same environment randomness until the first disagreement. The probability of first disagreement at time t is at most $\mathbb{E}_{\mu_t^T} \delta_t(h)$: discarding the indicator of previous agreement can only increase that expectation. Once a disagreement occurs, the total return difference is bounded by one. Consequently,

$$|F(S) - F(T)| \le \min\left\{1, \sum_{t=1}^{H} \mathbb{E}_{\mu_t^T} \delta_t(h)\right\}. \tag{56}$$

If $G = \sum_r \beta_r G_r$, where $G_r \in [0,1]$ is paid by round r, first disagreement at t can affect only rewards $r \ge t$. Multiplying each expectation in (56) by $\sum_{r \ge t} \beta_r$ proves the reward-timed version. Estimated occupancies and divergences would require their own simultaneous statistical upper bounds.

J.3 Why marginal occupancy averages do not yield a cubic remainder

A public context $Z \sim \mathrm{Bernoulli}(q)$ is observed before a one-step, three-agent episode. Each old policy emits zero. Each new policy also emits zero when $Z = 0$, and emits one when $Z = 1$.

These policies are deterministic, hence satisfy conditional independence. The common reward is one except when $Z = 1$ and all three actions are one. Its version table is

$$F(S) = 1 - q\mathbf{1}\{S = \{1, 2, 3\}\}.$$

The expected total variation of each old/new pair is exactly q under *every* configuration's occupancy law. Nonetheless the pairwise remainder is q, rather than $4q^3$. The adverse context activates all differences together. Uniform distances are one on that context, so the original theorem correctly gives no small-exposure conclusion. The example refutes replacement by separate occupancy means even when means under all mixed configurations are supplied.

More generally, suppose a single joint coupling and a version-independent random set C satisfy the pathwise inactive-version property: conditional on the coupled randomness, a return can depend on S only through $S \cap C$. The proof of the anchored bound then still gives

$$|F(S) - P_2(S)| \le \mathbb{E}\big[(|C \cap S| - 1)^2 \mathbf{1}\{|C \cap S| \ge 3\}\big] \le 4 \sum_{\{i,j,k\} \subseteq S} \Pr(i, j, k \in C).$$

Independence is needed only to replace joint activation probabilities by products, not for the conditional functional-norm argument. State-dependent residual probabilities generally make the activation set depend on the version-specific trajectory. They therefore supply neither independent activations nor a single version-independent active-set law automatically. A valid common envelope or verified joint activation probabilities would suffice, but the counterexample shows that marginal occupancy averages alone do not.

K Causal-cone bounds and polynomial target search

K.1 A checkable structural assumption

Assume $G = \sum_a \beta_a G_a$, $\beta_a \ge 0$, $\sum_a \beta_a \le 1$, and $0 \le G_a \le 1$. For each component give a set V_a such that, holding exogenous randomness fixed, changing versions outside V_a cannot change G_a. This is a property of the complete time-unrolled data flow, including policy observations and state transitions. A graphical support analysis can verify it when these dependencies are explicit. A graph of immediate pairwise rewards alone does not verify it: a remote policy can alter future local states.

Proof of Theorem 5. Write $F_a(S) = \mathbb{E}[G_a \mid S]$. By the structural assumption, $F_a(S)$ depends only on $S \cap V_a$. Its anchored differences involving an agent outside V_a vanish. Apply the activation bound to this restricted function, then use the triangle inequality and linearity of the anchored predictor:

$$|F(S) - P_2(S)| \le \sum_a \beta_a |F_a(S \cap V_a) - P_{2,a}(S \cap V_a)| \le \sum_a \beta_a B_2(p_{S \cap V_a}).$$

A component paid at time t_a can use prefix exposures $p_i(t_a)$. If every cone has size at most two, the remainder vanishes. If every cone has size at most q and $p_i \le p_{\max}$, it is at most $4\binom{q}{3} p_{\max}^3$. Also $h_{ij} = 0$ whenever no cone contains both i and j, because each component's mixed difference vanishes. Only $1 + n + |E|$ configurations are needed, where E is the union of cone-pairs. □

If a transition graph has maximum degree Δ, a reward's ancestor set can still grow like $1 + \Delta + \cdots + \Delta^H$ before meeting the system-size cap. The theorem uses that ancestor set, not Δ by itself. In the reservation-queue experiment, each queue has isolated state and three clients, so its cone size remains three throughout the episode. This structural conclusion cannot be transferred to a shared-resource task without verifying its dependencies.

K.2 A polynomial certificate with an explicit relaxation penalty

Let $x_i = \mathbf{1}\{i \in S\}$ and expand the coarse local bound as

$$B_{\mathrm{loc}}(S) \le \sum_{i<j<k} c_{ijk} x_i x_j x_k, \qquad c_{ijk} = 4 p_i p_j p_k \sum_{a : \{i,j,k\} \subseteq V_a} \beta_a.$$

Since $x_i x_j x_k \le (x_i + x_j + x_k)/3$ on binary vectors, put $b_i = \frac{1}{3}\sum_{\{j,k\}:i,j,k \text{ distinct}} c_{ijk}$. Then $B_{\mathrm{loc}}(S) \le \sum_i b_i x_i$. Minimize

$$\sum_i (g_i - b_i)x_i + \sum_{i<j} h_{ij} z_{ij} \quad \text{subject to} \quad 0 \le x_i, z_{ij} \le 1, \quad z_{ij} \le x_i, x_j, \quad z_{ij} \ge x_i + x_j - 1. \tag{57}$$

Each binary point with $z_{ij} = x_i x_j$ is feasible. Its objective is at most the actual gain because the subtracted unary penalty upper-bounds the remainder. The minimum in (57) is therefore a lower bound on the worst allowed gain. Extra valid linear restrictions on a deployment family can be included; omitting them gives a conservative full-family bound. There are $n + |E|$ variables and $3|E|$ nontrivial inequalities.

For uncertain means, simultaneously valid intervals give

$$\underline{g}_i = \ell_i - u_0, \qquad \underline{h}_{ij} = \ell_{ij} - u_i - u_j + \ell_0.$$

Replacing each coefficient by this lower endpoint preserves the bound, because x_i, z_{ij} are nonnegative. The intervals may be dependent. They contain the true coefficients coordinatewise on their joint coverage event. Known zero interactions should remain zero rather than acquiring artificial interval noise.

The relaxation is a standard McCormick construction (McCormick, 1976), closely related to roof-duality treatments of binary quadratic optimization (Hammer et al., 1984; Boros et al., 2008). Our contribution here is the sound reduction of a nonlinear activation penalty to that existing machinery. For any inequality multiplier $y \le 0$, the bound $b^{\mathsf{T}} y + \sum_j \min(0, c_j - (A^{\mathsf{T}} y)_j)$ is a dual lower bound on a unit-box LP. We reevaluate it in rational arithmetic and round downward, as in Appendix E.

General signed interactions create a hard target-search problem. For example, negating a weighted cut objective gives a quadratic minimization with positive pair coefficients. By contrast, a quadratic objective with all $h_{ij} \le 0$ is submodular and has a graph-cut formulation. Negative interactions alone are not the source of hardness. The local-game experiments and a separate signed-objective timing suite report the relaxation's cost and gap separately.

L Episode-level gates: a different, implementable deployment contract

An operator can enforce the exposure rather than infer it from arbitrary policy kernels. Let $f(T)$ be the expected return of the team using opaque candidate policies for T and old policies otherwise. For an *enabled* agent i, independently draw $B_i \sim \mathrm{Bernoulli}(\lambda_i)$ once at episode start; run its candidate for the entire episode if $B_i = 1$, and its old policy otherwise. Disabled agents always use their old policies. Candidates and old policies may have arbitrary internal complexity, but remain fixed and can be reset between episodes. Gates are independent of environment randomness and other gates.

The resulting wrapped policy table is

$$F_\lambda(S) = \sum_{T \subseteq S} w_S(T) f(T), \qquad w_S(T) = \prod_{i \in T} \lambda_i \prod_{i \in S \setminus T} (1 - \lambda_i). \tag{58}$$

This is the same function f on every active-set stratum, a stronger assumption than arbitrary $u_{C,T}$. Gate persistence matters: after an informative action reveals a candidate, its conditional per-history distance can be one. The horizon-free result follows from the once-per-episode independent gate, not from incorrectly claiming a per-step distance of λ_i.

Proof of Theorem 6. Subtracting the ungated baseline $f(\varnothing)$ from (58) gives

$$D_\lambda(S) = -(1 - w_S(\varnothing)) f(\varnothing) + \sum_{1 \le |T| \le 2, T \subseteq S} w_S(T) f(T) + \sum_{|T| > 2, T \subseteq S} w_S(T) f(T).$$

All nonbaseline coefficients are nonnegative. Substitute the lower query endpoints and zero for unqueried payoffs to get the lower bound in the theorem; the negative baseline coefficient uses its upper endpoint. Substituting opposite endpoints and one for unknown payoffs gives the upper

bound. Every assignment of the unknown table entries in $[0,1]$ is realizable in a one-step finite cooperative game with deterministic candidate-specific actions. Both endpoints are thus exact for the supplied query intervals. At exact data the width is $\tau_2(S) = \Pr(|S \cap B| > 2)$, giving minimax radius $\tau_2(S)/2$ by the same two-endpoint argument. □

The measured queries are the *ungated* values $f(T)$: an evaluator must be able to force candidate-only singleton and pair runs. This is an explicit gate-bypass testing capability. With exact gated means and nonzero gates, a triangular transform could recover these low-order values, but division by small gate probabilities amplifies sampling error. Our experiments use direct ungated tests instead of that inversion.

L.1 Polynomial family certification for equal gates

Suppose all enabled gates equal λ. For simultaneously valid query intervals let

$$\delta_j = \min_{|T|=j} (\ell_T - u_0), \quad j = 1, 2.$$

Every target of size m has a gain lower bound

$$L_m = \sum_{j=1}^{\min(2,m)} \binom{m}{j} \lambda^j (1-\lambda)^{m-j} \delta_j - \Pr(\mathrm{Bin}(m,\lambda) > 2) u_0. \tag{59}$$

Indeed, the contribution from known size-j subsets is at least their total probability times δ_j, and each unknown subset has gain at least $-u_0$. Checking $L_m \geq -\varepsilon$ for $m = 1, \ldots, n$ certifies the entire family. Candidate values can be heterogeneous: the cardinality simplification uses common externally chosen gates, not symmetry of the underlying task. The cost after computing the queried minima is polynomial and involves no array of size 2^n. Order one follows by replacing two with one throughout.

L.2 An operational example and its scope

Consider a task service containing 64 separately versioned agent components, with versions pinned for each request. Operators independently enable canaries or roll back components across tenants. A staging harness can force one or two candidate services while running all other original services, evaluate a bounded end-to-end metric, and reset the request. An outer router can independently draw the enable gates without inspecting any candidate's action probabilities. There are 2,081 ungated baseline/single/pair tests, but 2^{64} possible enable masks. This is a concrete access design, not a report of a deployed system or a large language model (LLM) benchmark. It remains expensive when each end-to-end evaluation is expensive; setup and episode budgets should both be counted.

The guarantee averages over gate draws separately for each enable mask. It does not certify each realized candidate subset on every request, nor the fully ungated system. This conservative deployment contract is appropriate only when the operator accepts randomized, request-pinned exposure. For a fixed rolling order with no independent rollback, the deployment family is a chain of $n + 1$ configurations, and direct chain evaluation is simpler.

M Exact reward-timed summary LP

Suppose $G = \sum_t \beta_t G_t$ with known nonnegative weights summing to at most one, and prefix exposures $p(t)$. The exact class retaining total means and reward times is

$$\mathcal{A}_{\mathrm{time}} = \sum_t \beta_t \mathcal{A}(p(t)),$$

a Minkowski sum. Necessity follows by applying the prefix coupling to each reward. For sufficiency, use one recording process with independent per-round flags. At time t, its state records the active agents and their old/new tags so far. Assign any specified $u_{t,C,T} \in [0,1]$ as the time-t reward. Every component assignment is jointly realizable, so there are no further restrictions on the retained means.

Monotone general prefix probabilities are realizable with hazards $(p_i(t)-p_i(t-1))/(1-p_i(t-1))$ until exposure reaches one; the constant-distance case gives the required prefix formula directly.

Introduce variables $v_{t,C,T} \in [0, \beta_t w_C(p(t))]$, concatenate each time's incidence matrix, and constrain the *sum* of queried reward means to the observed intervals. Minimizing a target gain gives the exact total-mean evaluation. The model has $H3^n$ variables. Observing separate component means would add constraints and define a richer information model; our comparison does not use them. The rational dual evaluation is the same box-support calculation as Appendix E, with each upper mass replaced by its time-weighted value.

N Budget planning and the role of the target family

For a chosen contrast let Δ_S be its exact expected mean minus its deterministic remainder and the required decision threshold. Suppose a fixed-budget statistical radius has the form $a_S/\sqrt{N}+b_S/N$. If $\Delta_S > 0$, the sufficient expected-margin batch count is

$$N_S \geq \left(\frac{a_S + \sqrt{a_S^2 + 4b_S\Delta_S}}{2\Delta_S} \right)^2 . \tag{60}$$

This follows by solving a quadratic inequality in $\sqrt{N}$. For the known-moment bound, $a_S = \sqrt{2V_S \log(K/\alpha)}$ and $b_S = W_S \log(K/\alpha)/3$. Pairwise cost is $|\mathcal{Q}_2| \max_S N_S$ for complete batches. Equal-batch exhaustive cost is $2^n \max_S N_S^{\mathrm{direct}}$. An optimistic direct allocation uses a distinct required count per target and a reusable baseline bank: its count is $\sum_{S\neq\varnothing} N_S^{\mathrm{direct}} + \max_S N_S^{\mathrm{direct}}$. These are calculations with exact expected margins, not probabilities of successful certification at those budgets and not sample-complexity lower bounds. The oracle allocation is not available without learning the relevant margins.

Our all-target workflow study is often limited by a small singleton margin, whereas a full-team-only calculation need not be. This explains why its crossover differs from a calculation inspecting only the final team. Known finite families replace K by their size. If the family is a fixed chain, direct evaluation uses only $n+1$ queries. Independent enables and rollbacks, tenant-specific component availability, and broad antichains can instead produce combinatorially many allowed masks. Additive configuration-setup cost c changes an evaluation's cost from qN to $qN + cq$; it favors smaller query families but does not remove their extrapolation uncertainty.

O Exact homogeneous minimax threshold

Optimal recovery of a scalar linear functional from linear information over a symmetric convex class has classical linear-optimality theory (Micchelli & Rivlin, 1977; Foucart & Paouris, 2023). Equation (8) specializes that principle to the realizable activation box. The result below instead characterizes exactly when its particular anchored recovery rule is already optimal. All exposures in this section equal p, and the target contains $m \geq 3$ agents. Queries include the baseline and every singleton and pair within the target.

Theorem 13 (Homogeneous threshold, complete statement)**.** *For $0 < p < 1$ and $m \geq 3$, with $q = 1-p$,*

$$R_* = B_2(p,\dots,p) \quad \Longleftrightarrow \quad q^{m-2} \geq \tfrac{1}{2}.$$

At $p = 0$ both radii are zero; at $p = 1$, $R_* = 1/2 < B_2 = (m-1)^2$. The equality statement concerns worst-data absolute recovery error, not the feasible width at every specific dataset.

Proof. Write $x_{C,T} = u_{C,T} - 1/2$, so $|x_{C,T}| \leq 1/2$. The query-null optimal-recovery problem in Theorem 3 maximizes the target functional over this box subject to zero queried means. On this null space it equals the anchored remainder. A maximizing point can reach B_2 only by reaching the support of the remainder on every positive-weight activation stratum.

For $|C| \geq 3$, these support conditions force $x_{C,T} = +1/2$ at $|T| = 1$ and $T = C$, and $x_{C,T} = -1/2$ at $|T| = 0, 2$. The remaining coordinates have zero remainder coefficient. Fix a pair i, j. Its second query difference must vanish. High strata containing this pair contribute

$$-2p^2(1-q^{m-2}).$$

The sole low stratum able to change this difference is $C = \{i, j\}$. Its second difference is at most 2, giving contribution at most $2p^2q^{m-2}$. Cancellation requires $q^{m-2} \geq 1/2$. Compactness ensures a maximizer, so if this condition fails the minimax radius is strictly below B_2.

For sufficiency let

$$\tau = 1 - q^m - mpq^{m-1} - \binom{m}{2}p^2q^{m-2},$$
$$a = p\{1 - q^{m-1} - (m-1)pq^{m-2}\},$$
$$b = p^2(1 - q^{m-2}), \qquad t = (m-1)b - a.$$

These are respectively the mass of high strata, of high strata containing a specified coordinate, and of high strata containing a specified pair. Define

$$x_0 = \frac{\tau - ma + \binom{m}{2}b}{2q^m}, \qquad x_1 = \frac{t}{2pq^{m-1}}, \qquad x_2 = \frac{b}{2p^2q^{m-2}}. \tag{61}$$

Use the forced signs above for high strata, and set their unspecified coordinates to zero. In the low strata, use $x_{\varnothing,\varnothing} = x_0$; for $C = \{i\}$ use $(-x_1, +x_1)$; for $C = \{i, j\}$ use $(+x_2, -x_2, -x_2, +x_2)$ in the order $\varnothing, \{i\}, \{j\}, \{i, j\}$.

We verify the box constraints. First $x_2 \leq 1/2$ is precisely $q^{m-2} \geq 1/2$. That condition also implies $p \leq 1/(m-1)$: for $r = m-2$, the binomial expansion gives $(1 + 1/r)^r \geq 2$, hence $2^{-1/r} \geq r/(r+1)$. Direct simplification gives $t = p\{(m-1)p - 1 + q^{m-1}\}$, and therefore $x_1 \leq 1/2$ follows from $(m-1)p \leq 1$. Also

$$2q^m - 1 + mp - \binom{m}{2}p^2 \geq (m-2)p\{1 - (m+1)p/2\} \geq 0,$$

where the first inequality uses $2q^m \geq q^2$ and the second uses $p \leq 1/(m-1)$. This is equivalent to $x_0 \leq 1/2$. Nonnegativity follows from

$$\tau - ma + \binom{m}{2}b = \mathbb{E}\left[\frac{(M-1)(M-2)}{2}\mathbf{1}\{M \geq 3\}\right] \geq 0$$

and $t = \sum_{j \neq i} \Pr(i, j \in C, |C| \geq 3) - \Pr(i \in C, |C| \geq 3) \geq 0$.

It remains to verify the data constraints, rather than merely the coordinate bounds. For any query Q of size $l \leq 2$, the high-stratum contribution is

$$-\tau/2 + la - 2\binom{l}{2}b.$$

The low-stratum contributions are

$$q^m x_0 + pq^{m-1}(2l - m)x_1 + p^2q^{m-2}\{\binom{m}{2} - 2l(m-l)\}x_2.$$

Substituting (61) cancels the high contribution at $l = 0, 1, 2$. Thus every queried value is $1/2$. The high strata attain the anchored support and the low strata have zero remainder, so the target value is $1/2 + B_2$. Complementing every payoff gives a second admissible world with the same data and target $1/2 - B_2$. These worlds prove the matching minimax lower bound. □

Relation to heterogeneous exposures. Corollary 12 remains a sufficient result for unequal exposures. Its homogeneous specialization is weaker than Theorem 13. For example, at $m = 5, p = .2$, $q^3 = 64/125$ satisfies the exact threshold. At $m = 16, p = .02$, $q^{14} \simeq .75$ and $B_2 \approx 1.60 \times 10^{-2}$; the anchored radius is minimax. At $m = 16, p = .05$, $q^{14} \simeq .49$ and equality fails. These statements concern homogeneous counterparts of the heterogeneous sampled generators.

Independent finite-dimensional check. Permutation symmetry reduces low-stratum corrections to six coordinates indexed by $(|C|, |T|) = (0,0), (1,0), (1,1), (2,0), (2,1), (2,2)$. Their three query constraints form a 3×6 matrix with entries

$$A_{l,(c,j)} = \binom{l}{j}\binom{m-l}{c-j}p^c q^{m-c}, \qquad l = 0, 1, 2.$$

Once high-stratum signs are forced, equality holds exactly when their negative query vector lies in $A[-1/2, 1/2]^6$. Facet normals are obtained from cross-products of pairs of the integer column shapes. The implementation checks all nine distinct facet directions using rational arithmetic. This check is separate from the explicit witness and from the symmetric LP solver; tests compare all three on small cases and check the closed threshold up to 64 agents.

Table 3: Range-aware minimax comparison from rationally bracketed endpoints. The constant predictor already attains radius $1/2$, so the last column isolates any benefit beyond choosing between the anchored and constant predictors.

m	p	B_2	R_*	$\min(B_2, 1/2)$	Ratio
4	0.20	0.12	0.12	0.12	1.00
4	0.30	0.38	0.37	0.38	0.98
4	0.40	0.84	0.50	0.50	1.00
5	0.20	0.27	0.27	0.27	1.00
5	0.30	0.82	0.50	0.50	1.00
5	0.40	1.78	0.50	0.50	1.00
6	0.15	0.22	0.22	0.22	1.00
6	0.20	0.49	0.42	0.49	0.85
6	0.30	1.46	0.50	0.50	1.00
6	0.40	3.04	0.50	0.50	1.00

Range-aware comparison. A constant predictor $1/2$ has worst-case error $1/2$. Selecting between it and the anchored predictor before seeing data gives the benchmark $R_{\text{simple}} = \min\{B_2, 1/2\}$. Table 3 reports R_*/R_{simple} as well as the untruncated comparison. An improvement from $B_2 > 1/2$ to $R_* = 1/2$ needs no queried information and is not evidence for optimized recovery. The equality theorem is the useful small-exposure conclusion; numerical gains outside its region can occur at large absolute uncertainty.

P Conditional validity of pilot allocation and a tighter target relaxation

Proposition 14 (Independent holdout allocation)**.** *An evaluator may use a pilot sample to choose a count $N_S \geq 2$ for each direct target contrast. Conditional on that pilot, generate independent holdout rows, using correct-marginal common random numbers across targets as desired. Apply empirical Bernstein to each target with error budget α/K, discarding all pilot observations from the final means and variances. The resulting simultaneous direct lower bounds have coverage at least $1 - \alpha$, even though N_S depends on the pilot.*

Proof. Conditional on the pilot, counts are fixed and each contrast has i.i.d. bounded holdout rows. The empirical-Bernstein inequality and a union bound give conditional failure probability at most α. Correlation between different target columns is unrestricted. Averaging over the pilot proves the unconditional assertion. Reusing pilot outcomes inside the final confidence interval, or stopping its holdout adaptively without a time-uniform correction, would not be covered by this argument. □

The implemented allocation pools pilot difficulty by target size to avoid fitting $2^n - 1$ noisy parameters. It charges the pilot, every final target episode, and a fresh baseline bank for each size group. Its sample-allocation formula is a heuristic for power; only the holdout interval establishes validity. This distinction makes the comparison stronger without attributing an optimality guarantee to the allocation heuristic.

Proposition 15 (Separate small targets before relaxation)**.** *For a pairwise objective $D_2(x)$ and valid unary upper bound $b^T x$ on its nonnegative remainder, let $v_{\geq 3}$ be any valid lower bound on*

$$\min\{D_2(x) - b^T x : (x, z) \text{ obey the McCormick constraints}, \sum_i x_i \geq 3\}.$$

If $v_{\leq 2}$ is the minimum of the gains for $|S| \leq 2$ obtained from direct values or structurally verified zero interactions, then $\min\{v_{\leq 2}, v_{\geq 3}\}$ is a valid family-wide lower bound. When the relaxation

optima are computed exactly, this value is at least as large as the analogous optimum without the cardinality cut.

Proof. The anchored remainder is zero for targets of size at most two. Each larger binary target is feasible in the cut relaxation and its true gain is bounded below by the penalized objective. Taking the minimum proves validity. Restricting the relaxation raises its optimum, while removing a non-negative penalty from small targets raises their values; both exact branches dominate the original unrestricted relaxation's optimum. Arbitrary approximate lower bounds need not preserve that comparison. Taking the maximum of the split certificate and a previously verified unrestricted certificate preserves monotonic improvement when numerical lower bounds are used. For inexact query values, use simultaneous signed-coefficient lower bounds in the same two branches. □

The implementation includes the cut in the rational dual lower-bound check. Its additional constraint is one sparse row. The number of small-target evaluations is already covered by the pairwise query design; the split requires no new environment evaluations.