WHICH ENVIRONMENTS TEACH COOPERATION? GUARANTEES, FEEDBACK, AND COMPETENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Choosing environments for independently trained agents involves two different questions: are their near-optimal self-play solutions compatible, and can the learners reach those solutions? We study this distinction while requiring competence on a fixed task distribution. For sequential protocols with shared preferences and action-independent paths, an exact balanced partition of diagnostic losses characterizes incompatible near-optimal pairs. Conflicting one-step sources instead induce signed exposures; conflicting sequential sources need not admit the same representation. We then show that observable failure locations lead to a polynomial singleton-exposure objective for a correction learner, even when a terminal-reward estimator requires much more data. Experiments separate solution-set guarantees from learned performance. On invariant tasks with delayed rewards, training on the competence distribution is a strong default. In small action-steering tasks, controls with matched target information reveal that the advantage of margin selection depends on the competence floor and nominal tie-breaking rule. Additional mixtures selected independently of the margin show a positive association within positive-margin regimes. The resulting framework gives conditional certificates and identifies which environmental, feedback, and competence assumptions are needed to interpret them; it does not make a large margin a universal curriculum objective.

1 INTRODUCTION

Two teams can each perform well in self-play and still disagree when their members are mixed. One team may assign a handoff to the nearest agent, another to a fixed role. Zero-shot coordination studies such mismatches without joint adaptation at deployment (Hu et al., 2020; Treutlein et al., 2021). Training across environments can expose conventions that work only in a narrow setting (Jha et al., 2025; You et al., 2025). The practical question is not simply how many environments to generate, but which distinctions their rewards teach.

Consider two training tasks with self-play scores $(1, .8, 1)$ and $(1, 1, .8)$ for three conventions. Equal weighting makes only the first convention .02-optimal. Adding eighteen neutral tasks and sampling uniformly makes both alternatives score .99, so incompatible conventions can again be near-optimal. The informative tasks have not disappeared; their training weight has been diluted. We call the reward separation supplied by weighted tasks *diagnostic exposure*. Adding optional environments cannot hurt the best attainable curriculum, but assigning them weight can reduce the information present at a fixed learning tolerance.

Three qualifications change how this example should guide practice. First, environments can prefer opposite conventions rather than merely distinguish them with different strength. Their effects can cancel. Second, compatibility alone can be cheap: installing any shared convention may ensure agreement while sacrificing task competence. Third, a learner may already receive a direct explanation of each failure. A curriculum optimized for a terminal-reward estimator need not help a learner that uses those explanations.

The natural default is therefore *train on the task distribution on which competence is judged*, denoted ν. We compare against this default, not only against a deliberately diluted distribution. A proposed teacher must explain what additional benefit comes from target knowledge, a larger near-optimal-

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.

set margin, or a learning-specific information objective. These choices have different effects. In particular, the existence of a competent optimum under ν does not ensure that different training runs choose compatible optima or reach them.

We formalize the first issue using a *compatibility margin*: the smallest self-play tolerance that admits a harmful pair. This is a reward-based form of controlling a surviving set of hypotheses, not a new general teaching principle (Tosh & Dasgupta, 2017; Zhu et al., 2018). The contribution is a structural calculation of that margin and an analysis of where it ceases to describe learning difficulty. In an invariant sequential protocol, a harmful disagreement can be divided between two learners, each paying only for its own mistakes. The exact path costs determine whether both shares fit inside their learning-error budgets. A competence constraint adds mistakes that cannot be left unresolved even if they rarely cause target disagreement.

Results and scope. The path identity assumes common observable labels, a shared preferred convention, and matched actions that cannot steer the path. We state these restrictions before using the theorem. Signed exposures handle conflicting one-step preferences, while a counterexample excludes a direct extension to sequential conflict. Exact evaluation already contains PARTITION; tractable sign-cell and approximate-cover results require additional restrictions. For a learner observing failure locations, we derive a rate governed by singleton exposure rather than the balanced margin. Finally, experiments compare competence thresholds and teachers with the same target access. They support the usefulness of the margin as a solution-set diagnostic, but also show that simpler learning objectives or target-informed controls can match or exceed margin-based teaching. All experiments use synthetic tasks; none establishes generalization to large policy classes.

2 Related work

Ad hoc teamwork and the Hanabi benchmark make convention dependence explicit (Stone et al., 2010; Bard et al., 2020). Other-Play, Off-Belief Learning, Fictitious Co-Play (FCP), trajectory diversity, Learning Incompatible Policies (LIPO), adversarial diversity, maximum-entropy populations, and diverse-convention methods modify objectives or partner populations (Hu et al., 2020; 2021; Strouse et al., 2021; Lupu et al., 2021; Charakorn et al., 2023; Cui et al., 2023; Zhao et al., 2023; Sarkar et al., 2023). Incompatibility graphs and human-model-based training address related partner mismatches (Li et al., 2024; Carroll et al., 2019). Our teacher chooses source frequencies for future independent learners; it does not introduce the bad-pair relation or claim that solution-set control is new.

Cross-Environment Cooperation, automatic coordination curricula, and multi-agent environment design already study environmental variation (Jha et al., 2025; You et al., 2025; Samvelyan et al., 2023). They connect to reverse curricula, level replay, Protagonist Antagonist Induced Regret Environment Design (PAIRED), Adversarially Compounding Complexity by Editing Levels (ACCEL), and procedural generalization (Florensa et al., 2017; Jiang et al., 2021; Dennis et al., 2020; Parker-Holder et al., 2022; Cobbe et al., 2020; Narvekar et al., 2020). Our analysis does not replace those algorithms. We isolate when environmental rewards exclude harmful solutions and use information-matched controls to avoid attributing representative target access to a particular objective.

Minimal-knowledge coordination asks which distinctions matter between collaborators (Lauffer et al., 2023). Machine teaching and diameter-based active learning likewise precede our exclusion viewpoint (Goldman & Kearns, 1995; Zhu, 2015; Brown & Niekum, 2019; Bharti et al., 2024; Tosh & Dasgupta, 2017). Our calculations concern two learners' separate reward gaps, not merely coverage of observations. Submodular cover, multiplicative weights, Boolean moments, arrangements, and statistical intervals are established tools (Iyer & Bilmes, 2013; Arora et al., 2012; O'Donnell, 2014; Edelsbrunner et al., 1986; Hoeffding, 1963; Clopper & Pearson, 1934). Short reductions and concentration arguments are credited as such. The experiments use sampled policy gradients but make no global convergence assumption (Williams, 1992; Agarwal et al., 2021a); independent libraries, rather than dependent cross-play pairs, are the unit of uncertainty estimation (Agarwal et al., 2021b).

3 What the teacher must guarantee

Let $\Pi = \{\pi_1, \ldots, \pi_K\}$ be a specified finite class of role-conditioned policies. Two role copies of one policy define self-play; complementary roles need not take identical physical actions. For M sources

and a stationary training distribution $\mu \in \Delta_M$, write

$$A_{ei} = J_e(\pi_i, \pi_i), \qquad S_i(\mu) = \sum_e \mu_e A_{ei}, \qquad g_i(\mu) = \max_k S_k(\mu) - S_i(\mu). \tag{1}$$

Returns are in $[0, 1]$. A globally η-accurate learner may output any member of $\mathcal{A}_\eta(\mu) = \{i : g_i(\mu) \leq \eta\}$. This describes achieved solution quality, not the behavior of every optimizer. Let T_{ij} be target cross-play for the ordered role assignment (i, j).

Definition 1 (Margin of a bad-pair relation)**.** For any relation $\mathcal{B} \subseteq [K]^2$, including self-loops when needed, define

$$\Gamma_A(\mu; \mathcal{B}) = \min_{(i,j)\in\mathcal{B}} \max\{g_i(\mu), g_j(\mu)\}, \qquad \Gamma_A(\mu; \varnothing) = +\infty. \tag{2}$$

Every pair in $\mathcal{A}_\eta(\mu)$ avoids $\mathcal{B}$ exactly when $\Gamma_A > \eta$. At least one member of each bad pair must be excluded. This is an immediate organizing lemma, not the main theorem. Strictness matters: a gap equal to η is still allowed. For example, a pair with gaps .03 and .08 cannot both survive tolerance .05; excluding one member suffices, and a unique surviving convention is unnecessary.

Competence and cooperation use common thresholds. Fix a common source evaluation law ν, competence floor c, and target threshold τ. Define

$$V_i = \sum_e \nu_e A_{ei}, \qquad \mathcal{B}_{\tau,c} = \{(i, j) : T_{ij} < \tau \text{ or } V_i < c \text{ or } V_j < c\}. \tag{3}$$

The teacher's robust objective is $\max_\mu \Gamma_A(\mu; \mathcal{B}_{\tau,c})$. The measured outcome is the probability that two independent training runs meet *all three* requirements. It is not determined by the two marginal average rewards. We report those averages, gaps, and infeasibility as diagnostics. Throughout, "train on ν" means sampling the competence tasks themselves; uniform sampling over an enlarged library is a distinct baseline.

A margin and a learning distribution answer different questions. Maximizing Γ enlarges the tolerance to adversarially chosen near-optimal solutions. It need not maximize success under the actual distribution of optimizer outputs. A teacher that selects the target-best source optimum uses target information but ignores suboptimal solutions. Comparing it with a margin teacher holds that information fixed and tests whether considering those solutions adds value. Likewise, a sampled evaluation concerns the output distribution of a frozen learning procedure, not every member of Π.

What the teacher and learner observe. The exact teacher receives source and target return information; a rollout-only teacher instead trains pilots and evaluates them on representative tasks. Acquisition costs are additional to final learning. Learners do not receive the exact tables or a preferred convention. We distinguish (i) delayed verification: only an episode-level convention and final binary reward, and (ii) trace feedback: the first failed node is visible. Private-action policy gradient is a third exploration mechanism, even when termination is visible. A shared convention installed without training is a valid control, but the teacher cannot install a *known competent* policy in the learning experiments. Claims outside a specified policy class, target family, or feedback interface require additional assumptions; Appendix B gives the precise correction when restricting the class lowers its optimum.

4 Diagnostic exposure: exact identities and their boundaries

4.1 Invariant sequential paths

A convention $b \in \{0, 1\}^d$ assigns complementary roles at observable nodes. All sources share a preferred vector θ. Each source samples an exogenous path ω of at most H visits; matched assignments do not steer it. Conditional on the path, correct matching assignments survive, wrong ones fail with specified independent hazard coins, and a mismatch fails. The target has no matching-assignment hazards, so its return isolates disagreement. Policies observe the node and role, not the source identity, and cannot choose a convention based on history or time.

This model includes node-level Markov paths $(\alpha_e, P_e, \Delta_{ej})$ and explicit path banks. The former admit a dynamic program; the latter use weighted sums. These representations are not interchangeable computationally. Let $r_{ej}(\omega)$ be the product of survival probabilities on all visits to j, and define

$$F_e(E) = 1 - \mathbb{E}_{\omega\sim e} \prod_{j\in E} r_{ej}(\omega), \quad F_\mu = \sum_e \mu_e F_e, \quad h_q(U) = \mathbb{P}_{\omega\sim q}(\omega \cap U \neq \varnothing). \tag{4}$$

For $E_b = \{j : b_j \neq \theta_j\}$, the source gap is $F_\mu(E_b)$ and target loss is $h_q(E_b \triangle E_{b'})$. Each F_e is normalized, monotone, and submodular: a further mistake cannot increase survival, and it matters less after other mistakes already cause failure. Two certain visits with hazard $1/2$ give loss $3/4$, not an additive loss of 1.

Theorem 2 (Balanced path exposure)**.** *For the full convention class and $\tau = 1 - \varepsilon$, $0 \le \varepsilon < 1$, the coordination-only margin is*

$$\Gamma(\mu) = \min_{U:h_q(U)>\varepsilon} \min_{V\subseteq U} \max\{F_\mu(V), F_\mu(U \setminus V)\}. \tag{5}$$

With $\kappa(\mu) = \min_{U:h_q(U)>\varepsilon} F_\mu(U)$, one has $\kappa/2 \le \Gamma \le \kappa$.

A harmful disagreement U is divided between two learners: one pays for errors in V, the other for errors in $U \setminus V$. Their common errors can be removed without changing disagreement or increasing either gap. Conversely every disjoint partition is realized by two policies. This short argument uses the error-cost representation, not a general property of cooperative Markov games. Appendix G gives the full proof.

Corollary 3 (Coverage and dilution)**.** *Let $Z = \{j : F_e(\{j\}) = 0 \text{ for all } e\}$. A curriculum with positive margin exists exactly when $h_q(Z) \le \varepsilon$. Adding nondiagnostic modes while retaining diagnostic probability p multiplies the margin by p.*

This is a coverage result, not source-conflict resolution. It permits harmless ambiguity but does not give a uniform positive margin: visible distinctions can be arbitrarily weak. Maximizing κ yields the path linear program (LP)

$$\max_{\mu\in\Delta_M, t} t \quad \text{subject to } F_\mu(U) \ge t \quad \text{for every } U \text{ with } h_q(U) > \varepsilon. \tag{6}$$

Its output has at least half the optimal margin, and the factor is tight already with two nodes and $H = 1$. Source vectors $(1/2, 1/2)$ and $(1 - h', 0)$ give ratio $1/[2(1 - h')] \to 1/2$ for a target where only the full set is harmful. Exact path information can outperform occupancy: a separate correlation family gives margin .51 for the occupancy-selected source, versus $1 - \binom{n}{n/2}^{-1}$ for path selection, approaching 1. Both constructions are in Appendix C.

4.2 Adding competence; allowing conflicts

For invariant paths, put $\beta = 1 - c$ and $\ell(\mu) = \min_{E:F_\nu(E)>\beta} F_\mu(E)$, with an empty minimum equal to $+\infty$. Then

$$\Gamma_{\text{joint}}(\mu) = \min\{\Gamma(\mu), \ell(\mu)\}. \tag{7}$$

The joint LP imposes $F_\mu(U) \ge 2t$ for harmful disagreements and $F_\mu(E) \ge t$ for incompetent error sets. Its certificate t is at least half the best joint margin. At a sufficiently strict competence floor, however, every error is disallowed: if $c > 1 - \min_j F_\nu(\{j\})$, both learners must recover θ exactly. This explains why a coordination advantage can vanish when competence is required (Appendix A.3).

One-step sources can instead prefer distinct vectors θ_e. Write $w_{ej} = \mathbb{P}_e(j)\Delta_{ej}$.

Theorem 4 (Signed conflict geometry)**.** *For arbitrary source-specific one-step preferences, set*

$$s_j(\mu) = \sum_e \mu_e w_{ej}(2\theta_{ej} - 1), \qquad \theta_{\mu j} = \mathbf{1}\{s_j(\mu) > 0\}. \tag{8}$$

The gap is $g_\mu(b) = \sum_j |s_j(\mu)|\mathbf{1}\{b_j \neq \theta_{\mu j}\}$, with arbitrary zero-score ties. The balanced formula holds with this modular cost and target disagreement $q(U)$.

Opposing exposures can cancel, and the optimal reference changes with μ. For any fixed number M of sources, sign-cell enumeration gives $O(d^{M-1})$ optimization LPs *if every singleton is harmful*. This condition requires $\varepsilon < 1/d$ and does not establish tractability at arbitrary fixed tolerance. Without it, even evaluating Γ for one source at $H = 1$ contains PARTITION (Karp, 1972). Appendix A.4 proves the extension and distinguishes final optimization LPs from additional cell-enumeration work.

The sequential shared-reference argument fails under conflict. On a sure three-node path with fatal wrong assignments, three sources prefer $000, 101, 011$. Equal training makes these vectors optimal. If the target visits node 0 or 1 uniformly, 101 and 011 form a bad optimal pair at $\varepsilon = .6$, so $\Gamma = 0$. Removing their common last-coordinate error raises both gaps to $1/3$; the disjoint formula is incorrect. The action-steering experiment below therefore uses general finite return tables, not this path identity or its perturbative extension.

5 Which learning objective follows from the feedback?

Delayed verification. Uniform episode-level convention probes estimate each preferred bit from the sign of $(2B_j - 1)R$. Define

$$m_{ej} = \mathbb{E}_{E \subseteq [d]\setminus\{j\} \text{ uniform}}[F_e(E \cup \{j\}) - F_e(E)], \qquad m_{\mu j} = \sum_e \mu_e m_{ej}. \tag{9}$$

For a common relevant set C, two learners' sign-error probability is at most $2\sum_{j\in C} \exp(-N m_{\mu j}^2/8)$. Leaving $U = C^c$ unresolved is sufficient if both $h_q(U) \le \varepsilon$ and $F_\nu(U) \le 1 - c$. Probe cover maximizes the minimum signal over such a set; probe all takes $C = [d]$. These are estimator-specific objectives, not margin maximization. Cover selection remains exponential in general (Appendix H).

Observable failures. A failure at a matched node certifies that its bit is wrong: preferred matched assignments never fail. Flip that bit and never undo a correct one. On *branching and revisiting paths*, while any relevant bit remains wrong, the failure probability is at least $s_C = \min_{j\in C} F_\mu(\{j\})$. At most d corrections occur, including any outside C. Hence

$$\mathbb{E}T_C \le d/s_C, \qquad \mathbb{P}(T_C > N) \le \mathbb{P}\{\mathrm{Bin}(N, s_C) < d\}. \tag{10}$$

A coupling to Bernoulli successes proves the bounds (Appendix A.2). The corresponding max–min singleton LP has only $|C|$ exposure constraints. It optimizes this upper bound, not necessarily mean correction time; the nonadditive path margin still characterizes worst-case near-optimal pairs but is not needed for this learner's rate.

Theorem 5 (Same margin, different identification complexity)**.** *On a deterministic d-node path with unknown θ and fatal wrong assignments, $\Gamma = 1$. With delayed binary verification, any adaptive learner making N queries succeeds in producing an η-optimal convention, $\eta < 1$, with probability at most $(N+1)/2^d$, averaged over uniform θ. With failure locations, at most d episodes suffice for exact correction. If wrong-assignment hazards are at least $c_0 > 0$, the expected correction cost is at most d/c_0.*

The lower bound is the standard equality-query counting argument, not a new information-theoretic method. It concerns competence identification; shared agreement alone does not require identifying θ. The feedback model, rather than the margin, creates the exponential separation. Ordinary private-action policy gradient receives different exploration feedback from episode-level shared probes or matched-bit correction. We test it separately instead of assuming it reaches the near-optimal set.

6 Experiments: defaults and information-matched controls

All outcomes below come from trained policies and exact evaluation on small tasks, except the stated sampled evaluation. We write cross-play as XP and self-play as SP. Library-bootstrap intervals use 10,000 resamples of independent task libraries, not dependent seed pairs. They are pointwise and descriptive, without a multiple-comparison correction.

Table 1: Delayed probes, 12 invariant libraries, 4,096 episodes. Joint success requires both common-source returns $\geq .98$ and target XP $\geq .8$. Entries include 95% library-bootstrap intervals. Competence cover equals probe all at this design floor.

Curriculum	Target XP	Common SP	Joint (%)
Train on ν	0.81 [0.77, 0.86]	0.99 [0.99, 0.99]	73.2 [69.0, 76.9]
Inverse singleton	0.84 [0.80, 0.88]	0.99 [0.99, 0.99]	76.8 [74.3, 79.3]
Path coverage	0.88 [0.85, 0.90]	0.96 [0.95, 0.96]	21.5 [16.3, 26.9]
Probe cover	0.88 [0.86, 0.90]	0.96 [0.95, 0.96]	21.0 [16.2, 25.8]
Competence cover / all	0.86 [0.82, 0.90]	0.99 [0.99, 0.99]	80.7 [76.2, 85.1]
Joint-margin LP	0.90 [0.87, 0.93]	0.98 [0.98, 0.98]	55.9 [50.5, 61.9]

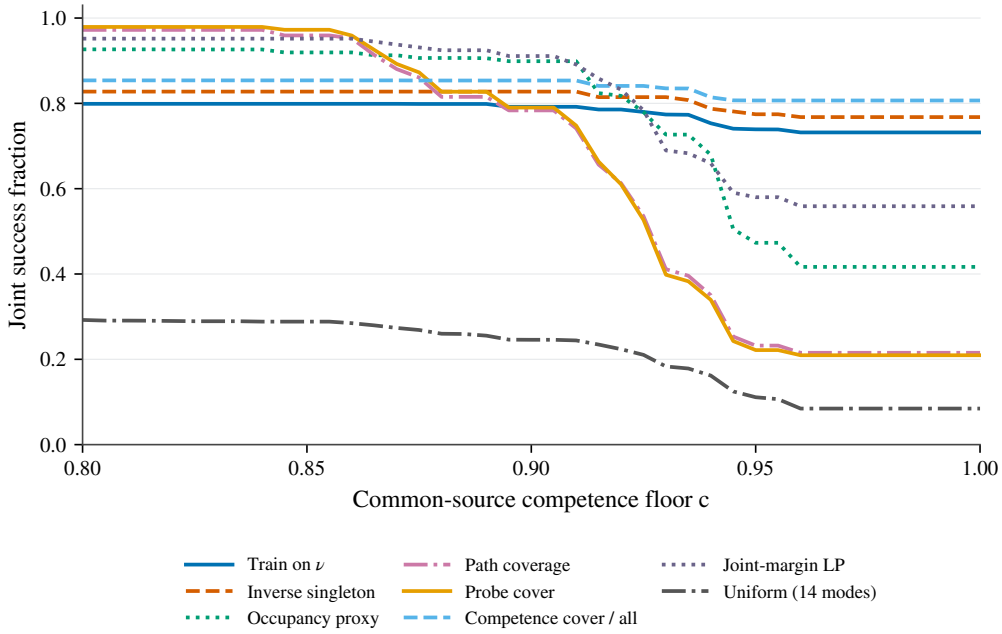


Figure 1: Joint success versus the competence floor, with target threshold .8. Eight curricula use fixed trained policies; the joint LP and competence cover were designed at .98, not redesigned along this curve. The grid spacing is $1/200$.

6.1 Invariant protocols under a common competence floor

Twelve six-node libraries have six diagnostic modes and eight neutral modes. Diagnostic source e starts mostly at node e and increases its wrong-assignment hazard. Paths branch and revisit nodes for six decisions. The target visits a low-probability branch with probability .12–.18; that branch can be ignored for coordination at $\tau = .8$, but not necessarily for competence. Here ν is uniform over the six diagnostic sources. Each delayed-probe method uses 24 seeds and 4,096 episodes. This suite measures the delayed interface; it does not claim that ordinary trace learning needs that budget.

Training on ν and inverse-singleton weighting outperform the joint-margin LP at $c = .98$ (Table 1). The joint LP minus inverse singleton changes joint success by -20.89 percentage points, interval $[-26.27, -15.10]$. Probe all minus inverse singleton is $+3.89$ points, $[-.91, 7.79]$, unresolved; probe all minus train-on-ν is $+7.49$ points, $[.60, 13.68]$. Probe all is a signal objective, not evidence that maximizing the margin wins. Coordination-only bad-pair improvements and conditional eligible-pair checks cannot substitute for this joint comparison.

Figure 1 shows results across competence floors. Above 95.60%, the competent set is exactly $\{\theta\}$ in every library, explaining the plateau through $c = 1$. The probe-cover ignored branch becomes competence-allowable at library-specific floors of approximately 84.18%–89.47%, not at a universal

Table 2: Action-steering task at $c = .60$, $\tau = .8$, all 16 libraries. Information is the teacher's additional final-library access; all actors share the same tuning protocol. Joint columns have 95% library-bootstrap intervals. Exact grid teachers share candidates and training budgets. FCP and pilot selection have additional costs.

Method	Teacher information	XP	SP	Joint (%)
Train on ν	ν only	0.32	0.64	16.8 [6.3, 29.0]
Small FCP (3× training)	ν only	0.48	0.67	31.1 [12.0, 54.5]
Nominal, near-ν ties	Exact tables	0.41	0.64	28.9 [15.8, 43.5]
Nominal, easy-source ties	Exact tables	0.57	0.65	43.1 [26.0, 59.8]
Joint-margin grid	Exact tables	0.52	0.65	43.2 [23.3, 63.6]
Frozen pilot	Target rollouts	0.83	0.62	43.7 [24.5, 62.8]
Aligned-source oracle	Source identity	0.86	0.62	47.4 [26.6, 67.8]

threshold. The plotted curves use the same trained policies at every point. Source-only and target-only occupancy ablations show that this learned suite does not resolve a benefit from exact source costs alone. Twelve randomized-target libraries and twelve complementary-support libraries provide additional checks: at 4,096 episodes, competence cover minus train-on-ν improves joint success by 6.97 points $[1.75, 12.23]$ in the first family and ties it in the second. Path coverage does not outperform the default there. The comparisons include ties at the larger training budget.

6.2 Conflicting sources with action-dependent paths

A depth-three binary tree has seven observable nodes. Matched binary assignments choose left or right successors, and wrong matching assignments may fail; mismatches fail. Four training sources prefer conflicting vectors, including one target-aligned and one fully opposed source. Their labels are independently permuted. The target also penalizes wrong matching assignments and starts with support on every node, so fixed agreement need not solve the task. The baseline ν is uniform over these four sources: a *multi-task competence requirement*, not a claim that this is the only reasonable deployment objective. We test two alternatives below. The no-training fixed-convention control installs seven zero bits in every actor; it has baseline target XP 0.42, common SP 0.52, and zero joint success at $.60$.

Each actor takes a one-hot node, an eight-unit hyperbolic tangent (tanh) layer, and a Bernoulli logit with a node-specific bias. It is a reparameterized tabular policy. Private actions are sampled during training; evaluation rounds the seven probabilities, yielding exactly the 128 deterministic conventions enumerated by the oracle. No observation transfer or nonenumerable-class validation is claimed. A shared 18-setting tuning sweep uses four separate libraries under uniform training, with target-based validation. Final results use 16 disjoint libraries and 24 seeds, each with 16,384 episodes. "No target" teacher controls therefore mean no *additional final-library* target information beyond that common tuning.

Equal access before comparing objectives. The exact teachers see the same source and target tables and choose among the same 35 quarter-grid mixtures. The nominal control selects the curriculum whose first-index source optimum has highest target self-play, subject to common competence; equal scores prefer weights nearest ν. An easy-source variant first breaks equal target scores by the largest optimal training return, without using suboptimal gaps. An optimal-set control evaluates only source maximizers. Margin selection instead considers all harmful pairs. These methods are reselected at *each* evaluation floor. A frozen pilot uses extra source/target rollouts to choose among uniform and four pure sources. An aligned-source oracle supplies the identity of the target-aligned source, but not its convention bits to the learner. Small FCP and a value-error replay heuristic remain additional baselines, not full reproductions of large-benchmark methods.

At $.60$, margin selection exceeds the near-ν nominal control by 14.29 points $[3.19, 27.45]$, but exceeds its easy-source tie variant by only $.14$ points $[-7.65, 7.31]$. Its difference from the aligned-source oracle is -4.23 points $[-14.72, 3.89]$. Thus an apparent advantage over one target-informed control is not robust to a reasonable alternative tie rule. Relative to train-on-ν, nominal selection itself contributes 12.09 points $[5.19, 19.75]$. The full margin-versus-default difference cannot be attributed solely to the margin objective.

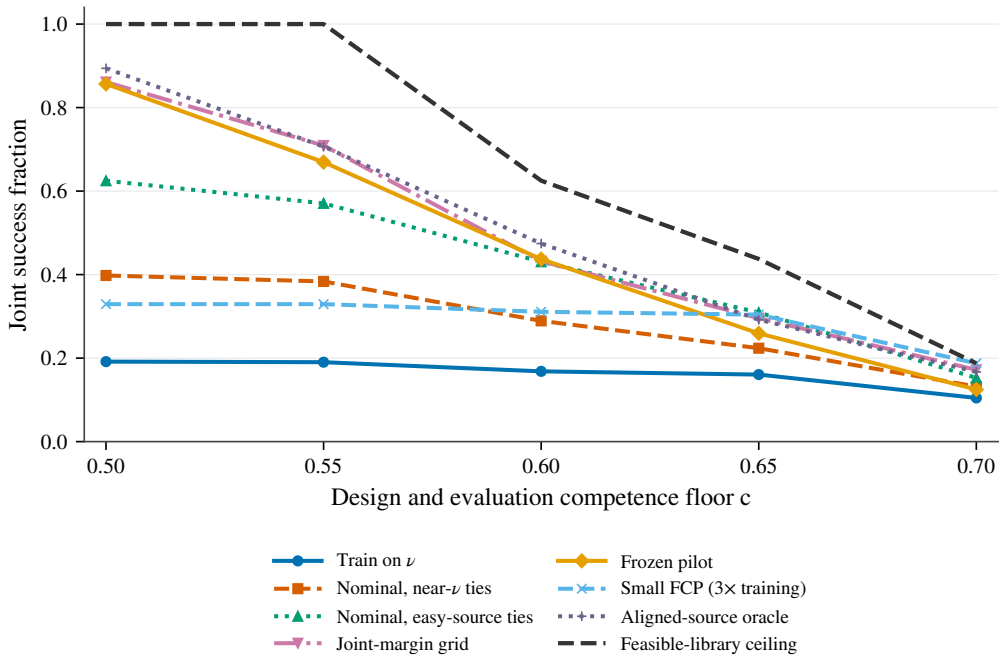


Figure 2: Matched-floor design and evaluation, uniform four-source ν. The margin and nominal teachers are reselected at each of the five marked floors; pilot and FCP stay fixed. The feasibility ceiling is the fraction of libraries admitting any acceptable deterministic pair. Lines connect evaluated points, not continuous guarantees.

The competence-floor sweep changes the comparison (Figure 2). At .50, redesigning the margin teacher raises joint success to 0.86, close to the frozen pilot's 0.86; using a teacher fixed at .60 understates that low-floor performance. The margin improves over the easy-source nominal control by 23.64 points $[9.58, 38.41]$ there, but not over the aligned-source oracle. At .70, only 3 of 16 libraries are feasible: FCP's 18.75% reaches the aggregate feasibility ceiling, rather than indicating a broadly unsolved feasible task. The .60 sensitivity is exploratory and pilot-informed, not preregistered. The analysis includes all five floors, including .70, and all libraries.

Changing ν also changes the question. Excluding the anti-target source makes all libraries feasible through .65; at .60, joint success is 0.86 for the margin, 0.64 for nominal-easy, and 0.37 for training on this new ν. With ν equal to the aligned source, simply training on ν reaches 0.89, versus 0.91 for margin selection; the small paired difference is not a general design advantage. Frozen pilot, FCP, and replay policies are only re-evaluated under alternative ν, whereas the exact teachers are reselected and the train-on-ν controls are retrained. At the baseline floor, the pilot selects a pure source in all 16 libraries, the aligned source in 12; the margin selects it in three.

Does numerical margin predict more than binary feasibility? For each library we additionally draw 24 Dirichlet mixtures independently of the margin and train the same 24 actor seeds per mixture. At $c = .60$, within-library Spearman correlation is 0.86 $[0.77, 0.92]$ for the margin, versus 0.79 $[0.72, 0.85]$ for the binary indicator that the nominal optimum is jointly acceptable, on ten nonconstant libraries. Restricting to positive-margin mixtures leaves nine informative libraries and correlation 0.64 $[0.36, 0.84]$; the binary predictor is then constant. These are exploratory associations on a small finite task, not calibrated predictions or proof of a causal margin advantage. Constant cases are excluded from correlation summaries rather than assigned zero correlation.

What can be claimed from sampled evaluation? A separate evaluation checks twelve disjoint trained pairs per method and library. At $c = .60$, it certifies 32/192 pairs for train-on-ν, 50 for nominal selection, 76 for margin selection, and 74 for the frozen pilot. Population lower bounds are nonzero in 7, 9, 9, and 9 libraries, respectively; even perfect counts yield a bound of only about

0.74 at this small population sample. Appendix F gives the population-evaluation procedure and its confidence guarantee. This does not validate nonenumerable policies: the evaluation formula requires no enumeration, but these evaluated actors remain a finite class.

The comparison also separates teaching from acquisition. The exact selectors use privileged expectation tables; the pilot spends 450,560 training-plus-evaluation episodes per library before final learners are trained. FCP uses three times the final self-play training budget including partner generation. The matched-grid control comparison holds model access and final learning fixed; comparisons with pilot or FCP do not establish an end-to-end sample saving. These distinctions are necessary before turning a solution-set guarantee into a practical efficiency claim.

7 Conclusion

The balanced margin certifies which near-optimal solutions are compatible, but does not by itself identify the best curriculum. Observable failures favor a polynomial singleton objective; source conflict can invalidate the shared-error model; and target-informed controls can explain gains over train-on-ν. The appendices give computational bounds, population evaluations, and teacher-acquisition costs. Our separator applies only to a special target law, and population guarantees do not cover entire parametric policy classes. The experiments are small and synthetic, with finite classes after rounding. These results support conditional guarantees, not universal curriculum superiority or end-to-end efficiency.

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

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021a. URL https://jmlr.org/papers/v22/19-736.html.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320, 2021b. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012. doi: 10.4086/toc.2012.v008a006.

Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling. The Hanabi challenge: A new frontier for AI research. *Artificial Intelligence*, 280:103216, 2020. doi: 10.1016/j.artint.2019.103216.

Shubham Bharti, Stephen Wright, Adish Singla, and Jerry Zhu. On the complexity of teaching a family of linear behavior cloning learners. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103945–103965, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bc46f5d72f432de4c58e1a311bad011b-Abstract-Conference.html.

Daniel S. Brown and Scott Niekum. Machine teaching for inverse reinforcement learning: Algorithms and applications. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 7749–7758, 2019. doi: 10.1609/aaai.v33i01.33017749.

Micah Carroll, Rohin Shah, Mark K. Ho, Thomas L. Griffiths, Sanjit A. Seshia, Pieter Abbeel, and Anca 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.

Rujikorn Charakorn, Poramate Manoonpong, and Nat Dilokthanakul. Generating diverse cooperative agents by learning incompatible policies. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=UkU05GOH7_6.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Karl Cobbe, Christopher Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2048–2056, 2020. URL https://proceedings.mlr.press/v119/cobbe20a.html.

Brandon Cui, Andrei Lupu, Samuel Sokota, Hengyuan Hu, David J. Wu, and Jakob N. Foerster. Adversarial diversity in Hanabi. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=uLE3WF3-H_5.

Michael Dennis, Natasha Jaques, Eugene Vinitsky, Alexandre Bayen, Stuart Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13049–13061, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/985e9a46e10005356bbaf194249f6856-Abstract.html.

Herbert Edelsbrunner, Joseph O'Rourke, and Raimund Seidel. Constructing arrangements of lines and hyperplanes with applications. *SIAM Journal on Computing*, 15(2):341–363, 1986. doi: 10.1137/0215024.

Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 482–495, 2017. URL https://proceedings.mlr.press/v78/florensa17a.html.

Sally A. Goldman and Michael J. Kearns. On the complexity of teaching. *Journal of Computer and System Sciences*, 50(1):20–31, 1995. doi: 10.1006/jcss.1995.1003.

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.

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

Hengyuan Hu, Adam Lerer, Brandon Cui, Luis Pineda, Noam Brown, and Jakob Foerster. Off-belief learning. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4369–4379, 2021. URL https://proceedings.mlr.press/v139/hu21c.html.

Rishabh K. Iyer and Jeff A. Bilmes. Submodular optimization with submodular cover and submodular knapsack constraints. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/a1d50185e7426cbb0acad1e6ca74b9aa-Abstract.html.

Kunal Jha, Wilka Carvalho, Yancheng Liang, Simon Shaolei Du, Max Kleiman-Weiner, and Natasha Jaques. Cross-environment cooperation enables zero-shot multi-agent coordination. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27198–27220, 2025. URL https://proceedings.mlr.press/v267/jha25b.html.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4940–4950, 2021. URL https://proceedings.mlr.press/v139/jiang21b.html.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger (eds.), *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49:209–232, 2002. doi: 10.1023/A:1017984413808.

Niklas Lauffer, Ameesh Shah, Micah Carroll, Michael D. Dennis, and Stuart Russell. Who needs to know? minimal knowledge for optimal coordination. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18599–18613, 2023. URL https://proceedings.mlr.press/v202/lauffer23a.html.

Yang Li, Shao Zhang, Jichen Sun, Wenhao Zhang, Yali Du, Ying Wen, Xinbing Wang, and Wei Pan. Tackling cooperative incompatibility for zero-shot human-AI coordination. *Journal of Artificial Intelligence Research*, 80:1139–1185, 2024. doi: 10.1613/jair.1.15884.

Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. Trajectory diversity for zero-shot coordination. In *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.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020. URL https://jmlr.org/papers/v21/20-212.html.

George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14:265–294, 1978. doi: 10.1007/BF01588971.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782. Corrected author version available as arXiv:2105.10386 (2021).

Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving curricula with regret-based environment design. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17473–17498, 2022. URL https://proceedings.mlr.press/v162/parker-holder22a.html.

Mikayel Samvelyan, Akbir Khan, Michael Dennis, Minqi Jiang, Jack Parker-Holder, Jakob Foerster, Roberta Raileanu, and Tim Rocktäschel. MAESTRO: Open-ended environment design for multi-agent reinforcement learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sKWlRDzPfd7.

Bidipta Sarkar, Andy Shih, and Dorsa Sadigh. Diverse conventions for human-AI collaboration. In *Advances in Neural Information Processing Systems*, volume 36, pp. 23115–23139, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/4818263715b25dc137d393af8af6d2fc-Abstract.html.

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

DJ Strouse, Kevin R. 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.

Christopher Tosh and Sanjoy Dasgupta. Diameter-based active learning. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3444–3452, 2017. URL https://proceedings.mlr.press/v70/tosh17a.html.

Johannes Treutlein, Michael Dennis, Caspar Oesterheld, and Jakob Foerster. A new formalism, method and open issues for zero-shot coordination. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10413–10423, 2021. URL https://proceedings.mlr.press/v139/treutlein21a.html.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. doi: 10.1007/BF00992696.

Won-Sang You, Tae-Gwan Ha, Seo-Young Lee, and Kyung-Joong Kim. Automatic curriculum design for zero-shot human-AI coordination. *IEEE Access*, 13:143606–143617, 2025. doi: 10.1109/ACCESS.2025.3596640.

Rui Zhao, Jinming Song, Yufeng Yuan, Haifeng Hu, Yang Gao, Yi Wu, Zhongqian Sun, and Wei Yang. Maximum entropy population-based training for zero-shot human-AI coordination. In *AAAI Conference on Artificial Intelligence*, volume 37, pp. 6145–6153, 2023. doi: 10.1609/aaai.v37i5.25758.

Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *AAAI Conference on Artificial Intelligence*, volume 29, 2015. doi: 10.1609/aaai.v29i1.9761.

Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. An overview of machine teaching. arXiv:1801.05927, 2018. URL https://arxiv.org/abs/1801.05927.

A FEEDBACK-AWARE RATES AND FINITE-DIMENSIONAL CONFLICT GEOMETRY

A.1 PATH LAWS AND WHAT IS OBSERVED

The invariant protocol can be defined for any exogenous distribution over paths of length at most H, not only a Markov chain on the observable nodes. Conditional on a latent path ω, let a wrong assignment at its tth visit fail with probability $\Delta_{e,t}(\omega)$, independently of the other hazard coins. Correct matched assignments survive, and a mismatch fails. Then

$$r_{ej}(\omega) = \prod_{t:j_t=j} (1 - \Delta_{e,t}(\omega)), \qquad F_e(E) = 1 - \mathbb{E}_\omega \prod_{j\in E} r_{ej}(\omega).$$

Every path proof in Appendix G uses this product representation. In particular, monotonicity, submodularity, removal of common errors, and the partition characterization remain valid for explicit path banks and fixed-size random subsets. This is an extension of the node-level Markov specification, not an assertion that every path bank has that specification. The implementation uses a Markov dynamic program for $(\alpha_e, P_e, \Delta_{ej})$ tasks and a finite weighted sum for explicit path banks. Complexity statements identify which representation is supplied. Matched actions still cannot steer the latent path, and all sources still share θ.

The two feedback interfaces expose different parts of this execution. Under delayed verification, the learner receives only its sampled convention and a final binary reward; the public execution length is fixed and neither a failure location nor a failure time is disclosed. Under trace feedback, the learner is told the first failed node, or that the episode survived. The policies being evaluated have the same returns in either case. Hiding naturally observed termination would change the learning problem, so the delayed experiments are labeled as such throughout.

A.2 A BRANCHING-PATH TRACE BOUND

Theorem 6 (Correction under arbitrary invariant paths)**.** *Fix an independent and identically distributed (iid) source curriculum μ, a common preferred convention θ, and a nonempty set C of relevant nodes. Write*

$$s_C(\mu) = \min_{j\in C} F_\mu(\{j\}) > 0.$$

Start from any convention and, after each episode, flip only the bit at the first failed matched assignment, if there is one. Let T_C be the first episode after which every bit in C is correct. Then

$$\mathbb{E}T_C \le d/s_C(\mu), \qquad \mathbb{P}(T_C > N) \le \mathbb{P}\{\mathrm{Bin}(N, s_C(\mu)) < d\}. \tag{11}$$

If $h_q(C^c) \le \varepsilon$ and $F_\nu(C^c) \le 1 - c$, correcting C suffices for target compatibility at $\tau = 1 - \varepsilon$ and common-source competence c for two learners. A union bound applies to their correction-time tails.

Proof. A preferred matched assignment never fails. A failure therefore identifies a wrong bit with certainty; flipping it corrects that bit, and it can never be undone. At most d corrections occur. Before T_C, the current error set E contains a node $j \in C$. Its episode failure probability satisfies

$$F_\mu(E) \ge F_\mu(\{j\}) \ge s_C(\mu)$$

by monotonicity. Importantly, the failed node need not itself belong to C. This is why the numerator in the bound is d, not automatically $|C|$.

Use independent uniforms to generate a success indicator with probability $s_C(\mu)$ underneath each actual failure event before T_C. After T_C, extend these indicators with independent Bernoulli variables of the same mean. If $T_C > N$, fewer than d actual failures have occurred in the first N episodes, and hence fewer than d of the coupled Bernoulli successes have occurred. This proves the binomial tail bound. Equivalently, T_C is stochastically bounded by the waiting time for d iid Bernoulli successes of probability $s_C(\mu)$, whose expectation is $d/s_C(\mu)$. This argument does not assume that paths are deterministic, do not branch, or do not revisit nodes.

After correction, each learner's remaining error set is a subset of C^c. Their disagreement is also a subset of C^c. Monotonicity of F_ν and h_q proves both requirements. Applying the two single-learner tail bounds proves the pair statement without treating its episodes as one learner's sample. □

The appropriate curriculum for this *upper bound* solves

$$\max_{\mu\in\Delta_M,s} s \quad \text{subject to} \quad \sum_e \mu_e F_e(\{j\}) \geq s \quad (j \in C). \tag{12}$$

It has $|C|$ exposure constraints. With an explicitly given Markov model, its coefficients are computable in polynomial time. With an explicit path bank they are finite sums. The result does not establish minimax optimality of this objective or imply that its curriculum has the smallest empirical correction time: the bound charges at most d corrections and uses only the weakest surviving singleton. It nevertheless identifies a simpler learning-relevant quantity than the near-optimal-set margin.

For the Markov tasks there is a second elementary bound. Let

$$p_j = \sum_e \mu_e \alpha_{ej} \Delta_{ej}.$$

Whenever bit j is still wrong, an episode starts and fails at j with probability p_j, independent of the other bits. Thus

$$\mathbb{P}(T_{[d]} > N) \leq \sum_j (1 - p_j)^N. \tag{13}$$

This bound does not exploit failures on later visits. When every $p_j > 0$, it gives an exponentially decreasing upper bound on the probability of any unresolved bit.

A.3 When a competence threshold forces exact identification

Proposition 7 (Strict competence region)**.** *In an invariant protocol, let $a_\nu = \min_j F_\nu(\{j\}) > 0$. If $c > 1 - a_\nu$, then the only convention with $V_\nu(b) \geq c$ is θ. Consequently a pair meets the joint requirement exactly when both learners recover θ, for any $\tau \leq 1$.*

Proof. Any nonempty error set contains a node j and has loss at least $F_\nu(\{j\}) \geq a_\nu$. Its return is at most $1 - a_\nu < c$. Conversely θ has source and target self-play return 1. The strictness matters: at equality a least-cost singleton error can remain competent. □

This observation explains flat sections of the empirical competence sweep; it is not a justification for selecting a floor after looking at results.

A.4 Any fixed number of conflicting one-step sources

Theorem 8 (Fixed-source sign-cell optimization)**.** *Consider the one-step conflict model with M sources, rational exposures, and target law q. Suppose every singleton is harmful, that is $\varepsilon < \min_j q_j$. Fix the common competence distribution ν and floor c. For fixed M, maximizing the positive joint margin reduces to at most*

$$\sum_{k=0}^{M-1} \binom{d}{k} = O(d^{M-1})$$

optimization linear programs, one per full-dimensional sign cell in the source simplex. Standard fixed-dimensional arrangement construction enumerates the cells in polynomial time. A straightforward incremental implementation also suffices, but can use $O(d^M)$ additional feasibility programs.

Proof. Each signed score $s_j(\mu)$ is affine on the $(M-1)$-dimensional source simplex. If it is identically zero, there are optimal conventions differing at the harmful singleton j, and the joint margin is everywhere zero. Otherwise the d zero hyperplanes divide the simplex into at most the displayed number of full-dimensional cells. Intersecting an arrangement with a convex simplex cannot split one sign cell into two. This is the classical fixed-dimensional hyperplane-arrangement bound and construction (Edelsbrunner et al., 1986).

On a sign cell with signs $\sigma_j \in \{-1, 1\}$, the source-optimal vector θ_σ is fixed. Positive margin requires strict signs and hence a unique optimum. If $V_\nu(\theta_\sigma) < c$, its incompetent self-loop has gap zero, and the cell cannot yield a positive joint margin. In a competent cell, every other convention differs on at least one harmful singleton. Its gap is at least $\min_j |s_j|$, and flipping a least-cost bit from

θ_σ attains that value as a coordination-bad pair. Thus the joint margin is $\min_j |s_j|$ throughout the cell.

For each competent sign cell solve

$$\max_{\mu,t} t \quad \text{subject to } \mu \in \Delta_M,\ t \geq 0, \quad \sigma_j s_j(\mu) \geq t \quad (j \in [d]).$$

Taking its closure is harmless: at a score-zero boundary the margin is zero. Every positive maximizer, including one on a simplex facet, belongs to the closure of a full-dimensional sign cell with the same strict signs. Therefore these programs miss no positive optimum. Rational LP solving in fixed or variable dimension is polynomial in input length; the number of cells is polynomial only for fixed M. Incrementally retaining feasible strict sign patterns at the first r hyperplanes costs $O(r^{M-1})$ tests per insertion and hence $O(d^M)$ in total; these are distinct from the final optimization programs. □

The condition explicitly implies $\varepsilon < 1/d$, since $\min_j q_j \leq 1/d$. This theorem does not solve arbitrary fixed-tolerance problems as d grows, and switches off target-harmless singleton ambiguity. Numerical checks use floating-point strict-feasibility tests and compare to independently enumerated sign programs on small instances; they do not establish exact-arithmetic correctness.

A.5 Weak hardness already with one source

Proposition 9 (Exact margin evaluation contains PARTITION)**.** *Evaluating the coordination margin exactly is weakly NP-hard (NP denotes nondeterministic polynomial time) even for one source, one-step episodes, and an invariant preferred convention, when the target tolerance is part of the input.*

Proof. Let $z_1, \ldots, z_d$ be positive integers of a PARTITION instance, and $Z = \sum_j z_j$. Use one source starting at j with probability z_j/Z, unit wrong-assignment hazards, and $\theta = 0$. Use a uniform target with $\varepsilon = (d-1)/d$. Only the full set of nodes is harmful. The balanced identity gives

$$\Gamma = \frac{1}{Z} \min_{V \subseteq [d]} \max \left\{ \sum_{j \in V} z_j,\ Z - \sum_{j \in V} z_j \right\}.$$

This number is at most $1/2$ exactly when the original integers can be partitioned into two equal sums. The probabilities and threshold have polynomial-length rational encodings. Hence exact margin evaluation solves PARTITION (Karp, 1972). On this subclass the subset-sum dynamic program is pseudo-polynomial in Z, which is consistent with weak, rather than strong, hardness. The reduction is not a hardness statement for every fixed ε, and not a hardness result for a particular approximation ratio. □

The elementary reduction justifies treating the partition and cover oracles as computational objects rather than free operations. The κ surrogate, the singleton-harmful restriction, and small exact instances address different parts of this difficulty; none supplies a general polynomial solution.

B Conflicting sources and restricted-class guarantees

B.1 Exact signed-exposure formula

Assume $w_{ej} \geq 0$, $\sum_j w_{ej} \leq 1$, and $\theta_e \in \{0,1\}^d$. The one-step source return is

$$J_e(b,b) = 1 - \sum_j w_{ej} \mathbf{1}\{b_j \neq \theta_{ej}\}.$$

This is realized by a source drawing node j with probability p_{ej} and failing a wrong matched assignment with probability Δ_{ej}, where $w_{ej} = p_{ej}\Delta_{ej}$. The target draws a node with probability q_j and fails only on disagreement. All 2^d deterministic conventions are admissible, and the policy does not observe source identity.

Proof of Theorem 4, gap identity. Let $L_{0j} = \sum_e \mu_e w_{ej}\theta_{ej}$ be the expected loss of choosing bit 0 at node j and $L_{1j} = \sum_e \mu_e w_{ej}(1-\theta_{ej})$ the loss of choosing bit 1. Then

$$S_\mu(b) = 1 - \sum_j \{(1-b_j)L_{0j} + b_j L_{1j}\}, \qquad S^*_\mu = 1 - \sum_j \min(L_{0j}, L_{1j}).$$

The difference $L_{0j} - L_{1j}$ equals $s_j(\mu)$. If $s_j > 0$, bit 1 is better by s_j; if $s_j < 0$, bit 0 is better by $-s_j$; if $s_j = 0$, either is optimal. Subtracting the two displayed scores gives exactly

$$g_\mu(b) = \sum_j |s_j(\mu)| \mathbf{1}\{b_j \neq \theta_{\mu j}\}.$$

For a fixed μ, this is a nonnegative modular error cost. Applying the disjoint-error argument to reference θ_μ, with target disagreement $q(E_b \triangle E_{b'})$, proves the balanced formula. The reference is curriculum-dependent: it cannot be fixed before varying μ. □

B.2 A POLYNOMIAL TWO-SOURCE OPTIMIZER

Let the weight on source 1 be $p \in [0,1]$. Each coordinate is affine,

$$s_j(p) = w_{0j}(2\theta_{0j} - 1) + p\{w_{1j}(2\theta_{1j} - 1) - w_{0j}(2\theta_{0j} - 1)\}.$$

Its nonconstant zero inside $(0,1)$ contributes at most one breakpoint. Sorting all breakpoints gives at most $d+1$ open sign chambers. If a coordinate is identically zero, the maximum margin is zero whenever its singleton is target-harmful, and this can be returned immediately.

Suppose $\varepsilon < \min_j q_j$, so any disagreement is harmful. For every fixed p,

$$\Gamma_{\mathrm{coord}}(p) = \min_j |s_j(p)|.$$

Indeed any two distinct conventions have at least one nonzero error between them; their larger gap is at least the smallest absolute exposure. Flipping a smallest-exposure bit from an optimal convention attains equality, including zero exposure. For positive margin the optimum θ_p is unique. Let $V(b) = \sum_e \nu_e J_e(b,b)$ use a fixed ν. If $V(\theta_p) < c$, its self-loop belongs to $\mathcal{B}_{\tau,c}$ and $\Gamma_{\mathrm{joint}}(p) = 0$. If $V(\theta_p) \geq c$, every incompetent convention differs from θ_p and has gap at least $\min_j |s_j(p)|$. Therefore

$$\Gamma_{\mathrm{joint}}(p) = \begin{cases} \min_j |s_j(p)|, & V(\theta_p) \geq c, \\ 0, & V(\theta_p) < c. \end{cases}$$

At a sign-changing breakpoint the coordination margin is zero; no positive maximizer is missed by considering adjacent chambers.

Inside a chamber $[l,u]$ the signs $\sigma_j \in \{-1,1\}$ and optimum convention are fixed. Discard the chamber if that convention violates competence. Otherwise solve

$$\max_{p,t} t \quad \text{subject to } l \leq p \leq u, \quad t \geq 0, \quad \sigma_j s_j(p) \geq t \quad (j = 1, \ldots, d).$$

There are two variables and d coordinate constraints. Maximizing the returned value across all chambers is exact for the maximum *positive* joint margin; if none has positive value, return zero. Roots, sign tests, and LP data are polynomial-sized for rational input. This is not an enumeration of 2^d orientations: two-source mixtures encounter at most $d+1$ of them. Endpoints with positive margin are covered because no sign vanishes there.

The implementation checks the identity against direct finite-class evaluation on 240 random conflicting instances. A separate 40-instance scaling sweep runs the chamber algorithm up to $d = 512$ and compares to a 2,001-point mixing grid. A dense grid is a numerical lower comparator, not the proof of exactness.

B.3 ARBITRATION AND FAILURE OF THE SEQUENTIAL EXTENSION

For a two-node one-step example, let the sources prefer 00 and 11, with exposures $(.4, .1)$ and $(.1, .4)$, respectively. Under equal source weights for common evaluation, the common evaluation score is .90 for 01, .75 for 00 or 11, and .60 for 10. Equal training has signed exposures $(-.15, +.15)$ and gaps

$0, .15, .15, .30$ relative to 01. Pure-source training has optimum 00 or 11 and fails competence floor .85, despite positive coordination margin. The mixture can therefore satisfy a joint requirement that neither pure optimum satisfies.

The sequential counterexample uses source scores that are 1 exactly at their respective vectors $000, 101, 011$ and 0 elsewhere. Such scores are realized by a sure path through all three nodes with hazard 1 for every wrong assignment. Equal weighting gives score $1/3$ to those three vectors and 0 to the other five. The target's one-step law is $q = (1/2, 1/2, 0)$; hence 101 and 011 disagree on target mass 1 and are bad at $\varepsilon = .6$. Their gaps are zero.

Relative to reference 000, their error sets are $\{0, 2\}$ and $\{1, 2\}$. Removing common error 2 gives policies 100 and 010, each with gap $1/3$. Thus the removal step increases both gaps. To check the claimed erroneous formula independently, any target-harmful U must contain nodes 0 and 1. A disjoint partition cannot simultaneously realize both zero-gap sets $\{0, 2\}$ and $\{1, 2\}$; the other zero-gap set is empty. Every such partition has at least one gap $1/3$, and some attain it. Therefore the disjoint formula gives $1/3$ while the actual margin is zero. A generic finite-table margin remains valid, but normalized monotone source-error costs need not exist. Action-dependent paths are a second, independent reason that the latent-path representation may fail.

B.4 Restricting the policy class changes the optimum

Proposition 10 (Reference-optimum correction)**.** *Fix μ and restrict a full class Π to a nonempty Π'. Write $D = \max_{i\in\Pi} S_i - \max_{i\in\Pi'} S_i \geq 0$ and $\mathcal{B}' = \mathcal{B} \cap (\Pi' \times \Pi')$. If $\mathcal{B}' \neq \varnothing$, then*

$$\Gamma_{\Pi'}(\mu; \mathcal{B}') = \min_{(i,j)\in\mathcal{B}'} \max(g_i^{\Pi}, g_j^{\Pi}) - D \geq \max\{0, \Gamma_{\Pi}(\mu; \mathcal{B}) - D\}.$$

If $\mathcal{B}'$ is empty, its margin is $+\infty$. In particular, $D = 0$ when Π' contains a full-class maximizer, and then the full-class margin is a lower bound.

Proof. For every retained policy, $g_i^{\Pi'} = g_i^{\Pi} - D$. Subtract D inside the maximum and minimum. Restricting the bad-pair set cannot lower the full-gap minimum, and restricted gaps are nonnegative. For necessity of the correction, take scores $(1, .6, .6)$ with only the pair of the last two policies bad. The full margin is $.4$. Restricting to those two policies drops the optimum by $.4$ and the margin to zero. □

C Computational guarantees and informative separating families

C.1 Tightness inside the restricted path family

The following one-step construction attains the path surrogate's worst-case approximation ratio.

Proposition 11 (Tight optimized path LP)**.** *The factor $1/2$ for the optimized path-coverage LP cannot be improved in general even for two one-step sources and two nodes.*

Proof. Let $q = (1/2, 1/2)$ and $\varepsilon = 3/4$, so only $U = \{1, 2\}$ is harmful. Source A starts uniformly and has unit hazards, giving exposures $(1/2, 1/2)$. Source B starts at node 1 and has hazard $1 - h'$, with $0 < h' < 1/2$, giving exposures $(1 - h', 0)$. At weight p on A,

$$a_1 = p/2 + (1 - p)(1 - h'), \quad a_2 = p/2, \quad \kappa(p) = 1 - (1 - p)h', \quad \Gamma(p) = a_1.$$

The last equality follows by splitting the two nodes, since $a_1 \geq a_2$; the alternative unsplit cost is no smaller. κ is strictly increasing in p, so its unique optimum is $p = 1$, with margin $1/2$. The actual margin is strictly decreasing, so its optimum is $p = 0$, with value $1 - h'$. The ratio $1/[2(1 - h')]$ tends to $1/2$ as $h' \downarrow 0$. No tie-breaking ambiguity explains the failure. Numerical checks evaluate four positive values of h'. □

C.2 Feasible approximate oracles and multiplicative weights

Assume there is at least one harmful set, all costs lie in $[0, 1]$, and the oracle returns $U(\mu)$ satisfying

$$h_q(U(\mu)) > \varepsilon, \qquad F_\mu(U(\mu)) \leq \rho\kappa(\mu), \qquad \rho \geq 1.$$

We require exact feasibility for the stated target threshold. An approximation that only satisfies a relaxed cover constraint requires a separately stated relaxed target guarantee. This distinction matters for submodular-cost submodular-cover algorithms (Iyer & Bilmes, 2013).

For source e, initialize log weight $z_{1e} = 0$. At iteration t, set $\mu_{te} \propto \exp(z_{te})$, query U_t, and update $z_{t+1,e} = z_{te} + \lambda F_e(U_t)$. Return $\bar{\mu} = T^{-1} \sum_t \mu_t$. For $M > 1$, use $\lambda = \sqrt{8 \log M/T}$; for $M = 1$ the only source is fixed.

For the returned average, the guarantee is

$$\Gamma(\bar{\mu}) \geq (\Gamma^* - \sqrt{\log M/(2T)})/(2\rho). \tag{14}$$

Proof of (14). The exponential-potential bound for gains in $[0, 1]$ gives

$$\frac{1}{T} \sum_t F_{\mu_t}(U_t) \geq \max_e \frac{1}{T} \sum_t F_e(U_t) - r_T, \qquad r_T = \sqrt{\log M/(2T)}.$$

One obtains this by summing $\log \mathbb{E}_{e \sim \mu_t} \exp(\lambda F_e(U_t)) \leq \lambda F_{\mu_t}(U_t) + \lambda^2/8$ and comparing the final potential with each source's cumulative gain. This is the usual multiplicative-weights argument (Arora et al., 2012), valid for the oracle's adaptive responses. The maximum cumulative source payoff is at least the cumulative payoff of any mixture μ^*. For a mixture maximizing κ, every feasible U_t has $F_{\mu^*}(U_t) \geq \kappa^*$. Hence

$$\rho \frac{1}{T} \sum_t \kappa(\mu_t) \geq \frac{1}{T} \sum_t F_{\mu_t}(U_t) \geq \kappa^* - r_T.$$

The pointwise minimum of linear functions κ is concave, giving $\kappa(\bar{\mu}) \geq T^{-1} \sum_t \kappa(\mu_t)$. Therefore $\Gamma(\bar{\mu}) \geq (\kappa^* - r_T)/(2\rho) \geq (\Gamma^* - r_T)/(2\rho)$. A negative lower bound is replaced by zero. The empirical dual quantity $\max_e T^{-1} \sum_t F_e(U_t)$ upper-bounds κ^*, while $F_{\bar{\mu}}(U(\bar{\mu}))/(2\rho)$ is an a-posteriori lower bound on the returned margin. □

For constraint generation, solve a restricted master LP with $0 \leq t \leq 1$. Its optimum t upper-bounds κ^*. If the approximate oracle returns value at least $t - \zeta$, then $\rho\kappa(\mu) \geq t - \zeta \geq \kappa^* - \zeta$, and the same half-factor gives $(\Gamma^* - \zeta)/(2\rho)$. If a violated constraint is returned, add it. This is a valid *termination certificate*; we do not infer polynomial iteration count from it. The multiplicative-weights construction supplies an explicit oracle-call bound instead.

C.3 A POLYNOMIAL ORACLE FOR UNIFORM ONE-VISIT TARGETS

Let the target visit one of d nodes uniformly, and put $k = \lfloor \varepsilon d \rfloor + 1$. Harmful sets have cardinality at least k. Since F_μ is monotone, an optimum can be chosen with cardinality exactly k. Compute all singleton costs $c_j = F_\mu(\{j\})$ and return the k smallest.

For a fixed path and a set U of size k, let $l \leq \min(k, H)$ be the number of its distinct visited nodes. Each nonzero singleton cost $1 - r_j$ is at most $1 - \prod_{i \in U} r_i$. The union bound supplies the other direction, so pathwise

$$1 - \prod_{j \in U} r_j \leq \sum_{j \in U} (1 - r_j) \leq \min(k, H) \left(1 - \prod_{j \in U} r_j \right).$$

After expectation, this becomes $F_\mu(U) \leq \sum_{j \in U} c_j \leq \rho F_\mu(U)$. If U^* is a minimum-cost harmful k-set and $\widehat{U}$ the returned set, then

$$F_\mu(\widehat{U}) \leq \sum_{j \in \widehat{U}} c_j \leq \sum_{j \in U^*} c_j \leq \rho F_\mu(U^*), \qquad \rho = \min(k, H).$$

Thus feasibility and the required approximation factor both hold. Computing d singleton costs and sorting is polynomial given value access. Model acquisition is not included in the oracle-call bound.

C.4 PATH CORRELATION, NOT JUST MARGINAL EXPOSURE

Let $n \geq 4$ be even, with n relevant nodes and one dummy. The target visits one relevant node uniformly, with $\varepsilon = (n-1)/n$. Its only harmful relevant set is the full union. Source A visits all n relevant nodes once with probability $p > 1/2$, otherwise only dummy nodes. Source B draws a uniformly random $n/2$-subset, visits each selected node once, and pads to the same horizon with dummy visits. Relevant hazards are 1 and dummy hazards zero.

Every relevant singleton has both occupancy and failure cost p under A, and $1/2$ under B. An additive source proxy, whether divided by the common horizon or not, strictly prefers pure A. For any nonempty relevant set, $F_A(U) = p$, so its balanced margin is p. Under B, $F_B(U)$ is the probability a random half-set intersects U. If a partition has one side larger than $n/2$, that side has cost 1. For an equal partition, both sides have cost $1 - 1/\binom{n}{n/2}$. This is the best partition, so

$$\Gamma_B = 1 - \binom{n}{n/2}^{-1}.$$

Whole-union cost is p under A and 1 under B, so exact path coverage selects B. The gain Γ_B/Γ_A approaches $1/p$, arbitrarily close to two when $p \downarrow 1/2$. The target risk is already exact and additive because $H_q = 1$; this separation isolates source path dependence. It does not rely on dropping a low-probability target branch. The closed-form values are checked by brute-force partition tests at $n = 4, 6, 8$.

D COMPETENCE-AWARE DESIGN AND LEARNING

Proposition 12 (Joint margin and joint LP)**.** *In the invariant path class, fix ν, $\beta = 1 - c$, and $\tau = 1 - \varepsilon$. Let $\ell(\mu) = \min_{E:F_\nu(E)>\beta} F_\mu(E)$. Then* (7) *holds. The joint LP specified after* (7) *returns a certified lower joint margin t and satisfies $t \geq \frac{1}{2}\max_\mu \Gamma_{\mathrm{joint}}(\mu)$.*

Proof. The bad relation is the union of target-incompatible pairs and pairs with at least one incompetent policy. The first part has margin Γ. For an incompetent error set E, every pair involving it has larger gap at least $F_\mu(E)$, and its self-loop attains that value. Minimizing over these sets gives ℓ, proving the identity.

The LP's harmful-union constraints imply $\kappa \geq 2t$ and hence $\Gamma \geq t$; its incompetent-set constraints imply $\ell \geq t$. For a mixture achieving the largest joint margin G^*, $\kappa \geq \Gamma \geq G^*$ and $\ell \geq G^*$. The same mixture with $t = G^*/2$ is LP-feasible. Therefore the LP optimum is at least $G^*/2$. The certificate admits optimization tolerances strictly below t, consistent with the closed near-optimal set. □

Theorem 13 (Reward-only learning with a common competence floor)**.** *In the invariant model under iid delayed-verification episodes, execute a uniformly random convention B with both self-play role copies and observe terminal $R \in [0,1]$. Put $Z_j = (2B_j - 1)R$. Then $\mathbb{E}Z_j = (2\theta_j - 1)m_{\mu j}/2$, and $m_{\mu j} > 0$ exactly when $F_\mu(\{j\}) > 0$. Estimate each bit by the sign of its empirical moment, using a private random bit at ties. If a predetermined common set C satisfies $h_q([d] \setminus C) \leq \varepsilon$ and $F_\nu([d] \setminus C) \leq \beta$, two learners obey*

$$\mathbb{P}(\textit{XP} < 1 - \varepsilon \textit{ or either SP}_\nu < 1 - \beta) \leq \min\left\{1, 2\sum_{j \in C} \exp(-N m_{\mu j}^2/8)\right\}. \tag{15}$$

In particular, $N \geq 8m_{\min}^{-2}\log(2|C|/\delta)$ suffices for failure probability at most δ when $m_{\min} = \min_{j \in C} m_{\mu j} > 0$.

The moment and probability proofs are in Appendix H. Requiring both constraints on the *same* ignored set prevents the two learners from each spending a full but different risk allowance. Competence-aware cover selection enumerates such sets and solves a min-signal LP on the complement. It optimizes a sufficient learning-bound surrogate, not exact finite-sample success. If the only allowed ignored set is empty, it reduces to probe all; that is exactly what happens at $\beta = .02$ in the twelve invariant libraries.

E FEEDBACK-DEPENDENT LOWER AND UPPER BOUNDS

E.1 THE TWO FEEDBACK INTERFACES

The delayed-verification task follows a fixed d-node path. Internal failure flags do not alter its length or public node sequence. Only the final reward reveals whether every matched assignment survived. A learner can choose each bit based on earlier episodes, but obtains no information about the current episode's correctness before its final reward. The failure-location interface instead ends the episode at the first failed node and reports that node. Both define the same expected return for each fixed convention. They differ in information available for future choices.

The uniform-probe learner observes only (B, R) under the delayed-verification interface; the latent pre-failure execution length is not exposed to it. For cost interpretation, a delayed episode uses H public decisions even if an internal failure occurs earlier. The latent pre-failure prefix must not be read as the full delayed-interface interaction cost. The neural-policy experiment uses ordinary, observed termination.

E.2 DELAYED BINARY FEEDBACK IS EQUALITY-QUERY SEARCH

At hazard 1, $R = 1$ exactly when the executed convention equals θ. For any prior history of failed guesses, the unqueried candidates remain indistinguishable under a uniform prior. Repeated guesses cannot help. Fix the learner's internal randomness; after N distinct failed queries there can be at most one additional candidate for its final output. Counting the N guessed candidates and that output proves

$$\mathbb{P}_{\theta\sim\mathrm{Unif}(\{0,1\}^d)}(\widehat{\theta} = \theta) \le \min\{1, (N+1)/2^d\}.$$

Averaging over internal randomness preserves the bound. Since every incorrect convention has source gap 1, producing an η-optimal convention for $\eta < 1$ is identical to identifying θ. A worst-case $(1-\delta)$ guarantee therefore requires $N \ge (1-\delta)2^d - 1$.

For exact identification with probability one, let $K = 2^d$. Under uniform θ, distinct queries encounter the correct candidate at a uniformly distributed rank. If it is last, it can be inferred after $K-1$ failures. No zero-error method can stop earlier while two unqueried candidates remain. The least expected number of queries is consequently

$$\frac{1 + 2 + \cdots + (K-1) + (K-1)}{K} = \frac{(K-1)(K+2)}{2K}.$$

A uniformly random query order achieves this expectation for *every* fixed θ. The uniform-prior lower bound and randomized-order upper bound establish the minimax expected value. This is a standard unstructured-search argument, included to state the stronger interface-level limitation explicitly. It is not a claim that target agreement alone requires exponential samples.

E.3 FAILURE LOCATIONS ALLOW MONOTONE CORRECTION

Execute a current deterministic convention with both self-play copies. A failure at node j proves $b_j \ne \theta_j$, because correct matched assignments never fail. Flip that bit; no later observation can require reverting it. At hazard 1, every episode with a remaining error fixes at least one bit. After at most d episodes all bits are correct, without knowing when that event occurred. An additional successful episode can confirm completion in the fatal-hazard case.

Now suppose every wrong-assignment hazard on the sure path is at least $c_0 > 0$. Whenever an error remains, the next episode fails with probability at least c_0: the earliest remaining wrong node is reached with no previous wrong node, and its failure probability is at least c_0. A later failure only helps. The waiting time for each correction is stochastically dominated by a geometric random variable with parameter c_0, conditional on the past. At most d corrections are required, so expected correction time is at most d/c_0. With $L = \lceil \log(d/\delta)/c_0 \rceil$, each waiting time exceeds L with probability at most $(1-c_0)^L \le e^{-c_0 L} \le \delta/d$. A union bound shows that dL fixed-budget episodes suffice with probability $1-\delta$. This proof does not require a learner to know the true stopping time.

Proposition 14 (Signal of uniform policy probes)**.** *On a deterministic path visiting d distinct nodes with constant hazard $c_0 \in (0, 1]$, the coordination margin is c_0, but each uniform-policy probe signal equals $c_0(1 - c_0/2)^{d-1}$.*

The proof in Appendix H is algebraic. This proposition concerns one estimator. The equality-query bound is learner-independent only for the delayed binary interface at $c_0 = 1$; it is not a lower bound for failure-location feedback, dense rewards, or arbitrary trajectory-feedback algorithms. Both interfaces have exactly the same Γ, so Γ alone cannot determine their sample complexity.

F Population evaluation without an enumerable policy class

Fix all teacher data and decisions first, including curriculum, learning algorithm, training budget, competence floor, and target threshold. Let the random variable π be the policy output by one fresh independent training run; it may have continuous parameters and need not belong to a known finite class. For two independent runs define the population probability

$$p_{\text{joint}} = \mathbb{P}\{V(\pi) \geq c,\ V(\pi') \geq c,\ T(\pi, \pi') \geq \tau\}.$$

This probability is conditional on the chosen task library and teacher decisions. It is not a worst-policy guarantee.

Train n disjoint pairs. For pair i evaluate three means, V_i, V_i', T_i, each using k independent bounded-return episodes. Put

$$r = \sqrt{\log(6n/\delta)/(2k)}, \qquad I_i = \mathbf{1}\{\widehat{V}_i - r \geq c,\ \widehat{V}_i' - r \geq c,\ \widehat{T}_i - r \geq \tau\}, \quad s = \sum_i I_i.$$

Let $L_\alpha(s, n)$ denote the one-sided Clopper–Pearson lower limit: 0 for $s = 0$, and the α quantile of $\text{Beta}(s, n - s + 1)$ otherwise. Report $L_{\delta/2}(s, n)$.

Proposition 15 (Model-free population certificate)**.** *Under independent fresh learner pairs and iid evaluation episodes, with probability at least* $1 - \delta$, $p_{\text{joint}} \geq L_{\delta/2}(s, n)$. *The result is valid conditionally on arbitrary prior pilot-based curriculum selection, provided the evaluation pairs and their episodes are fresh.*

Proof. One-sided Hoeffding bounds and a union bound over $3n$ means imply, except on an event of probability $\delta/2$, that every empirical lower bound is below its true mean. On this event, each certified pair truly satisfies the joint event, so $s \leq S$, where S is the number of truly jointly successful pairs. Conditional on the frozen teacher decisions, $S \sim \text{Binomial}(n, p_{\text{joint}})$. Exact binomial inversion gives $\mathbb{P}\{L_{\delta/2}(S, n) > p_{\text{joint}}\} \leq \delta/2$ (Clopper & Pearson, 1934). Since the lower limit is nondecreasing in its count, $L_{\delta/2}(s, n) \leq L_{\delta/2}(S, n)$ whenever $s \leq S$. A union bound proves the assertion; no independence between these two failure events is needed. Conditioning and then averaging establishes validity after pilot selection. □

Reusing one policy in many pairings does not produce n independent Bernoulli events. Repeated checkpoints or selection of thresholds after examining evaluation outcomes require further multiplicity or sequential corrections. The reported evaluation has 12 designated pairs per method and library, $k = 8192$, and $\delta = .05$. All intervals are per method/library, not simultaneous across methods and libraries. Even 12 certifications would yield only $(1/40)^{1/12} \approx .735$ as a population lower bound. This deliberately modest claim replaces neither a near-optimality bound nor the finite-class Γ certificate.

Pilot selection is not covered by the margin theorem. The pilot teacher trains eight policies under each of five candidate curricula, evaluates four disjoint pairs using 2,048 rollouts per return, and selects by empirical joint success with a tiny mean-XP tie breaker. It receives no preferred convention or transition table. Its selection score is not itself certified and can overfit those pilot pairs. The independent evaluation, not a claim of optimal pilot selection, supplies the subsequent confidence guarantee. Access to representative source and target rollouts remains essential.

G Definitions and complete path proofs

G.1 Notation and information available to each component

The model-aware teacher receives $(\alpha_e, P_e, \Delta_e)$ and the target path law, but not θ. These magnitudes determine F_e, h_q, and m_{ej} independently of the bit orientation. The probing learner receives its

Symbol	Meaning
μ	Stationary probability distribution over source modes
b, θ	Learned convention vector and preferred source convention
E_b	Nodes where b differs from θ
$F_e(E), F_\mu(E)$	Source and curriculum failure probabilities for error set E
$h_q(U)$	Probability that a target path visits at least one node in U
η	Allowed global self-play optimality gap
ε	Allowed target failure probability; $\tau = 1 - \varepsilon$
Γ, κ	Exact balanced margin and whole-set coverage surrogate
$m_{\mu j}$	Mean reward advantage of correcting node j under random probes
C	Common predetermined set of nodes whose bits must be resolved

chosen policy bits and the terminal reward; it does not receive a correctness label, the preferred bits, the hazard rates, or the source identity. The evaluator alone uses θ for exact self-play gaps. All methods in the sequential comparison share the same structural teacher access. The method is not an environment-discovery algorithm without simulator access.

The latent un-killed path is a proof device. In the immediate-trace interface the episode ends on failure. For terminal-only experiments, the learner is instead given a delayed-verification wrapper: the latent path continues to its fixed horizon after an internal failure and only the final reward reports success. The executed-feedback distinction is essential to Theorem 5. Conditional on a latent path, the wrong-assignment survival probability is the product of independent per-visit survival probabilities. Node labels and complementary roles are common across teams. The convention class excludes time-dependent and history-dependent rules; such rules cannot be silently included in the guarantee.

G.2 Finite-class equivalence

Lemma 16. *For a fixed finite class and $\eta \geq 0$, every ordered pair in $\mathcal{A}_\eta(\mu)$ has return at least τ if and only if $\Gamma_A(\mu; \mathcal{B}_\tau) > \eta$.*

Proof. A violation is a bad pair (i, j) with $g_i \leq \eta$ and $g_j \leq \eta$. This is equivalent to $\max(g_i, g_j) \leq \eta$ for some bad pair, or $\Gamma_A \leq \eta$. Negating the statement proves the result. An empty bad-pair set makes compatibility unconditional and gives margin $+\infty$. The finite class ensures that a nonempty minimum is attained. □

For unequal tolerances η_1, η_2, the direct condition is that no bad ordered pair has $g_i \leq \eta_1$ and $g_j \leq \eta_2$. The single scalar margin in the paper uses the same tolerance for the two roles.

G.3 Path representation and submodularity

Lemma 17. *For every source e, F_e in* (4) *is normalized, monotone, and submodular. The self-play gap of b is $F_\mu(E_b)$; target cross-play loss is $h_q(E_b \triangle E_c)$.*

Proof. On a fixed latent path, the survival probability of a policy with error set E is $\prod_{j \in E} r_{ej}$, with all factors in $[0, 1]$. Averaging gives the first expression in (4). The preferred policy θ survives with probability 1 in every source, so the best self-play score is 1 and its competitors' gaps equal these failure probabilities.

For a fixed path, $f(E) = 1 - \prod_{j \in E} r_{ej}$ has $f(\varnothing) = 0$. If $j \notin E$, its increment is

$$f(E \cup \{j\}) - f(E) = (1 - r_{ej}) \prod_{k \in E} r_{ek} \geq 0.$$

If $E \subseteq E'$ and $j \notin E'$, the increment at E' is no larger, since it contains more factors in $[0, 1]$. This proves monotonicity and diminishing returns. Expectations and nonnegative mixtures preserve both properties. In particular, $F_\mu(V) + F_\mu(W) \geq F_\mu(V \cup W)$ for disjoint sets. This is a use of standard submodularity, not a new general set-function theorem (Nemhauser et al., 1978).

Two policies select different assignments precisely on $E_b \triangle E_c$. The target has no hazard for a matching assignment, so an episode fails exactly when its latent path first hits that disagreement set. Thus its failure probability is $h_q(E_b \triangle E_c)$. The same path argument shows that h_q is a normalized monotone coverage function. □

A dynamic program starts with mass α_e and repeats H updates $v \leftarrow (v \odot s)P_e$, where $s_j = 1 - \Delta_{ej}\mathbf{1}\{j \in E\}$. The final mass sum is survival probability. Multiplying by P_e after the final survival step is harmless because $P_e\mathbf{1} = \mathbf{1}$. Replacing the hazard by 1 on U computes target avoidance.

G.4 Balanced exposure and the approximation constants

Proof of Theorem 2. A bad pair with error sets B, D has $h_q(B \triangle D) > \varepsilon$. Remove the common errors: set $V = B \setminus D$, $W = D \setminus B$. Then $V \cup W = B \triangle D$, while monotonicity gives $F_\mu(V) \le F_\mu(B)$ and $F_\mu(W) \le F_\mu(D)$. Removing common errors preserves target disagreement and cannot increase the larger gap. Conversely, for every harmful U and $V \subseteq U$, the full convention class contains the two policies with error sets V and $U \setminus V$. Taking the minimum over these pairs proves (5).

For every partition, subadditivity gives

$$\max\{F_\mu(V), F_\mu(U \setminus V)\} \ge \tfrac{1}{2}\{F_\mu(V) + F_\mu(U \setminus V)\} \ge \tfrac{1}{2}F_\mu(U).$$

Minimizing gives the lower bound. Choosing $V = U$ and the other part empty gives the upper bound. For pair sums, first remove common errors, then use the same subadditivity inequality to obtain a value at least κ. The pair with error sets $U, \varnothing$ attains $F_\mu(U)$, so minimizing over harmful U gives equality. □

Both constants are attained within the one-step subclass. For the upper constant, use one target-relevant node with positive exposure. For the lower constant, let $q = (1/2, 1/2)$, $\varepsilon = 3/4$, and source exposure $a = (s, s)$ with $0 < s \le 1/2$. Only the two-node set is harmful; splitting it gives $\Gamma = s = \kappa/2$. These examples concern the relationship between Γ and κ. Proposition 11 additionally proves tightness of the optimized path LP within this same one-step subclass.

The proof applies more generally to any normalized monotone subadditive error cost that is the exact self-play gap, with disagreement determined by symmetric difference. Submodularity is a convenient sufficient property supplied by the protocol. The sequential step is deriving that cost from the stochastic execution model; assuming an arbitrary return table is already additive would miss it.

G.5 Observability, dilution, and compatible ambiguity

Proof of Corollary 3. Let Z be invisible under all sources. Subadditivity and $F_e(\{j\}) = 0$ imply $F_e(Z) = 0$. If $h_q(Z) > \varepsilon$, the preferred policy and the policy flipping exactly Z form a bad pair with zero gaps under every curriculum.

Conversely, give every source positive weight and suppose $h_q(Z) \le \varepsilon$. Any harmful set contains a node outside Z, whose singleton cost has positive mixture weight. By monotonicity the whole harmful set has positive cost. There are finitely many such sets, hence $\kappa > 0$ and $\Gamma \ge \kappa/2 > 0$. The same proof characterizes any prescribed support by defining invisibility only over sources on that support.

If nondiagnostic modes have $F_e(E) = 0$ for all E, a mixture retaining diagnostic probability p has $F_{\mu'} = pF_\mu$. Every value in the balanced formula scales by p, proving the exact dilution law. A finite-policy analogue holds whenever the added mode assigns the same self-play return to every policy: its common score cancels from all gaps. □

Invisible nodes with total hitting risk at most ε can be left unresolved, so unique identification is unnecessary. Conversely, unseen observation labels with unrestricted behavior invalidate the assumed policy class. Two policies may then agree on every training observation and disagree elsewhere without any source penalty. The results give no unrestricted extrapolation guarantee.

H CONSTRUCTIVE LEARNING AND FIXED-SAMPLE CERTIFICATION

H.1 THE PROBE SIGNAL IDENTITY

Proof of the moment identity in Theorem 13. For uniform B, the other coordinates of E_B are uniform and independent of B_j. Conditional on those other errors, choosing the correct bit instead of the wrong bit increases expected terminal reward by $F_\mu(E \cup \{j\}) - F_\mu(E)$. Therefore

$$\mathbb{E}[(2B_j - 1)R] = \tfrac{1}{2}\{\mathbb{E}[R \mid B_j = 1] - \mathbb{E}[R \mid B_j = 0]\} = (2\theta_j - 1)m_{\mu j}/2.$$

Equivalently, conditional on a latent source path, averaging the marginal over all other error bits gives

$$m_{ej} = \mathbb{E}_{\omega \sim e}\left[(1 - r_{ej}(\omega)) \prod_{k \neq j} \frac{1 + r_{ek}(\omega)}{2}\right]. \tag{16}$$

Every factor in the product lies in $[1/2, 1]$, so $2^{-(d-1)}F_e(\{j\}) \leq m_{ej} \leq F_e(\{j\})$. This proves nonnegativity, orientation independence, and equivalence of positive singleton exposure and positive probe signal, also after mixing sources. The construction is the first-order Fourier correlation under a uniform Boolean input distribution (O'Donnell, 2014); our use of it exploits the common preferred convention and monotone survival model. □

H.2 LEARNING AND CERTIFICATE GUARANTEES

Proof of the probability bound in Theorem 13. Each Z_{tj} lies in $[-1, 1]$ and has mean of magnitude $m_{\mu j}/2$. A wrong sign, including an adverse tie, implies a one-sided deviation of at least that magnitude. Hoeffding's inequality gives probability at most $\exp(-Nm_{\mu j}^2/8)$. Union-bound over $j \in C$ and both learners. Outside this event, both learned vectors equal θ on C, so their disagreement is contained in $[d] \setminus C$. Monotonicity of h_q bounds the target failure probability by ε. This proves (15), with a cap at 1. Bounding every signal below by $m_{\min}$ gives the stated sufficient sample size. Independence across coordinates is unnecessary; episodes must satisfy the stated iid sampling assumption. Independence of the two training runs is used in the experimental design, although the union bound itself does not require it. □

On the same event, the error set of either learner is contained in $U = [d] \setminus C$. Monotonicity therefore gives $g_\mu(b) = F_\mu(E_b) \leq F_\mu(U)$. More generally, for a specified competence-evaluation distribution ν, $g_\nu(b) \leq F_\nu(U)$. Consequently the right side of (15) also bounds the probability that either source-competence condition or the target-compatibility condition fails, provided the ignored-set cost meets the chosen competence tolerance. This is a worst-case bound on all unresolved bits, not a prediction of their average empirical score.

If $h_q(Z) \leq \varepsilon$, choose $C = [d] \setminus Z$ and a full-support curriculum. Then every signal on C is positive, yielding a finite sample size for any fixed $\delta > 0$. This connects the observability criterion to an explicit learner, but the resulting bound can be very large.

For each learner and coordinate, the two-sided bound is $\mathbb{P}(|\bar{Z}_j - \mathbb{E}Z_j| > r) \leq 2\exp(-Nr^2/2)$. Taking $r = r_N = \sqrt{2\log(4d/\delta)/N}$ and union-bounding over $2d$ moments gives simultaneous coverage at least $1 - \delta$. On that event, $|\bar{Z}_j| > r_N$ forces the empirical sign to equal the nonzero true sign. Requiring this for both learners on the same predetermined C therefore certifies target compatibility. In particular, a coordinate with zero true signal cannot be certified on the coverage event. The method may abstain even when the learned pair is compatible.

The event concerns two specified runs at a specified N. For P designated pairs and L predetermined checkpoints, assigning failure budgets summing to δ would yield a simultaneous statement. The fixed-pair statement alone does not supply that simultaneous guarantee. Arbitrary optional stopping requires a different confidence construction. Likewise, individually certifying different relevant sets is insufficient unless the target risk of their combined unresolved set is controlled.

H.3 TWO CURRICULUM OBJECTIVES AND A SEPARATION

For fixed C, the probe LP maximizes $\min_{j \in C} m_{\mu j}$ because all $m_{\mu j}$ are linear in μ. Enumeration over safely ignorable sets solves the corresponding finite cover-selection problem. This optimizes

the simple worst-coordinate sufficient bound, not the full sum of exponentials or exact finite-sample performance. No source or target rollout outcomes from the learned seeds are used to select the curriculum.

Proof of Proposition 14. The path visits all nodes, so $h_q(U) = 1$ for every nonempty U; all different conventions are a bad pair. Source failure cost is $F(E) = 1 - (1-c)^{|E|}$. Any distinct pair has at least one nonempty error set and thus maximum gap at least c. The preferred vector paired with a single-bit error attains c. For a uniform subset E of the other $d-1$ nodes,

$$F(E \cup \{j\}) - F(E) = c(1-c)^{|E|}, \quad m_j = c \prod_{k \neq j} \left[\tfrac{1}{2} + \tfrac{1}{2}(1-c)\right] = c(1-c/2)^{d-1}.$$

This establishes a separation between an exact compatibility margin and this learner's signal. It does not prove that the sufficient Hoeffding rate is necessary or that every learner has exponential complexity. □

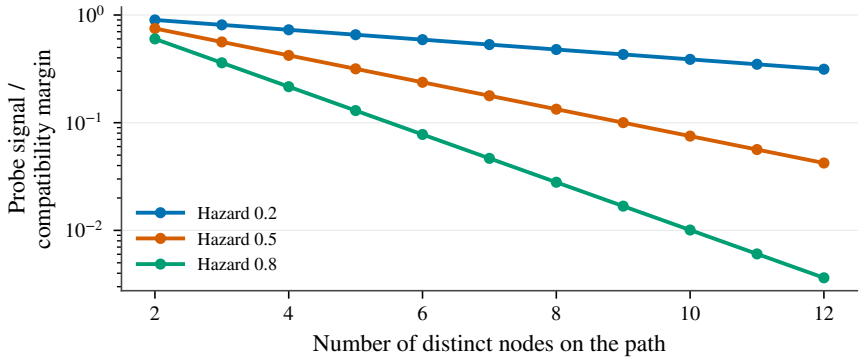


Figure 3: Exact signal-to-margin ratio in Proposition 14. The margin is constant in the number of distinct nodes, while the uniform-policy probing signal decreases geometrically. These are analytic values checked against exact evaluation, not fitted learning curves.

H.4 Independent-action local optimization

With one node and independent role actions, matching convention 0 occurs with probability $(1-p)^2$, matching convention 1 with probability p^2, and disagreement earns 0. Thus $f(p) = r_0(1-p)^2 + r_1 p^2$. Its derivative is $-2r_0 + 2(r_0 + r_1)p$. It is negative near 0 and positive near 1, so both boundaries are local maxima relative to $[0, 1]$, separated by an interior minimum. The preferred convention 0 is globally better when $r_0 > r_1$. For a logistic parameterization the boundaries are limiting deterministic policies; the reward-gradient direction still favors different limits on either side of the unstable interior point. We do not claim that every REINFORCE run converges or that finite-temperature stochastic exploration has identical behavior. This construction simply invalidates an unconditional inference from a positive margin to successful local optimization.

I Finite-class curriculum optimization

I.1 Reference-conditioned pair-sum linear programs

For each possible maximizing policy r, define $d_{ri}(\mu) = S_r(\mu) - S_i(\mu)$ and solve

$$\begin{aligned} \max_{\mu, t} \quad & t \\ \text{subject to} \quad & \mu \in \Delta_M, \quad d_{ri}(\mu) \geq 0 \quad \forall i, \\ & d_{ri}(\mu) + d_{rj}(\mu) \geq t \quad \forall (i,j) \in \mathcal{B}. \end{aligned} \tag{17}$$

Skip infeasible references and keep the largest objective. For a nonempty bad-pair set, a feasible reference exists because every curriculum has a maximizer. Nonnegative reference gaps are essential.

Theorem 18 (General pair-sum approximation)**.** *The curriculum returned by* (17) *has margin at least* $\Gamma^*/2$*, where* $\Gamma^* = \max_\mu \Gamma_A(\mu; \mathcal{B})$*. The factor is tight for this surrogate.*

Proof. For nonnegative x, y, $\max(x, y) \le x + y \le 2\max(x, y)$. Thus the optimized pair-sum objective is at least Γ^*, while the returned curriculum's margin is at least half its pair-sum value.

For sharpness use two source rows $(1, 0, 1)$ and $(1, 1/2 - h, 1/2 - h)$, $0 < h < 1/2$. Only policies 1 and 2 are incompatible; policy 0 is compatible with both. The first source alone has margin 1. If its mixture probability is p, the pair-sum value is $p + (1 - p)(1 + 2h)$, uniquely maximized by $p = 0$. The selected margin is $1/2 + h$, approaching $1/2$ as $h \downarrow 0$. Since no gap exceeds 1, the exact optimum is 1. This is not an inapproximability result for all curriculum algorithms. □

For the path family, Theorem 2 identifies the pair-sum objective with κ. Hence (6) attains half the optimal balanced margin. The general tight family is separate from the explicit path-family construction in Proposition 11; both claims have explicit constructions.

I.2 Exact mixed-integer oracle

For bounded return tables, an exact small-instance oracle uses $s \in [0, 1]$, $t \in [0, 1]$, maximizing-policy indicators $z_i \in \{0, 1\}$, and exclusion indicators $y_i \in \{0, 1\}$. Maximize t subject to

$$\begin{aligned} &\mu \in \Delta_M, \quad \sum_i z_i = 1, \quad s \ge S_i(\mu), \quad s \le S_i(\mu) + 1 - z_i && \forall i, \\ &y_i + y_j \ge 1 && \forall (i, j) \in \mathcal{B}, \\ &t \le s - S_i(\mu) + 1 - y_i && \forall i. \end{aligned}$$

The anchor constraints set $s = \max_i S_i$. Excluded policies must have gap at least t, and the exclusion set meets every bad pair. Thus feasible solutions have margin at least t. Conversely, for any curriculum, policies whose gaps are at least its margin form such a cover, giving equality of optima. A self-loop forces its single vertex to be excluded. An independent oracle enumerates maximizing references and vertex covers, solves the remaining LPs, and checks the mixed-integer linear program (MILP) on separate small cases. Empty bad-pair sets are handled separately because their margin is infinite.

The path program requires explicit evaluation of $M2^d$ subset losses. A dense matrix implementation costs $O(M2^d H d^2)$ for those dynamic programs. The balanced oracle enumerates $O(3^d)$ disjoint pairs. These are acceptable diagnostic costs for $d \le 6$ in learned experiments, not a scalable approach for a neural policy class. The source model and target relation must also be supplied or estimated; this information is not free in an unknown environment.

J Finite-data and perturbation proofs

J.1 Sharp rectangular return uncertainty

For a fixed finite policy class and simultaneous cell intervals $L \le A \le U$, put $\ell_i = \mu^T L_{:i}$ and $u_i = \mu^T U_{:i}$. For nonempty fixed $\mathcal{B}$,

$$\inf_{A \in [L, U]} \Gamma_A(\mu; \mathcal{B}) = \left[\max_k \ell_k - \max_{(i,j) \in \mathcal{B}} \min(u_i, u_j) \right]_+ . \tag{18}$$

This certificate concerns a finite rectangular table, not an arbitrary neural class. If the target or common competence returns are uncertain too, use simultaneous lower bounds to form a conservative bad-pair relation: include a pair when either lower source bound is below c or its lower target bound is below τ. On the simultaneous event this relation contains all actual bad pairs. Reusing that same event for curriculum selection is allowed; estimating a table for a post-selected policy class is not automatically covered.

Theorem 19 (Product-box certificate)**.** *For a nonempty bad-pair set and a fixed curriculum, equation* (18) *is the attained infimum over* $A \in [L, U]$.

Proof. For any score vector s, $\max(g_i, g_j) = \max_k s_k - \min(s_i, s_j)$. Let $z = \max_k \ell_k$. Since $s_k \geq \ell_k$ and $s_i \leq u_i$, every feasible vector has pair gap at least $[z - \min(u_i, u_j)]_+$. Minimizing over pairs gives the lower bound in (18).

Choose a pair (i, j) maximizing $c = \min(u_i, u_j)$. If $z \geq c$, set $s_i = \max(\ell_i, c)$, $s_j = \max(\ell_j, c)$, and every other score to its lower endpoint. These chosen pair values are feasible; at least one equals c because one pair upper endpoint equals c. All scores are at most z, and at least one lower endpoint equals z, so the pair gap is $z - c$. If $z < c$, set $s_i = s_j = c$ and the other scores to their lower endpoints. The pair is now maximizing and its gap is zero. For a self-loop take its one chosen score accordingly. This constructs a score vector attaining the lower bound.

Every chosen score $s_i \in [\ell_i, u_i]$ is attainable by a column of the cell box. If $u_i > \ell_i$, set $A_{ei} = L_{ei} + v_i(U_{ei} - L_{ei})$, with $v_i = (s_i - \ell_i)/(u_i - \ell_i) \in [0, 1]$. If the interval is degenerate, use its lower column. Columns are unconstrained relative to one another in a product box, so the whole witness is attainable. □

If returns must additionally satisfy a shared protocol model, this attaining table may not respect it. The formula is sharp for the stated uncertainty set, not for every stronger model-informed uncertainty set.

For predetermined sample counts $n_{ei} > 0$, bounded iid rewards within each cell yield simultaneous Hoeffding radii $b_{ei} = \sqrt{\log(2MK/\delta_A)/(2n_{ei})}$. Clip $\widehat{A}_{ei} \pm b_{ei}$ to $[0, 1]$. Independence across different cells is not required for the union bound. If a target lower-bound table L^T covers simultaneously with probability $1 - \delta_T$, set $\mathcal{B}^+ = \{(i, j) : L^T_{ij} < \tau\}$. On the joint event, every issued claim $\Gamma_\square(\mu; \mathcal{B}^+) > \eta$ is valid, including for any number of curricula selected from the same tables. No union bound over curricula is needed because coverage already holds for all table entries. This does not license unchecked adaptive sample counts, optional stopping, or new policies added after inspecting the data.

If empirical self-play scores are uniformly within b of their true values, an empirically η-optimal policy has true gap at most $\eta + 2b$. The certificate must be compared with that larger tolerance unless a separate bound on the true gap is available. A curriculum certificate alone does not certify the learner's achieved optimization accuracy.

J.2 A RESTRICTED CALIBRATION LOWER BOUND

Proposition 20. *Consider one source, known return $A_0 = 3/4$, and an unknown Bernoulli return $A_1 = 1/2 - h$ or $1/2 + h$, with $0 < h \leq 1/4$. Let target diagonal returns be 1, off-diagonal returns 0, $0 < \tau \leq 1$, and $\eta = 1/4$. If a procedure certifies with probability at least $1 - \delta$ in the first world and at most δ in the second, $0 < \delta < 1/2$, its predetermined number of observations of the unknown cell satisfies*

$$n \geq \frac{(1 - 2\delta)\log((1 - \delta)/\delta)}{16h^2}.$$

Proof. The first world has margin $1/4 + h > \eta$, while the second has margin $1/4 - h \leq \eta$. Let P, Q be the two observation distributions and D the event that the procedure certifies. Data processing for Kullback–Leibler (KL) divergence and monotonicity of binary KL on $P(D) \geq 1 - \delta$, $Q(D) \leq \delta$ give $\mathrm{KL}(P, Q) \geq \mathrm{kl}(1 - \delta, \delta) = (1 - 2\delta)\log((1 - \delta)/\delta)$. Independence gives $\mathrm{KL}(P, Q) = n\,\mathrm{kl}(1/2 - h, 1/2 + h)$. This Bernoulli KL is $2h\log((1 + 2h)/(1 - 2h)) \leq 16h^2$ for $h \leq 1/4$; the inequality follows from $\log((1 + x)/(1 - x)) \leq 2x/(1 - x)$ with $x = 2h$. Combining proves the bound. □

The proposition requires both validity and completeness near a threshold. It is not a lower bound on cooperation itself: two agents could agree on a fixed convention without learning which is self-play optimal. Its scope is deliberately narrower than a minimax theorem for all curricula or all learning methods.

J.3 Model perturbation and action-dependent transitions

The coordination-only transport implication is

$$\Gamma_{A^0}(\mu; \mathcal{B}_{\tau+\xi}(T^0)) > \eta + 2\beta \implies \text{all actual } \eta\text{-optimal pairs have target return at least } \tau. \quad (19)$$

It is not used to justify the action-steering experiment; that experiment evaluates the actual finite return table and a separate population evaluation.

Proposition 21. *Suppose* $|A_{ei} - A^0_{ei}| \le \beta$ *and* $|T_{ij} - T^0_{ij}| \le \xi$ *uniformly, for the same policy class. Then* (19) *holds. More generally, for nonempty relevant bad-pair sets,*

$$\Gamma_A(\mu; \mathcal{B}_\tau(T)) \ge \left[\Gamma_{A^0}(\mu; \mathcal{B}_{\tau+\xi}(T^0)) - 2\beta\right]_+ .$$

Proof. Curriculum scores and their maxima change by at most β. Thus gaps change by at most 2β, so every actual η-optimal policy is ideal $(\eta + 2\beta)$-optimal. For an actual bad pair, $T_{ij} < \tau$ implies $T^0_{ij} < \tau + \xi$. Applying the ideal margin to this larger bad-pair set proves both statements. If there are no actual bad pairs, compatibility is automatic.

For the kernel bound, couple the initial states identically. Whenever the two executions have the same history, couple their next transition and reward with disagreement probability at most λ. The probability they remain coupled for all H decisions is at least $(1-\lambda)^H$. A bounded terminal reward differs in expectation by at most the probability of a coupling failure, giving $\beta, \xi \le 1 - (1-\lambda)^H \le H\lambda$. This is the usual finite-horizon model-comparison argument; model-based reinforcement learning provides the broader context (Kearns & Singh, 2002). The uniform kernel assumption must hold for each joint action, including disagreements. □

The result is a sufficient robustness certificate, not a sharp transition-uncertainty theorem. With convention-dependent transitions, the exclusive-or (XOR) target formula need not hold; source-policy and target-pair returns must be evaluated under their actual transition laws.