STRATA: EQUILIBRIUM-RELATIVE STATE ABSTRACTIONS FOR MARKOV GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Which state abstractions of a Markov game preserve its equilibria? We answer with a hierarchy of three strategic-distortion criteria for N-player games: global model similarity (D0), model similarity under unilateral deviations from the played profile (D1), and best-response-value similarity (D2), and prove that at every level an approximate equilibrium of the abstract game lifts to an approximate Nash equilibrium of the ground game (at the model levels also coarse-correlated equilibria, with distortions under the device's conditional laws). The levels nest, and existing multi-agent abstraction criteria are profile-independent: the restriction D1 and D2 relax. We show where that restriction costs: in a constructed family where the played mixture cancels a variable's effect on the payoffs of played actions, every criterion that compares rewards or minimax values joint action by joint action (on all joint actions, the undominated ones or the minimax supports) keeps all of the variable's values, while D1 and D2 keep none; criteria that compare states through minimax strategies are not covered. Where discarding dominated actions suffices, it compresses more: on our two Markov Soccer games the prior value criterion restricted to undominated actions, a baseline we add, is lossless with fewer states than D2 (328 and 772). Because D1 and D2 are defined relative to the equilibrium being played, learning them is a fixed-point problem; we give a refinement loop that splits only the states that deviations from the current equilibrium tell apart, terminates by construction, and returns a joint (abstraction, equilibrium) fixed point at which the tolerance holds for the equilibrium actually played. Solving exactly lets us measure every claim against ground truth: on Markov Soccer, D0 cannot compress, D1 rediscovers the hand-designed abstraction, and D2 matches it with 15% fewer states, while the prior model criterion cannot compress either and the prior value criterion needs more states (805 vs. 649) at a $7.7\times$ larger duality gap; on the prior criterion's own benchmark, D2 is lossless (zero duality gap at every state) with fewer states (582 vs. 612), an advantage that holds on five of nine pitch sizes and reverses on four. One abstraction refined on three games, with $1.65\times$ the states of the hand design, matches or improves on it on every held-out game.

1 INTRODUCTION

Solving a Markov game gets harder with every state, so practitioners merge states. In a single-agent problem the merged states must agree on what matters to one decision maker; the conditions are well understood (Li et al., 2006; Abel et al., 2016). In a game the question changes character: *what matters* depends on how everyone else plays. A variable the opponents' equilibrium strategy never reacts to is irrelevant to me, even if it changes the kernel under joint actions nobody takes.

Existing criteria for Markov games do not use this freedom. Ishibashi et al. (2025) merge states whose rewards and transitions agree under *every joint action*, or whose minimax $Q^\star$ agrees under every joint action; both are, in their words, independent of any strategy profile. Comparing under every joint action is what stops them from discarding a variable that only dominated, unplayed actions react to (Proposition 6). Restricting the comparison to undominated actions removes that obstacle without reference to any profile, and on Markov Soccer it compresses more than our criteria do (§5.2). What no restriction of this kind removes is a variable that changes the payoffs of actions equilibrium play *does* use, in a way the played mixture cancels: every criterion that compares re-

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.

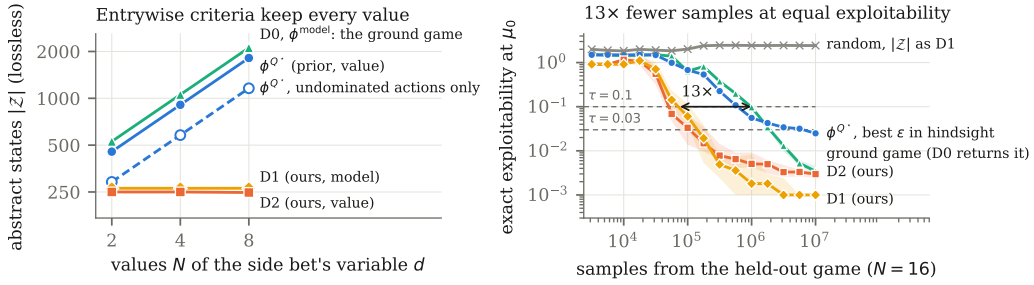


Figure 1: *Left:* lossless abstract states on the side-bet family H_N of Proposition 7 (plain 3×4 soccer core; the side bet's stakes depend on a variable d with N values; exact, §5.2). On G_N (Proposition 6) the undominated-action criterion is flat instead (Appendix H). *Right:* exact exploitability on a held-out game known only through samples, for partitions learned on a training game (G_N, $N = 16$; median and interquartile band, 20 seeds; values below 10^{-3} drawn at 10^{-3}; dashed: the tolerances fixed in advance; §5.3).

wards or minimax values joint action by joint action must then keep all of its values, while D1 and D2 keep none (Proposition 7). The result is a map, not a ranking. On the soccer games dominance does most of the work: the undominated-action criterion is lossless with fewer states than D1 and D2 (against D2, 328 vs. 582 and 772 vs. 1,243; §5.2); on Oshi-Zumo no criterion dominates (Appendix J). Profile-relativity is needed for a variable whose effect the played mixture cancels: it must be compared through strategies rather than joint actions, which D1 and D2 do with the profile played and, in two-player zero-sum games, the opponents' minimax strategies would do too. When there is no such variable, a profile-relative criterion still never needs more states than $\phi^{Q^\star}$ for a lossless abstraction of a two-player zero-sum game (Proposition 5), but finding them is a search problem that our algorithms solve in most cases, not all (§5.2).

Contributions. The picture is two-sided: profile-relativity buys what no criterion comparing states *entrywise over joint actions* can, and is *not* what compresses best where dominated actions can simply be discarded. (1) A hierarchy of strategic distortions for N-player general-sum Markov games along two axes, *level* (model vs. best-response value) and *profile-relativity*, with equilibrium-transfer theorems for Nash and coarse-correlated equilibria; nesting results that make every arrow between prior criteria and ours a relaxation; and two separations: criteria that compare states under every joint action must keep every value of a variable that only dominated actions react to, a factor 2^k for k such bits; and every criterion that compares rewards or minimax values joint action by joint action, restricted to undominated actions or not, must keep every value of a variable whose effect on played actions the equilibrium mixture cancels, while D1 and D2 keep none (§3). (2) Because profile-relative criteria depend on the equilibrium that the abstraction itself induces, learning them is a fixed-point problem. We give a monotone refinement loop that splits only the states that deviations from the current equilibrium tell apart, terminates by construction, and returns a joint (abstraction, equilibrium) fixed point (§4). (3) Exact verification: every claim of §5 is measured against exact exploitability, the transfer bounds included. Theorems 1–2 are admissibility conditions, not estimates: on our games they lie $26\times$ to $10^5\times$ above the exact values (§7). The occupancy-weighted distortion is 8–$17\times$ smaller than ε_Q (Appendix C), a measurement of what a bound using it could gain; we do not prove one. (4) Transfer: one partition refined on three games is less exploitable than the hand design on every held-out game with $\delta > 0$, at $1.65\times$ its size (§5.4); on a held-out game known only through samples, a learned partition that drops a constructed irrelevant variable cuts the samples a ground-model learner needs by a factor that grows with the variable, $13\times$ at 16 values on G_N, while on H_N the ground model needs fewer samples than the loop's partitions (§5.3, Figure 1); the loop's partitions transfer to the changed core, the dominance-aware ones, which drop the variable too, do not.

Why exact. All experiments in §5 solve the games exactly (Shapley value iteration with one linear program per state and sweep, exact best responses, exact duality gaps): the only setting in which a bracket on exploitability can be verified end to end, and what separates our comparison from

prior work, whose gaps are estimated from minimax Q-learning. Appendix K documents a deep instantiation, on which no claim of the paper rests: the profile-relative loop did not carry over to function approximation.

2 PRELIMINARIES

An N-player discounted Markov game is $G = (N, \mathcal{S}, \{\mathcal{A}_i\}, P, \{r_i\}, \gamma, \mu_0)$ with finite $\mathcal{S}$, rewards in an interval of length R, and $V_{\max} = R/(1-\gamma)$ the span of attainable values ($R = 1$ gives the usual normalization; our soccer games have $R \approx 2$). For a profile π_{-i} of the others, player i faces an MDP $M_i(\pi_{-i})$ with reward $r_i^{\pi_{-i}}(s, a_i) = \mathbb{E}_{a_{-i}\sim\pi_{-i}(s)} r_i(s, a_i, a_{-i})$ and kernel $P^{\pi_{-i}}$ defined likewise; $Q_i^\star(s, a_i \mid \pi_{-i})$ is its optimal action value. Exploitability is $\mathrm{EXP}_G(\pi) = \sum_i \big(V_i^\star(\mu_0 \mid \pi_{-i}) - V_i^\pi(\mu_0)\big)$. An abstraction is a surjection $\phi : \mathcal{S} \to \mathcal{Z}$ with a weighting $w(\cdot \mid z)$ on each class (uniform here); it induces the abstract game $\hat{G}_\phi$ by w-averaging rewards and pushing transitions forward. Appendix M collects the notation. An abstract profile $\hat{\pi}$ *lifts* to $\pi = \hat{\pi} \circ \phi$. Deviators in the ground game are *not* restricted to be ϕ-measurable; that is the content of the transfer theorems. A "player" need not be a single agent. A group that shares a reward and acts on a joint decision is one player over the product of its members' actions, so two competing teams are a two-player game; a team whose members cannot coordinate is a set of players with identical rewards, and its Nash equilibria are those studied for adversarial team Markov games (Kalogiannis et al., 2023). Theorems 1–2 are stated for N players and general sums and cover both readings; the predator–prey environment of Appendix K is of the second kind.

3 A HIERARCHY OF STRATEGIC DISTORTIONS

Definition 1 (Strategic distortions). *Write $s \sim_\phi s'$ for $\phi(s) = \phi(s')$. For a profile π with ϕ-measurable components,*

$$\textbf{\textit{D0:}}\quad \varepsilon_r^0 = \max_{i,a,s\sim_\phi s'} |r_i(s,a) - r_i(s',a)|, \qquad \varepsilon_P^0 = \max_{a,s\sim_\phi s'} \mathrm{TV}\big(\phi_\# P(\cdot|s,a), \phi_\# P(\cdot|s',a)\big),$$

$$\textbf{\textit{D1:}}\quad \varepsilon_r^1(\pi),\ \varepsilon_P^1(\pi):\ \textit{the same with } r_i^{\pi_{-i}}, P^{\pi_{-i}} \textit{ and the maximum over player } i\textit{'s} \text{ own } \textit{actions } a_i,$$

$$\textbf{\textit{D2:}}\quad \varepsilon_Q(\pi) = \max_{i,a_i,s\sim_\phi s'} \big|Q_i^\star(s, a_i|\pi_{-i}) - Q_i^\star(s', a_i|\pi_{-i})\big|.$$

D0 quantifies over joint actions and knows nothing about play. D1 and D2 quantify over a single player's deviations *against the profile being played.*

Two axes. Table 1 places prior criteria. D0 is the total-variation form of the same profile-independent model-level criterion as Ishibashi et al. (2025): their condition bounds the largest *coordinate* of the pushforward transition difference and ours its total variation, so the two coincide at zero distortion and at positive tolerance differ by the norm conversion $\|p - q\|_\infty \le \mathrm{TV}(p, q) \le \frac{|\mathcal{Z}|}{2}\|p - q\|_\infty$ (their condition at tolerance ε need not give D0 at ε; the implementation in the released code is faithful to their coordinatewise form). Their $Q^\star$ criterion is value-level but profile-independent: it compares minimax $Q^\star(s, \boldsymbol{a})$ over all joint actions. D1 and D2 occupy the profile-relative row.

Table 1: Criteria by level and profile-relativity. Prior work occupies the top row.

	model level	value level
profile-independent	ϕ^{model} (Ishibashi et al., 2025); D0 is its TV form	$\phi^{Q^\star}$ (Ishibashi et al., 2025)
profile-relative	**D1**	**D2**

Theorem 1 (Model-level transfer). *Let $\hat{\pi}$ be an $\hat{\varepsilon}$-NE of $\hat{G}_\phi$ in sup-norm, $\pi = \hat{\pi} \circ \phi$ and $\Delta_1 = \big(\varepsilon_r^1(\pi) + \gamma\varepsilon_P^1(\pi)V_{\max}\big)/(1-\gamma)$. Then π is an $(\hat{\varepsilon} + 2\Delta_1)$-NE of G, and $\mathrm{EXP}_G(\pi) \le \mathrm{EXP}_{\hat{G}_\phi}(\hat{\pi}) + 2N\Delta_1$. For an $\hat{\varepsilon}$-coarse-correlated equilibrium with a stationary ϕ-measurable device $\hat{\mu}(\cdot|z)$ the same holds once $\varepsilon_r^1, \varepsilon_P^1$ are maximized over both laws of the others that occur: the marginal $\hat{\mu}_{-i}(\cdot|z)$*

a deviator faces and the conditionals $\hat{\mu}(\cdot|a_i, z)$, $\hat{\mu}(a_i|z) > 0$, faced on path. With the marginal alone the statement is false (Appendix A, Remark 1).

Theorem 2 (Value-level transfer)**.** *Under the same hypotheses, for any weighting w and all i, s: $V_i^\star(s|\pi_{-i}) - V_i^\pi(s) \le \hat{\varepsilon}/(1-\gamma) + 2\varepsilon_Q(\pi)/(1-\gamma)^2$.*

Proposition 3 (Nesting)**.** *$\varepsilon_r^1 \le \varepsilon_r^0$, $\varepsilon_P^1 \le \varepsilon_P^0$ and $\varepsilon_Q \le 2\Delta_1$; hence an abstraction within tolerance ε at D0 is within ε at D1, and one within $(\varepsilon, \varepsilon)$ at D1 is within $2\Delta_1$ at D2.*

The second inclusion is coarse ($2\Delta_1 = 2\varepsilon(1 + \gamma V_{\max})/(1-\gamma)$ is 200ε at $\gamma = 0.9$, $R = 1$): a statement about order, not a usable conversion. The partitions the loop finds at D2 have $\varepsilon_P^1 = 1$ (Table 8): D2 merges states a model criterion must separate.

Corollary 4 (Two-sided bracket)**.** *For any ground deviators $\{\psi_i\}$: $\sum_i \big(V_i^{(\psi_i, \pi_{-i})}(\mu_0) - V_i^\pi(\mu_0)\big) \le \mathrm{EXP}_G(\pi) \le \min\{$Thm. 1, $N \cdot$ Thm. 2$\}$, with equality on the left for exact best responses. The Theorem 1 branch holds for any abstract profile; the Theorem 2 branch needs $\hat{\pi}$ to be an $\hat{\varepsilon}$-NE in sup-norm.*

The left side is the definition of exploitability and the right side sums Theorems 1–2; wherever the right side can be computed, so can $\mathrm{EXP}_G(\pi)$ itself. Its use is the left side: deviators give a lower bound where nothing is exact (Appendix K).

Proposition 5 (Value-level nesting)**.** *Let G be two-player zero-sum with minimax action values $Q^\star$, let $\varepsilon_{Q^\star}(\phi) = \max_{s \sim_\phi s', \boldsymbol{a}} |Q^\star(s, \boldsymbol{a}) - Q^\star(s', \boldsymbol{a})|$ be the distortion of the criterion of Ishibashi et al. (2025), and let π be any ϕ-measurable profile with duality gap $\mathrm{GAP}(\pi) = \max_s \sum_i \big(V_i^\star(s|\pi_{-i}) - V_i^\pi(s)\big)$. Then $\varepsilon_Q(\phi; \pi) \le \varepsilon_{Q^\star}(\phi) + \gamma\,\mathrm{GAP}(\pi)$. In particular a partition that is admissible for $\phi^{Q^\star}$ at tolerance ε and whose lifted equilibrium is exact is admissible for D2 at ε.*

With Proposition 3 this makes every arrow of Table 1 a relaxation: D0→D1 and D1→D2 in general, $\phi^{Q^\star}$ →D2 in two-player zero-sum games, where $\phi^{Q^\star}$ is defined. As a *criterion*, D2 never needs more states than $\phi^{Q^\star}$ for a lossless abstraction; whether a *search* finds them is a separate question, which §5.2 shows is not idle.

Proposition 6 (Succinctness separation)**.** *Let G_0 be a two-player zero-sum Markov game with m states in which each player has an action σ with $r(\cdot, \sigma, \sigma)$ constant, and let $c_1, \ldots, c_N$ be states of G_0 with pairwise distinct minimax values, $\Delta_{\min} = \min_{j \ne k} |V_0^\star(c_j) - V_0^\star(c_k)|$. Let G_N have states (c, d), $d \in [N]$, where d follows a Markov chain independent of play and each player gains one action τ that behaves as σ except that whoever plays it pays $\kappa > 0$ to the other and the joint action (τ, τ) sends the core state to c_d. Then: (i) no Markov perfect equilibrium of G_N plays τ, $V^\star(c, d) = V_0^\star(c)$, and $\phi(c, d) = c$ has $\varepsilon_r^1 = \varepsilon_P^1 = \varepsilon_Q = 0$ at every lifted Markov perfect equilibrium of its abstract game, so m abstract states suffice for D1 and D2, losslessly, for every N; (ii) every partition with $\varepsilon_{Q^\star}(\phi) < \gamma\Delta_{\min}$ keeps each class inside one value of d; (iii) so does every partition with $\varepsilon_r^0 = \varepsilon_P^0 = 0$, i.e. every exact bisimulation, and hence the model criterion of Ishibashi et al. (2025) at tolerance 0. Exact bisimulation, the model criteria, and $\phi^{Q^\star}$ at tolerances below $\gamma\Delta_{\min}$ therefore need at least N times the states they need for one value of d (a factor 2^k for k strategically irrelevant bits), while D1 and D2 need none of them.*

The mechanism is dominance: τ is iteratively strictly dominated, so a criterion that compares states only on the joint actions surviving iterated elimination of strictly dominated actions (IESDS) in the stage game $Q^\star(s, \cdot, \cdot)$ (which references no profile, as $Q^\star$ is unique) drops d too, and does (§5.2). In the next family the variable changes the payoffs of actions that equilibrium play *uses*, and only the mixture they are played with cancels the change.

Proposition 7 (Separation from entrywise criteria)**.** *Let G_0 be a two-player zero-sum Markov game with m states, $\Delta > 0$, and $B_1, \ldots, B_N$ distinct real $n \times n$ matrices with zero row and column sums, each of which, and whose mean, has the uniform strategies as its only equilibrium; $\beta = \min_{d \ne d'} \|B_d - B_{d'}\|_\infty$. Let H_N have states (c, d), d following a Markov chain independent of play; each player picks a core action and a side action $b \in [n]$, the row player receives $\Delta B_d(b_1, b_2)$ on top of the core reward, and side actions do not affect transitions. Then: (i) $V^\star(c, d) = V_0^\star(c)$, every Markov perfect equilibrium plays the side actions uniformly, and $\phi(c, d) = c$ with uniform weights has $\varepsilon_r^1 = \varepsilon_P^1 = \varepsilon_Q = 0$ at every lifted Markov perfect equilibrium of its abstract game, which is an exact equilibrium of H_N; (ii) let $K(s)$ contain, at every state, each joint action $(\boldsymbol{a}_1, \boldsymbol{a}_2)$ such*

that each $\boldsymbol{a}_i$ is played by some minimax strategy of player i in the stage game $Q^\star(s,\cdot,\cdot)$; e.g. all joint actions, the IESDS survivors, or the product of the minimax supports. For $f \in \{r, Q^\star\}$, every partition with $\max_{s\sim_\phi s'} \max_{\boldsymbol{a}\in K(s)} |f(s,\boldsymbol{a}) - f(s',\boldsymbol{a})| < \Delta\beta/2$ keeps each class inside one value of d. So D0, the model criterion, $\phi^{Q^\star}$, and each of them restricted to IESDS survivors or to minimax supports, at tolerances below $\Delta\beta/2$, need at least N times the states they need for one value of d, while D1 and D2 need none of them.

What the separation buys: samples. Suppose the ground model is unknown and only a generative model is available. Drawing s uniformly from a class $\phi^{-1}(z)$ and then $s' \sim P(\cdot|s,\boldsymbol{a})$, and recording $\phi(s')$, is *exactly* a generative model of the abstract game $\hat{G}_\phi$ with uniform weights. Any generative-model result for two-player discounted zero-sum Markov games therefore applies to $\hat{G}_\phi$ with $|\mathcal{Z}|$ in place of $|\mathcal{S}|$, and Theorem 1 carries its output back to G:

Proposition 8 (Samples, by composition)**.** *Let G be two-player zero-sum, let the reward function be* known *and ϕ-measurable, and let ϕ be fixed independently of the samples below. For $\varepsilon \le (1-\gamma)^{-1/2}$, with probability $1-\delta$, the model-based approach of Zhang et al. (2020) with a smooth planning oracle of planning error $O((1-\gamma)\varepsilon)$ returns an ε-equilibrium of $\hat{G}_\phi$, at every abstract state, from $\tilde{O}\big(|\mathcal{Z}||\mathcal{A}||\mathcal{B}|(1-\gamma)^{-3}\varepsilon^{-2}\big)$ samples (their Theorem 3.5); its lift is an $(\varepsilon + 2\Delta_1)$-equilibrium of G, Δ_1 measured at the lifted policies. For the abstraction of Proposition 6, $|\mathcal{Z}| = |\mathcal{S}|/N$ and $\Delta_1 = \gamma q V_{\max}/(1-\gamma)$, q the largest probability the learned policies put on τ: if they never play τ, as no exact equilibrium of $\hat{G}_\phi$ does, this upper bound on the samples falls by the factor N, which criteria comparing every joint action keep in theirs.*

A composition of a published bound with Theorem 1, not a new analysis: rescaling our rewards to their $[0,1]$ costs a constant; an exact solver need not be smooth, and for it their Corollary 3.3 gives $(1-\gamma)^{-5}$; an ε-equilibrium may play τ. §5.3 measures the factor directly.

Proofs are in Appendix A. With $N=1$ and $\hat{\varepsilon}=0$, Theorem 2 is Lemma 1 of Abel et al. (2016), their bound for approximate $Q^\star$-irrelevance, and Theorem 1 an approximate model-irrelevance bound in total variation; with zero distortions both reduce to the exact results of Li et al. (2006).

What the constants are for. The upper bounds are admissibility conditions: they say which level of the hierarchy licenses a given abstraction, and are not estimates of exploitability; Theorem 1 is 26–50$\times$ above the exact value on the hand-designed partitions and 10^2–$10^5\times$ elsewhere, Theorem 2 looser still (§5); the operative side of the bracket is the deviator lower bound. Constants against those of Ishibashi et al. (2025): Remark 2.

4 Learning a profile-relative abstraction is a fixed-point problem

D1 and D2 are defined relative to π, and π is the equilibrium of the game the abstraction induces. Partitioning against a fixed reference profile and re-solving once does not keep the tolerance: the re-solved equilibrium exposes new within-class differences, so the tolerance no longer holds for the profile actually played. On the distractor game, partitioning once against the exact ground equilibrium at $\varepsilon = 0.1$ gives 390 classes whose re-solved equilibrium has $\varepsilon_Q = 1.57$ and exact exploitability 0.336 (1.57 and 0.368 after three re-partitionings from scratch). The loop below returns 649 classes with $\varepsilon_Q = 0.097 \le \varepsilon$ and exploitability 0.058 (Table 4, Appendix B).

Monotone refinement. Start from any ϕ_0 (we use the one-state partition). Repeat: solve $\hat{G}_{\phi_t}$; lift; compute exact best responses to the lifted profile; split exactly those classes that contain $s \sim s'$ with $|Q_i^\star(s,a_i|\pi_{t,-i}) - Q_i^\star(s',a_i|\pi_{t,-i})| > \varepsilon$ (D2; analogously for D1). Partitions only refine and $\mathcal{S}$ is finite, so the loop terminates, and at termination $\varepsilon_Q(\phi_T;\pi_T) \le \varepsilon$ holds *for the equilibrium of the final abstraction*: (ϕ_T,π_T) is a joint fixed point to which Theorem 2 applies. Every run stopped within 11 iterations (soccer: D1 3, D2 6–11; three-player games: 2–5). Termination is cheap: the identity partition is always a fixed point, the limit depends on the solver's equilibrium selection and on the split order (§5.2), and we do not claim it is the coarsest admissible partition.

5 EXPERIMENTS: EXACT VERIFICATION

Setup. Littman's 4×5 Markov Soccer (Littman, 1994), $\gamma = 0.9$. The *plain* game (760 states, 5 actions) is the benchmark of Ishibashi et al. (2025). Our *distractor* game (1,520 states, 6 actions) adds a costly TAUNT action and a binary variable d that changes transitions only under (TAUNT,TAUNT) and adds a wind cost δ when $d = 1$: payoff-relevant to a model criterion, nearly irrelevant to equilibrium play. TAUNT is iteratively, not strictly, dominated (against a taunting opponent it can pay for the player the joint effect favors) and never played in equilibrium: the player it would hurt strictly prefers to stay, and then the other's taunt is a pure cost.

5.1 THE HIERARCHY IN NUMBERS

Table 2: Automatic granularity from the one-state partition, distractor game, $\delta = 0.05$. Exact exploitability. The hand design drops d: $|\mathcal{Z}| = 760$, EXP $= 0.0582$.

level	ε	$\lvert\mathcal{Z}\rvert$	EXP (exact)	
D0	0.1 / 0.3	1520	0.0000	cannot compress
D1	0.1 / 0.3	760	0.0582	rediscovers the hand design exactly
D2	0.05	1250	0.0120	$4.9\times$ lower exploitability at $1.6\times$ the states
D2	0.1	649	0.0582	same exploitability, 15% fewer states
D2	0.2 / 0.3 / 0.5	548 / 537 / 436	0.092 / 0.174 / 0.287	compression–exploitability curve

The bracket against ground truth. For the hand design across $\delta \in [0, 0.2]$, Theorem 1's bound (40δ) is never violated and sits 31–50$\times$ above the exact exploitability; Theorem 2's bound is 7.5 at $\delta = 0.01$ against an exact 0.011, 37 at $\delta = 0.05$, and from $\delta = 0.1$ on exceeds the largest attainable exploitability ($NV_{\max} \approx 40$). Deviators truncated to 10 value-iteration sweeps recover 88–94% of the exact exploitability for $\delta \geq 0.05$ but only 27–38% for $\delta \leq 0.02$; 3 sweeps give negative, useless lower bounds at every δ.

5.2 HEAD-TO-HEAD WITH PROFILE-INDEPENDENT CRITERIA

Table 3: Exact results under both measures: *gap* is the duality gap of Ishibashi et al. (2025), $\max_s$ over all states of the two players' summed best-response gains; EXP is the same quantity at the initial distribution μ_0. *Left:* the prior work's own benchmark. *Right:* the distractor game. *Bottom:* the dominance-aware baseline; "0" is 10^{-9}.

plain game, 760 states				distractor game, 1520 states			
criterion (ε)	$\lvert\mathcal{Z}\rvert$	gap	EXP	criterion (ε)	$\lvert\mathcal{Z}\rvert$	gap	EXP
ϕ^{model} (any ≤ 0.4)	760	0	0	ϕ^{model}, D0 (any ≤ 0.4)	1520	0	0
D1 (any ≤ 0.2)	760	0	0	D1 (0.05)	1520	0	0
$\phi^{Q^\star}$ (≤ 0.02)	636	0	0	$\phi^{Q^\star}$ (≤ 0.02)	1364	0	0
$\phi^{Q^\star}$ (0.03)	612	0	0	D2 (0.02)	1243	0	0
D2 (0.13)	582	0	0	$\phi^{Q^\star}$ (0.03)	1324	0.023	0
$\phi^{Q^\star}$ (0.05)	588	0.172	0	$\phi^{Q^\star}$ (0.05)	1249	0.115	0
D2 (0.15)	523	0.063	0	$\phi^{Q^\star}$ (0.1)	805	0.565	0.145
$\phi^{Q^\star}$ (0.1)	378	0.957	0.107	D1 (0.1)	760	0.073	0.058
D2 (0.2)	469	0.130	0.127	D2 (0.1)	649	0.073	0.058
profile-independent, on the IESDS survivors of $Q^\star(s,\cdot,\cdot)$ only (baseline added here)							
D0 (any ≤ 0.1)	390	0	0	D0 (any ≤ 0.1)	780	0	0
$\phi^{Q^\star}$ (0)	328	0	0	$\phi^{Q^\star}$ (0)	772	0	0
$\phi^{Q^\star}$ (0.01)	254	0.003	0.002	$\phi^{Q^\star}$ (0.01)	416	0.011	0.007
$\phi^{Q^\star}$ (0.05)	184	0.095	0.085	$\phi^{Q^\star}$ (0.03)	322	0.067	0.024

Table 3 and Figure 2 read along the two axes of Table 1. *Against the prior criteria, on these two games, within the value level and at light compression, the profile-relative criterion needs fewer*

states for the same loss: lossless at every state, 582 vs. 612 on the plain game and 1,243 vs. 1,364 on the distractor game (one common grid; Appendix I); at gap 0.11 or below on the distractor game, 649 vs. 1,324. *Across levels, value beats model*: on the plain game $\phi^{Q^\star}$ compresses where neither model-level criterion can, ours included; on the distractor game D1 removes d while both prior model criteria remain at the identity. The signs survive five random greedy orders (Appendix E).

A profile-independent criterion that ignores dominated actions beats both. Compared state by state only on the joint actions that survive IESDS in the stage game $Q^\star(s,\cdot,\cdot)$, $\phi^{Q^\star}$ is lossless with 328 states on the plain game and 772 on the distractor game, where IESDS removes TAUNT at every state (Appendix I: minimax supports, and D0). At matched gap D2 is ahead of the IESDS-restricted criterion at none of the 21 common tolerances on either game (against 17 and 20 of 21 for the unrestricted one). The difference is D2's criterion, not only its search: at the restricted partition's lifted equilibrium ε_Q is 1.49 on the plain game and 1.54 on the distractor game, all of it on the values of dominated own actions, which D2 compares and the restricted criterion does not ($\leq 10^{-10}$ over the surviving ones). On Markov Soccer, discarding dominated actions compresses more than profile-relativity does.

Two refinements of D2 do not change that: comparing only near-best deviations keeps Theorem 2's bound (Proposition 10, Appendix A) and gives lossless partitions never larger than D2's, smaller on seven of the ten soccer games, yet larger than the dominance-aware ones on all ten; one pass of D2's distance at the exact ground equilibrium, without the loop, is lossless on only five of the ten (Appendix I). On Oshi-Zumo, a game we did not design (Appendix J), no criterion dominates: the smallest lossless partitions are $\phi^{Q^\star}$'s on minimax supports (259 of 406 states) and one pass of D2 (626 of 848), and D2u beats the IESDS-restricted $\phi^{Q^\star}$ on the larger game (678 against 690); every margin is 1–10%.

The two separations, checked exactly (Appendix H). On G_N (two soccer pitches and 20 random zero-sum cores, N up to 16, 87 games), no equilibrium plays TAUNT, $V^\star(c,d) = V_0^\star(c)$ to 10^{-15}, dropping d has $\varepsilon_r^1 = \varepsilon_P^1 = \varepsilon_Q = 0$ while $\varepsilon_P^0 = 1$, no class of any D0, ϕ^{model} or $\phi^{Q^\star}$ partition built mixes two values of d, their sizes are linear in N, and the D1 and D2 loops are flat and lossless. Compared only on IESDS survivors or minimax supports, however, $\phi^{Q^\star}$ and D0 drop d losslessly on all 87, with fewer states than either loop (145 and 87 against D1's 264 on 3×4): G_N separates criteria that compare every joint action from the rest. On H_N, 20 random cores $\times\ N \in \{2,4,8,16\}$ and the plain 3×4 soccer core $\times\ N \in \{2,4,8\}$ (Figure 1, left), no class of any of 14 profile-independent partitions (D0, ϕ^{model}, $\phi^{Q^\star}$, and $\phi^{Q^\star}$ and D0 on IESDS survivors or minimax supports, at tolerance 0 and 0.05) contains two values of d, and their sizes grow in proportion to N, while the D1 and D2 loops return at most the core, losslessly, in all 83 games. Both families are constructed: neither shows that such variables occur naturally.

Other pitch sizes: criterion against search. On eight further pitch sizes, on one common grid of 21 tolerances, the advantage of D2 over $\phi^{Q^\star}$ reverses with the parity of the pitch: D2 is lossless with fewer states on the four with a two-cell goal (even rows, as the benchmark), $\phi^{Q^\star}$ on four of the five with a one-cell goal; 4–5 at μ_0 alone (Table 15, Appendix I). The criterion does not exclude what D2 fails to find: every lossless $\phi^{Q^\star}$ partition evaluated is D2-admissible at the same tolerance (342 of 342; Proposition 5 gives this when the lifted profile is ground-exact, at a cost $\gamma\,\mathrm{GAP}(\pi)$ otherwise), and refinement never undoes a split. The shortfall is in the search, though inclusion alone does not quantify it. Given the *same* registered search (greedy merges proposed by each criterion's distance, kept only if the exact gap stays zero, 400 oracle calls), D2's distance wins 11 of 20 cells, not the 14 registered (Table 18). That oracle admits partitions far below both criteria (30 states on 3×4 against 201 and 207); $\pi^\star$-irrelevance lifted to the game is lossless on none of ten games.

5.3 Does the abstraction pay for itself? Samples

Every partition above was computed with the model in hand, which costs as much as solving the game; an abstraction pays only if it is learned where a model is available and used where it is not. *Train:* G_N on the 3×4 pitch, where each criterion computes its partition with full model access. *Test:* the same state space with a different core game (the goal on another row, every move slipping to STAY with probability 0.2, faster dynamics for d), known only through a generative model. A

learner pools the samples within its classes, solves the estimated abstract game exactly, lifts the equilibrium, and is scored exactly in the true test game. All learners see the same sample stream, and the prior criterion gets its best tolerance in hindsight at every sample size. Protocol, hypotheses and falsifiers were fixed before the script was written (Appendix G).

Result (Figure 1, right). At $N = 16$, with 20 seeds, the ground-model learner needs 9.8×10^5 samples to bring the median exploitability to 0.1 and 2.0×10^6 to bring it to 0.03; D1 needs 7.6×10^4 and 1.5×10^5. The ratio is 13 at both tolerances (95% bootstrap intervals 9.2–17.4 and 10.6–16.4), whichever default the ground-model learner uses. The prior criterion, with its tolerance chosen in hindsight at every sample size, needs $8\times$ and $43\times$ D1's samples: the partitions it learned on the training game are no longer admissible on the new core. The ratio is the separation of Proposition 6 measured in samples: over $N = 2, 4, 8, 16$ it is 1.6, 2.9, 5.8, 13.0; the ground model's requirement grows with log-log slope 0.92 in N and D1's is flat (-0.09), both inside the bands fixed in advance. It is the price of keeping d, a constructed variable. The dominance-aware criteria drop d on G_N too (§5.2) but do not realize it: they also merge core states by the training game's values, which the new core no longer shares, so even with the exact test model their partitions leave an exploitability of 0.79–1.41, and they never reach $\tau = 0.1$; nor does one pass of D2 at the training equilibrium (1.74). The loop's partitions do: D2u needs $1.1\times$ D1's samples (Appendix G). On the side-bet family H_N ($N = 4$, 10 seeds, the same protocol) the order reverses: the ground model reaches $\tau = 0.1$ after 5.4×10^6 samples and $\phi^{Q^\star}$ after 2.9×10^6, while D1, D2 and D2u, whose partitions are exact on the test game up to 10^{-5}, do not within 10^7 (median 0.22; a registered test rules out the estimator's within-class weighting as the cause, Appendix G).

5.4 Transfer across games

One D2 partition refined jointly on $\delta \in \{0.02, 0.05, 0.1\}$ and evaluated on held-out games by re-solving the abstract game only is less exploitable than the hand design on every held-out game with $\delta > 0$ (0.039 against 0.159 at $\delta = 0.2$, outside the training range), but has $1.65\times$ its states (1,253 against 760): one partition can serve a family of games; it is not a smaller one (Appendix F).

5.5 Beyond two-player zero-sum: three players, general-sum

The transfer inequalities hold for *any* abstract profile $\hat{\pi}$ with $\hat{\varepsilon}$ its measured abstract gap, so they can be checked exactly without a general-sum equilibrium solver. On random three-player general-sum games with the same distractor construction (40 states, 27 joint actions; 8 games $\times$ 4 values of δ), Theorems 1–2 and Proposition 3 hold on all 768 (abstraction, profile) pairs, 640 of them random non-equilibrium profiles; D0 refinement returns the identity, D1 refinement recovers the 20-state hand design in 21 of 24 games with $\delta > 0$, and D2 does *not* reproduce its two-player advantage (Appendix L).

6 Related work

State abstraction with bounds, one agent. Li et al. (2006) order exact abstractions of MDPs from model- to policy-irrelevance; Abel et al. (2016) relax model- and $Q^\star$-irrelevance with value-loss bounds; bisimulation (Givan et al., 2003) is the model end and its metrics (Ferns et al., 2004) relax it. With $N = 1$, D2 is their approximate-$Q^\star$ criterion (§3); D1 is approximate model-irrelevance in total variation, where their model criterion bounds each abstract state's probability separately. The π-bisimulation of Castro (2020) also measures error under the behavior played, averaging under a fixed policy, and DBC (Zhang et al., 2021) couples such a metric with an improving policy; D1 and D2 average only over the others' play, keep the maximum over own actions, as equilibrium needs, and follow the abstract game's equilibrium.

Abstraction of stochastic and Markov games. Sandholm & Singh (2012) bound the regret of an abstract equilibrium of a stochastic game in terms of reward and transition errors of a state-and-action abstraction; Ishibashi et al. (2025) give model- and $Q^\star$-based criteria for two-player zero-sum Markov games with duality-gap bounds. Both measure error over joint actions, independently of the strategy being played: the top row of Table 1. Refinement loops in games and the payoff-relevant state of Markov perfect equilibria: Appendix B. In extensive-form games, abstraction with solution-

quality bounds (Kroer & Sandholm, 2014) and the non-monotonicity of finer abstractions (Waugh et al., 2009) are closest; the latter is one reason we evaluate abstractions by ground exploitability, not by fidelity. We do not compete with learned information-set abstraction for poker-like games.

Learned representations. Deep bisimulation-style objectives (Gelada et al., 2019; Zhang et al., 2021) learn model-level representations for one agent. Value equivalence (Grimm et al., 2020), policy similarity (Agarwal et al., 2021) and decision-theoretic divergences (Zhao et al., 2022) compare models, states or distributions only through the values or decisions that matter; D2 takes that stance in a game, where the decision is a best response and what it responds to is an equilibrium of the abstraction itself, the source of the fixed point.

Exploitability from best responses. Using trained best responses to lower-bound exploitability is standard in population-based training (Lanctot et al., 2017) and has been studied as an estimator in its own right (Timbers et al., 2022). Our bracket uses it as the operative side; Appendix K documents how strongly the estimate depends on the best-response budget. Team-game solvers: Appendix K.

7 Limitations

Three tiers of bounds. *A priori* (Theorems 1–2): admissibility conditions, $26\times$ to $10^5\times$ above the exact value; no decision uses them. *Measured* (Theorem $2'$; the residual form of Theorem 1, a standard contraction argument (cf. Pérolat et al., 2017); Appendix D): 10–$43\times$ and 27–$520\times$ of exact, on the plain game no better than the one-class partition. *Operative:* the deviator lower bound of Corollary 4, exact here and at scale only as good as the deviators (Appendix K). In matrix-game LPs the loop beats solving the ground game only when there is enough to discard (Table 5: $17\times$ and $4.0\times$ fewer at 32 values, costlier at 2). Compression experiments are two-player zero-sum; the three-player general-sum study (§5.5) verifies the inequalities and the D0/D1 separation exactly but relies on potential-structured rewards to find equilibria, as computing constant-accuracy perfect stationary equilibria is PPAD-hard in general (Daskalakis et al., 2023); the CCE statement is not verified empirically. D2's frontier does not dominate $\phi^{Q^\star}$'s; under one oracle-checked search its distance guides no better (11 cells to 9), and the search that does best starts from $\phi^{Q^\star}$'s partition, was designed after a registered repair failed, and needs the ground solution (§5.2). At a *fixed* profile the question is exact: every class must be a clique of the ε-compatibility graph, so a minimum clique cover *lower-bounds* the coarsest admissible partition (Definition 1 also requires a ϕ-measurable profile, which the graph does not enforce). At the lossless tolerance that bound is met (566 states on the plain game, $1{,}143$ on the distractor game, each returned by four independent searches) against the baseline's 328 and 772; minima at that profile and tolerance, not over all equilibria or lossless partitions. Open: the coarsest admissible abstraction for the *loop*, whose distance moves with the partition, so that no one graph describes its admissible partitions; the fixed-profile case is settled in Appendix N; a bound using the occupancy-weighted distortion (8–$17\times$ below ε_Q; Appendix C).

Reproducibility statement

Every number in the main text comes from a script in the supplementary code and carries a provenance record (commit, configuration, seeds, best-response budget); the registrations the appendices cite are in `REGISTRATIONS.md`. The tabular results were regenerated from a clean clone and agree to the last digit.

Use of generative AI

In this work, we used generative AI tools for identifying relevant literature, manuscript editing, implementing methods and experiment code, and interpreting results. We have not used generative AI tools for conceptual frameworks, generating synthetic data sets, data collection, cleaning or reformatting data, qualitative or thematic data analysis, or translation. We have reviewed all AI-assisted work: reported numerical results are checked against exact solvers, tests, and a provenance record (commit, configuration, seed, and best-response budget) rather than accepted as generated;

AI-suggested related work is checked against the primary source before citation; and AI-drafted or AI-edited text is reviewed and revised by the authors.

No AI system is an author. The authors take full responsibility for the contents of this submission.

REFERENCES

David Abel, D. Ellis Hershkowitz, and Michael L. Littman. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2915–2923, 2016.

Rishabh Agarwal, Marlos C. Machado, Pablo Samuel Castro, and Marc G. Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In *International Conference on Learning Representations*, 2021.

Anonymous. Reference withheld for double-blind review, 2026. A peer-reviewed conference paper, accepted and not yet public; it will be cited in full in the final version.

Noam Brown and Tuomas Sandholm. Simultaneous abstraction and equilibrium finding in games. In *International Joint Conference on Artificial Intelligence*, pp. 489–496, 2015.

Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic Markov decision processes. In *AAAI Conference on Artificial Intelligence*, volume 34, pp. 10069–10076, 2020.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 27952–27964, 2021.

Constantinos Daskalakis, Dylan J. Foster, and Noah Golowich. Independent policy gradient methods for competitive reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5527–5540, 2020.

Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. The complexity of Markov equilibrium in stochastic games. In *Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4180–4234, 2023.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Conference on Uncertainty in Artificial Intelligence*, 2004.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *International Conference on Machine Learning*, 2019.

Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147:163–223, 2003.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Hiroki Ishibashi, Kenshi Abe, and Atsushi Iwasaki. Approximate state abstraction for Markov games. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 17555–17563, 2025.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pp. 267–274, 2002.

Fivos Kalogiannis, Ioannis Anagnostides, Ioannis Panageas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Vaggos Chatziafratis, and Stelios Stavroulakis. Efficiently computing Nash equilibria in adversarial team Markov games. In *International Conference on Learning Representations*, 2023.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2–3):209–232, 2002.

Christian Kroer and Tuomas Sandholm. Extensive-form game abstraction with bounds. In *ACM Conference on Economics and Computation*, 2014.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

Lihong Li, Thomas J. Walsh, and Michael L. Littman. Towards a unified theory of state abstraction for MDPs. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2006.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 1994.

Eric Maskin and Jean Tirole. Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory*, 100(2):191–219, 2001.

H. Brendan McMahan, Geoffrey J. Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *International Conference on Machine Learning*, pp. 536–543, 2003.

Julien Pérolat, Florian Strub, Bilal Piot, and Olivier Pietquin. Learning Nash equilibrium for general-sum Markov games from batch data. In *International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 232–241, 2017.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. From Poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8525–8535, 2021.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, et al. JaxMARL: Multi-agent RL environments and algorithms in JAX. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Tuomas Sandholm and Satinder Singh. Lossy stochastic game abstraction with bounds. In *ACM Conference on Electronic Commerce*, pp. 880–897, 2012.

Satinder P. Singh and Richard C. Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *International Conference on Learning Representations*, 2023.

Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. Approximate exploitability: Learning a best response. In *International Joint Conference on Artificial Intelligence*, 2022.

Kevin Waugh, David Schnizlein, Michael Bowling, and Duane Szafron. Abstraction pathologies in extensive games. In *International Conference on Autonomous Agents and Multiagent Systems*, 2009.

Zelai Xu, Yancheng Liang, Chao Yu, Yu Wang, and Yi Wu. Fictitious cross-play: Learning global Nash equilibrium in mixed cooperative-competitive games. In *International Conference on Autonomous Agents and Multiagent Systems*, 2023.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Kaiqing Zhang, Sham M. Kakade, Tamer Başar, and Lin F. Yang. Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1166–1178, 2020.

Shengjia Zhao, Abhishek Sinha, Yutong He, Aidan Perreault, Jiaming Song, and Stefano Ermon. Comparing distributions by measuring differences that affect decision making. In *International Conference on Learning Representations*, 2022.

A PROOFS

A.1 TWO ELEMENTARY LEMMAS

Lemma 1 (Mixing and integration). *(a) For distributions p_k, q_k on a finite set and weights $\lambda \in \Delta$, $\mathrm{TV}(\sum_k \lambda_k p_k, \sum_k \lambda_k q_k) \le \sum_k \lambda_k \mathrm{TV}(p_k, q_k)$. (b) For any f with $\inf f \le f \le \sup f$, $|\mathbb{E}_p f - \mathbb{E}_q f| \le \mathrm{TV}(p, q)\,(\sup f - \inf f)$.*

Proof. (a) is the triangle inequality for $\|\cdot\|_1$. (b): for the constant $c = (\sup f + \inf f)/2$, $\mathbb{E}_p f - \mathbb{E}_q f = \sum_x (p(x) - q(x))(f(x) - c)$, and $|f - c| \le (\sup f - \inf f)/2$, so the absolute value is at most $\|p - q\|_1 (\sup f - \inf f)/2$. □

Lemma 2 (Aggregation commutes with ϕ-measurable opponents). *Let $\hat{\pi}_{-i}$ be an abstract profile of the other players and $\pi_{-i} = \hat{\pi}_{-i} \circ \phi$. Then the MDP faced by player i in $\hat{G}_\phi$ against $\hat{\pi}_{-i}$, call it $\hat{M}_i$, is exactly the w-aggregation of $M_i(\pi_{-i})$:*

$$\hat{r}_i^{\hat{\pi}_{-i}}(z, a_i) = \sum_{s \in \phi^{-1}(z)} w(s \mid z)\, r_i^{\pi_{-i}}(s, a_i),$$
$$\hat{P}^{\hat{\pi}_{-i}}(z' \mid z, a_i) = \sum_{s \in \phi^{-1}(z)} w(s \mid z)\, \phi_{\#} P^{\pi_{-i}}(z' \mid s, a_i).$$

Proof. $\hat{r}_i^{\hat{\pi}_{-i}}(z, a_i) = \sum_{a_{-i}} \hat{\pi}_{-i}(a_{-i} \mid z) \sum_s w(s \mid z) r_i(s, a_i, a_{-i}) = \sum_s w(s \mid z) \sum_{a_{-i}} \pi_{-i}(a_{-i} \mid s) r_i(s, a_i, a_{-i})$ because $\pi_{-i}(\cdot \mid s) = \hat{\pi}_{-i}(\cdot \mid z)$ for every $s \in \phi^{-1}(z)$. The transition identity is the same computation. □

A.2 MODEL-LEVEL TRANSFER (THEOREM 1)

Throughout this section fix i, an abstract profile $\hat{\pi}$, its lift π, and write $M = M_i(\pi_{-i})$ (with r, P its reward and kernel) and $\hat{M} = \hat{M}_i$ for the abstract induced MDP of Lemma 2. Set $\varepsilon_r = \varepsilon_r^1(\phi; \pi)$, $\varepsilon_P = \varepsilon_P^1(\phi; \pi)$ and $\Delta_1 = \Delta(\varepsilon_r, \varepsilon_P)$.

Lemma 3 (Lifted-policy value closeness). *For every abstract policy $\hat{\sigma}_i$ of player i, with $\sigma_i = \hat{\sigma}_i \circ \phi$,*

$$e := \sup_s \left| V_M^{\sigma_i}(s) - \hat{V}_{\hat{M}}^{\hat{\sigma}_i}(\phi(s)) \right| \;\le\; \Delta_1.$$

Proof. Fix s and $z = \phi(s)$. Writing $V = V_M^{\sigma_i}$, $\hat{V} = \hat{V}_{\hat{M}}^{\hat{\sigma}_i}$,

$$V(s) = \sum_a \hat{\sigma}_i(a \mid z) \Big[r(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V(s') \Big],$$
$$\hat{V}(z) = \sum_a \hat{\sigma}_i(a \mid z) \Big[\hat{r}(z, a) + \gamma \sum_{z'} \hat{P}(z' \mid z, a) \hat{V}(z') \Big].$$

For the reward terms, $\hat{r}(z,a) = \sum_{s''} w(s'' \mid z) r(s'', a)$ with every $s'' \sim_\phi s$, so $|r(s,a) - \hat{r}(z,a)| \le \varepsilon_r$. For the transition terms,

$$\sum_{s'} P(s' \mid s,a) V(s') - \sum_{z'} \hat{P}(z' \mid z,a) \hat{V}(z') = \underbrace{\sum_{s'} P(s' \mid s,a) \big[V(s') - \hat{V}(\phi(s'))\big]}_{|\cdot| \le e} + \underbrace{\sum_{z'} \big[\phi_\# P(z' \mid s,a) - \hat{P}(z' \mid z,a)\big] \hat{V}(z')}_{(*)}.$$

By Lemma 2, $\hat{P}(\cdot \mid z,a) = \sum_{s''} w(s'' \mid z) \phi_\# P(\cdot \mid s'', a)$; by Lemma 1(a) and Definition 1, $\mathrm{TV}(\phi_\# P(\cdot \mid s,a), \hat{P}(\cdot \mid z,a)) \le \sum_{s''} w(s'' \mid z) \mathrm{TV}(\phi_\# P(\cdot \mid s,a), \phi_\# P(\cdot \mid s'',a)) \le \varepsilon_P$, and since $\sup \hat{V} - \inf \hat{V} \le V_{\max}$ (rewards lie in an interval of length R and $V_{\max} = R/(1-\gamma)$), Lemma 1(b) gives $|(*)| \le \varepsilon_P V_{\max}$. Averaging over $a \sim \hat{\sigma}_i(z)$ and taking the supremum over s, $e \le \varepsilon_r + \gamma(e + \varepsilon_P V_{\max})$. Since $e < \infty$, $e \le (\varepsilon_r + \gamma \varepsilon_P V_{\max})/(1-\gamma) = \Delta_1$. □

Lemma 4 (Optimal-value closeness). $e' := \sup_{s,a} |Q^\star_M(s,a) - \hat{Q}^\star_{\hat{M}}(\phi(s),a)| \le \Delta_1$*; consequently* $\sup_s |V^\star_M(s) - \hat{V}^\star_{\hat{M}}(\phi(s))| \le \Delta_1$.

Proof. Identical to Lemma 3 with the Bellman optimality equations $Q^\star(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} Q^\star(s',a')$ and its abstract counterpart, using $|\max_{a'} Q^\star(s',a') - \max_{a'} \hat{Q}^\star(\phi(s'),a')| \le e'$ (the max is 1-Lipschitz in sup-norm) and that $\max_{a'} \hat{Q}^\star$ takes values in an interval of length $V_{\max}$, the only property Lemma 1(b) needs (rewards may be signed). The second claim follows from $V^\star = \max_a Q^\star$. □

Lemmas 3 and 4 are aggregation analogues of the simulation lemma (Kearns & Singh, 2002, Lemma 4), which compares two MDPs on one state space; the argument here is a direct Bellman recursion.

Theorem 1 (restated). *Let $\hat{\pi}$ be an $\hat{\varepsilon}$-NE of $\hat{G}_\phi$ in sup-norm, $\pi = \hat{\pi} \circ \phi$, and $\Delta_1 = \Delta(\varepsilon^1_r(\phi;\pi), \varepsilon^1_P(\phi;\pi))$. Then for all i and s,*

$$V_i^\star(s \mid \pi_{-i}) - V_i^\pi(s) \le \hat{\varepsilon} + 2\Delta_1,$$

so π is an $(\hat{\varepsilon} + 2\Delta_1)$-NE of G. Moreover, without any sup-norm assumption,

$$\mathrm{EXP}_G(\pi) \le \mathrm{EXP}_{\hat{G}_\phi}(\hat{\pi}) + 2N\Delta_1.$$

The same statements hold for $\hat{\varepsilon}$-coarse-correlated equilibria with a stationary ϕ-measurable device $\hat{\mu}(\cdot \mid z)$ (a deviator commits at the outset to ignore recommendations and plays any Markov policy; the $\hat{\varepsilon}$ condition is per state) provided $\varepsilon^1_r, \varepsilon^1_P$ *are replaced by $\varepsilon^{1,\mathrm{cce}}_r, \varepsilon^{1,\mathrm{cce}}_P$ defined in the proof, which also range over the conditionals $\hat{\mu}(\cdot \mid a_i, z)$ with $\hat{\mu}(a_i \mid z) > 0$.*

Proof. Fix i and s, $z = \phi(s)$. Lemma 3 with $\hat{\sigma}_i = \hat{\pi}_i$ gives $V_i^\pi(s) \ge \hat{V}_i^{\hat{\pi}}(z) - \Delta_1$; the abstract equilibrium condition gives $\hat{V}_i^{\hat{\pi}}(z) \ge \hat{V}_i^\star(z) - \hat{\varepsilon}$; Lemma 4 gives $\hat{V}_i^\star(z) \ge V_i^\star(s \mid \pi_{-i}) - \Delta_1$. Chaining the three inequalities proves the per-state bound. For the second claim, the first and third inequalities hold at every s without the middle one, so $V_i^\star(s \mid \pi_{-i}) - V_i^\pi(s) \le \hat{V}_i^\star(\phi(s)) - \hat{V}_i^{\hat{\pi}}(\phi(s)) + 2\Delta_1$; integrating over $s \sim \mu_0$ turns the abstract gap into the abstract gap under $\phi_\# \mu_0$, and summing over i gives the exploitability bound. For a ϕ-measurable correlation device recommending joint actions from $\hat{\mu}(\cdot \mid z)$ the distortions must be taken with respect to *both* laws of the other players that occur: the marginal $\hat{\mu}_{-i}(\cdot \mid z)$ faced by a deviator, and the conditionals $\hat{\mu}(a_{-i} \mid a_i, z)$ faced on path by a player recommended a_i (for a product profile these coincide, which is why Definition 1 suffices for Nash equilibria). Define $\varepsilon^{1,\mathrm{cce}}_r, \varepsilon^{1,\mathrm{cce}}_P$ as the maxima of the corresponding quantities over these laws and Δ_1 accordingly. On path, player i's value satisfies the recursion of Lemma 3 with, at each (z, a_i), opponents drawn from $\hat{\mu}(\cdot \mid a_i, z)$, so Lemma 3 applies with the conditional distortions; a deviating player i ignores recommendations while the others follow theirs, so player i faces M_i built from the marginal of $\hat{\mu}(\cdot \mid \phi(s))$ over a_{-i}, which is ϕ-measurable, and Lemmas 2 and 4 apply unchanged. The abstract $\hat{\varepsilon}$-CCE condition is precisely the middle inequality. □

Remark 1 (The marginal alone does not suffice for CCE)**.** *Take two absorbing states s_0, s_1, two players with actions $\{0,1\}$ and a common reward $\mathbf{1}[a_1 = a_2]$ in s_0 and $\mathbf{1}[a_1 \neq a_2]$ in s_1; merge the two states with w uniform, and let the device recommend $(0,0)$ or $(1,1)$ with probability $\frac{1}{2}$ each. Every marginal is uniform, so against the marginal each action earns $\frac{1}{2}$ in both states: $\varepsilon_r^1 = \varepsilon_P^1 = 0$ under Definition 1, and in the abstract game following earns $\frac{1}{2}$ and so does every deviation, so $\hat{\varepsilon} = 0$. The marginal-only statement would make the lifted device an exact CCE. It is not: in s_1 following earns 0 forever and any deviation earns $\frac{1}{2}$, a gap of $\frac{1}{2}/(1-\gamma)$ $(= 5$ at $\gamma = 0.9)$. With the conditionals the reward distortion is 1, $\Delta_1 = 1/(1-\gamma)$, and the bound $2\Delta_1 = 20$ holds. The example is a regression test in the supplementary code.*

A.3 Value-level transfer (Theorem 2)

Theorem 2 (restated). *Let $\hat{\pi}$ be an $\hat{\varepsilon}$-NE of $\hat{G}_\phi$ in sup-norm, $\pi = \hat{\pi} \circ \phi$, and $\varepsilon_Q = \varepsilon_Q(\phi; \pi)$. Then, for any weighting w, for all i and s,*

$$V_i^\star(s \mid \pi_{-i}) - V_i^\pi(s) \;\leq\; \frac{\hat{\varepsilon}}{1-\gamma} + \frac{2\varepsilon_Q}{(1-\gamma)^2}.$$

Proof. Fix i; let $M = M_i(\pi_{-i})$ with optimal $Q^\star, V^\star$, and $\hat{M}$ its w-aggregation (Lemma 2) with optimal $\hat{Q}^\star, \hat{V}^\star$ and Bellman optimality operator $\hat{\mathcal{T}}$.

Steps 2–4 follow Claims 1–3 in the proof of Lemma 1 of Abel et al. (2016) (the case $N = 1$, $\hat{\varepsilon} = 0$), extended to an $\hat{\varepsilon}$-optimal abstract policy and to the MDP each player faces against the lifted profile.

Step 1 (an approximate abstract fixed point). Define $\tilde{Q}(z,a) = \sum_{s \in \phi^{-1}(z)} w(s \mid z) Q^\star(s,a)$. Since $|Q^\star(s,a) - Q^\star(s'',a)| \leq \varepsilon_Q$ for $s'' \sim_\phi s$, we have $|\tilde{Q}(\phi(s),a) - Q^\star(s,a)| \leq \varepsilon_Q$ for all s, a, and hence $|\max_{a'} \tilde{Q}(\phi(s'),a') - \max_{a'} Q^\star(s',a')| \leq \varepsilon_Q$. Using Lemma 2,

$$\begin{aligned}
\hat{\mathcal{T}}\tilde{Q}(z,a) &= \sum_s w(s \mid z)\Big[r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} \tilde{Q}(\phi(s'),a')\Big] \\
&= \sum_s w(s \mid z)\Big[r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} Q^\star(s',a')\Big] \pm \gamma\varepsilon_Q \\
&= \tilde{Q}(z,a) \pm \gamma\varepsilon_Q.
\end{aligned}$$

Thus $\|\hat{\mathcal{T}}\tilde{Q} - \tilde{Q}\|_\infty \leq \gamma\varepsilon_Q$.

Step 2 (abstract optimal values are close to ground ones). Since $\hat{\mathcal{T}}$ is a γ-contraction with fixed point $\hat{Q}^\star$, $\|\hat{Q}^\star - \tilde{Q}\|_\infty \leq \|\hat{\mathcal{T}}\hat{Q}^\star - \hat{\mathcal{T}}\tilde{Q}\|_\infty + \|\hat{\mathcal{T}}\tilde{Q} - \tilde{Q}\|_\infty \leq \gamma\|\hat{Q}^\star - \tilde{Q}\|_\infty + \gamma\varepsilon_Q$, so $\|\hat{Q}^\star - \tilde{Q}\|_\infty \leq \gamma\varepsilon_Q/(1-\gamma)$ and

$$\sup_{s,a} |\hat{Q}^\star(\phi(s),a) - Q^\star(s,a)| \leq \varepsilon_Q + \frac{\gamma\varepsilon_Q}{1-\gamma} = \frac{\varepsilon_Q}{1-\gamma} =: \kappa.$$

Step 3 (the lifted policy is nearly greedy for $Q^\star$ at every state). Because $\hat{\pi}_i$ is $\hat{\varepsilon}$-optimal in $\hat{M}$ at every z and $\hat{Q}^\star \geq \hat{Q}^{\hat{\pi}}$ pointwise, $\mathbb{E}_{a \sim \hat{\pi}_i(z)} \hat{Q}^\star(z,a) \geq \mathbb{E}_{a \sim \hat{\pi}_i(z)} \hat{Q}^{\hat{\pi}}(z,a) = \hat{V}^{\hat{\pi}}(z) \geq \hat{V}^\star(z) - \hat{\varepsilon}$. Hence for every s with $z = \phi(s)$, using Step 2 twice,

$$\mathbb{E}_{a \sim \pi_i(s)} Q^\star(s,a) \;\geq\; \mathbb{E}_{a \sim \hat{\pi}_i(z)} \hat{Q}^\star(z,a) - \kappa \;\geq\; \hat{V}^\star(z) - \hat{\varepsilon} - \kappa \;\geq\; V^\star(s) - \hat{\varepsilon} - 2\kappa.$$

Step 4 (performance difference). For the ground policy π_i in M, the performance-difference lemma (Kakade & Langford, 2002, Lemma 6.1) gives

$$V^\star(s) - V^{\pi_i}(s) = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d_s^{\pi_i}} \mathbb{E}_{a \sim \pi_i(s')} \big[V^\star(s') - Q^\star(s',a)\big] \leq \frac{\hat{\varepsilon} + 2\kappa}{1-\gamma},$$

where $d_s^{\pi_i}$ is the normalized discounted state-visitation distribution of π_i started at s. Substituting κ proves the claim. □

A.4 Nesting (Proposition 3) and the two-sided bracket (Theorem 4)

Proposition 3 (restated). *For every ϕ and every profile π with ϕ-measurable components: $\varepsilon_r^1(\phi;\pi) \le \varepsilon_r^0(\phi)$, $\varepsilon_P^1(\phi;\pi) \le \varepsilon_P^0(\phi)$, and $\varepsilon_Q(\phi;\pi) \le 2\Delta(\varepsilon_r^1(\phi;\pi), \varepsilon_P^1(\phi;\pi))$. Consequently, writing $\mathcal{M}(\pi)$ for the partitions with respect to which π is measurable, $\Phi^0(\varepsilon) = \{\phi : \varepsilon_r^0, \varepsilon_P^0 \le \varepsilon\}$, $\Phi^1(\varepsilon;\pi) = \{\phi \in \mathcal{M}(\pi) : \varepsilon_r^1, \varepsilon_P^1 \le \varepsilon\}$ and $\Phi^Q(\varepsilon;\pi) = \{\phi \in \mathcal{M}(\pi) : \varepsilon_Q \le \varepsilon\}$,*

$$\Phi^0(\varepsilon) \cap \mathcal{M}(\pi) \subseteq \Phi^1(\varepsilon;\pi) \subseteq \Phi^Q\big(2\Delta(\varepsilon,\varepsilon);\pi\big);$$

in particular every $\phi \in \Phi^0(\varepsilon)$ lies in $\Phi^1(\varepsilon; \hat{\pi} \circ \phi)$ for every abstract profile $\hat{\pi}$ of $\hat{G}_\phi$.

Proof. For $s \sim_\phi s'$, $r_i^{\pi_{-i}}(s, a_i) - r_i^{\pi_{-i}}(s', a_i) = \sum_{a_{-i}} \pi_{-i}(a_{-i} \mid \phi(s))[r_i(s,a) - r_i(s',a)]$ has absolute value at most ε_r^0; the transition claim is Lemma 1(a) applied to the same mixture. For the third claim apply Lemma 4 at s and at s' (both map to the same z): $|Q^\star(s,a_i) - Q^\star(s',a_i)| \le |Q^\star(s,a_i) - \hat{Q}^\star(z,a_i)| + |\hat{Q}^\star(z,a_i) - Q^\star(s',a_i)| \le 2\Delta_1$. The inclusions restate the three inequalities. □

Corollary 4 (restated). *For any ϕ, abstract profile $\hat{\pi}$ with lift π, and any ground policies $\{\psi_i\}$,*

$$\begin{aligned}\sum_i \Big(V_i^{(\psi_i,\pi_{-i})}(\mu_0) - V_i^{\pi}(\mu_0)\Big) \;&\le\; \mathrm{EXP}_G(\pi) \\ &\le\; \min\Big\{\mathrm{EXP}_{\hat{G}_\phi}(\hat{\pi}) + 2N\Delta_1,\; N\Big(\frac{\hat{\varepsilon}}{1-\gamma} + \frac{2\varepsilon_Q}{(1-\gamma)^2}\Big)\Big\},\end{aligned}$$

where $\hat{\varepsilon}$ is the sup-norm equilibrium gap of $\hat{\pi}$ in $\hat{G}_\phi$. The lower bound is an equality when each ψ_i is an exact best response to π_{-i}.

Proof. The lower bound is the definition of $V_i^\star$ as a maximum over all policies; the two upper bounds are Theorem 1 (second statement) and Theorem 2 integrated over μ_0 and summed over i. □

Remark 2 (What the constants say)**.** *The transition term enters Δ_1 with the factor $\gamma V_{\max}/(1-\gamma)$, so the model-level bound is informative only when ε_P^1 is (near) zero – e.g. when the abstraction drops a variable that does not affect the pushforward dynamics under unilateral deviations. The value-level bound carries a $(1-\gamma)^{-2}$ factor. In the tabular experiments both a priori bounds are loose by one to five orders of magnitude, whereas the deviator lower bound equals the exploitability when best responses are exact; the a priori bounds serve as admissibility conditions (which level of the hierarchy licenses an abstraction) and the deviators as the operative stopping criterion. Between the two sit the measured forms (Theorem 2′ below and the residual form of Theorem 1 in Appendix D), which compute the quantity each proof bounds instead of bounding it; they stay 10–43× (Theorem 1) and 27–520× (Theorem 2′) above the exact value where it is not zero (Appendix D). As a function of tolerance, the bound of Theorem 2, $2N\varepsilon/(1-\gamma)^2$ (per player that of Abel et al. (2016, Lemma 1)), has a constant smaller than that of the best bound of Ishibashi et al. (2025), $12\varepsilon/(1-\gamma)^3$, by the factor $6/(N(1-\gamma))$: 30× for $N = 2$, $\gamma = 0.9$. This is a comparison of constants and no more. The tolerances measure different things (the spread of minimax $Q^\star$ over joint actions against that of best-response values over own actions); neither hypothesis is shown to imply the other, and at $\gamma = 0.9$ neither number is informative (the largest attainable gap is $NV_{\max} \approx 40$).*

A.5 Markov perfection, correlated devices, regularized equilibria, and a measured form of Theorem 2

Remark 3 (Markov perfection)**.** *The per-state statements of Theorems 1 and 2 bound $V_i^\star(s \mid \pi_{-i}) - V_i^\pi(s)$ at* every *state s, reached or not, and $V_i^\star$ is a maximum over all policies of player i, stationary or not. A stationary profile with this property for a tolerance ε is an ε-Markov perfect equilibrium: an ε-equilibrium of every subgame. So if $\hat{\pi}$ is an $\hat{\varepsilon}$-Markov perfect equilibrium of $\hat{G}_\phi$ (which is what "$\hat{\varepsilon}$-NE in sup-norm" means), then $\hat{\pi} \circ \phi$ is an $(\hat{\varepsilon} + 2\Delta_1)$-Markov perfect equilibrium of G, and an $\big(\hat{\varepsilon}/(1-\gamma) + 2\varepsilon_Q/(1-\gamma)^2\big)$-Markov perfect one. Only the exploitability statements, which average over μ_0 and need no sup-norm hypothesis, are statements about equilibrium from μ_0 alone.*

Remark 4 (Correlated equilibrium). *Let $\hat{\mu}(\cdot \mid z)$ be a stationary ϕ-measurable device recommending joint actions. A swap deviation of player i is a map from (history and) its own recommendation a_i to an action, so a deviating player faces the* same *law of the others as a player who follows: the conditional $\hat{\mu}(\cdot \mid z, a_i)$, not the product of marginals and, unlike the coarse-correlated case of Theorem 1, not the marginal either. Player i's problem is then an MDP on the augmented states (s, a_i), $\hat{\mu}_i(a_i \mid \phi(s)) > 0$, with reward $\mathbb{E}_{a_{-i}\sim\hat{\mu}(\cdot|\phi(s),a_i)} r_i(s, a'_i, a_{-i})$ for playing a'_i, the corresponding kernel on s', and the next recommendation drawn from $\hat{\mu}_i(\cdot \mid \phi(s'))$. Because that last factor depends on s' only through $\phi(s')$, the abstract device's problem is the w-aggregation of the ground one under $(s, a_i) \mapsto (\phi(s), a_i)$ (the computation of Lemma 2), and Lemmas 3 and 4 apply verbatim with $\varepsilon_r^{1,\mathrm{ce}}, \varepsilon_P^{1,\mathrm{ce}}$: the within-class spreads of that reward and of the pushforward of that kernel, maximized over i, a_i in the support and a'_i. Chaining the three inequalities of the proof of Theorem 1 at every (s, a_i): if no swap deviation gains more than $\hat{\varepsilon}$ at any (z, a_i) in $\hat{G}_\phi$, none gains more than $\hat{\varepsilon} + 2\Delta(\varepsilon_r^{1,\mathrm{ce}}, \varepsilon_P^{1,\mathrm{ce}})$ at any (s, a_i) in G. The example of Remark 1 shows that the conditionals cannot be replaced by marginals here either: there every marginal distortion is 0 and the swap that contradicts the recommendation in s_1 gains $1/(1-\gamma)$.*

Corollary 9 (Entropy-regularized transfer). *Fix a temperature $\tau > 0$ and give each player the regularized reward $r_i(s, a) + \tau\mathcal{H}(\sigma_i(\cdot \mid s))$ for its own policy σ_i; write $V^{\pi}_{i,\tau}$, $V^{\star}_{i,\tau}(\cdot \mid \pi_{-i})$ for the regularized values and $\mathrm{EXP}_{G,\tau}(\pi) = \sum_i \big(V^{\star}_{i,\tau}(\mu_0 \mid \pi_{-i}) - V^{\pi}_{i,\tau}(\mu_0)\big)$. Let $V^{\tau}_{\max} = V_{\max} + \tau \log|\mathcal{A}|/(1-\gamma)$ with $|\mathcal{A}| = \max_i |\mathcal{A}_i|$, and $\Delta^{\tau}_1 = \big(\varepsilon^1_r(\pi) + \gamma\varepsilon^1_P(\pi)V^{\tau}_{\max}\big)/(1-\gamma)$. For every abstract profile $\hat{\pi}$ with lift π,*

$$\mathrm{EXP}_{G,\tau}(\pi) \le \mathrm{EXP}_{\hat{G}_\phi,\tau}(\hat{\pi}) + 2N\Delta^{\tau}_1, \qquad \big|\mathrm{EXP}_{G,\tau}(\pi) - \mathrm{EXP}_G(\pi)\big| \le \frac{\tau N \log|\mathcal{A}|}{1-\gamma}.$$

In particular the lift of a τ-quantal response equilibrium of $\hat{G}_\phi$ (every player's policy the softmax of its own regularized action values, so that $\mathrm{EXP}_{\hat{G}_\phi,\tau}(\hat{\pi}) = 0$) has regularized exploitability at most $2N\Delta^{\tau}_1$ in G, and exploitability at most $2N\Delta^{\tau}_1 + \tau N \log|\mathcal{A}|/(1-\gamma)$.

Proof. A lifted policy has the same entropy at every state of a class, so on path the regularized reward of $M_i(\pi_{-i})$ differs from $r_i^{\pi_{-i}}$ by a term that is constant on classes: its within-class spread is still at most ε^1_r, the kernel is unchanged, and Lemma 2 holds as before. Lemma 3 then holds for the regularized evaluation equations with one change: regularized values span at most $V^{\tau}_{\max}$ instead of $V_{\max}$, because a step's reward now lies in an interval of length $R + \tau\log|\mathcal{A}|$; hence Δ^{τ}_1. (The change is needed: with all rewards zero, a class whose two states lead to absorbing classes of policy entropy 0 and $\log|\mathcal{A}|$ has $\Delta_1 = 0$ and regularized values that differ by $\gamma\tau\log|\mathcal{A}|/(1-\gamma)$.) For a deviator the entropy is its own choice and the optimality equations are the soft ones, $Q(s,a) = r(s,a) + \gamma\sum_{s'} P(s' \mid s,a)\,\tau\log\sum_{a'} e^{Q(s',a')/\tau}$. Log-sum-exp is a nonexpansion in sup-norm and takes values in an interval of length at most $V^{\tau}_{\max}$, which are the only two properties of the maximum used in the proof of Lemma 4; so Lemma 4 holds for the soft operators with Δ^{τ}_1. The first and third inequalities in the proof of Theorem 1 therefore hold for regularized values at every s, and integrating over μ_0 and summing over i gives the first claim. For the second, $V^{\pi}_{i,\tau} = V^{\pi}_i + \tau\bar{\mathcal{H}}_i$ with $\bar{\mathcal{H}}_i \in [0, \log|\mathcal{A}_i|/(1-\gamma)]$ the expected discounted entropy of π_i along the play, and for the same reason $V^{\star}_i \le V^{\star}_{i,\tau} \le V^{\star}_i + \tau\log|\mathcal{A}_i|/(1-\gamma)$; subtract and sum over i. □

Theorem 2′ (measured value-level transfer). *Let $\hat{\pi}$ be an $\hat{\varepsilon}$-NE of $\hat{G}_\phi$ in sup-norm, $\pi = \hat{\pi}\circ\phi$, and for each player let $\kappa'_i = \sup_{s,a_i}\big|\hat{Q}^{\star}_i(\phi(s), a_i) - Q^{\star}_i(s, a_i \mid \pi_{-i})\big|$, the distance between the optimal action values of $\hat{M}_i$ and of $M_i(\pi_{-i})$. Then for all i and s,*

$$V^{\star}_i(s \mid \pi_{-i}) - V^{\pi}_i(s) \;\le\; \frac{\hat{\varepsilon} + 2\kappa'_i}{1-\gamma},$$

so π is an $(\hat{\varepsilon} + 2\max_i \kappa'_i)/(1-\gamma)$-NE of G in sup-norm. Moreover $\kappa'_i \le \varepsilon_Q/(1-\gamma)$, so the bound is never larger than that of Theorem 2.

Proof. Steps 3 and 4 of the proof of Theorem 2 use ε_Q only through the inequality $\sup_{s,a}|\hat{Q}^{\star}(\phi(s), a) - Q^{\star}(s,a)| \le \kappa$ established in Step 2; they go through with κ'_i, which is that supremum itself, in place of κ. Step 2 is the statement $\kappa'_i \le \varepsilon_Q/(1-\gamma)$. □

At $\hat{\varepsilon} = 0$ Theorem 2′ is Corollary 2 of Singh & Yee (1994) applied to $\hat{Q}_i^\star \circ \phi$; $\hat{\varepsilon}$ allows an approximately greedy abstract policy. Unlike ε_Q, κ_i' is not a property of the partition and the ground best response alone: it also needs the abstract optimal action values, which are available once $\hat{G}_\phi$ is solved. It is the quantity tabulated in Appendix D.

A.6 Value-level nesting (Proposition 5) and the succinctness separation (Proposition 6)

Proof of Proposition 5. Take player 1; player 2 is symmetric with $-Q^\star, -V^\star$. Write $V^1 = V_1^\star(\cdot \mid \pi_2)$ and $V^2 = V_2^\star(\cdot \mid \pi_1)$ for the best-response values. Against any fixed strategy of the opponent a player secures at least the minimax value, so $V^1 \geq V^\star$ and $V^2 \geq -V^\star$ pointwise. Since $V_1^\pi + V_2^\pi = 0$, $0 \leq V^1 - V^\star \leq V^1 + V^2 = (V^1 - V_1^\pi) + (V^2 - V_2^\pi) \leq \mathrm{GAP}(\pi)$ at every state. Now $Q_1^\star(s, a_1 \mid \pi_2) = \sum_{a_2} \pi_2(a_2|s)\big[Q^\star(s, a_1, a_2) + \gamma \sum_{s''} P(s''|s, a_1, a_2)\,(V^1 - V^\star)(s'')\big]$. For $s \sim_\phi s'$ the mixture $\pi_2(\cdot|s) = \pi_2(\cdot|s')$ is the same because π is ϕ-measurable. The first terms then differ by at most $\varepsilon_{Q^\star}(\phi)$; each second term lies in $[0, \gamma\,\mathrm{GAP}(\pi)]$, so they differ by at most $\gamma\,\mathrm{GAP}(\pi)$. □

Proof of Proposition 6. (i) Fix any $V : \mathcal{S} \to \mathbb{R}$ as continuation and a state (c, d), and let Δ be player 1's continuation gain from the jump to c_d relative to (σ, σ). Suppose $\Delta \leq 0$. Against $a_2 \neq \tau$, player 1's τ moves as σ and costs κ; against $a_2 = \tau$, playing σ collects κ and avoids the jump while τ collects nothing and jumps, a difference of $-\kappa + \Delta < 0$. So τ is strictly worse than σ for player 1 against every action, and once player 1 never plays τ the jump cannot occur, so player 2's τ only costs κ. If $\Delta > 0$ the jump hurts player 2 and the same argument runs with the roles exchanged. Hence for every continuation the stage game at (c, d) has the value, and the equilibria, of its restriction to actions other than τ, which is the stage game of G_0 at c whenever V does not depend on d. Therefore $W(c, d) = V_0^\star(c)$ is a fixed point of the Shapley operator of G_N, which has only one: $V^\star = W$, and no Markov perfect equilibrium plays τ (a Nash equilibrium from μ_0 may prescribe τ at states it never reaches). The mechanism is iterated strict dominance, for every continuation. The same argument applies to the abstract game of $\phi(c, d) = c$, so its lifted Markov perfect equilibria never play τ; against such a profile neither a player's induced reward nor its induced kernel over classes depends on d, whatever that player does, which is $\varepsilon_r^1 = \varepsilon_P^1 = 0$, and $\varepsilon_Q = 0$ follows from $\varepsilon_Q \leq 2\Delta_1$. Theorem 1 then makes the lifted equilibrium exact. *(ii)* By (i), $Q^\star((c, d), \tau, \tau) = r(c, \sigma, \sigma) + \gamma V_0^\star(c_d)$: the payments cancel and the jump is certain. As $r(\cdot, \sigma, \sigma)$ is constant, two states with $d \neq d'$ differ in this entry by $\gamma|V_0^\star(c_d) - V_0^\star(c_{d'})| \geq \gamma\Delta_{\min}$, whatever their core states. *(iii)* If $\varepsilon_r^0 = \varepsilon_P^0 = 0$ the Shapley operator maps ϕ-measurable functions to ϕ-measurable functions, so $V^\star$ is constant on classes. Suppose a class contained (c, d) and (c', d') with $d \neq d'$. Under (τ, τ) their successors are $\{(c_d, e)\}_e$ and $\{(c_{d'}, e)\}_e$; these have different values by (i), hence lie in different classes, so the two pushforward distributions have disjoint supports and total-variation distance 1, contradicting $\varepsilon_P^0 = 0$. In (ii) and (iii), since no class contains two values of d, the partition is the union of N partitions, one per value of d. □

In the soccer games of §5, σ is STAY, τ is TAUNT and $r(\cdot, \sigma, \sigma) = 0$ because nobody scores by standing still. Appendix H checks the premises and the conclusions of Proposition 6 exactly on seven instances.

Proof of Proposition 7. Separable stage games. Let $M(\boldsymbol{a}, \boldsymbol{b}) = C(a_1, a_2) + \Delta B(b_1, b_2)$ on actions (a_i, b_i). Against a pure column (a_2, b_2) a row strategy x earns $\sum_{a_1} x^{\mathrm{core}}(a_1) C(a_1, a_2) + \Delta \sum_{b_1} x^{\mathrm{side}}(b_1) B(b_1, b_2)$, which depends on x only through its two marginals, and the minimum over columns separates. So x secures $\mathrm{val}(C) + \Delta\,\mathrm{val}(B)$ iff both marginals are optimal, and likewise for the column player. If B has zero row and column sums, the uniform strategies secure 0 for both players, so $\mathrm{val}(B) = 0$; if they are its only equilibrium, every optimal strategy of M has a uniform side marginal. Since optimality depends on the marginals only, $x^{\mathrm{core}} \otimes$uniform is optimal whenever x is, so every action (a_1, b_1) with a_1 in the support of an optimal core strategy is played by some optimal strategy of M. *(i)* Let $W(c, d) = V_0^\star(c)$. With continuation W the stage game at (c, d) is $Q_0^\star(c, \cdot, \cdot) + \Delta B_d$, separable, of value $V_0^\star(c)$, because the core transition ignores side actions and d's chain ignores play, so $\sum P\,W = \sum_{c'} P_0(c'|c, a_1, a_2) V_0^\star(c')$. W is thus a fixed point of the Shapley operator of H_N, which has only one, so $V^\star = W$, $Q^\star((c, d), \cdot) = Q_0^\star(c, \cdot) + \Delta B_d$,

and every Markov perfect equilibrium plays the side actions uniformly. With uniform weights the abstract game of $\phi(c,d) = c$ has rewards $r_0(c,\cdot) + \Delta\bar{B}$, $\bar{B}$ the mean of the B_d, and the class kernel P_0, since d is marginalised by the pushforward; by the same argument its value is $V_0^\star$, and every Markov perfect equilibrium plays, at every class, an optimal core strategy of $Q_0^\star(c,\cdot,\cdot)$ and, $\bar{B}$ having only the uniform equilibrium, uniform side actions (a Nash equilibrium from $\phi_\# \mu_0$ may do otherwise at classes it never reaches). Lifted, that profile is optimal in every stage game $Q_0^\star(c,\cdot,\cdot) + \Delta B_d$: an exact equilibrium of H_N. Against it, player 1's own action (a_1, b_1) earns $r_0(c, a_1, \pi_2^{\text{core}}) + \frac{\Delta}{n}\sum_{b_2} B_d(b_1, b_2) = r_0(c, a_1, \pi_2^{\text{core}})$ by the zero row sums, and its class kernel is $P_0(\cdot|c, a_1, \pi_2^{\text{core}})$; neither depends on d, and player 2 is symmetric by the zero column sums. So $\varepsilon_r^1 = \varepsilon_P^1 = 0$, and $\varepsilon_Q \le 2\Delta_1 = 0$ by Proposition 3. *(ii)* Let $(c,d) \sim_\phi (c',d')$ with $d \neq d'$, and $E = B_d - B_{d'} \neq 0$. By (i), $f((c,d),\boldsymbol{a}) - f((c',d'),\boldsymbol{a}) = D(a_1,a_2) + \Delta E(b_1,b_2)$, where D collects the core terms (for $f = Q^\star$ the continuations $\sum_{c''} P_0 V_0^\star$ do not depend on d). Fix a_1, a_2 in the supports of optimal core strategies at c; by the first paragraph $K(c,d)$ contains $((a_1,b_1),(a_2,b_2))$ for every b_1, b_2. A matrix with zero row sums has entries of both signs unless it is zero, so $\max E - \min E \ge \|E\|_\infty$, and $\max_{b_1,b_2} |D(a_1,a_2) + \Delta E(b_1,b_2)| \ge \frac{\Delta}{2}(\max E - \min E) \ge \Delta\beta/2$ (and $\ge \Delta\beta$ when $c = c'$, where $D = 0$). No class of a partition below that tolerance contains two values of d, so the partition is the union of N partitions, one per value of d. □

Such matrices exist for every N: $B_d = s_d\, C$ with C the 3×3 rock–paper–scissors matrix and distinct $s_d > 0$. Our instances draw them at random (Appendix H).

Comparing, for each pair of states, only the own actions whose best-response value is within a margin m of the best at either state keeps Theorem 2's bound at an exact abstract equilibrium once the margin is large enough:

Proposition 10 (D2 on near-best deviations)**.** *Let $\hat{\pi}$ be an exact Markov perfect equilibrium of $\hat{G}_\phi$, $\pi = \hat{\pi} \circ \phi$, $Q_i = Q_i^\star(\cdot,\cdot|\pi_{-i})$, $V_i = \max_{a_i} Q_i$ and $U_i^m(s) = \{a_i : Q_i(s,a_i) \ge V_i(s) - m\}$, and let ε_Q^m be the largest $|Q_i(s,a_i) - Q_i(s',a_i)|$ over players, $s \sim_\phi s'$ and $a_i \in U_i^m(s) \cup U_i^m(s')$. If $m \ge 2\varepsilon_Q^m/(1-\gamma)$, then $V_i^\star(s|\pi_{-i}) - V_i^\pi(s) \le 2\varepsilon_Q^m/(1-\gamma)^2$ for every i and s.*

Proof of Proposition 10. Fix player i and drop the index: $\varepsilon = \varepsilon_Q^m$, $U = U^m$, $a^*(s)$ a best action at s, and $\bar{Q}(z,a) = \sum_s w(s|z) Q(s,a)$. *(1)* For $s \sim_\phi s'$, $|V(s) - V(s')| \le \varepsilon$: $a^*(s) \in U(s)$, so $V(s') \ge Q(s', a^*(s)) \ge V(s) - \varepsilon$, and symmetrically. *(2)* By Lemma 2, player i's abstract best-response MDP against $\hat{\pi}_{-i}$ is the w-aggregation of its ground one against π_{-i}; let $\hat{Q}$ be its optimal action values and $\hat{T}$ its Bellman operator. Then $\hat{T}\bar{Q}(z,a) - \bar{Q}(z,a) = \gamma \sum_s w(s|z) \sum_{s'} P(s'|s,a)\,[\max_{a'} \bar{Q}(\phi(s'),a') - V(s')]$, and for $s' \in z'$, $\max_{a'} \bar{Q}(z',a') \ge \bar{Q}(z', a^*(s')) \ge V(s') - \varepsilon$ because $a^*(s') \in U(s')$, while $\bar{Q}(z',a') \le \max_{s'' \in z'} V(s'') \le V(s') + \varepsilon$ for every a'. So $\|\hat{T}\bar{Q} - \bar{Q}\|_\infty \le \gamma\varepsilon$ and $\|\hat{Q} - \bar{Q}\|_\infty \le \gamma\varepsilon/(1-\gamma)$, for every action, inside U or not. *(3)* Let $\hat{\pi}_i$ play $\hat{a}$ at z. As $\hat{\pi}$ is Markov perfect, $\hat{a}$ is greedy for $\hat{Q}(z,\cdot)$, so for every $s \in z$, $\bar{Q}(z,\hat{a}) \ge \bar{Q}(z, a^*(s)) - 2\gamma\varepsilon/(1-\gamma) \ge V(s) - (1+\gamma)\varepsilon/(1-\gamma)$. If $\hat{a} \notin U(s)$ for some $s \in z$, then every $s'' \in z$ has either $\hat{a} \in U(s'')$ and $Q(s'',\hat{a}) \le Q(s,\hat{a}) + \varepsilon < V(s) - m + \varepsilon$, or $\hat{a} \notin U(s'')$ and $Q(s'',\hat{a}) < V(s'') - m \le V(s) + \varepsilon - m$; hence $\bar{Q}(z,\hat{a}) < V(s) - m + \varepsilon \le V(s) - (1+\gamma)\varepsilon/(1-\gamma)$ when $m \ge 2\varepsilon/(1-\gamma)$, a contradiction. So $\hat{a} \in U(s)$ for every $s \in z$, the values $Q(s'',\hat{a})$, $s'' \in z$, lie within ε of each other, and $Q(s,\hat{a}) \ge \bar{Q}(z,\hat{a}) - \varepsilon \ge V(s) - 2\varepsilon/(1-\gamma)$. *(4)* Every action π_i plays is therefore $2\varepsilon/(1-\gamma)$-greedy for Q at every state, and the performance-difference argument of Step 4 in the proof of Theorem 2 gives $V^\star(s|\pi_{-i}) - V^\pi(s) \le 2\varepsilon/(1-\gamma)^2$. □

A.7 Algorithmic note: monotone refinement is a joint fixed point

Let SOL be any deterministic equilibrium solver. Consider the loop: ϕ_0 arbitrary; $\hat{\pi}_t = \text{SOL}(\hat{G}_{\phi_t})$; $\pi_t = \hat{\pi}_t \circ \phi_t$; ϕ_{t+1} = refinement of ϕ_t that splits exactly those classes containing $s \sim_{\phi_t} s'$ with $|Q_i^\star(s, a_i \mid \pi_{t,-i}) - Q_i^\star(s', a_i \mid \pi_{t,-i})| > \varepsilon$ for some i, a_i (level D2; analogously with $(\varepsilon_r^1, \varepsilon_P^1)$ at level D1). Since partitions only refine and $\mathcal{S}$ is finite, the loop terminates at some T with $\phi_{T+1} = \phi_T$, and then $\varepsilon_Q(\phi_T; \pi_T) \le \varepsilon$ with π_T the lifted equilibrium of $\hat{G}_{\phi_T}$: the pair (ϕ_T, π_T) is a joint fixed point to which Theorem 2 applies with tolerance ε. (Building the partition against a fixed reference profile and re-solving once does *not* give this: the re-solved equilibrium generally exposes new within-class differences; the tabular experiments show ε_Q jumping by an order of magnitude after a single re-solve.)

B IS THE LOOP NEEDED, AND WHAT DOES IT COST?

Table 4: One D2 partition against a fixed reference profile (the exact ground equilibrium $\pi^\star$, an oracle start) versus the monotone loop from the one-state partition. ε_Q is measured on the returned (partition, re-solved equilibrium) pair; Theorem 2 needs $\varepsilon_Q \le \varepsilon$. Exact.

		distractor game (1520)			plain game (760)		
ε	variant	$\lvert\mathcal{Z}\rvert$	ε_Q attained	EXP	$\lvert\mathcal{Z}\rvert$	ε_Q attained	EXP
0.05	one-shot at $\pi^\star$	738	1.478	0.053	445	0.777	0.012
	re-partition $\times 3$, not monotone	909	1.469	0.041	455	1.494	0.022
	monotone loop	1250	**0.040**	0.012	605	**0.028**	0.000
0.1	one-shot at $\pi^\star$	390	1.571	0.336	221	1.532	0.544
	re-partition $\times 3$, not monotone	455	1.569	0.368	301	1.527	0.300
	monotone loop	649	**0.097**	0.058	605	**0.088**	0.000
0.2	one-shot at $\pi^\star$	198	1.529	0.929	132	1.181	0.894
	re-partition $\times 3$, not monotone	314	1.510	0.768	217	0.968	0.692
	monotone loop	548	**0.192**	0.092	469	**0.195**	0.127

Without the loop the tolerance is violated by a factor of 4.8 to 30 in every cell, and repeating the partition without monotonicity does not repair it. What the loop buys is that ε means what Definition 1 says it means for the profile that is played. It does not follow that one-shot partitions are useless as heuristics: at $\varepsilon = 0.05$ on the distractor game the one-shot partition (738 states, $\mathrm{EXP} = 0.053$) is comparable to the loop's partition at $\varepsilon = 0.1$ (649 states, 0.058); they are simply not partitions to which either theorem applies at the tolerance they were built for.

Table 5: Cost of the loop against solving the ground game, as the payoff-irrelevant variable takes n_d values ($\delta = 0.05$, $\varepsilon = 0.1$). The dominant cost of exact solving is one matrix-game LP per state per sweep; the loop solves LPs on abstract states only and needs from the ground game exact best responses, which are single-agent solves without LPs. Reference implementation, cold-started from the one-state partition, every abstract game solved from $V = 0$. The ground solve returns an exact equilibrium, the loop an approximate one.

n_d	$\lvert\mathcal{S}\rvert$	method	$\lvert\mathcal{Z}\rvert$	matrix games solved	vs. ground	EXP (exact)
2	1520	ground solve	1520	112,480	–	0
		D1 loop / D2 loop	760 / 649	115,810 / 263,839	$1.03\times$ / $2.35\times$ more	0.058
4	3040	ground solve	3040	209,760	–	0
		D1 loop / D2 loop	760 / 680	109,794 / 311,370	$1.91\times$ fewer / $1.48\times$ more	0.049
8	6080	ground solve	6080	413,440	–	0
		D1 loop / D2 loop	760 / 680	106,786 / 252,500	$3.87\times$ / $1.64\times$ fewer	0.030
16	12160	ground solve	12160	826,880	–	0
		D1 loop / D2 loop	760 / 702	99,266 / 380,206	$8.33\times$ / $2.17\times$ fewer	0.014
32	24320	ground solve	24320	1,629,440	–	0
		D1 loop / D2 loop	760 / 709	94,754 / 409,468	$17.2\times$ / $3.98\times$ fewer	0.007

The loop's LP count does not grow with the ground state space when the growth is payoff-irrelevant (D1: 95–116 thousand for n_d from 2 to 32), while the ground solve grows linearly; the rows for $n_d = 16$ and 32 were predicted before they were run (D1 about $7.7\times$ and $15\times$: observed $8.3\times$ and $17.2\times$; D2 about $3\times$ and $6\times$: observed $2.2\times$ and $4.0\times$, because it needed ten and eleven iterations). The loop is therefore more expensive than solving the ground game when there is little to discard ($n_d = 2$; D2 up to $n_d = 4$) and cheaper beyond, at the price of an approximate equilibrium. We did not tune the loop for cost (no warm starts, abstract games re-solved from scratch at every iteration).

Table 6: The D2 loop driven by *truncated* deviators: splits use the exact action values of the greedy response after k value-iteration sweeps, which is what an under-trained deviator knows. Everything reported is exact; "true ε_Q" is measured with exact best responses on the returned pair. $\varepsilon = 0.1$.

	plain game (760)			distractor game (1520)		
k	$\lvert\mathcal{Z}\rvert$	gap	true ε_Q	$\lvert\mathcal{Z}\rvert$	gap	true ε_Q
1	556	0.043	0.604	1345	0.031	0.604
3	558	0.022	0.489	1213	0.027	0.218
10	598	0	0.113	690	0.073	0.100
30	605	0	0.088	649	0.073	0.097
exact	605	0	0.088	649	0.073	0.097

How good must the best responses be? Thirty sweeps reproduce the exact loop to the state, and ten are already within 13% of the target tolerance (Table 6). With one or three sweeps the loop stops at a pair whose true distortion is two to six times the tolerance it believes it has, the exact counterpart of what Appendix K found with learned deviators, where the measured exploitability moved by a factor of 25–30 with the deviator's budget. We had predicted, in advance, that weak deviators would also give *coarser* partitions; that is true on the plain game and false on the distractor game, where they over-split ($1{,}345$ states against 649) and fail to discard d.

Related loops, and the payoff-relevant state. Like double oracle (McMahan et al., 2003), the loop alternates restricted solves with full-game best responses and stops when these change nothing, but it refines a state partition instead of adding the best responses to the players' strategy sets. Brown & Sandholm (2015) also interleave abstraction and equilibrium finding, adding actions to an extensive-form game's action abstraction during CFR, with convergence results for two-player zero-sum games; the loop instead splits state classes whose ground best-response values differ. The payoff-relevant state of Maskin & Tirole (2001) is the coarsest partition of histories within whose cells each player's preferences agree, up to a positive affine transformation, against every measurable strategy of the others: being profile-independent, it can keep a variable that a given equilibrium ignores, which D1 and D2, defined against one profile, can drop.

C Measured distortion variants (no bound is claimed)

Why we measured them. ε_Q is a supremum over all within-class pairs, including states the equilibrium never reaches. On the partitions our loop returns, an occupancy-weighted distortion is 8–$17\times$ smaller (e.g. 0.0084 against 0.0969 at $\varepsilon = 0.1$ on the distractor game), which would place the Theorem 2 expression at 3.4 rather than 38.7, below the largest attainable gap, and so informative, though still $58\times$ the exact value. Restricting instead to the support does not help: where enough classes contain two reached states to make the quantity meaningful it is only 1.5–$1.6\times$ smaller. *We prove no bound with either quantity*; these are measurements on the returned partitions, offered as the reason we think a visitation-weighted transfer theorem, in the style of concentrability arguments in single-agent RL, is the most promising way to make the constants matter.

Table 7 reports all of them. For each partition the D2 loop returns, all three quantities are measured against the *same* lifted equilibrium: ε_Q as Definition 1 gives it (a supremum over all within-class pairs), the same supremum restricted to states with positive discounted occupancy under that equilibrium, and the occupancy-weighted mean spread. The last two appear in no theorem in this paper; the T_2 columns show what the Theorem 2 expression would give *if* a bound of the same shape held, next to the exact exploitability.

Table 7: Distortion variants and the Theorem 2 expression evaluated at each. "classes with 2+ reached" is the number of classes containing at least two states of positive occupancy: where it is small, the support column is zero for want of anything to compare and means nothing.

			reached	classes	ε_Q			T_2 evaluated at		
game	ε	$\lvert\mathcal{Z}\rvert$	states	2+ reached	sup	support	weighted	sup	support	weighted
plain	0.05	605	97/760	3	0.028	0.000	0.000	11.3	0.0	0.0
plain	0.1	605	86/760	2	0.088	0.000	0.000	35.3	0.0	0.0
plain	0.2	469	36/760	4	0.195	0.077	0.012	78.1	30.7	4.7
distractor	0.05	1250	222/1520	6	0.040	0.025	0.005	16.0	10.0	2.0
distractor	0.1	649	166/1520	82	0.097	0.064	0.008	38.7	25.4	3.4
distractor	0.2	548	156/1520	54	0.192	0.123	0.011	76.6	49.3	4.5

The exact exploitabilities of these six rows are 0, 0, 0.127, 0.012, 0.058 and 0.092, so even the weighted column remains 37–167$\times$ above the truth where the truth is non-zero. The point is only that it falls below $NV_{\max} \approx 40$, which the supremum version does not.

D A POSTERIORI (MEASURED) VERSIONS OF THE TWO UPPER BOUNDS

Both proofs bound an intermediate quantity that can simply be computed once the abstract game is solved. *Theorem 2:* Step 2 of its proof bounds $\kappa_i' = \sup_{s,a} |\hat{Q}_i^\star(\phi(s), a) - Q_i^\star(s, a)|$ by $\varepsilon_Q/(1-\gamma)$ and the remaining steps use only κ_i'; reading it off is Theorem 2′ (Appendix A), $V_i^\star(s|\pi_{-i}) - V_i^\pi(s) \le (\hat{\varepsilon}+2\kappa_i')/(1-\gamma)$. *Theorem 1:* Lemmas 3 and 4 bound $\|V_i^\pi - \hat{V}_i^{\hat{\pi}} \circ \phi\|_\infty$ and $\|V_i^\star - \hat{V}_i^\star \circ \phi\|_\infty$ by Δ_1. Let T_i^π and $T_i^\star$ be the policy-evaluation and optimality operators of the MDP that player i faces against π_{-i}. Each of the two errors is at most $1/(1-\gamma)$ times the Bellman residual of the lifted abstract value, $\delta_i = \|T_i^\pi(\hat{V}_i^{\hat{\pi}} \circ \phi) - \hat{V}_i^{\hat{\pi}} \circ \phi\|_\infty$ and $\delta_i^\star = \|T_i^\star(\hat{V}_i^\star \circ \phi) - \hat{V}_i^\star \circ \phi\|_\infty$, so $\mathrm{EXP}_G(\pi) \le \mathrm{EXP}_{\hat{G}_\phi}(\hat{\pi}) + \sum_i(\delta_i + \delta_i^\star)/(1-\gamma)$. Both inequalities are the standard contraction bound $\|x^\star - x\|_\infty \le \|Tx - x\|_\infty/(1-\gamma)$, applied to two different lifted functions and joined through the abstract gap. Pérolat et al. (2017, Theorem 1) bound an L_p version of the Nash gap of an N-player general-sum Markov game by the evaluation and best-response residuals of one value function per player, with a concentrability coefficient; the sup-norm analogue follows from the pointwise inequality in their proof. When $\hat{\pi}$ is an exact equilibrium of $\hat{G}_\phi$, $\hat{V}_i^{\hat{\pi}} = \hat{V}_i^\star$ and $\mathrm{EXP}_{\hat{G}_\phi}(\hat{\pi}) = 0$, and the bound above is that form at $v_i = \hat{V}_i^{\hat{\pi}} \circ \phi$; for other abstract profiles it is not an instance of their result. The second expression costs one backup per ground state and no ground best response; the first needs the ground best-response values.

What they give, and what they do not. We re-examined 64 partitions of §5 without re-running any experiment (18 of them rebuilt by the released code and accepted only when they reproduced the reported rows); Table 8 shows rows fixed by rule before it was built. On the 48 partitions with non-zero exploitability the measured Theorem 1 value is 10–43$\times$ the exact exploitability (median 17), where the a priori value is 26–1.6$\cdot 10^5\times$ (median 1,400); for Theorem 2 the figures are 27–520$\times$ (median 50) against 250–12,000$\times$ (median 640). What remains is at least of the order of $1/(1-\gamma)$, which a contraction argument does not remove. Two caveats. First, the one-class partition has exact exploitability 2.86 on every game used here, and the measured Theorem 1 value is at or above that on 20 of the 41 lossy partitions other than the one-class partition itself, including all 15 on the plain game (a priori: 34 of 41; Theorem 2: 32 measured, 41 a priori): there it does not tell the partition from no abstraction at all. Second, it vanishes on every partition that is lossless by construction, but not on every partition whose lifted equilibrium is unexploitable: $\phi^{Q^\star}$ at $\varepsilon = 0.05$ has exact exploitability 0 and a measured value of 1.15, because the residuals ask whether the lifted *values* solve the ground equations, not whether the lifted *strategies* are optimal. Neither quantity is used by the loop.

Table 8: Exact exploitability against the a priori and the measured value of each upper bound. Rows: every partition of Table 2 other than the identity, every head-to-head row of the original five-tolerance grid, and the two rows cited in the text ($\phi^{Q^\star}$ at 0 and 0.05). The one-class partition has exploitability 2.86 on both games.

				Theorem 1		Theorem 2	
game	partition (ε)	$\lvert\mathcal{Z}\rvert$	exact	a priori	measured	a priori	measured
distractor	D1 (0.1, 0.3)	760	0.058	2.00	1.00	37.2	2.86
distractor	D2 (0.05)	1,250	0.012	757	0.180	16.0	0.887
distractor	D2 (0.1)	649	0.058	758	1.00	38.7	3.11
distractor	D2 (0.2)	548	0.092	758	3.55	76.6	6.11
distractor	D2 (0.3)	537	0.174	758	5.11	118	8.83
distractor	D2 (0.5)	436	0.287	776	7.92	200	21.7
plain	$\phi^{Q^\star}$ (0.0)	636	0	720	0	0	0
plain	$\phi^{Q^\star}$ (0.05)	588	0	720	1.15	79.0	6.86
plain	$\phi^{Q^\star}$ (0.1)	378	0.107	760	4.61	348	33.7
plain	$\phi^{Q^\star}$ (0.2)	278	0.353	760	8.78	340	33.6
plain	$\phi^{Q^\star}$ (0.4)	217	0.549	760	10.6	302	32.1
plain	$\phi^{Q^\star}$ (0.6)	138	1.49	760	15.1	369	49.2
plain	$\phi^{Q^\star}$ (1.0)	20	2.66	760	30.4	776	70.8
plain	D2 (0.1)	605	0	720	0	35.3	1.44
plain	D2 (0.2)	469	0.127	720	2.97	78.1	6.34
plain	D2 (0.4)	366	0.476	720	7.28	157	14.8
plain	D2 (0.6)	316	0.525	760	12.8	237	23.5
plain	D2 (1.0)	37	1.49	760	19.8	397	47.8

E FRONTIERS, INSERTION ORDER, AND A METHODOLOGICAL DIFFERENCE

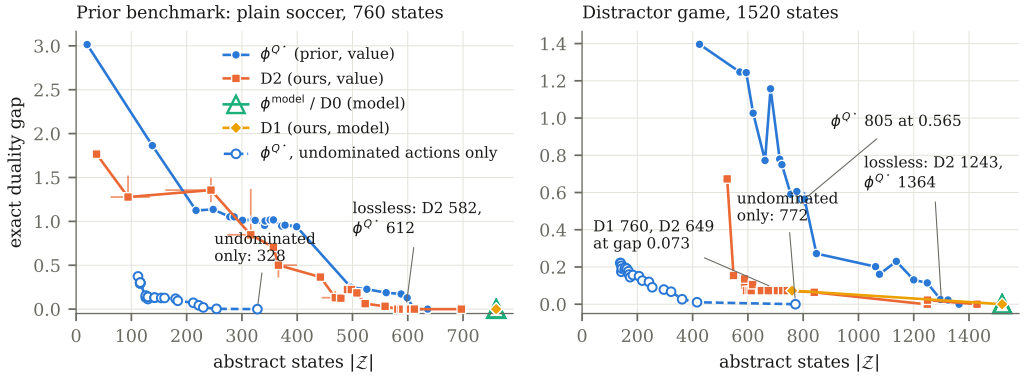


Figure 2: Exact duality gap against abstraction size. *Left:* the prior work's benchmark; D2 is lossless with fewer states and better at light and moderate compression, $\phi^{Q^\star}$ is better for $\lvert\mathcal{Z}\rvert \approx 220$–$280$. *Right:* with a payoff-irrelevant variable, both model-level profile-independent criteria stay at the identity (open triangle), D1 halves the state space and D2 goes further at the same gap; D2 lies below $\phi^{Q^\star}$ at every sampled size. *Dashed, open circles:* $\phi^{Q^\star}$ compared only on the joint actions surviving IESDS (profile-independent; §5.2), below both criteria on both games. Whiskers: range over five random greedy insertion orders, where measured.

Robustness to the greedy order. All partitions are built by greedy complete-linkage splitting, which depends on insertion order. Over five random orders the entries of Table 3 move little and never change sign: $\phi^{Q^\star}$ at $\varepsilon = 0$ is 636 in every order and D2 at $\varepsilon = 0.1$ is 590–609 at zero gap; on the distractor game D2 is 649–662 and D1 exactly 760, both at gap 0.073, while $\phi^{Q^\star}$ is 805–830

at gaps 0.53–0.60. The table reports the default order, which is the smallest draw for D2 on the distractor game (13–15% fewer states than the hand design across orders) and also the smallest for $\phi^{Q^\star}$.

Where the prior criterion wins. On the plain game the frontiers cross: for $|\mathcal{Z}| \approx 220$–$280$, $\phi^{Q^\star}$ attains gaps 1.05–1.14 where D2 has 1.36 at 244 states. The crossing is not an artifact of the greedy merge order: over five random insertion orders $\phi^{Q^\star}$ is stable (at each tolerance $|\mathcal{Z}|$ varies by at most 4 across random orders, the default order lying up to 11 states away; gaps 1.05–1.17), while D2 at the tolerances that land in this band is order-sensitive ($|\mathcal{Z}|$ from 163 to 225 at $\varepsilon = 0.7$, gaps 1.23–1.49) and under no order beats $\phi^{Q^\star}$ there. In this band the profile-independent criterion is both better and more stable. Below it the default-order frontiers cross back (D2: 94 states at gap 1.28; $\phi^{Q^\star}$: 138 at 1.86), but D2 there ranges over 64–132 states across orders and we draw no conclusion. The advantage of D2 that survives every order is at light compression.

A methodological difference. Ishibashi et al. (2025) compute $Q^\star$ and gaps by minimax Q-learning (10^6 iterations); their abstraction at $\varepsilon = 0$ has 760 states. With exact $Q^\star$ their criterion already merges to 636 at $\varepsilon = 0$, consistent with estimation noise removing exact ties. We implement two of their four criteria; the other two carry bounds exponential in $1/(1-\gamma)$.

F TRANSFER ACROSS GAMES

Table 9 gives the exact exploitability of the jointly refined partition and of the hand design on each held-out game.

Table 9: Exact exploitability on held-out games (§5.4).

held-out game	learned (joint)	hand design
$\delta = 0.2$ (outside training range)	**0.039**	0.159
$\delta = 0.05$, i.i.d. distractor	**0.030**	0.078
$\delta = 0.05$, permanent distractor	**0.012**	0.065
$\delta = 0$	0.000	0.000

The joint partition (1,253 states) is refined on $\delta \in \{0.02, 0.05, 0.1\}$ at once and evaluated on each held-out game by re-solving the abstract game only; both it and the hand design are exact at $\delta = 0$. Refining each held-out game on its own gives 596–1,239 states, and on $\delta = 0.2$ reaches $\mathrm{EXP} = 0$ with 1,239. Refining only against the currently worst game yields a smaller partition (932 states) that is 2.6–$4.8\times$ more exploitable on the held-out games than the joint one and barely improves on the hand design, which is why we split on every game of the family at once.

G SAMPLE EFFICIENCY: PROTOCOL AND CONTROLS

Fixed in advance. A test game that differs only in d is solved by the training equilibrium with no samples at all, so we required that the training equilibrium have a duality gap of at least 0.2 on the test game (it has 2.1). All learners see the same nested stream, drawn as random permutations of all state–action triples (the best design for the ground-model learner); a pair with no sample falls back to the learner's exact training row. $n^\star(\tau)$ is the first sample size from which the median over seeds stays at or below τ, interpolated in $\log n$; $\tau \in \{0.1, 0.03\}$; $\rho = n^\star_{\text{ground model}}/n^\star_{\text{D1}}$ with the ground-model learner credited with the better of its two defaults (exact training row, or an absorbing zero-reward state, for a pair with no sample); the prior criterion credited with its best tolerance of five at every n; intervals from a paired bootstrap over the 20 seeds. *H1:* $\rho \in [8, 32]$ at both tolerances under both metrics, against either default. *H2:* over $N \in \{2, 4, 8, 16\}$ the log-log slope of $n^\star$ in N lies in $[0.7, 1.3]$ for the ground model and in $[-0.15, 0.15]$ for D1. *H3:* the $\phi^{Q^\star}$ envelope needs at least $4\times$ D1's samples or never reaches τ. *Falsified if* $\rho < 2$ at either tolerance, any baseline comes within $2\times$ of D1, the ground model's slope is below 0.5, or the effect appears under one default or

one metric only. The gate (duality gap on the test game of the training game's ground equilibrium ≥ 0.2) reads 2.10–2.11 on the four games.

Table 10 reports the samples each learner needs on the held-out game, and Table 11 how that requirement grows with the number of values of the irrelevant variable.

Table 10: Samples needed on the held-out game, $N = 16$ (4,224 ground states; 20 seeds; 15 sample sizes from 3.2×10^3 to 10^7). "Asymptote" is the exact value with infinitely many samples; "never" means not within 10^7.

			exploitability at μ_0		max-state gap	
learner	$\lvert\mathcal{Z}\rvert$	asymptote	$n^\star(0.1)$	$n^\star(0.03)$	$n^\star(0.1)$	$n^\star(0.03)$
ground model (either default)	4,224	0	984,508	1,996,537	never	never
$\phi^{Q^\star}$, best tolerance in hindsight	433–3,632	0.016–1.29	613,329	6,513,238	never	never
random partition	264	2.44	never	never	never	never
D2	211	0.003	54,189	110,722	never	never
D1	264	0	75,701	153,010	443,332	3,492,557

Table 11: The ratio ρ and the requirement $n^\star(0.1)$ against the number N of values of the irrelevant variable (exploitability at μ_0; 95% bootstrap intervals).

	$N = 2$	$N = 4$	$N = 8$	$N = 16$	slope in N
ground model, $n^\star(0.1)$	146,785	253,960	489,724	984,508	0.92
D1, $n^\star(0.1)$	91,587	87,424	83,816	75,701	−0.09
ρ at $\tau = 0.1$	1.6 (1.3–2.8)	2.9 (2.5–5.0)	5.8 (4.0–9.5)	13.0 (9.2–17.4)	
ρ at $\tau = 0.03$	2.1 (1.1–3.1)	3.2 (2.2–4.1)	7.3 (3.5–8.4)	13.0 (10.6–16.4)	
ρ at $\tau = 0.1$, max-state gap	1.8 (1.4–2.9)	5.5 (3.3–7.9)	10.6 (8.3–15.6)	undefined	

Outcome, clause by clause. *H1* holds on the primary metric ($\rho = 13.0$ at both tolerances, identical under both defaults) and *cannot be evaluated* on the max-state gap at $N = 16$: the ground-model learner never reaches either tolerance within 10^7 samples while D1 does, which bounds ρ below by 22.6 and 2.9 but does not place it in the registered band. *H2* holds wherever it is defined: slopes 0.92 and 0.84 for the ground model and -0.09 and -0.07 for D1 on the primary metric, 1.24 and -0.09 at $\tau = 0.1$ on the max-state gap. *H3* holds at $N = 16$ (8.1× and 42.6×; never on the max-state gap). No falsifier fires at $N = 16$, the game for which the band was written; the registration did not say at which N the "within 2×" falsifier applies, and at $N = 2$ and 4 baselines are within 2× of D1 (1.3–2.0×), as a ratio that grows like $0.8N$ implies. D2 is not the arm to rely on: its training partition leaves an exploitability of 0.07 on the new core for $N \leq 8$, so it reaches 0.1 (with 1.5–2.5× D1's samples) and never 0.03; at $N = 16$ its partition is different, its asymptote 0.003, and it needs 0.7× D1's samples. On the max-state gap D2 and $\phi^{Q^\star}$ never reach either tolerance at any N. Seventeen of the 12,000 estimated games made the LP solver fail after two retries with explicit methods, all for $\phi^{Q^\star}$ (15) or the random partition (2), none for the ground model, D1 or D2; they enter the medians as missing.

Two controls, settings and predictions fixed in advance ($N = 16$, 20 seeds, the same sample streams). *Partitions computed from the test game's exact model*: a ceiling on what partition discovery could save, not evidence of paying for itself. D1's partition is the same core either way, so its requirement is unchanged (75,701 and 153,010) and ρ stays at 13.0 (9.1–16.9 and 9.7–16.1): the factor against the ground model owes nothing to where the partition was learned. The prior criterion, now with partitions that are admissible on the test game, reaches 0.03 about five times sooner (1.4×10^6 samples) and its handicap against D1 falls from 8.1× and 42.6× to 4.1× and 8.9×: part of it was a stale partition, and what remains is the price of keeping every value of d. Both predictions held. *A test game that differs from the training game only in the dynamics of d.* As declared in advance, this design is void: the training equilibrium is exactly optimal with no samples at all (gate 0.00). What it shows instead is that samples first *hurt* a learner that falls back on an exact prior: the median exploitability rises from 0 to 0.52 for D1 (at 1.8×10^4 samples) and to 0.74 for the ground model (at 1.8×10^5), and is back under 0.03 after 7.6×10^5 and 1.4×10^6 samples. We had predicted

that the ground model's peak would be at least three times D1's; it is 1.4 times, and the prediction failed. Nine of the $6{,}000$ estimated games of the two controls made the LP solver fail, seven for $\phi^{Q^\star}$ and two for the random partition.

Dominance-aware and loop-free baselines (G_N, $N = 16$, the same 20 seeds and streams; registered with its predictions before the runs). New learners: $\phi^{Q^\star}$ compared only on IESDS survivors (tolerances 10^{-9}, 0.05, 0.2), D0 on IESDS survivors, one pass of D2's distance at the training game's exact equilibrium (no loop), and the D2u loop (§5.2). All four drop d; the first three also merge core states by the training game's values, and on the changed core that is fatal: with the exact test model their partitions leave an exploitability of 0.79 (D0 on IESDS survivors), 1.07–1.41 ($\phi^{Q^\star}$ on IESDS survivors) and 1.74 (one pass), and none reaches $\tau = 0.1$ within 10^7 samples. D2u's partition leaves 7×10^{-5} and needs 8.3×10^4 and 1.8×10^5 samples for $\tau = 0.1$ and 0.03, against D1's 7.6×10^4 and 1.5×10^5; D2 needs 5.4×10^4 and 1.1×10^5. Of three predictions, the one about D2u held; the other two, which put 0.1 and 0.15 on a dominance-aware learner or the one-pass partition reaching $\tau = 0.1$, did not occur.

The side-bet family (H_N on the plain 3×4 core, test core changed as above; $N = 4$, because at $N = 16$ a game with 225 joint actions could not finish in time; 10 seeds; the same grid; registered with four predictions before the runs). The gate holds (2.31). The ground model reaches $\tau = 0.1$ after 5.4×10^6 samples, $\phi^{Q^\star}$ at tolerance 0.05 after 2.9×10^6 (at 10^{-9} and 0.2 never); D1, D2 and D2u never do within 10^7 (medians 0.216, 0.217, 0.217; D1 above 0.15 on every seed), although with the exact test model their partitions leave at most 10^{-5}. No learner reaches 0.03. Their medians first rise, from 1.6–2.0 with no samples to 5–6.6 up to 10^5 samples, then fall (1.5 at 10^6, 0.22 at 10^7, still falling). The dominance-aware partitions and one pass of D2 are not admissible on the test core (0.38–1.41) and never reach 0.1. The gate held; the three predictions that compared learners failed: that the best profile-independent learner and the ground model need at least twice D1's samples, and that D2u needs at most 1.5 times D1's. Why pooling costs the loop's partitions more here than on G_N is open. One candidate was registered and tested: the estimator weighs each ground member of a class by its number of visits, which differ by one within a class, and on H_N the cancellation the family rests on needs equal weights. Re-running every learner with equal weights within each class changes nothing material: D1, D2 and D2u still never reach $\tau = 0.1$ within 10^7 (medians 0.197, 0.209, 0.209 against 0.216, 0.217, 0.217), the ground model and $\phi^{Q^\star}$ still reach it (5.3×10^6 and 2.8×10^6), and on G_N D1's requirement is unchanged ($75{,}701$), so the weighting is not the cause. Differences of about a percent between the two runs are the solver's choice among equilibria: for singleton classes, where the two weightings coincide, the estimates differ by 10^{-16} and the ground model's requirement still moves by 1.2%. The constructions differ in that on G_N the variable matters only after TAUNT, which no equilibrium plays, while on H_N it changes the payoffs of played joint actions and only the played mixture cancels it.

H THE SEPARATION FAMILY, CHECKED EXACTLY

G_0 is soccer with TAUNT and no wind; G_N adds $d \in [N]$, which switches with probability 0.1 per step independently of play, and sends the game to a target core state c_d under (TAUNT,TAUNT). Table 12 gives the abstract state counts each criterion reaches on this family, and Table 13 the premises and conclusions of Proposition 6 measured exactly. Targets are chosen from the solved G_0 alone: among its distinct minimax values take the N nearest to N evenly spaced values between the smallest and the largest, and represent each by its first core state. The construction, the grid (3×4 with $N \in \{2, 4, 8, 16\}$, 4×4 with $N \in \{2, 4, 8\}$), the checks and what would refute the proposition were written down before the games were built. (The first target rule, value quantiles, failed its own guard: many states are worth exactly 0, so targets shared values. It was replaced before any affected result existed; all seven games use the rule above.)

Table 12: Abstract states on the separation family G_N (3×4 pitch, 264 core states; tolerances in parentheses; "0" is 10^{-9}). D1, D2 and the restricted criteria are lossless in every column.

criterion	$N = 2$	$N = 4$	$N = 8$	$N = 16$
ground game	528	1056	2112	4224
D0, ϕ^{model} (0 and 0.1)	528	1056	2112	4224
$\phi^{Q^\star}$ (0.02)	454	908	1816	3632
D1 loop (0.01)	264	264	264	264
D2 loop (0.05)	243	239	240	238
D2 loop with re-clustering (0.05)	196	196	202	211
compared only on IESDS survivors / on minimax supports				
$\phi^{Q^\star}$ (0)	**145 / 87**	**145 / 87**	**145 / 87**	**145 / 87**
D0 (0)	172	172	172	172

Table 13: Premises and conclusions of Proposition 6, measured exactly. "Mixed" counts classes containing two values of d over the six profile-independent partitions built per game.

pitch	N	$\lvert\mathcal{S}\rvert$	$\max \lvert V^\star - V_0^\star \rvert$	$\max P(\text{TAUNT})$	drop-d: $\varepsilon_r^1, \varepsilon_P^1, \varepsilon_Q$, gap	ε_P^0	$\gamma\Delta_{\min}$	mixed	$\phi^{Q^\star}$ (0.02)	D1 / D2 (0.05)
3×4	2	528	3×10^{-17}	0	all $\le 10^{-15}$	1.00	1.800	0	454	264 / 196
3×4	4	1056	3×10^{-16}	0	all $\le 10^{-15}$	1.00	0.572	0	908	264 / 196
3×4	8	2112	3×10^{-16}	0	all $\le 10^{-14}$	1.00	0.220	0	1816	264 / 202
3×4	16	4224	7×10^{-16}	0	all $\le 10^{-14}$	1.00	0.030	0	3632	264 / 211
4×4	2	960	1×10^{-16}	0	all $\le 10^{-8}$	1.00	1.800	0	856	480 / 404
4×4	4	1920	3×10^{-16}	0	all $\le 10^{-8}$	1.00	0.600	0	1712	480 / 409
4×4	8	3840	4×10^{-16}	0	all $\le 10^{-8}$	1.00	0.192	0	3424	480 / 408

One clause we had written down in advance fails, and is reported as a failure: that profile-independent sizes are *exactly* N times the one-slice size. It holds in 37 of 42 rows; the other five are $\phi^{Q^\star}$ at exactly $\varepsilon = 0$, where the size *exceeds* N times the one-slice size (926 rather than 4×227). The slices' $Q^\star$ differ by 3×10^{-16} from rounding in value iteration, which at $\varepsilon = 0$ breaks ties differently in each slice; at $\varepsilon = 10^{-12}$ every slice has 227 classes. Proposition 6 claims "at least", which holds in every row.

Random games. The construction takes any two-player zero-sum core game with a stay action; with the soccer core our generic implementation reproduces the family above (identical rewards, kernel within 10^{-15}). We drew 20 random cores (30 states, 4 actions per player of which one is σ with $r(\cdot, \sigma, \sigma) = 0$, three successors per state–action pair with Dirichlet weights) and built G_N for $N \in \{2, 4, 8, 16\}$: 80 games, none dropped or re-drawn, checks fixed in advance. In all 80, $\max \lvert V^\star(c, d) - V_0^\star(c) \rvert \le 1.3 \times 10^{-15}$, no equilibrium puts mass on τ, dropping d has $\varepsilon_r^1 = \varepsilon_P^1 = \varepsilon_Q = 0$ and zero gap while $\varepsilon_P^0 = 1$; none of the 480 partitions built by D0, ϕ^{model} (tolerances 0 and 0.05) or $\phi^{Q^\star}$ (tolerances 0 and $\gamma\Delta_{\min}/2$) contains a class with two values of d, and each has exactly as many classes as the ground game has states ($30N$); the D1 loop, started from one class, returns 30 classes and is lossless in all 80.

Restricted to undominated actions, G_N separates nothing profile-relative. On all 87 games above, $\phi^{Q^\star}$ and D0 compared only on IESDS survivors or only on minimax supports, at tolerance 10^{-9}, put every core state's N copies in one class and are lossless (largest gap 3×10^{-9}): 145, 87, 172 and 172 states on 3×4 and 210, 210, 248 and 248 on 4×4 at every N, and the 30-state core on the random games. That was predicted when the baseline was registered, from the proof of Proposition 6: TAUNT is iteratively strictly dominated and only the (TAUNT,TAUNT) entry depends on d.

The side-bet family H_N (Proposition 7; construction, checks and predictions fixed before any game was built). Each player gets three side actions and $\Delta = 1$; the N matrices are drawn at random, double-centered and scaled to sup-norm 1, and kept only if each of them and their mean have the uniform strategies as their only equilibrium (an exact linear-programming test) and they are pairwise at least 0.5 apart, so $\Delta\beta/2 \ge 0.25$; d switches with probability 0.1 per step. Cores:

the 20 random cores above, $N \in \{2, 4, 8, 16\}$, and plain 3×4 soccer, $N \in \{2, 4, 8\}$: 83 games, none dropped or re-drawn. Profile-independent criteria at tolerance 0 (10^{-9}) and 0.05, one fixed value below 0.25 (the first draft of the protocol used $\Delta\beta/4$, which merged core states lossily on a smoke-test game, and was replaced before any registered game was built). On all 83 games: $\max |V^\star(c, d) - V_0^\star(c)| \leq 1.6 \times 10^{-14}$ and every computed equilibrium plays the side actions uniformly (to 2.4×10^{-12}); dropping d has $\varepsilon_r^1 = \varepsilon_P^1 = \varepsilon_Q = 0$ and zero gap; IESDS and the minimax supports remove no side action anywhere; no class of any of the 14 profile-independent partitions contains two values of d. On the random cores every one of them equals the ground game ($30N$ states), and the D1 and D2 loops return the 30-state core, losslessly, in all 80 games; on the soccer core they return 264 and 248–250 states at every N (Table 14). All six registered predictions held, among them that every profile-independent size grows at least $6\times$ from $N = 2$ to 16 on the random cores (median 8.00) and at least $3\times$ from 2 to 8 on the soccer core (4.00), while the loops grow at most $1.25\times$ (1.00; 0.99).

Table 14: Abstract states on H_N, plain 3×4 soccer core (264 states). *: not lossless (gap in parentheses); every other entry has gap $\leq 10^{-13}$. From $N = 2$ to $N = 8$ every profile-independent partition grows by exactly $4\times$.

criterion (ε)	$N = 2$	$N = 4$	$N = 8$
ground game; D0, ϕ^{model} (0 and 0.05)	528	1056	2112
$\phi^{Q^\star}$ (0)	454	908	1816
$\phi^{Q^\star}$ (0.05)	422* (0.019)	844* (0.019)	1688* (0.020)
$\phi^{Q^\star}$ on IESDS survivors (0)	290	580	1160
$\phi^{Q^\star}$ on IESDS survivors (0.05)	246* (0.010)	492* (0.011)	984* (0.011)
$\phi^{Q^\star}$ on minimax supports (0)	174* (0.092)	348* (0.092)	696* (0.102)
D0 on IESDS survivors or supports (0 and 0.05)	344	688	1376
D1 loop (0.01)	**264**	**264**	**264**
D2 loop (0.05)	**250**	**250**	**248**

I THE PRIOR BENCHMARK AT OTHER PITCH SIZES

Plain Markov Soccer (no TAUNT, no distractor, $\gamma = 0.9$) at nine pitch sizes, default insertion order, in three pre-registered batches. First, five sizes on coarse grids ($\phi^{Q^\star}$ at $\varepsilon \in \{0, 0.02, 0.05, 0.1, 0.2\}$, D2 at $\varepsilon \in \{0.02, 0.05, 0.1, 0.2\}$); then, once 5×5 came out against D2, four more chosen to separate "odd number of rows" from "square pitch" (the generator puts the goal on rows $\{\lfloor r/2 \rfloor - 1, \lfloor r/2 \rfloor\}$ when r is even and on the middle row alone when it is odd, so parity of r is the width of the goal); finally both criteria on one common grid, $\varepsilon \in \{0.00, 0.01, \ldots, 0.20\}$, on all nine pitches and the distractor game, so that neither criterion is judged on a grid the other did not get, with the read-out and four predictions fixed before the runs (all four held). Table 15 is the common grid. "Lossless" means an exact duality gap of zero to solver precision (10^{-6}); the right-hand block asks the same of EXP at μ_0 instead. 5×3 is not a legal pitch in this generator: both players start on the same cell.

Table 15: Smallest lossless partition of each criterion on one common grid of 21 tolerances (the tolerance in parentheses): zero duality gap at every state (left block) and zero EXP at μ_0 (right block). **Bold**: the smaller. The split is by the width of the goal, which is the parity of the number of rows. The coarse-grid batches give the same direction on every pitch.

		zero gap at every state			zero EXP at μ_0		
pitch	$\lvert\mathcal{S}\rvert$	$\phi^{Q^\star}$	D2	fewer states	$\phi^{Q^\star}$	D2	fewer states
two-cell goal (even number of rows)							
4×4	480	424 (0.04)	**400 (0.16)**	D2, 5.7%	412 (0.06)	**400 (0.16)**	D2, 2.9%
4×5 (benchmark of prior work)	760	612 (0.03)	**582 (0.13)**	D2, 4.9%	588 (0.05)	**523 (0.15)**	D2, 11.1%
4×6	1104	876 (0)	**819 (0.1)**	D2, 6.5%	797 (0.03)	**784 (0.11)**	D2, 1.6%
6×5	1740	1392 (0.02)	**1366 (0.05)**	D2, 1.9%	1356 (0.03)	**1274 (0.12)**	D2, 6.0%
one-cell goal (odd number of rows)							
3×4	264	207 (0.06)	**201 (0.19)**	D2, 2.9%	**151 (0.19)**	201 (0.19)	$\phi^{Q^\star}$, 24.9%
3×5	420	**281 (0.07)**	286 (0.18)	$\phi^{Q^\star}$, 1.7%	**169 (0.19)**	286 (0.18)	$\phi^{Q^\star}$, 40.9%
5×4	760	**547 (0.08)**	591 (0.16)	$\phi^{Q^\star}$, 7.4%	**445 (0.19)**	567 (0.17)	$\phi^{Q^\star}$, 21.5%
5×5	1200	**755 (0.07)**	899 (0.13)	$\phi^{Q^\star}$, 16.0%	**507 (0.19)**	802 (0.17)	$\phi^{Q^\star}$, 36.8%
5×6	1740	**981 (0.06)**	1184 (0.11)	$\phi^{Q^\star}$, 17.1%	**553 (0.2)**	1000 (0.2)	$\phi^{Q^\star}$, 44.7%
distractor game (§5.2)	1520	1364 (0)	**1243 (0.02)**	D2, 8.9%	1249 (0.05)	**1243 (0.02)**	D2, 0.5%

Table 16 separates criterion from search: the same D2 distance under three different search procedures.

Table 16: Smallest sampled lossless partition under different D2 *searches* (same criterion, same tolerances, same greedy builder). "Loop" is the method of §4. "Re-cluster" re-clusters all states against the current lifted equilibrium and refines again; that it would reach or beat $\phi^{Q^\star}$ everywhere was predicted in advance and is false. The last column was designed after seeing the others: D2 refinement started from $\phi^{Q^\star}$'s own partition, a re-clustering round kept only if the exact gap does not grow. **Bold**: smaller than $\phi^{Q^\star}$. A dash: no sampled partition is lossless.

pitch	$\lvert\mathcal{S}\rvert$	$\phi^{Q^\star}$	loop	loop, re-cluster	from $\phi^{Q^\star}$, gap-safe
4×4	480	428	**410**	**404**	**410**
4×5	760	636	**605**	**605**	**614**
4×6	1104	876	**819**	–	**817**
6×5	1740	1392	**1366**	**1320**	**1315**
3×4	264	211	**201**	**202**	**168**
3×5	420	289	355	**285**	**192**
5×4	760	575	661	592	**481**
5×5	1200	795	983	915	795
5×6	1740	1001	1270	1165	1001
smaller than $\phi^{Q^\star}$			5 of 9	5 of 9	7 of 9 (2 equal)
tolerances ahead (of 36)			19	28	27

The inequality of Proposition 5 was evaluated on every partition built for this table, under every search and for both criteria: it holds on 342 of 342. At matched gap on the common grid D2 is ahead at 69 of 84 tolerances on the even pitches and at 2 of 105 on the odd ones. At μ_0 alone both criteria go further on the benchmark ($\phi^{Q^\star}$ to 588 states, D2 to 523) while leaving gaps of 0.17 and 0.06 off the equilibrium path; on a one-cell goal a profile-independent criterion at a large tolerance merges everything off the equilibrium path and is judged only on it. D2's tolerance is not monotone in size at the lossless end (1,243 states at $\varepsilon = 0.02$ on the distractor game, 1,249 at 0.03): the partition depends on the path of the loop. Admissibility at tolerance zero implies losslessness; at the positive tolerances where these partitions were found, losslessness is measured, not implied.

The dominance-aware baseline on the common grid. At every state, IESDS (strict dominance by pure or mixed strategies, one linear program per test) is run on the stage game $Q^\star(s, \cdot, \cdot)$ of the ground minimax values, or the union of the supports of its minimax strategies is taken; $\phi^{Q^\star}$ and D0 then compare two states only if their surviving joint-action sets are equal, and only on them. On the

plain game IESDS keeps between 1 and 20 of the 25 joint actions at a state, on the distractor game between 1 and 20 of 36, and it removes TAUNT, for both players, at all 1,520 states. Smallest lossless partitions: $\phi^{Q^\star}$ restricted to IESDS survivors 328 (plain) and 772 (distractor), to minimax supports 332 and 772, D0 restricted either way 390 and 780, all at tolerance 10^{-9} (at exactly 0 floating-point noise of order 10^{-16} breaks ties); D2 at matched gap is ahead of the IESDS-restricted criterion at 0 of 21 tolerances on both games. The registered predictions: that the restricted criterion's smallest lossless partition on the distractor game lies between D2's 1,243 and $\phi^{Q^\star}$'s 1,364, and on the plain game within 5% of $\phi^{Q^\star}$'s 612, both *failed*; so did D2 ahead at matched gap at 11 or more tolerances. That IESDS removes TAUNT everywhere held.

Every value-level criterion on the ten games, and equilibrium selection (Table 17; registered with read-out and predictions before the runs). D2u is the loop of §4 with the split rule of Proposition 10 and $m = 2\varepsilon/(1-\gamma)$. Registered: that D2u is smaller than D2 on the benchmark held; that it reaches the IESDS-restricted criterion there or on the distractor game, is ahead of it at matched gap at 11 or more tolerances, or is smaller on five or more pitches, all *failed* (0 of 21 tolerances, 0 of 9 pitches); that it drops d on every separation game held (7 of 7 and 83 of 83). An exact search over 400 random zero-sum games found no violation of Proposition 10 in 32,400 checks (655 with a positive spread; largest gain 0.024 of the bound) and no case where dropping the margin condition broke the conclusion either. *Selection:* the loops rerun on the same games with each player's actions listed in reverse order - only the LP solver's choice among equilibria can change - return different partitions at 15–21 of 21 tolerances; D2's smallest lossless partition moves by at most 2% ($582 \to 573$ on the benchmark), D2u's by up to 11% ($574 \to 513$; $1{,}149 \to 1{,}067$ on the distractor game).

D2u and one pass, against D2 and the dominance-aware baseline. The loop on the criterion of Proposition 10 (D2u, $m = 2\varepsilon/(1-\gamma)$) is never larger than D2's at the lossless end and is smaller on seven of the ten games (574 against 582 states on the benchmark, 1,149 against 1,243 on the distractor game, 1,170 against 1,366 on 6×5), and it still drops d on all 7 G_N and all 83 H_N games. But $\phi^{Q^\star}$ on IESDS survivors is smaller than D2u on all ten games, by 30–45%. One pass of D2's distance at the exact ground equilibrium, with no loop, is lossless at 566 states on the benchmark against the loop's 582, but lossless at all on only five of the ten games, none with a one-cell goal; on the distractor game the loop is ahead at matched gap at 15 of 21 tolerances. What the loop buys is the tolerance holding for the equilibrium played, not size.

Table 17: Smallest partition with zero duality gap at every state, per criterion, on the common grid of 21 tolerances (tolerance in parentheses; "0" is 10^{-9}; "–": no sampled partition is lossless). One pass: a single complete-linkage pass on D2's (D2u's) distance at the exact ground equilibrium, no loop.

	profile-relative				profile-independent			
game	D2 loop	D2u loop	one-pass D2	one-pass D2u	$\phi^{Q^\star}$	$\phi^{Q^\star}$, IESDS	$\phi^{Q^\star}$, supports	D0, IESDS
3×4	201 (0.19)	184 (0)	–	–	207 (0.06)	101 (0.09)	**87** (0)	172 (0)
3×5	286 (0.18)	286 (0.18)	–	–	281 (0.07)	195 (0)	**103** (0)	262 (0)
4×4	400 (0.16)	300 (0)	380 (0.01)	–	424 (0.04)	**210** (0)	**210** (0)	248 (0)
4×5	582 (0.13)	574 (0.01)	566 (0.01)	–	612 (0.03)	**328** (0)	332 (0)	390 (0)
4×6	819 (0.10)	786 (0)	791 (0.01)	674 (0.01)	876 (0)	**468** (0)	**468** (0)	564 (0)
5×4	591 (0.16)	560 (0)	–	–	547 (0.08)	**379** (0)	–	476 (0)
5×5	899 (0.13)	899 (0.13)	–	–	755 (0.07)	**513** (0)	–	726 (0)
5×6	1184 (0.11)	1184 (0.11)	–	–	981 (0.06)	**647** (0)	–	1026 (0)
6×5	1366 (0.05)	1170 (0)	1287 (0)	–	1392 (0.02)	**758** (0)	–	890 (0)
distractor	1243 (0.02)	1149 (0.01)	1143 (0.01)	978 (0)	1364 (0)	**772** (0)	**772** (0)	780 (0)

The same search for both criteria, and the policy-level floor. Two experiments registered on the same day with their read-outs and predictions, both exact. *(i) Gap-oracle agglomeration* (Table 18). Start: a criterion's smallest zero-gap partition on the common grid. Repeat: propose the pair of classes at the smallest complete-linkage distance under one criterion ($\phi^{Q^\star}$: the largest difference of ground minimax $Q^\star$ over joint actions, fixed; D2: the largest difference of exact best-response values over players and own actions against the *current* lifted equilibrium, recomputed after every kept merge), merge them, re-solve the abstract game from scratch, lift, and keep the merge iff the exact duality gap is at most 10^{-6}; a rejected pair is never proposed again; 400 oracle calls per cell;

two starts, two orders, ten games. Registered: D2's order ends smaller than $\phi^{Q^\star}$'s in at least 14 of the 20 (game, start) cells: *failed*, 11; from its own start with its own order D2 ends smaller on at least 6 of 9 pitches: *failed*, 5; merges kept per call higher under D2's order: held, 0.44 against 0.38. D2's order wins on every two-cell-goal pitch and on the distractor game, $\phi^{Q^\star}$'s on every one-cell-goal pitch but 5×6 from $\phi^{Q^\star}$'s start. Four cells (all D2's order from D2's start) met a block-LP failure of the solver on a merged game and were resumed from their checkpoints after the solver learned to split the block; nothing else changed. *(ii) The policy-level floor.* States are merged iff both players' exact equilibrium strategies agree to 10^{-6} (a game version of Li et al. (2006)' $\pi^\star$-irrelevance, which asks only for one optimal action shared by a class), and the abstract game is re-solved, lifted and scored as everywhere else: 18–140 classes, and a gap of 1.0–2.4 on every one of the ten games (rewards in $[-1, 1]$), with ε_Q at the ground equilibrium 1.8–1.9. D2's refinement started from that partition is lossless on all ten but smaller than the loop's own at the same tolerance on two. Two of four predictions failed: at most 40 classes on every plain pitch (57–132 on the two-cell-goal pitches, where the equilibrium is mixed at more states), and the refinement from the floor smaller than the loop's own on at least five games (two).

Table 18: One search for both criteria. Start: the criterion's smallest zero-gap partition on the common grid; order: whose distance proposes the next merge; a merge is kept iff the exact gap stays zero; 400 oracle calls per cell. Entries: final $|\mathcal{Z}|$ (merges kept). **Bold**: the smaller of the two orders from the same start. $\pi^\star$: classes and gap of the policy-level floor; last column: D2 refinement from it at $\varepsilon = 0.01$.

			start $\phi^{Q^\star}$			start D2			
pitch	$\lvert\mathcal{S}\rvert$	$\pi^\star$: classes, gap	$\lvert\mathcal{Z}\rvert$	order $\phi^{Q^\star}$	order D2	$\lvert\mathcal{Z}\rvert$	order $\phi^{Q^\star}$	order D2	D2 from $\pi^\star$
two-cell goal (even number of rows)									
4×4	480	57, 1.16	424	356 (68)	**295 (129)**	400	347 (53)	**293 (107)**	418
4×5	760	75, 1.32	612	499 (113)	**475 (137)**	582	466 (116)	**425 (157)**	623
4×6	1104	89, 1.01	876	804 (72)	**637 (239)**	819	695 (124)	**614 (205)**	860
6×5	1740	132, 1.74	1392	1297 (95)	**1112 (280)**	1366	1206 (160)	**1125 (241)**	1393
one-cell goal (odd number of rows)									
3×4	264	18, 1.55	207	**40 (167)**	56 (151)	201	**30 (171)**	89 (112)	241
3×5	420	22, 1.33	281	**140 (141)**	141 (140)	286	**107 (179)**	124 (162)	350
5×4	760	21, 2.38	547	**399 (148)**	449 (98)	591	**377 (214)**	414 (177)	689
5×5	1200	26, 2.14	755	**605 (150)**	621 (134)	899	**562 (337)**	766 (133)	1020
5×6	1740	25, 1.96	981	818 (163)	**752 (229)**	1184	**822 (362)**	963 (221)	1415
distractor game	1520	140, 1.20	1364	1247 (117)	**1090 (274)**	1243	1149 (94)	**1055 (188)**	1263

Both criteria were run from the same ground solution by the same code, so the reversal is not an implementation difference. It is also not the greedy insertion order: on the benchmark pitch $\phi^{Q^\star}$ moves by at most 4 states across five random orders (§5.2). Nor is it the tolerance grid: the common grid of Table 15 gives both criteria the same 21 tolerances and leaves the direction unchanged on all nine pitches relative to the coarse batches (on 4×5, 605 vs. 636 there and 582 vs. 612 here). The margins do move (5×4 from 13% to 7%), which is why the percentages should be read as properties of the grid. Every row of every batch is in the supplementary code.

J A GAME WE DID NOT DESIGN: OSHI-ZUMO

Oshi-Zumo is a two-player simultaneous-move bidding game from the computer-games literature; we use it because it has mixed equilibria and no structure we put there. Rules as implemented: each player starts with C coins and a wrestler stands at position 0 of a field $\{-K, \dots, K\}$. Every round both players bid simultaneously, $b_i \in \{0, \dots, C\}$, a bid above a player's remaining coins counting as all of them; the higher bid moves the wrestler one step towards the other player's end, equal bids leave it, and both pay their bids. The game ends when the wrestler is pushed off the field (the pusher wins, payoff ± 1) or when neither player has coins left (the player on whose opponent's side the wrestler stands wins; a draw at 0); the payoff is received on that transition, after which the game sits in an absorbing state with reward 0; $\gamma = 0.9$. Instances: $C = 8$, $K = 2$ (406 states, 81 joint actions, 122 states whose ground equilibrium is mixed) and $C = 10$, $K = 3$ (848 states, 121 joint actions, 274 mixed). Every value-level criterion of the paper was run on the grid of 21 tolerances of Appendix I, with its predictions registered before the runs (Table 19). No criterion dominates. Of

four predictions, three held - D2u below every profile-independent restricted criterion on one game, D2 or D2u ahead of the IESDS-restricted $\phi^{Q^\star}$ at matched gap at 11 or more tolerances on one, D2 below plain $\phi^{Q^\star}$ on both - and one failed: that the IESDS-restricted $\phi^{Q^\star}$ is smaller than D2 on both games (a tie at 315 on the smaller).

Table 19: Oshi-Zumo: smallest partition with zero duality gap at every state, per criterion (tolerance in parentheses; "0" is 10^{-9}; "–": none lossless), and on how many of the 21 tolerances D2 / D2u are ahead of the IESDS-restricted $\phi^{Q^\star}$ at matched gap.

game	$\phi^{Q^\star}$	$\phi^{Q^\star}$, IESDS	$\phi^{Q^\star}$, supports	D0, IESDS	one-pass D2	D2 loop	D2u loop	ahead: D2 / D2u
$C = 8$, field 5 (406 states)	322 (0.09)	315 (0.07)	**259** (0)	350 (0)	282 (0.06)	315 (0.08)	310 (0)	11 / 12
$C = 10$, field 7 (848 states)	698 (0.09)	690 (0)	–	750 (0)	**626** (0.07)	697 (0.09)	678 (0)	8 / 9

K DEEP INSTANTIATION: PROTOCOL AND FULL RECORD

Summary. We instantiated the loop with MAPPO (Yu et al., 2022) on JaxMARL's MPE `simple_tag` (Rutherford et al., 2024) (three predators against one prey, per-agent distractor channels, held-out layouts), with PPO deviators retrained against the current lifted profile at every outer step; the rest of this appendix holds the full record. **The profile-relative method does not survive function approximation**: learned re-embeddings never matched raw MAPPO, even at eight times its budget, and with learned deviators the D2-over-D1 advantage did not reproduce, as the exact analysis predicts (the D2 loop needs deviators good to about ten value-iteration sweeps; Table 6). What survives is a profile-independent (D0-style) deletion of observation variables, chosen without deviators: it beats raw MAPPO head-to-head (prey on all 24 held-back seeds and predators on 22 at raw's budget, where raw's prey has not yet learned by our own criterion; both roles on three fresh seeds at 2×10^7 steps, where it has), a two-timescale solver behind it is less exploitable at raw's budget only, and at 2×10^7 steps the deletion alone is more exploitable than raw. It is not the loop of §4.

The record in full.

We instantiated the loop with MAPPO (Yu et al., 2022) on JaxMARL's MPE `simple_tag` (Rutherford et al., 2024), three cooperating predators against one prey, per-agent distractor channels, held-out layouts: a latent solver that sees only z, PPO deviators on raw observations retrained against the current lifted profile at every outer step, and an encoder trained on their signals. **Re-embedding the observation never matched raw MAPPO; deleting observation variables with a profile-independent (D0-style) test did, and beat it.** Protocol, registrations, failures: Appendix K. Returns are held-out, lower bounds use $K' = 10^6$ steps per deviator, 3 seeds unless stated.

D1 needs no best responses; D2 needs good ones. Exactly, the D2 loop tolerates truncated deviators down to about ten value-iteration sweeps (Table 6); learned deviators were far from that regime (the lower bound of one fixed profile moved from ≈ 3 at $K' = 10^5$ to ≈ 80 at $K' = 10^6$), and adding the D2 signal to a D1-only encoder made it a weaker player and, on training layouts, more exploitable on every seed at two budgets (5.4 ± 0.7 against 3.3 ± 0.7; indistinguishable on held-out layouts). We read Table 2's D2-over-D1 advantage as a result about accurate best responses (an interpretation: deviator quality was not varied).

Re-embedding fails, and not for want of budget. Every learned encoder is a much weaker player than raw MAPPO (9.3–11.2 against 30.5 at eight times the budget). Their lower exploitability (4.8–8.9 against 14.2 ± 1.6 on training layouts) is no win: a weak profile is cheap to make unexploitable by one deviator at a time. When two or three predators deviate together, the D1-only encoder's bound rises from 8 to 20 to 44, while raw's and the filter's stay at 14 and 10 (Table 21).

Deletion carries over, and pays. Dropping a variable, the abstraction of Proposition 6, becomes the gate $\phi_g(o) = g \odot o$: MAPPO on z is the latent solver, and its lifted profile is the latent equilibrium that raw deviators attack (all arms at raw's budget of 3.1×10^6 steps, scored head-to-head against the raw profile of their seed; Table 20). The confirmed arm is policy-free and early: after 12.5% of the budget it keeps the inputs whose association with the reward, or with another input's next value, exceeds a permutation null; no threshold is chosen. On 24 seeds held back for it, it deletes all 64

distractors and at most three landmark coordinates; its prey beat raw's on all 24 and its predators on 22 ($+7.0$, $+3.4$), the criterion fixed beforehand (every seed, both roles), failing in one batch of four. Raw MAPPO is a weak reference at this budget: by the learning criterion of Appendix K its prey has not learned, and gets worse against uniform predators on all 24 seeds. At 2×10^7 steps raw passes the gate (3 of 3 fresh seeds), and the filter still beats it head-to-head in both roles on every seed (predators $+1.4$ to $+1.9$, prey $+0.9$ to $+5.0$) but is *more* exploitable (unilateral bound 1.8–5.0 against 0.7–1.1; Appendix K). At raw's budget the filter's bound is not reliably lower than raw's or the D1-only encoder's ($p = 0.06$, 0.87 on fifteen seeds; a significant gap on the first six vanished).

A better solver, which needs the abstraction. All of raw's and the filter's bound is the prey's: self-play leaves it far from its best response. A two-timescale schedule (cf. Daskalakis et al., 2020) (two small prey-only updates after each joint one, same budget) was run on nine fresh seeds with its tests fixed in advance. Behind the filter the bound falls to 3.0, below raw's 15.8 and the D1-only encoder's 10.4 ($t_8 = -4.8$, -3.8; 3.2 against 16.7 and 45.1 with team deviations). On raw observations it changes nothing (15.7): it works only behind the abstraction ($t_8 = -7.4$, nine seeds of nine). The prey's gain is most of every arm's bound and the extra updates train the prey, so part of the fall is the metric being trained on; not so the predators' team gain (-0.7; 39 for the D1-only encoder) or the arm's rank at each of three attack budgets (Appendix K). The price is competence: the prey beats raw's by 13.6 on every seed, the predators trail raw's by 2.0, on seven of nine; a gate on that had failed in development (Appendix K). All of this is at 3.1×10^6 steps: at 2×10^7 raw alone is less exploitable (0.7–1.1), and the solver was not run there.

What the deep result is not. It is not the loop of §4: the test that chose g asks whether an input matters to the model at all, which is D0, all a criterion needs on distractors independent of play; profile-relativity is not shown to pay, and deviators took no part in choosing the inputs (Appendix K).

Environment. JaxMARL MPE `simple_tag` (3 predators, 1 prey, 5 actions, 25-step episodes), subclassed so that a *game descriptor* (arena size, 1–3 landmarks, speeds) lives in the state; 20 training and 5 held-out descriptors from a fixed seed; m_d payoff-irrelevant sticky channels per agent (resampled with probability 0.1 per step). A unit test reproduces JaxMARL step for step on the default game; distractors provably never enter physics or rewards; descriptors never enter observations. The environment was frozen by a tag before any comparison, and every results directory records the tree hash of the environment code.

Protocol. Every gate criterion was written down before the corresponding run; no seed was dropped, re-run or selected; every attempt, including failures, is in the repository with commit, configuration, seeds and best-response budget. Best responses: PPO on raw observations against the frozen lifted profile, initialized at the incumbent plus a zero residual, best checkpoint by validation episodes, scored on disjoint test episodes with common random numbers. Returns are undiscounted episodic; "$\pm$" is the standard deviation over 3 seeds.

Build steps and their gates.

Raw MAPPO learns ($m_d = 8$)	Criterion: both sides improve against a fixed uniform opponent. Failed at 3×10^6 steps (prey: $0/3$ seeds), $2/3$ at 10^7, passed at 2×10^7 ($3/3$, $z \geq 9.8$); twelve remediation settings failed first and are reported.
Raw MAPPO learns ($m_d = 64$), applied afterwards	The same criterion, applied afterwards to the recorded curves of every $m_d = 64$ run (registered before it was applied): at 3.1×10^6 steps it fails on all 24 held-back seeds (the predators improve against a uniform prey, $z = 24$–29; the prey gets *worse* against uniform predators, $z = -4.5$ to -8.9) and at 2×10^7 it passes ($3/3$; prey $-14.5 \to -9.4$). The summary at the head of this appendix compares against the former.
Latent solver, frozen random encoder	Criterion: clearly worse than raw. VQ-256 passes (2.4–3.1 vs. 23.4–25.5); a continuous encoder *fails* it (21.9–23.7): it is near-lossless, not leaking; structural tests show the policy depends on an observation only through z. Quantizing a random embedding destroys locality, not information: a linear probe recovers the task coordinates from the codes with $R^2 = 0.75$–0.98, and a 4,096-code variant does no better than 256.
Deviators	Gain positive and non-decreasing in the budget on every seed ($5 \to 42 \to 77$ at $K' = 2, 5, 10 \times 10^5$); a collapsed encoder ($z \equiv 0$) has a measured lower bound of 85–88 where raw MAPPO at 2×10^7 steps and $m_d = 8$ has ≈ 0, also at $K' = 5\times10^6$; the comparisons of that summary use $m_d = 64$ at 3.1×10^6 steps, where raw fails the competence gate (second row).
Encoder losses	First attempt collapsed the codebook to one code in 100 updates (two-sided distance loss, distances ≈ 12 vs. targets ≈ 1). With one-sided losses: gain falls across outer steps for D1+D2 and D1+D2+IB but not for D2 alone, and the distractors-off control fails (latent 24 return units below raw in both roles); unchanged by $4\times$ and $16\times$ wider codebooks. The encoder must also be the slow timescale: performance recovers monotonically as encoder steps per outer step go from 100 to 5.

Matched-budget comparison ($m_d = 64$, observation 82-dim, latent 32; raw 3.1×10^6 steps, latent arms 2.4×10^6).

profile	exploitability lower bound, $K' = 10^6$ training layouts	held-out layouts	effective latent dims	held-out predators
raw MAPPO	14.2 ± 1.6	17.7 ± 0.9	–	29.4–31.2
frozen random projection	5.9 ± 0.4	9.6 ± 1.2	22.8	6.5 ± 0.6
reconstruction (VAE)	8.2 ± 0.4	not run	23.3	6.8 ± 0.3
D1 only (DBC-like)	4.8 ± 0.5	9.6 ± 0.6	19.2	7.2 ± 0.9
D2 only	8.9 ± 1.1	not run	13.1	6.5 ± 0.6
D1+D2+IB	8.1 ± 1.4	11.8 ± 4.1 (8.9, 9.0, 17.6)	12.7	6.7 ± 0.6
hard-gated ($z = g \odot o$), D1+D2+IB	23.9–36.4	not run	8–12 open gates of 82	13.7–24.4

Variable deletion at the raw baseline's budget (3.1×10^6 steps, $m_d = 64$, observation 82-dim). Each arm is compared with the *raw profile of its own seed* on held-out layouts: its predators against that profile's prey, and that profile's predators against its prey. "SE" is of the difference, from the two evaluations. Seeds 0–2 were used to develop these arms; seeds 3–5, 6–8, 9–11, 12–20 and 21–29 were each run once, for the rows that name them, after the criterion was fixed.

The three arms were designed in one night, in this order, each after seeing the previous one; only the last was scored on seeds held back for it. The "profile does not react" gate measures, every five updates, how much each input changes the played profile's action distribution and value, and closes at most eight of the inputs below both thresholds; it cannot be used before the profile is any good. The early test (given in full below, "The input filter in full") replaces it with a ridge regression of the reward and the other inputs' next values on $[o, o^2, \text{action}]$, pooled over the burn-in, and keeps input j when the Herfindahl index of its normalized association profile exceeds γ/K. It missed one live coordinate on one seed, which is reported rather than re-tuned. Both gates keep the distractor count they close stable across seeds; neither uses a deviator.

Table 20: Deleting observation variables, against the raw profile of the same seed. Differences are the arm minus that raw profile; "vs. uniform" is the predators' return against a uniform-random prey, a measure of competence rather than of the match-up.

gate, and when it is decided	inputs kept	predators vs. that seed's raw prey	prey vs. its raw predators	vs. uniform	$K' = 10^6$ bound
none (raw MAPPO), seeds 0–2	82	29.4, 31.2, 31.0	−44.2, −47.2, −44.0	40.7	17.7 ± 0.9
profile does not react, after 50%	25–30	+1.7, +2.2, +4.2 (2.1–5.0 SE)	+5.9, +9.6, +2.5 (3.3–11.9 SE)	44.1	13.3 ± 5.2
oracle (distractors), from the start	18	+5.5, +3.5, +5.0 (4.1–6.7 SE)	+12.4, +15.9, +12.1 (15–20 SE)	50.3	not run
none (raw MAPPO), seeds 3–5	82	31.4, 31.5, 31.0	−43.9, −44.6, −44.0	42.2	12.4 ± 3.6
association test, after 12.5%	18–24	+4.5, +0.3, +5.4 (mean 5.2 SE)	+4.2, +5.8, +5.2 (mean 8.3 SE)	44.3	13.0 ± 5.2
none (raw MAPPO), seeds 6–8	82	30.7, 31.7, 30.7	−41.7, −45.3, −42.6	40.7	13.7 ± 6.4
permutation null, after 12.5%	15–16	+4.3, +4.2, +4.1 (mean 6.4 SE)	+5.0, +9.3, +7.3 (mean 12.1 SE)	44.6	6.4 ± 1.5
D1-only encoder (DBC), seeds 6–8	latent 32	a much weaker player: predators 6.9–7.2 against the raw prey			11.6 ± 1.3
D1+D2+bottleneck encoder, seeds 6–8	latent 32	a much weaker player: predators 5.8–6.4 against the raw prey			11.5 ± 2.7
none (raw MAPPO), seeds 9–11	82	28.6, 27.9, 31.2	−44.2, −40.2, −44.8	40.9	13.0 ± 4.5
permutation null, after 12.5% (m_0 inputs)	14–17	+5.9, +6.3, +4.2 (mean 8.4 SE)	+8.7, +6.3, +5.8 (mean 11.3 SE)	45.6	7.4 ± 2.1
the $\lceil m_0/2 \rceil$ strongest inputs	7–9	−21.3, +2.6, −2.2	+17.8, −7.4, −0.3	38.9	24.2 ± 15.1
the $2m_0$ strongest inputs	28–34	+3.6, +5.4, +5.1 (mean 7.3 SE)	+4.0, +3.6, +5.7 (mean 7.2 SE)	43.5	12.2 ± 6.2
depth chosen by deviators (half, m_0, twice)	7, 15, 34	−21.3, +6.3, +5.1	+17.8, +6.3, +5.7	–	4.6 ± 1.4
D1-only encoder (DBC), seeds 9–11	latent 32	a much weaker player: predators 6.6–7.5 against the raw prey			8.6 ± 0.7
D1+D2+bottleneck encoder, seeds 9–11	latent 32	a much weaker player: predators 6.2–7.6 against the raw prey			9.8 ± 0.6
none (raw MAPPO), seeds 12–20	82	26.2–33.5	−44.7 to −41.2	39.5	14.2 ± 3.5
permutation null, after 12.5%	15–16	ahead on 7 of 9, mean +2.3 (6.1 SE)	ahead on 9 of 9, mean +6.3 (18.5 SE)	44.6	11.4 ± 8.8
D1-only encoder (DBC), seeds 12–20	latent 32	a much weaker player: predators 5.4–7.4 against the raw prey			9.9 ± 2.2
none (raw MAPPO), seeds 21–29	82	28.2–32.7	−46.7 to −41.8	39.9	15.8 ± 4.2
permutation null, after 12.5%	15–17	ahead on 9 of 9, mean +3.7 (9.6 SE)	ahead on 9 of 9, mean +7.5 (21.5 SE)	43.9	7.1 ± 4.2
with the two-timescale solver	15–17	ahead on 2 of 9, mean −2.0 (−5.5 SE)	ahead on 9 of 9, mean +13.6 (40.6 SE)	40.9	3.0 ± 4.4
none, with the two-timescale solver	82	ahead on 0 of 9, mean −7.9 (−23.0 SE)	ahead on 4 of 9, mean +0.2 (0.5 SE)	34.2	15.7 ± 3.0
D1-only encoder (DBC), seeds 21–29	latent 32	a much weaker player: predators 6.1–7.5 against the raw prey			10.4 ± 1.6

The permutation null. The exclusivity statistic, its threshold γ/K, the accumulation rate and the burn-in follow Anonymous (2026); none was tuned on this task. On the development seeds that threshold sits on the edge of a narrow useful window ($\gamma = 1.5$ keeps 57–61 distractors, $\gamma = 3$ starts to cut task coordinates). We replaced it by a rule without a threshold: input j is kept iff its largest association exceeds the largest association found in each of ten shuffles of the targets against the inputs, pooled over the burn-in. A different parameter-free rule that we had fixed first (cut at the widest gap in the sorted Herfindahl indices) closed four to six task coordinates on the development seeds and is reported as a failed idea. The permutation rule was run once on seeds 6–8 and once more on seeds 9–11, under the criterion of the row above it. One prediction made for it failed: the task coordinates it closes are not confined to the set seen in development (a landmark coordinate that never closed there closes on four of the six seeds). Over the first six seeds its exploitability lower bound was below raw's on every seed (paired $t_5 = -3.3$, $p = 0.02$). We then ran a third batch of nine seeds, registered in advance as the last look, with one test: the filter against DBC over all fifteen seeds, two-sided, at $\alpha = 0.025$. It failed ($t_{14} = -0.2$, $p = 0.87$; below DBC on ten seeds of fifteen), and the comparison with raw lost its significance too ($t_{14} = -2.1$, $p = 0.06$; below on twelve). For raw and the filter the measured exploitability is the prey's alone (no predator gains more than 0.3 by deviating on its own); against DBC each predator also gains, 1.2 on average. On three of the nine new seeds a learned prey gains 17–29 against the filter's predators, which beat raw's against a uniform-random prey on all fifteen seeds; so the filter trades a more competent predator team for one that is, on some seeds, easier to evade. On seeds 6–11 the full learned encoder (D1 and D2 signals with a bottleneck) measures 10.7 on average, indistinguishable from DBC's 10.1 ($t_5 = 0.4$) and above the filter's 6.9 ($t_5 = 3.4$); its predators score 5.8–7.6 against the raw prey, where raw's score 28–32.

Raw and the filter at 2×10^7 steps (fresh seeds 30–32; registered with its read-out and four predictions before the runs). Settings are those of the batch above with only the budget, the seeds and the output directories changed; the filter's burn-in stays 12.5% of the budget. Raw passes the learning criterion (first two rows of the build-step table at the start of this appendix) on all three seeds (predicted: at least two). The filter deletes all 64 distractors on every seed and keeps 18, 17 and 18 of the 18 task inputs. Head-to-head against raw's profile of the same seed, its predators gain $+1.41$, $+1.66$, $+1.87$ and its prey $+0.90$, $+4.99$, $+2.88$ (predicted: ahead on at least two seeds in each role; both held). Unilateral bounds (the attack protocol of this appendix, $K' = 1{,}001{,}472$, held-out layouts): raw 0.74, 0.84, 1.12; the filter 5.00, 1.83, 3.11; the prediction that the filter's would be below raw's on at least two seeds *failed* on all three. Every positive gain but one ($+0.24$, a predator) is the prey's. Three seeds are a description, not a test.

The input filter in full. The filter as it was run; nothing in it uses which inputs are distractors. Let $m = 82$ be the dimension of an agent's observation, $\mathcal{A}$ the five actions, and $\Gamma \in \{\text{predators}, \text{prey}\}$ a group.

Data. During the burn-in (the first 12.5% of the budget, 24 ordinary MAPPO updates with every input open), the rollout of each update yields, for every agent of Γ, transitions (o, a, r, o'). Transitions that end an episode are dropped, because their o' belongs to the next episode, and $20{,}000$ of the rest are drawn without replacement. Every such batch is standardized column by column; constant columns are set to zero.

Associations. With features $x = [\, o,\ o \odot o,\ e_a \,] \in \mathbb{R}^{2m+|\mathcal{A}|}$, where e_a is the one-hot action, and targets $y = [\, o',\ r \,] \in \mathbb{R}^{m+1}$, the batches are pooled through their sufficient statistics $S = \sum X^\top X$ and $C = \sum X^\top Y$, and

$$B = (S + \lambda n I)^{-1} C, \qquad \lambda = 10^{-3}, \qquad a_{jk} = \sqrt{B_{j,k}^2 + B_{m+j,k}^2}\ \ (k \neq j), \qquad a_{jj} = 0,$$

with n the number of pooled transitions. a_{jk} is the association of input j with target k: the coefficient of o_j and that of o_j^2 together, the square catching symmetric dependence such as a penalty on the distance from the center. An input's own next value is excluded: an autonomous noise channel predicts itself perfectly and nothing else.

Read-out (i): the permutation null, the arm confirmed on seeds 6–29. For $p = 1, \ldots, 10$ the rows of Y are permuted within every batch, $C^{(p)} = \sum X^\top Y_{\pi_p}$, and $a^{(p)}$ is computed from $(S, C^{(p)})$ as above. With $\nu = \max_p \max_{j,k} a_{jk}^{(p)}$, input j is kept iff $\max_k a_{jk} > \nu$ for at least one group. The

decision is taken once, at the end of the burn-in, from the statistics pooled over it. No threshold is chosen; the number of permutations sets the level of the test.

Read-out (ii): exclusivity, the earlier arm, seeds 3–5. Let

$$H_j = \sum_k \Big(a_{jk} \Big/ \sum_{k'} a_{jk'} \Big)^2 \in [1/K,\, 1]$$

be the Herfindahl index of input j's association profile over its K targets. After every burn-in update t the reward $r_{t,j} = \max_\Gamma \mathbf{1}[H_j > \gamma/K]$ is computed from the statistics pooled so far and accumulated as a bandit accumulates the reward of an arm, each input being an arm: $\rho_j \leftarrow \rho_j + \eta\,(r_{t,j} - \rho_j)$ from $\rho_j = 0$; input j is kept iff $\rho_j > \frac{1}{2}$ at the end of the burn-in. The statistic and its constants ($\gamma = 2$, $\eta = 0.1$, burn-in ratio 0.125) follow Anonymous (2026); none was tuned on this task. The statistic is pooled over the burn-in because, computed batch by batch, it closes live task coordinates; and the accumulated score is used once, as a hard gate.

Gate. With $g \in \{0,1\}^m$ the indicator of the kept inputs, each actor receives $g \odot o$ and the centralized critic the concatenation of all agents' $g \odot o$; one mask serves both groups, an input being kept if either group's test keeps it. Training continues to the end of the budget, and the filter spends no environment step. The stored profile has the mask folded into its first-layer weights (the rows of closed inputs are zero), so it is an ordinary raw-observation profile and is attacked and evaluated as one.

Depth variants (used only where deviators chose the depth). Inputs are ordered by $s_j = \max_\Gamma \max_k a_{jk}/\nu$, so that read-out (i) keeps $\{j : s_j > 1\}$, m_0 inputs; "half" keeps the $\lceil m_0/2 \rceil$ largest s_j and "twice" the $2m_0$ largest.

Procedure (early input filter with read-out (i)).

1. $g \leftarrow \mathbf{1}$; for each group, S, C, $C^{(1)}, \ldots, C^{(10)} \leftarrow 0$.
2. For $t = 1, \ldots, 24$: one MAPPO update on $g \odot o$; then, for each group, draw 20,000 non-terminal transitions of that update's rollout, standardize them, and add $X^\top X$, $X^\top Y$ and $X^\top Y_{\pi_p}$ to S, C and $C^{(p)}$.
3. For each group compute a, the $a^{(p)}$ and ν; set $g_j \leftarrow \mathbf{1}\big[\exists\, \Gamma :\ \max_k a^\Gamma_{jk} > \nu^\Gamma\big]$.
4. Continue MAPPO on $g \odot o$ until the budget is spent; fold g into the first layer.

Deviators and the gate. Two ways of letting deviators set the gate, each fixed before it ran. *(i) A forward gate.* Every input starts closed and opens only when the values of a one-step deviator (a per-action regression of the return-to-go on the raw observation, from the solver's own rollouts, at no cost in environment steps) vary with it across actions more than they vary with a *ghost* input: the same agent's observation in another, independent game of the same rollout, which has exactly the law of the real input and is independent of this game's play. Once each input is studentized by its own ghosts the null is calibrated (no distractor opened in ten tests of 64), but the test has no power where it matters: under a blind profile it never detects the opponents' offsets, and on the raw profile's own rollouts a third of the way through training the prey's test still does not. A policy gated this way starves (on a throw-away seed, held-out predator return 7.0 after 30 updates); a fitted-Q version with the same features is no better. *(ii) Depth selection.* The association test fixes the order of the inputs; candidates keep the strongest $\lceil m_0/2 \rceil$, the m_0 that pass the null, the strongest $2m_0$, or all of them; per seed, the candidate with the smallest exploitability lower bound on the *training* layouts is chosen (rows of seeds 9–11). The deviators chose a different depth on each seed, and each choice was also the least exploitable candidate on held-out layouts (2.9, 4.5, 6.3 against raw's 19.0, 8.2, 11.8). The half-depth profile chosen on seed 9 has predators that score 17 against a uniform-random prey where raw's score 42: one deviating predator cannot repair a weak team, so a unilateral lower bound does not see it. Choosing the depth costs three further trainings and four attacks per seed, which the raw baseline does not get.

Multilateral exploitability. §2 gives two readings of a team; every bound above uses the second: four players, unilateral deviations. After the unilateral comparison with DBC had failed, and before measuring anything else, we defined $\mathrm{EXP}_k = \max(0, G_k) + \max(0, g_{\mathrm{prey}})$, where G_k is the gain of k predators deviating *jointly* with one shared residual policy, everyone else frozen, under the same oracle and the same 10^6 steps per attack; G_1 is the largest unilateral predator gain, and EXP_3 is the exploitability of the two-player game between the team and the prey. On the 45 profiles of seeds 6–20 (held-out layouts; means):

Table 21: Exploitability lower bounds when coalitions of k predators may deviate together. EXP_1 differs from the bound reported elsewhere only in counting the largest predator gain once instead of summing the positive gains of all four players (9.99 for DBC).

profile	G_1	G_2	G_3	EXP_1	EXP_2	EXP_3
raw MAPPO	-0.7	-0.7	$+0.2$	13.8	13.9	14.4
permutation filter	-0.6	-1.2	-0.6	9.6	9.6	9.8
D1-only encoder (DBC)	$+1.9$	$+14.1$	$+37.5$	8.2	20.5	43.9

The two tests fixed in advance, on EXP_3 over the fifteen seeds at $\alpha = 0.025$ each: the filter is below DBC on 15 of 15 seeds ($t_{14} = -13.2$), and below raw on 12 of 15 without significance ($t_{14} = -2.2$, $p = 0.04$). Against DBC a team of deviators gains at least 25 on every seed; against raw and the filter it gains more than 1 on only 5 of 30 profiles, which we had not predicted, although swapping in the filter's whole team gains $+4.0$ against raw's prey on 13 seeds of 15, so a learned deviator that must itself see the 64 distractors is a weak attacker of raw (with that swap counted, raw's EXP_3 is 17.9). The measure was motivated by what the unilateral bounds had shown and was registered before it was run. Under it a profile is exploitable simply by being a weak team; that is its point, and the reason it does not replace the unilateral bound.

A two-timescale solver. For raw and the filter the whole bound is the prey's deviation gain: simultaneous self-play leaves the prey far from its best response to the predators it ends with. After the filter's burn-in, unchanged, every joint PPO update (256×64 steps) is followed by M updates of the prey alone against the frozen predators on small rollouts (64×32 steps), its optimizer state carried across, until exactly the baseline's 3,112,960 environment steps are spent; with $M = 2$ the predators take 157 joint updates where raw takes 190. The single adversary tracks its best response while the team moves slowly. *Development* (spent seeds 0–2, $M \in \{2, 4, 8\}$ declared beforehand): the held-out bound fell from 7.8 (filter, self-play) to 1.7, 1.4, 1.6 (raw: 17.7), and nothing we had (a learned team of deviators, raw's whole team, raw's prey) gained against the $M = 2$ profile; but the predators were behind raw's head-to-head on three seeds of three for every M (by 0.5–4.3 at $M = 2$), which failed the competence gate we had registered, so no fresh seed was spent. *Confirmation:* with those numbers known, a new registration was opened for $M = 2$ on seeds 21–29, its two tests unchanged from the first and competence reported instead of gated (rows above). Both tests hold at $\alpha = 0.025$: the bound is below DBC's on 8 seeds of 9 ($t_8 = -3.8$, $p = 0.005$) and below raw's on 8 of 9 ($t_8 = -4.8$, $p = 0.001$); the exception is seed 23 (14.3), reported as run. With team deviations it measures 3.2 against raw's 16.7 ($t_8 = -5.2$) and DBC's 45.1. Registered as description with a prediction: the same schedule on raw observations leaves raw's bound where it was (15.7 against 15.8) and costs its predators 7.9 head-to-head, while behind the filter the bound is 12.8 lower on nine seeds of nine ($t_8 = -7.4$): the solver works only behind the abstraction. Against self-play within the filter the gain is not significant at nine seeds (-4.1, $p = 0.12$). At the oracle's smaller attack budgets (2×10^5 and 5×10^5 steps per player, against 10^6 in the rows above) the bounds are raw 4.8 / 13.1, filter 3.0 / 4.7, filter with the schedule 1.1 / 1.9, raw with the schedule 2.7 / 12.3, DBC 3.0 / 6.1: the solver arm is lowest at every budget, and the order of the five arms is the same at the two larger ones. Two of seven predictions missed narrowly (predators against a uniform prey 3.1 below the ordinary filter's, not within 3; team-deviation bound 3.2, not 3 or less). No arm meets both of the conditions we had set ourselves: ordinary self-play gives the head-to-head, the two-timescale schedule gives the exploitability, and M moves between them. The schedule is a heuristic after Daskalakis et al. (2020): under independent policy gradient in two-player zero-sum stochastic games, the player with much smaller step sizes has approximately min-max iterates on average; their theorem covers neither team games nor PPO, and here the gain we measure is the fast player's.

Multiplicity. This appendix reports eleven pre-registered paired t tests in all: two on the first six seeds, three on the full encoder, two at the fifteen-seed last look, two on team deviations and two for the solver, each registered at 0.05 or, where a registration had two, at 0.025. A Holm correction over all eleven at a family level of 0.05 keeps exactly the three that were significant at their registered levels (the filter's team-deviation bound against DBC's, and the solver's bound against raw's and against DBC's) and none of the others; at a family level of 0.025 the last of these ($p = 0.005$) no longer survives.

Related work: teams, and solvers for the latent game. Games in which agents cooperate within a group and compete between groups are the setting of adversarial team Markov games (Kalogiannis et al., 2023), and self-play for them is an active area: naive simultaneous self-play need not reach an equilibrium, which population methods address for such games (Xu et al., 2023). For two players, regularization turns cycling into convergence: to a perturbed equilibrium in imperfect-information games (Perolat et al., 2021), to the quantal response equilibrium of entropy-regularized Markov games at a linear rate (Cen et al., 2021), and in normal- and extensive-form games (Sokota et al., 2023); none of these treats teams. We contribute nothing to that problem. Our theorems take the equilibrium of the abstract game as given and say what it is worth in the ground game; any of these solvers could play the role of the latent solver in this appendix, and Appendix K reports one substitution of our own, the two-timescale schedule.

Long-budget runs (1.9×10^7 solver steps, 12 outer steps, 3 seeds, $\approx$ 3 h per seed).

encoder	exploitability lower bound	held-out predators	training predators	latent codes used
reconstruction (VAE)	2.0 ± 0.5	9.6 ± 0.4	8.7 ± 0.2	22.2
D1 only (DBC-like)	3.3 ± 0.7	$\mathbf{11.2 \pm 0.5}$	9.7 ± 0.3	20.3
D1+D2+IB	5.4 ± 0.7	9.3 ± 0.6	8.3 ± 0.7	15.0
raw MAPPO (same environment)	–	**30.5**	27.6	–

Eight times the solver budget of the matched-budget table moves the latent predators from ≈ 6.5 to 9.3–11.2 and leaves them at a third of raw MAPPO's return; the D2 signal is again the most exploitable arm, on every seed. The exploitability column here is measured on training layouts.

Every latent profile except the hard-gated one has a lower measured exploitability than raw MAPPO, and every one is a far weaker player; low exploitability of a weak profile is cheap, so wedo not read the first column as a win. D2-containing encoders use a third fewer effective dimensions at comparable exploitability, and are *more* exploitable than the D1-only encoder on all three seeds under two independent attacks on training layouts. Attacked on held-out layouts every profile is more exploitable than on training layouts; there the D1+D2+IB encoder is more exploitable than the D1-only encoder in the mean but not on every seed (lower by 1.5 and 0.2 on two seeds, higher by 8.4 on the third, where it is also above raw MAPPO). The hard-gated encoder closes 61–63 of 64 distractor gates and doubles the latent predators' return, but also closes task coordinates the prey needs and fails our pre-registered criterion for matching raw MAPPO on every seed. Cross-play between independently trained runs shows no partner-specific conventions in this game (return drops of 0.1–0.6), so zero-shot coordination cannot discriminate either. IPPO is weaker than MAPPO in both settings (45.8 vs. 49.4 at $m_d = 8$; 36.2 vs. 40.8 at $m_d = 64$, against a uniform opponent).

L THREE PLAYERS, GENERAL-SUM: THE CONSTRUCTION AND THE RECORD

The theorems are stated for N-player general-sum games; the experiments of §5 are two-player zero-sum because exact equilibrium solving there is one linear program per state. The transfer inequalities, however, hold for *any* abstract profile $\hat{\pi}$ with $\hat{\varepsilon}$ its measured abstract gap, so exact policy evaluation and exact best responses suffice to check them. We build random three-player general-sum games (20 core states $\times$ a binary distractor $=$ 40 states, 27 joint actions, potential-structured rewards so that exact best-response iteration has pure equilibria to find) with the same distractor construction: TAUNT is penalized in reward and is not played jointly at the equilibria found, d changes transitions only when all three players taunt, and costs player 0 δ on one action. Across 8 games, $\delta \in \{0, 0.02, 0.05, 0.1\}$ and 768 (abstraction, profile) pairs (640 of them random, non-equilibrium profiles), Theorem 1 (per state and in exploitability), Theorem 2 and Proposition 3 hold in every case. The hierarchy reproduces: for the hand design $\varepsilon_P^0 = 1$ in every game, so D0 forbids it and D0 refinement returns the identity (40 states), while $\varepsilon_P^1 = 0$ against the equilibrium, and D1 refinement from the one-class partition recovers exactly the 20-state hand design in 21 of 24 games with $\delta > 0$ (mean exact exploitability 0.012 over those 21) and returns the identity in the other three (one base game at its three values of δ), although the hand design is D1-admissible there too: the fixed point reached depends on the path. D2 does *not* reproduce its two-player advantage here: at $\varepsilon = 0.2$ it returns 19–40 states, more than D1 in 17 of 32 games and with higher exploitability in 14. The solver reaches an exact pure Markov equilibrium ($\hat{\varepsilon} < 10^{-9}$) in 100 of 128 abstract games; the inequalities are verified on the remaining 28 too, with their measured $\hat{\varepsilon}$. As in the two-player

case the bound of Theorem 1 is loose (median $296\times$ the exact value over the hand-design equilibria with EXP > 0, $718\times$ over all 67 solver profiles with EXP > 0) and serves as an admissibility condition.

M NOTATION

Collected for reference; nothing here is new. Section numbers point to where each object is defined.

symbol	where	meaning
$G = (N, \mathcal{S}, \{\mathcal{A}_i\}, P, \{r_i\}, \gamma, \mu_0)$	§2	N-player discounted Markov game, finite $\mathcal{S}$
R, $V_{\max} = R/(1-\gamma)$	§2	reward span; span of attainable values ($R \approx 2$ in our soccer games)
π_i, π_{-i}, π	§2	player i's policy, the others', the profile
$M_i(\pi_{-i})$, $r_i^{\pi_{-i}}$, $P^{\pi_{-i}}$	§2	the MDP player i faces against π_{-i}, and its reward and kernel
$Q_i^\star(s, a_i \mid \pi_{-i})$	§2	optimal action value in $M_i(\pi_{-i})$, the *best-response value* D2 compares
$Q^\star(s, \boldsymbol{a})$, $V^\star(s)$	§5.2	minimax action value and value of the ground game (two-player zero-sum)
$\mathrm{EXP}_G(\pi)$	§2	exploitability, $\sum_i (V_i^\star(\mu_0 \mid \pi_{-i}) - V_i^\pi(\mu_0))$
$\mathrm{GAP}(\pi)$	App. A	$\max_s \sum_i (V_i^\star(s \mid \pi_{-i}) - V_i^\pi(s))$, the duality gap over all states
$\phi : \mathcal{S} \to \mathcal{Z}$, $\mathcal{Z}$, $w(\cdot \mid z)$	§2	abstraction, its classes, the within-class weighting (uniform throughout)
$s \sim_\phi s'$	Def. 1	$\phi(s) = \phi(s')$: s and s' share a class
$\hat{G}_\phi$, $\hat{\pi}$, $\pi = \hat{\pi} \circ \phi$	§2	abstract game, an abstract profile, its *lift* to the ground game
$\hat{\varepsilon}$	Thm. 1	the abstract profile's own equilibrium gap in $\hat{G}_\phi$
ε_r^0, ε_P^0	Def. 1	D0 distortions: reward and pushforward-TV, over *all* joint actions
$\varepsilon_r^1(\pi)$, $\varepsilon_P^1(\pi)$	Def. 1	D1 distortions: the same under unilateral deviations from π, own actions only
$\varepsilon_Q(\pi)$	Def. 1	D2 distortion: within-class spread of $Q_i^\star(\cdot, a_i \mid \pi_{-i})$
ε_Q^m, $U_i^m(s)$, m	Prop. 10	D2u: distortion over *near-best* own actions, that set, and the margin
$\phi^{Q^\star}$	§5.2	the prior value criterion: $Q^\star$ compared entrywise over joint actions
γ, μ_0	§2	discount, initial distribution
N (in G_N, H_N)	§3	number of values of the constructed nuisance variable d
Δ, β, B_d	Prop. 7	side-bet scale, pattern separation $\min_{d \neq d'} \|B_d - B_{d'}\|_\infty$, the patterns
τ	§5.3	the exploitability level a sampled learner must reach
$n^\star(\tau)$, ρ	§5.3	samples needed to reach τ; the ratio between two learners' requirements

N THE COARSEST PARTITION AT A FIXED PROFILE: CRITERION OR SEARCH?

§5.2 leaves a question the frontier alone cannot answer: when a dominance-aware criterion compresses further than ours, is that the criterion or our greedy search? At a *fixed* profile it can be answered exactly, because the criterion's admissible partitions have an exact combinatorial description.

The equivalence. Fix a profile and let $d(s, s') = \max_k |X_{sk} - X_{s'k}|$ with $X = (Q_A^{\mathrm{BR}}, Q_B^{\mathrm{BR}})$ the best-response values at that profile, so d does not move with the partition. A partition is admissible at tolerance ε exactly when every within-class pair is within ε, i.e. when every class is a clique of the graph $G(\varepsilon)$ on states with an edge where $d(s, s') \leq \varepsilon$. Definition 1 also asks the profile to be ϕ-measurable, so the admissible partitions are the clique covers of $G(\varepsilon)$ that keep it so; a minimum clique cover ignores that second condition and therefore *lower-bounds* the coarsest admissible partition, which is the direction every comparison below needs. In the sup-norm a set is pairwise within

ε if and only if it lies in one axis-parallel cube of side ε (take the coordinatewise extremes), so a minimum clique cover is also the fewest side-ε cubes covering the points X_s. Any set of pairwise-incompatible states needs one class each, so an independent set of $G(\varepsilon)$ bounds every cover from below, with no search and no oracle.

What is exact, and what is not. On the ten games of Table 15 at the exact ground equilibrium, D2's distance and the same 21 tolerances, the lower bound meets our one-pass cover on 101 of the 210 instances, so the exact minimum is known there and our search attains it. Four further searches (saturation-ordered coloring, a width-16 beam over the same insertion decisions, and a cross-entropy policy over insertion orders at three seeds) reach the bound on 40 of the remaining 109, giving the exact minimum on 141 of 210; on the rest the cover is within 9.3% of a bound that holds for every cover. No cover below the bound was ever returned (909 method-instances checked). Every partition is scored as in §5: aggregated, re-solved from scratch, lifted, and given its exact duality gap.

Criterion, not search. At the tolerance where the one-pass criterion is lossless, the minimum cover is 566 states on the plain game and $1{,}143$ on the distractor game; both are instances where the bound is met, so these are minima *for that profile and tolerance* and not heuristic outputs, and all nine search variants return exactly them. They are not minima over all equilibria of the ground game, nor over all lossless partitions: a different equilibrium or a different tolerance may admit a coarser one. The IESDS-restricted $\phi^{Q^\star}$ is lossless at 328 and 772 (§5.2). The 42% and 32% gaps are therefore not closed by any search *under this distance at this profile*, which is what makes them properties of the criterion rather than of our greedy.

Registered outcomes and limits. Of the four predictions registered before the covers were computed, the two about the comparison held (the gap between our cover and the bound stays under 10%, and the optimal cover stays above the dominance-aware baseline) and two did not: the bound is met at the lossless tolerance on five of the ten games rather than the eight predicted, and $G(\varepsilon)$ *is* chordal there on three games (and on 61 of the 210 instances, all 61 of which are instances where the bound is met, as perfection requires). Of the five registered for the searches, three held and two did not: the best search improves on ours by 2.2% on the open instances rather than the predicted $<2\%$, and the width-16 beam is worse than our greedy on four instances by one class, its ranking key being a heuristic and not an admissible bound. Two limits bound what the above establishes. The exact-minimum solver closed no instance that the lower bound had not already closed, so the 69 instances without a minimum are open rather than hard-won. And the equivalence holds for criteria evaluated at a fixed profile; the loop of §4 recomputes its distance from the partition it is building, so its admissible partitions are not the clique covers of any one graph, and the coarsest partition for the loop remains open. That distinction is not merely formal. Substituting the saturation-ordered search (the one that is never worse at a fixed profile) for our static order *inside* the loop leaves its smallest lossless partition smaller on four of the ten games and larger on six, by 8.0% on one, because a coarser partition at one round changes the signatures the next round splits on. Per-round optimality is not fixed-point optimality, and we do not have the second.