A Unified Framework for Locality in Scalable MARL

Sourav Chakraborty[†] SOURAV.CHAKRABORTY@COLORADO.EDU
Amit Kiran Rege[†*] AMIT.REGE@COLORADO.EDU
Claire Monteleoni[†‡] CLAIRE.MONTELEONI@COLORADO.EDU
Lijun Chen[†] LIJUN.CHEN@COLORADO.EDU
[†] *University of Colorado, Boulder, USA.*
[‡] *INRIA Paris, France.*

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Scalable methods for networked multi-agent reinforcement learning let each agent plan using only a small neighborhood of the agent graph. This works only when the system is value-local, meaning a perturbation at one agent affects the long-run value at another agent weakly when the two are far apart. In the average-reward setting, the standard way to certify locality is the Dobrushin row-sum bound on a single matrix C^π that captures how each agent's next state depends on each other agent's current state. To make this matrix easy to work with, prior work bounds it by a supremum over joint actions. The resulting bound is independent of the policy, but it is loose whenever the policy never picks the worst-case action. We split C^π into pieces that separately track environment sensitivity and policy sensitivity, $C^\pi \preceq E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$, where E^{s} measures how the next state moves with the current state, E^{a} measures how it moves with the current action, and $\Pi(\pi)$ measures how reactive the policy is to changes in state. The spectral radius of $H^\pi := E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$ then controls the decay of the average-reward Poisson solution, and the spectral certificate $\rho(H^\pi) < 1$ is strictly weaker than the row-sum condition $\|H^\pi\|_\infty < 1$ on the same matrix and applies in regimes where policy-independent action-supremum bounds used in prior Dobrushin-style work cannot. For temperature-τ softmax policies we get $\Pi(\pi) \le L/(2\tau)$, so the softmax temperature directly controls locality. We use this decay result to give a deterministic oracle guarantee for a block-coordinate KL-proximal policy-improvement template whose truncation bias decays exponentially in the message-passing radius κ.

Keywords: Multi-agent Reinforcement Learning

1. Introduction

Cooperative multi-agent reinforcement learning on networked systems faces a curse of dimensionality: the joint state and action spaces grow exponentially in the number of agents n, so even when the model factors into local interactions, centralized planning is infeasible (Blondel and Tsitsiklis, 2000; Papadimitriou and Tsitsiklis, 1999). Scalable networked MARL methods address this by letting each agent plan using only a κ-hop neighborhood of the agent graph (Qu et al., 2020, 2019; Lin et al., 2020). Their complexity scales with neighborhood size rather than network size. They are sound, though, only when the system itself is local, meaning that a perturbation at one agent affects the long-run value at a distant agent by an amount that decays exponentially in their graph distance. We call this property *value-locality*. When it holds, κ-hop truncation costs an error that vanishes exponentially in κ.

[*] Equal Contribution with S. Chakraborty.

So the question is when value-locality holds. In the standard γ-discounted setting it is automatic, because the discount factor $\gamma < 1$ multiplies every step of the Bellman backup and produces decay regardless of the agent interaction structure. The average-reward setting ($\gamma = 1$) has no such temporal multiplier. The role of the value function there is played by the *bias function* h^π, which measures, for each starting state s, the long-run advantage of starting in s relative to starting from the stationary distribution of π. It is defined up to an additive constant by the average-reward Poisson equation $h^\pi - T^\pi h^\pi = r^\pi - \bar{r}^\pi$, where T^π is the one-step expectation operator under π and $\bar{r}^\pi$ is the average reward. Whatever decay h^π has must come from the decay of one-step influence under T^π. Existing average-reward guarantees (Qu et al., 2020) get this via a Dobrushin coupling condition. The idea is to build an $n \times n$ matrix C^π whose entry $C^\pi_{j\leftarrow i}$ is the largest change a perturbation of s_i can cause in the next-state distribution of agent j. If every row sum of C^π is below one, then T^π is a contraction in the seminorm $\|\delta(f)\|_\infty = \max_i \sup_{x_{-i}=y_{-i}} |f(x) - f(y)|$, the largest single-coordinate oscillation of f, and iterating the contraction gives exponential decay of h^π in graph distance.

The Dobrushin condition is convenient because the bound on C^π does not depend on the policy: one takes a supremum over joint actions when measuring how much an action change moves the next state. That is also why the condition is conservative. Consider two agents in which agent 2's next state copies agent 1's last action, regardless of state. The worst-case action move is maximal, so the supremum says the system is globally coupled. But if agent 1's policy barely reacts to its own state, then flipping s_1 only weakly moves the distribution of a_1, which only weakly moves the next-state marginal of agent 2. In closed loop, s_1 and s_2' are nearly independent, so the system is value-local. The Dobrushin test cannot see this because it threw the policy away.

To recover that information, we decompose the policy-induced one-step interdependence matrix into an environment piece and a policy piece. Using the total variation distance $\mathrm{TV}(\mu, \nu) = \frac{1}{2}\sum_x |\mu(x) - \nu(x)|$, define $E^{\mathrm{s}}_{j\leftarrow i}$ as the worst-case TV between $P_j(\cdot \mid s, a)$ and $P_j(\cdot \mid s', a)$ over pairs (s, s') that differ only on coordinate i, with the action fixed. Define $E^{\mathrm{a}}_{j\leftarrow i}$ the same way for an action change with the state fixed. Define $\Pi_{k\leftarrow i}(\pi)$ as the TV-sensitivity of agent k's action distribution to a change in s_i. Of these, E^{s} and E^{a} depend only on the environment; $\Pi(\pi)$ depends on the policy. Our first result (Proposition 1) is the entrywise bound

$$C^\pi \preceq E^{\mathrm{s}} + E^{\mathrm{a}}\,\Pi(\pi).$$

The product $E^{\mathrm{a}}\Pi(\pi)$ makes the cancellation visible. The action channel E^{a} can be large, but if the policy is smooth ($\Pi(\pi)$ small) the closed-loop influence is small anyway. Write $H^\pi := E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$. Our second result (Theorem 3) is that whenever the spectral radius $\rho(H^\pi)$ (the largest eigenvalue magnitude of H^π) is below one, the Poisson solution has $\delta(h^\pi) \le (I - (H^\pi)^\top)^{-1}\delta(r^\pi)$, where the right-hand side is the matrix-geometric (Neumann) series $\sum_{t\ge 0}((H^\pi)^\top)^t\delta(r^\pi)$. Our certificate is strictly weaker than prior ones in two senses. First, on the same comparison matrix, replacing the row-sum condition by the spectral-radius condition is strictly weaker since $\rho(M) \le \|M\|_\infty$ for any nonnegative M. Second, H^π is policy-dependent and can be much smaller than the policy-independent action-supremum bounds used in prior Dobrushin-style guarantees, so the resulting certificate applies in regimes where those prior conditions are silent.

The policy factor $\Pi(\pi)$ is something a learning algorithm has control over. For temperature-τ softmax policies (Geist et al., 2019; Haarnoja et al., 2018), the policy class used in entropy-regularized control and KL-proximal updates, Lemma 6 gives $\Pi_{k\leftarrow i}(\pi) \le \min\{1,\, L_{k\leftarrow i}/(2\tau)\}$, where $L_{k\leftarrow i}$ is the coordinatewise Lipschitz constant of the logit. Raising τ therefore directly tightens the certificate

on H^π, and Section 4 works out the tradeoff between certifying locality (higher τ) and approaching the unregularized optimum (lower τ).

Section 5 uses the decay result to give a deterministic oracle guarantee for a block-coordinate KL-proximal improvement template. The same Neumann tail $\lambda^{\kappa+1}$ appears in two places: in the locally computable certificate an agent uses to decide whether κ hops are enough, and in the per-step improvement bound. The certificate is genuinely local. The truncated Poisson surrogate used in the analysis is an oracle object in the general model; turning it into something locally computable needs additional structure on observation scopes or function approximation, which we leave to follow-up work. We summarize related work next; the appendix has more.

Related work. Exponential decay of value on networked MDPs was first studied in the scalable networked MARL line of work (Qu et al., 2020, 2019; Lin et al., 2021, 2020), which proves discounted and average-reward locality results under a graph-local transition assumption $P_i(s_i' \mid s_{N_i}, a_i)$ and a Dobrushin row-sum bound on C^π. Our setting allows P_i to depend on the full (s, a) and assumes no graph upfront, since we derive a graph from the support of H^π after the fact. Older work on factored MDPs (Kearns and Koller, 1999; Guestrin et al., 2003) and weakly coupled MDPs (Meuleau et al., 1998) uses different forms of locality, typically asking for local rewards or independent transitions. The policy-dependent angle here parallels work that ties decay rates to policy regularization in single-agent control, but to our knowledge gives the first policy-dependent spectral certificate of value-locality for average-reward networked MARL. Function-approximation MARL methods (Zhang et al., 2018; Lowe et al., 2017) and independent learners (Tan, 1993; Matignon et al., 2012) attack scalability differently and do not certify locality. Appendix A expands on this.

2. Setup and preliminaries

Consider a system of n agents. Each agent $i \in [n]$ has a finite state space $\mathcal{S}_i$ and a finite action space $\mathcal{A}_i$. The joint spaces are $\mathcal{S} = \prod_i \mathcal{S}_i$ and $\mathcal{A} = \prod_i \mathcal{A}_i$, with elements written $s = (s_1, \dots, s_n)$ and $a = (a_1, \dots, a_n)$, and s_{-i}, a_{-i} for the profiles that exclude agent i.

At time t the joint state is s^t, a joint action $a^t \sim \pi(\cdot \mid s^t)$ is drawn, and the next state $s^{t+1} \sim P(\cdot \mid s^t, a^t)$. We assume the kernel factors as $P(s' \mid s, a) = \prod_{i=1}^n P_i(s_i' \mid s, a)$, so that given (s, a) the next-state coordinates are conditionally independent across agents. Each P_i is still allowed to depend on the full (s, a). This is more permissive than the typical networked-MARL assumption $P_i(s_i' \mid s_{N_i}, a_i)$ used in (Qu et al., 2020; Lin et al., 2021), which fixes a graph N_i ahead of time. We do not, since we want the locality structure to come out of the analysis rather than the modeling assumptions.

Policies are of product form $\pi(a \mid s) = \prod_{i=1}^n \pi_i(a_i \mid s_{O_i})$, where the observation scope $O_i \subseteq [n]$ is the set of state coordinates agent i's policy actually depends on. Taking $O_i = [n]$ recovers globally conditioned product policies; smaller O_i models agents that observe only locally. The kernel induced by π is

$$P^\pi(s' \mid s) = \sum_{a \in \mathcal{A}} \Big(\prod_{i=1}^n P_i(s_i' \mid s, a) \Big) \Big(\prod_{j=1}^n \pi_j(a_j \mid s_{O_j}) \Big).$$

Average reward and the Poisson equation. For the average-reward analysis we restrict to policies π under which P^π is irreducible on $\mathcal{S}$, so that it has a unique stationary distribution d^π. Given a per-step reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, write $r^\pi(s) = \sum_a r(s, a) \prod_k \pi_k(a_k \mid s_{O_k})$ for the expected one-step reward in state s and $\bar{r}^\pi = \sum_s d^\pi(s) r^\pi(s)$ for the average reward under π. The object

that plays the role of the value function in this setting is the *bias function* h^π, which measures the long-run advantage of starting in s relative to starting from stationarity and is defined up to a constant by the Poisson equation

$$h^\pi - T^\pi h^\pi \;=\; r^\pi - \bar{r}^\pi, \qquad T^\pi f(s) := \mathop{\mathbb{E}}_{S' \sim P^\pi(\cdot|s)}[f(S')].$$

This h^π is used by policy-gradient and policy-improvement updates the same way the discounted value function is. When we say the system is value-local, we mean exactly that the coordinatewise oscillations of h^π decay quickly in graph distance. Discounted-setting corollaries are in the appendix.

Coordinatewise oscillations. The natural way to quantify how much an agent's reward or value depends on every other agent is the coordinatewise oscillation. For a bounded $f : \mathcal{S} \to \mathbb{R}$, the i-oscillation

$$\delta_i(f) \;=\; \sup_{x,y \in \mathcal{S}:\, x_{-i} = y_{-i}} |f(x) - f(y)|$$

is the largest change f can undergo when only coordinate i changes; equivalently, it is the Lipschitz constant of f under a single-coordinate Hamming change. Write $\delta(f) = (\delta_i(f))_{i=1}^n$ for the vector of oscillations and $\|\delta(f)\|_\infty = \max_i \delta_i(f)$ for its maximum. The seminorm $\|\delta(\cdot)\|_\infty$ is the standard object in Dobrushin-type arguments for interacting particle systems and Glauber dynamics (Dobruschin, 1968; Martinelli, 1999); basic properties are in Appendix E.3. Spatial truncation works precisely when the entries of $\delta(h^\pi)$ are small at agents far from any source of reward variation.

Three sensitivity matrices. We measure one-step coupling with three $n \times n$ nonnegative matrices. The first two depend only on the environment; the third depends on the policy.

$$E^{\mathrm{s}}_{j \leftarrow i} \;=\; \sup_{\substack{s,s' \in \mathcal{S}:\, s_{-i} = s'_{-i} \\ a \in \mathcal{A}}} \mathrm{TV}\big(P_j(\cdot \mid s, a),\, P_j(\cdot \mid s', a)\big), \qquad \text{(state channel)}$$

$$E^{\mathrm{a}}_{j \leftarrow i} \;=\; \sup_{\substack{s \in \mathcal{S} \\ a,a' \in \mathcal{A}:\, a_{-i} = a'_{-i}}} \mathrm{TV}\big(P_j(\cdot \mid s, a),\, P_j(\cdot \mid s, a')\big), \qquad \text{(action channel)}$$

$$\Pi_{j \leftarrow i}(\pi) \;=\; \sup_{s,s':\, s_{-i} = s'_{-i}} \mathrm{TV}\big(\pi_j(\cdot \mid s_{O_j}),\, \pi_j(\cdot \mid s'_{O_j})\big). \qquad \text{(policy reactivity)}$$

In words, $E^{\mathrm{s}}_{j \leftarrow i}$ is the largest jump in j's next-state law that a change of s_i can cause with the action held fixed; $E^{\mathrm{a}}_{j \leftarrow i}$ is the corresponding quantity for an action change at i with the state held fixed; and $\Pi_{j \leftarrow i}(\pi)$ is the largest jump in j's action distribution that a change of s_i can cause. The last is zero whenever $i \notin O_j$. The policy-induced closed-loop influence of i on the next-state marginal of j is

$$C^\pi_{j \leftarrow i} = \sup_{s,s':\, s_{-i} = s'_{-i}} \mathrm{TV}\big(P^\pi_j(\cdot \mid s), P^\pi_j(\cdot \mid s')\big), \qquad P^\pi_j(\cdot \mid s) = \sum_a P_j(\cdot \mid s, a) \prod_k \pi_k(a_k \mid s_{O_k}),$$

and C^π is the matrix that prior work bounds by Dobrushin row sums. The next section decomposes it into the environment and policy pieces and replaces the row-sum bound with a spectral one.

3. Policy-induced influence and locality

This section gives the decomposition of C^π that powers everything else, the spectral condition that controls how the Poisson solution decays, and the resulting locality of the average-reward bias function. Proofs are in the appendix.

A perturbation of s_i reaches the next-state marginal of j along two routes in one step. The first is direct through the environment: even with the action fixed, a change in s_i can move $P_j(\cdot \mid s, a)$. The second is indirect, through the policy and back into the environment, since changing s_i shifts the action distribution of each agent k, and a change in a_k can in turn move j's next-state law. The direct route is bounded by $E^{\mathrm{s}}_{j\leftarrow i}$. The indirect route factors as $\sum_k E^{\mathrm{a}}_{j\leftarrow k}\Pi_{k\leftarrow i}(\pi)$, picking up the action-channel weight $E^{\mathrm{a}}_{j\leftarrow k}$ from the policy-action-state hop and the policy weight $\Pi_{k\leftarrow i}(\pi)$ from the state-policy hop.

Proposition 1 (Decomposition of policy-induced influence) *For any product policy π and any factorized synchronous dynamics on a finite state space, $C^\pi \preceq E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$ entrywise. Equivalently, $C^\pi_{j\leftarrow i} \le E^{\mathrm{s}}_{j\leftarrow i} + \sum_k E^{\mathrm{a}}_{j\leftarrow k}\Pi_{k\leftarrow i}(\pi)$ for every i, j.*

The proof inserts an intermediate distribution that uses the new policy weights but the old kernel. It then bounds the two halves separately, once by a TV-convexity step (which gives the E^{s} term) and once by a maximal coupling on the actions (which gives the $E^{\mathrm{a}}\Pi$ term). Full details are in Appendix B.1.

A two-agent example. Take $n = 2$ binary agents, $\mathcal{S}_i = \mathcal{A}_i = \{0, 1\}$. Let P_1 be constant (agent 1's next state is independent of everything), and let agent 2's next state copy agent 1's action: $P_2(s_2' = 1 \mid s, a) = \mathbf{1}\{a_1 = 1\}$. Then $E^{\mathrm{s}} = 0$ and $E^{\mathrm{a}}_{2\leftarrow 1} = 1$ with all other entries of E^{a} equal to zero. An action-supremum bound stops here and reports the system as globally coupled, since $E^{\mathrm{a}}_{2\leftarrow 1} = 1$ is as large as it can be. Now suppose agent 1's policy changes by at most α in TV when s_1 flips, and that π_1 does not observe s_2 (so $\Pi_{1\leftarrow 2} = 0$). Then $\Pi_{1\leftarrow 1}(\pi) = \alpha$, and Proposition 1 gives

$$C^\pi_{2\leftarrow 1} \;\le\; E^{\mathrm{s}}_{2\leftarrow 1} + E^{\mathrm{a}}_{2\leftarrow 1}\Pi_{1\leftarrow 1} + E^{\mathrm{a}}_{2\leftarrow 2}\Pi_{2\leftarrow 1} \;=\; \alpha.$$

The closed-loop coupling is as small as the policy. The same bound handles the failure mode: if π_1 is sharp in s_1 (α near 1), it returns α and recovers the action-supremum answer. Smoothing the policy only helps when there is policy reactivity to smooth out. Appendix E.1 works through a complementary instance.

We now convert the entrywise decomposition into a one-step contraction.

Lemma 2 (Oscillation bound via H^π) *Let $T^\pi f(s) = \mathbb{E}[f(S_{t+1}) \mid S_t = s]$ under the synchronous update with product policy and factorized kernel. Then $\delta_i(T^\pi f) \le \sum_j H^\pi_{j\leftarrow i}\delta_j(f)$ for every i and every bounded f, where $H^\pi := E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$. In vector form, $\delta(T^\pi f) \le (H^\pi)^\top \delta(f)$.*

Iterating Lemma 2 turns a spectral bound on H^π into global decay of T^π, which in turn bounds the Poisson solution.

Theorem 3 (Policy-uniform contraction and Poisson decay) *Let Π_{pol} be a compact class of product policies, set $H^\pi := E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$ and $\lambda_\star := \sup_{\pi\in\Pi_{\mathrm{pol}}} \rho(H^\pi)$, and assume $\lambda_\star < 1$. For every $\pi \in \Pi_{\mathrm{pol}}$, every bounded f, and every $t \ge 0$,*

$$\delta\big((T^\pi)^t f\big) \le \big((H^\pi)^\top\big)^t \delta(f),$$

and for every $\bar\lambda \in (\lambda_\star, 1)$ there exists $C_{\bar\lambda} \ge 1$ independent of t, π such that $\|\delta((T^\pi)^t f)\|_\infty \le C_{\bar\lambda}\bar\lambda^t \|\delta(f)\|_\infty$. If P^π is irreducible for every $\pi \in \Pi_{\mathrm{pol}}$, with stationary distribution d^π, then the

Poisson equation $h^\pi - T^\pi h^\pi = r^\pi - \bar{r}^\pi$ has a solution unique up to an additive constant, and every solution satisfies

$$\delta(h^\pi) \le \sum_{t=0}^{\infty} \left((H^\pi)^\top\right)^t \delta(r^\pi) \le \left(I - (H^\pi)^\top\right)^{-1} \delta(r^\pi).$$

Theorem 3 does not assume any agent graph upfront. A graph does emerge, though, through the directed support graph G_H^π of H^π, in which $i \to j$ is an edge whenever $H^\pi_{j\leftarrow i} > 0$ (agent i's state directly affects agent j's next state in one step under π). The matrix power $((H^\pi)^\top)^t$ then has a clean path interpretation: $((H^\pi)^\top)^t_{ij}$ is a path-weighted sum, where each contributing path runs from i to j in exactly t edges of G_H^π and has weight equal to the product of H^π entries along it. Three consequences follow.

- *Path accumulation.* $\delta_i(h^\pi)$ is bounded by a sum over all directed paths in G_H^π starting at i and ending at any agent j where the reward varies, weighted by H^π along the path.
- *Exponential attenuation.* The contribution of paths of length t is $O(\bar{\lambda}^t)$, so agents far downstream of i in G_H^π matter exponentially less.
- *Truncation error.* Ignoring agents more than κ hops downstream of i drops the tail of the Neumann series, which is bounded by $\frac{C}{1-\bar{\lambda}} \bar{\lambda}^{\kappa+1} \|\delta(r^\pi)\|_\infty$.

If $E^{\mathrm{s}}, E^{\mathrm{a}}, \Pi(\pi)$ are each local with respect to some underlying graph G, then G_H^π is a finite-radius closure of G. It need not equal G, because $E^{\mathrm{a}}\Pi(\pi)$ can create new closed-loop edges. An action of agent k that affects j, combined with a policy at k that reacts to s_i, produces a closed-loop edge $i \to j$ even when i and j are not graph-neighbors in G.

Theorem 3 recovers and strengthens the locality results of (Qu et al., 2020; Lin et al., 2021): the spectral-radius condition is strictly weaker than the row-sum one on any single comparison matrix, and the policy-dependent matrix H^π can be much smaller than the policy-independent action-supremum bounds those works use, so $\rho(H^\pi) < 1$ can hold in regimes where their row-sum conditions fail. The two-agent example above is one such case; Appendix B.4 works out the formal embedding. The reward enters Theorem 3 only through $\delta(r^\pi)$, so a reward that decomposes locally produces a sparse $\delta(r^\pi)$ and tightens the bound automatically; no separate local-reward assumption is needed.

A small numerical instance of the gap. The gap between $\rho(H^\pi)$ and $\|H^\pi\|_\infty$ can be large even on small matrices, and it is easy to build closed-loop matrices that our condition certifies but the Dobrushin row-sum bound does not. For $n = 2$ agents, take

$$H^\pi = \begin{pmatrix} 0.6 & 0 \\ 0.6 & 0.6 \end{pmatrix}.$$

With our convention $H^\pi_{j\leftarrow i} = H^\pi[j, i]$, this models a system in which agent 1 influences agent 2's next state ($H^\pi_{2\leftarrow 1} = 0.6$) but agent 2 does not influence agent 1 ($H^\pi_{1\leftarrow 2} = 0$). The row sums are 0.6 and 1.2, so $\|H^\pi\|_\infty = 1.2$ and the row-sum condition is violated. The eigenvalues are both 0.6, so $\rho(H^\pi) = 0.6$ and Theorem 3 certifies locality with decay rate arbitrarily close to 0.6. A κ-hop truncation costs error $O(0.6^{\kappa+1})$, so $\kappa = 10$ already drives the truncation error below 10^{-2} while the row-sum diagnostic offers no guarantee at any κ. The asymmetry is what creates the gap: closed-loop influence flows one way ($1 \to 2$) and decays in one hop in the reverse direction, even though one row sum is above one.

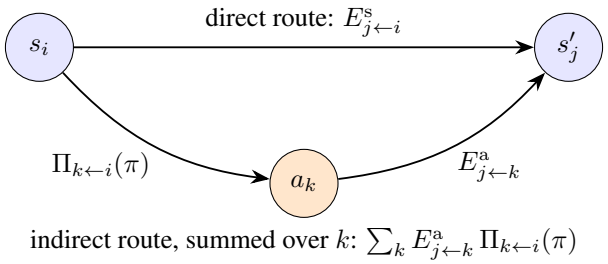

Figure 1: The decomposition $C^{\pi}_{j\leftarrow i} \le E^{\mathrm{s}}_{j\leftarrow i} + \sum_k E^{\mathrm{a}}_{j\leftarrow k}\Pi_{k\leftarrow i}(\pi)$, written as two routes from a perturbation of s_i to the next-state marginal of j. The direct route through the environment has weight $E^{\mathrm{s}}_{j\leftarrow i}$. The indirect route passes through each intermediate action a_k, picking up a factor $\Pi_{k\leftarrow i}(\pi)$ for the state-to-action hop and $E^{\mathrm{a}}_{j\leftarrow k}$ for the action-to-next-state hop, then sums over k.

Discounted setting. For the standard γ-discounted Bellman operator T^{π}_{γ}, the one-step bound becomes $\delta(T^{\pi}_{\gamma} f) \le \gamma (H^{\pi})^{\top}\delta(f)$, so locality is certified whenever $\gamma\rho(H^{\pi}) < 1$. The decay rate is set jointly by the temporal factor γ and the structural factor $\rho(H^{\pi})$. With $\gamma = 0.99$ and $\rho(H^{\pi}) = 0.5$, for instance, taking $\tilde{\lambda} = 0.5$ as a certificate gives a neighborhood of order 7 to reach error 10^{-2}, against roughly 458 from a γ-only bound. The same analysis carries over to the asynchronous Glauber-style update; see Appendix B.6.

4. Controlling policy sensitivity via softmax temperature

The decay rate in Theorem 3 is $\rho(H^{\pi}) = \rho(E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi))$. The environment terms $E^{\mathrm{s}}, E^{\mathrm{a}}$ are fixed once the MDP is given. The policy term $\Pi(\pi)$ is set by the policy class, so it is what a learning algorithm can change. A natural way to keep $\Pi(\pi)$ small is to restrict the policy class to state-to-action maps that are smooth. The temperature-τ softmax class (Geist et al., 2019; Haarnoja et al., 2018) is the standard such class, and it is the policy class used in entropy-regularized control, soft actor-critic, and KL-proximal updates, where τ is a hyperparameter of the algorithm. We show in this section that the same τ that controls the reward-versus-entropy tradeoff in those methods also bounds $\Pi(\pi)$ entrywise, and so bounds the locality certificate of Theorem 3. All proofs are in Appendix C.

Setup. The entropy-regularized average-reward objective is

$$J_{\tau}(\pi) \;=\; \mathop{\mathbb{E}}_{s\sim d^{\pi},\; a\sim\pi(\cdot|s)}\Big[r(s,a) - \tau\sum_{k=1}^{n}\log\pi_k(a_k \mid s_{O_k})\Big],$$

the average-reward analogue of the discounted soft-Bellman objective. We study the temperature-τ softmax policy class commonly used in entropy-regularized and KL-proximal methods.

Definition 4 (Softmax policy) *A policy π is a temperature-τ softmax policy if for each agent k there is a logit function $g_k : \mathcal{S}_{O_k}\times\mathcal{A}_k \to \mathbb{R}$ such that $\pi_k(a_k \mid s_{O_k}) \propto \exp(g_k(s_{O_k}, a_k)/\tau)$.*

How reactive a softmax policy is to its state input is controlled by τ together with how reactive the logit itself is.

Definition 5 (Logit Lipschitz constant) *For a logit function g_k, the one-coordinate logit constant with respect to state i is* $L_{k\leftarrow i} := \sup_{s_{-i}=s'_{-i}} \|g_k(s_{O_k}, \cdot) - g_k(s'_{O_k}, \cdot)\|_\infty$.

$L_{k\leftarrow i}$ is a property of the logit parameterization alone and does not involve τ. For a linear logit $g_k(s,a) = \langle w_{k,a}, \phi(s)\rangle$ with features ϕ of single-coordinate oscillation $\delta_i(\phi)$, one can take $L_{k\leftarrow i} \le \sup_a 2\|w_{k,a}\|\delta_i(\phi)$, so L is set entirely by the feature design and the weight magnitudes and can be precomputed.

Lemma 6 (Softmax temperature controls $\Pi(\pi)$) *For a temperature-τ softmax policy, $\Pi_{k\leftarrow i}(\pi) \le \min\{1,\, L_{k\leftarrow i}/(2\tau)\}$ for all $k, i \in [n]$.*

The proof bounds the TV distance between two softmax distributions by a sigmoidal function of the difference of their logits, then linearizes at zero with Lipschitz constant $1/(2\tau)$. The cap at 1 is the trivial TV bound. Writing $L = [L_{k\leftarrow i}]$ for the matrix of logit constants and applying the lemma entrywise gives $\Pi(\pi) \preceq \min\{\mathbf{1}, L/(2\tau)\}$. The spectral radius $\rho(M)$ of a nonnegative matrix M is entrywise monotone (Perron–Frobenius), so an entrywise upper bound on H^π gives an upper bound on $\rho(H^\pi)$:

$$H^\pi \preceq E^{\mathrm{s}} + E^{\mathrm{a}}\frac{L}{2\tau} \quad\Longrightarrow\quad \rho(H^\pi) \le \rho\left(E^{\mathrm{s}} + E^{\mathrm{a}}\frac{L}{2\tau}\right).$$

The right-hand side is the quantity a practitioner actually computes. It is built from the (known) environment matrices $E^{\mathrm{s}}, E^{\mathrm{a}}$ and the (designed) logit constants L, and its spectral radius is monotone nonincreasing in τ.

Locality–optimality tradeoff. The behavior at the two ends of the temperature range is easy to describe. As $\tau \to \infty$, the softmax policy approaches uniform on every state, $\Pi(\pi) \to 0$, and $H^\pi \to E^{\mathrm{s}}$. Whenever $\rho(E^{\mathrm{s}}) < 1$, a high enough τ certifies locality regardless of how strong the action channel is. The cost is that the policy is far from greedy and the objective J_τ deviates from the unregularized average reward by $O(\tau \log|\mathcal{A}|)$. As $\tau \to 0$ the policy becomes deterministic, $\Pi(\pi)$ saturates near 1 wherever the logit ordering is sensitive to the state, and $\rho(H^\pi)$ can cross one and break the certificate.

In an algorithmic loop this gives a clean recipe. Pick a minimum temperature $\tau_{\min}$ at which $\rho(E^{\mathrm{s}} + E^{\mathrm{a}}L/(2\tau_{\min})) < 1$ holds with margin, run the entropy-regularized improvement at $\tau = \tau_{\min}$ with a truncation radius κ chosen to push the bias below the desired accuracy, and either stop there (accepting the regularization bias) or anneal τ downward in stages, increasing κ at each stage to compensate. Appendix E.2 works through a quantitative instance.

5. An oracle framework for localized evaluation and block-coordinate improvement

We now use the decay result of Section 3 to give a deterministic oracle guarantee for a block-coordinate KL-proximal improvement template. The section deliberately separates the structural effect of locality from the orthogonal issues of statistical estimation and function approximation, since the latter can be addressed by standard tools once the structural picture is clear. The only object in our framework that is guaranteed to be locally computable is the *certificate* below; the truncated Poisson surrogate is an oracle quantity used to bound bias.

Recall the support graph G_H^π of $H^\pi = E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$, with edge $i \to j$ whenever $H^\pi_{j\leftarrow i} > 0$. When H^π is sparse, the κ-hop neighborhoods in G_H^π are small.

Phase 1: locality certificate. Set $b^\pi = \delta(r^\pi)$. Theorem 3 gives $\delta(h^\pi) \le \sum_{t\ge 0}((H^\pi)^\top)^t b^\pi$. Truncating this Neumann series at depth κ yields the computable certificate

$$\widehat{\delta}^{(\kappa)} := \sum_{t=0}^{\kappa} \big((H^\pi)^\top\big)^t b^\pi.$$

The i-th component $\widehat{\delta}_i^{(\kappa)}$ is a sum over directed paths in G_H^π of length at most κ starting at i (weighted by H^π entries along the path) and accumulating b_j^π at each endpoint j. It depends only on the κ-hop *out-ball* of i in G_H^π (the agents reachable from i in at most κ directed edges) and is computable in κ rounds of message passing along the reverse edges of that ball. For the bias analysis we also use the truncated Poisson surrogate $\widehat{h}_\kappa^\pi := \sum_{t=0}^{\kappa}(T^\pi)^t(r^\pi - \bar{r}^\pi)$.

Theorem 7 (Localized certificate and truncation bias) *Let π be a product policy with P^π irreducible, and suppose $\|((H^\pi)^\top)^t\|_{\infty\to\infty} \le C\lambda^t$ for some $C \ge 1$, $\lambda \in (0,1)$. Let $R^{(\kappa)} = \sum_{t>\kappa}((H^\pi)^\top)^t b^\pi$. Every solution h^π of the Poisson equation satisfies $\delta(h^\pi) \le \widehat{\delta}^{(\kappa)} + R^{(\kappa)}$ with $\|R^{(\kappa)}\|_\infty \le \frac{C}{1-\lambda}\lambda^{\kappa+1}\|b^\pi\|_\infty$, and the oracle bias of $\widehat{h}_\kappa^\pi$ satisfies*

$$B_\kappa^\pi := \inf_{c\in\mathbb{R}} \|\widehat{h}_\kappa^\pi - h^\pi - c\mathbf{1}\|_\infty \;\le\; \tfrac{1}{2}\,\|R^{(\kappa)}\|_1 \;\le\; \tfrac{nC}{2(1-\lambda)}\lambda^{\kappa+1}\|b^\pi\|_\infty.$$

The two bounds play different roles. $\|R^{(\kappa)}\|_\infty$ is the residual error of the certificate the agent actually computes. B_κ^π is the oracle-side bias of the truncated Poisson surrogate, which appears only in the proof of the improvement step below. Both decay at the same rate $\lambda^{\kappa+1}$.

Phase 2: block KL-prox improvement. We use the surrogate $\widehat{h}_\kappa^\pi$ to drive a one-agent policy update. The KL-proximal (or mirror-descent) update is a standard step in entropy-regularized policy gradient, NPG, and TRPO-style methods: it picks a new policy that maximizes the predicted improvement (a linearized advantage) minus η times the KL divergence from the old policy, where the temperature $\eta > 0$ controls how aggressive the step is. We apply this update one agent at a time. This per-step η is separate from the softmax temperature τ of Section 4, which is a property of the policy class. Fix a baseline product policy π and an agent k. The exact advantage and its truncated counterpart are

$$A^\pi(s,a) = r(s,a) - \bar{r}^\pi + \sum_y P(y \mid s,a)h^\pi(y) - h^\pi(s), \qquad \widehat{A}_\kappa^\pi = A^\pi\big|_{h^\pi\to\widehat{h}_\kappa^\pi}.$$

The exact and approximate block logits for agent k are

$$g_{k,\star}^\pi(s,a_k) = \mathop{\mathbb{E}}_{a_{-k}\sim\pi_{-k}(\cdot|s)}\big[A^\pi(s,(a_k,a_{-k}))\big], \qquad \widehat{g}_{k,\kappa}^\pi(s,a_k) = \mathop{\mathbb{E}}_{a_{-k}\sim\pi_{-k}(\cdot|s)}\big[\widehat{A}_\kappa^\pi(s,(a_k,a_{-k}))\big],$$

and the KL-prox update is $\mu_k(\cdot \mid s) \propto \pi_k(\cdot \mid s)\exp(\widehat{g}_{k,\kappa}^\pi(s,\cdot)/\eta)$ with $\mu_{-k} = \pi_{-k}$.

Theorem 8 (One-block oracle improvement with logit error) *Under the assumptions of Theorem 7, and assuming additionally that π has full action support at every state so that P^μ inherits irreducibility from P^π (the KL-prox update preserves the action support of π), the update μ above satisfies*

$$\bar{r}(\mu) - \bar{r}(\pi) \;\ge\; \eta \mathop{\mathbb{E}}_{S\sim d^\mu}\Big[\mathrm{KL}\big(\mu_k(\cdot \mid S) \,\big\|\, \pi_k(\cdot \mid S)\big)\Big] \;-\; \frac{2nC}{1-\lambda}\lambda^{\kappa+1}\|b^\pi\|_\infty.$$

The same Neumann tail $\lambda^{\kappa+1}$ appears in the certificate of Theorem 7 and in the improvement bound, so a single truncation radius κ controls both. The bound is additive in the KL-prox step (first term) and the truncation bias (second term), so the proximal temperature η and the radius κ can be tuned independently. Algorithm 1 interleaves the two phases over outer iterations and inner agents, recomputing H and b at every step. Using stale (H, b) over a full pass is also valid at the cost of a staleness term we omit.

What is local, what is oracle. The framework separates two layers of approximation that scalable MARL methods often conflate. The certificate $\widehat{\delta}^{(\kappa)}$ is genuinely local: each $\widehat{\delta}_i^{(\kappa)}$ depends only on entries of H^π and of b^π supported on the κ-hop out-ball of agent i in G_H^π. The surrogate $\widehat{h}_\kappa^\pi$ and the truncated logit $\widehat{g}_{k,\kappa}^\pi$, by contrast, are oracle objects, since in the general model they depend on the full reward vector and on global action sums. The improvement bound treats them as accessible because the analysis only compares the truncated logit to the exact one. Making the surrogate locally computable needs either (i) restricted observation scopes O_i that turn r^π and the action expectations into κ-local sums, or (ii) function approximation of $\widehat{h}_\kappa^\pi$ on a parametric family with locality structure, in the spirit of the localized critic in (Qu et al., 2020; Lin et al., 2021). The analysis here gives the structural part of the bound; an additional approximation-error term then enters additively.

Algorithm 1 Oracle block-coordinate improvement with locality certificates

Input: baseline $\pi^{(0)}$, radius κ, prox temperature η
for $\ell = 0, 1, \dots$ **do**
  $\pi^{(\ell,0)} \leftarrow \pi^{(\ell)}$
  for $k = 1, \dots, n$ **do**
    compute $b^{(\ell,k-1)} = \delta(r^{\pi^{(\ell,k-1)}})$ and $H^{(\ell,k-1)} = E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi^{(\ell,k-1)})$
    certificate $\widehat{\delta}^{(\kappa,\ell,k-1)} = \sum_{t=0}^{\kappa} \big((H^{(\ell,k-1)})^\top\big)^t b^{(\ell,k-1)}$ (κ rounds message-pass)
    oracle surrogate $\widehat{h}_\kappa^{(\ell,k-1)} = \sum_{t=0}^{\kappa} (T^{\pi^{(\ell,k-1)}})^t (r^{\pi^{(\ell,k-1)}} - \bar{r}^{\pi^{(\ell,k-1)}})$
    block logit $\widehat{g}_{k,\kappa}^{(\ell,k-1)}$ via Phase 2 with $h^\pi \to \widehat{h}_\kappa^{(\ell,k-1)}$
    update $\pi_k^{(\ell,k)}(\cdot \mid s) \propto \pi_k^{(\ell,k-1)}(\cdot \mid s) \exp(\widehat{g}_{k,\kappa}^{(\ell,k-1)}(s,\cdot)/\eta)$; $\pi_j^{(\ell,k)} = \pi_j^{(\ell,k-1)}$ for $j \neq k$
  end for
  $\pi^{(\ell+1)} \leftarrow \pi^{(\ell,n)}$
end for

6. Conclusion

We gave a policy-dependent spectral certificate of value-locality, $\rho(H^\pi) < 1$ for $H^\pi = E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$, that is strictly weaker than the row-sum condition on the same matrix and that applies in regimes where the policy-independent action-supremum bounds used in prior Dobrushin-style work cannot. The decomposition splits the closed-loop influence into an environment piece $(E^{\mathrm{s}}, E^{\mathrm{a}})$ and a policy piece $(\Pi(\pi))$. The split makes precise the intuition that a smooth policy can neutralize a strong action channel, since $E^{\mathrm{a}}\Pi(\pi)$ shrinks linearly in policy smoothness even when E^{a} is large. For temperature-τ softmax policies the bound $\Pi(\pi) \leq L/(2\tau)$ makes τ a direct way to tighten the certificate, and the same Neumann tail $\lambda^{\kappa+1}$ that governs the decay of the bias function also governs the truncation bias of an oracle block-coordinate KL-prox improvement template, whose decentralized realization is the natural next step.

References

Vincent D Blondel and John N Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.

P. L. Dobruschin. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability and Its Applications*, 13:197–224, 1968. URL https://api.semanticscholar.org/CorpusID:122528571.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:59523693.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and Andreas Krause, editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 1856–1865. PMLR, 2018. URL http://dblp.uni-trier.de/db/conf/icml/icml2018.html#HaarnojaZAL18.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored MDPs. In *IJCAI*, volume 16, pages 740–747, 1999.

Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. Distributed reinforcement learning in multi-agent networked systems. *arXiv preprint arXiv:2006.06555*, 2020.

Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. Multi-agent reinforcement learning in stochastic networked systems. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7825–7837. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/412604be30f701b1b1e3124c252065e6-Paper.pdf.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.

Fabio Martinelli. *Lectures on Glauber Dynamics for Discrete Spin Models*, pages 93–191. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999. ISBN 978-3-540-48115-7. doi: 10.1007/978-3-540-48115-7_2. URL https://doi.org/10.1007/978-3-540-48115-7_2.

Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas L Dean, and Craig Boutilier. Solving very large weakly coupled Markov decision processes. In *AAAI/IAAI*, pages 165–172, 1998.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. *arXiv preprint arXiv:1912.02906*, 2019.

Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable multi-agent reinforcement learning for networked systems with average reward. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2074–2086. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/168efc366c449fab9c2843e9b54e2a18-Paper.pdf.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5872–5881. PMLR, 2018.

Appendix A. Related work

Our work addresses the curse of dimensionality in multi-agent reinforcement learning (MARL). The challenge is that the global state and action spaces ($\mathcal{S} = \prod_i \mathcal{S}_i$, $\mathcal{A} = \prod_i \mathcal{A}_i$) grow exponentially with the number of agents, n. This problem falls into the category of "succinctly described" MDPs (Blondel and Tsitsiklis, 2000), which are known to be computationally intractable in the general case, even with structural assumptions (Papadimitriou and Tsitsiklis, 1999).

To achieve scalability, researchers have explored various structural assumptions. One common approach in general MARL is to use function approximation (Zhang et al., 2018; Lowe et al., 2017) or assume independent learners (Tan, 1993), though the latter can suffer from non-stationarity (Matignon et al., 2012). Other related areas include *Factored MDPs* (Kearns and Koller, 1999; Guestrin et al., 2003), which assume local states but typically a global action, and *Weakly Coupled MDPs* (Meuleau et al., 1998), which assume agents' transitions are independent and coupling only occurs through the reward. Our work differs from these by focusing on systems with local, coupled transitions, which is common in networked systems.

This paper is most related to the line of work on scalable Networked MARL (Qu et al., 2020, 2019; Lin et al., 2021). These foundational papers were the first to show that if the system's local transitions and rewards depend only on a local graph neighborhood, the system exhibits an Exponential Decay Property (EDP). This property, where influence decays exponentially with graph distance, justifies κ-hop truncation and enables scalable algorithms.

In the average-reward setting these works establish the EDP via a Dobrushin coupling condition (Qu et al., 2020), which bounds environment influence by a supremum over joint actions and asks $\|C\|_\infty < 1$ for the resulting policy-independent matrix C. This bound is tight only when the policy realizes the worst-case action; it fails to certify locality whenever the action channel is strong but the policy in use does not excite it. The present paper replaces this row-sum condition with the spectral condition $\rho(H^\pi) < 1$ for $H^\pi = E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$, which is policy-dependent, recovers the prior condition as a special case, and certifies locality in regimes where the prior one cannot.

Appendix B. Proofs from Section 3

This appendix contains complete proofs of the results in the main text and an extension to asynchronous updates. All total variation distances are normalized: for probability measures μ and ν on a finite set, $\mathrm{TV}(\mu,\nu) = \frac{1}{2}\sum_x |\mu(x) - \nu(x)|$. For a bounded function f on a product space, the coordinatewise oscillations are $\delta_i(f) = \sup_{x,y\in\mathcal{S},\, x_{-i}=y_{-i}} |f(x) - f(y)|$ and $\delta(f)$ is the vector with entries $\delta_i(f)$.

B.1. Proof of Proposition 1

Given $i,\ j \in [n]$ and two states s, s' that agree off coordinate i, define

$$\mu_j = \sum_{a\in\mathcal{A}} \Big(\prod_k \pi_k(a_k \mid s_{O_k})\Big) P_j(\cdot \mid s, a), \qquad \nu_j = \sum_{a\in\mathcal{A}} \Big(\prod_k \pi_k(a_k \mid s'_{O_k})\Big) P_j(\cdot \mid s', a).$$

We claim

$$\mathrm{TV}(\mu_j, \nu_j) \ \le\ E^{\mathrm{s}}_{j\leftarrow i} \ + \ \sum_k E^{\mathrm{a}}_{j\leftarrow k}\, \Pi_{k\leftarrow i}(\pi).$$

The decomposition uses the triangle inequality and a coupling argument.

Define an intermediate distribution

$$\tilde{\mu}_j = \sum_{a \in \mathcal{A}} \Big(\prod_k \pi_k(a_k \mid s'_{O_k}) \Big) P_j(\cdot \mid s, a).$$

By the triangle inequality,

$$\mathrm{TV}(\mu_j, \nu_j) \;\le\; \mathrm{TV}(\mu_j, \tilde{\mu}_j) \;+\; \mathrm{TV}(\tilde{\mu}_j, \nu_j).$$

We bound $\mathrm{TV}(\tilde{\mu}_j, \nu_j)$ first. The measures $\tilde{\mu}_j$ and ν_j are mixtures with the same mixing weights $w(a) = \prod_k \pi_k(a_k \mid s'_{O_k})$. By convexity of total variation,

$$\mathrm{TV}(\tilde{\mu}_j, \nu_j) \le \sum_{a \in \mathcal{A}} w(a)\, \mathrm{TV}\big(P_j(\cdot \mid s, a), P_j(\cdot \mid s', a)\big) \le \sup_{a \in \mathcal{A}} \mathrm{TV}\big(P_j(\cdot \mid s, a), P_j(\cdot \mid s', a)\big) \le E^{\mathrm{s}}_{j \leftarrow i}.$$

We now bound $\mathrm{TV}(\mu_j, \tilde{\mu}_j)$. This is the difference between two mixtures with the same components and different mixing weights $p(a \mid s) = \prod_k \pi_k(a_k \mid s_{O_k})$ and $p(a \mid s') = \prod_k \pi_k(a_k \mid s'_{O_k})$, but the components $P_j(\cdot \mid s, a)$ can be far apart for different a. We upper bound this difference by a coupling that changes actions along one coordinate at a time and uses the action sensitivity of the kernel.

Let (A, A') be a coupling of $p(\cdot \mid s)$ and $p(\cdot \mid s')$ constructed as follows. For each agent k, couple the marginals $\pi_k(\cdot \mid s_{O_k})$ and $\pi_k(\cdot \mid s'_{O_k})$ by a maximal coupling so that

$$\mathbb{P}[A_k \neq A'_k] \;=\; \mathrm{TV}\big(\pi_k(\cdot \mid s_{O_k}), \pi_k(\cdot \mid s'_{O_k})\big) \;\le\; \Pi_{k \leftarrow i}(\pi).$$

Take these couplings independent across k, which is possible because the policy is of product form and the coordinates are independent under each product marginal.

Conditional on (A, A'), couple next–state coordinates for agent j by a maximal coupling of $P_j(\cdot \mid s, A)$ and $P_j(\cdot \mid s, A')$. Then

$$\mathbb{P}[S'^{(j)} \neq \tilde{S}'^{(j)} \mid A, A'] \;=\; \mathrm{TV}\big(P_j(\cdot \mid s, A), P_j(\cdot \mid s, A')\big).$$

Taking expectation,

$$\mathrm{TV}(\mu_j, \tilde{\mu}_j) \;\le\; \mathbb{E}\, \mathrm{TV}\big(P_j(\cdot \mid s, A), P_j(\cdot \mid s, A')\big).$$

For any a, a' differing on a set $D \subseteq [n]$ of action coordinates, the triangle inequality and the definition of E^{a} imply

$$\mathrm{TV}\big(P_j(\cdot \mid s, a), P_j(\cdot \mid s, a')\big) \;\le\; \sum_{k \in D} E^{\mathrm{a}}_{j \leftarrow k}.$$

Applying this bound with the random pair (A, A') and taking expectations yields

$$\mathrm{TV}(\mu_j, \tilde{\mu}_j) \;\le\; \sum_{k=1}^{n} E^{\mathrm{a}}_{j \leftarrow k}\, \mathbb{P}[A_k \neq A'_k] \;\le\; \sum_{k=1}^{n} E^{\mathrm{a}}_{j \leftarrow k}\, \Pi_{k \leftarrow i}(\pi),$$

where the last inequality uses $\mathbb{P}[A_k \neq A'_k] \le \Pi_{k \leftarrow i}(\pi)$ from the maximal coupling above. Combining the two pieces finishes the proof:

$$\mathrm{TV}(\mu_j, \nu_j) \;\le\; E^{\mathrm{s}}_{j \leftarrow i} \;+\; \sum_{k=1}^{n} E^{\mathrm{a}}_{j \leftarrow k}\, \Pi_{k \leftarrow i}(\pi).$$

By the definition of $C^{\pi}_{j \leftarrow i}$ as a supremum over s, s' with $s_{-i} = s'_{-i}$, the same bound holds for $C^{\pi}_{j \leftarrow i}$.

B.2. Proof of Lemma 2

Fix i and two states s, s' with $s_{-i} = s'_{-i}$. We will couple one synchronous step from s and from s'.

First couple the actions. For each $k \in [n]$, let (A_k, A'_k) be a maximal coupling of $\pi_k(\cdot \mid s_{O_k})$ and $\pi_k(\cdot \mid s'_{O_k})$, chosen independently across k. Then $(A, A') = ((A_k)_k, (A'_k)_k)$ is a coupling of the product measures $\pi(\cdot \mid s)$ and $\pi(\cdot \mid s')$, and

$$\mathbb{P}[A_k \neq A'_k] \;=\; \mathrm{TV}\big(\pi_k(\cdot \mid s_{O_k}), \pi_k(\cdot \mid s'_{O_k})\big) \;\leq\; \Pi(\pi)_{k \leftarrow i}.$$

Next, conditional on $(A, A') = (a, a')$, draw the next state coordinates independently across j: let $X_j \sim P_j(\cdot \mid s, a)$ and $Y_j \sim P_j(\cdot \mid s', a')$, and couple (X_j, Y_j) by a maximal coupling of these two marginals. By construction,

$$\mathbb{P}[X_j \neq Y_j \mid A = a, A' = a'] \;=\; \mathrm{TV}\big(P_j(\cdot \mid s, a), P_j(\cdot \mid s', a')\big).$$

Taking expectation over (A, A') gives

$$\mathbb{P}[X_j \neq Y_j] \;=\; \mathbb{E}_{A,A'}\Big[\mathrm{TV}\big(P_j(\cdot \mid s, A), P_j(\cdot \mid s', A')\big)\Big].$$

For any a, a', apply the triangle inequality:

$$\mathrm{TV}\big(P_j(\cdot \mid s, a), P_j(\cdot \mid s', a')\big) \;\leq\; \underbrace{\mathrm{TV}\big(P_j(\cdot \mid s, a), P_j(\cdot \mid s, a')\big)}_{\text{action change at fixed } s} + \underbrace{\mathrm{TV}\big(P_j(\cdot \mid s, a'), P_j(\cdot \mid s', a')\big)}_{\text{state change at fixed } a'}.$$

By the definitions of E^{a} and E^{s},

$$\mathrm{TV}\big(P_j(\cdot \mid s, a), P_j(\cdot \mid s, a')\big) \;\leq\; \sum_{k:\, a_k \neq a'_k} E^{\mathrm{a}}_{j \leftarrow k}, \qquad \mathrm{TV}\big(P_j(\cdot \mid s, a'), P_j(\cdot \mid s', a')\big) \;\leq\; E^{\mathrm{s}}_{j \leftarrow i}.$$

Therefore,

$$\mathbb{P}[X_j \neq Y_j] \;\leq\; E^{\mathrm{s}}_{j \leftarrow i} \;+\; \sum_{k=1}^{n} E^{\mathrm{a}}_{j \leftarrow k}\, \mathbb{P}[A_k \neq A'_k] \;\leq\; E^{\mathrm{s}}_{j \leftarrow i} \;+\; \sum_{k=1}^{n} E^{\mathrm{a}}_{j \leftarrow k}\, \Pi(\pi)_{k \leftarrow i}.$$

We now relate $|f(X) - f(Y)|$ to the coordinate-disagreement indicators. Define a sequence $Z_0 = X,\ Z_1, \ldots, Z_n = Y$ by changing coordinates one at a time, with Z_j obtained from Z_{j-1} by replacing its j-th coordinate with Y_j. Then Z_{j-1} and Z_j agree off coordinate j, so by definition of $\delta_j(f)$,

$$|f(Z_{j-1}) - f(Z_j)| \;\leq\; \delta_j(f)\, \mathbf{1}\{X_j \neq Y_j\}.$$

Summing along the chain and using the triangle inequality,

$$|f(X) - f(Y)| \;\leq\; \sum_{j=1}^{n} \delta_j(f)\, \mathbf{1}\{X_j \neq Y_j\}.$$

Taking expectations,

$$|(T^\pi f)(s) - (T^\pi f)(s')| \;\leq\; \sum_{j=1}^{n} \delta_j(f)\, \mathbb{P}[X_j \neq Y_j] \;\leq\; \sum_{j=1}^{n} \delta_j(f)\, \Big(E^{\mathrm{s}}_{j \leftarrow i} + \sum_{k=1}^{n} E^{\mathrm{a}}_{j \leftarrow k}\, \Pi(\pi)_{k \leftarrow i}\Big).$$

Taking the supremum over all s, s' that agree off i yields $\delta_i(T^\pi f) \le \sum_j H^\pi_{j\leftarrow i}\,\delta_j(f)$, and the vector form follows.

The multi-step bound follows by induction. The base case $t = 0$ is trivial. For the inductive step, applying the one-step bound to $g = (T^\pi)^t f$ gives $\delta(T^\pi g) \le (H^\pi)^\top \delta(g) \le (H^\pi)^\top ((H^\pi)^\top)^t \delta(f) = ((H^\pi)^\top)^{t+1}\delta(f)$, using the inductive hypothesis $\delta(g) \le ((H^\pi)^\top)^t\delta(f)$ and entrywise nonnegativity of $(H^\pi)^\top$.

B.3. Proof of Theorem 3

We note that the map $\pi \mapsto H^\pi = E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)$ is continuous on the finite-state finite-action setting: $\Pi(\pi)$ is a finite maximum of total variations between π-marginals, which are continuous in π. Therefore $\rho(H^\pi)$ is upper semicontinuous on π, and the supremum $\lambda_\star = \sup_{\pi\in\Pi_{\mathrm{pol}}} \rho(H^\pi)$ is attained on the compact class Π_{pol}.

By Lemma 2, for every $\pi \in \Pi_{\mathrm{pol}}$ and every bounded f,

$$\delta(T^\pi f) \;\le\; (H^\pi)^\top \delta(f).$$

Iterating gives

$$\delta\big((T^\pi)^t f\big) \;\le\; \big((H^\pi)^\top\big)^t \delta(f), \qquad t \ge 0.$$

Fix any $\bar\lambda \in (\lambda_\star, 1)$, where

$$\lambda_\star = \sup_{\pi\in\Pi_{\mathrm{pol}}} \rho(H^\pi) < 1.$$

For each $\pi \in \Pi_{\mathrm{pol}}$, define

$$w^\pi := (\bar\lambda I - (H^\pi)^\top)^{-1}\mathbf{1} \in \mathbb{R}^n_{++}.$$

This is well-defined because $\rho((H^\pi)^\top) = \rho(H^\pi) < \bar\lambda$. Moreover,

$$(H^\pi)^\top w^\pi = \bar\lambda w^\pi - \mathbf{1} \;\le\; \bar\lambda w^\pi.$$

Define the weighted sup norm on $\mathbb{R}^n$ by

$$\|x\|_{w^\pi,\infty} := \max_{i\in[n]} \frac{|x_i|}{w^\pi_i}.$$

We claim that for every nonnegative vector $x \in \mathbb{R}^n_+$,

$$\|(H^\pi)^\top x\|_{w^\pi,\infty} \;\le\; \bar\lambda\, \|x\|_{w^\pi,\infty}.$$

Indeed, for each i,

$$\big((H^\pi)^\top x\big)_i = \sum_j H^\pi_{j\leftarrow i}\, x_j = \sum_j H^\pi_{j\leftarrow i}\, w^\pi_j \cdot \frac{x_j}{w^\pi_j} \le \|x\|_{w^\pi,\infty} \sum_j H^\pi_{j\leftarrow i}\, w^\pi_j = \|x\|_{w^\pi,\infty} \big((H^\pi)^\top w^\pi\big)_i \le \|x\|_{w^\pi,\infty}\, \bar\lambda\, w^\pi_i,$$

where the last step uses $(H^\pi)^\top w^\pi \le \bar\lambda w^\pi$ from above. Dividing by w^π_i and taking the max over i gives the claim. By induction,

$$\big\|\big((H^\pi)^\top\big)^t x\big\|_{w^\pi,\infty} \;\le\; \bar\lambda^t\, \|x\|_{w^\pi,\infty}, \qquad t \ge 0.$$

Because the state and action spaces are finite, the map $\pi \mapsto H^\pi$ is continuous. Hence $\pi \mapsto w^\pi$ is continuous on the compact set Π_{pol}, so

$$m_{\bar{\lambda}} \coloneqq \inf_{\pi \in \Pi_{\mathrm{pol}}} \min_i w_i^\pi > 0, \qquad M_{\bar{\lambda}} \coloneqq \sup_{\pi \in \Pi_{\mathrm{pol}}} \max_i w_i^\pi < \infty.$$

Therefore, for every nonnegative vector x,

$$\|x\|_{w^\pi,\infty} \le \frac{1}{m_{\bar{\lambda}}} \|x\|_\infty, \qquad \|x\|_\infty \le M_{\bar{\lambda}} \|x\|_{w^\pi,\infty},$$

and thus

$$\left\| \left((H^\pi)^\top \right)^t x \right\|_\infty \;\le\; \frac{M_{\bar{\lambda}}}{m_{\bar{\lambda}}} \, \bar{\lambda}^t \, \|x\|_\infty.$$

Applying this with $x = \delta(f)$ yields

$$\|\delta((T^\pi)^t f)\|_\infty \;\le\; C_{\bar{\lambda},\Pi_{\mathrm{pol}}} \, \bar{\lambda}^t \, \|\delta(f)\|_\infty, \qquad C_{\bar{\lambda},\Pi_{\mathrm{pol}}} \coloneqq \frac{M_{\bar{\lambda}}}{m_{\bar{\lambda}}}.$$

Now assume in addition that, for each $\pi \in \Pi_{\mathrm{pol}}$, the Markov chain with kernel P^π is irreducible, and let d^π be its stationary distribution. Define

$$g^\pi \coloneqq r^\pi - \bar{r}^\pi \mathbf{1}, \qquad \bar{r}^\pi = \sum_{s \in \mathcal{S}} d^\pi(s) \, r^\pi(s).$$

Since $\delta(g^\pi) = \delta(r^\pi)$, the bound above implies

$$\sum_{t=0}^{\infty} \|\delta((T^\pi)^t g^\pi)\|_\infty < \infty.$$

To handle the additive-constant ambiguity, work on the quotient space

$$\mathcal{B}_0(\mathcal{S}) \coloneqq B(\mathcal{S}) / \mathrm{span}\{\mathbf{1}\},$$

and write $[f]$ for the equivalence class of f. Define

$$\|[f]\|_\delta \coloneqq \|\delta(f)\|_\infty.$$

Because $\delta(f) = 0$ if and only if f is constant, this is a well-defined norm on $\mathcal{B}_0(\mathcal{S})$. The operator T^π induces a linear map

$$\widetilde{T}^\pi[f] \coloneqq [T^\pi f]$$

satisfying

$$\|(\widetilde{T}^\pi)^t [f]\|_\delta \;\le\; C_{\bar{\lambda},\Pi_{\mathrm{pol}}} \, \bar{\lambda}^t \, \|[f]\|_\delta.$$

Hence the Neumann series converges in operator norm on $\mathcal{B}_0(\mathcal{S})$, and we may define

$$[h^\pi] \;\coloneqq\; \sum_{t=0}^{\infty} (\widetilde{T}^\pi)^t [g^\pi].$$

Then

$$(I - \widetilde{T}^{\pi})[h^{\pi}] = [g^{\pi}].$$

Equivalently, for any representative h^{π} of the class $[h^{\pi}]$, there exists a constant c such that

$$h^{\pi} - T^{\pi}h^{\pi} = g^{\pi} + c\mathbf{1}.$$

Applying the stationary distribution d^{π} to both sides gives

$$0 = d^{\pi}(g^{\pi}) + c = c,$$

since $d^{\pi}(g^{\pi}) = 0$ by definition of $\bar{r}^{\pi}$. Therefore

$$h^{\pi} - T^{\pi}h^{\pi} = g^{\pi} = r^{\pi} - \bar{r}^{\pi}.$$

Uniqueness up to an additive constant follows similarly: if $h - T^{\pi}h = 0$, then

$$(I - \widetilde{T}^{\pi})[h] = 0.$$

Since $I - \widetilde{T}^{\pi}$ is invertible on the quotient space, $[h] = 0$, so h is constant.

Finally, using the triangle inequality for the oscillation seminorm and the multi-step bound,

$$\delta(h^{\pi}) \;\le\; \sum_{t=0}^{\infty} \delta((T^{\pi})^{t} g^{\pi}) \;\le\; \sum_{t=0}^{\infty} \big((H^{\pi})^{\top}\big)^{t} \delta(g^{\pi}) \;=\; \sum_{t=0}^{\infty} \big((H^{\pi})^{\top}\big)^{t} \delta(r^{\pi}).$$

Since $(H^{\pi})^{\top}$ is entrywise nonnegative and $\rho(H^{\pi}) < 1$, the Neumann series converges entrywise and

$$\sum_{t=0}^{\infty} \big((H^{\pi})^{\top}\big)^{t} = \big(I - (H^{\pi})^{\top}\big)^{-1}.$$

Hence

$$\delta(h^{\pi}) \;\le\; \big(I - (H^{\pi})^{\top}\big)^{-1} \delta(r^{\pi}).$$

This completes the proof.

B.4. Spatial decay as a corollary of sparsity

Suppose there is an underlying undirected graph G on $[n]$ such that the environment and the policy are local with respect to G in the following sense:

- $E^{\mathrm{s}}_{j \leftarrow i} = 0$ unless i lies in a fixed-radius neighborhood of j in G;
- $E^{\mathrm{a}}_{j \leftarrow k} = 0$ unless k lies in a fixed-radius neighborhood of j in G;
- $\Pi(\pi)_{k \leftarrow i} = 0$ unless $i \in O_k$, and each observation scope O_k is contained in a fixed-radius neighborhood of k in G.

In general, the product $E^{\mathrm{a}}\Pi(\pi)$ need not have exactly the same sparsity pattern as G; rather, it induces a derived directed support graph

$$G_H^\pi : \qquad i \to j \text{ whenever } H^\pi_{j\leftarrow i} > 0, \qquad H^\pi = E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi).$$

This graph captures one-step closed-loop influence. Under the locality assumptions above, G_H^π is a sparse finite-radius closure of the underlying graph.

For any pair of coordinates i, j, the entry

$$\Big[\big((H^\pi)^\top\big)^t\Big]_{ij}$$

can be nonzero only if there is a directed path of length at most t from i to j in G_H^π, where edges are oriented as $i \to j$ whenever $H^\pi_{j\leftarrow i} > 0$. Hence the Neumann-series bound

$$\delta(h^\pi) \;\le\; \sum_{t=0}^{\infty} \big((H^\pi)^\top\big)^t \, \delta(r^\pi)$$

shows that the contribution of coordinates outside a κ-hop neighborhood in G_H^π is controlled by the tail

$$\sum_{t=\kappa+1}^{\infty} \big((H^\pi)^\top\big)^t \, \delta(r^\pi),$$

which decays exponentially whenever $\rho(H^\pi) < 1$.

B.5. Average-reward localized certificates and oracle truncation (synchronous)

Under the sparsity conditions above, the truncated certificate

$$\widehat{\delta}^{(\kappa)} = \sum_{t=0}^{\kappa} \big((H^\pi)^\top\big)^t \delta(r^\pi)$$

from Theorem 7 is computable by local message passing on κ-neighborhoods in the support graph of H^π. The associated oracle truncated Poisson series $\widehat{h}^\pi_\kappa$ has bias modulo constants bounded by the same exponentially decaying Neumann tail. Using this oracle surrogate in a block KL-prox update yields the one-block improvement guarantee of Theorem 8. Establishing a fully local policy-improvement algorithm requires an additional approximation or projection step beyond the present structural analysis and is left to future work.

B.6. Asynchronous updates

For completeness we state the analogue of the main results when, at each step, all agents keep their states except for a randomly selected coordinate J_t which is updated according to a site–selection distribution ν with full support.

In this model the one–step operator is $K^\pi f(s) = \mathbb{E}_\pi[f(S_{t+1}) \mid S_t = s]$ with

$$\mathbb{P}[S_{t+1} = s^{(j\to y)} \mid S_t = s] = \nu_j \sum_a \Big(\prod_k \pi_k(a_k \mid s)\Big) P_j(y \mid s, a) \quad \text{for } y \in \mathcal{S}_j.$$

Define E^{s}, E^{a}, and $\Pi(\pi)$ as before and set

$$M^\pi = (I - \operatorname{diag}\nu) + \operatorname{diag}\nu\,(E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi)).$$

Then for all bounded f,

$$\delta(K^\pi f) \le (M^\pi)^\top \delta(f).$$

Consequently, if

$$\lambda_\star^{\mathrm{async}} := \sup_{\pi\in\Pi_{\mathrm{pol}}} \rho(M^\pi) < 1,$$

then for every $\bar\lambda \in (\lambda_\star^{\mathrm{async}}, 1)$ there exists a constant $C_{\bar\lambda}$ such that

$$\|\delta((K^\pi)^t f)\|_\infty \le C_{\bar\lambda}\bar\lambda^t \|\delta(f)\|_\infty.$$

Under the same irreducibility assumptions as in Theorem 3, the average-reward Poisson equation has a solution satisfying

$$\delta(h^\pi) \le \left(I - (M^\pi)^\top\right)^{-1}\delta(r^\pi).$$

The entrywise bound follows by conditioning on the selected coordinate $J \sim \nu$. Fix i and s, s' that agree off coordinate i, and couple the actions A, A' exactly as in the proof of Lemma 2. If $J = i$, then S_{t+1} and S'_{t+1} inherit all coordinates other than i from s, s', which already agree, so they differ only at coordinate i, with $\mathbb{P}[S_{t+1,i} \ne S'_{t+1,i}] \le H^\pi_{i\leftarrow i}$ by the same maximal-coupling argument used in Lemma 2, and $|f(S_{t+1}) - f(S'_{t+1})| \le \delta_i(f)\,H^\pi_{i\leftarrow i}$ in expectation. If $J = j \ne i$, then coordinate i is not updated, so $S_{t+1,i} = s_i \ne s'_i = S'_{t+1,i}$, contributing $\delta_i(f)$; and coordinate j is updated, contributing $H^\pi_{j\leftarrow i}\,\delta_j(f)$ in expectation. Averaging over $J \sim \nu$ gives

$$\delta_i(K^\pi f) \le (1-\nu_i)\,\delta_i(f) + \nu_i H^\pi_{i\leftarrow i}\,\delta_i(f) + \sum_{j\ne i}\nu_j H^\pi_{j\leftarrow i}\,\delta_j(f) = \sum_{j=1}^n M^\pi_{j\leftarrow i}\,\delta_j(f),$$

which is the entrywise form of $\delta(K^\pi f) \le (M^\pi)^\top\delta(f)$. The remaining steps (Poisson decay, exponential bound on iterates) repeat the proof of Theorem 3 with H^π replaced by M^π.

B.7. Extension of Theorem 3 to the weighted case

The same argument also yields weighted oscillation bounds. We record the statement because it is often useful for certifying spectral-radius conditions through weighted norms.

Fix $w \in \mathbb{R}^n_{++}$ and $W = \operatorname{diag}(w)$. For any bounded f and $i \in [n]$,

$$\delta_i^w(T^\pi f) \;=\; w_i\,\delta_i(T^\pi f) \;\le\; w_i \sum_{j=1}^n H^\pi_{j\leftarrow i}\,\delta_j(f) \;=\; \sum_{j=1}^n \left(W^{-1}H^\pi W\right)_{j\leftarrow i}\delta_j^w(f).$$

This is the entrywise weighted one–step contraction. Iterating gives the multi–step bound. For the Poisson resolvent, one repeats the quotient-space proof of Theorem 3 with the seminorm $\|f\|_{\delta^w} = \|\delta^w(f)\|_\infty$. Let

$$\widetilde{H}^{\pi,w} := W^{-1}H^\pi W.$$

With our convention $H^\pi_{j\leftarrow i} = H^\pi[j,i]$, the entrywise bound above becomes the vector inequality $\delta^w(T^\pi f) \le (\widetilde{H}^{\pi,w})^\top\delta^w(f)$, with a transpose for the same reason as in the unweighted Lemma 2. The

Neumann series $\sum_{t\geq 0}((\widetilde{H}^{\pi,w})^\top)^t$ converges entrywise when $\rho(\widetilde{H}^{\pi,w}) < 1$ (noting $\rho((\widetilde{H}^{\pi,w})^\top) = \rho(\widetilde{H}^{\pi,w})$), and

$$\delta^w(h^\pi) \;\leq\; \sum_{t=0}^{\infty} \big((\widetilde{H}^{\pi,w})^\top\big)^t \delta^w(r^\pi) \;=\; \big(I - (\widetilde{H}^{\pi,w})^\top\big)^{-1} \delta^w(r^\pi).$$

Since $\widetilde{H}^{\pi,w} = W^{-1}H^\pi W$ is similar to H^π, the spectral radius is unchanged: $\rho(\widetilde{H}^{\pi,w}) = \rho(H^\pi)$. The weighted version is therefore useful for sharpening operator-norm certificates, not for changing the spectral condition itself. The truncated series bound and the uniform power bound over a compact policy class are identical to the unweighted case after replacing H^π by $\widetilde{H}^{\pi,w}$.

Appendix C. Proofs from Section 4

We bound the policy sensitivity matrix $\Pi(\pi)$ for entropy-regularized (temperature-τ) softmax policies in terms of per-agent logit Lipschitz constants. Throughout, total variation is normalized: $\mathrm{TV}(\mu,\nu) = \frac{1}{2}\|\mu-\nu\|_1$.

Lemma 9 (Softmax Lipschitz constant in total variation) *Fix $m \in \mathbb{N}$ and $\tau > 0$. For $u, v \in \mathbb{R}^m$, define*

$$\mathrm{soft}_\tau(u)_a \;\coloneqq\; \frac{\exp(u_a/\tau)}{\sum_{b=1}^m \exp(u_b/\tau)}, \qquad a \in [m].$$

Then

$$\mathrm{TV}\big(\mathrm{soft}_\tau(u),\, \mathrm{soft}_\tau(v)\big) \;\leq\; \frac{1}{2\tau}\,\|u-v\|_\infty.$$

The constant $1/(2\tau)$ is optimal: for $m \geq 2$, the supremum of $\mathrm{TV}(\mathrm{soft}_\tau(u), \mathrm{soft}_\tau(v))/\|u-v\|_\infty$ over $u \neq v$ equals $1/(2\tau)$, approached (but not attained) as $\|u-v\|_\infty \to 0$ along a balanced ± 1 direction starting from a uniform softmax. Sharpness in this asymptotic sense is exactly what makes $1/(2\tau)$ the best uniform Lipschitz constant.

Proof Write $p = \mathrm{soft}_\tau(u)$ and $q = \mathrm{soft}_\tau(v)$. By the mean value theorem on the line segment $w(t) = v + t(u-v)$, $t \in [0,1]$,

$$\|p-q\|_1 \;=\; \left\| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{soft}_\tau(w(t))\,\mathrm{d}t \right\|_1 \;\leq\; \int_0^1 \|D\mathrm{soft}_\tau(w(t))\,(u-v)\|_1\,\mathrm{d}t,$$

where $D\mathrm{soft}_\tau(w)$ is the Jacobian. For $w \in \mathbb{R}^m$ with $r = \mathrm{soft}_\tau(w)$,

$$D\mathrm{soft}_\tau(w) \;=\; \frac{1}{\tau}\big(\mathrm{diag}(r) - rr^\top\big).$$

Hence, using the induced operator norm from ℓ_∞ to ℓ_1,

$$\|p-q\|_1 \;\leq\; \frac{1}{\tau}\left(\sup_{t\in[0,1]} \big\|\mathrm{diag}(r(t)) - r(t)r(t)^\top\big\|_{\infty\to 1}\right) \|u-v\|_\infty, \qquad r(t) = \mathrm{soft}_\tau(w(t)).$$

We claim that for every probability vector r,

$$\big\|\mathrm{diag}(r) - rr^\top\big\|_{\infty\to 1} \;\leq\; 1. \qquad (\star)$$

This yields $\|p-q\|_1 \le \frac{1}{\tau}\|u-v\|_\infty$ and therefore $\mathrm{TV}(p,q) \le \frac{1}{2\tau}\|u-v\|_\infty$.

It remains to prove $(\star)$. Let $J(r) = \mathrm{diag}(r) - rr^\top$ and fix $\delta \in \mathbb{R}^m$ with $\|\delta\|_\infty \le 1$. Then

$$(J(r)\delta)_i \;=\; r_i\big(\delta_i - \langle r, \delta\rangle\big), \qquad i \in [m],$$

hence

$$\|J(r)\delta\|_1 \;=\; \sum_{i=1}^{m} r_i\,\big|\delta_i - \mu\big| \;=\; \mathbb{E}_{I\sim r}\big[\,|\delta_I - \mu|\,\big], \qquad \mu := \langle r, \delta\rangle \in [-1,1].$$

Let $X = \delta_I$ with $I \sim r$; then $X \in [-1,1]$ and $\mathbb{E}[X] = \mu$. By Jensen's inequality and the variance bound for bounded random variables,

$$\mathbb{E}\,|X-\mu| \;\le\; \sqrt{\mathrm{Var}(X)} \;\le\; \sqrt{1-\mu^2} \;\le\; 1,$$

where the middle step uses $\mathrm{Var}(X) \le \mathbb{E}[X^2] - \mu^2 \le 1-\mu^2$ since $X^2 \le 1$. Therefore $\|J(r)\delta\|_1 \le 1$ for all δ with $\|\delta\|_\infty \le 1$, proving $(\star)$. The bound is tight in the limit $\mu \to 0$ with X taking values ± 1 equally. ■

We now translate Lemma 9 into a bound on the policy sensitivity matrix $\Pi(\pi)$ for product-form, temperature-τ softmax policies with local logits.

Definition 10 (Local logits and per-coordinate logit Lipschitz constants) *For each agent k, suppose there is a logit function $g_k : \mathcal{S}_{O_k} \times \mathcal{A}_k \to \mathbb{R}$ such that*

$$\pi_k(a_k \mid s_{O_k}) \;\propto\; \exp\big(g_k(s_{O_k}, a_k)/\tau\big), \qquad a_k \in \mathcal{A}_k,$$

with temperature $\tau > 0$. For $i \in [n]$, define the one-coordinate logit Lipschitz constant

$$L_{k\leftarrow i} \;:=\; \sup_{s_{-i}=s'_{-i}} \big\|g_k(s_{O_k},\cdot) \;-\; g_k(s'_{O_k},\cdot)\big\|_\infty,$$

where the sup is over $s, s' \in \mathcal{S}$ that differ only on coordinate i.

Lemma 11 (Softmax temperature controls $\Pi(\pi)$) *Under the setup above, for all $k, i \in [n]$,*

$$\Pi_{k\leftarrow i}(\pi) \;\le\; \min\left\{1,\; \frac{L_{k\leftarrow i}}{2\tau}\right\}.$$

In particular, if $i \notin O_k$ then $L_{k\leftarrow i} = 0$ and $\Pi_{k\leftarrow i}(\pi) = 0$.

Proof Fix k and i. If $i \notin O_k$ then $g_k(s_{O_k},\cdot)$ is unchanged when s_i varies, hence $L_{k\leftarrow i} = 0$ and $\pi_k(\cdot \mid s_{O_k}) = \pi_k(\cdot \mid s'_{O_k})$ for all $s_{-i} = s'_{-i}$, giving $\Pi_{k\leftarrow i}(\pi) = 0$.

Assume $i \in O_k$. For s, s' with $s_{-i} = s'_{-i}$, apply Lemma 9 with $u = g_k(s_{O_k},\cdot)$ and $v = g_k(s'_{O_k},\cdot)$ to obtain

$$\mathrm{TV}\big(\pi_k(\cdot \mid s_{O_k}),\; \pi_k(\cdot \mid s'_{O_k})\big) \;\le\; \frac{1}{2\tau}\,\big\|g_k(s_{O_k},\cdot) - g_k(s'_{O_k},\cdot)\big\|_\infty \;\le\; \frac{L_{k\leftarrow i}}{2\tau}.$$

Taking the supremum over such s, s' yields $\Pi_{k\leftarrow i}(\pi) \le L_{k\leftarrow i}/(2\tau)$. The bound is trivially capped by 1 because total variation lies in $[0,1]$. ■

Remark 12 (Sharpness) *The constant $1/(2\tau)$ inherited from Lemma 9 is optimal in the asymptotic sense described there: it cannot be improved as $\|g_k(s_{O_k},\cdot) - g_k(s'_{O_k},\cdot)\|_\infty \to 0$. Consequently, controlling $\Pi(\pi)$ uniformly over a policy class amounts to lower-bounding the entropy temperature τ and upper-bounding the one-coordinate logit oscillations $L_{k\leftarrow i}$.*

Appendix D. Proofs from Section 5

All state and action spaces are finite. Total variation is normalized as $\mathrm{TV}(\mu,\nu) = \frac{1}{2}\sum_x |\mu(x)-\nu(x)|$. Coordinatewise oscillations are $\delta_i(f) = \sup\{|f(x) - f(y)| : x_{-i} = y_{-i}\}$ and $\delta(f) = (\delta_i(f))_{i=1}^n$. We use Theorem 3 from the main text.

D.1. Proof of Theorem 7

Fix a policy π. Let

$$g^\pi := r^\pi - \bar{r}^\pi, \qquad b^\pi := \delta(r^\pi) = \delta(g^\pi), \qquad H^\pi := E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi).$$

By Theorem 3,

$$\delta(h^\pi) \;\le\; \sum_{t=0}^{\infty} \left((H^\pi)^\top\right)^t b^\pi.$$

Define

$$\widehat{\delta}^{(\kappa)} := \sum_{t=0}^{\kappa} \left((H^\pi)^\top\right)^t b^\pi, \qquad R^{(\kappa)} := \sum_{t=\kappa+1}^{\infty} \left((H^\pi)^\top\right)^t b^\pi.$$

Then immediately

$$\delta(h^\pi) \;\le\; \widehat{\delta}^{(\kappa)} + R^{(\kappa)},$$

which proves the certificate statement.

For the tail bound,

$$\|R^{(\kappa)}\|_\infty \;\le\; \sum_{t=\kappa+1}^{\infty} \left\|\left((H^\pi)^\top\right)^t\right\|_{\infty\to\infty} \|b^\pi\|_\infty \;\le\; \sum_{t=\kappa+1}^{\infty} C\lambda^t \,\|b^\pi\|_\infty \;=\; \frac{C}{1-\lambda}\lambda^{\kappa+1}\|b^\pi\|_\infty.$$

For locality, let G_H^π be the support graph of H^π (edge $i \to j$ when $H^\pi_{j\leftarrow i} > 0$). Expanding the matrix power,

$$\left[\left((H^\pi)^\top\right)^t b^\pi\right]_i \;=\; \sum_{i=k_0,k_1,\dots,k_t} H^\pi_{k_1\leftarrow k_0}\, H^\pi_{k_2\leftarrow k_1} \cdots H^\pi_{k_t\leftarrow k_{t-1}}\, b^\pi_{k_t},$$

which sums over directed length-t paths $k_0 = i \to k_1 \to \cdots \to k_t$ in G_H^π. The contribution is nonzero only when every $H^\pi_{k_{\ell+1}\leftarrow k_\ell}$ is positive, i.e. when k_t is reachable from i in t steps. Hence the truncated sum $\widehat{\delta}_i^{(\kappa)}$ depends only on entries of H^π and b^π supported on the κ-hop *out-ball* of i in G_H^π (agents reachable from i in at most κ directed edges), and it can be computed by κ rounds of message passing along the reverse edges of that ball.

For the oracle truncation bias, define

$$\widehat{h}^\pi_\kappa := \sum_{t=0}^{\kappa} (T^\pi)^t g^\pi.$$

From the quotient-space construction used in the proof of Theorem 3,

$$[h^\pi] = \sum_{t=0}^{\infty} (\widetilde{T}^\pi)^t [g^\pi], \qquad [\widehat{h}^\pi_\kappa] = \sum_{t=0}^{\kappa} (\widetilde{T}^\pi)^t [g^\pi],$$

where $\widetilde{T}^\pi$ denotes the induced operator on $B(\mathcal{S})/\mathrm{span}\{\mathbf{1}\}$. Therefore

$$[\widehat{h}_\kappa^\pi - h^\pi] = -\sum_{t=\kappa+1}^{\infty} (\widetilde{T}^\pi)^t [g^\pi].$$

Applying the oscillation bound gives

$$\delta(\widehat{h}_\kappa^\pi - h^\pi) \;\le\; \sum_{t=\kappa+1}^{\infty} \left((H^\pi)^\top\right)^t b^\pi = R^{(\kappa)}.$$

Now use the standard identity

$$\inf_{c\in\mathbb{R}} \|f - c\mathbf{1}\|_\infty = \frac{1}{2}\,\mathrm{osc}(f), \qquad \mathrm{osc}(f) \le \sum_{i=1}^{n} \delta_i(f) = \|\delta(f)\|_1.$$

Applying this to $f = \widehat{h}_\kappa^\pi - h^\pi$ yields

$$B_\kappa^\pi = \inf_{c\in\mathbb{R}} \|\widehat{h}_\kappa^\pi - h^\pi - c\mathbf{1}\|_\infty \;\le\; \frac{1}{2}\,\|\delta(\widehat{h}_\kappa^\pi - h^\pi)\|_1 \;\le\; \frac{1}{2}\sum_{i=1}^{n} R_i^{(\kappa)}.$$

Finally,

$$\sum_{i=1}^{n} R_i^{(\kappa)} \le n\|R^{(\kappa)}\|_\infty \le \frac{nC}{1-\lambda}\lambda^{\kappa+1}\|b^\pi\|_\infty,$$

which gives

$$B_\kappa^\pi \le \frac{nC}{2(1-\lambda)}\lambda^{\kappa+1}\|b^\pi\|_\infty.$$

This completes the proof.

D.2. Preliminaries for the block-improvement step

Throughout this subsection, fix a baseline product policy π and write

$$A^\pi(s,a) = r(s,a) - \bar{r}^\pi + \sum_y P(y \mid s,a)h^\pi(y) - h^\pi(s).$$

Lemma 13 (Advantage perturbation via value perturbation modulo constants) *Let $\widehat{h}$ be any function on $\mathcal{S}$ and define*

$$\widehat{A}(s,a) = r(s,a) - \bar{r}^\pi + \sum_y P(y \mid s,a)\widehat{h}(y) - \widehat{h}(s).$$

Then

$$\sup_{s\in\mathcal{S},\, a\in\mathcal{A}} |\widehat{A}(s,a) - A^\pi(s,a)| \;\le\; 2\inf_{c\in\mathbb{R}} \|\widehat{h} - h^\pi - c\mathbf{1}\|_\infty.$$

Proof For any constant $c \in \mathbb{R}$, replacing $\widehat{h}$ by $\widehat{h} + c\mathbf{1}$ does not change $\widehat{A}$, because the additive constant cancels between the transition term and the state term. Hence, for every c,

$$\begin{aligned}|\widehat{A}(s,a) - A^\pi(s,a)| &= \left|\sum_y P(y \mid s,a)\big(\widehat{h}(y) + c - h^\pi(y)\big) - \big(\widehat{h}(s) + c - h^\pi(s)\big)\right| \\ &\le \sum_y P(y \mid s,a)\,|\widehat{h}(y) + c - h^\pi(y)| + |\widehat{h}(s) + c - h^\pi(s)| \\ &\le 2\|\widehat{h} - h^\pi - c\mathbf{1}\|_\infty.\end{aligned}$$

Taking the supremum over (s,a) and then the infimum over c proves the claim. ■

Lemma 14 (Per-state KL-prox duality) *Fix a finite action set $\mathcal{A}$, a reference distribution $q \in \Delta(\mathcal{A})$ with $q(a) > 0$ for all a (or, equivalently, restrict the maximization to distributions absolutely continuous with respect to q), a score vector $g \in \mathbb{R}^{\mathcal{A}}$, and $\eta > 0$. Then*

$$\max_{p \in \Delta(\mathcal{A})} \Big\{\langle p, g\rangle - \eta\,\mathrm{KL}(p\|q)\Big\} = \eta \log \sum_{a \in \mathcal{A}} q(a) e^{g(a)/\eta},$$

attained uniquely at

$$p^\star(a) \propto q(a) e^{g(a)/\eta}.$$

Moreover,

$$\langle p^\star - q,\ g\rangle \ \ge\ \eta\,\mathrm{KL}(p^\star\|q).$$

Proof The maximization is the standard log-sum-exp / Donsker–Varadhan variational formula. The Lagrangian for the constraint $\sum_a p(a) = 1$ has gradient $g(a) - \eta\log(p(a)/q(a)) - \eta - \lambda = 0$, which gives $p(a) \propto q(a)e^{g(a)/\eta}$. The value of the objective at $p^\star$ is $\eta \log \sum_a q(a) e^{g(a)/\eta}$ by direct substitution, and uniqueness follows from strict convexity of $\mathrm{KL}(\cdot\|q)$ on the simplex.

For the final inequality, evaluate the objective at $p = q$: $\langle q, g\rangle - \eta\mathrm{KL}(q\|q) = \langle q, g\rangle$. By optimality of $p^\star$,

$$\langle p^\star, g\rangle - \eta\mathrm{KL}(p^\star\|q) \ \ge\ \langle q, g\rangle,$$

which rearranges to $\langle p^\star - q, g\rangle \ge \eta\mathrm{KL}(p^\star\|q)$. ■

Lemma 15 (One-block performance difference) *Let π be a product policy and let μ be another product policy such that $\mu_{-k} = \pi_{-k}$ for some agent k. Assume the Markov chain under μ is irreducible with stationary distribution d^μ. Define*

$$g^\pi_{k,\star}(s, a_k) := \mathop{\mathbb{E}}_{a_{-k} \sim \prod_{j \ne k} \pi_j(\cdot|s)} \big[A^\pi(s, (a_k, a_{-k}))\big].$$

Then

$$\bar{r}(\mu) - \bar{r}(\pi) = \mathop{\mathbb{E}}_{S \sim d^\mu}\Big[\langle \mu_k(\cdot \mid S) - \pi_k(\cdot \mid S),\ g^\pi_{k,\star}(S, \cdot)\rangle\Big].$$

Proof Using the Poisson equation for h^π,

$$h^\pi - T^\pi h^\pi = r^\pi - \bar{r}^\pi,$$

one obtains the standard average-reward performance-difference identity

$$\bar{r}(\mu) - \bar{r}(\pi) = \mathop{\mathbb{E}}_{S\sim d^\mu,\ A\sim\mu(\cdot|S)} \big[A^\pi(S,A)\big].$$

Indeed,

$$\begin{aligned}
\mathop{\mathbb{E}}_{d^\mu,\mu}[A^\pi] &= \mathop{\mathbb{E}}_{d^\mu,\mu}\left[r(S,A) - \bar{r}^\pi + \sum_y P(y \mid S,A)h^\pi(y) - h^\pi(S)\right] \\
&= \mathop{\mathbb{E}}_{d^\mu,\mu}[r(S,A)] - \bar{r}^\pi + \mathop{\mathbb{E}}_{S'\sim d^\mu}[h^\pi(S')] - \mathop{\mathbb{E}}_{S\sim d^\mu}[h^\pi(S)] \\
&= \bar{r}(\mu) - \bar{r}(\pi),
\end{aligned}$$

where the middle two terms cancel by stationarity of d^μ under μ.

Now use $\mu_{-k} = \pi_{-k}$:

$$\mathop{\mathbb{E}}_{A\sim\mu(\cdot|s)}[A^\pi(s,A)] = \sum_{a_k} \mu_k(a_k \mid s)\, g^\pi_{k,\star}(s,a_k) = \langle \mu_k(\cdot \mid s), g^\pi_{k,\star}(s,\cdot)\rangle.$$

Likewise,

$$\langle \pi_k(\cdot \mid s), g^\pi_{k,\star}(s,\cdot)\rangle = \mathop{\mathbb{E}}_{A\sim\pi(\cdot|s)}[A^\pi(s,A)] = 0.$$

Therefore

$$\mathop{\mathbb{E}}_{A\sim\mu(\cdot|s)}[A^\pi(s,A)] = \langle \mu_k(\cdot \mid s) - \pi_k(\cdot \mid s),\, g^\pi_{k,\star}(s,\cdot)\rangle.$$

Averaging over $S \sim d^\mu$ proves the claim. ∎

D.3. Proof of Theorem 8

Fix π, k, and κ. First, by Lemma 13 with $\widehat{h} = \widehat{h}^\pi_\kappa$,

$$\sup_{s,a} |\widehat{A}^\pi_\kappa(s,a) - A^\pi(s,a)| \le 2B^\pi_\kappa.$$

Taking expectation over $a_{-k} \sim \prod_{j\neq k} \pi_j(\cdot \mid s)$ preserves the sup norm, so

$$\|\widehat{g}^\pi_{k,\kappa} - g^\pi_{k,\star}\|_\infty \le 2B^\pi_\kappa.$$

Now consider the policy μ defined by

$$\mu_k(\cdot \mid s) \propto \pi_k(\cdot \mid s) \exp\big(\widehat{g}^\pi_{k,\kappa}(s,\cdot)/\eta\big), \qquad \mu_{-k} = \pi_{-k}.$$

For each fixed state s, Lemma 14 with

$$q = \pi_k(\cdot \mid s), \qquad g = \widehat{g}^\pi_{k,\kappa}(s,\cdot), \qquad p^\star = \mu_k(\cdot \mid s)$$

gives

$$\langle \mu_k(\cdot \mid s) - \pi_k(\cdot \mid s),\, \widehat{g}^{\pi}_{k,\kappa}(s,\cdot)\rangle \geq \eta\, \mathrm{KL}\big(\mu_k(\cdot \mid s) \,\|\, \pi_k(\cdot \mid s)\big).$$

Subtract and add the exact logit:

$$\begin{aligned}
\langle \mu_k - \pi_k,\, g^{\pi}_{k,\star}\rangle &= \langle \mu_k - \pi_k,\, \widehat{g}^{\pi}_{k,\kappa}\rangle + \langle \mu_k - \pi_k,\, g^{\pi}_{k,\star} - \widehat{g}^{\pi}_{k,\kappa}\rangle \\
&\geq \eta\, \mathrm{KL}(\mu_k \| \pi_k) - \|\mu_k - \pi_k\|_1\, \|\widehat{g}^{\pi}_{k,\kappa} - g^{\pi}_{k,\star}\|_\infty \\
&\geq \eta\, \mathrm{KL}(\mu_k \| \pi_k) - 2\|\widehat{g}^{\pi}_{k,\kappa} - g^{\pi}_{k,\star}\|_\infty \\
&\geq \eta\, \mathrm{KL}(\mu_k \| \pi_k) - 4B^{\pi}_{\kappa},
\end{aligned}$$

where all distributions are evaluated at the same state s and we used $\|\mu_k - \pi_k\|_1 \leq 2$.

Finally, average over $S \sim d^{\mu}$ and apply Lemma 15:

$$\begin{aligned}
\bar{r}(\mu) - \bar{r}(\pi) &= \mathop{\mathbb{E}}_{S\sim d^{\mu}}\Big[\langle \mu_k(\cdot \mid S) - \pi_k(\cdot \mid S),\, g^{\pi}_{k,\star}(S,\cdot)\rangle\Big] \\
&\geq \eta \mathop{\mathbb{E}}_{S\sim d^{\mu}}\Big[\mathrm{KL}\big(\mu_k(\cdot \mid S) \,\|\, \pi_k(\cdot \mid S)\big)\Big] - 4B^{\pi}_{\kappa}.
\end{aligned}$$

Substituting the bound from Theorem 7,

$$B^{\pi}_{\kappa} \leq \frac{nC}{2(1-\lambda)}\lambda^{\kappa+1}\|b^{\pi}\|_\infty,$$

yields

$$\bar{r}(\mu) - \bar{r}(\pi) \geq \eta \mathop{\mathbb{E}}_{S\sim d^{\mu}}\Big[\mathrm{KL}\big(\mu_k(\cdot \mid S) \,\|\, \pi_k(\cdot \mid S)\big)\Big] - \frac{2nC}{1-\lambda}\lambda^{\kappa+1}\|b^{\pi}\|_\infty.$$

This proves the theorem.

Appendix E. Miscellaneous Results

E.1. Example: A Coupled System where Policy Smoothing is Ineffective

Our framework also captures the failure mode where policy smoothing cannot create locality because the environment itself has strong direct state-to-state feedback. Consider two agents with binary states and arbitrary actions. Suppose the next states deterministically copy each other:

$$P_1(s_1' = 1 \mid s, a) = \mathbf{1}\{s_2 = 1\}, \qquad P_2(s_2' = 1 \mid s, a) = \mathbf{1}\{s_1 = 1\}.$$

Actions have no effect on the next state. Hence

$$E^{\mathrm{a}} = 0, \qquad E^{\mathrm{s}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

where rows and columns are indexed by the influence convention $j \leftarrow i$. Therefore

$$H^{\pi} = E^{\mathrm{s}} + E^{\mathrm{a}}\Pi(\pi) = E^{\mathrm{s}}, \qquad \rho(H^{\pi}) = 1.$$

The policy sensitivity matrix $\Pi(\pi)$ is irrelevant because the action-coupling channel is absent. Thus no amount of policy smoothing can make $\rho(H^{\pi}) < 1$ in this example. This illustrates the complementary failure mode to the policy-induced locality examples: if the environment has an unstable direct state-feedback loop, the policy cannot remove it unless actions actually mediate that coupling.

E.2. Example: policy-induced locality in a nonlocal-looking system.

Consider a hub–and–spoke system with $n \geq 3$ agents.

Let P_1 be constant and, for $j > 1$, $P_j(s'_j = 1 \mid s, a) = \mathbf{1}\{a_1 = 1\}$. Then $E^{\mathrm{s}} = 0$ and $E^{\mathrm{a}}_{j \leftarrow 1} = 1$ for all $j > 1$ (others zero). A policy-independent, action-supremum certificate declares the system non-local ($\|C\|_\infty = 1$) (Qu et al., 2020).

Under a temperature-τ softmax for agent 1 with logit Lipschitz constants $\{L_{1 \leftarrow i}\}$, Lemma 6 gives $\Pi_{1 \leftarrow i} \leq L_{1 \leftarrow i}/(2\tau)$ and hence

$$\rho(H^\pi) \;\leq\; \rho\Big(E^{\mathrm{a}} \frac{L}{2\tau}\Big) \;=\; \frac{1}{2\tau} \sum_{i>1} L_{1 \leftarrow i} \;\leq\; \frac{(n-1)L_{\max}}{2\tau},$$

Thus locality is certified whenever $\tau > (n-1)L_{\max}/2$, and if the hub ignores its inputs ($L \equiv 0$) then $\rho(H^\pi) = 0$. This illustrates how locality can be *policy-induced*, while worst-case, policy-independent tests cannot detect it.

E.3. Properties of the Coordinatewise Oscillation Seminorm

Let $V = B(\mathcal{S})$ be the vector space of all bounded real-valued functions on the finite state space $\mathcal{S} = \prod_{i=1}^n \mathcal{S}_i$.

Definition 16 (Coordinatewise Oscillation) *For a function $f \in V$ and a coordinate $i \in [n]$, the i-oscillation is:*

$$\delta_i(f) = \sup \Big\{ |f(x) - f(y)| : x, y \in \mathcal{S},\ x_{-i} = y_{-i} \Big\}.$$

We define the function $p : V \to \mathbb{R}$ as the maximum oscillation:

$$p(f) = \|\delta(f)\|_\infty = \max_{i \in [n]} \delta_i(f).$$

Proposition 17 *The function $p(f) = \|\delta(f)\|_\infty$ is a seminorm on the vector space V.*

Proof To prove that $p(f)$ is a seminorm, we must verify three properties:

1. **Non-negativity:** $p(f) \geq 0$ for all $f \in V$.
2. **Absolute Homogeneity:** $p(cf) = |c|p(f)$ for all $f \in V$ and scalar $c \in \mathbb{R}$.
3. **Subadditivity (Triangle Inequality):** $p(f + g) \leq p(f) + p(g)$ for all $f, g \in V$.

1. Non-negativity: The absolute value $|f(x) - f(y)|$ is always non-negative. The supremum of a set of non-negative numbers, $\delta_i(f)$, is also non-negative. The maximum of a set of non-negative numbers, $p(f)$, is therefore non-negative.

2. Absolute Homogeneity: For any $f \in V$ and $c \in \mathbb{R}$:

$$\begin{aligned} \delta_i(cf) &= \sup_{x_{-i} = y_{-i}} |(cf)(x) - (cf)(y)| \\ &= \sup_{x_{-i} = y_{-i}} |c \cdot (f(x) - f(y))| \\ &= |c| \cdot \sup_{x_{-i} = y_{-i}} |f(x) - f(y)| = |c| \cdot \delta_i(f). \end{aligned}$$

Taking the maximum over all i:

$$p(cf) = \max_{i\in[n]} \delta_i(cf) = \max_{i\in[n]} \big(|c| \cdot \delta_i(f)\big) = |c| \cdot \max_{i\in[n]} \delta_i(f) = |c| \cdot p(f).$$

3. Subadditivity: For any $f, g \in V$:

$$\begin{aligned}
\delta_i(f+g) &= \sup_{x_{-i}=y_{-i}} |(f+g)(x) - (f+g)(y)| \\
&= \sup_{x_{-i}=y_{-i}} |(f(x) - f(y)) + (g(x) - g(y))| \\
&\le \sup_{x_{-i}=y_{-i}} \big(|f(x) - f(y)| + |g(x) - g(y)|\big) \quad \text{(by the triangle inequality for } \mathbb{R}) \\
&\le \sup_{x_{-i}=y_{-i}} |f(x) - f(y)| + \sup_{x_{-i}=y_{-i}} |g(x) - g(y)| \quad \text{(by a standard property of suprema)} \\
&= \delta_i(f) + \delta_i(g).
\end{aligned}$$

Now, taking the maximum over all i:

$$p(f+g) = \max_{i\in[n]} \delta_i(f+g) \le \max_{i\in[n]} (\delta_i(f) + \delta_i(g)).$$

For any i, we know $\delta_i(f) \le \max_j \delta_j(f) = p(f)$ and $\delta_i(g) \le \max_j \delta_j(g) = p(g)$. Thus:

$$p(f+g) \le \max_{i\in[n]} (p(f) + p(g)) = p(f) + p(g).$$

This completes the proof. ■

Remark 18 (Why it is a seminorm, not a norm) *A norm requires $p(f) = 0 \iff f = 0$. For our $p(f)$, if $f(x) = c$ for some non-zero constant c, then $f \neq 0$. However, for any i and any pair x, y with $x_{-i} = y_{-i}$, $f(x) = c$ and $f(y) = c$, so $|f(x) - f(y)| = 0$. This implies $\delta_i(f) = 0$ for all i, and thus $p(f) = 0$. Since $p(f) = 0$ for non-zero constant functions, $p(f)$ is a seminorm. In fact, $p(f) = 0 \iff f$ is a constant function.*

Proposition 19 *The value $p(f) = \|\delta(f)\|_\infty$ is the (best) Lipschitz constant of f with respect to the Hamming distance $d_H(\cdot,\cdot)$ on $\mathcal{S}$.*

Proof Let $K = p(f)$. We must show that for any $x, z \in \mathcal{S}$, $|f(x) - f(z)| \le K \cdot d_H(x, z)$.

Case 1: $d_H(x, z) = 1$. If $d_H(x, z) = 1$, then x and z differ in exactly one coordinate, say j. This means $x_{-j} = z_{-j}$. By the definition of $\delta_j(f)$:

$$|f(x) - f(z)| \le \sup_{x'_{-j}=y'_{-j}} |f(x') - f(y')| = \delta_j(f).$$

By the definition of $p(f)$, $\delta_j(f) \le \max_i \delta_i(f) = p(f) = K$. Therefore, $|f(x) - f(z)| \le K = K \cdot d_H(x, z)$. This also shows that K is the best Lipschitz constant for pairs at distance 1.

Case 2: $d_H(x, z) = k > 1$. Let the coordinates where x and z differ be $I = \{i_1, \dots, i_k\}$. We can construct a path from x to z by changing one coordinate at a time. Let $z^{(0)} = x$, and let $z^{(t)}$ be the state obtained by changing the first t coordinates in I from their values in x to their values

in z. For example, $z^{(1)}$ is identical to x except $z^{(1)}_{i_1} = z_{i_1}$. In general, $z^{(t)}$ and $z^{(t+1)}$ differ only in coordinate i_{t+1}. Thus, $d_H(z^{(t)}, z^{(t+1)}) = 1$. The full path is $x = z^{(0)}, z^{(1)}, \ldots, z^{(k)} = z$.

Using the triangle inequality, we "telescope" the sum:

$$\begin{aligned} |f(x) - f(z)| &= |f(z^{(0)}) - f(z^{(k)})| \\ &= \left| \sum_{t=0}^{k-1} \left(f(z^{(t)}) - f(z^{(t+1)}) \right) \right| \\ &\leq \sum_{t=0}^{k-1} |f(z^{(t)}) - f(z^{(t+1)})|. \end{aligned}$$

For each term in the sum, $z^{(t)}$ and $z^{(t+1)}$ differ in exactly one coordinate, so $d_H(z^{(t)}, z^{(t+1)}) = 1$. From Case 1, we know:

$$|f(z^{(t)}) - f(z^{(t+1)})| \leq p(f) = K.$$

Substituting this into the sum:

$$|f(x) - f(z)| \leq \sum_{t=0}^{k-1} K = k \cdot K.$$

Since $k = d_H(x, z)$, we have shown:

$$|f(x) - f(z)| \leq K \cdot d_H(x, z).$$

This holds for all $x, z \in \mathcal{S}$, so $p(f)$ is the Lipschitz constant of f w.r.t. d_H.

Optimality. Let L be any constant such that $|f(x) - f(y)| \leq L, d_H(x, y)$ for all x, y. Fix i and x, y with $x_{-i} = y_{-i}$; then $d_H(x, y) = 1$, so $|f(x) - f(y)| \leq L$. Taking the supremum over such pairs gives $\delta_i(f) \leq L$, hence $p(f) = \max_i \delta_i(f) \leq L$. Therefore $p(f)$ is the least Lipschitz constant. ■