TAYLOR REPRESENTATIONS FOR MODEL-FREE RL IN NETWORKED MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

In Networked Markov Decision Processes, transition dynamics are often unknown and the state–action space grows rapidly with the number of agents. In this setting, Taylor representations naturally approximate Q-functions, but a naive order-n expansion over N agents requires $\Theta(N^n)$ coefficients. We justify these expansions under smooth expected future local rewards with controlled derivatives. Under this condition, finite-speed information propagation and discounting imply that local-critic Taylor coefficients decay exponentially with the graph distance to the farthest agent involved. Discarding distant-agent coefficients and marginalizing then yield scalable local Taylor representations with a bound controlled by graph locality. Building on these representations, we propose a scalable model-free actor–critic algorithm, establishing finite-sample critic and near-stationarity guarantees for a linear LSTD critic. We then introduce a more expressive neural TD parameterization. Unlike prior constructive spectral methods, our approach covers settings without access to a known local dynamics map, such as hidden switched linear–quadratic regulation. Across three control benchmarks, our method matches or outperforms spectral baselines while scaling efficiently to large graphs.

1 INTRODUCTION

Networked Markov Decision Processes, in the context of multi-agent reinforcement learning (MARL), provide a natural framework for cooperative control in large interconnected systems, including communication networks, traffic systems, queueing networks, and other cyber-physical systems (Qu et al., 2020b; 2022; Chu et al., 2020). They nevertheless face two central challenges: the transition kernel is often unknown and the joint state–action space grows rapidly with the number of agents, making large-scale control challenging. To address the latter, in finite state–action spaces, Qu et al. (2020b; 2022) exploit the exponential decay property of local Q-functions, whereby the influence of distant agents decreases exponentially with graph distance. This enables localized Q-function and policy-gradient approximations. Lin et al. (2021) extend this principle to stochastic non-local dependencies through a more general decay property, while related guarantees also hold for average-reward objectives (Qu et al., 2020a). Related locality phenomena have also been studied in mean-field MARL and networked policy optimization (Gu et al., 2025; Zhang et al., 2023; Liang et al., 2025) and exponential sensitivity decay arises in graph-structured optimization and nonlinear optimal control (Shin et al., 2022; Na & Anitescu, 2020; Na et al., 2022). In continuous state–action spaces, however, localization alone is insufficient: even a local critic remains an unknown continuous function that must be represented and learned.

For continuous networked control, recent work addresses this representation problem through spatial structure (Olsson et al., 2024) and spectral representations (Ren et al., 2025b; Wang et al., 2026). A main setting targeted by the spectral approaches is one in which each local transition follows a known dynamics map with additive Gaussian noise. Spectral representations, together with finite-dimensional features such as random Fourier features and the exponential decay property mentioned above, then yield scalable critic approximations with explicit guarantees (Ren et al., 2023; 2025a; Rahimi & Recht, 2007). However, this approach can be restrictive: the transition kernel is typically unknown, and canonical network-control problems need not belong to this class of dynamics. Switched linear–quadratic regulation (LQR) with an unobserved global mode (Costa et al., 2005; Ogura et al., 2018), for instance, is not captured by such dynamics as it induces state– and action-

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.

dependent noise, shared across agents and temporally correlated through the Markov mode. Ren et al. (2025b) also propose a model-free variant named Spectral-SAC, but without the corresponding guarantees of their random Fourier features approach. In contrast, we focus here on the fully model-free setting, where the transition dynamics are unknown and only rollout data are available.

Beyond the setting of networked MDPs, a complementary route to tractable function approximation is smoothness. Taylor representations have been developed for continuous-state MDPs (Braverman et al., 2020; Xu et al., 2020; Maran et al., 2024a), with Maran et al. (2024a;b) showing how regularizing properties of the Bellman operator can justify polynomial representations. However, the poor scalability of a naive Taylor representation stems from its isotropy: the analysis places all Taylor coefficients on equal footing, so an order-n expansion over N agents generally requires $\Theta(N^n)$ coefficients. Moreover, much like the dynamics assumptions of prior constructive spectral methods, the smoothness assumption of (Maran et al., 2024a) need not hold in canonical network-control settings such as switched LQR with unobserved operating modes, as we show in Appendix E. In our analysis, we instead require smoothness of expected future local rewards in the initial state.

Our contribution. Our key observation is that, in networked MDPs, Taylor coefficients are not equally important: they themselves inherit graph locality. Finite propagation and discounting imply that derivatives involving an agent at graph distance d decay exponentially with d. This allows us to truncate the global Taylor expansion to monomials supported on $\mathcal{N}_i^\kappa$, the κ-hop neighborhood of agent i. After marginalizing distant variables, the resulting polynomial is equivalently a Taylor representation of a localized critic with only $\Theta(|\mathcal{N}_i^\kappa|^n)$ state monomials. Using rollout-based centroids as expansion points, we first obtain a linear-in-parameters Taylor Centroid critic and a scalable actor–critic with finite-sample critic and near-stationarity guarantees. We then generalize the parameterization with neural networks. Unlike prior constructive spectral approaches, our method requires neither knowledge nor a structural specification of the transition kernel. We benchmark it across three control problems of increasing difficulty. Table 1 summarizes how our contributions relate to prior work.

Table 1: Comparison with representative prior representation-based approaches.

	Qu et al.	**Olsson et al.**	**Ren et al.**	**Wang et al.**	**Maran et al.**	**Ours**
Networked MDP	✓	✓[‡]	✓	✓	–	✓
Continuous state–action	–	✓	✓	✓	✓	✓
Critic representation	Tabular	Local quadratic	Spectral	Spectral	Taylor	Local Taylor
Scalable representation	✓	✓	✓	✓	–	✓
Model-free construction	✓	✓	–[†]	–	✓	✓
End-to-end guarantee	Stationarity	–	Stationarity	Stationarity	No-regret	Stationarity

[†]Ren et al. (2025b) also propose a model-free alternative to their random Fourier features but without guarantees. [‡]Olsson et al. (2024) specialize to LQR.

2 Problem Setting

We consider N agents on an undirected graph $G = (V, E)$, $V = [N]$. Let $d_G(i, j)$ denote graph distance and, for $\kappa \geq 0$, define $\mathcal{N}_i^\kappa := \{j : d_G(i, j) \leq \kappa\}$, $\mathcal{N}_i := \mathcal{N}_i^1$, and $\mathcal{N}_{-i}^\kappa := V \setminus \mathcal{N}_i^\kappa$. Agent i has state $s_i \in \mathcal{S}_i \subset \mathbb{R}^{d_s}$ and takes action $a_i \in \mathcal{A}_i \subset \mathbb{R}^{d_a}$, where the state spaces are compact and convex and the action spaces are compact. We write $s_I = (s_j)_{j\in I}$ and $a_I = (a_j)_{j\in I}$. Our framework also allows a global context $z \in \mathcal{Z}$ which may be discrete or continuous. We consider the class of factorized transition kernels and policies:

$$P(z', s' \mid z, s, a) = P_g(z' \mid z) \prod_{i=1}^{N} P_i(s_i' \mid s_{\mathcal{N}_i}, a_{\mathcal{N}_i}, z), \quad \pi_\theta(a \mid z, s) = \prod_{i=1}^{N} \pi_{\theta_i}\Big(a_i \mid s_{\mathcal{N}_i^{\kappa_\pi}}, h_i(z)\Big).$$

Here, the policy observation radius is κ_π, and $h_i(z) \in \mathbb{R}^{d_g^i}$ denotes the global context observable to agent i. Thus z may be unobserved, partially observed, or fully observed by agent i. Furthermore, z affects every local transition but cannot transmit physical information between distant agents. Let Δ be the maximum degree of the graph. Assume that the rewards satisfy $r(s, a) = \frac{1}{N}\sum_{i=1}^{N} r_i(s_i, a_i)$ with $|r_i(s_i, a_i)| \leq R_{\max}$. Our analysis extends directly to local rewards depending on any fixed-radius neighborhood of agent i. The discount factor satisfies $\gamma \in (0, 1)$. We assume that the initial distribution ν_0 of (Z_0, S_0) is fixed and does not depend on the policy parameters θ. The cooperative

objective to be maximized is

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{(Z_0,S_0)\sim\nu_0,\pi_\theta}\left[\sum_{t\geq 0}\gamma^t r_i(S_{i,t},A_{i,t})\right].$$

To maximize this objective, we adopt the actor-critic framework (Konda & Tsitsiklis, 1999; Sutton et al., 1999; Sutton & Barto, 2018). To this end, define the local and global critics

$$Q_i^\pi(z,s,a) := \mathbb{E}_\pi\left[\sum_{t\geq 0}\gamma^t r_i(S_{i,t},A_{i,t})\,\middle|\, Z_0=z, S_0=s, A_0=a\right], \quad Q^\pi := \frac{1}{N}\sum_{i=1}^{N} Q_i^\pi. \tag{1}$$

To make our targeted Taylor representations well-defined, we use the following assumption:

Assumption 1 (Expected reward regularity) *There exist an integer $q \geq 2$ and a constant $\bar{L} < \infty$, both independent of the network size N, such that, uniformly over agents i and all policies π of the above factorized form, for every $t \geq 0$ the map $s \mapsto \mathbb{E}_\pi[r_i(S_{i,t},A_{i,t}) \,|\, Z_0=z, S_0=s, A_0=a]$ is of class C^q and*

$$\max_{|\beta|\leq q}\sup_{z,s,a}\left|\partial_s^\beta \mathbb{E}_\pi[r_i(S_{i,t},A_{i,t}) \,|\, Z_0=z,\ S_0=s,\ A_0=a]\right| \leq \bar{L}.$$

Assumption 1 is an expected reward regularity condition: a small perturbation of the current physical state may lead to very different trajectories, but should not create arbitrarily sharp changes in the expected future local reward. Appendix E gives *concrete sufficient conditions* for Assumption 1, introduces a *rollout-based stress test* for assessing it empirically, and shows that the uniform-in-t derivative bound can be weakened to a *discounted-summability condition*. We retain the uniform formulation in the main text for simplicity.

Assumption 1 plays two roles: it makes each Q_i^π of class C^q (Proposition 1), enabling Taylor expansions of any order $n \leq q-1$, and, together with finite graph propagation, yields the exponential derivative decay of Lemma 1. In contrast, the standard exponential decay of the value-localization error with the localization radius (Ren et al., 2025b; Wang et al., 2026) requires only bounded rewards and graph-local factorization.

Proposition 1 *The individual action-value function $s \mapsto Q_i^\pi(z,s,a)$ is of class C^q for all i, z, a.*

Maran et al. (2024a) also construct local polynomial value-function approximations through Taylor expansions, but rely on Mild smoothness, a regularity assumption on the Bellman operator, whereas Assumption 1 only requires smoothness of the future reward sequence defining Q_i^π. Hence, under smooth rewards, sufficiently high-order Mild smoothness implies Assumption 1, but not conversely: Appendix E shows this separation on switched LQR (Costa et al., 2005; Ogura et al., 2018), while also showing that the additive-Gaussian setting of Ren et al. (2025b); Wang et al. (2026) satisfies Assumption 1 under uniformly C^q local dynamics and rewards when Δ is independent of N.

The smoothness property of Proposition 1 naturally leads to our Taylor representations for the actor–critic structure developed in the remainder of the paper: we first construct a scalable Taylor representation of the critic in Section 3, and then use it to derive an actor-critic algorithm in Section 4.

3 Taylor Centroid Critics

3.1 Taylor expansions and scalability

Having established the required smoothness of the critic, we now turn to its representation. As in Maran et al. (2024a), we begin with an order-n Taylor expansion around an expansion point μ, for any chosen integer $1 \leq n \leq q-1$.[1]

$$Q_i^\pi(z,s,a) \simeq \sum_{\substack{\beta\in\mathbb{N}^{Nd_s}\\ |\beta|\leq n}} \frac{\partial_s^\beta Q_i^\pi(z,\mu,a)}{\beta!}(s-\mu)^\beta \tag{2}$$

[1]Write $\beta = (\beta_{j,r})_{j\in[N],\, r\in[d_s]} \in \mathbb{N}^{Nd_s}$. We use $|\beta| = \sum_{j=1}^N \sum_{r=1}^{d_s} \beta_{j,r}$, $\beta! = \beta_{1,1}!\cdots\beta_{N,d_s}!$, $x^\beta = x_{1,1}^{\beta_{1,1}}\cdots x_{N,d_s}^{\beta_{N,d_s}}$, $\partial_s^\beta = \frac{\partial^{|\beta|}}{\partial s_{1,1}^{\beta_{1,1}}\cdots\partial s_{N,d_s}^{\beta_{N,d_s}}}$ and $\mathrm{supp}_G(\beta) = \{i : \exists l \mid \beta_{i,l} \neq 0\}$.

The drawback of this representation is dimensionality: an order-n expansion contains $\binom{Nd_s+n}{n} = \Theta(N^n)$ coefficients, while μ and a are themselves ambient-dimensional. Importantly, existing locality results show that changing states and actions outside the κ-hop neighborhood of any agent i changes its local value Q_i^π by an amount that decays exponentially with κ (Qu et al., 2020b; 2022; Ren et al., 2025b; Wang et al., 2026), but do not characterize its Taylor coefficients. Our key observation is that expected reward regularity further allows exponential decay of the derivatives with graph distance. Lemma 1 formalizes this decay.

Lemma 1 (Exponential decay of critic derivatives) *Define $\rho := \gamma^{1/(1+\kappa_\pi)} \in (0,1)$, and the graph reach $d_i(\beta) := \max_{j \in \mathrm{supp}_G(\beta)} d_G(i,j)$ with $d_i(0) := 0$. Then, for every $|\beta| \le n+1$,*

$$\left|\partial_s^\beta Q_i^\pi(z,s,a)\right| \le L_Q \rho^{d_i(\beta)}, \qquad L_Q := \frac{\bar{L}}{1-\gamma}$$

Hence any derivative involving an agent at distance d is $O(\rho^d)$: indeed, finite propagation delays its influence on agent i by at least $d/(1+\kappa_\pi)$ steps, after which discounting contributes the factor ρ^d. The decay can be substantially faster in practice (see Appendix D.1); henceforth, ρ may denote the best valid geometric rate. This motivates discarding derivatives with respect to agents outside $\mathcal{N}_i^\kappa$. We denote their multi-indices by:

$$\mathcal{M}^n_{\mathcal{N}_i^\kappa} := \{\beta \in \mathbb{N}^{Nd_s} : |\beta| \le n,\ \mathrm{supp}_G(\beta) \subseteq \mathcal{N}_i^\kappa\}.$$

This removes coordinates of agents at distance more than κ from the polynomial terms. However, the remaining Taylor coefficients still depend on their values. To obtain coefficients that require only local states and actions, fix an arbitrary probability measure $\nu_{i,\mathrm{far}}$ on $\mathcal{S}_{\mathcal{N}_{-i}^\kappa} \times \mathcal{A}_{\mathcal{N}_{-i}^\kappa}$ and define, for $\beta \in \mathcal{M}^n_{\mathcal{N}_i^\kappa}$, the marginalized Taylor coefficients:

$$\mathscr{D}_{i,\beta}(z, s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) := \int_{\mathcal{S}_{\mathcal{N}_{-i}^\kappa} \times \mathcal{A}_{\mathcal{N}_{-i}^\kappa}} \partial_s^\beta Q_i^\pi(z,s,a)\, d\nu_{i,\mathrm{far}}(s_{\mathcal{N}_{-i}^\kappa}, a_{\mathcal{N}_{-i}^\kappa}). \tag{3}$$

We use expansion points $\mu_1, \dots, \mu_K \in \mathcal{S}$, with $k(s) \in [K]$ selecting the point $\mu_{k(s)}$ closest to s in L^1 norm, in the spirit of classical local polynomial and locally weighted approximation methods (Fan & Gijbels, 1996; Atkeson et al., 1997). We define the clusters in the global state space, so that the cluster index $k(s)$ can inject global-state information into otherwise local critics, in particular helping recover latent global regimes when z is unobserved. Figure 8 illustrates this effect. Discarding distant derivatives and marginalizing the retained ones then yields the local Taylor representation:

$$\widehat{Q}_i^\pi(z,s,a) = \sum_{\beta \in \mathcal{M}^n_{\mathcal{N}_i^\kappa}} \frac{\mathscr{D}_{i,\beta}(z, \mu_{k(s),\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})}{\beta!} \left(s_{\mathcal{N}_i^\kappa} - \mu_{k(s),\mathcal{N}_i^\kappa}\right)^\beta. \tag{4}$$

At each fixed $k(s)$, this Taylor representation requires $\Theta(|\mathcal{N}_i^\kappa|^n)$ learned coefficients. Notice that $\mathscr{D}_{i,0}$ is precisely the critic obtained by marginalizing Q_i^π over states and actions outside $\mathcal{N}_i^\kappa$; denoting it by $\bar{Q}_i^\pi$, this recovers the localized critic of Ren et al. (2025b). Moreover, $\mathscr{D}_{i,\beta} = \partial_s^\beta \bar{Q}_i^\pi$, so Equation 4 is exactly the order-n Taylor expansion of $\bar{Q}_i^\pi$ around the selected centroid. Combining this viewpoint with Lemma 1 yields the following error estimate:

Theorem 1 (Scalable local Taylor representation) *Under Assumption 1, define, for $0 \le d \le \kappa$ and every integer $p \ge 0$,*

$$W_d := \sup_s \left\| s_{\mathcal{N}_i^d} - \mu_{k(s),\mathcal{N}_i^d} \right\|_1, \qquad \mathcal{R}_p := \rho^\kappa W_\kappa^p + (1-\rho) \sum_{d=0}^{\kappa-1} \rho^d W_d^p.$$

Then the truncated Taylor critic $\widehat{Q}_i^\pi$ satisfies

$$\left\| Q_i^\pi - \widehat{Q}_i^\pi \right\|_\infty \le \frac{2R_{\max}}{1-\gamma} \rho^{\kappa+1} + \frac{L_Q}{(n+1)!} \mathcal{R}_{n+1} =: \Delta_{\mathrm{TC}}$$

Moreover, $\widehat{Q}_i^\pi$ contains $\binom{|\mathcal{N}_i^\kappa| d_s + n}{n} = \Theta(|\mathcal{N}_i^\kappa|^n) = O(\Delta^{n\kappa})$ state monomials.

Theorem 1 decomposes the approximation error into the standard $O(\rho^{\kappa+1})$ localization term (Qu et al., 2020b; Ren et al., 2025b; Wang et al., 2026) and a Taylor remainder. Lemma 1 sharpens the usual isotropic remainder W_κ^{n+1} to $\mathcal{R}_{n+1}$, which attenuates contributions from the d-hop neighborhood by ρ^d through graph locality. Due to the $1/(n+1)!$ term, the Taylor order n typically improves the local approximation. The effect of κ is subtler: enlarging the neighborhood reduces localization error but increases the Taylor radius W_κ and introduces additional distant coordinates into $\mathcal{R}_{n+1}$. Corollary F.1 replaces the worst-case radii W_d by average distances to the centroids, motivating the K-medians (Cardot et al., 2012) procedure used below to reduce the average L^1 distance to expansion points. The key takeaway is a controlled local Taylor approximation with an N-independent number of monomials for fixed (κ, n, d_s).

3.2 Linear and Neural Taylor Centroid critic estimation

The local Taylor representation of Equation 4 is scalable, but its marginalized coefficients $\mathscr{D}_{i,\beta}$ must still be learned from rollouts. Here, we first consider a linear parameterization which we name Linear Taylor Centroid (LTC), and a neural one, Neural Taylor Centroid (NTC). Letting $y_i = h_i(z)$, define the coefficient features[2] and local Taylor basis as

$$\psi_i(k, y_i, a_{\mathcal{N}_i^\kappa}) := [1; \mu_{k,\mathcal{N}_i^\kappa}; a_{\mathcal{N}_i^\kappa}; y_i], \qquad T_i(s) := \left[\frac{(s_{\mathcal{N}_i^\kappa} - \mu_{k(s),\mathcal{N}_i^\kappa})^\beta}{\beta!}\right]_{\beta \in \mathcal{M}^n_{\mathcal{N}_i^\kappa}}.$$

Linear Taylor Centroid. LTC approximates each marginalized Taylor coefficient with a function linear in ψ_i:

$$\widetilde{Q}_i^{\mathrm{LTC}}(y_i, s, a_{\mathcal{N}_i^\kappa}; w_i) = \sum_{\beta \in \mathcal{M}^n_{\mathcal{N}_i^\kappa}} \left(\psi_i(k(s), y_i, a_{\mathcal{N}_i^\kappa})^\top w_{i,\beta}\right) \frac{(s_{\mathcal{N}_i^\kappa} - \mu_{k(s),\mathcal{N}_i^\kappa})^\beta}{\beta!}$$

Alternatively, setting $\Phi_i = \psi_i \otimes T_i$, $\widetilde{Q}_i^{\mathrm{LTC}} = \Phi_i^\top w_i$ is linear in w_i, enabling us to use the Least-Squares Temporal Differences (LSTD) algorithm (Bradtke & Barto, 1996; Lazaric et al., 2012; Ren et al., 2025b; Qu et al., 2020b; Wang et al., 2026) to fit the critic. Moreover, the number of learnable parameters in this representation is $p_i := \binom{|\mathcal{N}_i^\kappa| d_s + n}{n}\left[1 + |\mathcal{N}_i^\kappa|(d_s + d_a) + d_g^i\right] = \Theta\left(|\mathcal{N}_i^\kappa|^{n+1}\right) = O\left(\Delta^{\kappa(n+1)}\right)$ which is independent of N. Importantly, we share the coefficient map w_i across clusters rather than fitting cluster-specific weights, since rarely visited clusters would otherwise have too little data to estimate their coefficients reliably. Cluster-specific behavior is still captured through the centroid μ_k in the feature map and behavior relative to the global state s is captured by $k(s)$. Since the ideal Taylor Centroid coefficient maps $\mathscr{D}_{i,\beta}$ need not be affine, we quantify the resulting approximation error by defining

$$\bar{\varepsilon}_{\mathrm{lin}} := \max_{\beta \in \mathcal{M}^n_{\mathcal{N}_i^\kappa}} \rho^{-d_i(\beta)} \inf_{w \in \mathbb{R}^{d_\psi^i}} \sup_{z,k,a} \left|\mathscr{D}_{i,\beta}(z, \mu_{k,\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) - w^\top \psi_i(k, h_i(z), a_{\mathcal{N}_i^\kappa})\right| \le L_Q.$$

The quantity $\bar{\varepsilon}_{\mathrm{lin}}$ measures misspecification of the coefficient feature map, normalized by the natural scale $\rho^{d_i(\beta)}$ from Lemma 1. Importantly, localization reduces the effective dimension of coefficient estimation from the global state dimension to that of a local neighborhood, making this estimation problem substantially more tractable. Furthermore, adding adequate features to the feature map ψ_i may lower this misspecification. Notice also that increasing n introduces additional coefficients to fit, so $\bar{\varepsilon}_{\mathrm{lin}}$ is nondecreasing in n. Increasing κ similarly enlarges the representation and its parameter count but may allow for better estimation through a better feature map. Thus, although larger n and κ can capture richer critic structure, they can also make estimation harder, yielding an approximation–estimation tradeoff in practice. This leads to the following finite-sample bound.

Theorem 2 (Finite-sample error of LTC) *Fix the centroids and collect M_c independent rollouts of length T. Define*

$$\mathcal{E}_i^{(\kappa)} := \bar{\varepsilon}_{\mathrm{lin}} \sum_{p=0}^{n} \frac{\mathcal{R}_p}{p!}, \qquad \mathcal{G}_i^{(\kappa)} := R_\psi^2 \sum_{p=0}^{n} \frac{W_\kappa^{2p}}{p!}, \qquad \mathcal{V}_i^{(\kappa)} := \frac{R_{\max}}{1-\gamma} + L_Q \sum_{p=1}^{n} \frac{\mathcal{R}_p}{p!},$$

[2]The coefficient feature map ψ may be chosen arbitrarily and may include nonlinearities.

where $R_\psi := \sup_{k,y,a} \|\psi_i(k,y,a)\|$. *Conditionally on the centroids and on* $\bigcap_{i=1}^N \{\sigma_{\min}(\widehat{A}_i) \geq c_{\mathrm{LSTD}} > 0\}$*, with probability at least* $1-\delta$*, simultaneously for all* $i \in [N]$*,*

$$\left\| Q_i^\pi - \widetilde{Q}_i^{\mathrm{LTC}}\big(h_i(\cdot),\cdot,\cdot_{\mathcal{N}_i^\kappa};\widehat{w}_i\big)\right\|_\infty \leq \left(1+\frac{(1+\gamma)\mathcal{G}_i^{(\kappa)}}{c_{\mathrm{LSTD}}}\right)\left(\Delta_{\mathrm{TC}}+\mathcal{E}_i^{(\kappa)}\right) + \frac{2\gamma\mathcal{G}_i^{(\kappa)}\big(\mathcal{V}_i^{(\kappa)}+\mathcal{E}_i^{(\kappa)}\big)}{c_{\mathrm{LSTD}}}\sqrt{\frac{2\log(2N/\delta)}{M_c(T-1)}}.$$

Theorem 2 separates the LTC error into the intrinsic Taylor Centroid approximation Δ_{TC}, the estimation error $\mathcal{E}_i^{(\kappa)}$, and a standard statistical $O((M_cT)^{-1/2})$ term modulated by the local feature factor $\mathcal{G}_i^{(\kappa)}$, critic scale $\mathcal{V}_i^{(\kappa)}$, and LSTD conditioning. The term $\mathcal{E}_i^{(\kappa)}$ captures the following centroid tradeoff: increasing K reduces the local radii in $\mathcal{R}_p$, but can increase $\bar{\varepsilon}_{\mathrm{lin}}$ as a single shared coefficient map must fit more centroid-dependent coefficients. Appendix D.2 examines these two effects empirically. As in Ren et al. (2025b), our guarantee is conditional on the LSTD matrices being well conditioned. We choose the number of rollouts empirically to ensure this condition (see Appendix C.2); ridge regularization can further improve conditioning in practice.
This parallels prior finite-sample critic bounds of (Qu et al., 2020b; Ren et al., 2025b), with Ren et al. (2025b) also incurring spectral approximation error. Ours instead obtains Taylor and estimation errors, motivating the following neural extension when the latter dominates.

Neural Taylor Centroid. NTC replaces the affine coefficients by a neural network $f_{\xi_i} : \mathbb{R}^{\dim(\psi_i)} \longrightarrow \mathbb{R}^{|\mathcal{M}^n_{\mathcal{N}_i^\kappa}|}$.

$$\widetilde{Q}_i^{\mathrm{NTC}} = \langle f_{\xi_i}(\psi_i), T_i(s)\rangle = \sum_{\beta\in\mathcal{M}^n_{\mathcal{N}_i^\kappa}} [f_{\xi_i}(\psi_i)]_\beta \frac{\left(s_{\mathcal{N}_i^\kappa}-\mu_{k(s),\mathcal{N}_i^\kappa}\right)^\beta}{\beta!}.$$

We train f_{ξ_i} by standard semi-gradient TD regression (Sutton & Barto, 2018) (see Appendix H). NTC preserves the structural guarantees of Theorem 1 and changes only the coefficient parameterization. This can reduce the analogous error to $\bar{\varepsilon}_{\mathrm{lin}}$, at the cost of the guarantees of Theorem 2.

4 Learning and Policy Optimization

To construct a scalable policy-gradient estimator, let agent i's observation be $o_i(z,s) := (s_{\mathcal{N}_i^{\kappa_\pi}}, h_i(z))$ and define its policy score $\mathsf{s}_i(o_i,a_i) := \nabla_{\theta_i}\log\pi_{i,\theta_i}(a_i \mid o_i)$, which we assume to be uniformly bounded as standard in localized policy-gradient analyses (Qu et al., 2020a; Ren et al., 2025b): $\sup_{o_i,a_i}\|\mathsf{s}_i(o_i,a_i)\| \leq B_{\mathsf{s},i}$. By the global reward decomposition and the policy gradient theorem (Sutton et al., 1999):

$$\nabla_{\theta_i} J(\theta) = \frac{1}{N}\mathbb{E}_{\pi_\theta}\left[\sum_{t\geq 0}\gamma^t \mathsf{s}_i(O_{i,t},A_{i,t})\sum_{\ell=1}^N A_\ell^{\pi_\theta}(Z_t,S_t,A_t)\right].$$

Estimating the advantage through the critic. At actor iteration m, let $\widetilde{Q}_{\ell,m}$ denote the LTC/NTC critic of agent ℓ and suppose that, for every ℓ,z,s,a, $|\widetilde{Q}_{\ell,m}(h_\ell(z),s,a_{\mathcal{N}_\ell^\kappa}) - Q_\ell^{\pi_{\theta_m}}(z,s,a)| \leq \Delta_Q^{(m)}$. For LTC, Theorem 2 provides such a finite-sample certificate; for NTC, $\Delta_Q^{(m)}$ denotes the approximation error achieved by the neural critic. Let $\widetilde{V}_{\ell,m}$ be the value function induced by $\widetilde{Q}_{\ell,m}$. We use the one-step TD (Sutton, 1988; Sutton & Barto, 2018; Konda & Tsitsiklis, 1999) residual as the local advantage estimate $\widehat{A}_{\ell,t} := \widehat{\delta}_{\ell,t} = r_{\ell,t} + \gamma\widetilde{V}_{\ell,m}(Z_{t+1},S_{t+1}) - \widetilde{V}_{\ell,m}(Z_t,S_t)$. In practice, we use M_s additional actor rollouts to form agent i's empirical policy-gradient estimator:

$$\widehat{g}_{i,m} = \frac{1}{NM_s}\sum_{j=1}^{M_s}\sum_{t=0}^{T-1}\gamma^t \mathsf{s}_i\big(O_{i,t}^{(j)},A_{i,t}^{(j)}\big)\sum_{\ell\in\mathcal{N}_i^{\kappa+\kappa_\pi}}\widehat{A}_{\ell,t}^{(j)}. \tag{5}$$

Equation 5 estimates agent i's policy-gradient contribution by retaining only advantages of agents $\ell \in \mathcal{N}_i^{\kappa+\kappa_\pi}$ which removes the dependence in N. This follows the truncated policy gradient principle of Qu et al. (2020b); Ren et al. (2025b), here using localized one-step TD advantages rather than localized Q-values. We then obtain the following *near-stationarity* guarantee.

Theorem 3 (End-to-end policy optimization) *Let* $\Gamma_T := (1-\gamma^T)/(1-\gamma)$ *and define*

$$\tau_i := \frac{1}{N} \sum_{\ell:\, d_G(i,\ell) > \kappa+\kappa_\pi} \rho^{d_G(i,\ell)}, \quad \mathcal{T} := \Big(\sum_{i=1}^{N} B_{\mathbf{s},i}^2 \tau_i^2\Big)^{1/2}, \quad \mathcal{B} := \frac{1}{N}\Big(\sum_{i=1}^{N} (B_{\mathbf{s},i}|\mathcal{N}_i^{\kappa+\kappa_\pi}|)^2\Big)^{1/2}.$$

Assume that J *is* L_J*-smooth. Let* $\hat{g}_m = (\hat{g}_{1,m}, \ldots, \hat{g}_{N,m})$*; with probability at least* $1-\delta$*, for every* $m < M$*,*

$$\begin{aligned} \|\widehat{g}_m - \nabla J(\theta_m)\| \le \epsilon_m := \frac{2R_{\max}}{(1-\gamma)^2}\mathcal{T} + \mathcal{B}\left(\gamma\Gamma_T \Delta_Q^{(m)} + \frac{2R_{\max}\gamma^T}{(1-\gamma)^2}\right) \\ + G_m\sqrt{\frac{2\log(4M/\delta)}{M_s}}, \end{aligned} \tag{6}$$

where G_m *bounds the localized policy-gradient samples (details are given in Appendix F.5). If* $\epsilon_m \le \bar{\epsilon}$ *and* $G_m \le \bar{G}$ *for all* $m < M$*, then* $\theta_{m+1} = \theta_m + \eta(m+1)^{-1/2}\widehat{g}_m$ *satisfies*

$$\min_{0\le m<M} \|\nabla J(\theta_m)\|^2 \le O(M^{-1/2}) + \bar{\epsilon}^2 + \bar{G}^2 O\left(\frac{\log M}{\sqrt{M}}\right). \tag{7}$$

Theorem 3 separates localization, critic, rollout-truncation, and actor-sampling errors, which decrease with localization radius, critic accuracy, rollout length, and actor sample size, respectively, yielding a near-stationarity guarantee with an explicit error floor. The term $\mathcal{T}$ sharpens the worst-case truncation bound of Ren et al. (2025b) by accounting for each discarded agent ℓ with its own decay factor $\rho^{d_G(i,\ell)}$. This can substantially improve the bound when ρ is small: if $\sum_\ell \rho^{d_G(i,\ell)} = O(1)$, then $\mathcal{T} = O(N^{-1/2})$. In contrast, discarded agents were previously assigned the same worst-case decay factor $\rho^{\kappa+\kappa_\pi}$ which can produce an $O(1)$ error in each agent's policy-gradient estimate and a corresponding term $\mathcal{T}$ of order $O(\sqrt{N})$.

5 Algorithm and Complexity

We present the Taylor Centroid actor–critic for LTC and NTC, which uses only rollout data.

Algorithm 1 Taylor Centroid Actor–Critic

Require: Policy θ_0, stepsizes $\eta_m = \eta(m+1)^{-1/2}$, radii κ, κ_π, Taylor order n, centroids K, number of epochs M

1: **for** $m = 0, \ldots, M-1$ **do**
2: Refit the K centroids from previously observed states.
3: Collect M_c critic rollouts; in parallel over agents i, fit $\widetilde{Q}_{i,m}$ using LTC or NTC.
4: Collect M_s additional actor rollouts; in parallel over i, compute $\widehat{A}_{\ell,t}$ for $\ell \in \mathcal{N}_i^{\kappa+\kappa_\pi}$ and form $\widehat{g}_{i,m}$.
5: Set $\widehat{g}_m = (\widehat{g}_{1,m}, \ldots, \widehat{g}_{N,m})$ and update $\theta_{m+1} \leftarrow \theta_m + \eta_m \widehat{g}_m$.
6: **end for**

Motivated by Corollary F.1, we fit K-medians centroids and refit them at each iteration to track policy-induced shifts in the state distribution (see Appendix D.5). Importantly, the time complexity of each LTC iteration scales linearly with N and is largely parallelizable across agents and samples; a detailed complexity analysis is provided in Appendix F.6.

6 Experiments

We evaluate LTC and NTC on networked LQR (Kalman, 1960; Fazel et al., 2018; Ren et al., 2025b), switched LQR (Costa et al., 2005; Ogura et al., 2018), and nonlinear Kuramoto synchronization (Kuramoto, 1975; Dörfler & Bullo, 2014; Ren et al., 2025b). Baselines include random Fourier features (RF on LQR and RF-SAC for Kuramoto) and Spectral-SAC (Ren et al., 2025b), CDCPG[3] (Wang et al., 2026), and generic multilayer perceptron (MLP) and graph neural network (GNN) critics. Recall that RF and CDCPG do not directly apply to hidden switched LQR, which lacks their required known local dynamics model. We describe the environments in Appendix B; the evaluation and training protocols are given in Appendix C.

[3]To satisfy the assumption $\kappa > \kappa_\pi$ of Wang et al. (2026), we choose $\kappa' = \max(\kappa, \kappa_\pi + 1)$ in each panel of the plots below for CDCPG only.

Critic estimation quality. Figure 1 describes the effect of Taylor order on the critic estimation. On LQR, where V^π is quadratic for the affine-mean Gaussian policies considered here, the sharp improvement from $n = 1$ to $n = 2$ and saturation at $n = 3$ match the expected second-order structure. Switched LQR shows the same pattern despite the additional difficulty of predicting the mode-conditioned value V^π_σ without observing σ; here NTC significantly improves over LTC, consistent with its greater expressivity. Regarding the Kuramoto benchmark, we observe a near-systematic improvement with n, consistent with the fact that the critic is non-quadratic. Increasing κ or n can improve expressivity but makes estimation harder by increasing the number of parameters. Nevertheless, for every (κ, κ_π), we observe that Taylor Centroid is competitive with the other baselines. Another parameter affecting the critic estimation quality is the number of clusters. We study this effect in Appendix D.2.

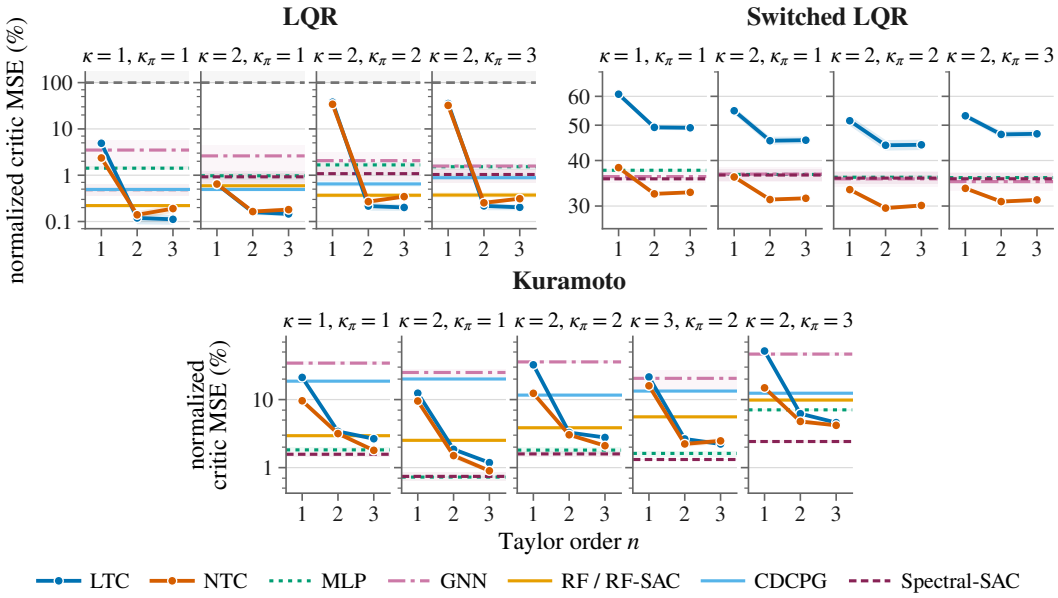

Figure 1: **Critic estimation quality.** Normalized critic MSE across Taylor orders and locality settings on LQR, switched LQR and the Kuramoto oscillator.

Policy performance and convergence. Figure 2a shows the evolution of the gap $|J - J^\star|/|J^\star|$ at each epoch on both LQR settings. Grey lines give the best value of $J(\theta)$ found by multistart optimization over each class of affine-mean Gaussian policies with radius κ_π (a local optimum in general, see Paragraph C.1). We observe that LTC/NTC matches or outperforms other baselines across various configurations (κ, κ_π). In particular, on both LQR settings and all configurations (κ, κ_π) but one, LTC/NTC converge to the maximum value mentioned above. Similarly to Figure 1, we observe that for $\kappa_\pi \geq 2$ the performance improves significantly from $n = 1$ to $n = 2$ (Figure 6 in Appendix D). Figure 2b shows the evolution of the undiscounted return at each epoch on Kuramoto. Importantly, NTC matches or outperforms LTC, highlighting the benefit of more expressivity under nonlinear dynamics. Furthermore, in this setting, our methods converge near the greedy controller's value. It is also noteworthy that for both LQR benchmarks, LTC ($n = 2$) matches or beats the baselines with few learnable parameters. For instance, LTC ($n = 2$) matches the MLP baseline on LQR and systematically outperforms it on switched LQR, while learning 231 coefficients per agent at $\kappa = 2$ against 17,409 for the MLP (Table 2). Finally, our model-free approach systematically matches or outperforms prior spectral baselines across all three environments despite not having access to the transition kernel. Appendix D.4 provides empirical evidence that these policy improvements translate into the intended behavior in the resulting trajectories.

Scalability. Figure 3 shows the predicted $O(N)$ runtime scaling and near-reference performance up to $N = 1000$; in this experiment LTC outperforms RF in runtime.

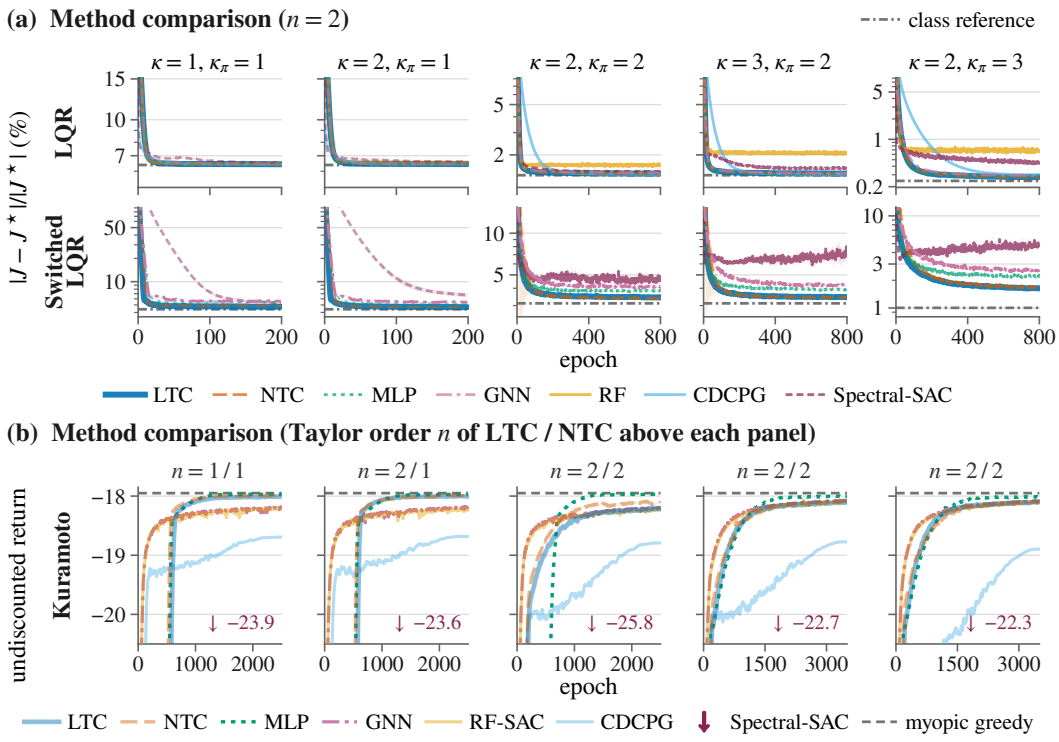

Figure 2: **Policy convergence**. Method comparison LQR and switched LQR (relative optimality gap), and policy convergence on Kuramoto (undiscounted return).

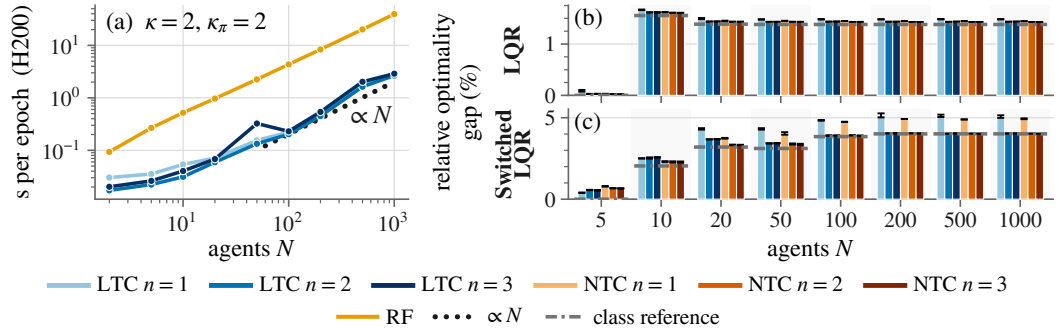

Figure 3: **Scalability. Left:** runtime per epoch on an H200 GPU as a function of the number of agents N. **Right:** relative optimality gap as a function of N on LQR and switched LQR.

7 Conclusion and Limitations

We showed that graph locality turns otherwise ambient-dimensional Taylor representations into scalable local critics without the transition-model access required by prior constructive spectral approaches. This yields a model-free actor–critic for continuous networked MDPs applicable beyond prior frameworks. While experiments show competitive, scalable performance, some limitations also suggest directions for future work. Our LTC guarantee assumes well-conditioned LSTD systems, whose characterization requires complementary analysis. Moreover, the Taylor-coefficient function class is unknown a priori, precluding sharp control of $\bar{\varepsilon}_{\text{lin}}$ and requiring empirical feature selection. The policy-optimization guarantee reaches a stationary point of J, which may not be a global maximum. K-medians can also be challenging in high-dimensional state spaces. Finally, extending finite-sample critic guarantees to NTC is an important but more challenging direction.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing (grammar, syntax...), LaTeX drafting, to assist with debugging as well as to assist with retrieval and discovery of related relevant literature. We have not used generative AI tools for research ideation or execution, generating synthetic datasets nor for proving mathematical claims. We have reviewed all AI-assisted work. We have reviewed all references provided by AI and made sure of their relevance for our work. We also reviewed and validated our code. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The empirical study uses simulated control benchmarks and does not involve human subjects or personal data. The work is methodological and theoretical; any deployment of the proposed control methods in safety-critical systems would require application-specific validation beyond the scope of this paper.

REPRODUCIBILITY STATEMENT

Several parts of this article allow for reproducibility. The algorithmic specification is given in Section 5. Appendix B, *"Definition of the environments"*, provides the benchmark definitions. Appendix C, *"Training and Evaluation Protocols"*, collects the evaluation protocol, implementation and training details. Complete proofs of the main claims are provided in Appendix F, *"Proofs of the Main-Text Results"*, and Appendix G, *"Proofs of the Additional Appendix Results"*. Appendix H, *"Neural Network Implementation Details"*, specifies the neural network architectures and training procedures, including those used for NTC.

REFERENCES

Christopher G. Atkeson, Andrew W. Moore, and Stefan Schaal. Locally weighted learning. *Artificial Intelligence Review*, 11(1–5):11–73, 1997. doi: 10.1023/A:1006559212014.

Steven J. Bradtke and Andrew G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1–3):33–57, 1996. doi: 10.1007/BF00114723.

Anton Braverman, Itai Gurvich, and Junfei Huang. On the taylor expansion of value functions. *Operations Research*, 68(2):631–654, 2020. doi: 10.1287/opre.2019.1903.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995. doi: 10.1137/0916069.

Bruno Didier Olivier Capron and Darci Odloak. An extended linear quadratic regulator with zone control and input targets. *Journal of Process Control*, 29:33–44, 2015. doi: 10.1016/j.jprocont.2015.03.005.

Hervé Cardot, Peggy Cénac, and Jean-Marie Monnez. A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics & Data Analysis*, 56(6):1434–1449, 2012. doi: 10.1016/j.csda.2011.11.019.

Luis Carvalho, Jozias R. L. Neto, Jefferson C. Rezende, Marcus V. S. Costa, Elenilson V. Fortes, and Leonardo H. Macedo. Linear quadratic regulator design via metaheuristics applied to the damping of low-frequency oscillations in power systems. *ISA Transactions*, 134:322–335, 2023. doi: 10.1016/j.isatra.2022.08.024.

Tianshu Chu, Sandeep Chinchali, and Sachin Katti. Multi-agent reinforcement learning for networked system control. In *International Conference on Learning Representations*, 2020.

Oswaldo Luiz Valle Costa, Ricardo Paulino Marques, and Marcelo Dutra Fragoso. *Discrete-Time Markov Jump Linear Systems*. Probability and Its Applications. Springer, 2005. doi: 10.1007/b138575.

Florian Dörfler and Francesco Bullo. Synchronization and transient stability in power networks and non-uniform kuramoto oscillators. *SIAM Journal on Control and Optimization*, 50(3):1616–1642, 2012. doi: 10.1137/110851584.

Florian Dörfler and Francesco Bullo. Synchronization in complex networks of phase oscillators: A survey. *Automatica*, 50(6):1539–1564, 2014. doi: 10.1016/j.automatica.2014.04.012.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1996.

Maryam Fazel, Rong Ge, Sham M. Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1466–1475. PMLR, 2018.

W. Steven Gray, Hong Zhang, and Oscar R. González. Closed-loop performance measures for flight controllers subject to neutron-induced upsets. In *Proceedings of the 42nd IEEE Conference on Decision and Control*, volume 3, pp. 2465–2470, 2003. doi: 10.1109/CDC.2003.1272990.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field multiagent reinforcement learning: A decentralized network approach. *Mathematics of Operations Research*, 50(1):506–536, 2025. doi: 10.1287/moor.2022.0055.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

Davorin Hrovat. Survey of advanced suspension developments and related optimal control applications. *Automatica*, 33(10):1781–1817, 1997. doi: 10.1016/S0005-1098(97)00101-5.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2137–2143, 2020.

Rudolf E. Kalman. Contributions to the theory of optimal control. *Boletin de la Sociedad Matematica Mexicana*, 5:102–119, 1960.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1008–1014, 1999.

Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics*, volume 39 of *Lecture Notes in Physics*, pp. 420–422. Springer, 1975. doi: 10.1007/BFb0013365.

Alessandro Lazaric, Mohammad Ghavamzadeh, and Rémi Munos. Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research*, 13:3041–3074, 2012.

Hao Liang, Shuqing Shi, Yudi Zhang, Biwei Huang, and Yali Du. Causality meets locality: Provably generalizable and scalable policy learning for networked systems. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5311.

Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. Multi-agent reinforcement learning in stochastic networked systems. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7825–7837, 2021.

Zhixin Lu, Kevin Klein-Cardeña, Steven Lee, Thomas M. Antonsen, Michelle Girvan, and Edward Ott. Resynchronization of circadian oscillators and the east-west asymmetry of jet-lag. *Chaos*, 26(9):094811, 2016. doi: 10.1063/1.4954275.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restelli. Local linearity: The key for no-regret reinforcement learning in continuous MDPs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 75986–76029, 2024a. doi: 10.52202/079017-2419.

Davide Maran, Alberto Maria Metelli, Matteo Papini, and Marcello Restelli. No-regret reinforcement learning in smooth MDPs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34760–34789. PMLR, 2024b.

Sen Na and Mihai Anitescu. Exponential decay in the sensitivity analysis of nonlinear dynamic programming. *SIAM Journal on Optimization*, 30(2):1527–1554, 2020. doi: 10.1137/19M1265065.

Sen Na, Sungho Shin, Mihai Anitescu, and Victor M. Zavala. On the convergence of overlapping schwarz decomposition for nonlinear optimal control. *IEEE Transactions on Automatic Control*, 67(11):5996–6011, 2022. doi: 10.1109/TAC.2022.3194087.

Masaki Ogura, Ahmet Cetinkaya, Tomohisa Hayakawa, and Victor M. Preciado. State-feedback control of markov jump linear systems with hidden-markov mode observation. *Automatica*, 89: 65–72, 2018. doi: 10.1016/j.automatica.2017.11.022.

Johan Olsson, Runyu Zhang, Emma Tegling, and Na Li. Scalable reinforcement learning for linear-quadratic control of networks. In *2024 American Control Conference*, pp. 1813–1818, 2024. doi: 10.23919/ACC60939.2024.10644413.

Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, 22(4):1679–1706, 1994. doi: 10.1214/aop/1176988477.

Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable multi-agent reinforcement learning for networked systems with average reward. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2074–2086, 2020a.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 256–266. PMLR, 2020b.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 70(6):3601–3628, 2022. doi: 10.1287/opre.2021.2226.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184, 2007.

Tongzheng Ren, Tianjun Zhang, Lisa Lee, Joseph E. Gonzalez, Dale Schuurmans, and Bo Dai. Spectral decomposition representation for reinforcement learning. In *International Conference on Learning Representations*, 2023.

Zhaolin Ren, Tongzheng Ren, Haitong Ma, Na Li, and Bo Dai. Stochastic nonlinear control via finite-dimensional spectral dynamics embedding. *IEEE Transactions on Automatic Control*, 70 (12):7907–7922, 2025a. doi: 10.1109/TAC.2025.3580296.

Zhaolin Ren, Runyu Zhang, Bo Dai, and Na Li. Scalable spectral representations for multiagent reinforcement learning in network MDPs. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 550–558. PMLR, 2025b.

Sungho Shin, Mihai Anitescu, and Victor M. Zavala. Exponential decay of sensitivity in graph-structured nonlinear programs. *SIAM Journal on Optimization*, 32(2):1156–1183, 2022. doi: 10.1137/21M1391079.

Hisham Soliman and Muhammad Shafiq. Robust stabilisation of power systems with random abrupt changes. *IET Generation, Transmission & Distribution*, 9(15):2159–2166, 2015. doi: 10.1049/iet-gtd.2014.1111.

Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988. doi: 10.1007/BF00115009.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018.

Richard S. Sutton, David A. McAllester, Satinder P. Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pp. 1057–1063, 1999.

Dongming Wang, Pengcheng Dai, Wenwu Yu, and Wei Ren. Scalable policy optimization for networked multi-agent reinforcement learning with continuous state-action spaces. *arXiv preprint arXiv:2607.18554*, 2026.

Kurt Wiesenfeld, Pere Colet, and Steven H. Strogatz. Frequency locking in josephson arrays: Connection with the kuramoto model. *Physical Review E*, 57(2):1563–1569, 1998. doi: 10.1103/PhysRevE.57.1563.

Li Xie and Lihua Xie. Stabilizing sampled-data linear systems with markovian packet losses and random sampling. In *Proceedings of the 17th IFAC World Congress*, volume 41, pp. 599–604, 2008. doi: 10.3182/20080706-5-KR-1001.00101.

Junhong Xu, Kai Yin, and Lantao Liu. Kernel taylor-based value function approximation for continuous-state markov decision processes. In *Robotics: Science and Systems*, 2020. doi: 10.15607/RSS.2020.XVI.050.

Yaguang Yang. Analytic LQR design for spacecraft control system based on quaternion model. *Journal of Aerospace Engineering*, 25(3):448–453, 2012. doi: 10.1061/(ASCE)AS.1943-5525.0000142.

Xunyuan Yin, Lixian Zhang, Zepeng Ning, Dapeng Tian, Ahmed Alsaedi, and Bashir Ahmad. State estimation via markov switching-channel network and application to suspension systems. *IET Control Theory & Applications*, 11(3):411–419, 2017. doi: 10.1049/iet-cta.2016.1108.

Yizhou Zhang, Guannan Qu, Pan Xu, Yiheng Lin, Zaiwei Chen, and Adam Wierman. Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–51, 2023. doi: 10.1145/3579443.

Pixuan Zhou, Fei-Yue Wang, Weinong Chen, and Paul Lever. Optimal construction and control of flexible manipulators: A case study based on LQR output feedback. *Mechatronics*, 11(1):59–77, 2001. doi: 10.1016/S0957-4158(00)00004-0.

A Appendix

Overview Appendix B, *"Definition of the environments"*, describes the three benchmark environments and the role played by each.

Appendix C, *"Training and Evaluation Protocols"*, gives the evaluation protocol, implementation details, and hyperparameters of Taylor Centroid and of the competing baselines.

Appendix D, *"Further Experimental Results"*, reports the additional experiments and ablations.

Appendix E, *"Understanding the Expected Reward Regularity Assumption"*, studies the regularity condition used throughout the theoretical analysis. It gives its interpretation, describes a rollout-based empirical stress test, presents a weaker discounted-summability formulation, and provides concrete sufficient conditions under which the assumption holds. It also positions the condition relative to the smooth-MDP hierarchy of Maran et al. (2024a;b).

Appendix F, *"Proofs of the Main-Text Results"*, contains the complete proofs of the results stated in the main text. It proceeds through finite information propagation and the exponential derivative decay, the localized Taylor Centroid approximation, the finite-sample LTC–LSTD guarantee, the localized policy-gradient and one-step TD analysis, the end-to-end policy-optimization guarantee, and the computational-complexity result.

Appendix G, *"Proofs of the Additional Appendix Results"*, contains the proofs of the auxiliary results used in Appendix E, *"Understanding the Expected Reward Regularity Assumption"*. These include the discounted-summability extension, the sufficient conditions based on smooth stochastic

and spectral transition models, and the comparison with the smoothness assumptions of Maran et al. (2024a;b).

Finally, Appendix H, *"Neural Network Implementation Details"*, describes the implementation of neural network methods, including NTC.

B Definition of the environments

This appendix provides the complete description and motivation of the experimental benchmarks used in Section 6. The three environments are deliberately complementary. Networked LQR provides a controlled setting in which the exact critic is quadratic and the assumptions underlying the constructive spectral baseline are satisfied. It is also a canonical environment for the random Fourier feature setting studied by Ren et al. (2025b). Switched LQR introduces potentially unobserved operating regimes through σ_t and therefore tests the ability of the representation to exploit regime structure without access to the transition model. Finally, the Kuramoto oscillator removes the linear–quadratic structure altogether and tests Taylor Centroid critics on nonlinear graph-coupled dynamics.

Remark. As in Ren et al. (2025b), although some of the benchmarks are naturally defined on unbounded state spaces, over any finite experimental horizon their trajectories lie in a sufficiently large compact set with arbitrarily high probability. One can therefore view the experiments as evolving inside such a compact set; equivalently, the dynamics could be clipped at its boundary without affecting the reported trajectories except on an arbitrarily low-probability event.

B.1 Environment 1: Networked linear–quadratic regulation

Environment. We first consider a networked linear–quadratic regulator (LQR) (Kalman, 1960). Here, we consider a ring of N scalar subsystems, with dynamics

$$s_{t+1} = As_t + Ba_t + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_N), \qquad \gamma = 0.75 \tag{8}$$

over an episode length of $T = 20$. The matrices A and B are cyclic and banded, so each subsystem interacts only with its two immediate neighbors. With indices understood modulo N, the nonzero entries are

$$A_{ii} = 0.8, \qquad A_{i,i\pm1} = 0.1, \qquad B_{ii} = 1, \qquad B_{i,i\pm1} = 0.7.$$

All remaining entries are zero. The local stage reward is

$$r_i(s_i, a_i) = -(3s_i^2 + a_i^2). \tag{9}$$

Thus, agents are encouraged to minimize their state value. LQR is one of the classical models of optimal control and is routinely used either directly or as a local approximation of more complicated physical systems. Applications include spacecraft attitude control (Yang, 2012), automotive suspension (Hrovat, 1997), robotic manipulators (Zhou et al., 2001), industrial process control (Capron & Odloak, 2015), and power-system stabilization (Carvalho et al., 2023). In many such applications, the LQR model results from linearizing nonlinear dynamics around a nominal operating point. This environment was also used as a benchmark in Ren et al. (2025b)'s released code.

Role of LQR for the benchmark. This benchmark serves two purposes. First, under our policy class the true value function is exactly quadratic. It therefore provides a controlled test of the Taylor mechanism.

Second, Equation 8 is a known local dynamics map with additive, state-independent Gaussian noise, precisely the constructive case for which both the random Fourier features of Ren et al. (2025b) and CDCPG of Wang et al. (2026) are naturally defined. LQR therefore provides a comparison in which the spectral method is evaluated inside its intended model class, while Taylor Centroid uses only sampled trajectories.

B.2 Environment 2: Switched LQR with an unobserved operating mode

Many engineered systems do not evolve around a single stationary operating point. Their dynamics can change abruptly because of faults, topology changes, communication conditions, workloads, or external operating modes. Switched LQR systems with Markovian switching provide a classical model for such phenomena (Costa et al., 2005). They have been used, for example, in fault-tolerant flight control (Gray et al., 2003), power systems subject to abrupt faults or topology changes (Soliman & Shafiq, 2015), automotive systems with switching communication conditions (Yin et al., 2017), and networked control under correlated packet losses (Xie & Xie, 2008).

This setting is particularly relevant to the motivation behind Taylor Centroid critics: a physical system may repeatedly move between a small number of qualitatively different operating regimes even though the underlying state space is continuous. This is the motivation behind the centroids around which our Taylor expansions are performed.

Environment. We consider the same sparse ring structure as in the LQR benchmark, but let a persistent mode

$$Z_t := \sigma_t \in \{1, \dots, M_{\text{mode}}\}, \qquad M_{\text{mode}} = 5,$$

modify the dynamics:

$$s_{t+1} = A^{(\sigma_t)} s_t + B^{(\sigma_t)} a_t + C^{(\sigma_t)} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, I_N). \tag{10}$$

over an episode length of $T = 20$. Each of the five matrices induces graph-local interactions, so switching changes the local physical behavior without changing the interaction graph.

The mode follows an autonomous Markov chain: at each step, with probability h, the mode is redrawn (possibly to its current value). When redrawing, σ' is drawn with probability $(6 - \sigma')/15$. Hence

$$\Pr(\sigma_{t+1} = \sigma' \mid \sigma_t = \sigma) = \begin{cases} 1 - h + h\,\dfrac{6-\sigma}{15}, & \sigma' = \sigma, \\ h\,\dfrac{6-\sigma'}{15}, & \sigma' \neq \sigma. \end{cases} \tag{11}$$

We choose $h \approx 0.139$ so that a length-20 trajectory contains two mode changes on average. The local rewards and discount factor are identical to those of the LQR benchmark. Our main experiments keep σ_t completely *unobserved*: neither the actor nor the critic receives the current mode.

The constructive RF and CDCPG baselines are not directly applicable here. The distinction between observed and unobserved switching is important. Conditional on a known mode $\sigma_t = \sigma$, Equation 10 is indeed linear–Gaussian. However, in our experiments, the controller does not observe σ_t. After marginalizing the persistent mode, there is no single observable transition of the form

$$s_{t+1} = f(s_t, a_t) + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma) \quad \text{independently over time}, \tag{12}$$

with a known deterministic local map f and fixed homoskedastic noise.

This matters directly for the constructive spectral methods of Ren et al. (2025b) and Wang et al. (2026). Both approaches build their local critic representations from explicit random features of the transition kernel and require access to the local dynamics map under structured additive noise. We therefore do not include the constructive RF baseline of Ren et al. (2025b) nor CDCPG of Wang et al. (2026) on this benchmark.

Role of switched LQR for the benchmark. Switched LQR highlights the flexibility of the model-free approach by evaluating Taylor Centroid in a setting where the dynamics are unknown to the learner. It also tests whether the method can extract from rollout data the information relevant for critic estimation and control when important aspects of the environment, such as the operating regime, are not directly observed.

B.3 Environment 3: Kuramoto oscillator network

Our final benchmark removes the linear–quadratic structure entirely. The Kuramoto model (Kuramoto, 1975; Dörfler & Bullo, 2014) is a canonical model of synchronization in networks of coupled oscillators. Closely related models arise in electrical power networks (Dörfler & Bullo, 2012),

Josephson-junction arrays (Wiesenfeld et al., 1998), and biological oscillator networks (Lu et al., 2016).

Following the implementation of the Kuramoto benchmark of Ren et al. (2025b), the state of oscillator i is represented by

$$s_i = (s_{i,1}, s_{i,2}) = (\cos\theta_i, \sin\theta_i) \in \mathbb{R}^2.$$

The phase is recovered as

$$\theta_i = \mathrm{atan2}(s_{i,2}, s_{i,1}),$$

where $\mathrm{atan2}(y, x) \in (-\pi, \pi]$ is the polar angle of the vector (x, y). Its pulsation is

$$\dot{\theta}_i = \omega_i + a_i - \sum_{j \in \mathcal{N}_i} P_{ij} \sin(\theta_i - \theta_j),$$

where ω_i is the natural frequency, a_i is the control input, and P_{ij} is the coupling strength. The state dynamics are then

$$\begin{aligned} s_{i,1}^{+} &= \cos\Big(\theta_i + dt\,\dot{\theta}_i\Big) + \sigma\xi_{i,1}, \\ s_{i,2}^{+} &= \sin\Big(\theta_i + dt\,\dot{\theta}_i\Big) + \sigma\xi_{i,2}, \qquad (\xi_{i,1}, \xi_{i,2}) \sim \mathcal{N}(0, I_2). \end{aligned} \tag{13}$$

As in Ren et al. (2025b), we clip the two state coordinates to $[-1, 1]$ in the implementation. We additionally impose a negligible radial safeguard before evaluating $\mathrm{atan2}$ to avoid its singularity at the origin. These are purely numerical safeguards and are essentially never active in practice, since each deterministic update lies on the unit circle and the process noise is small. We use $dt = 0.01$, $\gamma = 0.99$, and episodes of length $T = 500$. Following the released code of Ren et al. (2025b), the oscillators form a ring, the natural frequencies $\omega_i \sim \mathcal{U}[-0.5, 0.5]$ and the symmetric couplings $P_{ij} \sim \mathcal{U}[0.2, 1.2]$ are drawn once per network size and shared by all runs, $\omega^\star = 0.2$, $\sigma = 0.0025$, and actions take values in $[-1, 1]$. Thus the Kuramoto dynamics are evolved in phase coordinates, mapped back through $(\cos, \sin)$, and then perturbed by isotropic Gaussian noise directly in the two state coordinates.

The control objective is to synchronize the oscillators at a prescribed angular frequency $\omega^\star$. Since $\dot{\theta}_i$ is precisely the instantaneous angular frequency of oscillator i, we use the smooth frequency-mismatch reward

$$r_i(s_{\mathcal{N}_i}, a_i) = -\sqrt{(\dot{\theta}_i - \omega^\star)^2 + \varepsilon^2},$$

where $\varepsilon = 10^{-6}$ smooths the absolute-value loss; Spectral-SAC and RF-SAC keep $\varepsilon = 0$, as in Ren et al. (2025b)'s code. Maximizing this reward therefore drives $\dot{\theta}_i$ toward $\omega^\star$. This reward depends on the neighbors of agent i, a case covered by the remark of Section 2.

Role of the Kuramoto oscillator for the benchmark. Unlike LQR, the transition dynamics contain nonlinear sinusoidal interactions and the critic is non-quadratic. The benchmark therefore tests whether a *local* Taylor representation remains useful when the value function is more complex than in LQR. It also provides a natural stress test of the regime mechanism. Because the true critic is non-polynomial, the number of clusters plays a significant role.

C Training and Evaluation Protocols

This appendix gives the training, implementation, and evaluation details for the experiments of Section 6. Environment definitions and their role in the benchmark are given in Appendix B. Detailed neural architectures are given separately in Appendix H.

C.1 Evaluation protocols

Experimental outline. We compare LTC and NTC with five reference baselines whenever applicable: (i) the constructive random-feature (RF) algorithm of Ren et al. (2025b), (ii) their model-free Spectral-SAC alternative, (iii) CDCPG of Wang et al. (2026), (iv) a local MLP critic with our one-step TD actor, and (v) a GNN critic paired with the actor of Ren et al. (2025b) on LQR and Kuramoto (SAC) and our actor on Switched LQR. The first three baselines compare Taylor Centroid

against methods designed specifically for networked MARL. The MLP provides a generic function-approximation baseline while the GNN uses information from the graph. For the Kuramoto oscillator, we use the SAC variant of RF (RF-SAC) following the approach of Ren et al. (2025b). Unless stated otherwise, we use $N = 50$ agents on LQR and switched LQR and $N = 40$ on Kuramoto.

Critic estimation. Figure 1 reports the normalized MSE of the estimated global value $V(s) = N^{-1} \sum_i V_i(s)$,

$$100 \, \frac{\mathbb{E}[(\widetilde{V}(S) - V(S))^2]}{\mathrm{Var}(V(S))},$$

evaluated on an independent rollout. For LQR, the target is the exact V^π; for switched LQR, it is the exact mode-conditioned value (i.e. the value function that can observe σ_t and which we want to recover); and for Kuramoto it is estimated by Monte Carlo.

For each environment and (κ, κ_π), we first train a policy with Algorithm 1 for 80 epochs on LQR, 120 on switched LQR, and 20 on Kuramoto, and then fit the critic with this policy. For each configuration, we train five critics with independent random seeds and evaluate each of them on a separate rollout.

Policy performance. For LQR and switched LQR, Figure 2 reports the relative optimality gap

$$100 \, \frac{|J - J^\star|}{|J^\star|},$$

where J is the expected sum of discounted rewards of the deterministic policy. For LQR, $J^\star$ is given by the discounted Riccati solution; for switched LQR, where σ_t is unobserved, it is obtained by directly maximizing J over mode-blind affine controllers with circulant gains. Each κ_π-class reference is similarly obtained by multistart L-BFGS-B (Byrd et al., 1995) on the exact J and its exact gradient (483 starts per κ_π), and is a strict local maximum (vanishing gradient, negative-definite Hessian on the admissible gains). Structured static feedback is non-convex, so these values, and $J^\star$ on switched LQR, are only lower bounds on the corresponding optima. For Kuramoto, where $J^\star$ is not available in closed form, we instead report the undiscounted return of the deterministic policy, following Ren et al. (2025b), with the myopic greedy controller (which chooses the actions so as to cancel the instantaneous frequency error of each oscillator) as reference. Results are averaged over five independent seeds on LQR and switched LQR, and over three seeds on Kuramoto.

We evaluate the following configurations

$$(\kappa, \kappa_\pi) \in \{(1,1), (2,1), (2,2), (3,2), (2,3)\},$$

covering $\kappa < \kappa_\pi$, $\kappa = \kappa_\pi$, and $\kappa > \kappa_\pi$. We note that CDCPG requires $\kappa \geq \kappa_\pi + 1$ for its policy-gradient estimator (Wang et al., 2026). Hence, for CDCPG only, we use $\kappa' = \max\{\kappa, \kappa_\pi + 1\}$, the smallest radius allowed by its analysis, while keeping κ_π and therefore the policy class unchanged across methods. For this figure, we use $N = 50$ for LQR and switched LQR and $N = 40$ for the Kuramoto oscillator.

Scalability. Runtime measurements use a single dedicated NVIDIA H200 GPU at $(\kappa, \kappa_\pi) = (2, 2)$ and exclude the first epoch. Policy-scaling experiments use three seeds, up to $N = 1000$ for LQR and switched LQR. Performance is compared with $J^\star$ and with the $\kappa_\pi = 2$ class reference described above.

C.2 Common actor and training protocol

On LQR, we use a Gaussian policy with linear mean

$$\mu_\theta(s) = Fs,$$

and fixed covariance $\sigma_{\mathrm{pol}}^2 I$, where the ith row of F depends only on $s_{\mathcal{N}_i^{\kappa_\pi}}$. On switched LQR, we instead use a Gaussian policy with affine mean

$$\mu_\theta(s) = Fs + f,$$

with the same locality restriction on F and fixed covariance $\sigma_{\text{pol}}^2 I$. The switching mode σ_t remains unobserved by both the actor and the critic. In both LQR environments, $\sigma_{\text{pol}} = 0.3$. On Kuramoto, each agent uses a Gaussian policy with local affine mean $\mu_{\theta_i}(o_i)$, and actions are sampled as

$$a_i = \text{clip}\big(\mu_{\theta_i}(o_i) + \sigma_{\text{pol}}\varepsilon_i, -1, 1\big), \qquad \varepsilon_i \sim \mathcal{N}(0,1),$$

where σ_{pol} decreases log-linearly from 0.1 to 0.012 over the first 55% of iterations, then stays constant. Actor updates use $\eta_m = \eta/\sqrt{m+1}$; the policy gradient is normalized to unit norm on LQR and Kuramoto and clipped at norm 1 on switched LQR. The runs of Figure 2 last 200 actor iterations for $\kappa_\pi = 1$ and 800 for $\kappa_\pi \geq 2$ on LQR and switched LQR, and 2500 on Kuramoto (3500 at $(3,2)$ and $(2,3)$).

Advantages are computed using the one-step TD estimator described in Section 4. Critic and actor rollouts are sampled independently. Unless stated otherwise, each LQR and switched-LQR iteration uses 2500 episodes in total, split between fitting the critic and updating the actor. LTC, NTC, MLP, and RF use 1250 episodes for each. The GNN uses $1670/830$ critic/actor episodes for $\kappa \leq 2$ and $1750/750$ for $\kappa = 3$. On Kuramoto, each iteration uses 300 critic (raised when needed to reach 50 critic transitions per critic parameter) and 2000 actor episodes of length 500; Spectral-SAC, RF-SAC, and the GNN keep the training schedule of the released code of Ren et al. (2025b).

C.3 Method-specific details

Architectures and training objectives of the neural critics are described in Appendix H.

Taylor Centroid critics. LTC and NTC use all monomials of order $n \in \{1,2,3\}$ supported on $\mathcal{N}_i^\kappa$, expanded around K-medians centroids. LTC is fitted by LSTD with ridge $\lambda = 10^{-6}$ on LQR and switched LQR, and with the ridge of Table 5 on Kuramoto.NTC keeps the same basis, with its own number of centroids (Tables 3–5), and replaces the affine coefficient map by a neural network trained by semi-gradient TD. On LQR and switched LQR, its value estimates average 16 action samples, and it uses 1500 TD updates in the standalone critic-estimation experiment.

Generic critics. The MLP critic is paired with the same model-free actor as Taylor Centroid. It has two hidden layers of width 128 on LQR and switched LQR and 64 on Kuramoto, and is trained with Adam (Kingma & Ba, 2015) at learning rate 2×10^{-3}, using 200 TD updates per iteration on batches of 4096 transitions (50 updates on batches of 1024 on Kuramoto) and a target network synchronized every 25 updates; on LQR and switched LQR, gradients are clipped at norm 1. The GNN critic, whose receptive field matches κ on LQR and switched LQR and is $\min\{\kappa, 2\}$ on Kuramoto, is paired with the actor of the RF baseline on LQR and Kuramoto (and ours on switched LQR): RF's model-based policy gradient with stepsize $\eta = 0.2$ on LQR, and the SAC training loop of RF-SAC on Kuramoto, where the GNN replaces the fixed random-feature map of its critic. On LQR and switched LQR, it has $\max(2,\kappa)$ layers of width 128 shared across agents and it is trained like the MLP (200 TD updates, batches of up to 4096 transitions, target synchronization every 25 updates, gradient clipping at 1); on Kuramoto, it has width 64 and follows the RF-SAC training loop.

Networked-MARL baselines. Following (Ren et al., 2025b), on LQR, RF uses 200 cosine features at $\kappa = 1$ and 800 at $\kappa \geq 2$, with actor stepsize $\eta = 0.2$; on Kuramoto, its SAC variant RF-SAC uses 1024 features. Spectral-SAC uses 512 random features, hidden width 256, batch size 128, 128 parallel environments, one update per environment step, and automatic entropy tuning with initial coefficient 0.1. On LQR and switched LQR, its actor and critic learning rates are 2×10^{-4}, its actor is linear and unbounded with action scale 3, and its replay capacity is 10^7 for $\kappa_\pi = 1$ and 4×10^7 for $\kappa_\pi \geq 2$. On Kuramoto, its actor is a tanh-Gaussian MLP with two hidden layers of width 256 (three for RF-SAC and the GNN), its replay capacity is 2×10^7, and the learning rates are 10^{-3} for $(\kappa,\kappa_\pi) \in \{(1,1),(2,1),(2,2)\}$ and, at $(3,2)$ and $(2,3)$, 5×10^{-4} for Spectral-SAC and 3×10^{-4} for RF-SAC and the GNN. On LQR, CDCPG uses ridge 10^{-4}, policy standard deviation 0.3, $\theta_{\max} = 1$ and actor stepsize 3×10^{-4}; its 12,500 trajectories per iteration, of mean length $1/(1-\gamma) = 4$, match our 2500×20 transitions. On Kuramoto, it uses policy standard deviation 0.05, $\theta_{\max} = 1.5$ and stepsize 10^{-5}. Its critic radius is $\max\{\kappa, \kappa_\pi + 1\}$, the smallest value allowed by its analysis, while κ_π, and hence the policy class, is unchanged.

Parameter count Table 2 reports the number of learned parameters of every critic.

Table 2: Learned parameters of each critic in the policy experiments of Figure 2, per agent unless marked as total ($N = 50$ on LQR and switched LQR, $N = 40$ on Kuramoto). LTC counts its coefficients, RF, RF-SAC and CDCPG their linear weights, and the other critics their network weights. The GNNs share their weights across agents and are reported as totals; twin critics count both copies.

(κ, κ_π)	$(1,1)$	$(2,1)$	$(2,2)$	$(3,2)$	$(2,3)$
LQR and switched LQR					
LTC, $n = 1/2/3$	28/70/140	66/231/616	66/231/616	120/540/1,800	66/231/616
NTC, $n = 1$	4,932	5,318	5,318	5,704	5,318
NTC, $n = 2$	5,322	6,293	6,293	7,524	6,293
NTC, $n = 3$	5,972	8,568	8,568	12,984	8,568
MLP	17,153	17,409	17,409	17,665	17,409
GNN (total)	17,025	33,409	33,409	66,305	33,409
RF (LQR)	200	800	800	800	800
CDCPG (LQR)	120	120	190	190	276
Spectral-SAC	200,192	201,216	201,216	202,240	201,216
Kuramoto					
LTC, $n = 1$	70	176	176	330	176
LTC, $n = 2$	280	1,056	1,056	2,640	1,056
LTC, $n = 3$	840	4,576	4,576	14,960	4,576
NTC, $n = 1$	5,319	5,963	5,963	6,607	5,963
NTC, $n = 2$	6,684	9,538	9,538	13,432	9,538
NTC, $n = 3$	10,324	23,838	23,838	49,832	23,838
MLP	4,673	4,929	4,929	5,185	4,929
GNN (total)	17,282	17,282	17,282	17,282	17,282
RF-SAC	4,096	4,096	4,096	4,096	4,096
CDCPG	257	257	257	257	257
Spectral-SAC	200,960	202,496	202,496	204,032	202,496

C.4 Hyperparameter selection

Tables 3–5 report the selected hyperparameters. On both LQR settings, the GNN keeps the fixed stepsize $\eta = 0.2$, while we follow the configuration designs of Ren et al. (2025b) for RF, RF-SAC and Spectral-SAC.

Table 3: Selected (η, K) on LQR.

(κ, κ_π)	$(1,1)$	$(2,1)$	$(2,2)$, $(3,2)$	$(2,3)$
LTC $n = 1$	$(0.50, 40)$	$(0.35, 40)$	$(0.50, 100)$	$(0.50, 100)$
LTC $n = 2$	$(0.35, 25)$	$(0.35, 25)$	$(0.50, 40)$	$(0.50, 40)$
LTC $n = 3$	$(0.35, 25)$	$(0.35, 40)$	$(0.50, 40)$	$(0.50, 40)$
NTC $n = 1$	$(0.50, 400)$	$(0.50, 400)$	$(0.50, 2000)$	$(0.50, 2000)$
NTC $n = 2$	$(0.35, 40)$	$(0.35, 25)$	$(0.50, 40)$	$(0.50, 40)$
NTC $n = 3$	$(0.35, 40)$	$(0.35, 25)$	$(0.50, 25)$	$(0.50, 25)$
MLP η	0.35	0.35	0.50	0.50

Table 4: Selected (η, K) on switched LQR.

(κ, κ_π)	$(1,1)$	$(2,1)$	$(2,2)$, $(3,2)$	$(2,3)$
LTC $n = 1$	$(0.35, 40)$	$(0.35, 40)$	$(0.50, 40)$	$(0.50, 40)$
LTC $n = 2$	$(0.50, 25)$	$(0.50, 25)$	$(0.50, 25)$	$(0.50, 25)$
LTC $n = 3$	$(0.50, 25)$	$(0.50, 25)$	$(0.50, 25)$	$(0.50, 25)$
NTC $n = 1$	$(0.35, 40)$	$(0.35, 40)$	$(0.50, 40)$	$(0.50, 40)$
NTC $n = 2$	$(0.50, 40)$	$(0.50, 40)$	$(0.50, 25)$	$(0.50, 40)$
NTC $n = 3$	$(0.50, 40)$	$(0.50, 40)$	$(0.50, 40)$	$(0.50, 40)$
MLP η	0.35	0.35	0.50	0.50

Table 5: Selected hyperparameters on Kuramoto, at the Taylor order n shown in each panel of Figure 2; λ is the LSTD ridge of LTC.

(κ, κ_π)	$(1,1)$	$(2,1)$	$(2,2)$	$(3,2)$	$(2,3)$
LTC n	1	2	2	2	2
LTC (η, K)	$(0.25, 16)$	$(0.25, 16)$	$(0.50, 4)$	$(0.50, 16)$	$(0.50, 16)$
LTC λ	10^{-3}	10^{-1}	10^{-3}	10^{-1}	10^{-1}
NTC n	1	1	2	2	2
NTC (η, K)	$(0.25, 16)$	$(0.25, 16)$	$(0.50, 16)$	$(0.50, 16)$	$(0.50, 16)$
MLP η	0.25	0.25	0.25	0.50	0.50

D Further Experimental Results

This appendix provides additional experiments supporting the structural claims of the paper. We first verify empirically the exponential decay of critic derivatives predicted by Lemma 1. We then study the main design choices of the Taylor Centroid representation: the number and location of expansion points, the dependence of approximation error on distance to the assigned centroid, and the effect of the Taylor order. We next report representative trajectories on all three environments to complement the aggregate policy metrics. Finally, on hidden switched LQR, we examine whether centroids fitted from rollout data alone recover structure associated with the unobserved operating regimes, and how this structure evolves under the learned policy.

D.1 Empirical exponential derivative decay

We empirically examine the central structural prediction of Lemma 1: derivatives of an agent's action-value function should become smaller when they involve state variables located farther away in the interaction graph. For a derivative order p and graph reach d, define

$$H_p(d) := \max_i \max_{\substack{|\beta|=p \\ d_i(\beta)=d}} \sup_{z,s,a} \left|\partial_s^\beta Q_i^\pi(z,s,a)\right|.$$

Lemma 1 gives

$$H_p(d) \le L_Q \rho^d, \qquad \rho = \gamma^{1/(1+\kappa_\pi)}.$$

Since the multiplicative constant L_Q is unknown, the experiment is designed to test the *geometric dependence on graph reach*, rather than the absolute magnitude of the bound. We therefore plot $H_p(d)/H_p(0)$ and fit an empirical rate $\widehat{\rho}_p$ through

$$\log H_p(d) = \alpha_p + d \log \widehat{\rho}_p.$$

Normalizing by $H_p(0)$ only fixes the common value at $d = 0$ and does not force the measured curves to satisfy the theoretical rate.

For LQR and switched LQR, we evaluate the converged LTC($n = 2$) affine-mean Gaussian policy at each $\kappa_\pi \in \{0, 1, 2, 3\}$. Under such a policy, the corresponding action-value functions are quadratic (conditionally on σ for switched LQR). In fact, we can compute the real critics, and therefore their derivatives. The first-order supremum is evaluated exactly over the declared state–action box, second-order derivatives are constant in (s, a), and third- and higher-order derivatives vanish. For Kuramoto, we evaluate the converged LTC $(n = 2)$ policy at $\kappa_\pi = 1$; in this experiment, its critic uses $\gamma = 0.75$, as on LQR, instead of the benchmark value 0.99, hence $\rho = \sqrt{0.75}$. Since no closed-form critic is available, we enumerate all distinct state-derivative multi-indices up to order three and numerically maximize their magnitudes over states and actions using multi-start projected ascent. The resulting Kuramoto statistics are therefore numerical lower bounds on the corresponding suprema.

Figure 4 exhibits the exponential decay structure predicted by the theory across all three environments. At every nonzero derivative order, sensitivities involving increasingly distant state variables decrease approximately geometrically with graph reach. The phenomenon therefore extends beyond ordinary spatial localization of Q_i^π: graph distance influences the individual higher-order sensitivities of the critic, including the derivatives that form its Taylor coefficients.

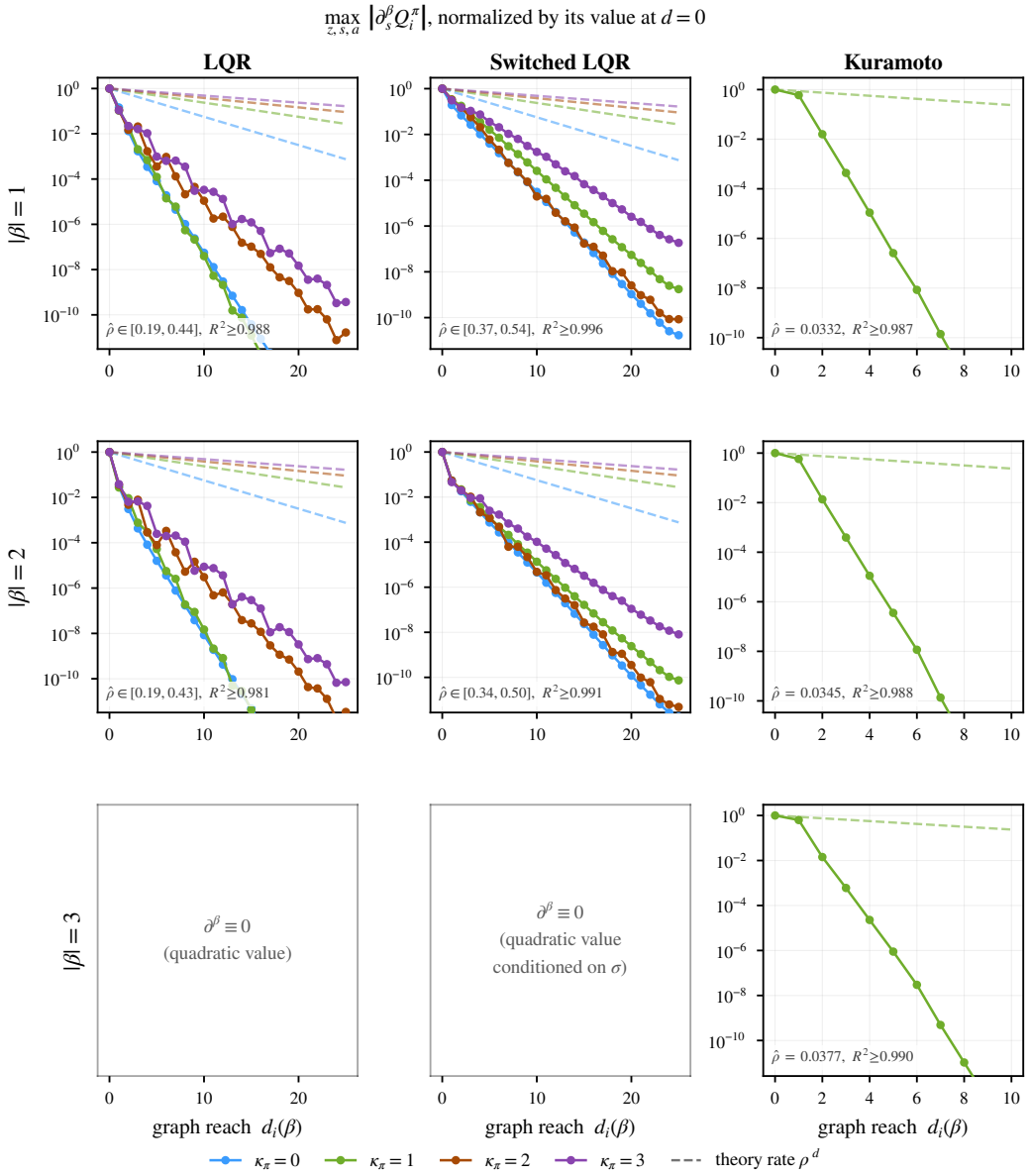

Figure 4: **Empirical exponential decay of action-value derivatives.** For each derivative order $p = |\beta|$, we plot $H_p(d)/H_p(0)$ against the graph reach $d_i(\beta)$. Solid curves show measured derivative magnitudes and dashed curves show the explicit theoretical rate ρ^d, with $\rho = \gamma^{1/(1+\kappa_\pi)}$; colors indicate the policy locality radius κ_π. LQR and switched LQR derivatives are computed exactly from their quadratic and quadratic conditioned on σ value functions, so their third-order derivatives vanish. For Kuramoto ($N = 20$), every derivative up to order three is maximized over states and actions: all multi-indices are enumerated at 256 visited states, and the 4 best starts per graph reach are refined by projected signed-gradient ascent (step 0.05, fixed budget, no tolerance), giving lower bounds on the suprema. Reported $\widehat{\rho}$ values are least-squares fits of $\log H_p(d)$ over reaches above the numerical precision threshold.

Across the LQR experiments, the fitted rates lie in $\widehat{\rho} \in [0.19, 0.44]$, while switched LQR gives $\widehat{\rho} \in [0.34, 0.54]$.

We emphasize that the explicit rate

$$\rho = \gamma^{1/(1+\kappa_\pi)}$$

is not intended to be sharp. The measured curves lie far below the dashed ones: the theoretical rate is conservative, and the fitted rates $\widehat{\rho}$ are much smaller in practice.

The gap is largest on the nonlinear Kuramoto system, where

$$\widehat{\rho}_1 = 0.0332, \qquad \widehat{\rho}_2 = 0.0345, \qquad \widehat{\rho}_3 = 0.0377,$$

with $R^2 \geq 0.987$, whereas the explicit theorem gives $\rho = \sqrt{0.75} \simeq 0.866$ for $\kappa_\pi = 1$. We observe additionally that the fitted rates are also nearly unchanged across derivative orders.

D.2 Centroids and local approximation quality

The finite-sample LTC bound of Theorem 2 contains the estimation error term

$$\mathcal{E}_i^{(\kappa)} = \bar{\varepsilon}_{\text{lin}} \sum_{p=0}^{n} \frac{\mathcal{R}_p}{p!},$$

which exposes two competing effects of the centroid construction. Increasing the number of centroids K makes the Taylor expansion more local, thereby reducing the centroid radii and thus $\mathcal{R}_p$. At the same time, LTC uses a single shared affine coefficient map across all clusters, so increasing K can make the collection of centroid-dependent Taylor coefficients harder to represent and thus can increase $\bar{\varepsilon}_{\text{lin}}$. Hence $\mathcal{E}_i^{(\kappa)}$ need not decrease monotonically with K.

Figure 5a directly illustrates this tradeoff. For LTC, increasing K initially improves the held-out critic fit, consistent with smaller local Taylor regions, but performance eventually deteriorates as estimation error becomes dominant. The turnover occurs earlier at larger Taylor orders: richer local polynomials require fewer expansion points.

In contrast, NTC does not exhibit the same pronounced degradation at large K, consistent with its more expressive neural coefficient map being able to accommodate the variation of the Taylor coefficients across centroids.

Figure 5b isolates the complementary geometric effect underlying the factors $\mathcal{R}_p$. On LQR, critic error increases with the distance to the assigned centroid, as expected from a local Taylor approximation, while increasing the Taylor order typically attenuates this dependence. Together, the two panels empirically illustrate the two components entering $\mathcal{E}_i^{(\kappa)}$: centroids control the size of the local Taylor regions, whereas the coefficient parameterization controls how accurately the corresponding local Taylor coefficients can be represented.

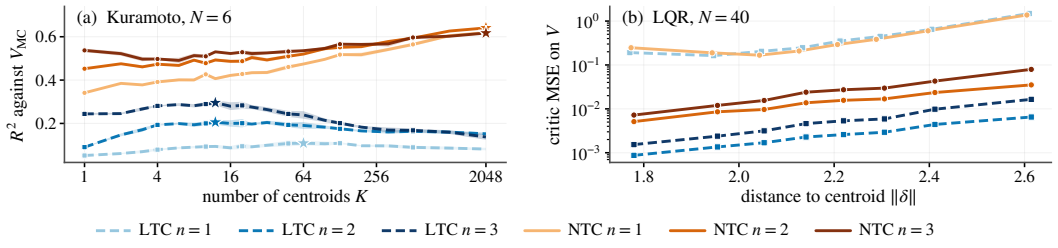

Figure 5: **Centroid and local-approximation ablations.** (a) held-out critic fit versus the number of centroids K on Kuramoto ($N = 6$, with encoding $s_i = \theta_i$), measured by $R^2 = 1 - \sum_{s,i}\big(\tilde{V}_i(s) - V_i^{\text{MC}}(s)\big)^2 / \sum_{s,i}\big(V_i^{\text{MC}}(s) - \bar{V}^{\text{MC}}\big)^2$ over 4096 held-out states, where $V_i^{\text{MC}}(s)$ averages the discounted return of agent i over 20 Monte Carlo continuations of 500 steps from s. (b) Mean squared error of the critic on the exact agent-mean value $\bar{V} = \frac{1}{N}\sum_i V_i$ on LQR ($N = 40$), as a function of the distance of the evaluated state to its assigned centroid, for LTC and NTC with $n = 1, 2, 3$.

D.3 Taylor Order Ablation

In LQR and switched LQR, under our affine-mean Gaussian policies, the critic is a quadratic (more precisely, quadratic conditioned on σ for switched LQR) function. As a result, we expect $n = 2$ to yield far better results than $n = 1$ when training our policy and $n = 3$ to provide relatively little improvement. We made this observation in the case of critic fitting in Figure 1. Figure 6 shows the same pattern during policy training as soon as $\kappa_\pi \geq 2$: the gap to the optimum is much lower for $n = 2$ than for $n = 1$, and $n = 3$ provides no improvement.

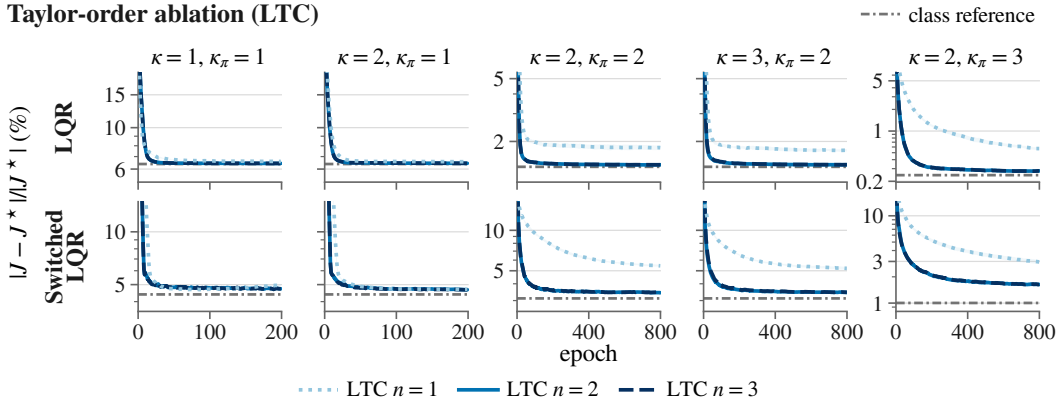

Figure 6: **Taylor order ablation.** Convergence of LTC for $n \in \{1, 2, 3\}$ for different configurations (κ, κ_π) on LQR and switched LQR.

D.4 Episode trajectories on the three environments

Figure 7 complements the aggregate performance metrics with representative trajectories. The trajectories are run on one seed simply as a qualitative overview of the policy's behavior in practice. On LQR, all methods stabilize the network rapidly around the target. On switched LQR, the trajectories exhibit the expected abrupt deviations at mode changes, yet LTC and NTC repeatedly recover stable behavior despite the switching mode being unobserved by the controller. Finally, on Kuramoto, Taylor Centroid rapidly synchronizes the agents' frequencies and maintains synchronization over the rollout. Hence, the policy obtained at convergence leads to the expected trajectories for these environments.

D.5 The latent modes appear as data-driven operating regimes.

A useful feature of switched LQR is that clusters can often be identified with interpretable regimes. In this environment, one can think of each σ_t as an operating regime for the system. Here, the meaning of a "regime" is known to the experimenter while remaining hidden from the learning algorithm. This allows us to check whether the data-driven centroids introduced in our method recover meaningful structure rather than merely producing an arbitrary partition of state space.

In this regard, Figure 8 provides a direct empirical motivation for the cluster-based representation. The K-medians centroids are fitted solely from physical states, yet, after control, the regions of state space visited under different latent modes become clearly separated and the learned centroids lie inside these recurrent regions. Thus the centroids used by the critic can acquire an interpretable operational meaning even without access to the variable generating the regimes. Figure 8 also reveals a significant state-distribution shift during training. Crucially, this motivates refitting the centroids throughout training.

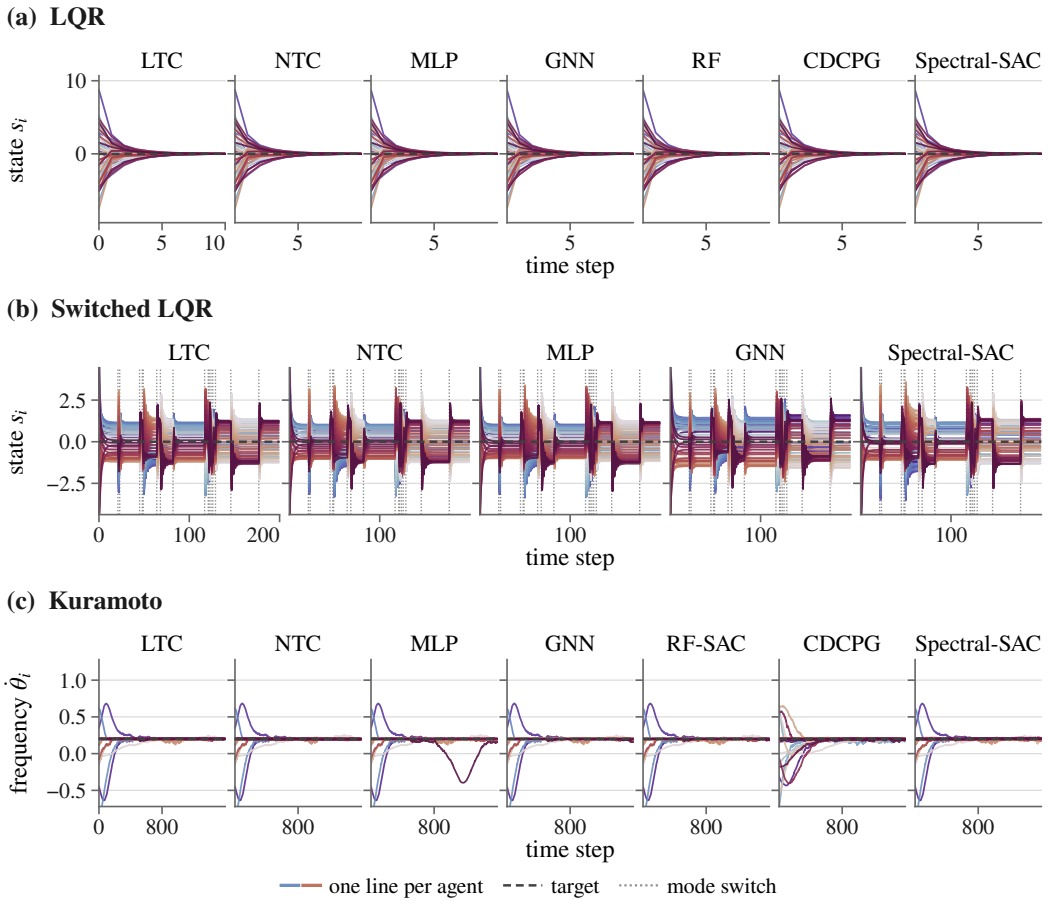

Figure 7: **Representative trajectories.** Trajectories of the learned policies on networked LQR, hidden-mode switched LQR, and the Kuramoto oscillator.

E Understanding the Expected Reward Regularity Assumption

Throughout this appendix, $q \geq 2$ denotes the regularity order in Assumption 1, while $1 \leq n \leq q-1$ denotes the Taylor order used in the critic representation. For a sufficiently differentiable function $f(z, s, a)$, define

$$\|f\|_{C_s^q} := \max_{|\beta| \leq q} \sup_{z,s,a} \left|\partial_s^\beta f(z, s, a)\right|. \tag{14}$$

For convenience, recall the expected rewards at time t:

$$g_{i,t}^\pi(z, s, a) := \mathbb{E}_\pi \left[r_i(S_{i,t}, A_{i,t}) \mid Z_0 = z, S_0 = s, A_0 = a\right]. \tag{15}$$

Assumption 1 can then be written as

$$\sup_{t \geq 0} \|g_{i,t}^\pi\|_{C_s^q} \leq \bar{L} \tag{16}$$

uniformly in the network size, agent, and policy in the considered class.

We first give an operational interpretation of Equation 16 and describe how it can be stress-tested from rollouts. We then show that its uniform-in-time form can be weakened to a discounted-summability condition. Finally, we provide several sufficient conditions and position the assumption relative to the smooth-MDP hierarchy of Maran et al. (2024a;b).

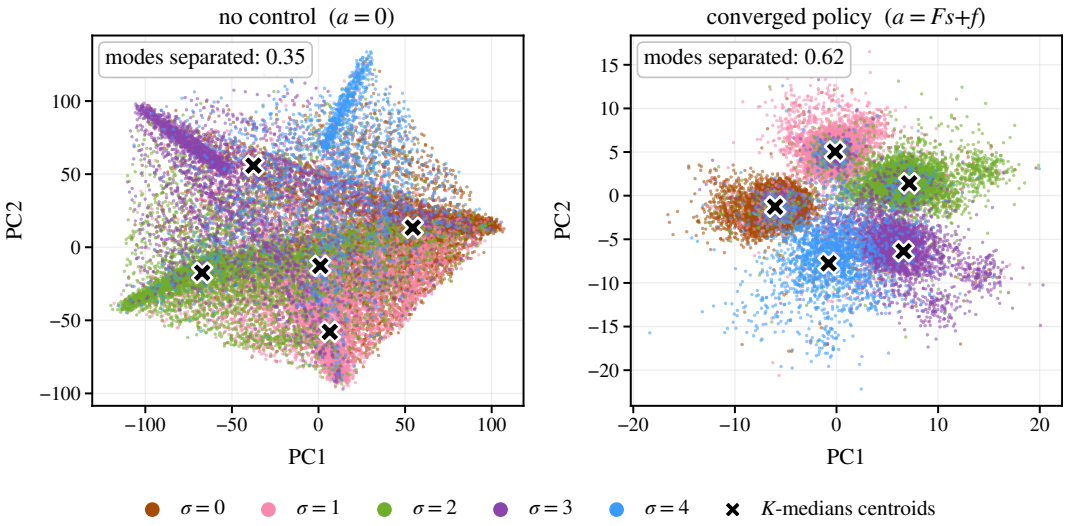

Figure 8: **Operating regimes in hidden switched LQR.** PCA projection of state trajectories, colored by latent mode σ_t for visualization; crosses denote the $K = 5$ Taylor Centroid medians. Mode separation, the fraction of the variance of the two principal-component scores z_t explained by the latent mode, $\sum_m n_m \|\bar{z}_m - \bar{z}\|^2 / \sum_t \|z_t - \bar{z}\|^2$ (with $\bar{z}_m$ the mean of the n_m points with $\sigma_t = m$ and $\bar{z}$ the overall mean; 0 when all modes share the same mean, 1 when each mode collapses to a point), increases from 0.35 without control (left) to 0.62 after training with σ_t unobserved (right), as trajectories form clearer recurrent regions and the centroids align with the latent regimes.

E.1 Interpretation

Assumption 1 has a direct physical interpretation. Consider a small perturbation of the current physical state. The assumption requires that, as this perturbation propagates through the system dynamics, it does not create arbitrarily sharp variations in the expected local reward of agent i. Thus, nearby physical configurations may evolve differently, but the expected reward received after a finite number of steps must remain a sufficiently smooth function of the current state.

Letting

$$g^{\pi}_{i,t}(z, s, a) = \mathbb{E}_{\pi}[r_i(S_{i,t}, A_{i,t}) \mid Z_0 = z, S_0 = s, A_0 = a]$$

The assumption requires for the derivatives of $g^{\pi}_{i,t}$ in s to remain bounded for all t.

E.2 A rollout-based stress test

A practical advantage of Assumption 1 is that it is stated in terms of a physically interpretable quantity, the future reward predictions $g^{\pi}_{i,t}$ of Equation 15, that can be probed directly from rollouts. This is considerably easier to inspect empirically than smoothness conditions imposed on an entire Bellman operator, such as Mild smoothness in Maran et al. (2024a).

Figure 9 evaluates $\left|\partial^{\beta}_{s} g^{\pi}_{i,t}(z, s, a)\right|$ for derivative orders $|\beta| \in \{1, 2, 3\}$ on LQR, switched LQR, and Kuramoto. Each curve follows one sampled initial value (z, s, a), reference agent i, and multi-index β over the prediction horizon. Finite differences perturb only the initial physical state s, while the initial context z and action a are held fixed.

Derivatives are estimated by symmetric finite differences around states visited by the policy. Each state coordinate is perturbed by h times its empirical standard deviation, where h is a dimensionless step size. We repeat the computation for $h \in \{0.01, 0.02, 0.05\}$ and display $h = 0.02$ in the figure. The estimates are generally stable across these choices. Relative differences across h should be interpreted cautiously when a derivative is itself extremely small, since tiny absolute changes can then correspond to large relative variations; this is most visible for the third-order Kuramoto derivatives, which are somewhat less stable across h than the other panels while remaining controlled over the tested horizon.

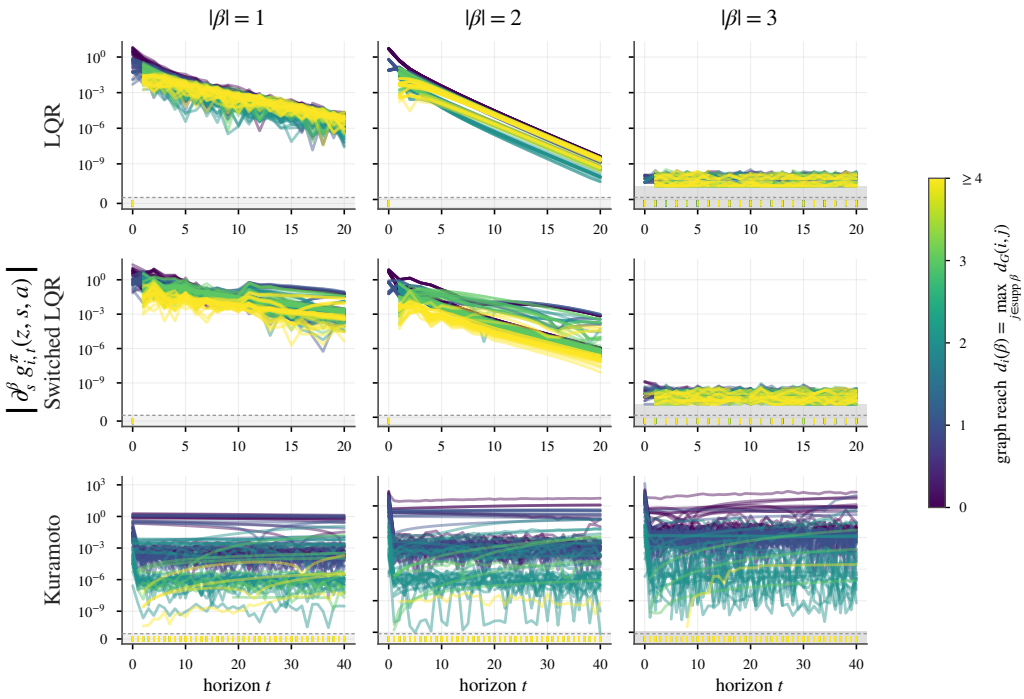

Figure 9: **Empirical stress test of expected reward regularity.** $|\partial_s^\beta g_{i,t}^\pi(z,s,a)|$ versus horizon t for $|\beta| \in \{1,2,3\}$ on LQR and switched LQR ($N = 50$, learned policies with $\kappa_\pi = 2$) and Kuramoto ($N = 40$, greedy controller), at 16 visited anchors (z,s,a) and 5 agents per environment, with 256 rollouts per expectation. Colors: graph reach $d_i(\beta) = \max_{j \in \mathrm{supp}_G(\beta)} d_G(i,j)$; marks in the bottom strip: numerically zero derivatives; gray: finite-difference resolution range.

Curves are colored by the graph reach $d_i(\beta) := \max_{j \in \mathrm{supp}_G(\beta)} d_G(i,j)$, from derivatives involving agent i itself to derivatives involving more distant agents. Colored marks in the lower strip indicate numerically zero derivatives, while the gray region marks values below the numerical resolution of the finite-difference estimate and is not interpreted quantitatively.

The results are consistent with the regularity assumption. Across all three environments, derivative magnitudes remain controlled as the horizon grows, with no systematic amplification in t. On LQR and switched LQR, first- and second-order sensitivities decrease over the prediction horizon, while third-order derivatives lie at the numerical resolution floor, as expected from the quadratic (and quadratic conditioned on σ) dependence of the expected rewards on the initial state under the affine-mean policies considered here. Kuramoto, whose dynamics are nonlinear, retains nonzero derivatives through third order while their magnitudes remain controlled over the tested horizon. The plots also exhibit the expected spatial structure: derivatives involving more distant agents are typically smaller.

Because Assumption 1 is uniform over states, policies, agents, and prediction horizons, no finite experiment can verify it formally. Figure 9 should instead be viewed as a stress test: systematic growth of these expected reward sensitivities with the prediction horizon on highly visited states would provide direct evidence against the assumption. We observe no such behavior on any of the three benchmarks; instead, expected reward sensitivities remain controlled over the tested horizons, together with the distance-dependent attenuation expected from graph-local dynamics.

E.3 A WEAKER DISCOUNTED-SUMMABILITY CONDITION

For simplicity of exposition, Assumption 1 is stated as a uniform-in-t bound. This is stronger than mathematically necessary: it suffices that the smoothness of the expected reward predictions grows slowly enough to be dominated by discounting as shown in Assumption E.1.

Assumption E.1 (Discounted-summable expected reward regularity) *There exists a nonnegative sequence $\bar{L}(t)$, uniform in the network size, agent, initial global state, and policy in the considered class, such that*

$$\|g^{\pi}_{i,t}\|_{C^q_s} \le \bar{L}(t), \qquad t \ge 0, \tag{17}$$

and

$$\sum_{t=0}^{\infty} \gamma^t \bar{L}(t) < \infty. \tag{18}$$

Assumption 1 is recovered by taking $\bar{L}(t) \equiv \bar{L}$. More generally, the condition allows finite-horizon sensitivities to grow with time, provided that discounting dominates this growth. We now show that we recover the main results of our main theory under this weaker assumption.

Proposition E.1 (Regularity under a summable envelope) *Under Assumption E.1, for every multi-index $|\beta| \le q$,*

$$\partial^{\beta}_s Q^{\pi}_i(z,s,a) = \sum_{t=0}^{\infty} \gamma^t \partial^{\beta}_s g^{\pi}_{i,t}(z,s,a), \tag{19}$$

and

$$\sup_{z,s,a} \left|\partial^{\beta}_s Q^{\pi}_i(z,s,a)\right| \le \sum_{t=0}^{\infty} \gamma^t \bar{L}(t). \tag{20}$$

The exponential decay extends to this setting. For $d \ge 0$, define

$$\tau_d := \left\lceil \frac{d}{1+\kappa_\pi} \right\rceil, \qquad \Lambda_d := \sum_{t=\tau_d}^{\infty} \gamma^t \bar{L}(t). \tag{21}$$

The sequence $(\Lambda_d)_{d\ge 0}$ is nonincreasing and tends to 0 as $d \to +\infty$.

Proposition E.2 (Exponential decay under a summable envelope) *Under Assumption E.1, for every nonzero multi-index $|\beta| \le q$,*

$$\left|\partial^{\beta}_s Q^{\pi}_i(z,s,a)\right| \le \Lambda_{d_i(\beta)}. \tag{22}$$

Thus the geometric envelope used in the main text is only one particular case. The decay rate is determined by the discounted sum of $\bar{L}(t)$.

For the uniform envelope $\bar{L}(t) \equiv \bar{L}$, this reduces to

$$\Lambda_d = \frac{\bar{L}}{1-\gamma}\gamma^{\tau_d} \le \frac{\bar{L}}{1-\gamma}\rho^d = L_Q \rho^d,$$

recovering Lemma 1.

The Taylor Centroid approximation can be extended explicitly to this general envelope. For $0 \le d \le \kappa$, define the same local centroid radii as in Theorem 1,

$$W_d := \sup_s \left\| s_{\mathcal{N}^d_i} - \mu_{k(s),\mathcal{N}^d_i} \right\|_1,$$

and, for every integer $m \ge 1$, define

$$\mathcal{R}^{\Lambda}_m := \Lambda_\kappa W^m_\kappa + \sum_{d=0}^{\kappa-1} \big(\Lambda_d - \Lambda_{d+1}\big) W^m_d. \tag{23}$$

Because $(\Lambda_d)_d$ is nonincreasing, all coefficients in Equation 23 are nonnegative.

Proposition E.3 (Scalable local Taylor representation under a summable envelope) *Under Assumption E.1, let $1 \le n \le q-1$ and let $\widehat{Q}^{\pi}_i$ be the order-n Taylor Centroid approximation defined in Equation 4. Then*

$$\sup_{z,s,a} \left|Q^{\pi}_i(z,s,a) - \widehat{Q}^{\pi}_i(z,s,a)\right| \le \frac{2R_{\max}}{1-\gamma}\rho^{\kappa+1} + \frac{\mathcal{R}^{\Lambda}_{n+1}}{(n+1)!}, \qquad \rho := \gamma^{1/(1+\kappa_\pi)}. \tag{24}$$

Moreover, $\widehat{Q}_i^\pi$ contains

$$\binom{|\mathcal{N}_i^\kappa| d_s + n}{n}$$

state monomials and therefore remains independent of N for bounded graph degree and fixed (κ, n, d_s).

The two terms in Equation 24 have different origins. The localization term is unchanged because spatial decay uses only bounded rewards, graph locality, and discounting. The Taylor term instead inherits the derivative envelope $(\Lambda_d)_d$. The latter is obtained by grouping Taylor monomials according to their graph reach and applying summation by parts:

$$\sum_{d=0}^{\kappa} \Lambda_d \big(A_{m,d} - A_{m,d-1}\big) = \Lambda_\kappa A_{m,\kappa} + \sum_{d=0}^{\kappa-1} (\Lambda_d - \Lambda_{d+1}) A_{m,d}, \tag{25}$$

where

$$A_{m,d} = \frac{\|h_{\mathcal{N}_i^d}\|_1^m}{m!}, \qquad A_{m,-1} := 0.$$

For the uniform envelope $\bar{L}(t) \equiv \bar{L}$, the geometric majorant $\Lambda_d \le L_Q \rho^d$ gives

$$\mathcal{R}_m^\Lambda \le L_Q \left[\rho^\kappa W_\kappa^m + (1-\rho) \sum_{d=0}^{\kappa-1} \rho^d W_d^m \right] = L_Q \mathcal{R}_m,$$

so Proposition E.3 recovers the bound of Theorem 1.

The remaining results adapt in the same way. Value-localization terms retain their geometric factor ρ, while every use of the exponential decay $L_Q \rho^{d_i(\beta)}$ is replaced by the corresponding term $\Lambda_{d_i(\beta)}$. The LSTD concentration argument itself is unchanged, and the actor analysis only uses the fitted critic through its final error certificate Δ_Q.

E.4 When does expected reward regularity hold?

We now give sufficient conditions under which Assumption 1 holds.

E.4.1 Smooth stochastic dynamics.

In many cases, the regularity we require is provided by the transition kernel itself. Indeed, a broad instance of problems satisfying Assumption 1 consists of graph-local systems with smooth stochastic transitions. Consider

$$S_i' = f_i(Z, S_{\mathcal{N}_i}, A_{\mathcal{N}_i}) + M_i(Z, S_{\mathcal{N}_i}, A_{\mathcal{N}_i})\, \eta_i, \tag{26}$$

where η_i is independent of the current state.

Proposition E.4 (Smooth stochastic dynamics are sufficient) *Suppose first that, uniformly in i and a_i, $r_i(\cdot, a_i)$ is C^q with bounded derivatives through order q. Suppose further that, uniformly in i and z, the maps*

$$f_i(z, \cdot, \cdot), \qquad M_i(z, \cdot, \cdot), \qquad M_i(z, \cdot, \cdot)^{-1}$$

are C^q in the physical-state variables with bounded derivatives through order q, and that M_i is uniformly invertible. Suppose finally that the density p_i of η_i satisfies

$$\max_{|\alpha| \le q} \int_{\mathbb{R}^{d_s}} (1 + \|u\|)^q \, |\partial^\alpha p_i(u)| \, du < \infty$$

uniformly in i. Then Assumption 1 holds with a constant independent of N whenever the graph degree is bounded.

Here, noise acts as a smoothing mechanism: even if individual trajectories are sensitive, expected rewards remain uniformly smooth functions of the current state due to the regularizing effect of the kernel.

Low-noise limit. For isotropic noise $M_i = \sigma I$, the constants produced by the above certificate may deteriorate as $\sigma \to 0$: at derivative order q, the proof gives a worst-case scaling of order $O(\sigma^{-q})$ (up to local smoothness and graph-degree factors). This reflects a loss of smoothing as the noise vanishes. Importantly, it is the sufficient certificate that degrades, not necessarily the true derivatives of the critic; for instance, in linear–quadratic systems the higher-order critic derivatives need not depend on σ. The proposition covers several standard stochastic systems. The additive-noise model

$$S_i' = f_i(Z, S_{\mathcal{N}_i}, A_{\mathcal{N}_i}) + \varepsilon_i$$

is recovered by taking $M_i \equiv I$, and includes graph-local systems with Gaussian noise as well as many smooth heavy-tailed noise distributions. Allowing M_i to depend on the current state additionally covers smooth heteroskedastic and state-dependent uncertainty. The required density condition is satisfied, for example, by Gaussian and Gaussian-mixture noise, smooth compactly supported densities, and smooth heavy-tailed distributions with sufficiently many finite moments. The main additive-noise settings studied by Wang et al. (2026); Ren et al. (2025b) are therefore special cases of Proposition E.4. Thus our regularity class contains their principal model classes while also allowing state-dependent stochastic dynamics. In particular, LQR, switched LQR and the Kuramoto oscillator all satisfy Assumption 1 through Proposition E.4.

E.4.2 Smooth linear transition models.

Our assumption also follows from sufficiently regular linear factorizations of the transition kernel. Suppose that a local transition density admits

$$p_j\left(s_j' \mid s_{\mathcal{N}_j}, a_{\mathcal{N}_j}, z\right) = \sum_{h=1}^{d} \varphi_{j,h}\left(s_{\mathcal{N}_j}, a_{\mathcal{N}_j}, z\right) \mu_{j,h}(s_j'). \tag{27}$$

Proposition E.5 (Smooth spectral representations are sufficient) *Suppose first that, uniformly in i and a_i, $r_i(\cdot, a_i)$ is C^q with bounded derivatives through order q. Suppose further that the current-state feature maps $\varphi_{j,h}$ have uniformly bounded physical-state derivatives through order q and that the next-state factors satisfy*

$$\sup_{j,h} \|\mu_{j,h}\|_{L^1} < \infty.$$

Then Assumption 1 holds with an N-independent constant for bounded graph degree.

Thus exact finite-dimensional spectral models of the type used in continuous networked MDPs provide another one-step smoothing certificate whenever their current-state features are sufficiently smooth. This is, in particular, the main setting targeted by (Ren et al., 2025b; Wang et al., 2026) and prior Linear MDP works (Jin et al., 2020).

E.4.3 Comparison between smoothness notions

We finally compare Assumption 1 with the smooth-MDP hierarchy of Maran et al. (2024a;b). In contrast to our assumption, Mild smoothness is a model-level one-step smoothing certificate: it requires a Bellman update to smooth every bounded continuation value.

For $x = (z, s)$, let

$$(\mathcal{T}^\star f)(x, a) := r(x, a) + \int \sup_{a'} f(x', a')\, P(dx' \mid x, a)$$

denote the one-step Bellman optimality operator, written without discount to match the finite-horizon convention of Maran et al. (2024a;b). Including a discount factor only changes the corresponding constants.

The original smooth-MDP hierarchy of Maran et al. (2024b) distinguishes Strongly and Weakly Smooth MDPs. In what follows, $C^{m,1}$ denotes the class of m-times differentiable functions whose derivatives of order m are Lipschitz continuous.

Definition E.1 (Strongly Smooth MDP, Maran et al. (2024b)) *For an integer $m \geq 0$, an MDP is m-Strongly Smooth if the reward and transition density are themselves smooth in the current state–action pair:*

$$r(\cdot, \cdot) \in C^{m,1}, \qquad p(x' \mid \cdot, \cdot) \in C^{m,1} \quad \textit{for every } x',$$

with the corresponding $C^{m,1}$ norms bounded uniformly in x' (and, in the finite-horizon formulation, uniformly over the time step).

Thus Strong smoothness imposes regularity directly on the model. In particular, it requires the transition kernel to admit a sufficiently regular density.

Definition E.2 (Weakly Smooth MDP, Maran et al. (2024b)) *For an integer $m \geq 0$, an MDP is m-Weakly Smooth if the Bellman optimality operator is bounded from $C^{m,1}$ to itself: there exists $C_{\mathcal{T}} < \infty$ such that*

$$\|\mathcal{T}^\star f\|_{C^{m,1}} \leq C_{\mathcal{T}}\big(\|f\|_{C^{m,1}} + 1\big)$$

for every $f \in C^{m,1}$.

Weak smoothness therefore does not require the transition density itself to be smooth. Instead, it requires Bellman updates to preserve smoothness of functions that are already smooth. This is intuitively similar to our assumption which requires the expected value of the rewards (a smooth function) to remain smooth.

Maran et al. (2024a) introduce an intermediate notion, Mild smoothness. For $\nu > 0$, let

$$\nu_\star := \lceil \nu - 1 \rceil.$$

Their space $\mathcal{C}^\nu$ consists of functions with $\nu_\star$ continuous derivatives whose derivatives of order $\nu_\star$ are Hölder continuous with exponent $\nu - \nu_\star$. In particular, at integer indices,

$$\mathcal{C}^\nu = C^{\nu-1,1}$$

up to equivalence of the corresponding norms.

Definition E.3 (Mildly Smooth MDP, (Maran et al., 2024a)) *An MDP is ν-Mildly Smooth if the Bellman optimality operator is bounded from L^∞ to $\mathcal{C}^\nu$: there exists $C_{\mathcal{T}} < \infty$ such that*

$$\|\mathcal{T}^\star f\|_{\mathcal{C}^\nu} \leq C_{\mathcal{T}}\big(\|f\|_\infty + 1\big)$$

for every bounded measurable f.

The key distinction is that Mild smoothness does not require the input to be smooth: a Bellman update must produce a smooth function from an arbitrary bounded input. At integer orders, Appendix E.3 (Relation between the settings) of Maran et al. (2024a) proves

$$m\text{-Strong} \quad \Longrightarrow \quad (m+1)\text{-Mild} \quad \Longrightarrow \quad m\text{-Weak}. \tag{28}$$

Mild smoothness provides a particularly simple sufficient condition for our assumption.

Proposition E.6 (Smooth MDPs satisfy expected reward regularity) *Suppose that the local rewards are C^q with uniformly bounded derivatives. If the augmented MDP is ν-Mildly Smooth for some*

$$\nu > q,$$

then Assumption 1 of order q holds.

Consequently, at integer indices,

$$q\text{-Strong} \quad \Longrightarrow \quad (q+1)\text{-Mild} \quad \Longrightarrow \quad \text{Assumption 1 of order } q. \tag{29}$$

In particular, Maran et al. (2024a;b) show that Mild smoothness captures a wide variety of MDPs which therefore satisfy Assumption 1. The reason is that every finite-horizon local reward prediction remains uniformly bounded:

$$\|g^\pi_{i,t}\|_\infty \leq R_{\max}.$$

The condition $\nu > q$ comes from our choice to formulate Assumption 1 using C^q following the classical notation. This one-order shift is therefore a matter of smoothness convention rather than a substantive difference in the Taylor approximation.

The implication is strict. Mild smoothness is substantially stronger in scope than what our analysis requires. It controls the Bellman operator on every bounded input function, whereas Assumption 1 only controls the particular expected reward sequence.

To understand why that is the case, let us consider an example with a single-agent MDP and a singleton action space,

$$s = (x, y) \in [-1, 1]^2, \qquad r(x, y) = x^2,$$

We consider the following deterministic dynamics

$$F(x, y) = \left(x, \mathbf{1}_{\{x \geq 0\}}\right).$$

The reward depends only on the first coordinate, which is preserved by the dynamics. Consequently,

$$g_t(x, y) = x^2 \qquad \text{for every } t \geq 0,$$

so the entire expected reward sequence is smooth at every finite order and Assumption 1 holds.

The Bellman operator, however, also probes parts of the dynamics that never affect this reward sequence. Consider the smooth bounded function

$$f(x', y') = y'.$$

Then

$$\mathcal{T}^\star f(x, y) = x^2 + \mathbf{1}_{\{x \geq 0\}},$$

which is discontinuous at $x = 0$. Thus the Bellman operator *does not even preserve continuity* of smooth inputs. The discontinuity in the second state coordinate is entirely irrelevant to every future reward and therefore to the critic itself. Assumption 1 correctly ignores it. Mild smoothness cannot make this distinction because it asks the Bellman operator to regularize arbitrary functions, including functions specifically chosen to probe that irrelevant coordinate.

Beyond this informative example, the following proposition shows that this distinction in notions also arises in switched LQR, a well-established control model.

Proposition E.7 (A continuous-context switched LQR separation) *Consider a single-agent switched LQR on compact state and action domains, with a continuous global context $Z_t \in [-1, 1]$,*

$$Z_{t+1} = -Z_t, \qquad \sigma(z) := \mathbf{1}_{\{z \geq 0\}},$$

and physical dynamics

$$S_{t+1} = A^{(\sigma(Z_t))} S_t + B^{(\sigma(Z_t))} A_t + \varepsilon_t.$$

Assume that ε_t is centered and has a C^∞ full-dimensional density satisfying the conditions of Proposition E.4 at every finite order, and choose the domains, mode matrices, and noise support so that the physical state remains in its compact domain. Let

$$r(s, a) = -(s^\top Q s + a^\top R a),$$

and assume that the two regimes are distinct on the domain, i.e., there exist $(\bar{s}, \bar{a})$ such that

$$d := \left(A^{(1)} - A^{(0)}\right) \bar{s} + \left(B^{(1)} - B^{(0)}\right) \bar{a} \neq 0.$$

Then Assumption 1 holds at every finite order q, whereas the augmented MDP on (Z, S) is not Mildly Smooth at any positive order.

Here Z_t is a continuous switching variable: except at $Z_t = 0$, its sign flips at every step, so the physical dynamics alternate between the two linear regimes selected by $\sigma_t = \mathbf{1}_{\{Z_t \geq 0\}}$. This example shows a second way in which expected reward regularity can be strictly more targeted than operator-wide smoothness. The mode switch is discontinuous as a function of the continuous context z, so Bellman outputs on the full augmented state need not be smooth in z.

Discrete and hidden operating modes. There is a further distinction concerning the global context Z allowed in our model. The standard Mild-smoothness formulation is stated on a continuous state–action domain equipped with a C^ν smoothness structure. Consequently, when the Markov state is hybrid,

$$(Z, S),$$

with Z taking values in a finite set, Mild smoothness is not directly defined in its standard form without first specifying how the smoothness space should be extended to the discrete coordinate.

This situation arises naturally in switched LQR (Costa et al., 2005; Ogura et al., 2018). Assumption 1 requires no such extension: derivatives are taken only with respect to the continuous physical state s, while the corresponding bounds are required uniformly over the context z.

F Proofs of the main-text results

This section contains the proofs of the results stated in the main text. We first establish the finite-propagation and decay properties, then derive the Taylor Centroid approximation bounds, prove the finite-sample LTC–LSTD guarantee, and finally analyze the localized actor and the end-to-end policy-optimization update.

Conditioning convention. The algorithm is adaptive across actor iterations. Fix one such iteration. All critic-side statements are understood conditionally on the sigma-field generated before the current critic rollouts. Under this conditioning, the current policy, centroids, cluster assignment, and feature map are fixed, while the M_c critic rollouts are independent trajectories generated by the current policy. After fitting the critic, actor-side statements are understood conditionally on the enlarged sigma-field containing the fitted critics and any randomness used to construct action-independent baselines. Under this conditioning, the M_s fresh actor rollouts are independent. Conditional failure-probability bounds below are uniform in the realized history, so the corresponding unconditional statements follow from the tower property. We suppress these conditionings whenever no ambiguity can arise.

For $t \geq 0$, let

$$g^\pi_{i,t}(z, s, a) := \mathbb{E}_\pi[r_i(S_{i,t}, A_{i,t}) \,|\, Z_0 = z, S_0 = s, A_0 = a]\,. \tag{30}$$

Then

$$Q^\pi_i(z, s, a) = \sum_{t=0}^{\infty} \gamma^t g^\pi_{i,t}(z, s, a), \tag{31}$$

and the bounded-reward assumption gives

$$|g^\pi_{i,t}(z, s, a)| \leq R_{\max}, \qquad |Q^\pi_i(z, s, a)| \leq \frac{R_{\max}}{1-\gamma}. \tag{32}$$

F.1 Finite propagation and exponential decay

Proof of Proposition 1. *Proof.* By Equation 32,

$$\sup_{z,s,a} \left|\gamma^t g^\pi_{i,t}(z, s, a)\right| \leq \gamma^t R_{\max},$$

and $\sum_{t\geq 0} \gamma^t R_{\max} < \infty$. Hence Equation 31 converges uniformly.

Fix a multi-index β with $|\beta| \leq q$. Assumption 1 gives

$$\sup_{z,s,a} \left|\partial^\beta_s g^\pi_{i,t}(z, s, a)\right| \leq \bar{L} \qquad \text{for every } t \geq 0.$$

Therefore

$$\sum_{t=0}^{\infty} \sup_{z,s,a} \left|\gamma^t \partial^\beta_s g^\pi_{i,t}(z, s, a)\right| \leq \frac{\bar{L}}{1-\gamma}.$$

Thus the derivative series converges uniformly for every derivative order up to q. Applying the standard theorem on termwise differentiation of uniformly convergent series, successively in each state coordinate, shows that $s \mapsto Q_i^\pi(z, s, a)$ is of class C^q and yields

$$\partial_s^\beta Q_i^\pi(z, s, a) = \sum_{t=0}^{\infty} \gamma^t \partial_s^\beta g_{i,t}^\pi(z, s, a), \qquad |\beta| \le q. \tag{33}$$

The derivatives are continuous because, by Assumption 1, each $g_{i,t}^\pi$ is of class C^q, and the derivative series converges uniformly. Moreover,

$$\sup_{z,s,a} \left|\partial_s^\beta Q_i^\pi(z, s, a)\right| \le \frac{\bar{L}}{1-\gamma} = L_Q, \qquad |\beta| \le q. \tag{34}$$

This proves Proposition 1. □

Lemma F.1 (Finite propagation) *Let*

$$R_t := t(1 + \kappa_\pi).$$

Conditionally on the global trajectory $(Z_0, \dots, Z_t)$, the conditional law of $(S_{i,t}, A_{i,t})$ given $(S_0, A_0) = (s, a)$ depends on the initial distributed state–action pair only through its restriction to $\mathcal{N}_i^{R_t}$.

Proof. We argue by induction on t, simultaneously for all agents. At $t = 0$, $(S_{i,0}, A_{i,0}) = (s_i, a_i)$, so the statement holds with $R_0 = 0$.

Assume that it holds at time $t - 1$. Conditional on $(Z_{t-1}, S_{t-1}, A_{t-1})$, the factorized transition kernel implies that the law of $S_{i,t}$ depends only on

$$\big(S_{j,t-1}, A_{j,t-1}\big)_{j \in \mathcal{N}_i}.$$

For $j \in \mathcal{N}_i$, the induction hypothesis restricts the dependence of these variables on the initial distributed coordinates to $\mathcal{N}_j^{R_{t-1}}$, and

$$\mathcal{N}_j^{R_{t-1}} \subseteq \mathcal{N}_i^{R_{t-1}+1}.$$

Hence the conditional law of $S_{i,t}$ depends on the initial distributed variables only through $\mathcal{N}_i^{R_{t-1}+1}$.

Next,

$$A_{i,t} \sim \pi_{\theta_i}\Big(\cdot \mid S_{\mathcal{N}_i^{\kappa_\pi}}(t), h_i(Z_t)\Big).$$

After conditioning on the global trajectory, $h_i(Z_t)$ is fixed. For every $j \in \mathcal{N}_i^{\kappa_\pi}$, the preceding state argument shows that $S_{j,t}$ depends on the initial distributed coordinates only through $\mathcal{N}_j^{R_{t-1}+1}$, and

$$\mathcal{N}_j^{R_{t-1}+1} \subseteq \mathcal{N}_i^{R_{t-1}+1+\kappa_\pi}.$$

Thus the conditional law of $(S_{i,t}, A_{i,t})$ depends on the initial distributed state–action pair only through $\mathcal{N}_i^{R_{t-1}+1+\kappa_\pi}$. Since

$$R_{t-1} + 1 + \kappa_\pi = (t-1)(1+\kappa_\pi) + 1 + \kappa_\pi = t(1+\kappa_\pi) = R_t,$$

the induction closes. □

Lemma F.2 (Spatial decay) *If (s, a) and (s', a') agree on all state–action coordinates indexed by $\mathcal{N}_i^\kappa$, then*

$$|Q_i^\pi(z, s, a) - Q_i^\pi(z, s', a')| \le \frac{2R_{\max}}{1-\gamma} \rho^{\kappa+1}.$$

Proof. Set

$$\tau_\kappa := \left\lfloor \frac{\kappa}{1+\kappa_\pi} \right\rfloor.$$

Since $P_g(z' \mid z)$ does not depend on the distributed state–action variables, the global process $(Z_t)_{t\geq 0}$ has the same law under the two initializations. For every $t \leq \tau_\kappa$,

$$R_t = t(1+\kappa_\pi) \leq \kappa.$$

By Lemma F.1, conditionally on any realization of the global trajectory, $(S_{i,t}, A_{i,t})$ therefore has the same law under the two initializations. Hence:

$$g^\pi_{i,t}(z,s,a) = g^\pi_{i,t}(z,s',a') \qquad (0 \leq t \leq \tau_\kappa).$$

Put simply, at time t, information propagating from outside $\mathcal{N}_i^\kappa$ has not yet reached agent i, so only $\mathcal{N}_i^\kappa$ is relevant to $g_{i,t}$. Using Equation 31 and Equation 32,

$$\begin{aligned}
|Q^\pi_i(z,s,a) - Q^\pi_i(z,s',a')| &\leq \sum_{t=\tau_\kappa+1}^{\infty} \gamma^t \left| g^\pi_{i,t}(z,s,a) - g^\pi_{i,t}(z,s',a') \right| \\
&\leq 2R_{\max} \sum_{t=\tau_\kappa+1}^{\infty} \gamma^t \\
&= \frac{2R_{\max}}{1-\gamma} \gamma^{\tau_\kappa+1}.
\end{aligned}$$

Moreover,

$$\tau_\kappa + 1 = \left\lfloor \frac{\kappa}{1+\kappa_\pi} \right\rfloor + 1 \geq \frac{\kappa+1}{1+\kappa_\pi}.$$

Since $\gamma \in (0,1)$,

$$\gamma^{\tau_\kappa+1} \leq \gamma^{(\kappa+1)/(1+\kappa_\pi)} = \rho^{\kappa+1},$$

which proves the claim. □

Proof of Lemma 1. *Proof.* Let $\beta \neq 0$ with $|\beta| \leq n+1$, and set

$$d := d_i(\beta), \qquad \tau_\beta := \left\lceil \frac{d}{1+\kappa_\pi} \right\rceil.$$

If $t < \tau_\beta$, then $R_t = t(1+\kappa_\pi) < d$. Conditional on the global trajectory $(Z_t)_{t\geq 0}$, Lemma F.1 implies that $g^\pi_{i,t}$ depends on the initial distributed variables only through $\mathcal{N}_i^{R_t}$. By the definition of $d_i(\beta)$, the support of β contains a state coordinate belonging to an agent at distance d from i. Since this coordinate lies outside $\mathcal{N}_i^{R_t}$, $g^\pi_{i,t}$ is constant in that coordinate, and therefore

$$\partial_s^\beta g^\pi_{i,t}(z,s,a) = 0 \qquad (t < \tau_\beta).$$

The law of the global trajectory $(Z_t)_{t\geq 0}$ does not depend on the initial distributed state, so averaging over it does not reintroduce such dependence.

Using Equation 33,

$$\begin{aligned}
\left|\partial_s^\beta Q^\pi_i(z,s,a)\right| &= \left| \sum_{t=\tau_\beta}^{\infty} \gamma^t \partial_s^\beta g^\pi_{i,t}(z,s,a) \right| \\
&\leq \bar{L} \sum_{t=\tau_\beta}^{\infty} \gamma^t \\
&= L_Q \gamma^{\tau_\beta}.
\end{aligned}$$

Since $\tau_\beta \geq d_i(\beta)/(1+\kappa_\pi)$ and $\gamma \in (0,1)$,

$$\gamma^{\tau_\beta} \leq \gamma^{d_i(\beta)/(1+\kappa_\pi)} = \rho^{d_i(\beta)}.$$

This proves the claimed bound for $\beta \neq 0$. For $\beta = 0$, $d_i(0) = 0$ and Equation 34 gives $|Q^\pi_i| \leq L_Q = L_Q \rho^0$. □

F.2 Localized Taylor Centroid approximation

Lemma F.3 (Localization error) *For every (z, s, a),*

$$\left|Q_i^\pi(z,s,a) - \bar{Q}_i^\pi(z, s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})\right| \le \frac{2R_{\max}}{1-\gamma}\rho^{\kappa+1}. \tag{35}$$

Proof. Write

$$x := (s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}), \qquad y := (s_{\mathcal{N}_{-i}^\kappa}, a_{\mathcal{N}_{-i}^\kappa}).$$

By definition of $\bar{Q}_i^\pi = \mathscr{D}_{i,0}$,

$$\bar{Q}_i^\pi(z,x) = \int Q_i^\pi(z,x,\widetilde{y})\,\nu_{i,\mathrm{far}}(d\widetilde{y}).$$

Hence

$$Q_i^\pi(z,x,y) - \bar{Q}_i^\pi(z,x) = \int [Q_i^\pi(z,x,y) - Q_i^\pi(z,x,\widetilde{y})]\,\nu_{i,\mathrm{far}}(d\widetilde{y}).$$

For every $\widetilde{y}$, the two arguments inside the brackets agree on all state–action coordinates in $\mathcal{N}_i^\kappa$. Lemma F.2 therefore bounds the absolute value of the integrand by $2R_{\max}(1-\gamma)^{-1}\rho^{\kappa+1}$. Since $\nu_{i,\mathrm{far}}$ is a probability measure, the same bound holds after integration. □

This is the same idea as the one used by (Qu et al., 2020b; Ren et al., 2025b; Wang et al., 2026).

Lemma F.4 (Localized derivatives) *For every multi-index β supported in $\mathcal{N}_i^\kappa$ with $|\beta| \le n+1$,*

$$\partial_s^\beta \bar{Q}_i^\pi(z, s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) = \mathscr{D}_{i,\beta}(z, s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}), \tag{36}$$

and

$$\left|\mathscr{D}_{i,\beta}(z, s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})\right| \le L_Q \rho^{d_i(\beta)}. \tag{37}$$

Proof. For every such β, Lemma 1 gives

$$\left|\partial_s^\beta Q_i^\pi(z,s,a)\right| \le L_Q\rho^{d_i(\beta)} \le L_Q$$

uniformly in all far coordinates. This uniform bound is integrable with respect to the probability measure $\nu_{i,\mathrm{far}}$. Repeated application of dominated convergence therefore permits differentiation under the integral defining $\bar{Q}_i^\pi$ and yields Equation 36. The same bound gives

$$|\mathscr{D}_{i,\beta}| \le \int \left|\partial_s^\beta Q_i^\pi\right| d\nu_{i,\mathrm{far}} \le L_Q\rho^{d_i(\beta)},$$

which proves Equation 37. □

Lemma F.5 (Graph-shell identity) *Let*

$$h := s_{\mathcal{N}_i^\kappa} - \mu_{k(s),\mathcal{N}_i^\kappa}.$$

For every integer $p \ge 0$ and $0 \le d \le \kappa$, define

$$A_{p,d} := \sum_{\substack{|\beta|=p \\ d_i(\beta)\le d}} \frac{|h^\beta|}{\beta!}, \qquad A_{p,-1} := 0.$$

Then

$$A_{p,d} = \frac{\|h_{\mathcal{N}_i^d}\|_1^p}{p!}, \tag{38}$$

and

$$\sum_{d=0}^{\kappa} \rho^d (A_{p,d} - A_{p,d-1}) = \rho^\kappa A_{p,\kappa} + (1-\rho)\sum_{d=0}^{\kappa-1}\rho^d A_{p,d}. \tag{39}$$

Consequently,

$$\sum_{\substack{|\beta|=p \\ \mathrm{supp}_G(\beta)\subseteq\mathcal{N}_i^\kappa}} \rho^{d_i(\beta)}\frac{|h^\beta|}{\beta!} \le \frac{\mathcal{R}_p}{p!}. \tag{40}$$

Proof. The condition $d_i(\beta) \le d$ is equivalent to $\mathrm{supp}_G(\beta) \subseteq \mathcal{N}_i^d$. The multinomial theorem therefore gives

$$A_{p,d} = \sum_{\substack{|\beta|=p \\ \mathrm{supp}_G(\beta)\subseteq\mathcal{N}_i^d}} \frac{|h^\beta|}{\beta!} = \frac{1}{p!}\left(\sum_{j\in\mathcal{N}_i^d}\sum_{c=1}^{d_s}|h_{j,c}|\right)^p = \frac{\|h_{\mathcal{N}_i^d}\|_1^p}{p!},$$

which proves Equation 38.

The difference $A_{p,d} - A_{p,d-1}$ is precisely the contribution of monomials of graph reach d. Hence

$$\sum_{d=0}^{\kappa}\rho^d(A_{p,d}-A_{p,d-1}) = \sum_{d=0}^{\kappa}\rho^d A_{p,d} - \sum_{d=0}^{\kappa-1}\rho^{d+1}A_{p,d} = \rho^\kappa A_{p,\kappa} + (1-\rho)\sum_{d=0}^{\kappa-1}\rho^d A_{p,d},$$

which is Equation 39. Finally, if $s \in C_{k(s)}$ (where $C_1, \dots, C_K$ are the fitted clusters), the definition of W_d gives $\|h_{\mathcal{N}_i^d}\|_1 \le W_d$. Substituting Equation 38 into Equation 39 and applying these bounds yields

$$\sum_{\substack{|\beta|=p \\ \mathrm{supp}_G(\beta)\subseteq\mathcal{N}_i^\kappa}} \rho^{d_i(\beta)}\frac{|h^\beta|}{\beta!} \le \frac{1}{p!}\left[\rho^\kappa W_\kappa^p + (1-\rho)\sum_{d=0}^{\kappa-1}\rho^d W_d^p\right] = \frac{\mathcal{R}_p}{p!}.$$

□

Proposition F.1 (Approximation error of the ideal Taylor Centroid critic) *Under the assumptions of Lemma 1,*

$$\sup_{z,s,a}\left|Q_i^\pi(z,s,a) - \widehat{Q}_i^\pi(z,s,a)\right| \le \Delta_{\mathrm{TC}} = \frac{2R_{\max}}{1-\gamma}\rho^{\kappa+1} + \frac{L_Q}{(n+1)!}\mathcal{R}_{n+1}. \tag{41}$$

Proof. Decompose

$$Q_i^\pi - \widehat{Q}_i^\pi = \left(Q_i^\pi - \bar{Q}_i^\pi\right) + \left(\bar{Q}_i^\pi - \widehat{Q}_i^\pi\right).$$

The first term is bounded by Lemma F.3.

For the second term, fix (z,s,a) and set

$$\mu := \mu_{k(s),\mathcal{N}_i^\kappa}, \qquad h := s_{\mathcal{N}_i^\kappa} - \mu, \qquad p := n+1.$$

By Lemma F.4, Equation 4 is exactly the order-n Taylor polynomial of $\bar{Q}_i^\pi$ in the local state variables at μ. Since the state spaces are convex, the line segment $\mu + uh$, $u \in [0,1]$, remains in the state domain. The multivariate Taylor formula with integral remainder therefore gives

$$\bar{Q}_i^\pi(z, s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}) - \widehat{Q}_i^\pi(z,s,a) = p\sum_{\substack{|\beta|=p \\ \mathrm{supp}_G(\beta)\subseteq\mathcal{N}_i^\kappa}} \frac{h^\beta}{\beta!}\int_0^1 (1-u)^{p-1}\partial_s^\beta \bar{Q}_i^\pi(z, \mu+uh, a_{\mathcal{N}_i^\kappa})\,du.$$

Using Equation 37 and $p\int_0^1(1-u)^{p-1}\,du = 1$,

$$\left|\bar{Q}_i^\pi - \widehat{Q}_i^\pi\right| \le L_Q \sum_{\substack{|\beta|=p \\ \mathrm{supp}_G(\beta)\subseteq\mathcal{N}_i^\kappa}} \rho^{d_i(\beta)}\frac{|h^\beta|}{\beta!} \le \frac{L_Q}{(n+1)!}\mathcal{R}_{n+1},$$

where the last inequality is Lemma F.5. Adding the localization error proves Equation 41. □

Corollary F.1 (Distribution-weighted Taylor Centroid error) *Let ν be any distribution over states and put $p = n + 1$. Define*

$$\mathcal{R}_p(s) := \rho^{\kappa} \left\| s_{\mathcal{N}_i^{\kappa}} - \mu_{k(s),\mathcal{N}_i^{\kappa}} \right\|_1^p + (1-\rho) \sum_{d=0}^{\kappa-1} \rho^d \left\| s_{\mathcal{N}_i^{d}} - \mu_{k(s),\mathcal{N}_i^{d}} \right\|_1^p .$$

Then, for every $z \in \mathcal{Z}$ and $a \in \mathcal{A}$,

$$\mathbb{E}_{S\sim\nu}\left[\left| Q_i^{\pi}(z,S,a) - \widehat{Q}_i^{\pi}(z,S,a) \right| \right] \le \frac{2R_{\max}}{1-\gamma} \rho^{\kappa+1} + \frac{L_Q}{(n+1)!} \mathbb{E}_{S\sim\nu}[\mathcal{R}_{n+1}(S)] .$$

Moreover, if, for every $0 \le d \le \kappa$,

$$\left\| S_{\mathcal{N}_i^{d}} - \mu_{k(S),\mathcal{N}_i^{d}} \right\|_1 \le D_d \qquad \textit{almost surely},$$

then

$$\begin{aligned} \mathbb{E}[\mathcal{R}_p(S)] \le{}& \rho^{\kappa} D_{\kappa}^{p-1} \mathbb{E} \left\| S_{\mathcal{N}_i^{\kappa}} - \mu_{k(S),\mathcal{N}_i^{\kappa}} \right\|_1 \\ &+ (1-\rho) \sum_{d=0}^{\kappa-1} \rho^d D_d^{p-1} \mathbb{E} \left\| S_{\mathcal{N}_i^{d}} - \mu_{k(S),\mathcal{N}_i^{d}} \right\|_1 . \end{aligned}$$

Proof. Repeat the proof of Theorem 1 without replacing the state-dependent distances to the assigned centroid by their uniform bounds W_d, and then take expectations. The second inequality follows from $x^p \le D_d^{p-1} x$ for $0 \le x \le D_d$. □

The K-medians objective used by Taylor Centroid therefore reduces a valid upper bound on the average Taylor remainder under the policy-induced state distribution: it aims to minimize the Taylor representation error *on average*.

Proposition F.2 (Approximation by the affine coefficient map) *Let $\mathcal{E}_i^{(\kappa)}$ be as defined in Theorem 2. For every $\zeta > 0$, there exists an LTC parameter $w_i^{(\zeta)} \in \mathbb{R}^{p_i}$ such that*

$$\sup_{z,s,a} \left| \widehat{Q}_i^{\pi}(z,s,a) - \widetilde{Q}_i^{\mathrm{LTC}}(h_i(z), s, a_{\mathcal{N}_i^{\kappa}}; w_i^{(\zeta)}) \right| \le \mathcal{E}_i^{(\kappa)} + \zeta. \tag{42}$$

In particular, $\bar{\varepsilon}_{\mathrm{lin}} \le L_Q$.

Proof. For $\beta \in \mathcal{M}_{\mathcal{N}_i^{\kappa}}^n$, let

$$\varepsilon_{\mathrm{lin}}(\beta) := \inf_{v \in \mathbb{R}^{d_\psi^i}} \sup_{z,k,a} \left| \mathscr{D}_{i,\beta}(z, \mu_{k,\mathcal{N}_i^{\kappa}}, a_{\mathcal{N}_i^{\kappa}}) - v^{\top} \psi_i(k, h_i(z), a_{\mathcal{N}_i^{\kappa}}) \right| .$$

By Lemma F.4,

$$\sup_{z,k,a} \left| \mathscr{D}_{i,\beta}(z, \mu_{k,\mathcal{N}_i^{\kappa}}, a_{\mathcal{N}_i^{\kappa}}) \right| \le L_Q \rho^{d_i(\beta)}.$$

The zero vector is an admissible affine coefficient, hence

$$\varepsilon_{\mathrm{lin}}(\beta) \le L_Q \rho^{d_i(\beta)}.$$

Multiplying by $\rho^{-d_i(\beta)}$ and taking the maximum over β proves $\bar{\varepsilon}_{\mathrm{lin}} \le L_Q$.

Now fix $\zeta > 0$. Because $\mathcal{M}_{\mathcal{N}_i^{\kappa}}^n$ is finite, for every $\varepsilon > 0$ and every β one can choose $v_{i,\beta}^{(\varepsilon)}$ such that

$$\begin{aligned} &\sup_{z,k,a} \left| \mathscr{D}_{i,\beta}(z, \mu_{k,\mathcal{N}_i^{\kappa}}, a_{\mathcal{N}_i^{\kappa}}) - (v_{i,\beta}^{(\varepsilon)})^{\top} \psi_i(k, h_i(z), a_{\mathcal{N}_i^{\kappa}}) \right| \\ &\qquad \le \varepsilon_{\mathrm{lin}}(\beta) + \varepsilon \rho^{d_i(\beta)} \le (\bar{\varepsilon}_{\mathrm{lin}} + \varepsilon) \rho^{d_i(\beta)}. \end{aligned}$$

Stack these coefficient vectors into $w_i^{(\varepsilon)} \in \mathbb{R}^{p_i}$. With $h = s_{\mathcal{N}_i^\kappa} - \mu_{k(s),\mathcal{N}_i^\kappa}$, the definitions of LTC and $\widehat{Q}_i^\pi$ imply

$$\begin{aligned}
&\left|\widehat{Q}_i^\pi(z,s,a) - \widetilde{Q}_i^{\mathrm{LTC}}(h_i(z), s, a_{\mathcal{N}_i^\kappa}; w_i^{(\varepsilon)})\right| \\
&\quad\le (\bar{\varepsilon}_{\mathrm{lin}} + \varepsilon) \sum_{\beta \in \mathcal{M}^n_{\mathcal{N}_i^\kappa}} \rho^{d_i(\beta)} \frac{|h^\beta|}{\beta!} \\
&\quad\le (\bar{\varepsilon}_{\mathrm{lin}} + \varepsilon) \sum_{p=0}^{n} \frac{\mathcal{R}_p}{p!},
\end{aligned}$$

where the second inequality follows by applying Lemma F.5 degree by degree. Since

$$\mathcal{E}_i^{(\kappa)} = \bar{\varepsilon}_{\mathrm{lin}} \sum_{p=0}^{n} \frac{\mathcal{R}_p}{p!},$$

choosing $\varepsilon > 0$ so that

$$\varepsilon \sum_{p=0}^{n} \frac{\mathcal{R}_p}{p!} \le \zeta$$

gives Equation 42. □

Lemma F.6 (Feature and critic envelopes) *Let $\mathcal{G}_i^{(\kappa)}$ and $\mathcal{V}_i^{(\kappa)}$ be as defined in Theorem 2. For every state s,*

$$\|T_i(s)\|_2^2 \le \sum_{p=0}^{n} \frac{W_\kappa^{2p}}{p!}, \qquad \|\Phi_i\|^2 \le \mathcal{G}_i^{(\kappa)}. \tag{43}$$

Moreover,

$$\sup_{z,s,a} |\widehat{Q}_i^\pi(z,s,a)| \le \mathcal{V}_i^{(\kappa)}, \tag{44}$$

and the approximant in Proposition F.2 may be chosen so that

$$\left\| \widetilde{Q}_i^{\mathrm{LTC}}(\cdot; w_i^{(\zeta)}) \right\|_\infty \le \mathcal{V}_i^{(\kappa)} + \mathcal{E}_i^{(\kappa)} + \zeta. \tag{45}$$

Proof. Let $h = s_{\mathcal{N}_i^\kappa} - \mu_{k(s),\mathcal{N}_i^\kappa}$. For every $p \le n$,

$$\begin{aligned}
\sum_{|\beta|=p} \frac{|h^\beta|^2}{(\beta!)^2} &\le \sum_{|\beta|=p} \frac{(|h|^2)^\beta}{\beta!} \\
&= \frac{\|h\|_2^{2p}}{p!} \le \frac{\|h\|_1^{2p}}{p!} \le \frac{W_\kappa^{2p}}{p!}.
\end{aligned}$$

Summing over $p = 0, \dots, n$ proves the first inequality in Equation 43. Since the Euclidean norm is multiplicative under Kronecker products,

$$\|\Phi_i\|^2 = \|\psi_i\|^2 \, \|T_i\|_2^2 \le R_\psi^2 \sum_{p=0}^{n} \frac{W_\kappa^{2p}}{p!} = \mathcal{G}_i^{(\kappa)}.$$

For the ideal critic, the degree-zero coefficient satisfies $|\bar{Q}_i^\pi| \le R_{\max}/(1-\gamma)$ by Equation 32. For the positive-degree terms, Lemma F.4 and Lemma F.5 give

$$\begin{aligned}
|\widehat{Q}_i^\pi| &\le \frac{R_{\max}}{1-\gamma} + L_Q \sum_{p=1}^{n} \sum_{\substack{|\beta|=p \\ \mathrm{supp}_G(\beta) \subseteq \mathcal{N}_i^\kappa}} \rho^{d_i(\beta)} \frac{|h^\beta|}{\beta!} \\
&\le \frac{R_{\max}}{1-\gamma} + L_Q \sum_{p=1}^{n} \frac{\mathcal{R}_p}{p!} = \mathcal{V}_i^{(\kappa)}.
\end{aligned}$$

Finally, Equation 45 follows from Equation 44, Equation 42, and the triangle inequality. □

F.3 FINITE-SAMPLE ANALYSIS OF LTC–LSTD

For rollout $j \in [M_c]$ and $0 \le t \le T-2$, write

$$\Phi_i^{j,t} := \Phi_i\Big(h_i(Z_t^{(j)}), S_t^{(j)}, A_{t,\mathcal{N}_i^\kappa}^{(j)}\Big), \qquad n_c := M_c(T-1).$$

The empirical LSTD system is

$$\widehat{A}_i \widehat{w}_i = \widehat{b}_i, \tag{46}$$

with

$$\widehat{A}_i = \frac{1}{n_c}\sum_{j=1}^{M_c}\sum_{t=0}^{T-2} \Phi_i^{j,t}\left(\Phi_i^{j,t} - \gamma\Phi_i^{j,t+1}\right)^\top, \qquad \widehat{b}_i = \frac{1}{n_c}\sum_{j=1}^{M_c}\sum_{t=0}^{T-2}\Phi_i^{j,t} r_i^{j,t}. \tag{47}$$

Proof of Theorem 2. *Proof.* Fix an agent i and condition as described at the beginning of this section. The policy, centroids, cluster assignment, and feature map are therefore fixed.

Fix $\zeta > 0$ and let $w_i^{(\zeta)}$ be supplied by Proposition F.2. Define

$$f^{(\zeta)}(y_i, s, a_{\mathcal{N}_i^\kappa}) := \widetilde{Q}_i^{\mathrm{LTC}}(y_i, s, a_{\mathcal{N}_i^\kappa}; w_i^{(\zeta)})$$

and abbreviate

$$\Delta_\zeta := \Delta_{\mathrm{TC}} + \mathcal{E}_i^{(\kappa)} + \zeta, \qquad \mathcal{V}_\zeta := \mathcal{V}_i^{(\kappa)} + \mathcal{E}_i^{(\kappa)} + \zeta.$$

Propositions F.1 and F.2 imply

$$\sup_{z,s,a}\left|Q_i^\pi(z,s,a) - f^{(\zeta)}(h_i(z), s, a_{\mathcal{N}_i^\kappa})\right| \le \Delta_\zeta. \tag{48}$$

Step 1: reduction to the empirical LSTD residual. The triangle inequality gives

$$\begin{aligned}\sup_{z,s,a}&\left|Q_i^\pi(z,s,a) - \widetilde{Q}_i^{\mathrm{LTC}}(h_i(z), s, a_{\mathcal{N}_i^\kappa}; \widehat{w}_i)\right| \\ &\le \Delta_\zeta + \sup_{z,s,a}\left|f^{(\zeta)}(h_i(z), s, a_{\mathcal{N}_i^\kappa}) - \widetilde{Q}_i^{\mathrm{LTC}}(h_i(z), s, a_{\mathcal{N}_i^\kappa}; \widehat{w}_i)\right|.\end{aligned} \tag{49}$$

Both functions in the second term belong to the same linear feature class. By Lemma F.6,

$$\sup_{z,s,a}\left|f^{(\zeta)} - \widetilde{Q}_i^{\mathrm{LTC}}(\cdot; \widehat{w}_i)\right| \le \sqrt{\mathcal{G}_i^{(\kappa)}}\left\|w_i^{(\zeta)} - \widehat{w}_i\right\|.$$

On the event $\sigma_{\min}(\widehat{A}_i) \ge c_{\mathrm{LSTD}} > 0$, $\widehat{A}_i$ is invertible and

$$\begin{aligned}\left\|w_i^{(\zeta)} - \widehat{w}_i\right\| &= \left\|\widehat{A}_i^{-1}\left(\widehat{A}_i w_i^{(\zeta)} - \widehat{b}_i\right)\right\| \\ &\le \frac{1}{c_{\mathrm{LSTD}}}\left\|\widehat{A}_i w_i^{(\zeta)} - \widehat{b}_i\right\|.\end{aligned} \tag{50}$$

Consequently,

$$\sup_{z,s,a}\left|f^{(\zeta)} - \widetilde{Q}_i^{\mathrm{LTC}}(\cdot; \widehat{w}_i)\right| \le \frac{\sqrt{\mathcal{G}_i^{(\kappa)}}}{c_{\mathrm{LSTD}}}\left\|\widehat{A}_i w_i^{(\zeta)} - \widehat{b}_i\right\|. \tag{51}$$

Step 2: Bellman bias and martingale fluctuation. For each (j,t) set

$$f_{j,t}^{(\zeta)} := f^{(\zeta)}\left(h_i(Z_t^{(j)}), S_t^{(j)}, A_{t,\mathcal{N}_i^\kappa}^{(j)}\right).$$

By Equation 47,

$$\widehat{A}_i w_i^{(\zeta)} - \widehat{b}_i = \frac{1}{n_c}\sum_{j=1}^{M_c}\sum_{t=0}^{T-2}\Phi_i^{j,t}\left[f_{j,t}^{(\zeta)} - r_i^{j,t} - \gamma f_{j,t+1}^{(\zeta)}\right]. \tag{52}$$

Let

$$\widetilde{f}^{(\zeta)}(z,s,a) := f^{(\zeta)}(h_i(z), s, a_{\mathcal{N}_i^\kappa})$$

and define the local Bellman evaluation operator

$$(\mathcal{T}_i^\pi u)(z,s,a) := r_i(s_i,a_i) + \gamma \mathbb{E}_\pi \left[u(Z_1,S_1,A_1) \mid Z_0 = z, S_0 = s, A_0 = a \right].$$

Adding and subtracting the one-step conditional mean of $f_{j,t+1}^{(\zeta)}$ decomposes

$$\widehat{A}_i w_i^{(\zeta)} - \widehat{b}_i = R_{\mathrm{Bell}}^{(\zeta)} + R_{\mathrm{MG}}^{(\zeta)},$$

where

$$R_{\mathrm{Bell}}^{(\zeta)} := \frac{1}{n_c} \sum_{j,t} \Phi_i^{j,t} \left[f_{j,t}^{(\zeta)} - r_i^{j,t} - \gamma \mathbb{E}\left[f_{j,t+1}^{(\zeta)} \mid Z_t^{(j)}, S_t^{(j)}, A_t^{(j)} \right] \right],$$

$$R_{\mathrm{MG}}^{(\zeta)} := \frac{\gamma}{n_c} \sum_{j,t} \Phi_i^{j,t} \left[\mathbb{E}\left[f_{j,t+1}^{(\zeta)} \mid Z_t^{(j)}, S_t^{(j)}, A_t^{(j)} \right] - f_{j,t+1}^{(\zeta)} \right].$$

Because $Q_i^\pi = \mathcal{T}_i^\pi Q_i^\pi$ and conditional expectation is a contraction in the sup norm,

$$\begin{aligned} \left\| \widetilde{f}^{(\zeta)} - \mathcal{T}_i^\pi \widetilde{f}^{(\zeta)} \right\|_\infty &\le \left\| \widetilde{f}^{(\zeta)} - Q_i^\pi \right\|_\infty + \left\| \mathcal{T}_i^\pi \widetilde{f}^{(\zeta)} - \mathcal{T}_i^\pi Q_i^\pi \right\|_\infty \\ &\le (1+\gamma)\Delta_\zeta. \end{aligned}$$

Using $\|\Phi_i^{j,t}\| \le \sqrt{\mathcal{G}_i^{(\kappa)}}$,

$$\left\| R_{\mathrm{Bell}}^{(\zeta)} \right\| \le \sqrt{\mathcal{G}_i^{(\kappa)}} (1+\gamma)\Delta_\zeta. \tag{53}$$

For the stochastic term, define

$$Z_{j,t}^{(\zeta)} := \gamma \Phi_i^{j,t} \left[\mathbb{E}\left[f_{j,t+1}^{(\zeta)} \mid Z_t^{(j)}, S_t^{(j)}, A_t^{(j)} \right] - f_{j,t+1}^{(\zeta)} \right].$$

Order the pairs (j,t) by revealing each rollout sequentially and then moving to the next independent rollout. Relative to the resulting natural filtration, $Z_{j,t}^{(\zeta)}$ is a martingale difference: the factor $\Phi_i^{j,t}$ is measurable at the current step and the term in brackets has conditional mean zero. Moreover, by Equation 45,

$$\left\| Z_{j,t}^{(\zeta)} \right\| \le 2\gamma \sqrt{\mathcal{G}_i^{(\kappa)}}\, \mathcal{V}_\zeta.$$

The Hilbert-space Azuma–Hoeffding inequality (Pinelis, 1994) therefore yields, for this fixed agent, with probability at least $1 - \delta/N$,

$$\left\| R_{\mathrm{MG}}^{(\zeta)} \right\| \le 2\gamma \sqrt{\mathcal{G}_i^{(\kappa)}}\, \mathcal{V}_\zeta \sqrt{\frac{2\log(2N/\delta)}{M_c(T-1)}}. \tag{54}$$

In particular, no independence assumption is made between transitions from the same rollout.

Step 3: conclusion. Combining Equation 53 and Equation 54, then using Equation 51 and Equation 49, gives

$$\begin{aligned} &\sup_{z,s,a} \left| Q_i^\pi(z,s,a) - \widetilde{Q}_i^{\mathrm{LTC}}(h_i(z), s, a_{\mathcal{N}_i^\kappa}; \widehat{w}_i) \right| \\ &\quad \le \left(1 + \frac{(1+\gamma)\mathcal{G}_i^{(\kappa)}}{c_{\mathrm{LSTD}}} \right) \Delta_\zeta + \frac{2\gamma \mathcal{G}_i^{(\kappa)} \mathcal{V}_\zeta}{c_{\mathrm{LSTD}}} \sqrt{\frac{2\log(2N/\delta)}{M_c(T-1)}}. \end{aligned}$$

The inequality holds for every $\zeta > 0$. Letting $\zeta \downarrow 0$ gives

$$\begin{aligned} &\sup_{z,s,a} \left| Q_i^\pi(z,s,a) - \widetilde{Q}_i^{\mathrm{LTC}}(h_i(z), s, a_{\mathcal{N}_i^\kappa}; \widehat{w}_i) \right| \\ &\quad \le \left(1 + \frac{(1+\gamma)\mathcal{G}_i^{(\kappa)}}{c_{\mathrm{LSTD}}} \right) (\Delta_{\mathrm{TC}} + \mathcal{E}_i^{(\kappa)}) \\ &\qquad + \frac{2\gamma \mathcal{G}_i^{(\kappa)} (\mathcal{V}_i^{(\kappa)} + \mathcal{E}_i^{(\kappa)})}{c_{\mathrm{LSTD}}} \sqrt{\frac{2\log(2N/\delta)}{M_c(T-1)}}. \end{aligned}$$

For each fixed i, the bound fails with probability at most δ/N. A union bound over $i \in [N]$ therefore yields the stated simultaneous probability at least $1 - \delta$. This is exactly the bound stated in Theorem 2. □

F.4 LOCALIZED POLICY GRADIENT WITH ONE-STEP TD ADVANTAGES

Throughout this subsection, in addition to the assumptions stated in the main text, we use the standard regularity conditions for the likelihood-ratio policy-gradient identity: the policy is differentiable in its parameters, its support is parameter-independent, and differentiation may be interchanged with the relevant action integrals. The bounded-score assumption supplies the required domination.

Define

$$\mathcal{E}_i := \mathcal{N}_i^{\kappa+\kappa_\pi}, \qquad n_i := |\mathcal{E}_i|.$$

For a local reward stream ℓ, let

$$V_\ell^\pi(z,s) := \mathbb{E}_{a\sim\pi_\theta(\cdot|z,s)}\left[Q_\ell^\pi(z,s,a)\right], \qquad A_\ell^\pi := Q_\ell^\pi - V_\ell^\pi.$$

Lemma F.7 (Policy-gradient identity) *For every agent i,*

$$\nabla_{\theta_i} J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t \mathsf{s}_i(O_{i,t},A_{i,t})\frac{1}{N}\sum_{\ell=1}^{N} A_\ell^\pi(Z_t,S_t,A_t)\right]. \tag{55}$$

More generally, any bounded baseline depending on (Z_t,S_t) but not on the current action may be subtracted inside the expectation.

Proof. See (Sutton et al., 1999). We also use that the global advantage is the average of individual advantages. □

Localization of discarded agents. Set $R := \kappa + \kappa_\pi$ and $\mathcal{E}_i := \mathcal{N}_i^R$. For any integer $r \geq 0$, let $\bar{Q}_{\ell,r}^\pi$ denote the marginalized critic of agent ℓ localized to $\mathcal{N}_\ell^r$, defined exactly as the localized critic used above but with localization radius r. The localization bound of Lemma F.3 applies for every such r.

For each discarded agent $\ell \notin \mathcal{E}_i$, write

$$d_{i\ell} := d_G(i,\ell), \qquad r_{i\ell} := d_{i\ell} - 1,$$

and define the auxiliary localized advantage

$$\bar{A}_{\ell|i}^\pi(z,s,a) := \bar{Q}_{\ell,r_{i\ell}}^\pi\big(z, s_{\mathcal{N}_\ell^{r_{i\ell}}}, a_{\mathcal{N}_\ell^{r_{i\ell}}}\big) - V_\ell^\pi(z,s).$$

These auxiliary critics are used only in the analysis; the algorithm still uses the common critic radius κ.

Lemma F.8 (Vanishing localized contribution of far reward streams) *For every t,*

$$\mathbb{E}\left[\mathsf{s}_i(O_{i,t},A_{i,t})\frac{1}{N}\sum_{\ell\notin\mathcal{E}_i}\bar{A}_{\ell|i}^\pi(Z_t,S_t,A_t)\right] = 0. \tag{56}$$

Proof. Fix $\ell \notin \mathcal{E}_i$. By construction, $r_{i\ell} = d_G(i,\ell) - 1$, hence $i \notin \mathcal{N}_\ell^{r_{i\ell}}$. Therefore $\bar{Q}_{\ell,r_{i\ell}}^\pi$ does not depend on the current action a_i; neither does $V_\ell^\pi(z,s)$. Thus $\bar{A}_{\ell|i}^\pi$ is independent of a_i.

Condition on $(Z_t,S_t,A_{-i,t})$. We have,

$$A_{i,t} \sim \pi_{\theta_i}\left(\cdot \mid S_{t,\mathcal{N}_i^{\kappa_\pi}}, h_i(Z_t)\right).$$

Under this conditioning, $\bar{A}_{\ell|i}^\pi(Z_t,S_t,A_t)$ is constant as a function of $A_{i,t}$, whereas the conditional expectation of $\mathsf{s}_i(O_{i,t},A_{i,t})$ is zero. Hence

$$\mathbb{E}\left[\mathsf{s}_i(O_{i,t},A_{i,t})\bar{A}_{\ell|i}^\pi(Z_t,S_t,A_t)\right] = 0.$$

Summing over $\ell \notin \mathcal{E}_i$ proves Equation 56. □

Define the distance-resolved tail

$$\tau_i(R) := \frac{1}{N}\sum_{\ell:\, d_G(i,\ell)>R} \rho^{d_G(i,\ell)}. \tag{57}$$

Lemma F.9 (Spatial truncation bias of the actor) *Let*

$$g_i^{\mathrm{loc}} := \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t \mathsf{s}_i(O_{i,t},A_{i,t})\frac{1}{N}\sum_{\ell\in\mathcal{E}_i} A_\ell^\pi(Z_t,S_t,A_t)\right].$$

Then

$$\left\|\nabla_{\theta_i} J(\theta) - g_i^{\mathrm{loc}}\right\| \le \frac{2B_{\mathsf{s},i}R_{\max}}{(1-\gamma)^2}\tau_i(R). \tag{58}$$

Proof. By Lemma F.7, the difference between the exact and localized gradients is the contribution of agents $\ell \notin \mathcal{E}_i$. For each such agent,

$$A_\ell^\pi = \bar{A}_{\ell|i}^\pi + \big(A_\ell^\pi - \bar{A}_{\ell|i}^\pi\big).$$

The contribution of $\bar{A}_{\ell|i}^\pi$ vanishes by Lemma F.8. Moreover,

$$A_\ell^\pi - \bar{A}_{\ell|i}^\pi = Q_\ell^\pi - \bar{Q}_{\ell,r_{i\ell}}^\pi.$$

Applying Lemma F.3 at radius $r_{i\ell} = d_{i\ell} - 1$ gives

$$\left|A_\ell^\pi - \bar{A}_{\ell|i}^\pi\right| \le \frac{2R_{\max}}{1-\gamma}\rho^{r_{i\ell}+1} = \frac{2R_{\max}}{1-\gamma}\rho^{d_G(i,\ell)}.$$

Consequently,

$$\left|\frac{1}{N}\sum_{\ell\notin\mathcal{E}_i}\big(A_\ell^\pi - \bar{A}_{\ell|i}^\pi\big)\right| \le \frac{2R_{\max}}{1-\gamma}\tau_i(R).$$

Using $\|\mathsf{s}_i\| \le B_{\mathsf{s},i}$ and $\sum_{t\ge 0}\gamma^t = (1-\gamma)^{-1}$ proves Equation 58. □

For the remainder of the actor analysis, suppose that the fitted critics satisfy

$$\sup_{z,s,a}\left|\widetilde{Q}_\ell(h_\ell(z),s,a_{\mathcal{N}_\ell^\kappa}) - Q_\ell^\pi(z,s,a)\right| \le \Delta_Q, \qquad \ell\in[N]. \tag{59}$$

Define

$$\widetilde{V}_\ell(z,s) := \mathbb{E}_{a\sim\pi_\theta(\cdot|z,s)}\left[\widetilde{Q}_\ell(h_\ell(z),s,a_{\mathcal{N}_\ell^\kappa})\right]. \tag{60}$$

Lemma F.10 (One-step TD advantage error) *Let*

$$e_\ell(z,s) := \widetilde{V}_\ell(z,s) - V_\ell^\pi(z,s).$$

Then $\|e_\ell\|_\infty \le \Delta_Q$. *Writing* $X_t = (Z_t,S_t)$, *the one-step TD advantage satisfies*

$$\mathbb{E}\left[\widehat{A}_{\ell,t}\,\middle|\,X_t,A_t\right] = A_\ell^\pi(X_t,A_t) + \gamma\,\mathbb{E}[e_\ell(X_{t+1})\mid X_t,A_t] - e_\ell(X_t). \tag{61}$$

Consequently, for every agent i,

$$\left\|\mathbb{E}\Big[\mathsf{s}_i(O_{i,t},A_{i,t})\big(\widehat{A}_{\ell,t} - A_\ell^\pi(X_t,A_t)\big)\Big]\right\| \le \gamma B_{\mathsf{s},i}\Delta_Q. \tag{62}$$

Proof. By Equations 60 and 59,

$$|e_\ell(z,s)| \le \mathbb{E}_{A\sim\pi_\theta(\cdot|z,s)}\left|\widetilde{Q}_\ell(h_\ell(z),s,A_{\mathcal{N}_\ell^\kappa}) - Q_\ell^\pi(z,s,A)\right| \le \Delta_Q.$$

Let

$$\delta_{\ell,t}^\pi := r_{\ell,t} + \gamma V_\ell^\pi(X_{t+1}) - V_\ell^\pi(X_t).$$

Since $\widetilde{V}_\ell = V_\ell^\pi + e_\ell$,

$$\widehat{A}_{\ell,t} = \delta_{\ell,t}^\pi + \gamma e_\ell(X_{t+1}) - e_\ell(X_t).$$

The Bellman identity gives $\mathbb{E}[\delta_{\ell,t}^\pi \mid X_t,A_t] = A_\ell^\pi(X_t,A_t)$, proving Equation 61.

Multiplying by the policy score and taking expectations, the term involving $e_\ell(X_t)$ vanishes because it is action-independent and $\mathbb{E}[\mathsf{s}_i(O_{i,t},A_{i,t}) \mid X_t] = 0$. Hence

$$\begin{aligned}&\left\|\mathbb{E}\Big[\mathsf{s}_i(O_{i,t},A_{i,t})\big(\widehat{A}_{\ell,t} - A_\ell^\pi(X_t,A_t)\big)\Big]\right\| \\ &\qquad \le \gamma\,\mathbb{E}[\|\mathsf{s}_i(O_{i,t},A_{i,t})\|\,|\mathbb{E}[e_\ell(X_{t+1})\mid X_t,A_t]|] \le \gamma B_{\mathsf{s},i}\Delta_Q,\end{aligned}$$

which proves Equation 62. □

Lemma F.11 (Bias of the localized finite-horizon actor estimator) *Let*

$$\Gamma_T := \sum_{t=0}^{T-1} \gamma^t = \frac{1-\gamma^T}{1-\gamma}.$$

Then

$$\|\mathbb{E}[\widehat{g}_i] - \nabla_{\theta_i} J(\theta)\| \le \frac{2B_{\mathsf{s},i} R_{\max}}{(1-\gamma)^2} \tau_i(R) + \frac{B_{\mathsf{s},i} n_i}{N} \gamma \Gamma_T \Delta_Q + \frac{2B_{\mathsf{s},i} n_i R_{\max} \gamma^T}{N(1-\gamma)^2}. \tag{63}$$

The expectation is over the fresh actor-rollout randomness under the conditioning convention above.

Proof. Define the exact localized policy gradient

$$g_i^{\mathrm{loc}} := \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \mathsf{s}_i(O_{i,t}, A_{i,t}) \frac{1}{N} \sum_{\ell \in \mathcal{E}_i} A_\ell^\pi(Z_t, S_t, A_t)\right],$$

and its finite-horizon counterpart

$$g_{i,T}^{\mathrm{loc}} := \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t \mathsf{s}_i(O_{i,t}, A_{i,t}) \frac{1}{N} \sum_{\ell \in \mathcal{E}_i} A_\ell^\pi(Z_t, S_t, A_t)\right].$$

By the definition of the actor estimator,

$$\mathbb{E}[\widehat{g}_i] = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t \mathsf{s}_i(O_{i,t}, A_{i,t}) \frac{1}{N} \sum_{\ell \in \mathcal{E}_i} \widehat{A}_{\ell,t}\right].$$

Hence,

$$\begin{aligned}\|\mathbb{E}[\widehat{g}_i] - \nabla_{\theta_i} J(\theta)\| \le{}& \left\|\nabla_{\theta_i} J(\theta) - g_i^{\mathrm{loc}}\right\| \\ &+ \left\|g_i^{\mathrm{loc}} - g_{i,T}^{\mathrm{loc}}\right\| \\ &+ \left\|g_{i,T}^{\mathrm{loc}} - \mathbb{E}[\widehat{g}_i]\right\|.\end{aligned}$$

The first term is the spatial-localization error. By Lemma F.9,

$$\left\|\nabla_{\theta_i} J(\theta) - g_i^{\mathrm{loc}}\right\| \le \frac{2B_{\mathsf{s},i} R_{\max}}{(1-\gamma)^2} \tau_i(R).$$

For the finite-horizon truncation term, using $|A_\ell^\pi| \le 2R_{\max}/(1-\gamma)$, $\|\mathsf{s}_i\| \le B_{\mathsf{s},i}$, and $|\mathcal{E}_i| = n_i$,

$$\begin{aligned}\left\|g_i^{\mathrm{loc}} - g_{i,T}^{\mathrm{loc}}\right\| &\le \sum_{t=T}^{\infty} \gamma^t B_{\mathsf{s},i} \frac{n_i}{N} \frac{2R_{\max}}{1-\gamma} \\ &= \frac{2B_{\mathsf{s},i} n_i R_{\max} \gamma^T}{N(1-\gamma)^2}.\end{aligned}$$

Finally, for the critic-approximation term, Equation 62 gives, for every $\ell \in \mathcal{E}_i$ and $t < T$,

$$\left\|\mathbb{E}\left[\mathsf{s}_i(O_{i,t}, A_{i,t})\big(\widehat{A}_{\ell,t} - A_\ell^\pi(Z_t, S_t, A_t)\big)\right]\right\| \le \gamma B_{\mathsf{s},i} \Delta_Q.$$

Therefore,

$$\begin{aligned}\left\|g_{i,T}^{\mathrm{loc}} - \mathbb{E}[\widehat{g}_i]\right\| &\le \frac{1}{N} \sum_{t=0}^{T-1} \gamma^t \sum_{\ell \in \mathcal{E}_i} \gamma B_{\mathsf{s},i} \Delta_Q \\ &= \frac{B_{\mathsf{s},i} n_i}{N} \gamma \Gamma_T \Delta_Q.\end{aligned}$$

Combining the three bounds gives Equation 63. □

Lemma F.12 (Actor-rollout concentration) *Let*

$$D := \frac{2R_{\max}}{1-\gamma} + (1+\gamma)\Delta_Q, \qquad G_i := \frac{2B_{\mathbf{s},i}n_i D}{N}\,\Gamma_T.$$

For a single actor rollout j, let $\widehat{g}_i^{(j)}$ denote its contribution to Equation 5, so that $\widehat{g}_i = M_s^{-1}\sum_{j=1}^{M_s}\widehat{g}_i^{(j)}$. Then

$$\|\widehat{g}_i^{(j)}\| \le \frac{G_i}{2}.$$

Consequently, with probability at least $1-\delta$,

$$\|\widehat{g}_i - \mathbb{E}[\widehat{g}_i]\| \le G_i\sqrt{\frac{2\log(2/\delta)}{M_s}}. \tag{64}$$

Proof. Equation 59 implies

$$\|\widetilde{V}_\ell\|_\infty \le \frac{R_{\max}}{1-\gamma} + \Delta_Q.$$

Hence every one-step TD advantage satisfies

$$|\widehat{A}_{\ell,t}| \le R_{\max} + (1+\gamma)\left(\frac{R_{\max}}{1-\gamma} + \Delta_Q\right) = D.$$

Therefore, for one rollout,

$$\|\widehat{g}_i^{(j)}\| \le \frac{B_{\mathbf{s},i}n_i D}{N}\sum_{t=0}^{T-1}\gamma^t = \frac{G_i}{2}.$$

Thus

$$\left\|\widehat{g}_i^{(j)} - \mathbb{E}[\widehat{g}_i^{(j)}]\right\| \le G_i.$$

The M_s actor rollouts are independent under the conditioning convention. Applying the Hilbert-space Hoeffding inequality to their centered contributions gives Equation 64. □

F.5 End-to-end policy optimization

For actor iteration m, define

$$\epsilon_{\mathrm{bias},i,m} := \frac{2B_{\mathbf{s},i}R_{\max}}{(1-\gamma)^2}\tau_i(R) + \frac{B_{\mathbf{s},i}n_i}{N}\gamma\Gamma_T\Delta_Q^{(m)} + \frac{2B_{\mathbf{s},i}n_iR_{\max}\gamma^T}{N(1-\gamma)^2}.$$

Let $G_{i,m}$ denote the quantity G_i in Lemma F.12 with $\Delta_Q = \Delta_Q^{(m)}$, and set

$$G_m := \left(\sum_{i=1}^N G_{i,m}^2\right)^{1/2}.$$

Also recall

$$\mathcal{T}_R = \left(\sum_{i=1}^N B_{\mathbf{s},i}^2\tau_i(R)^2\right)^{1/2}, \qquad \mathcal{B}_R = \left(\sum_{i=1}^N\left(\frac{B_{\mathbf{s},i}n_i}{N}\right)^2\right)^{1/2}.$$

Then the componentwise bias bounds and Minkowski's inequality give

$$\left(\sum_{i=1}^N \epsilon_{\mathrm{bias},i,m}^2\right)^{1/2} \le \frac{2R_{\max}}{(1-\gamma)^2}\mathcal{T}_R + \mathcal{B}_R\left(\gamma\Gamma_T\Delta_Q^{(m)} + \frac{2R_{\max}\gamma^T}{(1-\gamma)^2}\right). \tag{65}$$

Consequently, the quantity ϵ_m appearing in Theorem 3 may be taken as

$$\epsilon_m = \frac{2R_{\max}}{(1-\gamma)^2}\mathcal{T}_R + \mathcal{B}_R\left(\gamma\Gamma_T\Delta_Q^{(m)} + \frac{2R_{\max}\gamma^T}{(1-\gamma)^2}\right) + G_m\sqrt{\frac{2\log(4M/\delta)}{M_s}}.$$

Lemma F.13 (Simultaneous gradient control) *Conditionally on the critic certificates required by Theorem 3, with probability at least $1-\delta$ over the fresh actor rollouts, simultaneously for all $m = 0, \dots, M-1$,*

$$\|\widehat{g}_m - \nabla J(\theta_m)\| \le \epsilon_m, \qquad \|\widehat{g}_m\| \le \frac{G_m}{2}. \tag{66}$$

Proof. Fix m and condition on the history preceding its fresh actor rollouts. For one rollout, concatenate the agent-wise contributions:

$$\widehat{g}_m^{(j)} := \big(\widehat{g}_{1,m}^{(j)}, \dots, \widehat{g}_{N,m}^{(j)}\big).$$

Lemma F.12 gives

$$\|\widehat{g}_{i,m}^{(j)}\| \le \frac{G_{i,m}}{2}.$$

Therefore

$$\|\widehat{g}_m^{(j)}\|^2 = \sum_{i=1}^N \|\widehat{g}_{i,m}^{(j)}\|^2 \le \frac{1}{4}\sum_{i=1}^N G_{i,m}^2 = \frac{G_m^2}{4}.$$

Consequently,

$$\Big\|\widehat{g}_m^{(j)} - \mathbb{E}[\widehat{g}_m^{(j)}]\Big\| \le G_m.$$

Hilbert-space Hoeffding, applied conditionally with failure probability $\delta/(2M)$, gives

$$\|\widehat{g}_m - \mathbb{E}[\widehat{g}_m]\| \le G_m\sqrt{\frac{2\log(4M/\delta)}{M_s}}.$$

By Lemma F.11,

$$\|\mathbb{E}[\widehat{g}_m] - \nabla J(\theta_m)\| \le \left(\sum_{i=1}^N \epsilon_{\mathrm{bias},i,m}^2\right)^{1/2}.$$

The triangle inequality gives the first inequality in Equation 66. The second is deterministic under the conditioning, since $\widehat{g}_m$ is the average of vectors of norm at most $G_m/2$.

A union bound over m gives conditional failure probability at most $\delta/2$, hence in particular at most δ. Because each conditional bound holds uniformly over the realized past, the tower property yields the same unconditional conclusion. If the critic certificates themselves are obtained from Theorem 2, their failure probabilities may be allocated separately and combined by another union bound. □

Proof of Theorem 3. *Proof.* Work on the event in Lemma F.13 and define

$$e_m := \widehat{g}_m - \nabla J(\theta_m), \qquad \|e_m\| \le \epsilon_m.$$

Set

$$\eta_m := \frac{\eta}{\sqrt{m+1}}.$$

The L_J-smoothness of J and the ascent update $\theta_{m+1} = \theta_m + \eta_m \widehat{g}_m$ imply

$$\begin{aligned} J(\theta_{m+1}) &\ge J(\theta_m) + \eta_m \langle \nabla J(\theta_m), \widehat{g}_m\rangle - \frac{L_J}{2}\eta_m^2\|\widehat{g}_m\|^2 \\ &= J(\theta_m) + \eta_m\|\nabla J(\theta_m)\|^2 + \eta_m\langle \nabla J(\theta_m), e_m\rangle - \frac{L_J}{2}\eta_m^2\|\widehat{g}_m\|^2 \\ &\ge J(\theta_m) + \eta_m\|\nabla J(\theta_m)\|^2 - \eta_m\|\nabla J(\theta_m)\|\epsilon_m - \frac{L_J}{8}\eta_m^2 G_m^2, \end{aligned}$$

where the last inequality uses $\|\widehat{g}_m\| \le G_m/2$. Young's inequality gives

$$\|\nabla J(\theta_m)\|\epsilon_m \le \frac{1}{2}\|\nabla J(\theta_m)\|^2 + \frac{1}{2}\epsilon_m^2.$$

Hence

$$\frac{\eta_m}{2}\|\nabla J(\theta_m)\|^2 \le J(\theta_{m+1}) - J(\theta_m) + \frac{\eta_m}{2}\epsilon_m^2 + \frac{L_J}{8}\eta_m^2 G_m^2.$$

Summing over $m = 0, \dots, M-1$, using $J(\theta_M) \le J_{\max} = R_{\max}/(1-\gamma)$, and dividing by $\frac{1}{2}\sum_m \eta_m$ gives

$$\frac{\sum_{m=0}^{M-1} \eta_m \|\nabla J(\theta_m)\|^2}{\sum_{m=0}^{M-1} \eta_m} \le \frac{2(J_{\max} - J(\theta_0))}{\sum_{m=0}^{M-1} \eta_m} + \frac{\sum_{m=0}^{M-1} \eta_m \epsilon_m^2}{\sum_{m=0}^{M-1} \eta_m} + \frac{L_J}{4} \frac{\sum_{m=0}^{M-1} \eta_m^2 G_m^2}{\sum_{m=0}^{M-1} \eta_m}. \tag{67}$$

Now

$$\sum_{m=0}^{M-1} \eta_m = \eta \sum_{m=1}^{M} m^{-1/2} \ge \eta\sqrt{M},$$

and

$$\sum_{m=0}^{M-1} \eta_m^2 = \eta^2 \sum_{m=1}^{M} \frac{1}{m} \le \eta^2(1 + \log M).$$

If $\epsilon_m \le \bar{\epsilon}$ and $G_m \le \bar{G}$ for all m, then

$$\frac{\sum_m \eta_m \epsilon_m^2}{\sum_m \eta_m} \le \bar{\epsilon}^2$$

and

$$\frac{\sum_m \eta_m^2 G_m^2}{\sum_m \eta_m} \le \eta \bar{G}^2 \frac{1 + \log M}{\sqrt{M}}.$$

Substituting these estimates into Equation 67 gives

$$\frac{\sum_{m=0}^{M-1} \eta_m \|\nabla J(\theta_m)\|^2}{\sum_{m=0}^{M-1} \eta_m} \le \frac{2(J_{\max} - J(\theta_0))}{\eta\sqrt{M}} + \bar{\epsilon}^2 + \frac{L_J \eta \bar{G}^2}{4} \frac{1 + \log M}{\sqrt{M}}.$$

The minimum of the nonnegative quantities $\|\nabla J(\theta_m)\|^2$ is bounded by any positive weighted average of them, which proves the theorem. □

F.6 Computational complexity

We derive the cost of one LTC actor–critic iteration. Let

$$p_i = \binom{|\mathcal{N}_i^\kappa| d_s + n}{n} \left[1 + |\mathcal{N}_i^\kappa|(d_s + d_a) + d_g^i\right]$$

denote the LTC feature dimension for agent i, and let

$$n_c = M_c(T-1), \qquad n_s = M_s T.$$

Critic fitting. For each agent i, LSTD forms a $p_i \times p_i$ linear system $\widehat{A}_i w_i = \widehat{b}_i$ from the n_c critic transitions. Constructing $\widehat{A}_i$ costs $O(n_c p_i^2)$, while forming the right-hand-side vector $\widehat{b}_i$ costs only $O(n_c p_i)$. A dense solve then costs $O(p_i^3)$. Hence, fitting all local critics costs

$$\mathcal{O}\left(\sum_{i=1}^{N} \left(n_c p_i^2 + p_i^3\right)\right),$$

while one evaluation of critic i costs $\mathcal{O}(p_i)$.

Actor update. For each actor transition, agent i uses the local advantages $\widehat{A}_{\ell,t}$ for $\ell \in \mathcal{N}_i^{\kappa+\kappa_\pi}$. The actor phase therefore costs

$$\mathcal{O}\left(n_s \sum_{i=1}^{N} \sum_{\ell \in \mathcal{N}_i^{\kappa+\kappa_\pi}} p_\ell\right).$$

For fixed local interaction and policy radii, rollout generation is linear in N, since both the dynamics and policies access only local neighborhoods.

Centroid update. Suppose that K-medians is fitted on n_K sampled global states for I_K clustering iterations. Comparing all samples with all K centroids in $\mathbb{R}^{Nd_s}$ costs

$$\mathcal{O}\left(I_K n_K K N d_s\right),$$

with coordinatewise median updates of the same or lower order.

Combining these terms, one complete LTC iteration has time complexity

$$\mathcal{O}\left(\sum_{i=1}^{N}\left(n_c p_i^2 + p_i^3\right) + n_s \sum_{i=1}^{N} \sum_{\ell \in \mathcal{N}_i^{\kappa+\kappa_\pi}} p_\ell + (M_c + M_s)TN + I_K n_K K N d_s\right).$$

To make the dependence on the network structure explicit, suppose the graph has maximum degree Δ. For fixed state, action, and context dimensions,

$$|\mathcal{N}_i^{\kappa}| = \mathcal{O}(\Delta^{\kappa}), \qquad p_i = \mathcal{O}\left(\Delta^{\kappa(n+1)}\right),$$

and

$$|\mathcal{N}_i^{\kappa+\kappa_\pi}| = \mathcal{O}\left(\Delta^{\kappa+\kappa_\pi}\right).$$

Therefore, one LTC iteration costs

$$\mathcal{O}\left(N\left[n_c \Delta^{2\kappa(n+1)} + \Delta^{3\kappa(n+1)} + n_s \Delta^{\kappa(n+2)+\kappa_\pi} + (M_c + M_s)T + I_K n_K K\right]\right),$$

where fixed state, action, and context dimensions are absorbed into the constants.

For fixed sampling and clustering budgets, this reduces to

$$\boxed{\mathcal{O}\left(N\left[\Delta^{3\kappa(n+1)} + \Delta^{\kappa(n+2)+\kappa_\pi}\right]\right)}.$$

Thus, for fixed local parameters $(\Delta, \kappa, \kappa_\pi, n)$, the total work of one LTC iteration scales linearly with N. The critic and actor computations are also largely parallelizable across agents and samples.

G Proofs of the additional appendix results

This section contains the proofs of the auxiliary results stated in Appendix E. We first treat the discounted-summability extension and its generalized Taylor Centroid approximation bound, then the sufficient conditions based on stochastic and spectral transition models, and finally the comparison with Mild smoothness.

G.1 Discounted-summability results

Proof of Proposition E.1. *Proof.* Fix a multi-index β with $|\beta| \le q$. By Assumption E.1,

$$\sum_{t=0}^{\infty} \sup_{z,s,a} \left|\gamma^t \partial_s^\beta g_{i,t}^\pi(z,s,a)\right| \le \sum_{t=0}^{\infty} \gamma^t \bar{L}(t) < \infty.$$

Hence the series of derivatives converges uniformly. The value series

$$Q_i^\pi(z,s,a) = \sum_{t=0}^{\infty} \gamma^t g_{i,t}^\pi(z,s,a)$$

also converges uniformly because $|g_{i,t}^\pi| \le R_{\max}$. Standard term-by-term differentiation therefore gives

$$\partial_s^\beta Q_i^\pi(z,s,a) = \sum_{t=0}^{\infty} \gamma^t \partial_s^\beta g_{i,t}^\pi(z,s,a),$$

which is Equation 19. Taking absolute values yields

$$\sup_{z,s,a} \left|\partial_s^\beta Q_i^\pi(z,s,a)\right| \le \sum_{t=0}^{\infty} \gamma^t \bar{L}(t),$$

proving Equation 20. □

Proof of Proposition E.2. *Proof.* Fix a nonzero multi-index β and let

$$d := d_i(\beta), \qquad \tau_d := \left\lceil \frac{d}{1+\kappa_\pi} \right\rceil.$$

By Lemma F.1, before time τ_d the expected reward of agent i cannot depend on every state coordinate appearing in β. Therefore

$$\partial_s^\beta g_{i,t}^\pi(z,s,a) = 0, \qquad t < \tau_d.$$

Using Proposition E.1,

$$\begin{aligned}\left|\partial_s^\beta Q_i^\pi(z,s,a)\right| &= \left|\sum_{t=\tau_d}^{\infty} \gamma^t \partial_s^\beta g_{i,t}^\pi(z,s,a)\right| \\ &\le \sum_{t=\tau_d}^{\infty} \gamma^t \bar{L}(t) = \Lambda_d,\end{aligned}$$

which is Equation 22. Since τ_d is nondecreasing in d, Λ_d is nonincreasing. Moreover, $\tau_d \to \infty$ as $d \to \infty$, so $\Lambda_d \to 0$ because $\sum_t \gamma^t \bar{L}(t) < \infty$. □

Proof of Proposition E.3. *Proof.* The value-localization part does not use expected reward smoothness. Hence Lemma F.3 still gives

$$\sup_{z,s,a} \left|Q_i^\pi(z,s,a) - \bar{Q}_i^\pi(z, s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})\right| \le \frac{2R_{\max}}{1-\gamma}\rho^{\kappa+1}. \tag{68}$$

It remains to bound the Taylor remainder of the localized critic. For any multi-index β supported in $\mathcal{N}_i^\kappa$, Proposition E.2 gives, uniformly in the far coordinates,

$$\left|\partial_s^\beta Q_i^\pi(z,s,a)\right| \le \Lambda_{d_i(\beta)}.$$

Differentiation under the probability integral defining the marginalized critic is therefore justified exactly as in Lemma F.4, and yields

$$\left|\partial_s^\beta \bar{Q}_i^\pi(z, s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa})\right| \le \Lambda_{d_i(\beta)}. \tag{69}$$

Fix (z,s,a), let

$$\mu := \mu_{k(s),\mathcal{N}_i^\kappa}, \qquad h := s_{\mathcal{N}_i^\kappa} - \mu, \qquad m := n+1,$$

and use the integral form of the order-n Taylor remainder of $\bar{Q}_i^\pi$ along the segment from μ to $s_{\mathcal{N}_i^\kappa}$. Taking absolute values and applying Equation 69 gives

$$\left|\bar{Q}_i^\pi - \widehat{Q}_i^\pi\right| \le \sum_{\substack{|\beta|=m \\ \mathrm{supp}_G(\beta)\subseteq\mathcal{N}_i^\kappa}} \Lambda_{d_i(\beta)} \frac{|h^\beta|}{\beta!}. \tag{70}$$

For $0 \le d \le \kappa$, define

$$A_{m,d} := \sum_{\substack{|\beta|=m \\ \mathrm{supp}_G(\beta)\subseteq\mathcal{N}_i^d}} \frac{|h^\beta|}{\beta!} = \frac{\|h_{\mathcal{N}_i^d}\|_1^m}{m!}, \qquad A_{m,-1} := 0,$$

where the equality is the multinomial theorem. The contribution of multi-indices with graph reach exactly d is $A_{m,d} - A_{m,d-1}$, so

$$\begin{aligned}\sum_{\substack{|\beta|=m \\ \mathrm{supp}_G(\beta)\subseteq\mathcal{N}_i^\kappa}} \Lambda_{d_i(\beta)} \frac{|h^\beta|}{\beta!} &= \sum_{d=0}^{\kappa} \Lambda_d \big(A_{m,d} - A_{m,d-1}\big) \\ &= \Lambda_\kappa A_{m,\kappa} + \sum_{d=0}^{\kappa-1} (\Lambda_d - \Lambda_{d+1}) A_{m,d} \\ &\le \frac{1}{m!}\left[\Lambda_\kappa W_\kappa^m + \sum_{d=0}^{\kappa-1} (\Lambda_d - \Lambda_{d+1}) W_d^m\right] \\ &= \frac{\mathcal{R}_m^\Lambda}{m!}.\end{aligned}$$

Combining this with Equation 68 and setting $m = n + 1$ proves Equation 24. The monomial count is unchanged from Theorem 1, since the generalized envelope changes only the coefficient bounds and not the set $\mathcal{M}^n_{\mathcal{N}_i^\kappa}$. □

G.2 Smooth stochastic and spectral transition models

Proof of Proposition E.4. *Proof.* Fix an agent j and abbreviate

$$x_j := (s_{\mathcal{N}_j}, a_{\mathcal{N}_j}).$$

Under Equation 26, the conditional density of S'_j is

$$p_j^{\mathrm{tr}}(s'_j \mid z, x_j) = \frac{1}{|\det M_j(z, x_j)|}\, p_j\big(M_j(z, x_j)^{-1}\big(s'_j - f_j(z, x_j)\big)\big)\,,$$

where p_j is the density of η_j from the statement of the proposition. Set

$$u := M_j(z, x_j)^{-1}\big(s'_j - f_j(z, x_j)\big).$$

Because f_j, M_j, and M_j^{-1} are C^q with uniformly bounded derivatives, repeated use of the product and chain rules gives the following standard bound: for every state multi-index α with $|\alpha| = m \le q$,

$$\big|\partial_s^\alpha p_j^{\mathrm{tr}}(s'_j \mid z, x_j)\big| \le \frac{C_m}{|\det M_j(z, x_j)|}(1 + \|u\|)^m \sum_{|\nu| \le m} |\partial^\nu p_j(u)|\,,$$

where C_m is uniform in j, z, x_j and in the network size.

Now change variables from s'_j to u. The Jacobian $|\det M_j|$ cancels the determinant factor above, so

$$\begin{aligned}\int \big|\partial_s^\alpha p_j^{\mathrm{tr}}(s'_j \mid z, x_j)\big|\, ds'_j &\le C_m \sum_{|\nu| \le m} \int (1 + \|u\|)^m\, |\partial^\nu p_j(u)|\, du \\ &\le C'_m,\end{aligned}$$

where $C'_m < \infty$ is uniform by assumption. Thus every local transition density has state derivatives through order q whose L^1 norms are uniformly bounded.

We next use the graph factorization. Conditional on the current global state, the density of the distributed next state is

$$\prod_{j=1}^{N} p_j^{\mathrm{tr}}\left(s'_j \mid z, s_{\mathcal{N}_j}, a_{\mathcal{N}_j}\right).$$

Let β be a state multi-index of total order $m \le q$. A scalar state coordinate of agent ℓ appears only in transition factors associated with agents j such that $\ell \in \mathcal{N}_j$. Since the graph degree is bounded by Δ, there are at most $\Delta + 1$ such factors. Applying the product rule one differentiation at a time therefore produces at most $(\Delta + 1)^m$ terms. In each term, the undifferentiated factors integrate to one, while the differentiated factors have uniformly bounded L^1 norms by the preceding calculation. By Fubini's theorem, there is therefore a constant $C_q < \infty$, depending only on q, Δ, and the local bounds, such that for every bounded measurable H,

$$\sup_{z,s,a} \big|\partial_s^\beta \mathbb{E}[H(Z_1, S_1) \mid Z_0 = z, S_0 = s, A_0 = a]\big| \le C_q \|H\|_\infty, \qquad |\beta| \le q. \tag{71}$$

The uniform L^1 bounds above also justify differentiating under the integral.

For $t \ge 1$, define

$$H_{i,t-1}(z', s') := \mathbb{E}_\pi[r_i(S_{i,t}, A_{i,t}) \mid Z_1 = z', S_1 = s']\,.$$

Since $|r_i| \le R_{\max}$, we have $\|H_{i,t-1}\|_\infty \le R_{\max}$. By the Markov property,

$$g^\pi_{i,t}(z, s, a) = \mathbb{E}[H_{i,t-1}(Z_1, S_1) \mid Z_0 = z, S_0 = s, A_0 = a]\,.$$

Applying Equation 71 gives

$$\sup_{z,s,a} \big|\partial_s^\beta g^\pi_{i,t}(z, s, a)\big| \le C_q R_{\max}, \qquad t \ge 1, \qquad |\beta| \le q.$$

At $t = 0$,

$$g^{\pi}_{i,0}(z, s, a) = r_i(s_i, a_i),$$

and its state derivatives through order q are uniformly bounded by the first hypothesis of the proposition. Taking the larger of these two bounds gives Assumption 1, with a constant independent of N.

If Z takes finitely many values, the same proof applies mode by mode; uniform modewise bounds give the same conclusion for hidden or observed finite-mode systems. □

Proof of Proposition E.5. *Proof.* For the local spectral representation Equation 27, a state derivative acts only on the current-state factors $\varphi_{j,h}$. Thus, for every local state multi-index α with $|\alpha| \le q$,

$$\partial^{\alpha}_{s} p_j(s'_j \mid s_{\mathcal{N}_j}, a_{\mathcal{N}_j}, z) = \sum_{h=1}^{d} \partial^{\alpha}_{s} \varphi_{j,h}(s_{\mathcal{N}_j}, a_{\mathcal{N}_j}, z)\, \mu_{j,h}(s'_j).$$

Consequently,

$$\begin{aligned}\int \left|\partial^{\alpha}_{s} p_j(s'_j \mid s_{\mathcal{N}_j}, a_{\mathcal{N}_j}, z)\right| ds'_j &\le \sum_{h=1}^{d} \left|\partial^{\alpha}_{s} \varphi_{j,h}(s_{\mathcal{N}_j}, a_{\mathcal{N}_j}, z)\right| \|\mu_{j,h}\|_{L^1} \\ &\le C_{\mathrm{spec}},\end{aligned}$$

where C_{spec} is uniform by the assumptions of the proposition.

We can now repeat the same product-rule argument as above. Because a state coordinate appears in at most $\Delta + 1$ local transition factors, the L^1 norm of every derivative through order q of the factorized joint next-state density is bounded by a constant independent of N. Therefore, for every bounded measurable H,

$$\sup_{z,s,a} \left|\partial^{\beta}_{s} \mathbb{E}[H(Z_1, S_1) \mid Z_0 = z, S_0 = s, A_0 = a]\right| \le C^{\mathrm{spec}}_{q} \|H\|_{\infty}, \qquad |\beta| \le q.$$

For $t \ge 1$, use the bounded continuation function

$$H_{i,t-1}(z', s') := \mathbb{E}_{\pi}[r_i(S_{i,t}, A_{i,t}) \mid Z_1 = z', S_1 = s'], \qquad \|H_{i,t-1}\|_{\infty} \le R_{\max}.$$

The Markov property gives

$$g^{\pi}_{i,t}(z, s, a) = \mathbb{E}[H_{i,t-1}(Z_1, S_1) \mid Z_0 = z, S_0 = s, A_0 = a],$$

and hence

$$\sup_{z,s,a} \left|\partial^{\beta}_{s} g^{\pi}_{i,t}(z, s, a)\right| \le C^{\mathrm{spec}}_{q} R_{\max}, \qquad t \ge 1.$$

At $t = 0$, $g^{\pi}_{i,0} = r_i$, whose derivatives through order q are uniformly bounded by assumption. This proves Assumption 1 with an N-independent constant. □

G.3 Relation to Mild smoothness

Proof of Proposition E.6. *Proof.* Let $x = (z, s)$ denote the augmented state. For a bounded measurable function $H(x')$, define the state–action function

$$F_H(x', a') := H(x').$$

Since F_H does not depend on the action, $\sup_{a'} F_H(x', a') = H(x')$. Hence, with the Bellman operator used in Appendix E,

$$(\mathcal{T}^{\star} F_H)(x, a) - (\mathcal{T}^{\star} 0)(x, a) = \mathbb{E}[H(X_1) \mid X_0 = x, A_0 = a].$$

If the MDP is ν-Mildly Smooth, both Bellman terms belong to $\mathcal{C}^{\nu}$ and

$$\|\mathbb{E}[H(X_1) \mid X_0 = \cdot, A_0 = \cdot]\|_{\mathcal{C}^{\nu}} \le C_{\mathcal{T}}(\|H\|_{\infty} + 2).$$

Thus one transition maps an arbitrary bounded continuation function to a uniformly smooth function of the current state–action pair.

For $t \geq 1$, take

$$H_{i,t-1}(x') := \mathbb{E}_\pi[r_i(S_{i,t}, A_{i,t}) \mid X_1 = x'] .$$

This function is bounded by $R_{\max}$, and the Markov property gives

$$g^\pi_{i,t}(x, a) = \mathbb{E}[H_{i,t-1}(X_1) \mid X_0 = x, A_0 = a] .$$

Therefore

$$\sup_{t\geq 1} \|g^\pi_{i,t}\|_{\mathcal{C}^\nu} \leq C_{\mathcal{T}}(R_{\max} + 2).$$

Since $\nu > q$, the $\mathcal{C}^\nu$ norm controls the classical state derivatives through order q on the compact domain. Hence

$$\sup_{t\geq 1} \|g^\pi_{i,t}\|_{C^q_s} < \infty.$$

At $t = 0$, $g^\pi_{i,0} = r_i$, which is C^q with uniformly bounded derivatives by assumption. This proves Assumption 1 of order q. The integer-order implication Equation 29 follows immediately from Equation 28. □

Proof of Proposition E.7. *Proof.* For fixed z, the physical drift

$$f(z, s, a) = A^{(\sigma(z))}s + B^{(\sigma(z))}a$$

is linear in the physical state and action. Since there are only two modes, its physical-state derivatives are uniformly bounded over z. With $M \equiv I$, the noise assumptions are exactly those of Proposition E.4, uniformly over the two regimes. That proposition therefore yields Assumption 1 at every finite order q. Importantly, no smoothness of the policy is needed: after the first transition, the future policy is absorbed into a bounded continuation value, while differentiation with respect to the initial physical state acts only on the smooth transition density. Hence the bound is uniform over all admissible policies.

It remains to show that the augmented MDP is not Mildly Smooth. Let $x = (z, s)$ and use the nonzero vector d from the statement to define the smooth bounded Bellman input

$$F(x', a') = d^\top s'.$$

The state domain is compact, so F is bounded, and it is independent of the next action. Since the noise is centered,

$$(\mathcal{T}^\star F)((z, s), a) = r(s, a) + d^\top \left(A^{(\sigma(z))}s + B^{(\sigma(z))}a\right).$$

Evaluating this function at $(s, a) = (\bar{s}, \bar{a})$, its one-sided values at $z = 0$ differ by

$$d^\top \left[\left(A^{(1)} - A^{(0)}\right)\bar{s} + \left(B^{(1)} - B^{(0)}\right)\bar{a}\right] = \|d\|^2 > 0.$$

Thus $\mathcal{T}^\star F$ is discontinuous in the continuous context coordinate z. In particular, it does not belong to $\mathcal{C}^\nu$ for any $\nu > 0$. Therefore the augmented MDP is not Mildly Smooth at any positive order. □

H Neural Network Implementation Details

H.1 NTC Implementation

NTC retains the graph-local Taylor structure of LTC and only changes the parameterization of the Taylor coefficients. Recall the local Taylor feature vector

$$T_i(s) := \left[\frac{\left(s_{\mathcal{N}_i^\kappa} - \mu_{k(s),\mathcal{N}_i^\kappa}\right)^\beta}{\beta!}\right]_{\beta\in\mathcal{M}^n_{\mathcal{N}_i^\kappa}},$$

and the coefficient context

$$\psi_i\big(k(s), y_i, a_{\mathcal{N}_i^\kappa}\big) = \left[1; \mu_{k(s),\mathcal{N}_i^\kappa}; a_{\mathcal{N}_i^\kappa}; y_i\right].$$

LTC models each Taylor coefficient as an affine function of this context. NTC replaces this affine map by a neural coefficient map

$$f_{\xi_i} : \mathbb{R}^{d_\psi^i} \longrightarrow \mathbb{R}^{|\mathcal{M}^n_{\mathcal{N}_i^\kappa}|},$$

and defines

$$\widetilde{Q}_i^{\mathrm{NTC}}\big(y_i, s, a_{\mathcal{N}_i^\kappa}; \xi_i\big) = \big\langle f_{\xi_i}\big(\psi_i(k(s), y_i, a_{\mathcal{N}_i^\kappa})\big), T_i(s)\big\rangle . \tag{72}$$

At augmented state (z, s), we set $y_i = h_i(z)$.

Thus, the critic remains an order-n polynomial in the local state around the current centroid, while its coefficients may depend nonlinearly on the local actions, operating regime, and observable context. LTC is recovered as the special case in which f_{ξ_i} is affine.

Architecture. We use one independent coefficient network for each agent. The N networks have disjoint parameters but are stored in a single batched implementation, so that all agents can be evaluated in parallel. Each coefficient network is a two-hidden-layer MLP with hidden width 64, $\tanh$ activations, and agent-wise Kaiming initialization (He et al., 2015).

Importantly, the network does not receive the raw current state as input. The state dependence enters explicitly through the Taylor vector $T_i(s)$; the neural network only receives the coefficient context ψ_i. Hence, increasing the Taylor order changes the output dimension of the network, rather than its input dimension.

TD training. NTC is trained by on-policy semi-gradient TD using SARSA targets and a periodically synchronized target network. Let $\bar{\xi}_i$ denote the target parameters for agent i. For a transition

$$(Z_t, S_t, A_t, r_{i,t}, Z_{t+1}, S_{t+1}, A_{t+1}),$$

the target is

$$\tau_{i,t} = r_{i,t} + \gamma \widetilde{Q}_i^{\mathrm{NTC}}\left(h_i(Z_{t+1}), S_{t+1}, A_{t+1,\mathcal{N}_i^\kappa}; \bar{\xi}_i\right), \tag{73}$$

where the target-network evaluation is detached from the gradient. Given M_c critic rollouts of length T, the online parameters minimize

$$\mathcal{L}_i^{\mathrm{NTC}}(\xi_i) = \frac{1}{M_c(T-1)} \sum_{j=1}^{M_c} \sum_{t=0}^{T-2} \left[\widetilde{Q}_i^{\mathrm{NTC}}\left(h_i(Z_t^{(j)}), S_t^{(j)}, A_{t,\mathcal{N}_i^\kappa}^{(j)}; \xi_i\right) - \tau_{i,t}^{(j)}\right]^2 . \tag{74}$$

We optimize this objective with Adam using a learning rate of 2×10^{-3}. The target network is synchronized every 25 TD updates,

$$\bar{\xi}_i \leftarrow \xi_i,$$

and remains fixed between synchronization steps. We use 50, 200, and 200 TD optimization steps for Taylor orders $n = 1, 2, 3$, respectively as higher parameter counts require more steps to fit. On Kuramoto, NTC uses learning rate 10^{-3}, 50 TD updates per iteration on batches of 1024 transitions, and target synchronization every 50 updates. Agents are processed in blocks during training; each block operates only on the corresponding disjoint agent parameters.

State-value estimate for one-step TD advantages. NTC parameterizes an action-value function whose dependence on the action is generally nonlinear through f_{ξ_i}. Its state-value estimate is therefore obtained by marginalizing the fitted critic under the current policy,

$$\widetilde{V}_i(h_i(z), s) = \mathbb{E}_{a\sim\pi(\cdot|z,s)}\left[\widetilde{Q}_i^{\mathrm{NTC}}\big(h_i(z), s, a_{\mathcal{N}_i^\kappa}\big)\right].$$

In the implementation, this expectation is approximated using $n_{\mathrm{mc}} = 16$ independent action samples from the current policy. The resulting $\widetilde{V}_i$ is then used in the localized one-step TD advantage estimator described in Section 4. With independent action samples, this Monte Carlo marginalization is unbiased and can be absorbed into the actor-rollout randomness in the analysis. For LTC, this expectation is simpler because the critic is affine in the actions; therefore, only the expected action is needed to compute the value function.

H.2 LOCAL MLP CRITIC

As a generic local function-approximation baseline, we use one independent multilayer perceptron for each agent. The critic directly approximates the local state-value function

$$\widetilde{V}_i(s) = f_{\xi_i}\left(s_{\mathcal{N}_i^\kappa}\right),$$

where the input is the same κ-hop state neighborhood used by the Taylor Centroid critics. On the scalar ring benchmarks, this corresponds to input dimensions 3, 5, and 7 for $\kappa = 1, 2, 3$, respectively.

Each f_{ξ_i} is a two-hidden-layer MLP with $\tanh$ activations and hidden width 128 in the policy experiments on LQR and switched LQR, and 64 on Kuramoto and in the critic-estimation experiment. Parameters are not shared across agents: the implementation stores N separate sets of weights and evaluates them in parallel through a batched MLP.

The critic is trained on-policy using TD(0) with a target network. For a transition (s_t, r_t, s_{t+1}), the target for agent i is

$$y_{i,t} = r_{i,t} + \gamma \widetilde{V}_i^{\,\mathrm{targ}}(s_{t+1}),$$

and the training objective is

$$\mathcal{L}_{\mathrm{MLP}} = \frac{1}{N|\mathcal{B}|} \sum_{i=1}^{N} \sum_{t \in \mathcal{B}} \left(\widetilde{V}_i(s_t) - y_{i,t}\right)^2.$$

Training hyperparameters are given in Appendix C. Local state inputs are standardized using rollout statistics, and value targets are rescaled using a fixed scale initialized from the first rollout.

The critic directly outputs $\widetilde{V}_i(s)$ and is combined with the same localized actor and one-step TD advantage estimator used for the Taylor Centroid methods.

H.3 GRAPH NEURAL-NETWORK CRITIC

As a generic graph-based baseline, we use a message-passing neural network to approximate the local state-value functions. Each node is initialized from its own state,

$$h_i^{(0)} = s_i,$$

and neighborhood information is propagated through mean aggregation. A message-passing layer takes the form

$$h_i^{(\ell+1)} = \tanh\left(W_{\mathrm{self}}^{(\ell)} h_i^{(\ell)} + W_{\mathrm{nbr}}^{(\ell)} \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} h_j^{(\ell)} + b^{(\ell)}\right).$$

A shared linear readout produces the local value estimate

$$\widetilde{V}_i(s) = w_{\mathrm{out}}^\top h_i + b_{\mathrm{out}}.$$

The network has $L = \max\{2, \kappa\}$ message-passing layers, each with its own weights $(W_{\mathrm{self}}^{(\ell)}, W_{\mathrm{nbr}}^{(\ell)}, b^{(\ell)})$: weights are not reused across layers, and each layer performs at most one round of one-hop aggregation. The neighbor term is included in the first κ layers only, so that $\widetilde{V}_i$ depends only on $s_{\mathcal{N}_i^\kappa}$: every layer aggregates for $\kappa \in \{2, 3\}$, whereas for $\kappa = 1$ the second layer acts on each node separately. All weights are shared across agents.

In the policy-optimization experiments on LQR and switched LQR, the hidden width is 128, which gives $17{,}025$, $33{,}409$ and $66{,}305$ parameters for $\kappa = 1, 2, 3$; since all weights are shared across agents, this number does not depend on N. In the critic-estimation experiment, the width is chosen so that the parameter count matches that of the MLP used there. For policy optimization on Kuramoto, the GNN replaces the random-feature map inside the RF-SAC training loop; there it has two layers of width 64 at every κ, hence a receptive field of $\min\{\kappa, 2\}$, and SAC's twin critics are two such networks.

On LQR and switched LQR, the GNN critic is trained on-policy by TD(0) with a target network. For each agent i,

$$y_{i,t} = r_{i,t} + \gamma \widetilde{V}_{i,\mathrm{targ}}(s_{t+1}),$$

and the training loss is

$$\mathcal{L}_{\mathrm{GNN}} = \frac{1}{N|\mathcal{B}|} \sum_{i=1}^{N} \sum_{t \in \mathcal{B}} \left(\widetilde{V}_i(s_t) - y_{i,t} \right)^2 .$$

Training hyperparameters are given in Appendix C. State inputs are standardized using statistics from the rollout distribution, and value targets use the same fixed rescaling as the local MLP baseline.

For policy optimization, this critic is paired with the actor of Ren et al. (2025b) on LQR and Kuramoto, taken from their released code. On switched LQR, it is paired with our actor.