STABILITY AND DECISION FIDELITY IN TREE-STRUCTURED BELLMAN UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Tree-structured sequential protocols aggregate child continuations before local Bellman updates. Normalizing this channel can reduce its sensitivity while changing the decision problem. Tree-SeBIL distinguishes stability of a declared semantic operator from preservation of its decisions, separating fixed-point-preserving averaging from target-changing surrogates. On a closed invariant domain, explicit reward, transition, and policy-recovery bounds yield a structured small-gain certificate; a restricted leaf-injection family gives a sharp unweighted norm threshold. On an action-rich class, universal local Gibbs-policy fidelity at fixed continuation inputs and unchanged rewards holds exactly for aggregation translations. Reducing the aggregation Lipschitz budget below the semantic aggregation's minimal modulus yields a worst-case continuation mismatch and binary policy-separation bound, sharp on the stated linear family. An independent 900-map interception study illustrates low surrogate residual with substantial original-objective decision change. Neural results provide limited, mixed wrapper evidence, not equilibrium certificates. Fixed points solve the declared induced-node problems; transfer to a game's regularized equilibrium additionally requires policy consistency and compatible reward ownership, transitions, and histories.

Keywords: tree-structured games, soft Bellman operators, small-gain stability, policy fidelity, semantic aggregation

1 INTRODUCTION

Sequential MARL imposes a decision protocol to manage non-stationarity (17; 25; 28). A tree protocol adds a distinct issue: each parent combines several child continuations. Replacing a sum or another semantic rule by a normalized average changes not only the update scale but potentially the objective that an agent optimizes.

We study when soft backward induction is contractive for a declared continuation rule and when reducing continuation gain sacrifices local decision fidelity. Tree-SeBIL extends the chain SeBIL sweep to a rooted tree: children are solved before parents, with ancestor policies held from the input profile. In a restricted leaf-injection template, the child sensitivity $\gamma L_{\mathcal{A}}/(1-\gamma)$ compounds with depth. General trees also have direct ancestor feedback and require a matrix certificate. Neither a large unweighted norm nor failure of a sufficient certificate proves divergence.

Our contributions are threefold. First, we specify the local Bellman and policy-recovery interfaces and derive the tree path-sum certificate $M = (I - B)^{-1}C$, including its weighted form and an exact unweighted leaf-injection example. Second, we characterize universal local fidelity of reward-preserving aggregation replacements and quantify the price of lowering their Lipschitz budget. Third, we separate target-relative numerical checks, independent semantic decision diagnostics, and approximate neural wrapper results. These evidence types answer different questions.

The central restriction is local but useful: a nonconstant aggregation gap admits a two-action decision reversal even when a surrogate solves its own objective accurately. This is not a generic equilibrium algorithm for general-sum games. The declared node problem must be shown to equal the moving player's true deviation problem before a game-level interpretation is available.

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.

2 Related Work

Our primary object is the continuation operator and its audit protocol, with downstream learners used only as candidate generators. Concurrent CTDE MARL (5; 15; 17) suffers from non-stationarity; Tree-SeBIL restructures updates into layerwise backward sweeps (Gauss–Seidel) on a fixed protocol. Equilibrium-oriented RL (14; 16; 27) targets gaps or regret under different information models. CFR-style methods (3; 30) operate on extensive-form trees; we use Markov value recurrence with explicit $L_{\mathcal{A}}$–depth sensitivity.

Tree and hierarchical games. A close tree-structured game-theoretic antecedent is structured hierarchical games with Differential Backward Induction (11). There, players occupy a one-shot hierarchy, moves proceed root-to-leaf, utilities depend on local and leaf decisions, and the solver is gradient/backpropagation-style. Tree-SeBIL instead studies discounted Markov or proxy protocols whose child subtrees enter Bellman continuations through an aggregation operator $\mathcal{A}$. The main object is the semantic continuation operator, its small-gain stability boundary, and the invariance/bias of normalized surrogate continuations; we do not claim priority over tree-structured games or classical backward induction.

Sequential and factorized MARL. HARL (28) motivates SeBIL+X tactical solvers on shallow trees (teams sequential, agents simultaneous), and maximum-entropy HARL/HASAC (13) studies stochastic cooperative MARL with entropy-regularized objectives. Multi-Agent Transformer casts cooperative MARL as autoregressive sequence modeling using advantage decomposition (26); more recent sequential-update methods such as HAQO (12) and K-Level Policy Gradients (21) also address non-stationarity through ordered or recursive policy updates. These works are closest to the wrapper and credit-assignment side of SeBIL+X. Our theoretical claims are different in scope: they concern tree-valued soft Bellman operators under a fixed protocol, including a universal local fidelity restriction for aggregation-only replacements. Value-decomposition and factored-critic methods (18; 20; 24) aggregate cooperative value information; our analysis studies induced-node optimality and the sensitivity of declared continuation propagation. An ordered protocol alone does not identify those node objectives with a general-sum game's utilities. The contraction perturbation lemma, positive-matrix criterion, softmax shift invariance, and ordered advantage telescoping are standard ingredients (2; 7; 9). Our contribution is their joint restriction on tree continuation design, not a new Bellman contraction theorem or a new equilibrium concept. Small-gain certificates have also been used for differentiable-game convergence (23); our small-gain matrix acts on Bellman continuation errors across a rooted protocol tree rather than on pseudo-gradient dynamics in strategy space.

3 Sequential structure: trees and the chain degeneracy

3.1 Rooted tree-structured sequential games

Let $T = (V, E)$ be a rooted directed tree with root r. For $u, v \in V$, write $u \preceq v$ if a directed path from u to v exists, and $u \prec v$ if additionally $u \neq v$. Define $\mathrm{Anc}(u) = \{v : v \prec u\}$, $\mathrm{Ch}(u) = \{v : (u, v) \in E\}$, $d(u)$ as path length from r to u, leaves $\mathcal{L} = \{u : \mathrm{Ch}(u) = \emptyset\}$, and layers $\mathcal{L}_d = \{u : d(u) = d\}$, $D_{\max} = \max_u d(u)$.

The **history** at node u is

$$\mathcal{H}_u = \mathcal{S} \times \prod_{v \in \mathrm{Anc}(u)} \mathcal{A}_v. \tag{1}$$

Node u chooses $a_u \in \mathcal{A}_u$ after observing $h_u = (s, a_{\mathrm{Anc}(u)})$; the Markov transition from s occurs after a full joint action profile along a stage is committed (see Assumption B.3 for the chain-style intra-stage semantics used in benchmarks). Incomparable nodes do not observe or condition on sibling actions; once their ancestor actions are fixed, nodes in the same layer may act in parallel.

Declared continuation semantics. Each node has a finite action set, a bounded local reward, and an induced kernel on its history space. A supplied aggregation $\mathcal{A}_u$ maps child messages to a scalar and satisfies $|\mathcal{A}_u(x) - \mathcal{A}_u(y)| \leq L_u \|x - y\|_\infty$; $L_{\mathcal{A}} = \max_u L_u$. These are model inputs. In

particular, adding a child's value to a parent's reward does not generally represent the parent's utility in a general-sum game.

Definition 3.1 (Semantic, stabilized, and surrogate operators)**.** The semantic map uses the continuation rule declared by the model. Global KM averaging $(1-\eta)I + \eta\mathcal{T}^s$, $0 < \eta \leq 1$, preserves its fixed points. A surrogate changes that rule, for example to scaled average, and has its own target. Local policy fidelity and game-level equilibrium transfer are separate requirements.

3.2 Chain as a path tree

A **chain** with N agents (or nodes) is a tree in which each non-leaf node has exactly one child; equivalently $|\operatorname{Ch}(u)| \leq 1$. Ordering nodes $u_1, \ldots, u_N$ along the path recovers the original SeBIL protocol $\sigma = (1, \ldots, N)$ with histories $\mathcal{H}_i = \mathcal{S} \times \prod_{j<i} \mathcal{A}_j$. In this case $\mathcal{A}$ is the identity on a scalar ($L_{\mathcal{A}} = 1$), and the Tree-SeBIL layer sweep coincides with the SeBIL backward sweep (Corollary 4.15 and Theorem C.4).

Protocol scope. A simulator with simultaneous actions can be equipped with a sequential proxy protocol. Its information structure need not match the original game. The ordered telescoping identity and chain notation are given in Appendix B; they do not enlarge sibling information sets or prove equilibrium transfer.

4 Tree-SeBIL: operators and conditional statements

We first specify the oracle whose sensitivities are certified. Proofs and auxiliary chain results are in Appendices E and H. Superscripts s and u distinguish semantic and surrogate operators, respectively.

Definition 4.1 (Local oracle and recovery interface)**.** Let p_u collect recovered ancestor policies, $z = (z_v)_{v \in \operatorname{Ch}(u)}$ be an independently held child tuple, and W a local state value. A declared joint kernel $K_u^{p_u}(dh', dx \mid h, a)$ samples the next local history and a tuple of child histories $x = (x_v)_v$. Set $G_u(z, x) = \mathcal{A}_u((z_v(x_v))_v)$, with $G_u = 0$ at leaves. The local equations are

$$Q_u[W; p_u, z](h, a) = r_u^{p_u}(h, a) + \gamma \int [W(h') + G_u(z, x)]\, dK_u^{p_u}, \tag{2}$$

$$(\mathcal{B}_u^{p_u, z} W)(h) = \alpha \log \sum_{a \in \mathcal{A}_u} \exp(Q_u[W; p_u, z](h, a)/\alpha). \tag{3}$$

Let $F_u(p_u, z)$ be the unique fixed point of $\mathcal{B}_u^{p_u,z}$, and set $\Psi_u(p_u, z) = \operatorname{softmax}_\alpha Q_u[F_u(p_u, z); p_u, z]$. A supplied bounded action-value reconstruction $\mathcal{Q}_u(V)$ defines the input policy $\mathcal{R}_u(V) = \operatorname{softmax}_\alpha \mathcal{Q}_u(V)$; scalar state values alone do not define this map. The sweep sets $V'_u = F_u(\mathcal{R}_{\operatorname{Anc}(u)}(V), V'_{\operatorname{Ch}(u)})$ from leaves to root.

The augmented reward is $r_u + \gamma G_u$; W is the subsequent value for this *induced* problem. This equation is not an assertion that arbitrary inter-player utilities can be added or that one continuation should be counted twice. Numerically, log-sum-exp is $m + \alpha \log \sum_a \exp((Q_a - m)/\alpha)$ with $m = \max_a Q_a$.

Complete invariant domain. Use finite histories, or compact histories with continuous rewards/aggregates and Feller lifted kernels, finite actions, and continuous recovery maps. Fix $0 < \gamma < 1$ and $\alpha \in (0, \bar{\alpha}]$. If $|r_u^p| \leq R_u$ uniformly, recursively set

$$\bar{G}_u = |\mathcal{A}_u(0)| + L_u \max_{v \in \operatorname{Ch}(u)} H_v, \qquad H_u = \frac{R_u + \gamma \bar{G}_u + \bar{\alpha} \log |\mathcal{A}_u|}{1 - \gamma},$$

with $\bar{G}_u = 0$ at leaves. Then $\mathcal{D} = \prod_u \{W \in C(\mathcal{H}_u) : \|W\|_\infty \leq H_u\}$ is closed, convex, and complete. Each local map and the sweep preserve these bounds, including the entropy term. All sensitivity bounds below are uniform on this domain and on independently held child tuples.

Assumption 4.2 (Held-child ancestor sensitivity)**.** For every admissible z, the partial map satisfies

$$\|F_u(\mathcal{R}_{\operatorname{Anc}(u)}(V^a), z) - F_u(\mathcal{R}_{\operatorname{Anc}(u)}(V^b), z)\|_\infty \leq \sum_t c_{ut} \|V_t^a - V_t^b\|_\infty.$$

For ancestor-local recovery, only ancestor columns are nonzero. More general recovery dependencies must be included in C.

Proposition 4.3 (Primitive sensitivity bound)**.** *Suppose changing a parameter p changes r_u, K_u, and G_u with constants $L_{r,u}$, $L_{K,u}$, and $L_{G,p,u}$, where kernel distance is the dual norm $\sup_{\|f\|_\infty \le 1} |\int f\, d(K^a - K^b)|$. Uniformly in held children, the fixed-point sensitivity is at most*

$$A_u = \frac{L_{r,u} + \gamma L_{K,u}(H_u + \bar{G}_u) + \gamma L_{G,p,u}}{1-\gamma}.$$

If $\|p_u(V^a) - p_u(V^b)\| \le \sum_t \ell_{ut} \|V_t^a - V_t^b\|_\infty$, take $c_{ut} = A_u \ell_{ut}$. A fixed kernel and reward give child sensitivity $\gamma L_u/(1-\gamma)$; child-dependent kernels/rewards require additional terms (Appendix E.2).

The proof of Proposition 4.3 is in Appendix E.2.

Theorem 4.4 (Leaf-injection scalar boundary)**.** *Consider the leaf-injection template in which input-profile perturbations enter leaves with gain at most γ and internal nodes receive perturbations only through post-sweep child values. If $|\mathcal{A}(x) - \mathcal{A}(y)| \le L_{\mathcal{A}} \|x - y\|_\infty$, then $L_{\mathcal{A}} \le (1-\gamma)/\gamma$ makes the template a γ-contraction. For depth D, the relaxed sufficient scalar condition is $L_{\mathcal{A}} < \frac{1-\gamma}{\gamma}(1/\gamma)^{1/D}$, and the boundary is tight on the exact linear depth-D chain family. General ancestor-sensitive trees instead require the structured certificate in Theorem 4.7 (or its weighted form).*

The proof of Theorem 4.4 is in Appendix F.1.

4.1 Structured small-gain view

Node-wise errors satisfy the nonnegative recursion

$$e_u \le \sum_{z \in V} c_{uz} d_z + \sum_{v \in \mathrm{Ch}(u)} b_{uv} e_v, \qquad u \in V, \tag{4}$$

where $d_u = \|V_u^a - V_u^b\|_\infty$ and $e_u = \|(\mathcal{T}_T^s V^a)_u - (\mathcal{T}_T^s V^b)_u\|_\infty$. Stacking gives $e \le Be + Cd$ and $M = (I-B)^{-1}C$. If $Cd \le a\|d\|_\infty$, define $\rho_T^s \triangleq \|(I-B)^{-1}a\|_\infty$.

Definition 4.5 (Small-gain certificate)**.** A certificate is $\mathcal{C}(T, \mathcal{A}, \gamma, \alpha) = (B, C, M, \rho_c)$ with $M = (I-B)^{-1}C$ and $\rho_c = \|M\|_\infty$ for any nonnegative B, C making (4) valid. Since T is acyclic, $(I-B)^{-1}$ is a finite path-sum matrix.

Theorem 4.6 (Weighted small-gain and spectral criterion)**.** *If $Mw \le \lambda w$ for some $w > 0$ and $\lambda < 1$, then $\mathcal{T}_T^s$ is a λ-contraction in the weighted sup norm. Here $\|V\|_{\infty,w} = \max_u \|V_u\|_\infty / w_u$. Such a positive-weight certificate exists iff $\rho(M) < 1$.*

The proof of Theorem 4.6 is in Appendix E.4.

Theorem 4.7 (Structured small-gain contraction)**.** *If $\rho_T^s < 1$, then $\mathcal{T}_T^s$ is a ρ_T^s-contraction in $\|\cdot\|_\infty$.*

The scalar corollaries and exact-gain assumptions are stated in Corollary C.7 and Assumption F.1. The scaled average $(1-\gamma)\sum_k x_k/(\gamma K)$ gives modulus γ in the leaf-injection template, but does not remove general ancestor feedback. A failed unweighted certificate is not a divergence test: the depth-2 example in Appendix E.7 has $\|M\|_\infty = 3.6$ but $\rho(M) = 0.9$.

The proof of Theorem 4.7 is in Appendix E.5.

Proposition 4.8 (Stabilization and surrogate boundaries)**.** *For a $\rho < 1$ contraction, global KM averaging preserves fixed points and has upper bound $q_\eta = 1 - \eta(1-\rho)$, which is larger than ρ for $0 < \eta < 1$. Damping can reduce oscillatory modes, but does not improve this worst-case bound. For an exact nonnegative linear map it cannot cross the spectral boundary below. Per-layer damping is a different update (Appendix F.4).*

The proof of Proposition 4.8 is in Appendix F.5.

Proposition 4.9 (Spectral KM boundary)**.** *For $0 < \eta \le 1$ and nonnegative gain matrix M, $M_\eta = (1-\eta)I + \eta M$ has $\rho(M_\eta) = 1 - \eta + \eta\rho(M)$; hence $\rho(M_\eta) < 1$ iff $\rho(M) < 1$.*

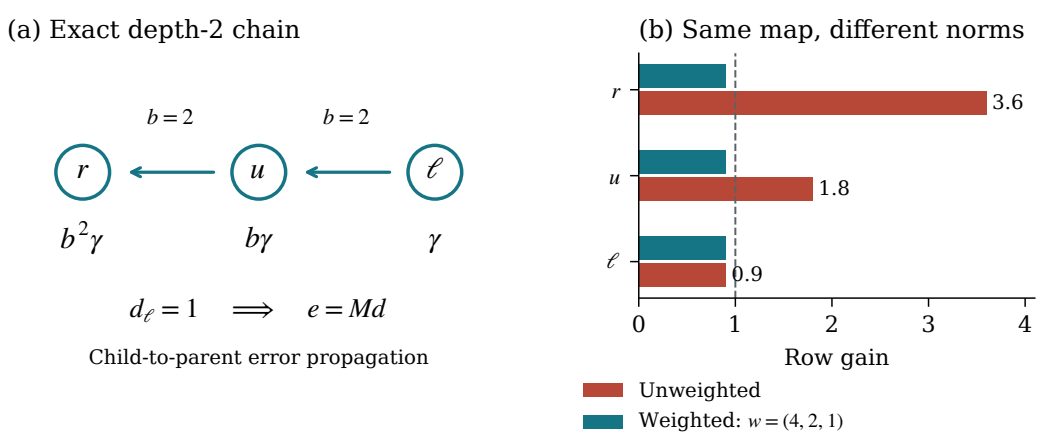


Figure 1: A worked depth-2 certificate from Appendix E.7. With $\gamma = 0.9$ and child gain $b = 2$, the exact path-sum map has unweighted row gains up to 3.6, while the positive weights $w = (4, 2, 1)$ give $Mw = 0.9w$. Thus the unweighted certificate fails even though the weighted spectral criterion certifies contraction. This is an algebraic diagnostic, not a claim about arbitrary cyclic trees.

The proof of Proposition 4.9 is in Appendix F.6.

Proposition 4.10 (Surrogate-to-semantic fixed-point bias)**.** *If the semantic operator is a $\rho_s < 1$ contraction and the surrogate has uniform mismatch ϵ_u, then $\|V_u^* - V_s^*\|_\infty \le \epsilon_u/(1-\rho_s)$, provided both fixed points exist in the common domain. For recovered policies, add the direct difference between recovery maps to their Lipschitz value-error term. Common local shifts preserve a fixed-input Gibbs policy, but do not alone prove global recursive policy equivalence.*

The proof of Proposition 4.10 is in Appendix F.7.

Definition 4.11 (Universal local semantic fidelity)**.** Let $A, S : \mathcal{X} \to \mathbb{R}$ be semantic and surrogate aggregations. For actions a, rewards r_a, and realizable action-conditioned inputs $x_a \in \mathcal{X}$, set $Q_A(a) = r_a + \gamma A(x_a)$ and $Q_S(a) = r_a + \gamma S(x_a)$. The reward-preserving replacement S is *universally faithful* at temperature $\alpha > 0$ if $\mathrm{softmax}_\alpha(Q_S) = \mathrm{softmax}_\alpha(Q_A)$ for every finite reward vector and every realizable $\{x_a\}$.

Theorem 4.12 (Universal fidelity characterization and decision flip)**.** *On an action-rich class containing every two-action deterministic continuation pair, S is universally faithful iff $S(x) = A(x) + c$ for a constant c. Allowing a fixed rescaled temperature $\alpha_S = \kappa\alpha$ while leaving rewards unchanged still forces $\kappa = 1$ and the same translation. If $S - A$ is nonconstant, a two-action instance exists whose unique top action is reversed by S.*

The proof of Theorem 4.12 is in Appendix F.8.

Theorem 4.13 (Sharp quantitative gain–fidelity frontier)**.** *Let $(\mathcal{X}, d)$ be a metric continuation space, let $A, S : \mathcal{X} \to \mathbb{R}$, and suppose S is L_S-Lipschitz. Writing $g = S - A$, every pair x, y satisfies*

$$|g(x) - g(y)| \ge \big(|A(x) - A(y)| - L_S d(x, y)\big)_+.$$

Hence $\mathrm{osc}(g) \ge M(A, L_S)$, the supremum of the right-hand side. If every pair is realizable, the supremum of total-variation distances over reward-centered binary instances is at least $\tanh(\gamma M(A, L_S)/(4\alpha))$. Attainment is not implied on an arbitrary metric space. Both bounds are attained on a closed bounded interval by $A(x) = L_A x$, $S(x) = L_S x$, $0 \le L_S \le L_A$.

The proof of Theorem 4.13 is in Appendix F.8.

Corollary 4.14 (Stability–fidelity impossibility)**.** *Let $L_A^\star = \mathrm{Lip}_\infty(A)$ be the minimal aggregation modulus and $\tau_\gamma = (1-\gamma)/\gamma$. If $L_A^\star > \tau_\gamma$, no reward-preserving aggregation replacement can be both universally faithful and satisfy the depth-uniform leaf-injection certificate $\mathrm{Lip}_\infty(S) \le \tau_\gamma$. For the exact linear-gain chain family, every universally faithful replacement is non-contractive in the ordinary sup norm at some finite depth. Positive affine Q transforms with matched temperature remain valid fixed-game exceptions when the full local Q, including rewards, is transformed.*

Algorithm 1 Tree-SeBIL: gated sweep and fidelity audit

Require: Declared local models, recovery map $\mathcal{R}$, $\mathcal{A}_s$, optional $\mathcal{A}_u$, fixed $\alpha > 0$, $0 < \eta \leq 1$

1: Construct valid B, C on $\mathcal{D}$; check an unweighted or weighted certificate
2: Use the semantic rule if certified; otherwise declare an uncertified run or separately certify a supplied surrogate and report its fidelity status
3: Initialize $V \in \mathcal{D}$
4: **for** $k = 0, 1, \ldots$ **do**
5: Freeze $p_u = \mathcal{R}_{\mathrm{Anc}(u)}(V)$ for every node
6: **for** layers $d = D_{\max}, \ldots, 0$ **do**
7: For $u \in \mathcal{L}_d$ in parallel, solve $(\widetilde{V}_u, \widetilde{\pi}_u) = (F_u, \Psi_u)(p_u, \widetilde{V}_{\mathrm{Ch}(u)})$
8: **end for**
9: $V \leftarrow (1-\eta)V + \eta\widetilde{V}$
10: **end for**
11: Report oracle policies, recovery-consistency residuals, and semantic status

The proof of Corollary 4.14 is in Appendix F.8.

Interpretation. These are local fixed-input fidelity statements. They do not by themselves identify two recursively coupled fixed-point policies. A surrogate needs a full-operator equivalence argument or a semantic mismatch report.

Corollary 4.15 (Chain SeBIL (path tree))**.** *When T is a path, Tree-SeBIL reduces to the SeBIL backward update; Theorem C.4 gives the chain modulus ρ_b.*

The proof of Corollary 4.15 follows from the path specialization in Appendix H.3.

Proposition 4.16 (Fixed-point profile and protocol-optimality scope)**.** *At a value fixed point V^*, the oracle policies $\Psi_u(p_u^*, V^*_{\mathrm{Ch}(u)})$ solve their declared induced MDPs. They form a self-consistent nodewise-optimal profile only if $\mathcal{R}_u(V^*) = \Psi_u(p_u^*, V^*_{\mathrm{Ch}(u)})$ for every u. Transfer to an ordered game's regularized equilibrium additionally requires each induced problem to equal that player's actual deviation problem, with compatible reward ownership, entropy accounting, and history/transition lift (Appendix H.12). Ordering alone is insufficient. Contraction gives a unique value fixed point, not uniqueness of all game equilibria.*

The proof of Proposition 4.16 is in Appendix H.12.

Remark 4.17 (Tractability vs. operator statements). These are structural statements about fixed proxy protocols: they do not give a polynomial-time Nash solver for arbitrary games (4).

Theorem 4.18 (Inexact outer-loop convergence)**.** *For a certified contraction $\mathcal{F}$ with modulus $\rho < 1$ and stepsizes $\eta_k \in (0, 1]$ with $\sum_k \eta_k = \infty$, implemented updates $V^{k+1} = (1-\eta_k)V^k + \eta_k(\mathcal{F}(V^k) + \xi_k)$ converge to V^* under vanishing slack and satisfy $\limsup_k \|V^k - V^*\| \leq \bar{\xi}/(1-\rho)$ under uniformly bounded slack $\bar{\xi}$, provided the iterates stay in the certified domain.*

The local invariant-tube version in Theorem C.12 distinguishes errors inside averaging from errors added after it: their asymptotic radii are respectively $\bar{\xi}/(1-\rho)$ and $\bar{\xi}/[\eta(1-\rho)]$ for fixed η. Measured neural update norms do not certify either error bound.

The proof of Theorem 4.18 is in Appendix H.10.

5 Algorithms

5.1 Tree-SeBIL meta-algorithm (primary)

Remark 5.1 (Implementation contract). Algorithm 1 averages the completed value sweep once. At every iteration its input policies are reconstructed from that value profile. A neural wrapper that instead refreshes independent actors is not this value-only map; its discrepancy must be included in the error of Theorem 4.18. Temperature annealing is a tracking problem, not convergence to one fixed-temperature operator (Appendix F.3).

Table 1: Exact multi-pursuer interception bridge at $\gamma = 0.99$ (mean $\pm$ sample SD over 900 map instances). All policies are evaluated under the route-aware semantic union. Flip is the fraction whose top assignment differs from the semantic optimum.

Backup	Capture	Semantic residual	Flip (%)
Semantic union	0.548 ± 0.230	$< 10^{-10}$	0.0
KM semantic	0.548 ± 0.230	$< 10^{-10}$	0.0
Plain average	0.412 ± 0.196	0.142 ± 0.106	79.2
Scaled surrogate	0.366 ± 0.156	0.135 ± 0.076	68.6
Max child	0.393 ± 0.154	0.124 ± 0.084	69.3
Independent	0.363 ± 0.156	0.138 ± 0.077	69.7

Convergence point. A certified exact sweep has a unique value fixed point on $\mathcal{D}$. Its policy meaning is precisely Proposition 4.16. A certified surrogate targets its own fixed point; local shift invariance alone does not establish global equivalence, and a value-bias bound does not establish exact equilibrium.

5.2 SeBIL+X (shallow tree)

SeBIL+X combines a sequential team protocol with a tactical solver such as MAPPO, HAPPO, or HASAC; agents within a team may act simultaneously. Applying the operator theory to this hierarchy requires an explicit team-level oracle and its sensitivity bounds. Neural tactical updates approximate that oracle, so their discrepancy enters the slack model of Theorems 4.18 and C.12.

6 Experiments

The experiments examine three consequences of the theory: local decision changes under aggregation replacement, the distinction between norm expansion and finite settling, and the dependence of neural wrapper behavior on the task.

6.1 Independent semantic decision audit

The 900-map interception bridge enumerates finite pursuer-to-exit assignments and evaluates every policy under the same route-aware capture objective. For $J_A(\pi) = \langle \pi, Q_A \rangle + \alpha H(\pi)$, semantic residual is $\max_\pi J_A(\pi) - J_A(\pi_u)$; own residual uses Q_S instead. At $\gamma = 0.99$, the scaled surrogate has own-objective residual 5.2×10^{-18} but semantic residual 0.1347 (95% CI $[0.1297, 0.1396]$), and changes 68.6% of top assignments (Table 1). This is local decision mismatch, not a capture-rate reduction or a neural Pursuit result. Its KM control averages toward an already computed semantic policy, so preservation of that target does not test general Bellman convergence.

A held-out 180-map control separates route omission from gain reduction: scaling the route-aware union itself, with rewards and temperature unchanged, still gives semantic residual 0.1277 and 57.8% top-assignment changes. Route-weighting and temperature controls are in Appendix I.2.

6.2 Target-relative tabular checks

The soft-MDP coefficient study measures fixed-point distance to a predeclared scaled-$1\times$ target (Appendix G.1). Zero distance at $1\times$ is definitional. Without ancestor feedback, a complete postorder solve gives a constant outer map for every coefficient; the distances therefore measure target drift, not a stability threshold.

In a separate hard, old-child recurrence, all 2,200 held-out cells reach finite fixed points within six sweeps, although only 200 satisfy the unweighted contraction test. Of these fixed points, 1,400 exceed magnitude 10^6. Large values and norm expansion can therefore coexist with finite settling on an acyclic graph (Appendix G.2).

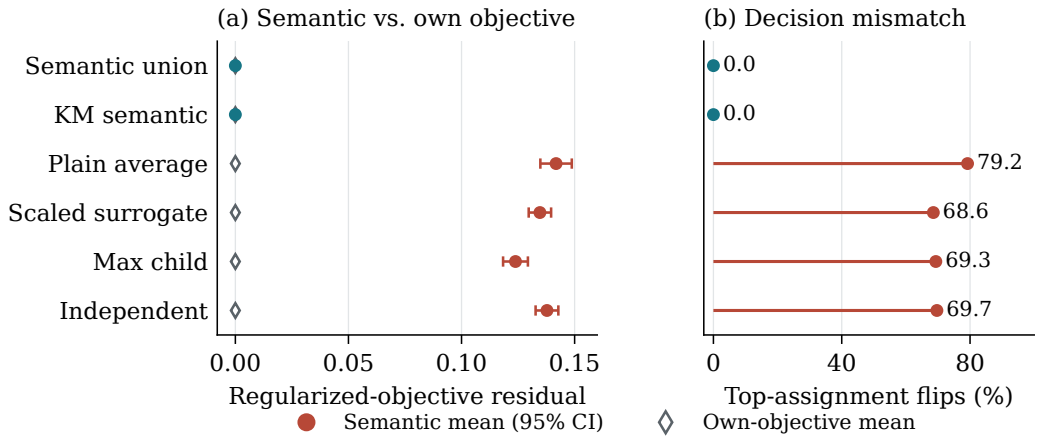


Figure 2: Independent semantic decision audit on 900 generated interception maps at $\gamma = 0.99$. Points show means and 95% normal intervals computed from the packaged per-map rows; diamonds are residuals under each method's own objective, while circles are evaluated under the route-aware semantic objective. The scaled surrogate is internally accurate but has nonzero semantic residual and decision flips. This is an assignment diagnostic, not learned capture performance.

6.3 Bounded neural wrapper evidence

Simulator stress and MPE assignment comparisons document wrapper behavior under their stated budgets (Appendix I.3). The 12 task/aggregation combinations exhibit different return rankings. The assignment intervention is conditional on its wrapper and mean-return convention. These descriptive results establish neither certified sensitivities nor an equilibrium guarantee.

Empirical interpretation. The experiments support semantic preservation of the declared joint objective and structure-dependent compute, not a tree-over-chain return advantage. The evidence is heterogeneous and is classified in the supplementary provenance record: the finite-tree and semantic-bridge rows are replayable from packaged inputs, the neural rows are replayable from frozen logs but not from a complete training environment, and the remaining summary-only observations are not re-executable from this package. Legal tree recursion recovers the exact joint value on the audited family (maximum absolute error $< 10^{-13}$; replayable diagnostic), while oracle-call counts and wall-clock time both track the widest sibling group— most sharply on star topologies under positive coupling—so both metrics must be reported. The latter compute result is supported only by aggregate summaries. On the independent 900-map interception bridge, the semantic union preserves the route-aware capture objective, whereas plain-average and scaled surrogates keep low own-objective residual yet induce nontrivial semantic residual and top-assignment flips.

Same-objective exact-versus-marginal-greedy mean capture gaps of about 0.010 (n=2) and 0.028 (n=4) are model-probability assignment diagnostics under that bridge; they are not executions of the learned simulator wrapper and not simulator-observed captures. Their source rows are unavailable in this package, so these values are summary-only observations and are not used as replayable table claims. An exploratory screen over three environments yields $0/60$ seed passes; a multi-step LBF probe (H=25) yields $0/40$ under an all-zero reward window; and a native-length pistonball probe (H=125, n=80) surfaces nonzero team return on $48/80$ seeds yet still yields $0/80$ unique tree-versus-path decisions. These counts have no packaged raw-run manifest and are supported only by aggregate summaries. Those three probes found no unique tree-versus-path decision on the evaluated instances. They do not show that every task suited to a tree protocol lacks a return advantage, and the exact recovery above is a numerical preservation check, not a learning-performance result. A fair independent real-capture executor for the interception objective under a frozen budget was not implemented, so that loop is unfinished rather than a completed negative result. We do not claim that tree protocols outperform chain protocols on return.

Limitations. The contraction results require valid uniform constants and an invariant domain. The sharp scalar threshold concerns an unweighted norm in a restricted family. The fidelity theorem holds at fixed continuation inputs; game-level transfer needs the additional compatibility contract. Finite assignment enumeration does not validate a learned dynamic game. Neural runs use few seeds and fixed protocols, and the reproduction record has legacy gaps (Appendix I).

7 Conclusion

Tree-SeBIL distinguishes contraction of a declared continuation map from preservation of its decisions. The tree path-sum certificate makes ancestor feedback and child propagation explicit; the local fidelity theorem identifies when gain reduction admits a binary instance with a changed action distribution at fixed continuation inputs. These results support auditable continuation design under stated interfaces, not automatic equilibrium recovery for arbitrary sequential games.

Reproducibility Statement

The local equations, assumptions, and proofs are in Section 4 and Appendices E and H; metrics, seed units, and budgets are in Appendix I. Reproducing the numerical studies requires their frozen generators and inputs. Selected neural traces support descriptive summaries, but a complete retraining environment and historical compute record are unavailable. Appendix I.1 documents corrections and exclusions.

Ethics Statement

This work studies algorithmic game-solving and MARL training on public simulators (OpenSpiel, PettingZoo, MuJoCo) without human-subject experiments. Strategic automation can affect fairness and safety when deployment imposes sequential protocols on agents; the protocol dependence discussed in Appendix I.3.1 precludes a claim of universally fair ordering. We do not release datasets containing personally identifiable information.

AI Use Statement

Generative AI tools assisted with mathematical revisions, implementation, and analysis of experimental results. The authors take responsibility for the final content.

References

[1] Abraham Berman and Robert J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. SIAM, 1994.

[2] Dimitri P Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 4th edition, 2012.

[3] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 793–802. PMLR, 2019.

[4] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

[5] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 2974–2982, 2018.

[6] Drew Fudenberg and Jean Tirole. *Game Theory*. MIT Press, 1991.

[7] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 2019.

[8] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 1861–1870. PMLR, 2018.

[9] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

[10] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

[11] Zun Li, Feiran Jia, Aditya Mate, Shahin Jabbari, Mithun Chakraborty, Milind Tambe, and Yevgeniy Vorobeychik. Solving structured hierarchical games using differential backward induction. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1107–1117. PMLR, 2022.

[12] Bor-Jiun Lin and Chun-Yi Lee. Heterogeneous agent q-weighted policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

[13] Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, Qiang Fu, Xiaojun Chang, and Yaodong Yang. Maximum entropy heterogeneous-agent reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] Qinghua Liu, Yuanhao Wang, and Chi Jin. Learning Markov games with adversarial opponents: Efficient algorithms and fundamental limits. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14036–14053. PMLR, 2022.

[15] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6382–6393, Red Hook, NY, USA, 2017. Curran Associates Inc.

[16] Weixin Mao, Kaiqing Zhang, Ray Srikant, Tamer Başar, and Zhuoran Yang. $\tilde{O}(t^{-1})$ convergence to (coarse) correlated equilibria in full-information general-sum markov games. In *Proceedings of the 41st International Conference on Machine Learning*, volume 242 of *Proceedings of Machine Learning Research*, pp. 35426–35456. PMLR, 2024.

[17] Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

[18] Bei Peng, Tabish Rashid, Christian Schroeder de Witt, Pierre-Alexandre Kamienny, Philip Torr, Wendelin Boehmer, and Shimon Whiteson. FACMAC: Factored multi-agent centralised policy gradients. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12208–12221, 2021.

[19] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[20] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 4295–4304. PMLR, 2018.

Symbol	Definition
$T = (V, E)$	Rooted directed tree; $\mathrm{Ch}(u)$, $\mathrm{Anc}(u)$, $\mathcal{L}_d$, $D_{\max}$
$\mathcal{H}_u$, $\mathcal{H}_i$	History at tree node u / chain agent i
$\mathcal{A}$, $L_{\mathcal{A}}$	Child-value aggregation; Lipschitz constant
$\mathcal{T}^s_{T,\alpha}$, $\mathcal{T}^u_{T,\alpha}$, $\mathcal{T}$	Semantic Tree-SeBIL / surrogate Tree-SeBIL / chain SeBIL operator
ρ_b	Backward-composed chain modulus (6)
ρ^s_T	Primary tree modulus $\|(I - B)^{-1}a\|_\infty$
M, $\rho(M)$	Matrix gain certificate $(I - B)^{-1}C$ and its spectral radius
ρ^ℓ_T	Maximum per-node scalar gain in the leaf-injection template
g_u	Per-node layer gain in Appendix F.1
B, C, a	Child-gain matrix, component-input matrix, and scalarized input vector in (4)
α, $\alpha_{\min}$, η	Entropy temperature (floor), KM damping
ϵ_k	One-step implementation slack
ϵ_u	Uniform surrogate-to-semantic operator mismatch

Table 2: Key notation (Tree-SeBIL and chain degeneracy).

[21] Aryaman Reddi, Gabriele Tiboni, Jan Peters, and Carlo D'Eramo. k-level policy gradients for multi-agent reinforcement learning. *arXiv preprint arXiv:2509.12117*, 2025.

[22] Reinhard Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4(1):25–55, 1975.

[23] Vedansh Sharma. Small-gain nash: Certified contraction to nash equilibria in differentiable games. *arXiv preprint arXiv:2512.06791*, 2025.

[24] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech M Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 2085–2087, 2018.

[25] Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning (ICML)*, pp. 330–337. Morgan Kaufmann, 1993.

[26] Muning Wen, Jakub Grudzien Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16509–16521, 2022.

[27] Han Zhong, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Can reinforcement learning find Stackelberg-Nash equilibria in general-sum markov games with myopically rational followers? *Journal of Machine Learning Research*, 24(35):1–52, 2023.

[28] Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32): 1–67, 2024.

[29] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, Carnegie Mellon University, 2010.

[30] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, pp. 1729–1736, 2007.

A NOMENCLATURE

B ADVANTAGE DECOMPOSITIONS

B.1 ORDERED TELESCOPING IDENTITY (TREE-GAD DIAGNOSTIC NOTATION)

Fix any forward topological order $\sigma = (u_1, \ldots, u_N)$ with $u_1 = r$ as an auxiliary linear extension for algebraic telescoping. For joint policy π, define $V_k^\pi(s, a_{<u})$ and $Q_k^\pi(s, a_{\leq u})$ as pre- and post-action values from agent k's perspective, and $A_{k,\pi}^u = Q_k^\pi(s, a_{\leq u}) - V_k^\pi(s, a_{<u})$. For this identity only, $a_{<u_i}$ is the revealed prefix in the order induced by σ. This is the standard advantage telescoping identity after imposing an auxiliary total reveal order. We call it Tree-GAD diagnostic notation because the identity itself does not assert that sibling actions are visible in the ancestor-only tree protocol. The chain version appears in HARL-style sequential update analyses (28).

Remark B.1 (Ordered history in Tree-GAD). For branching trees, $a_{<u_i}$ in the sum is the order σ-prefix $(a_{u_1}, \ldots, a_{u_{i-1}})$ so that $Q_k^\pi(s, a_{\leq u_i}) = V_k^\pi(s, a_{<u_{i+1}}^{(\sigma)})$. This auxiliary reveal order does not enlarge the policy information sets $\mathcal{H}_u$: sibling actions remain unobserved in the ancestor-only tree protocol and the layerwise algorithm.

Lemma B.2 (Ordered GAD telescoping)**.** *For any $k \in V$, state s, and joint action a along σ,*

$$Q_k^\pi(s, a) = V_k^\pi(s) + \sum_{i=1}^{N} A_{k,\pi}^{u_i}(s, a_{<u_i}, a_{u_i}). \tag{5}$$

Proof. Appendix E.1.

B.2 CHAIN GAD (SPECIAL CASE)

Under a chain with agent indices $i = 1, \ldots, N$ and Assumption B.3, Lemma B.2 reduces to the chain lemma below (credit-assignment notation used in SeBIL+X).

Assumption B.3 (Intra-stage sequential semantics (chain benchmarks))**.** Within a single decision stage at fixed $s \in \mathcal{S}$, agents act in order $1, \ldots, N$ before the Markov transition; stage rewards for agent k are evaluated under the same s. For all k, π, and $i < N$, the bridge identity $Q_k^\pi(s, a_{\leq i}) = V_k^\pi(s, a_{<i+1})$ holds.

Lemma B.4 (Generalized Advantage Decomposition (chain))**.** *Under Assumption B.3, for all k, s, and $a = (a_1, \ldots, a_N)$, $Q_k^\pi(s, a) = V_k^\pi(s) + \sum_{i=1}^{N} A_{k,\pi}^i(s, a_{<i}, a_i)$.*

Proof. Appendix H.1.

Definition B.5 (Ordered ME-SPNE and ancestor-only tree optimality)**.** Under Assumption B.3, fix a reachable intra-stage prefix of a chain or another explicitly ordered protocol. Agents $i, i{+}1, \ldots, N$ move in fixed order before the Markov transition; each prefix is an information set. **NE** means no unilateral strategy deviation is profitable. **SPNE** requires this property in every well-defined subgame, including off-path subgames; one-shot deviations suffice only when the relevant deviation principle applies. At temperature $\alpha > 0$, **ME-SPNE** denotes the corresponding regularized subgame-perfect profile for an explicitly specified entropy objective. Transferring induced-node solutions to this concept requires the compatibility conditions in Appendix H.12. For an ancestor-only tree T, **node-wise soft optimality** means that every node policy is an entropy-regularized best response at every reachable ancestor history $h_u \in \mathcal{H}_u$, holding the declared child-continuation semantics fixed. This property is relative to declared induced objectives. A full order is necessary for the ordered specialization but is not sufficient for game-level equilibrium transfer.

Remark B.6 (Scope of equilibrium and complexity claims). Contraction statements characterize idealized operators on G^σ / T; they are **not** polynomial-time Nash solvers for arbitrary games (4). Deep RL reports returns under matched interfaces, not certified equilibrium gaps (Section 6.2).

C AUXILIARY OPERATOR STATEMENTS

C.1 CHAIN SEBIL (PATH DEGENERACY)

Definition C.1 (Idealized chain SeBIL operator)**.** For protocol σ, $\mathcal{T}$ is Definition 4.1 on a path, using the declared recovery map $\mathcal{R}_{<i}(V)$ and the already solved child values. An extended successor tuple is allowed only with the stated uniform sensitivity bound.

Assumption C.2 (Parametric sensitivity (chain))**.** Agent-level fixed points satisfy

$$\|V_i^*(\theta^a) - V_i^*(\theta^b)\|_\infty \le C_a\|\theta_a^a - \theta_a^b\|_\infty + \frac{\gamma L_c}{1-\gamma}\|\theta_c^a - \theta_c^b\|_\infty.$$

Corollary C.3 (Explicit sensitivity template)**.** *On the common invariant domain, uniform parameter sensitivities of reward, lifted transition kernel, child integrand, and policy recovery give the constants in Assumption C.2, as detailed in Appendix H.4.*

Theorem C.4 (Conditional contraction of idealized chain SeBIL)**.** *With $\beta_c = \gamma L_c/(1-\gamma)$ and non-expansive predecessor embedding, define*

$$\rho_b = \begin{cases} C_a \dfrac{1-\beta_c^N}{1-\beta_c}, & \beta_c \neq 1, \\ NC_a, & \beta_c = 1. \end{cases} \tag{6}$$

If $\rho_b < 1$, then $\mathcal{T}$ is a ρ_b-contraction.

The proof is in Appendix H.3.

Remark C.5 (Naive $\max$ bounds are insufficient). Successor noise composes through the backward pass, so $\max\{C_a, \beta_c\} < 1$ need not imply $\rho_b < 1$.

Proposition C.6 (Certificate-filtered finite protocol selection)**.** *For a finite candidate set of protocols Σ, let $A(\sigma) \in \{0,1\}$ be a declared admissibility predicate that checks the available stability certificate and semantic-fidelity status for protocol σ. Let $\Sigma_a = \{\sigma \in \Sigma : A(\sigma) = 1\}$. If $\Sigma_a \neq \emptyset$ and each admissible fixed protocol has value $J(\sigma)$ and validation estimate $\widehat{J}(\sigma)$ with $|\widehat{J}(\sigma) - J(\sigma)| \le \varepsilon_v$ simultaneously for all $\sigma \in \Sigma_a$, then*

$$\widehat{\sigma} \in \arg\max_{\sigma \in \Sigma_a} \widehat{J}(\sigma)$$

satisfies $J(\widehat{\sigma}) \ge \max_{\sigma\in\Sigma_a} J(\sigma) - 2\varepsilon_v$. If $\Sigma_a = \emptyset$, the fail-closed rule returns no certified semantic protocol rather than selecting by return alone.

The proof of Proposition C.6 is in Appendix H.11.

Corollary C.7 (Depth-aware scalar sufficient bounds)**.** *If $a_u \le \bar{a}$ and the child-gain row sums satisfy $\sum_v b_{uv} \le \bar{s}$, then depth D gives the sufficient condition $\bar{a}\sum_{k=0}^{D}\bar{s}^k < 1$. In particular, $b_{uv} \le \bar{b}$ and maximum branching $K_{\max}$ allow $\bar{s} = K_{\max}\bar{b}$. In the separate leaf-injection max-child template this sharpens to $\gamma(\gamma L_{\mathcal{A}}/(1-\gamma))^D < 1$ without a branching multiplier.*

The proof is in Appendix E.6.

Remark C.8 (Iff only under Assumption F.1). The converse is not a general-tree statement: it holds only for the exact linear aggregation/gain recursion in Assumption F.1.

Theorem C.9 (Linear-gain iff)**.** *Under Assumption F.1, $\mathcal{T}_T^s$ is a contraction in $\|\cdot\|_\infty$ iff $\rho_T^s < 1$.*

The proof of Theorem C.9 is in Appendix F.

Proposition C.10 (Sharpness on a depth-D chain embed)**.** *For the exact linear leaf-injection chain of depth D with $\mathcal{A}(x) = L_{\mathcal{A}}x$, $L_{\mathcal{A}} \ge \frac{1-\gamma}{\gamma}(1/\gamma)^{1/D}$ makes the template non-contractive in $\|\cdot\|_\infty$.*

The proof of Proposition C.10 is in Appendix F.1.

Proposition C.11 (Scaled average normalized surrogate)**.** *The rule $\mathcal{A}(x) = \frac{1-\gamma}{\gamma K}\sum_{k=1}^{K} x_k$ has $L_{\mathcal{A}} = (1-\gamma)/\gamma$ and gives depth-independent modulus γ in the leaf-injection template for any D, K; a general tree still uses its structured certificate, and exact equivalence to a semantic game requires a full-model invariance argument; a bias bound only controls approximation.*

The proof of Proposition C.11 is in Appendix F.2.

Theorem C.12 (Local inexact small-gain tube)**.** *Let V^* be a fixed point of $\mathcal{F}$ and suppose $\mathcal{F}$ is a $\rho_\ell < 1$ contraction on the tube $\mathcal{U}_r = \{V : \|V - V^*\|_{\infty,w} \le r\}$. If the implemented update is $V^{k+1} = \mathcal{F}(V^k) + \xi_k$ with $\|\xi_k\|_{\infty,w} \le \bar{\xi}$ and $\rho_\ell r + \bar{\xi} \le r$, then every iterate initialized in $\mathcal{U}_r$ remains in $\mathcal{U}_r$ and*

$$\limsup_{k\to\infty} \|V^k - V^*\|_{\infty,w} \le \bar{\xi}/(1-\rho_\ell).$$

For $V^{k+1} = (1-\eta)V^k + \eta(\mathcal{F}(V^k) + \xi_k)$, replace the recurrence by $q_\eta\delta_k + \eta\bar{\xi}$, where $q_\eta = 1-\eta(1-\rho_\ell)$; the radius remains $\bar{\xi}/(1-\rho_\ell)$. For post-averaging error $V^{k+1} = (1-\eta)V^k + \eta\mathcal{F}(V^k) + \xi_k$, invariance requires $q_\eta r + \bar{\xi} \le r$ and the radius is $\bar{\xi}/[\eta(1-\rho_\ell)]$. Updates must remain in the stated domain.

The proof of Theorem C.12 is in Appendix H.9.

D Proof roadmap (logical dependency chain)

The dependency chain below distinguishes standard primitives, tree composition, and the additional game-transfer contract.

1. **Credit assignment notation (no contraction yet).** Lemmas B.2 and B.4 (Appendices E.1 and H.1) are standard ordered telescoping/GAD identities; the tree linear extension is an algebraic reveal order, not an enlargement of sibling information sets.
2. **Single-node well-posedness.** Lemma H.1 (Appendix H.2) cites the standard discounted soft Bellman contraction, hence a unique oracle fixed point exists at each backward step.
3. **Parametric stability of one fixed point.** Lemma E.3 (Appendix E.2) cites the standard contraction-map perturbation bound for Lipschitz dependence of $x^*(\theta)$.
4. **Verifiable sensitivity and structured tree small-gain.** Proposition 4.3 (Appendix E.2) derives a primitive sufficient route to Assumption 4.2. Theorems 4.6 and 4.7 and Corollary C.7 (Appendices E.4 to E.6): the tree-specific recursion gives $e \le Be + Cd$; the weighted contraction step uses standard nonnegative-matrix small-gain / Perron–Frobenius facts.
5. **Restricted linear-gain iff.** Theorems C.9 and F.2 (Appendix F): iff *only under* Assumption F.1.
6. **Tree spatial composition.** Lemma E.2 and Theorem 4.4 and Propositions C.10 and C.11 (Appendices F.1 and F.2): under the additional leaf-injection template, layer gains g_u give the scalar depth boundary and scaled average gives $\rho_T^\ell = \gamma$ for any depth. General ancestor-sensitive trees remain governed by the structured certificate in Item 4.
7. **Chain small-gain composition.** Corollary C.3 and Assumption C.2 (Appendix H.4) $\Rightarrow$ Theorem C.4 (Appendix H.3): backward recurrence $e_i \le C_a + \beta_c e_{i+1}$ yields ρ_b; Remark C.5 records why a naive $\max$ bound fails.
8. **Stabilization, inexact updates, surrogate gap, and finite protocol choice.** Propositions 4.8 to 4.10, C.6 and G.2, Theorems 4.12, 4.13, 4.18 and C.12, and Corollary 4.14 (Appendices F.5 to F.8 and H.8 to H.11): fixed-point-preserving averaging, spectral KM boundary, the stability–fidelity impossibility, the sharp quantitative gain–mismatch frontier, surrogate bias, local inexact stability, one-step slack, and certificate-filtered finite protocol model selection.
9. **Equilibrium meaning (regularized).** Proposition 4.16 (Appendix H.12): oracle solutions give induced-node optimality. A self-consistent profile additionally needs policy recovery compatibility, and game transfer needs the verified reward/transition/history contract. Ordering alone proves neither.
10. **Implementation gap (algorithms).** Definitions 4.1 and C.1 and Remark 5.1: sampled-action and Gauss–Seidel sweeps realize approximate backups, not the ideal oracle map.

E PROOFS FOR TREE STRUCTURE

E.1 VERIFICATION OF TREE-GAD (ORDERED TELESCOPING)

Verification. Fix $\sigma = (u_1, \dots, u_N)$ with $u_1 = r$ and ordered prefixes $a^{(\sigma)}_{<u_i}, a_{\le u_i}$. Because σ is topological, $a^{(\sigma)}_{<u_{i+1}} = a_{\le u_i}$ for $i < N$, hence $Q^\pi_k(s, a_{\le u_i}) = V^\pi_k(s, a^{(\sigma)}_{<u_{i+1}})$. Expanding $\sum_{i=1}^N A^{u_i}_{k,\pi}$ telescopes to $Q^\pi_k(s,a) - V^\pi_k(s)$ using $a^{(\sigma)}_{<u_1} = \emptyset$ and $a_{\le u_N} = a$. □

E.2 PRIMITIVE SENSITIVITY AND INVARIANT DOMAINS

The node model is exactly (2)–(3). The local recurrence $W(h')$ appears once, inside Q; log-sum-exp then gives a state value. No extra expectation of W is added outside log-sum-exp.

Envelope and completeness. For $m_u = |\mathcal{A}_u|$, $\max_a Q_a \le \alpha \log \sum_a e^{Q_a/\alpha} \le \max_a Q_a + \alpha \log m_u$. For held child values bounded by H_v, $|G_u| \le \bar{G}_u$. Thus

$$\|\mathcal{B}^{p,z}_u W\|_\infty \le R_u + \gamma \bar{G}_u + \gamma \|W\|_\infty + \bar{\alpha} \log m_u \le H_u$$

whenever $\|W\|_\infty \le H_u$. These bounds are obtained from leaves upward, so there is no circular choice of envelopes. On finite histories the spaces are Euclidean; on compact histories the stated continuity/Feller assumptions make the maps act on $C(\mathcal{H}_u)$. Its closed sup-norm ball is complete. The finite product $\mathcal{D}$ is complete in ordinary or positively weighted sup norm. The local Bellman map is a γ-contraction, has a unique fixed point in its ball, and reverse-depth composition preserves $\mathcal{D}$. This also gives a convex domain for global averaging.

Policy reconstruction is part of the model. A reconstruction map must be supplied as a deterministic function, not inferred from one scalar value. For example, a finite model can declare $\mathcal{Q}_u(V)(h,a) = \bar{r}_u(h,a) + \gamma \sum_{t,y} J_{ut}(y \mid h, a) V_t(y)$ with fixed bounded coefficients J_{ut}. Its support and absolute row sums specify its dependencies and Lipschitz bounds. This example closes the iteration but does not guarantee compatibility with the oracle's policy. Such compatibility must be checked separately at the fixed point. In general, if

$$\|\mathcal{Q}_u(V^a) - \mathcal{Q}_u(V^b)\|_\infty \le \sum_t q_{ut} \|V^a_t - V^b_t\|_\infty,$$

then the softmax Jacobian gives $\|\mathcal{R}_u(V^a) - \mathcal{R}_u(V^b)\|_{\infty,1} \le (2/\alpha) \sum_t q_{ut} \|V^a_t - V^b_t\|_\infty$. Indeed, for logit perturbation d, the sum of absolute derivatives is at most $2\|d\|_\infty/\alpha$. Stack policies in a declared norm and retain the resulting constants in ℓ_{ut}; a max norm on a tuple does not silently remove sums across predecessor policies.

Lemma E.1 (Uniform partial-parameter bound)**.** *For a node, hold z independently of p. Assume $\|r^a - r^b\|_\infty \le L_r \|p^a - p^b\|$, $\|K^a - K^b\|_{\infty,d} \le L_K \|p^a - p^b\|$, and $\|G^{p^a}(z,\cdot) - G^{p^b}(z,\cdot)\|_\infty \le L_{G,p} \|p^a - p^b\|$. These bounds must hold for every z in the child balls. Then*

$$\|F_u(p^a, z) - F_u(p^b, z)\|_\infty \le A_u \|p^a - p^b\|, \quad A_u = \frac{L_r + \gamma L_K (H_u + \bar{G}_u) + \gamma L_{G,p}}{1 - \gamma}.$$

For the kernel representation in the main text, G has no direct p dependence and $L_{G,p} = 0$.

Proof. Insert the same W and z in both backups. Splitting the expectation into a kernel difference and an integrand difference bounds the logit change by $[L_r + \gamma L_K (H_u + \bar{G}_u) + \gamma L_{G,p}] \|p^a - p^b\|$. Log-sum-exp is 1-Lipschitz. The contraction perturbation bound divides this by $1 - \gamma$. The reward term is not discounted. At a leaf, $\bar{G}_u = L_{G,p} = 0$, so policy Lipschitzness alone still does not determine A_u. □

Proof of Proposition 4.3. Compose the preceding partial bound with the declared recovery/parameter bound ℓ_{ut}. This gives $c_{ut} = A_u \ell_{ut}$ for every independently held child tuple. In the special ancestor-local case, a max-ancestor bound with factor γ is valid only when the resulting row sum is at most γ; it is not a consequence of softmax regularity alone. □

Lemma E.2 (Child-channel sensitivity). *With p fixed, suppose r and K do not depend on z. If $\mathcal{A}_u$ has modulus L_u, then*

$$\|\mathcal{B}_u^{p,z^a} W - \mathcal{B}_u^{p,z^b} W\|_\infty \le \gamma L_u \max_{v \in \mathrm{Ch}(u)} \|z_v^a - z_v^b\|_\infty,$$

and the fixed-point bound has coefficient $\gamma L_u/(1-\gamma)$. More generally, if r, K, G have coordinate sensitivities $L_{r,z,v}, L_{K,z,v}, L_{G,z,v}$ uniformly on the same domain, a valid child coefficient is

$$b_{uv} = \frac{L_{r,z,v} + \gamma L_{K,z,v}(H_u + \bar{G}_u) + \gamma L_{G,z,v}}{1-\gamma}.$$

A common invariant envelope is required for that extended model as well.

Proof. Apply the same splitting as above with p held. For a fixed kernel, expectation is non-expansive in its integrand, so its direct value channel remains γL_u even when the kernel has zero parameter sensitivity. Coordinate bounds follow by replacing child coordinates one at a time and summing; contraction perturbation yields the fixed-point coefficients. □

Lemma E.3 (Standard contraction-map perturbation bound, cited primitive). *Let $T(\cdot;\theta)$ be a γ-contraction and $T(x;\cdot)$ be L_θ-Lipschitz in θ. Then $\|x^*(\theta_1) - x^*(\theta_2)\| \le \frac{L_\theta}{1-\gamma}\|\theta_1 - \theta_2\|$.*

Citation. This bound is an immediate consequence of the standard contraction-mapping argument used in discounted dynamic programming (2; 19); we use it as a cited primitive rather than as a new lemma.

E.3 Structured small-gain proofs

Lemma E.4 (Component-wise node recursion). *For $V^a, V^b \in \mathcal{D}$, define $d_t = \|V_t^a - V_t^b\|_\infty$ and $e_u = \|(\mathcal{T}V^a)_u - (\mathcal{T}V^b)_u\|_\infty$. Under the uniform partial sensitivities, $e \le Be + Cd$. If $Cd \le a\|d\|_\infty$, then $e \le Be + a\|d\|_\infty$.*

Proof. Let p^a, p^b be the recovered node parameters and z^a, z^b the already solved child tuples. Insert $F_u(p^b, z^a)$ between $F_u(p^a, z^a)$ and $F_u(p^b, z^b)$. The first difference is bounded by $\sum_t c_{ut} d_t$ because the ancestor bound holds uniformly for held z^a. The second is bounded by $\sum_v b_{uv} e_v$. Leaves have no second term. A max-child bound can conservatively be replaced by its coordinate sum; retain tighter coordinate bounds where available. Stacking proves both recursions. In particular, $a = C\mathbf{1}$ always works. Unlike the separate leaf-injection template, internal direct-input terms need not vanish or equal γ. □

E.4 Weighted small-gain proof

Proof. Since T is acyclic, B is nilpotent after a topological ordering, so $(I-B)^{-1} = \sum_{k \ge 0} B^k$ is finite and nonnegative. Lemma E.4 implies $e \le (I-B)^{-1}Cd = Md$. If $Mw \le \lambda w$ and $\|V^a - V^b\|_{\infty,w} \le r$, then $d \le rw$ componentwise. Therefore $e \le Md \le rMw \le r\lambda w$, which is exactly

$$\|\mathcal{T}_T^s V^a - \mathcal{T}_T^s V^b\|_{\infty,w} \le \lambda \|V^a - V^b\|_{\infty,w}.$$

For a nonnegative matrix M, the equivalence between the positive-vector certificate $Mw \le \lambda w$ and $\rho(M) < 1$ is the standard Perron–Frobenius/Collatz–Wielandt characterization (1; 9). Thus the weighted-norm certificate is feasible exactly inside the spectral region $\rho(M) < 1$. If an actual finite-dimensional affine operator has linear part M on a domain containing a neighborhood of its fixed point, $\rho(M) \ge 1$ precludes contraction in any norm. Thus the spectral boundary is necessary for contraction of that exact gain map in any equivalent finite-dimensional norm; it is not claimed as a necessity theorem for arbitrary nonlinear semantic operators. □

E.5 STRUCTURED SMALL-GAIN PROOF

Proof. Let $e \in \mathbb{R}^{|V|}$ stack node errors and $B \in \mathbb{R}_+^{|V|\times|V|}$ with $B_{uv} = b_{uv}$. From Lemma E.4, $e \le Be + a\,\delta$. Since T is acyclic, index nodes by nondecreasing depth (leaves last); then B is strictly upper triangular in that order, hence B is nilpotent and $(I-B)^{-1} = \sum_{k=0}^{|V|-1} B^k$ exists with nonnegative entries. Therefore $e \le (I-B)^{-1} a\,\delta$ entrywise, and

$$\|\mathcal{T}_T^s V^a - \mathcal{T}_T^s V^b\|_\infty = \max_u e_u = \|e\|_\infty \le \|(I-B)^{-1}a\|_\infty\,\delta.$$

If $\rho_T^s = \|(I-B)^{-1}a\|_\infty < 1$, the map is a ρ_T^s-contraction. □

E.6 SCALAR DEPTH BOUND AND CERTIFICATE COMPUTATION

Proof. If $a_u \le \bar{a}$ and $\sum_v b_{uv} \le \bar{s}$ for every row, then $B\mathbf{1} \le \bar{s}\mathbf{1}$ and inductively $B^k\mathbf{1} \le \bar{s}^k\mathbf{1}$. Every directed path from a node to a descendant has length at most D, so $B^k = 0$ for $k > D$. Hence $(I-B)^{-1}a \le \bar{a}\sum_{k=0}^{D}\bar{s}^k\mathbf{1}$ entrywise and $\rho_T^s \le \bar{a}\sum_{k=0}^{D}\bar{s}^k$. If only $b_{uv} \le \bar{b}$ and maximum branching $K_{\max}$ are known, take $\bar{s} = K_{\max}\bar{b}$. For the leaf-injection template in Theorem 4.4, a leaf perturbation contributes $\gamma\delta$ at the leaf and is multiplied by $\bar{b} = \gamma L_{\mathcal{A}}/(1-\gamma)$ along each edge. The root gain is therefore at most $\gamma \max_{\ell\in\mathcal{L}} \bar{b}^{\mathrm{dist}(r,\ell)}$, and the global gain is at most $\max_{0\le k\le D}\gamma\bar{b}^k$. When $\bar{b} \le 1$, this global bound is $\gamma < 1$; a shallower leaf can dominate the root bound. When $\bar{b} > 1$, requiring $\gamma\bar{b}^D < 1$ gives $L_{\mathcal{A}} < \frac{1-\gamma}{\gamma}(1/\gamma)^{1/D}$ for $D \ge 1$. The same sufficient condition therefore holds without assuming equal leaf depths. □

For the scalarized certificate obtained from $Cd \le a\|d\|_\infty$, no matrix inverse is needed in implementation. Process nodes in reverse topological order and compute

$$h_u = a_u + \sum_{v\in\mathrm{Ch}(u)} b_{uv}h_v.$$

Because the graph is a tree, $h = (I-B)^{-1}a$ exactly, and $\rho_c = \max_u h_u$. For the matrix certificate, run the same dynamic program once per input component z with $a_u = C_{uz}$; the resulting columns form $M = (I-B)^{-1}C$. This certificate can use global conservative constants, empirical Lipschitz upper bounds from tabular models, or problem-specific analytic constants. Only certified upper bounds are used for theorem claims; empirical estimates are reported separately as mechanism diagnostics.

E.7 WORKED WEIGHTED CERTIFICATE

Order the depth-2 chain as (r, u, ℓ) and let the child-gain matrix be

$$B = \begin{pmatrix} 0 & b & 0 \\ 0 & 0 & b \\ 0 & 0 & 0 \end{pmatrix}, \qquad C = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \gamma \end{pmatrix}.$$

Since $B^3 = 0$,

$$(I-B)^{-1} = I + B + B^2 = \begin{pmatrix} 1 & b & b^2 \\ 0 & 1 & b \\ 0 & 0 & 1 \end{pmatrix},$$

and therefore

$$M = (I-B)^{-1}C = \begin{pmatrix} 0 & 0 & b^2\gamma \\ 0 & 0 & b\gamma \\ 0 & 0 & \gamma \end{pmatrix}.$$

The ordinary sup-norm certificate is $\|M\|_\infty = \max\{b^2\gamma, b\gamma, \gamma\}$, which equals 3.6 when $\gamma = 0.9$ and $b = 2$. However M is upper triangular with diagonal $(0, 0, \gamma)$, so $\rho(M) = \gamma = 0.9$. With $w = (4, 2, 1)^\top$,

$$Mw = (3.6, 1.8, 0.9)^\top = 0.9(4, 2, 1)^\top = 0.9w.$$

Hence Theorem 4.6 certifies 0.9-contraction in $\|\cdot\|_{\infty,w}$ even though the unweighted certificate fails.

Remark E.5 (Why the scalar bound is not a global iff). Different child weights or inactive branches can yield the same global $L_{\mathcal{A}}$ but different ρ_T^s; see Remark E.6.

Remark E.6 (Counterexample: same $(\gamma, D, L_{\mathcal{A}})$, different contraction). Let $D = 1$, with root r and two leaf children u, v. The same aggregation modulus can bound two different actual gain matrices, one attaining the bound on the active branch and one with much smaller realized sensitivity. Both satisfy the same global Lipschitz bound $L_{\mathcal{A}}$, but ρ_T^s can differ, so a single scalar threshold cannot be necessary for all trees.

F RESTRICTED LINEAR-GAIN CLASS (IFF)

Assumption F.1 (Linear aggregation and exact linear gains)**.** At each u, $\mathcal{A}_u(x) = \sum_{v \in \mathrm{Ch}(u)} w_{uv} x_v$ with $w_{uv} \geq 0$. A fixed nonnegative recursion $e \leq Be + a\delta$ bounds every input pair, with $B_{uv} > 0$ only for child edges. In addition, a nonzero admissible pair attains $e = Be + a\delta$, where $\delta = \|V^a - V^b\|_\infty > 0$. This attainability is an extra hypothesis; linear aggregation alone does not establish it.

Theorem F.2 (Contraction iff $\rho_T^s < 1$ *within* Assumption F.1 only)**.** *Under Assumption F.1 (linear aggregation and exact linear gain recursion;* ***not*** *a general-tree claim), with B nilpotent and $\rho_T^s = \|(I - B)^{-1}a\|_\infty$:*

(i) If $\rho_T^s < 1$, then $\mathcal{T}_T^s$ is a ρ_T^s-contraction.

(ii) If $\rho_T^s \geq 1$, then $\mathcal{T}_T^s$ is not a contraction in $\|\cdot\|_\infty$ on the constructed family attaining the gain template.

Hence, within Assumption F.1, contraction is equivalent to $\rho_T^s < 1$.

Proof. **(i)** is Theorem 4.7. **(ii)** Assume $\rho_T^s \geq 1$. Let $e^* := (I - B)^{-1}a$ (nonnegative because B is nilpotent with nonnegative entries). Then $e^* = Be^* + a$ and $\|e^*\|_\infty = \rho_T^s \geq 1$. By Assumption F.1, there is an admissible nonzero pair with input distance δ whose output errors are δe^*. Normalize the error ratio by δ; no unit-separated pair is required. Hence $\|\mathcal{T}_T^s V^a - \mathcal{T}_T^s V^b\|_\infty = \delta\|e^*\|_\infty \geq \delta$, so no uniform contraction modulus in $(0, 1)$ exists on this family. Combining (i)–(ii), contraction within Assumption F.1 is equivalent to $\rho_T^s < 1$. □

F.1 TREE CONTRACTION PROOF

Let $\delta = \|V^a - V^b\|_\infty$, $V'^a = \mathcal{T}_T^s(V^a)$, and $V'^b = \mathcal{T}_T^s(V^b)$. This subsection concerns the additional *leaf-injection template*: input-profile perturbations enter each leaf with gain at most γ, while an internal node has no direct input term and changes only through its post-sweep child tuple. This restriction is stronger than Assumption 4.2; without it, every internal ancestor-channel term must be retained in C and Theorem 4.7 applies instead.

Set $b = \gamma L_{\mathcal{A}}/(1 - \gamma)$, define $g_u = \gamma$ on leaves, and for each internal node set

$$g_u = b \max_{v \in \mathrm{Ch}(u)} g_v.$$

The leaf bound holds by the template definition. At an internal node, Lemmas E.2 and E.3 and the absence of a direct input term give

$$\|V_u'^a - V_u'^b\|_\infty \leq b \max_{v \in \mathrm{Ch}(u)} \|V_v'^a - V_v'^b\|_\infty \leq g_u \delta.$$

Reverse-depth induction therefore proves the per-node bounds, with $\rho_T^\ell \triangleq \max_{u \in V} g_u$. If $L_{\mathcal{A}} \leq (1 - \gamma)/\gamma$, then $b \leq 1$ and every $g_u \leq \gamma$, so the leaf-injection map is a γ-contraction.

For a tree of depth D,

$$\rho_T^\ell \leq \max_{0 \leq k \leq D} \gamma b^k.$$

Thus $L_{\mathcal{A}} < (1 - \gamma)/\gamma \cdot (1/\gamma)^{1/D}$ is sufficient; when $b \leq 1$ this is automatic from $\gamma < 1$, and when $b > 1$ it is exactly $\gamma b^D < 1$. On the exact linear depth-D chain, the bound is attained and gives the stated sharp threshold.

Sharpness on a depth-D chain embed. On the leaf-injection chain embed with $\mathcal{A}(x) = L_{\mathcal{A}}x$, a leaf perturbation ϵ propagates upward with factor $\gamma L_{\mathcal{A}}/(1-\gamma)$ per edge, so the root sensitivity is $\gamma(\gamma L_{\mathcal{A}}/(1-\gamma))^D$. If $L_{\mathcal{A}} \geq (1-\gamma)/\gamma \cdot (1/\gamma)^{1/D}$, then $\rho_T^{\ell} \geq 1$ and the operator is not a contraction on that family.

An explicit exact-gain witness. For clarity, the sharp family can be defined directly on scalar messages $X_0, \ldots, X_D$ along a path. Set $Y_D = \gamma X_D$ and $Y_j = bY_{j+1}$ for $j < D$, where $b = \gamma L/(1-\gamma)$. A one-state, one-action local oracle realizes the leaf with reward $(1-\gamma)\gamma X_D$ and the internal equations with $W = \gamma W + \gamma L Y_{j+1}$. This is an auxiliary leaf-input template, not an ancestor-only policy-feedback game. On the box $|X_j| \leq b^{D-j}H$ for $b > 0$, the map is invariant and its ordinary sup-norm modulus is exactly $\max_{0\leq k\leq D} \gamma b^k$, attained by perturbing only X_D a sufficiently small amount. Its linear matrix has spectral radius γ, even beyond the ordinary-norm threshold. The $b = 0$ case is immediate on a nondegenerate box. This witness proves norm sharpness without asserting divergence or a general game embedding.

F.2 Scaled-average surrogate proof

Proof. For $\mathcal{A}(x) = \frac{1-\gamma}{\gamma K}\sum_k x_k$, $\|\mathcal{A}(x) - \mathcal{A}(y)\|_\infty \leq \frac{1-\gamma}{\gamma}\|x-y\|_\infty$, so $L_{\mathcal{A}} = (1-\gamma)/\gamma$ and $\gamma L_{\mathcal{A}}/(1-\gamma) = 1$. In the leaf-injection template, each internal gain satisfies $g_u = \max_{v\in \mathrm{Ch}(u)} g_v$, hence $g_u = \gamma$ for all u by induction from leaves, and $\rho_T^{\ell} = \gamma$. For a general ancestor-sensitive tree, scaled average changes the child-gain matrix B but does not remove the direct-input matrix C; contraction must therefore be checked with $M = (I-B)^{-1}C$. □

F.3 Annealing-KM tracking (diagnostic)

Under scaled average in the leaf-injection template, $\mathcal{T}^u_{T,\alpha}$ is a γ-contraction for fixed α. With $V^{(k+1)} = (1-\lambda_k)V^{(k)} + \lambda_k \mathcal{T}^u_{T,\alpha_k}(V^{(k)})$ and $d_k = \|V^{(k)} - V^*(\alpha_k)\|_\infty$, if $\|V^*(\alpha) - V^*(\alpha')\|_\infty \leq K_\alpha|\alpha - \alpha'|$ on $[\alpha_{\min}, \alpha_0]$,

$$d_{k+1} \leq [1 - \lambda_k(1-\gamma)]d_k + K_\alpha|\alpha_k - \alpha_{k+1}|.$$

Quasi-static annealing can drive $d_k \to 0$ only under additional uniqueness/regularity of $\alpha \mapsto V^*(\alpha)$; we use this as a diagnostic, not an unconditional exact-SPNE guarantee.

F.4 Per-layer damping versus global averaging

For a separately defined locally damped sweep, set

$$Y_u = (1-\eta)X_u + \eta F_u(p_u(X), Y_{\mathrm{Ch}(u)})$$

in reverse depth order. Fixed points obey $X_u = F_u(p_u(X), X_{\mathrm{Ch}(u)})$, so this sweep preserves fixed points for $\eta > 0$. However its complete error recursion includes $(1-\eta)d_u$, ancestor feedback, and ηb_{uv} child propagation. It is not $(1-\eta)X + \eta\mathcal{T}(X)$.

In a leaf-injection experiment where all other old-value and parameter differences are held at zero, an injected leaf-output difference traverses D edges with factor $(\eta b)^D$, $b = \gamma L_{\mathcal{A}}/(1-\gamma)$. Thus $\eta b \leq 1$ prevents depth amplification of that isolated channel. This is neither a contraction criterion for the whole locally damped sweep nor a threshold for Algorithm 1.

F.5 Fixed-point-preserving boundary proof

Proof. For $\eta > 0$, $X = (1-\eta)X + \eta F(X)$ iff $X = F(X)$. Triangle inequality gives the bound $q_\eta = 1 - \eta + \eta\rho$. For $0 < \eta < 1$ and $0 < \rho < 1$, $q_\eta - \rho = (1-\eta)(1-\rho) > 0$. Exact equality holds, for example, for a nonnegative affine matrix in the ordinary sup norm: each old row sum s_u becomes $1 - \eta + \eta s_u$. It need not hold for an arbitrary nonlinear or signed map. A signed mode μ becomes $1 - \eta + \eta\mu$, which can reduce oscillation; this does not improve the generic bound. □

The unweighted leaf-injection norm threshold is distinct from spectral instability. In Appendix E.7, the exact linear map has spectral radius γ for every finite b, while its ordinary norm can be arbitrarily

large. It can have transient amplification and still converge. A cyclic nonnegative gain matrix with spectral radius b is an algebraic diagnostic; it is not a realization of that feed-forward leaf example without a separately specified feedback model.

F.6 Spectral KM boundary proof

Proof. Let $M \geq 0$ be the exact error matrix and $M_\eta = (1-\eta)I + \eta M$. If μ is an eigenvalue of M, then $1 - \eta + \eta\mu$ is an eigenvalue of M_η. By Perron–Frobenius for nonnegative matrices (1; 9), M has a nonnegative eigenvector with eigenvalue $\rho(M)$, so $\rho(M_\eta) \geq 1 - \eta + \eta\rho(M)$. Conversely, for any eigenvalue μ of M, $|1 - \eta + \eta\mu| \leq 1 - \eta + \eta|\mu| \leq 1 - \eta + \eta\rho(M)$ because $\eta \in (0, 1]$. Hence $\rho(M_\eta) = 1 - \eta + \eta\rho(M)$. The condition $\rho(M_\eta) < 1$ is therefore equivalent to $\rho(M) < 1$. □

F.7 Surrogate bias proof

Proof. Using $V_s^* = \mathcal{T}^s_{T,\alpha}(V_s^*)$ and $V_u^* = \mathcal{T}^u_{T,\alpha}(V_u^*)$,

$$\begin{aligned}\|V_u^* - V_s^*\|_\infty &\leq \|\mathcal{T}^u_{T,\alpha}(V_u^*) - \mathcal{T}^s_{T,\alpha}(V_u^*)\|_\infty \\ &\quad + \|\mathcal{T}^s_{T,\alpha}(V_u^*) - \mathcal{T}^s_{T,\alpha}(V_s^*)\|_\infty \\ &\leq \epsilon_u + \rho_s\|V_u^* - V_s^*\|_\infty.\end{aligned}$$

Rearranging gives $\|V_u^* - V_s^*\|_\infty \leq \epsilon_u/(1-\rho_s)$. For recovery maps $\mathcal{R}_s, \mathcal{R}_u$, assume $\|\mathcal{R}_s(U) - \mathcal{R}_s(V)\| \leq L_R\|U - V\|_\infty$ and $\sup_{V\in\mathcal{D}} \|\mathcal{R}_u(V) - \mathcal{R}_s(V)\| \leq \epsilon_R$. Then

$$\|\mathcal{R}_u(V_u^*) - \mathcal{R}_s(V_s^*)\| \leq \epsilon_R + L_R\frac{\epsilon_u}{1-\rho_s}.$$

For a shared recovery map, $\epsilon_R = 0$. The local translation theorem concerns equal child inputs; recursively changed child values may not be common shifts after a nonlinear parent aggregation. Consequently local invariance cannot be subtracted as zero global policy error without an additional full-operator argument. This bound controls distortion, not exact equilibrium transfer. □

F.8 Stability–fidelity characterization and impossibility

Lemma F.3 (Softmax equivalence)**.** *For finite vectors p, q and temperatures $\alpha, \beta > 0$,* $\mathrm{softmax}_\alpha(p) = \mathrm{softmax}_\beta(q)$ *iff* $q = (\beta/\alpha)p + b\mathbf{1}$ *for some scalar b.*

Proof. Equality of the policies gives, for every pair i, j, $(p_i - p_j)/\alpha = \log(\pi_i/\pi_j) = (q_i - q_j)/\beta$. Thus $q_i - (\beta/\alpha)p_i$ is independent of i; the converse follows because softmax is invariant to a common shift. □

Proof of Theorem 4.12. Apply Lemma F.3 with equal temperatures to a two-action instance with arbitrary realizable inputs x, y. Since rewards are unchanged, equality of policies implies

$$r_1 - r_2 + \gamma(S(x) - S(y)) = r_1 - r_2 + \gamma(A(x) - A(y)),$$

so $\delta(x) = S(x) - A(x)$ satisfies $\delta(x) = \delta(y)$ for all x, y; hence $S = A + c$. Conversely, such a translation adds $\gamma c\mathbf{1}$ to every local Q vector and leaves softmax unchanged.

Now allow $\alpha_S = \kappa\alpha$. The lemma requires

$$r_1 - r_2 + \gamma(S(x) - S(y)) = \kappa\{r_1 - r_2 + \gamma(A(x) - A(y))\}$$

for every reward difference. Comparing its coefficient forces $\kappa = 1$, after which the translation characterization applies.

Finally suppose $\Delta_\delta = \delta(x) - \delta(y) \neq 0$. Choose $r_1 - r_2 = -\gamma(A(x) - A(y)) - \frac{\gamma}{2}\Delta_\delta$. Then $Q_A(1) - Q_A(2) = -\frac{\gamma}{2}\Delta_\delta$ while $Q_S(1) - Q_S(2) = +\frac{\gamma}{2}\Delta_\delta$, so the unique top action flips. A depth-one deterministic tree realizes this construction, hence a claim over all tree games must include it. □

Algorithm 2 Exact chain specialization of Algorithm 1

Require: Path $(1, \dots, N)$, declared recovery and local models, fixed $\alpha > 0$, $0 < \eta \leq 1$
1: Initialize $V \in \mathcal{D}$
2: **while** not converged **do**
3: Freeze $p_i = \mathcal{R}_{<i}(V)$ for all i
4: **for** $i = N, \dots, 1$ **do**
5: Solve $\widetilde{V}_i = F_i(p_i, \widetilde{V}_{i+1})$ (void child for $i = N$)
6: **end for**
7: $V \leftarrow (1-\eta)V + \eta\widetilde{V}$
8: **end while**

Proof of Theorem 4.13. For $g = S - A$, the reverse triangle inequality and Lipschitzness of S give

$$\begin{aligned} |g(x) - g(y)| &= |(S(x) - S(y)) - (A(x) - A(y))| \\ &\geq |A(x) - A(y)| - |S(x) - S(y)| \\ &\geq |A(x) - A(y)| - L_S d(x, y). \end{aligned}$$

The left-hand side is nonnegative, so taking the positive part and then the supremum over pairs proves $\operatorname{osc}(g) \geq M(A, L_S)$.

For any pair, choose the reward difference exactly as in the final construction of Theorem 4.12. If $\Delta = g(x) - g(y)$, the semantic and surrogate logit differences become $-\gamma\Delta/2$ and $+\gamma\Delta/2$. Their binary Gibbs probabilities are therefore $\sigma(-\gamma\Delta/(2\alpha))$ and $\sigma(+\gamma\Delta/(2\alpha))$, whose total-variation distance is

$$|2\sigma(\gamma|\Delta|/(2\alpha)) - 1| = \tanh\left(\frac{\gamma|\Delta|}{4\alpha}\right).$$

For finite $M > 0$ and every $0 < \varepsilon < M$, a realizable pair whose defining gap exceeds $M - \varepsilon$ yields TV at least $\tanh(\gamma(M-\varepsilon)/(4\alpha))$. Taking suprema and then $\varepsilon \downarrow 0$ proves the claim; $M = 0$ is trivial and $M = \infty$ gives supremum one by a sequence. Compactness, continuity, and pair realizability suffice for attainment of a maximizing pair when applicable; arbitrary metric spaces do not.

For sharpness, take $\mathcal{X} = [0, D]$ with Euclidean distance, $A(x) = L_A x$, and $S(x) = L_S x$, where $0 \leq L_S \leq L_A$. Then $\operatorname{osc}(S - A) = (L_A - L_S)D = M(A, L_S)$; choosing the endpoints in the centered binary construction also attains the policy bound. □

Proof of Corollary 4.14. Adding a constant does not change the minimal Lipschitz modulus, so Theorem 4.12 gives $\operatorname{Lip}_\infty(S) = \operatorname{Lip}_\infty(A) = L_A^\star$ for every universally faithful replacement. If $L_A^\star > \tau_\gamma$, such an S cannot satisfy the sufficient depth-uniform leaf-injection certificate $\operatorname{Lip}_\infty(S) \leq \tau_\gamma$.

For the exact linear-gain chain family, let $A(x) = L_A^\star x$. Because $\tau_\gamma \gamma^{-1/D} \downarrow \tau_\gamma$, some finite D satisfies $L_A^\star \geq \tau_\gamma \gamma^{-1/D}$. Every universally faithful $S = A + c$ has the same gain, and Proposition C.10 makes its depth-D template non-contractive.

For the stated exception, if a fixed game satisfies $Q_S = cQ_A + b\mathbf{1}$ with $c > 0$ and the temperature is changed to $c\alpha$, Lemma F.3 preserves its Gibbs policy. This transforms the full local Q; an aggregation-only replacement with arbitrary unchanged rewards cannot realize $c \neq 1$ universally. □

G Chain SeBIL algorithm and supplementary results

Approximate solvers and independent actor refreshes incur the implementation slack of Theorem 4.18; annealing additionally requires tracking assumptions.

Corollary G.1 (Damped chain operator)**.** *For $0 < \eta \leq 1$, if $\rho_b < 1$, then $\mathcal{T}_\eta(V) = (1-\eta)V + \eta\mathcal{T}(V)$ has contraction upper bound $\rho_\eta = 1 - \eta(1 - \rho_b) < 1$.*

The proof of Corollary G.1 is in Appendix H.7.

Table 3: Target-relative coefficient sensitivity with corrected soft Bellman values. Each row averages 50 paired seeds; brackets are 95% seed-bootstrap intervals. D counts edges. The reference is the same-seed $1\times$ normalized target, so its zero distance is by construction, not evidence of unique semantic correctness. All variants are finite acyclic postorder solves without an ancestor channel.

D	K	γ	Coefficient	Target distance (95% CI)
5	2	0.990	0.75x	329 [326, 331]
5	2	0.990	1x	0 [0, 0]
5	2	0.990	1.25x	637 [633, 641]
5	2	0.990	plain	1.16×10^{12} [1.15×10^{12}, 1.17×10^{12}]
5	2	0.995	0.75x	657 [653, 662]
5	2	0.995	1x	0 [0, 0]
5	2	0.995	1.25×	1.27×10^{3} [1.27×10^{3}, 1.28×10^{3}]
5	2	0.995	plain	7.59×10^{13} [7.54×10^{13}, 7.64×10^{13}]
5	3	0.990	0.75x	328 [327, 330]
5	3	0.990	1x	0 [0, 0]
5	3	0.990	1.25x	637 [635, 639]
5	3	0.990	plain	1.16×10^{12} [1.16×10^{12}, 1.17×10^{12}]
5	3	0.995	0.75x	657 [654, 659]
5	3	0.995	1x	0 [0, 0]
5	3	0.995	1.25×	1.27×10^{3} [1.27×10^{3}, 1.28×10^{3}]
5	3	0.995	plain	7.59×10^{13} [7.58×10^{13}, 7.61×10^{13}]
7	2	0.990	0.75x	533 [530, 535]
7	2	0.990	1x	0 [0, 0]
7	2	0.990	1.25×	1.43×10^{3} [1.43×10^{3}, 1.44×10^{3}]
7	2	0.990	plain	1.14×10^{16} [1.14×10^{16}, 1.15×10^{16}]
7	2	0.995	0.75×	1.07×10^{3} [1.06×10^{3}, 1.07×10^{3}]
7	2	0.995	1x	0 [0, 0]
7	2	0.995	1.25×	2.87×10^{3} [2.86×10^{3}, 2.88×10^{3}]
7	2	0.995	plain	3.01×10^{18} [3.00×10^{18}, 3.02×10^{18}]

Proposition G.2 (Steady-state error under one-step slack)**.** *Let $\mathcal{F}$ be a $\rho < 1$ contraction in a declared norm $\|\cdot\|_*$, with fixed point V^*, and assume all implemented iterates stay in its certified domain. If $\|V^{(k+1)} - \mathcal{F}(V^{(k)})\|_* \le \epsilon_k$ for all k, then $\limsup_{k\to\infty} \|V^{(k)} - V^*\|_* \le (\limsup_{k\to\infty} \epsilon_k)/(1-\rho)$. The same norm is used for the certificate, slack, and conclusion. In particular, a weighted sup-norm certificate with weights $w > 0$ and an ordinary sup-norm slack bound ϵ_∞ implies the converted ordinary bound $\limsup_k \|V^{(k)} - V^*\|_\infty \le (\max_u w_u / \min_u w_u)\epsilon_\infty/(1-\rho)$.*

The proof of Proposition G.2 is in Appendix H.8.

Corollary G.3 (Protocol-wise contraction)**.** *For permutation σ, if $\rho_{b,\sigma} < 1$, then $\mathcal{T}_\sigma$ is contractive with fixed point $V^*(\sigma)$.*

The proof of Corollary G.3 is in Appendix H.6.

G.1 Target-relative coefficient sensitivity

With the environment fixed and no ancestor channel, a complete postorder solve is independent of the input profile. Its target-relative distances measure objective drift. The six cells and ten rules yield 3,000 seed-level rows at $\alpha = 1$; the displayed four-rule subset uses 10,000 seed-bootstrap resamples. Values remain finite, including very large plain-rule distances. Floating-point Bellman residuals and their propagated estimates are numerical diagnostics, not interval-verified error certificates.

G.2 Finite settling despite norm expansion

The held-out study evaluates a hard, old-child recurrence distinct from the soft oracle in (3). It uses four branching families, $\gamma = 0.99$, $\beta \in \{0, 0.1, \ldots, 1\}$, seeds 42–91, four states, three actions, 50

Table 4: Repaired held-out finite recurrence, selected β slices (50 seeds per family and slice). Residual is the current fixed-point residual. Legacy rejects count the withdrawn magnitude rule; they are not divergent runs. The full 44-cell summary contains 2,200 seed-level evaluations.

Family	β	Fixed points	Residual	Max. magnitude	Sweeps	Legacy rejects
2x2x2x2	0.0	50/50	0	55.3	6	0
2x2x2x2	0.5	50/50	0	3.92×10^6	6	50
2x2x2x2	1.0	50/50	0	5.90×10^7	6	50
2x3x2x4	0.0	50/50	0	55.2	6	0
2x3x2x4	0.5	50/50	0	3.54×10^6	6	50
2x3x2x4	1.0	50/50	0	5.31×10^7	6	50
2x4x3x2	0.0	50/50	0	54.9	6	0
2x4x3x2	0.5	50/50	0	3.54×10^6	6	50
2x4x3x2	1.0	50/50	0	5.32×10^7	6	50
3x2x4x2	0.0	50/50	0	55.9	6	0
3x2x4x2	0.5	50/50	0	3.53×10^6	6	50
3x2x4x2	1.0	50/50	0	5.30×10^7	6	50

scalar inner steps, and no transition or ancestor-policy channel. If $S_{50} = \sum_{j=0}^{49} \gamma^j$, its update is

$$V_u^+ = S_{50} \max_a \{R_u(s,a) + \gamma \mathcal{A}((\bar{V}_v)_v)\}, \qquad \bar{V}_v = \frac{1}{|\mathcal{S}|} \sum_{s' \in \mathcal{S}} V_v(s').$$

The child offset is action-independent. The resulting dependency matrix is strictly triangular, hence nilpotent; every node becomes constant after its descendants settle. The runner's depth convention counts five node levels, so six sweeps include the final zero-difference check.

All 2,200 cells reach finite fixed points in six sweeps. Maximum current residual and gap to a bottom-up evaluation of the *same recurrence* are zero in the recorded arithmetic. The largest magnitude is $59{,}033{,}480.37$. Exactly 1,400 cells exceed the old 10^6 value cutoff, reproducing every historical failure label; none exceeds the old 10^{10} step cutoff. Only 200 cells satisfy the actual map's unweighted contraction test. Thus finite settling, unweighted contraction, and cutoff acceptance are different properties. This reference is not an independent game oracle, and the absent ancestor channels cannot validate the earlier certificate.

H Proofs for chain SeBIL (path tree)

H.1 Verification of chain GAD

Verification. Expand $\sum_{i=1}^N A^i_{k,\pi}(s, a_{<i}, a_i) = \sum_{i=1}^N (Q^\pi_k(s, a_{\le i}) - V^\pi_k(s, a_{<i}))$. For $i < N$, Assumption B.3 gives $Q^\pi_k(s, a_{\le i}) = V^\pi_k(s, a_{<i+1})$. With $a_{<1} = \emptyset$ and $a_{\le N} = a$, telescoping yields $Q^\pi_k(s,a) - V^\pi_k(s)$. □

Lemma H.1 (Single-agent soft Bellman contraction)**.** *At fixed $\alpha > 0$, suppose the node action set is finite, or more generally that the entropy-regularized soft value functional is finite and 1-Lipschitz under the sup norm on bounded logits. Then the soft Bellman map on $C_b(\mathcal{H}_i)$ is a γ-contraction in $\|\cdot\|_\infty$.*

The standard argument and references are given in Appendix H.2.

H.2 Citation for single-agent soft Bellman contraction

This is the standard discounted Bellman contraction argument (2; 19), specialized to regularized Bellman operators whose soft value map is non-expansive in sup norm (7). The Gibbs/log-sum-exp form is the usual maximum-entropy control primitive (8; 29); the tree-specific analysis begins when these nodewise contractions are coupled through the continuation channel.

H.3 CHAIN CONTRACTION PROOF

Proof. Write $(\mathcal{T}(V))_i$ for the ith backward-sweep component. Fix V^a, V^b and set $\Delta V = \|V^a - V^b\|$, $\Delta_i = \|(\mathcal{T}(V^a))_i - (\mathcal{T}(V^b))_i\|_\infty$, $\beta_c = \gamma L_c/(1-\gamma)$.

Step 1 ($i = N$). No successors: Assumption C.2 and $\|\theta_a^a - \theta_a^b\|_\infty \le \Delta V$ give $\Delta_N \le C_a \Delta V$.

Step 2 (induction). For $i < N$, if $\Delta_j \le e_j \Delta V$ for all $j > i$, then $\|\theta_c^a - \theta_c^b\|_\infty \le \max_{j>i} \Delta_j \le (\max_{j>i} e_j)\Delta V$. Assumption C.2 and non-expansive ancestor embedding give $\Delta_i \le (C_a + \beta_c \max_{j>i} e_j)\Delta V$. Choose the deterministic majorants $e_i = C_a \sum_{k=0}^{N-i} \beta_c^k$. These are nonincreasing in i, so $\max_{j>i} e_j = e_{i+1}$ and satisfy the required recurrence. This monotonicity is a property of the chosen majorants, not an assumption about actual node errors. Unrolling yields $e_1 \le C_a \frac{1-\beta_c^N}{1-\beta_c}$ (or NC_a if $\beta_c = 1$), i.e. ρ_b in (6). Thus $\|\mathcal{T}(V^a) - \mathcal{T}(V^b)\| \le \rho_b \Delta V$. If $\rho_b < 1$, $\mathcal{T}$ is a ρ_b-contraction. □

H.4 SENSITIVITY-BOUNDS PROOF

Use (2)–(3) on a path. In particular,

$$(\mathcal{B}_i^{p,z} W)(h) = \alpha \log \sum_a \exp\left(\frac{r_i^p(h,a) + \gamma \int [W(h') + G_i(z,x)]\, dK_i^p}{\alpha} \right).$$

The entropy envelope is H_i, not $R_{\max}/(1-\gamma)$ when entropy or child augmentation is present. History regularity is not a substitute for sensitivity to p or z.

Lemma H.2 (Two-channel fixed-point sensitivity)**.** *Suppose the local Bellman map is a γ-contraction on its common invariant complete ball. Let $L_a^{\mathcal{B}}$ and $L_c^{\mathcal{B}}$ be its ancestor- and child-parameter moduli, respectively, uniform over independently held other parameters and local values in that ball. Then*

$$\|F_i(p^a, z^a) - F_i(p^b, z^b)\|_\infty \le \frac{L_a^{\mathcal{B}}}{1-\gamma} \|p^a - p^b\| + \frac{L_c^{\mathcal{B}}}{1-\gamma} \|z^a - z^b\|.$$

Proof. Insert $F_i(p^b, z^a)$, apply the single-channel perturbation bound twice, and add. The independent held-parameter domain ensures the intermediate problem exists. □

Proof of Corollary C.3. The explicit bounds in Lemmas E.1 and E.2 give

$$L_a^{\mathcal{B}} = L_r + \gamma L_K (H_i + \bar{G}_i) + \gamma L_{G,p}, \qquad L_c^{\mathcal{B}} = \gamma L_i$$

when reward and kernel are child-independent. Set $C_a = L_a^{\mathcal{B}}/(1-\gamma)$ and $L_c = L_c^{\mathcal{B}}/\gamma$. Compose C_a with any non-unit predecessor embedding/recovery constant exactly once. If children affect reward or transition, use their additional terms from Lemma E.2. Even for a constant kernel, the direct successor sensitivity need not vanish. □

H.5 REMARK ON VERIFIABLE CONSTANTS

Tighter, game-specific constants only decrease ρ_b in (6); Remark C.5 explains why $\max\{C_a, \beta_c\} < 1$ is not sufficient.

H.6 PROTOCOL CONVERGENCE PROOF

Proof. Re-index predecessor/successor channels by σ and repeat Steps 1–2 of Appendix H.3 with $(C_{a,\sigma}, L_{c,\sigma})$. □

H.7 DAMPED CONVERGENCE PROOF

Proof. For $\mathcal{T}_\eta = (1-\eta)I + \eta\mathcal{T}$ with $\|\mathcal{T}(V^a) - \mathcal{T}(V^b)\| \le \rho_b \|V^a - V^b\|$, $\|\mathcal{T}_\eta(V^a) - \mathcal{T}_\eta(V^b)\| \le (1-\eta)\|V^a - V^b\| + \eta\rho_b \|V^a - V^b\| = (1 - \eta(1-\rho_b))\|V^a - V^b\|$. □

H.8 Robustness proof

Proof. Let $\delta_k = \|V^{(k)} - V^*\|_*$ in the norm of the contraction certificate. Since $V^* = \mathcal{F}(V^*)$,

$$\delta_{k+1} \le \|V^{(k+1)} - \mathcal{F}(V^{(k)})\|_* + \|\mathcal{F}(V^{(k)}) - \mathcal{F}(V^*)\|_* \le \epsilon_k + \rho\delta_k.$$

Let $\bar{\epsilon} = \limsup_{k\to\infty} \epsilon_k$. If $\bar{\epsilon} < \infty$, then for every $\delta > 0$ all sufficiently late errors are at most $\bar{\epsilon} + \delta$. Unrolling from that index gives $\limsup_{k\to\infty} \delta_k \le (\bar{\epsilon} + \delta)/(1-\rho)$; let $\delta \downarrow 0$. The case $\bar{\epsilon} = \infty$ is vacuous. For the converted bound, use $\|x\|_{\infty,w} \le \|x\|_\infty / \min_u w_u$ on the slack and $\|x\|_\infty \le (\max_u w_u)\|x\|_{\infty,w}$ on the limiting error. □

H.9 Local inexact tube proof

Proof. Assume $V^k \in \mathcal{U}_r$. Since $V^* = \mathcal{F}(V^*)$ and $\mathcal{F}$ is a ρ_ℓ-contraction on $\mathcal{U}_r$,

$$\|V^{k+1} - V^*\|_{\infty,w} \le \|\mathcal{F}(V^k) - \mathcal{F}(V^*)\|_{\infty,w} + \|\xi_k\|_{\infty,w} \le \rho_\ell \|V^k - V^*\|_{\infty,w} + \bar{\xi} \le \rho_\ell r + \bar{\xi} \le r.$$

Thus the tube is forward invariant by induction. The same recursion gives

$$\delta_{k+1} \le \rho_\ell \delta_k + \bar{\xi}, \qquad \delta_k = \|V^k - V^*\|_{\infty,w},$$

and unrolling yields $\limsup_k \delta_k \le \bar{\xi}/(1-\rho_\ell)$. For fixed $0 < \eta \le 1$, let $q_\eta = 1 - \eta(1-\rho_\ell)$. Inside-averaging noise gives $\delta_{k+1} \le q_\eta \delta_k + \eta\bar{\xi}$ and invariance $q_\eta r + \eta\bar{\xi} \le r$, hence radius $\bar{\xi}/(1-\rho_\ell)$. Post-averaging noise instead gives $\delta_{k+1} \le q_\eta \delta_k + \bar{\xi}$, condition $q_\eta r + \bar{\xi} \le r$, and radius $\bar{\xi}/[\eta(1-\rho_\ell)]$. The tube is contained in the declared domain; the perturbations must define admissible updates. □

H.10 Inexact outer-loop proof

Proof. Let $\delta_k = \|V^k - V^*\|_{\infty,w}$. Since $V^* = \mathcal{F}(V^*)$ and $\mathcal{F}$ is a ρ-contraction,

$$\delta_{k+1} \le (1-\eta_k)\delta_k + \eta_k \|\mathcal{F}(V^k) - \mathcal{F}(V^*)\|_{\infty,w} + \eta_k \|\xi_k\|_{\infty,w} \le [1 - \eta_k(1-\rho)]\delta_k + \eta_k \|\xi_k\|_{\infty,w}.$$

For constant $\eta_k = \eta$ and $\|\xi_k\|_{\infty,w} \le \bar{\xi}$, unrolling gives

$$\limsup_k \delta_k \le \frac{\eta\bar{\xi}}{1 - [1 - \eta(1-\rho)]} = \frac{\bar{\xi}}{1-\rho}.$$

For bounded errors and variable stepsizes, compare with $R = \bar{\xi}/(1-\rho)$: $(\delta_{k+1} - R)_+ \le [1 - \eta_k(1-\rho)](\delta_k - R)_+$. The product tends to zero because $\sum_k \eta_k = \infty$, proving the same limsup bound. For the vanishing-error case, fix any $\varepsilon > 0$ and choose K so that $\|\xi_k\|_{\infty,w} \le (1-\rho)\varepsilon/2$ for all $k \ge K$. Whenever $\delta_k \ge \varepsilon$, the recursion gives

$$\delta_{k+1} \le \delta_k - \eta_k(1-\rho)\varepsilon/2.$$

Since $\sum_k \eta_k = \infty$, the sequence cannot remain above ε forever. If $\delta_k < \varepsilon$ for some $k \ge K$, then the same recursion gives

$$\delta_{k+1} \le [1 - \eta_k(1-\rho)]\varepsilon + \eta_k(1-\rho)\varepsilon/2 < \varepsilon.$$

Thus, after the first entrance below ε, all later iterates remain below ε. Because $\varepsilon > 0$ was arbitrary, $\delta_k \to 0$. □

H.11 Protocol selection proof

Proof. If $\Sigma_a = \emptyset$, the stated fail-closed rule returns no certified semantic protocol, so there is no selected protocol to analyze. Otherwise let $\sigma^* \in \arg\max_{\sigma\in\Sigma_a} J(\sigma)$. By definition of $\widehat{\sigma}$,

$$\widehat{J}(\widehat{\sigma}) \ge \widehat{J}(\sigma^*).$$

Using the simultaneous validation event twice,

$$J(\widehat{\sigma}) \ge \widehat{J}(\widehat{\sigma}) - \varepsilon_v \ge \widehat{J}(\sigma^*) - \varepsilon_v \ge J(\sigma^*) - 2\varepsilon_v.$$

This is ordinary finite model selection after certificate filtering over a declared protocol set Σ. The proposition does not imply that an admissible protocol exists in Σ, that the certificates are automatically available, or that the topology is learned end-to-end. □

H.12 FIXED-POINT PROFILE AND PROTOCOL-OPTIMALITY PROOF

Lemma H.3 (Soft Bellman fixed point and induced optimality)**.** *For the declared induced MDP with parameters (p, z), let $W = F_u(p, z)$ and $Q = Q_u[W; p, z]$. The unique local maximizer is $\pi = \mathrm{softmax}_\alpha Q$.*

Proof. At every history, $\alpha \log \sum_a e^{Q_a/\alpha} = \max_{\pi \in \Delta}\{\langle \pi, Q\rangle + \alpha H(\pi)\}$. The finite-action Gibbs optimizer is unique. Bellman contraction gives optimality for this discounted induced objective. This statement is about the induced reward and kernel, not an unidentified player's game utility. □

Compatibility required for game transfer. For a claimed ordered-game specialization, the following must be supplied and verified.

1. *Policy closure.* For every node at the candidate fixed point, $\mathcal{R}_u(V^*) = \Psi_u(\mathcal{R}_{\mathrm{Anc}(u)}(V^*), V^*_{\mathrm{Ch}(u)})$. The value equation by itself does not imply this equality.
2. *Reward ownership and entropy.* At each history, the complete local return in (2) must equal the moving player's actual regularized deviation return for every candidate action. Any child message must represent a component of that same player's utility, with the same discount and entropy convention. Another player's utility is not interchangeable. The child augmentation and the local future term must not count the same continuation twice.
3. *Transition and history lift.* The induced kernel must be the exact pushforward of the declared game's transitions and all unobserved actions under the recovered profile. This includes siblings and descendants where they are not observed. All states and histories affected by deviations must be covered, including off-path histories; an auxiliary linear extension must not reveal hidden actions. If the induced model needs extra dependence on descendant policies or state variables, that dependence belongs in its parameters and in B, C.
4. *Ordered deviation principle.* The target game must have the full-order information structure and regularity needed for its one-shot deviation principle (e.g., finite actions, bounded discounted payoffs, and perfect recall with the stated entropy objective). Local objectives must cover that game's deviations, not only on-path comparisons.

These conditions restrict the eligible model class. A chain topology or a cooperative reward label alone does not verify them.

Verification of Proposition 4.16. At a value fixed point, the solved child values equal the corresponding input components. The oracle policies solve their induced MDPs by Lemma H.3. Policy closure is additionally needed to identify the policies used as parameters with those output by the oracle. This proves self-consistent induced-node optimality under the first condition.

If the remaining compatibility conditions have been established for a particular ordered game, induced optimality is the moving player's true regularized optimality at every history. That game's one-shot deviation principle then permits equilibrium transfer (6; 22). Without an explicit embedding satisfying these conditions, no ME-SPNE conclusion is made. Banach uniqueness concerns the value fixed point on $\mathcal{D}$ and, with a deterministic recovery map, its recovered profile; it does not show uniqueness of all equilibria in the game. □

I EXPERIMENTAL DETAILS AND NUMERICAL RESULTS

I.1 ENVIRONMENT SUMMARY

The empirical suite is split by evidentiary role: finite assignment games test local semantic decisions, corrected coefficient and held-out studies audit their explicitly declared numerical operators, OpenSpiel CFR (10) provides information-set-game reference diagnostics, and selected neural traces document wrapper behavior. CFR is not the Tree-SeBIL learning interface.

Coefficient study. The environment is evaluated at $(D, K) \in \{(5,2),(5,3),(7,2)\}$, $\gamma \in \{0.99, 0.995\}$, with seeds $0, \ldots, 49$. The reference is the same-seed scaled-$1\times$ solution; the table compares $0.75\times$, $1\times$, $1.25\times$, and plain aggregation. Soft policy iteration solves the local MDPs, with the full log-sum-exp value. Appendix G.1 gives the numerical and inferential scope.

Evidence provenance and corrections. The earlier core-tabular, coefficient/scaling, and held-out boundary tables and figures are withdrawn. The inspected legacy coefficient generator omitted the log-sum-exp offset; its immutable historical execution identity is unresolved. The present coefficient results supersede those values and explicitly use a chosen $1\times$ reference. The held-out rerun reproduces all 1,400 historical rejection labels through the 10^6 magnitude cutoff while all 2,200 cells settle, withdrawing the former convergence-boundary interpretation. Cyclic damping plots are also excluded as tree validation because no matching Bellman feedback model was established. Neural exclusions comprise mixed-budget MPE aggregates (including endpoints 661760 versus 499200), unresolved HARL estimators and trace mapping, missing exact neural Pursuit post-hoc rows and Ant source records, the three untraceable comprehensive tables, and incomplete compute totals. These exclusions limit reproducibility claims; they are not assertions that every omitted return is incorrect.

Exact multi-pursuer semantic bridge. We generate connected obstacle grids with side length in $\{7, 9, 11\}$, obstacle density in $\{0, 0.15, 0.30\}$, and 2 or 4 pursuers, using 50 independently seeded maps per condition (900 maps total). Each map has three possible evader exits. An action assigns each pursuer to an exit; shortest-path travel determines an interception probability, and the original continuation is the route-probability-weighted probability that at least one assigned pursuer intercepts the evader. We compare semantic union, fixed-target KM, plain average, scaled average, max-child, and independent backups at $\gamma \in \{0.9, 0.99, 0.995\}$. Policies use $\alpha = 0.03$ and immediate travel penalty -0.06 times the travel metric. Semantic continuation enters with coefficient γ; the scaled surrogate uses $(1-\gamma)$ times mean child interception. Capture in the table is the policy-weighted model capture probability, not an observed neural success rate. Every policy is re-evaluated under the semantic regularized objective.

The same maps are reused across methods, discounts, and the balanced control, yielding 32,400 rows rather than that many independent instances. The reported 95% interval is mean ± 1.96 sample standard errors over the 900 map-level residuals; it is a descriptive normal-approximation interval for this equally weighted condition mixture. The 18 condition means range from 0.0909 to 0.1810 in residual and from 40% to 92% in top-assignment flips. The balanced control replaces each action metric by its within-map mean, constructing an action-independent-gap control rather than another map sample. KM uses $\eta = 0.2$ and an already computed semantic target. These are finite local decisions, not learned Pursuit or repeated-game Bellman convergence.

I.2 Matched semantic controls

We use ten new seeds in each of the original 18 map conditions, giving 180 held-out maps with no seed overlap with the 900-map study. Let q_e be the route probability, p_{ie} the interception probability, and a_i an exit assignment. The semantic union is $A(a) = \sum_e q_e[1 - \prod_{i:a_i=e}(1 - p_{ie})]$. To isolate route weighting, compare $n^{-1}\sum_i p_{i,a_i}$ with $n^{-1}\sum_i q_{a_i} p_{i,a_i}$ at the same coefficient. To isolate scaling, compare $r + \gamma A$ with $r + (1-\gamma)A$. All primary comparisons keep the same travel reward, $\gamma = 0.99$, and $\alpha = 0.03$.

Route weighting reduces the unscaled average's mismatch, but scaling the semantic union still changes decisions (Table 5). Thus route omission and gain reduction are distinct sources of mismatch. Translation and full-Q scaling with matched temperature preserve policies to numerical precision. In contrast, scaling only the semantic continuation and its temperature leaves residual 0.0633 when the travel reward is unchanged; setting that reward to zero restores policy agreement.

We also cross $\alpha \in \{0.01, 0.03, 0.1\}$ and travel coefficients $\{0, 0.06, 0.2\}$ with two interception models: independent successes and shared-threshold successes within each route, replacing the union by the largest assigned interception probability. Across these 18 cells, the mean residual of the scaled semantic continuation ranges from 0.0686 to 0.1657. These are sensitivity checks on declared models, not measured capture laws. An independent route/Bernoulli executor draws 4,096 paired outcomes per map under the independent model. The semantic-minus-scaled-plain capture

Table 5: Held-out assignment controls. All rows use the same rewards and temperature. Intervals are 95% percentile intervals from 5,000 map resamples within the 18 design strata; flips compare top assignments with the semantic union. Scaling uses coefficient $1-\gamma$ instead of γ.

Continuation	Semantic residual [95% interval]	Flips (%)
Semantic union	$< 10^{-12}$	0.0
Plain average	$0.1427\ [0.1299, 0.1555]$	78.3
Route-weighted average	$0.0637\ [0.0571, 0.0701]$	63.9
Route-weighted sum	$0.0854\ [0.0744, 0.0969]$	62.2
Scaled plain average	$0.1307\ [0.1215, 0.1402]$	72.2
Scaled route-weighted average	$0.1316\ [0.1224, 0.1412]$	72.2
Scaled semantic union	$0.1277\ [0.1188, 0.1370]$	57.8

gap is 0.1784 (95% stratified map-bootstrap interval $[0.1671, 0.1901]$), compared with enumerated probability gap 0.1779. This validates synthetic one-shot outcomes, not dynamic pursuit performance.

Ancestor-feedback control. To test a nonzero C using actual local Bellman solves, consider two one-state, two-action induced MDPs with $\gamma = 0.8$, $\alpha = 0.2$, and child aggregation $z/8$. Both actions have the same reward: $-\alpha \log 2$ at the root and $0.8(p-1/2) - \alpha \log 2$ at the leaf, where the root recovery probability is $p = \sigma(2gV_r)$, where $\sigma(t) = (1+e^{-t})^{-1}$. The entropy term cancels the reward offset, giving

$$V'_\ell = 4(p - 1/2), \qquad V'_r = V'_\ell/2 = \tanh(gV_r).$$

The box $|V_r| \le 1, |V_\ell| \le 2$ is invariant, and $B_{r\ell} = 1/2, C_{\ell r} = 2|g|$ give $M = \begin{bmatrix} |g| & 0 \\ 2|g| & 0 \end{bmatrix}$. For $g = 0.8$, its unweighted norm is 1.6, whereas weights $(1, 2)$ give modulus 0.8. Removing C would incorrectly predict a constant sweep. Across $g \in \{0, 0.4, 0.8, 0.99, 1.01, 1.2, -1.2\}$, $\eta \in \{1, 0.5, 0.2\}$, and two initial signs (42 cells), iterative local Bellman solves agree with the closed form within 5.7×10^{-14}. At $g = 1.2$, both initial signs settle at nonzero fixed points with $|V_r| = 0.65857$, but recovered and oracle root policies differ by 0.65857 in ℓ_1: value convergence does not imply policy closure. At $g = -1.2$, the undamped map cycles, while $\eta = 0.5$ converges to zero, illustrating the signed-mode exception to the nonnegative exact-gain boundary. Near $g = 0.99$, $\eta = 0.2$ leaves residual 3.2×10^{-5} after 2,000 sweeps; the finite budget is not treated as convergence. This is an induced-MDP control, not an equilibrium embedding of a general-sum game.

I.3 Neural wrapper stress-test details

Simulator stress. The LBF/MPE aggregation study uses three seeds per variant and reports all 12 task/variant combinations with sample standard deviations. Effects differ across tasks: scaled propagation, plain propagation, and independent subtrees do not induce a common return ranking. Logged endpoints are 960, 1440, 1800, and 320 training counters for the four families; these are not established environment-step counts. Normalized AUC divides trapezoidal return area by the observed counter span, without extrapolation to zero. Frozen raw traces support the descriptive summaries, while exact retraining additionally requires the original environment/solver configurations and dependencies. No theoretical gain constant is measured by these returns.

MPE assignment ablation. The nearest/optimal comparison has three seeds per assignment rule under the same wrapper and mean-return convention. Its budget is 200000 logged episodes, with no environment-step conversion. Tail return averages the last 20 of 100 evaluation checkpoints. AUC integrates return against episodes over $[2000, 200000]$, has units of return times episodes, and is displayed divided by 10^6. Both statistics report sample standard deviations across seeds. The comparison describes an assignment intervention within this wrapper; it does not isolate an aggregation mechanism or certify equilibrium.

Budget-limited negative follow-up. A historical one-seed scaled-average neural Pursuit exploration stopped at 4M steps and did not improve its matched-prefix independent-backup comparison.

Table 6: Finite-tree nodewise residual diagnostics for the declared finite objective (mean ± sample SD over five seeds). These are not general-game ME-SPNE gaps. Local gap is the maximum one-node regularized deviation; subtree gap re-optimizes a descendant semantic subtree; reach gap is the induced-reach-weighted subtree residual.

Case	(γ, D)	Local	Subtree	Reach	Policy
affine-invariant	$(0.990, 5)$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$
affine-invariant	$(0.990, 8)$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$
affine-invariant	$(0.995, 5)$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$
affine-invariant	$(0.995, 8)$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$
action-dependent	$(0.990, 5)$	0.148 ± 0.0664	0.205 ± 0.111	0.07 ± 0.0266	0.457 ± 0.117
action-dependent	$(0.990, 8)$	1.56 ± 0.322	2.33 ± 0.667	0.451 ± 0.127	0.832 ± 0.126
action-dependent	$(0.995, 5)$	0.151 ± 0.0679	0.21 ± 0.113	0.0718 ± 0.0272	0.461 ± 0.118
action-dependent	$(0.995, 8)$	1.62 ± 0.339	2.43 ± 0.685	0.469 ± 0.13	0.836 ± 0.125

Table 7: OpenSpiel deterministic equilibrium-oracle sanity checks. Lower is better.

Game	Solver	Iter.	NashConv	Exploitability	Iter. to 10^{-3}
Kuhn poker	CFR	10000	2.27×10^{-4}	1.13×10^{-4}	1500
Kuhn poker	CFR+	10000	1.93×10^{-5}	9.63×10^{-6}	500
Leduc poker	CFR	10000	4.08×10^{-3}	2.04×10^{-3}	–
Leduc poker	CFR+	10000	1.29×10^{-5}	6.46×10^{-6}	1000

Because the exact post-hoc result chain for the broader Pursuit table is unresolved, neither that table nor a numerical neural Pursuit effect is retained. This observation is a limitation of exploratory coverage, not a primary comparative result.

I.3.1 Protocol dependence

Fixed protocols can privilege late movers and change the information supplied to tactical learners. We do not infer a universal ordering advantage from the available neural traces. Proposition C.6 concerns finite model selection after valid certificate filtering and a simultaneous validation bound; the neural results do not establish those premises.

Compute and reproducibility limits. We do not claim complete historical compute accounting. Episode counts, logged training counters, environment steps, and wall time are distinct units. Missing records are left unavailable. Frozen log subsets can support selected descriptive summaries without providing a complete neural retraining environment. The analytic results and independent interception enumeration do not depend on these neural workloads.

I.4 Numerical results

The target-relative coefficient table appears in Appendix G.1, the finite-settling results in Appendix G.2, and the semantic bridge in Table 1.

Finite-tree nodewise controls. The independently implemented finite-tree residual diagnostic uses five seeds, $D \in \{2, 3, 5, 8\}$, $K = 2$, $\gamma \in \{0.9, 0.99, 0.995\}$, $\alpha = 0.2$, and two cases (120 rows). Its local and subtree residuals are defined by that finite declared objective. The common-shift construction is a compatible finite-instance control, not a test of general-game ME-SPNE; the table therefore reports finite-tree nodewise residuals rather than an ME-SPNE gap. The generating soft value retains the maximum-logit offset.

OpenSpiel reference check. CFR and CFR+ use 10000 iterations for both Kuhn and Leduc with evaluation every 500 iterations. The three seed labels denote deterministic repeated runs, not independent stochastic replications. These diagnostics concern their own information-set games.

Table 8: Descriptive simulator tree-protocol comparisons, showing all three variants for each task. Entries are mean ± sample SD across seeds $\{0, 1, 2\}$ (SD divisor $n - 1$). The counter is the final logged training counter, not environment steps. Final and Best are evaluation returns; normalized AUC is trapezoidal return area divided by the observed counter span, without extrapolation to zero. These comparisons do not certify convergence or benchmark dominance.

Env.	Task	Variant	Counter	Final	Best	Norm. AUC
LBF	2p2f diagnostic tree	independent	960	0.437 ± 0.042	0.533 ± 0.011	0.404 ± 0.025
LBF	2p2f diagnostic tree	plain	960	0.432 ± 0.075	0.488 ± 0.038	0.422 ± 0.027
LBF	2p2f diagnostic tree	scaled	960	0.461 ± 0.008	0.515 ± 0.029	0.447 ± 0.027
LBF	4p3f tree	independent	1440	0.231 ± 0.035	0.289 ± 0.013	0.228 ± 0.004
LBF	4p3f tree	plain	1440	0.222 ± 0.028	0.307 ± 0.031	0.223 ± 0.003
LBF	4p3f tree	scaled	1440	0.195 ± 0.011	0.256 ± 0.017	0.205 ± 0.006
MPE	Simple Spread true tree	independent	1800	-105.938 ± 24.232	-69.549 ± 3.642	-80.264 ± 2.114
MPE	Simple Spread true tree	plain	1800	-100.383 ± 16.831	-68.189 ± 2.548	-79.642 ± 2.587
MPE	Simple Spread true tree	scaled	1800	-103.690 ± 23.338	-70.491 ± 1.421	-81.581 ± 3.005
MPE	Simple Tag true tree	independent	320	11.250 ± 9.922	20.000 ± 4.330	9.911 ± 2.555
MPE	Simple Tag true tree	plain	320	17.500 ± 11.456	45.000 ± 22.810	17.768 ± 3.349
MPE	Simple Tag true tree	scaled	320	15.000 ± 9.922	27.500 ± 12.055	14.821 ± 3.273

Table 9: Descriptive MPE Simple Spread assignment ablation at 200000 logged episodes, with the same tree wrapper and mean reward aggregation. Entries are mean ± sample SD across seeds $\{0, 1, 2\}$ (SD divisor $n - 1$). Tail averages the last 20% of evaluation checkpoints (20 of 100). AUC integrates return against logged episodes over the observed interval [2000,200000], without extrapolation to zero; displayed values are divided by 10^6 and have units of return × episodes. No environment-step conversion is assumed.

Assignment	Seeds	Final	Tail 20%	AUC $/10^6$
nearest	3	-17.99 ± 1.97	-18.18 ± 0.30	-3.59 ± 0.05
optimal	3	-12.27 ± 0.03	-12.77 ± 0.25	-2.56 ± 0.02

J COOPERATIVE GAD SPECIAL CASE

In a fully cooperative team game, $Q_k^\pi = Q^\pi$ for all k and Lemma B.4 recovers the HARL-style advantage decomposition (28).