# LOCALIZED DEEP NEURAL ACTOR-CRITIC FOR DECENTRALIZED MARL: FINITE-TIME GLOBAL CONVERGENCE GUARANTEES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Reinforcement learning on decentralized multi-agent systems is limited by a global state–action space that grows exponentially with the number of agents. Even a local action–value can depend on the whole of it, but local dynamics and a localized policy limit how fast that dependence spreads outward. Over such a policy class, the influence of a distant agent on a local action-value diminishes exponentially with graph distance, so value components and policy-gradient directions admit finite-hop approximations. Existing finite-time analyses exploit this structure with tabular, aggregated, or spectral critics, whereas deep critics have so far been analyzed with global inputs. In this paper, we propose a localized deep neural actor–critic method in which each agent trains a private finite-hop critic and exchanges scalar value estimates within its neighborhood. Because the finite-hop observation process is in general not Markovian, we use a conditional Bellman operator under the discounted occupancy distribution, whose fixed point is learnable by local temporal-difference updates and stays close to the value that the policy gradient requires. We establish a finite-sample bound for the projected neural critic in terms of its linearization and tangent-approximation errors, and propagate it into an $\mathcal{O}(1/\sqrt{T})$ rate for the projected gradient mapping and, under a gradient-domination condition on the localized class, for the policy gap above an explicit error floor, with each agent's critic input dimension and message count scaling with its $\kappa$-hop neighborhood rather than with the network. Experiments on ring and wireless random-access networks corroborate our theoretical findings.

## 1 INTRODUCTION

Many large-scale control systems consist of agents coupled through a physical or communication network (Olfati-Saber et al., 2007; Cao et al., 2013). Examples include wireless power allocation (Nasir & Guo, 2019), traffic signal coordination (Chu et al., 2020), voltage regulation (Wang et al., 2021), and edge-computing resource allocation (Liu et al., 2020). Networked Markov decision processes (MDPs) model such systems by attaching local states, actions, rewards, and transition dependencies to the nodes and edges of a graph (Qu et al., 2020b; Zhang et al., 2021). One-step interactions are local, but their effects propagate over time. A local action-value may therefore depend on the entire state–action configuration, whose size grows exponentially with the number of agents, and global value estimation becomes prohibitive at scale.

Local dynamics and a localized policy nonetheless bound how fast that dependence spreads. Over such a class, the exponential decay property states that the influence of a distant agent on a local action-value diminishes exponentially with graph distance (Qu et al., 2020a;b), so each agent has a window beyond which the rest of the network barely registers (Figure 1, panel 1). Value components and policy-gradient directions therefore admit finite-hop approximations. A line of localized methods for networked MDPs builds on this structure (Qu et al., 2020a;b; Lin et al., 2021; Zhang et al., 2023; Zhou et al., 2023; Ren et al., 2025).

That line and the deep critics used in practice have developed apart. The localized methods keep the finite-hop structure, but the critics they analyze are tabular (Qu et al., 2020a;b), aggregated (Lin et al., 2021), or spectral (Ren et al., 2025; Wang et al., 2026), so the representation is fixed in advance



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.

Table 1: **Representative related guarantees.** Settings and comparators differ across these works.

| Work | Agent network | Finite-hop critic | Trainable critic | Guarantee |
|---|---|---|---|---|
| Qu et al. (2020a;b) | ✓ | ✓ | × | Stationarity |
| Zhang et al. (2023) | ✓ | ✓ | × | Policy gap‡ |
| Ren et al. (2025) | ✓ | ✓ | × | Stationarity |
| Gu et al. (2025) | × | ✓ | ✓ | Policy gap |
| Zhang et al. (2025) | ✓ | × | ✓ | Policy gap |
| Wang et al. (2026) | ✓ | ✓ | × | Stationarity |
| **This work** | ✓ | ✓ | ✓ | Policy gap |

"Agent network" is a fixed set of heterogeneous agents on a graph, "finite-hop critic" bounds the critic input to a graph neighborhood, "trainable critic" means a learned rather than fixed representation, and "policy gap" denotes suboptimality relative to each work's own comparator (Appendix A.1).
‡ Polynomial rather than geometric locality term in the hop radius.

rather than learned; the last of these reaches the conditional Bellman target with a linear random-feature critic. The analyses that do train a neural critic give up the finite-hop structure instead. Zhang et al. (2025) obtain a finite-time policy guarantee for decentralized MARL from a critic that reads the joint state, and Gu et al. (2025) analyze a two-layer critic on a mean-field network of states rather than on a fixed set of agents. Deep MARL as practiced sits on the same side of this divide, since MADDPG (Lowe et al., 2017), QMIX (Rashid et al., 2018), and MAPPO (Yu et al., 2022) all train neural value models that read centralized state, which is precisely the cost that finite-hop inputs avoid. Table 1 places these guarantees side by side, and no entry on a fixed agent network combines finite-hop inputs with a critic whose representation is learned. This paper asks whether such a critic can support a finite-time policy guarantee in which the localization, critic-learning, and distributed-implementation errors are kept explicit, with the architecture entering through named network-geometry constants rather than being assumed away.

We answer it with a localized deep neural actor–critic, in which each agent trains a private critic on the state–action variables in its $\kappa$-hop neighborhood and exchanges only scalar critic values within that neighborhood (Figure 1, panels 2 and 3). What such a critic should be trained toward, however, is not immediate. The global state–action process is Markov, but its restriction to a finite-hop neighborhood is generally not, since variables just outside the window enter it at the next step (Figure 1, panel 4). Consequently, the local observation process does not admit an autonomous Bellman equation for standard TD learning. Moreover, although the conditional expectation of the global action-value is finite-hop, it need not coincide with the fixed point induced by local TD samples.

The policy-dependent conditional Bellman operator resolves this mismatch. Its fixed point is induced by the local TD samples, while spatial decay bounds its discrepancy from the global local-reward value required by the policy gradient. We control the neural critic-learning error around this target at finite width and finitely many samples, and propagate the resulting localization and estimation errors into a finite-time policy-gap guarantee.

In this paper, we overcome the aforementioned challenges by establishing a series of new theoretical understandings on localized deep neural actor-critic MARL. We summarize our key contributions in this paper as follows:

- **Localized neural actor–critic.** Each agent's critic reads only the state–action variables in its $\kappa$-hop neighborhood, and the actor aggregation step exchanges only scalar critic outputs within that neighborhood. The conditional Bellman operator induced by the true global transition process gives this critic a well-defined and learnable target even though the local observation process is not Markovian.
- **Finite-time theory with an explicit error budget.** We analyze projected neural TD by separating its temporal-difference optimization error from the architecture-dependent linearization and approximation errors. The resulting critic accuracy holds at finite width and finitely many TD samples, stated in terms of those errors rather than assuming the learned critic has already converged. We propagate it into an $\mathcal{O}(1/\sqrt{T})$ rate for the projected gradient mapping and, under a gradient-domination condition, into a guarantee for the expected gap at a uniformly sampled

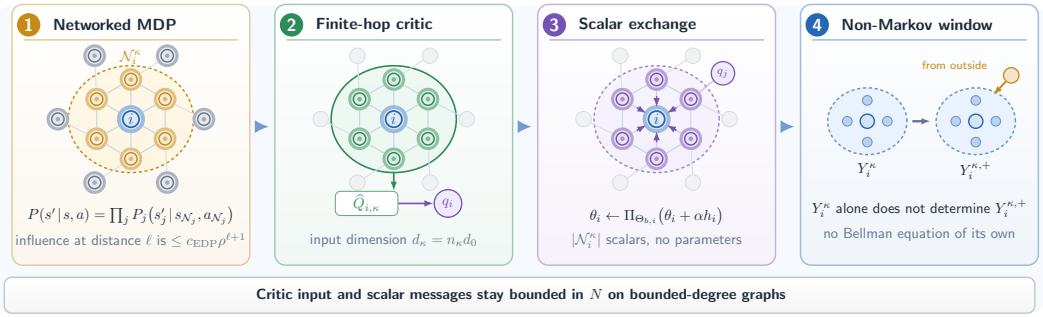



Figure 1: **Localized neural actor–critic.** Panel 1 fixes the networked MDP and the $\kappa$-hop window of agent $i$. Panels 2 and 3 light up the state–action variables the critic reads and the scalar critic values the neighbours send. Panel 4 shows why that window is not a Markov state on its own. The bar below records the per-agent cost.

iterate, measured against the best policy in the $b$-localized class, above an explicit error floor that separates the localization, critic-estimation, actor-sampling, and compatible-advantage contributions.

- **Per-agent locality and numerical study.** For a fixed radius on uniformly bounded-degree graph families, each agent's critic input dimension and value-message count scale with its $\kappa$-hop neighborhood rather than with the network. Experiments on ring MDPs and wireless random-access networks isolate the effect of critic input scope at matched budgets.

## 2 Problem Formulation and Preliminaries

In this section, we first formulate the networked MDP and the $b$-localized policy class, and then state the value decomposition, the localized policy gradient, and the spatial decay condition.

### 2.1 Networked MDP and Localized Policies

Consider $N$ agents associated with the nodes of a connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} := \{1, \dots, N\}$, and let $\mathrm{dist}(i, j)$ denote the graph distance between agents $i$ and $j$. For a nonnegative integer $r$, the $r$-hop neighborhood of agent $i$, including $i$ itself, is $\mathcal{N}_i^r := \{j \in \mathcal{V} : \mathrm{dist}(i, j) \leq r\}$, and we write $\mathcal{N}_i := \mathcal{N}_i^1$. Agent $i$ has a local state $S_i \in \mathcal{S}_i$ and a local action $A_i \in \mathcal{A}_i$, where the spaces are finite. We write $S = (S_1, \dots, S_N)$ and $A = (A_1, \dots, A_N)$ for the global state and action, with global spaces $\mathcal{S} := \prod_{i=1}^N \mathcal{S}_i$ and $\mathcal{A} := \prod_{i=1}^N \mathcal{A}_i$. For any node set $\mathcal{B} \subseteq \mathcal{V}$, we write $s_{\mathcal{B}} := (s_j)_{j\in\mathcal{B}}$ and $a_{\mathcal{B}} := (a_j)_{j\in\mathcal{B}}$ for the corresponding local configurations, and $A_{-i} := (A_j)_{j\neq i}$.

Conditioned on the current global state-action pair, the next local states are independent, and the transition of each agent depends only on its one-hop neighborhood:

$$P(s' \mid s, a) = \prod_{i=1}^{N} P_i(s'_i \mid s_{\mathcal{N}_i}, a_{\mathcal{N}_i}) \,. \tag{1}$$

Agent $i$ receives a bounded local reward $r_i(s_{\mathcal{N}_i^{r_R}}, a_{\mathcal{N}_i^{r_R}})$, where the reward radius $r_R$ is a nonnegative integer and $|r_i| \leq R_{\max}$. We write $r_i(s, a)$ when the local arguments are clear. The cooperative stage reward is the average $\bar{r}(s, a) := \frac{1}{N}\sum_{i=1}^N r_i(s, a)$.

At execution time, each agent selects its action using only local state information. Given a nonnegative integer policy radius $b$, we therefore consider the $b$-localized policy class

$$\pi_\theta(a \mid s) = \prod_{i=1}^{N} \pi_{\theta_i}\left(a_i \mid s_{\mathcal{N}_i^b}\right), \qquad \theta = (\theta_1, \dots, \theta_N) \in \Theta_b, \tag{2}$$

where $\Theta_b := \prod_{i=1}^N \Theta_{b,i}$ with blocks $\Theta_{b,i} := \{\theta_i \in \mathbb{R}^{d_{\pi,i}} : \|\theta_i\|_2 \leq R_\Theta\}$ and a fixed radius $R_\Theta < \infty$. Let $\mu_0$ be the initial-state distribution and let $\gamma \in (0, 1)$ be the discount factor. The objective

is to maximize the discounted team return (Sutton & Barto, 2018), and we measure performance against the best return achievable within the parameterized $b$-localized class:

$$J(\theta) \coloneqq \mathbb{E}_{\mu_0,\pi_\theta,P}\left[\sum_{t=0}^{\infty}\gamma^t \bar{r}(S_t,A_t)\right], \qquad \text{and} \qquad J_b^\star \coloneqq \sup_{\theta\in\Theta_b} J(\theta). \tag{3}$$

### 2.2 Value decomposition and spatial decay

For a fixed policy $\pi_\theta$, let $Q_i^\theta$ denote the action-value associated with agent $i$'s reward stream and let $Q^\theta$ denote the action-value associated with the team reward:

$$Q_i^\theta(s,a) \coloneqq \mathbb{E}_{\pi_\theta,P}\left[\sum_{t=0}^{\infty}\gamma^t r_i(S_t,A_t)\,\middle|\, S_0=s,\ A_0=a\right], \tag{4}$$

$$Q^\theta(s,a) \coloneqq \mathbb{E}_{\pi_\theta,P}\left[\sum_{t=0}^{\infty}\gamma^t \bar{r}(S_t,A_t)\,\middle|\, S_0=s,\ A_0=a\right] = \frac{1}{N}\sum_{j=1}^{N} Q_j^\theta(s,a). \tag{5}$$

The second equality follows from the linearity of $\bar{r}$. Note that $Q_i^\theta$ itself may depend on the entire global state–action pair, even though $r_i$ does not. Writing $Q_{\max} \coloneqq R_{\max}/(1-\gamma)$, bounded rewards imply $|Q_i^\theta(s,a)| \le Q_{\max}$ and $|Q^\theta(s,a)| \le Q_{\max}$. Let $d_{\theta,t}$ be the law of $S_t$ under $(\mu_0,P,\pi_\theta)$, let $d_\theta \coloneqq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t d_{\theta,t}$ be the normalized discounted state visitation distribution, and let $\nu_\theta$ be the induced discounted state–action occupancy distribution, so that $(S,A)\sim\nu_\theta$ means $S\sim d_\theta$ followed by $A\sim\pi_\theta(\cdot\mid S)$, and we write $\nu_\theta(s,a) = d_\theta(s)\pi_\theta(a\mid s)$.

Assume each $\pi_{\theta_i}$ is differentiable in $\theta_i$ with $\theta_i$-independent support, and define the policy score of agent $i$ by $\psi_i^\theta(s,a_i) \coloneqq \nabla_{\theta_i}\log\pi_{\theta_i}(a_i \mid s_{\mathcal{N}_i^b})$ (Williams, 1992). The product form in (2) gives the score identity $\mathbb{E}[\psi_i^\theta(S,A_i)\mid S,A_{-i}] = 0$, and the policy-gradient theorem (Sutton et al., 1999), together with (5), then yields

$$\nabla_{\theta_i} J(\theta) = \frac{1}{N(1-\gamma)}\mathbb{E}_{(S,A)\sim\nu_\theta}\left[\psi_i^\theta(S,A_i)\sum_{j=1}^{N} Q_j^\theta(S,A)\right]. \tag{6}$$

Therefore, despite localized policy execution, the exact gradient block of agent $i$ depends on the global team action-value, written here as the sum of the local reward values.

Spatial decay controls this long-range dependence. For any nonnegative integer $\ell$, let $y_i^\ell(s,a) \coloneqq (s_{\mathcal{N}_i^\ell}, a_{\mathcal{N}_i^\ell})$ denote the $\ell$-hop state–action configuration around agent $i$.

**Definition 2.1** (Exponential decay property)**.** The local action-values satisfy the $(c_{\mathrm{EDP}},\rho)$ exponential decay property over $\Theta_b$ if there exist constants $c_{\mathrm{EDP}} > 0$ and $\rho\in(0,1)$ such that, for every $\theta\in\Theta_b$, agent $i$, integer $\ell \ge r_R$, and any two global state–action pairs $(s,a)$ and $(\widetilde{s},\widetilde{a})$ satisfying $y_i^\ell(s,a) = y_i^\ell(\widetilde{s},\widetilde{a})$, we have

$$\left|Q_i^\theta(s,a) - Q_i^\theta(\widetilde{s},\widetilde{a})\right| \le c_{\mathrm{EDP}}\rho^{\ell+1}. \tag{7}$$

We impose this property throughout. It is standard in scalable networked MARL (Qu et al., 2020a;b; Lin et al., 2021; Ren et al., 2025) and is closely related to correlation decay in large decision networks (Gamarnik et al., 2014). Under the local transition, reward, and policy structures considered here, a finite-propagation argument gives the sufficient condition $\rho = \gamma^{1/(1+b)}$, with a constant that absorbs the reward radius (Appendix B.1).

## 3 The Localized Neural Actor–Critic Method

In this section, we first localize the policy-gradient direction, then define its finite-hop critic target, and finally specify the neural critic and actor update.

### 3.1 LOCALIZED POLICY-GRADIENT DIRECTION

Fix a critic radius $\kappa \geq \max\{r_R, b\}$, so that agent $i$'s $\kappa$-hop configuration carries both the arguments of $r_i$ and the state block $s_{\mathcal{N}_i^b}$ on which its own score depends. For $(S, A) \sim \nu_\theta$, write $Y_i^\kappa := y_i^\kappa(S, A)$ and let $\nu_{i,\kappa,\theta}$ denote the marginal distribution of $Y_i^\kappa$. We first project the global local-reward action-value onto the information available in agent $i$'s $\kappa$-hop neighborhood, $\widetilde{Q}_{i,\kappa}^\theta(y) := \mathbb{E}_{\nu_\theta}[Q_i^\theta(S, A) \mid Y_i^\kappa = y]$. Under the exponential decay property, this conditional value approximates the global local-reward value (Lemma C.1):

$$\left| Q_i^\theta(s, a) - \widetilde{Q}_{i,\kappa}^\theta(y_i^\kappa(s, a)) \right| \leq c_{\mathrm{EDP}} \rho^{\kappa+1}. \tag{8}$$

We now replace the global value components in (6) by their conditional approximations and retain only the agents within $\kappa$ hops, which gives the ideal localized direction

$$g_{i,\kappa}(\theta) := \frac{1}{N(1-\gamma)} \mathbb{E}_{\nu_\theta} \left[ \psi_i^\theta(S, A_i) \sum_{j \in \mathcal{N}_i^\kappa} \widetilde{Q}_{j,\kappa}^\theta(Y_j^\kappa) \right]. \tag{9}$$

After replacing every global value component by its conditional truncation, the terms with $j \notin \mathcal{N}_i^\kappa$ vanish in expectation. Indeed, for such $j$ undirectedness gives $i \notin \mathcal{N}_j^\kappa$, so $\widetilde{Q}_{j,\kappa}^\theta(Y_j^\kappa)$ is measurable with respect to $(S, A_{-i})$ and the conditional score identity applies. The cancellation concerns the truncated components; the original global values can still depend on $A_i$. The only remaining discrepancy is the finite-hop approximation error, which we bound next.

**Lemma 3.1** (Localized-gradient bias)**.** *Suppose $\kappa \geq r_R$, Definition 2.1 holds, and $\|\psi_i^\theta(s, a_i)\|_2 \leq L_{\psi,i}$ for every $\theta \in \Theta_b$. With $L_\psi := (\sum_i L_{\psi,i}^2)^{1/2}$ and $g_\kappa(\theta) := (g_{1,\kappa}(\theta), \ldots, g_{N,\kappa}(\theta))$,*

$$\|g_\kappa(\theta) - \nabla J(\theta)\|_2 \leq \frac{L_\psi c_{\mathrm{EDP}}}{1-\gamma} \rho^{\kappa+1}. \tag{10}$$

The proof is deferred to Appendix C.2; the lemma isolates the spatial approximation error before any critic or sampling error enters.

### 3.2 CONDITIONAL FINITE-HOP BELLMAN TARGET

The local process $Y_i^\kappa$ is in general not Markov, since its transition law can depend on variables outside $\mathcal{N}_i^\kappa$. It therefore admits no autonomous Bellman equation built from a local transition kernel alone. Instead, let $(S, A) \sim \nu_\theta$, draw $S^+ \sim P(\cdot \mid S, A)$ and $A^+ \sim \pi_\theta(\cdot \mid S^+)$, and set $Y_i^{\kappa,+} = y_i^\kappa(S^+, A^+)$. For any bounded function $f$, we define the conditional finite-hop Bellman operator as

$$(T_{i,\kappa}^\theta f)(y) := \mathbb{E}[r_i(S, A) + \gamma f(Y_i^{\kappa,+}) \mid Y_i^\kappa = y]. \tag{11}$$

**Proposition 3.2** (Conditional finite-hop Bellman target)**.** *For every agent $i$, policy $\theta \in \Theta_b$, and integer $\kappa \geq 0$, the operator $T_{i,\kappa}^\theta$ is a $\gamma$-contraction on $L_\infty(\nu_{i,\kappa,\theta})$, and its unique fixed point $Q_{i,\kappa}^\theta$ satisfies $\|Q_{i,\kappa}^\theta\|_{L_\infty(\nu_{i,\kappa,\theta})} \leq Q_{\max}$. If Definition 2.1 also holds and $\kappa \geq r_R$, then*

$$\left\| Q_{i,\kappa}^\theta - \widetilde{Q}_{i,\kappa}^\theta \right\|_{L_\infty(\nu_{i,\kappa,\theta})} \leq \frac{\gamma c_{\mathrm{EDP}}}{1-\gamma} \rho^{\kappa+1}. \tag{12}$$

The proof is deferred to Appendix C.1. Combining (8) with (12) shows that the learnable fixed point approximates the value required by the localized actor, $|Q_i^\theta(s, a) - Q_{i,\kappa}^\theta(y_i^\kappa(s, a))| \leq c_{\mathrm{EDP}} \rho^{\kappa+1}/(1-\gamma)$. Each agent can therefore learn its own target by local TD updates, which we specify next.

### 3.3 NEURAL CRITIC AND ACTOR UPDATE

We assume a simulator that can restart the global system from $\mu_0$ (Appendix D.1). Each agent observes only the neighborhood variables required by its policy and critic, but the samples are generated by the true coupled system rather than by an autonomous simulator for each neighborhood.

**Algorithm 1** Localized Neural Actor–Critic

1: **Input:** $T, K, M, D, m, \kappa, H_{\mathrm{a}}$ with $\kappa \geq \max\{r_R, b\}$; actor step $\alpha \leq 1/L_J$; critic step $\beta$; critic projection radius $B$; actor domain radius $R_\Theta$.
2: Initialize the localized policy parameter $\theta_0 \in \Theta_b$.
3: **for** $t = 0, \ldots, T-1$ **do**
4: Run the fresh agent-local neural-TD subroutine (Algorithm 2) for $K$ steps on independent geometric-horizon tuples to obtain $W_{i,t}^{\mathrm{out}}$ for every agent $i$.
5: Initialize the actor restart chain independently of critic training, burn in for $H_{\mathrm{a}}$ steps, and collect $M$ samples.
6: Relay the clipped scalars $q_{j,t,\ell}$ within $\kappa$ hops; each agent $i$ forms $h_{i,t}$ by (15).
7: Each agent updates $\theta_{i,t+1}$ by the projected ascent step (16).
8: **end for**
9: **Output:** $\theta_\tau$ with $\tau$ drawn uniformly from $\{0, \ldots, T-1\}$, the iterate of Theorem 4.5.

**1) Neural Critic:** Let $d_0$ be the encoding dimension of a single agent's local state–action pair, let $n_\kappa := \max_i |\mathcal{N}_i^\kappa|$ be the largest $\kappa$-hop neighborhood size, and set $d_\kappa := n_\kappa d_0$. Agent $i$ maps its local configuration to a normalized injective encoding $X_i^\kappa = \xi_i(Y_i^\kappa) \in \mathbb{R}^{d_\kappa}$, zero-padded when $|\mathcal{N}_i^\kappa| < n_\kappa$, and approximates $Q_{i,\kappa}^{\theta_t}$ with a depth-$D$, width-$m$ ReLU network $\widehat{Q}_{i,\kappa,t}(\cdot; W_i)$. Here $\mathcal{B}_{i,t}(B)$ is the set of weights lying within Frobenius distance $B$ of the round-$t$ initialization in every layer; the parameterization and that initialization are specified in Appendix D.2. The analysis of Section 4 uses this architecture only through named network-geometry constants at the initialization.

At outer iteration $t$, the critic performs $K$ projected semi-gradient TD steps (Sutton, 1988; Tsitsiklis & Van Roy, 1997) on tuples drawn independently from the discounted occupancy $\nu_{\theta_t}$ through geometric-horizon sampling. Independent tuples, rather than the Markov sampling analyzed in the linear case (Bhandari et al., 2018; Srikant & Ying, 2019), are what make the critic error explicit in $K$ and $m$. Writing $X_{i,t,k}^\kappa$ and $X_{i,t,k}^{\kappa,+}$ for the current and true-successor encodings of the $k$-th tuple, agent $i$ updates

$$\delta_{i,t,k} = \widehat{Q}_{i,\kappa,t}\left(X_{i,t,k}^\kappa; W_{i,t,k}\right) - r_i(S_{t,k}, A_{t,k}) - \gamma \widehat{Q}_{i,\kappa,t}\left(X_{i,t,k}^{\kappa,+}; W_{i,t,k}\right),$$
$$W_{i,t,k+1} = \Pi_{\mathcal{B}_{i,t}(B)}\left[W_{i,t,k} - \beta \delta_{i,t,k} \nabla_{W_i} \widehat{Q}_{i,\kappa,t}\left(X_{i,t,k}^\kappa; W_{i,t,k}\right)\right]. \tag{13}$$

The actor-facing critic uses the averaged parameter $W_{i,t}^{\mathrm{out}} := K^{-1} \sum_{k=0}^{K-1} W_{i,t,k}$ (Polyak & Juditsky, 1992). The TD recursion uses the raw network output, and clipping is applied only after critic training; Algorithm 2 in the appendix gives the complete sampling and update order.

**2) Localized Actor:** Conditionally on $\theta_t$, the actor batch is sampled independently of critic training. After $H_{\mathrm{a}}$ burn-in steps, the actor restart chain supplies $M$ samples $\{(S_{t,\ell}^{\mathrm{a}}, A_{t,\ell}^{\mathrm{a}})\}_{\ell=1}^M$. The clipped critic and the scalar it transmits within $\kappa$ hops are

$$\overline{Q}_{j,\kappa,t}(y) := \mathrm{clip}_{[-Q_{\max}, Q_{\max}]}\left(\widehat{Q}_{j,\kappa,t}\left(\xi_j(y); W_{j,t}^{\mathrm{out}}\right)\right), \quad q_{j,t,\ell} := \overline{Q}_{j,\kappa,t}\left(y_j^\kappa(S_{t,\ell}^{\mathrm{a}}, A_{t,\ell}^{\mathrm{a}})\right), \tag{14}$$

from which agent $i$ forms the empirical localized direction

$$h_{i,t} := \frac{1}{MN(1-\gamma)} \sum_{\ell=1}^{M} \psi_i^{\theta_t}\left(S_{t,\ell}^{\mathrm{a}}, A_{i,t,\ell}^{\mathrm{a}}\right) \sum_{j \in \mathcal{N}_i^\kappa} q_{j,t,\ell}. \tag{15}$$

Writing $h_t = (h_{1,t}, \ldots, h_{N,t})$ for the stacked direction, each agent then takes one projected ascent step:

$$\theta_{i,t+1} = \Pi_{\Theta_{b,i}}(\theta_{i,t} + \alpha\, h_{i,t}). \tag{16}$$

The update is unnormalized, so each block moves by its own raw direction $h_{i,t}$ and no network-wide coordination of the step is needed. The Euclidean projection $\Pi_{\Theta_{b,i}}$ onto the parameter ball of (2) acts within the $\theta_i$-block, and stacking the blocks gives the Euclidean projection $\Pi_{\Theta_b}$ onto $\Theta_b$. Every iterate therefore stays in $\Theta_b$, and the product, $b$-localized structure is preserved without further communication.

For a fixed $\kappa$ on a uniformly bounded-degree graph family, $n_\kappa$ is bounded uniformly in $N$, and hence so are the critic input dimension $d_\kappa = n_\kappa d_0$ and the number of finite-hop value messages each agent receives. That count refers to critic-value exchange only. Constructing critic samples uses policy observations within $\kappa + b$ hops (Appendix D.3).

# 4 FINITE-TIME CONVERGENCE ANALYSIS

In this section, we first bound the critic and policy-gradient estimation error, and then establish our main finite-time guarantee.

## 4.1 CRITIC AND POLICY-GRADIENT ACCURACY

We begin with the policy-side conditions that convert gradient accuracy into a policy-gap guarantee.

**Assumption 4.1** (Policy regularity)**.** There exist constants $L_{\psi,i}, L_J, \mu, C_{\mathrm{hc}} > 0$ and $\varepsilon_{\mathrm{bias}} \geq 0$ such that the following hold: (i) each agent's score is bounded, $\|\psi_i^\theta(s,a_i)\|_2 \leq L_{\psi,i}$ for every $\theta \in \Theta_b$; (ii) $J$ is $L_J$-smooth on the compact convex set $\Theta_b$; and (iii) for every $\theta \in \Theta_b$,

$$\sqrt{\mu}\big(J_b^\star - J(\theta)\big) \leq \big\|\mathcal{G}_\alpha(\theta)\big\|_2 + C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}}, \qquad \mathcal{G}_\alpha(\theta) \coloneqq \tfrac{1}{\alpha}\big(\Pi_{\Theta_b}(\theta + \alpha\nabla J(\theta)) - \theta\big), \tag{17}$$

where $\mathcal{G}_\alpha$ is the projected gradient mapping at the actor step size $\alpha$ of (16).

Clauses (i)–(ii) are standard (Fatkhullin et al., 2023; Gaur et al., 2024; Zhang et al., 2025): bounded scores control the localization and sampling errors, and smoothness supplies the one-step increase. Clause (iii) is the gradient-domination condition of that literature, stated through the gradient mapping $\mathcal{G}_\alpha$ so that constrained stationary points on $\partial\Theta_b$ are handled through the normal cone rather than through $\nabla J$ alone. Wherever the unconstrained step $\theta + \alpha\nabla J(\theta)$ stays in $\Theta_b$ one has $\mathcal{G}_\alpha(\theta) = \nabla J(\theta)$, so clause (iii) is exactly that condition there; on the active boundary it is the corresponding proximal form (Karimi et al., 2016). Appendix E.4 derives the unconstrained form from Fisher nondegeneracy and compatible-advantage approximation (Sutton et al., 1999; Kakade, 2001; Agarwal et al., 2021). It also exhibits a network family on which clause (iii) holds in closed form, with no compatible-advantage residual (Proposition E.14).

Let $\mathcal{H}_t$ denote the algorithmic history after $\theta_t$ has been determined but before the fresh round-$t$ critic initialization and samples are drawn, so $\theta_t$ is $\mathcal{H}_t$-measurable. We use the following conditional interface between the critic and actor analyses, in which $\|f\|_{L_2(\nu)}^2 \coloneqq \mathbb{E}_{Z\sim\nu}[f(Z)^2]$.

**Assumption 4.2** (Per-iteration local-critic accuracy)**.** There exists a deterministic $\varepsilon_{\mathrm{TD}} \geq 0$ such that, for every agent $i$ and every outer iteration $t$ with $\theta_t \in \Theta_b$, almost surely

$$\left(\mathbb{E}\left[\left\|\overline{Q}_{i,\kappa,t} - Q_{i,\kappa}^{\theta_t}\right\|_{L_2(\nu_{i,\kappa,\theta_t})}^2 \,\middle|\, \mathcal{H}_t\right]\right)^{1/2} \leq \varepsilon_{\mathrm{TD}}. \tag{18}$$

**Proposition 4.3** (Conditional neural-critic interface)**.** *Suppose that, on an event of probability at least $1 - \delta_{\mathrm{net}}$ over the critic initialization, the initial output and the tangent features of the network are bounded by $Q_0$ and $G_0$, and the network and its gradient differ from their linearizations at the initialization by at most $\varepsilon_f(m)$ and $\varepsilon_\phi(m)$, uniformly over inputs and over the projection ball $\mathcal{B}_{i,t}(B)$; Appendix E.2 states these conditions and the measurability requirements. At every outer iteration, use a fresh critic initialization, $K$ independent geometric-horizon tuples, and the step size $\beta \propto K^{-1/2}$ given there. Then Assumption 4.2 holds with $\varepsilon_{\mathrm{TD}} \leq A_{\mathrm{crit}}K^{-1/4} + \chi_m$.*

The first term is an optimization error that the TD budget drives to zero, with $A_{\mathrm{crit}} = \big(2B\sqrt{D}\,G\,[R_{\max}+(1+\gamma)Q_B]/(1-\sqrt{\gamma})\big)^{1/2}$, where $G = G_0+\varepsilon_\phi(m)$ and $Q_B = Q_0+G_0B\sqrt{D}+\varepsilon_f(m)$ bound the tangent features and the network output on the projection ball. The floor $\chi_m$ is not reduced by $K$: it is built from $\varepsilon_f(m)$, $\varepsilon_\phi(m)$, the error $\varepsilon_{\mathrm{app}}(m)$ of approximating $Q_{i,\kappa}^\theta$ by the tangent class at width $m$, and $\delta_{\mathrm{net}}$, with both constants given in closed form in Appendix E.2. Appendix E.8 separates the two kinds of input to $\chi_m$: the network-geometry constants, which are architectural and for which bounds of this form exist in the multilayer neural actor–critic literature, and $\varepsilon_{\mathrm{app}}(m)$, which is a regularity condition on the target and is assumed here as it is there.

We next propagate critic accuracy into the actor direction. We collect the localization, critic, and actor-sampling errors in a single quantity,

$$\varepsilon_{\mathrm{AC}} \coloneqq \underbrace{C_{\mathrm{loc},\kappa}\rho^{\kappa+1}}_{\text{localization}} + \underbrace{C_{\mathrm{TD},\kappa}\varepsilon_{\mathrm{TD}}}_{\text{critic}} + \underbrace{F_\kappa\left(2\gamma^{H_{\mathrm{a}}/2} + \sqrt{\frac{C_\gamma}{M}}\right)}_{\text{actor burn-in and sampling}}. \tag{19}$$

Here $C_\gamma \coloneqq 1 + 8\gamma/(1-\gamma)$, and the other three constants, displayed in (69), touch the graph only through the largest $\kappa$-hop neighborhood size $n_\kappa$: each is proportional to $L_\psi \coloneqq (\sum_i L_{\psi,i}^2)^{1/2}$, and $C_{\mathrm{TD},\kappa}, F_\kappa$ carry a further factor $n_\kappa/N$. With per-agent scores bounded uniformly in $N$ this gives $C_{\mathrm{loc},\kappa} = \Theta(\sqrt{N})$ against $C_{\mathrm{TD},\kappa}, F_\kappa = \Theta(n_\kappa/\sqrt{N})$.

**Proposition 4.4** (Policy-gradient estimation error)**.** *Suppose $\kappa \geq \max\{r_R, b\}$, that $\theta_t \in \Theta_b$ almost surely, and that Definition 2.1, clauses (i)–(ii) of Assumption 4.1, and Assumption 4.2 hold. Suppose also that, conditionally on $\mathcal{H}_t$, the actor restart chain is independent of all critic-side randomness. Then, with $h_t \coloneqq (h_{1,t}, \ldots, h_{N,t})$,*

$$\left(\mathbb{E}\left[\|h_t - \nabla J(\theta_t)\|_2^2 \,\middle|\, \mathcal{H}_t\right]\right)^{1/2} \leq \varepsilon_{\mathrm{AC}}. \tag{20}$$

The proof is deferred to Appendix E.3.

### 4.2 Finite-time policy guarantee

We now combine Proposition 4.4 with the projected actor update. Let $J_{\mathrm{rng}} \coloneqq 2R_{\max}/(1-\gamma)$, which bounds the range of $J$ on $\Theta_b$.

**Theorem 4.5** (Finite-time suboptimality over the localized policy class)**.** *Fix an integer $\kappa \geq \max\{r_R, b\}$ and a horizon $T$. Suppose Definition 2.1 and Assumptions 4.1–4.2 hold on $\Theta_b$ with constants uniform in $T$, the actor-sampling condition of Proposition 4.4 holds at every iteration, and $\theta_0 \in \Theta_b$. Run Algorithm 1 with the constant actor step $\alpha \leq 1/L_J$; the blockwise projection* (16) *then keeps every iterate in $\Theta_b$ (Lemma E.15). Let $\tau$ be drawn uniformly from $\{0, \ldots, T-1\}$, independently of the run. Then*

$$\mathbb{E}\left[J_b^\star - J(\theta_\tau)\right] \leq \frac{1}{\sqrt{\mu}} \left( \sqrt{\frac{8J_{\mathrm{rng}}}{\alpha T} + 10\,\varepsilon_{\mathrm{AC}}^2} + C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}} \right). \tag{21}$$

First, the $T$-dependent term inside the square root vanishes at $\mathcal{O}(1/T)$, so the gap at the random iterate $\theta_\tau$ decreases at rate $\mathcal{O}(1/\sqrt{T})$ above the explicit error floor $\left(\sqrt{10}\,\varepsilon_{\mathrm{AC}} + C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}}\right)/\sqrt{\mu}$ set by (19). The same one-step analysis bounds the expected squared projected gradient mapping at $\mathcal{O}(1/T)$ without invoking gradient domination (Proposition E.17). Second, the comparator is $J_b^\star$, the best return within the localized class $\Theta_b$; we discuss this point in Appendix A.2.

**Corollary 4.6** (Accuracy above the error floor)**.** *Fix $\kappa$, the critic architecture and projection domain, and valid constants for the neural and policy conditions. Define the fixed part of the floor as*

$$B_{\mathrm{fixed}} \coloneqq \sqrt{10}\left(C_{\mathrm{loc},\kappa}\rho^{\kappa+1} + C_{\mathrm{TD},\kappa}\chi_m\right) + C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}}. \tag{22}$$

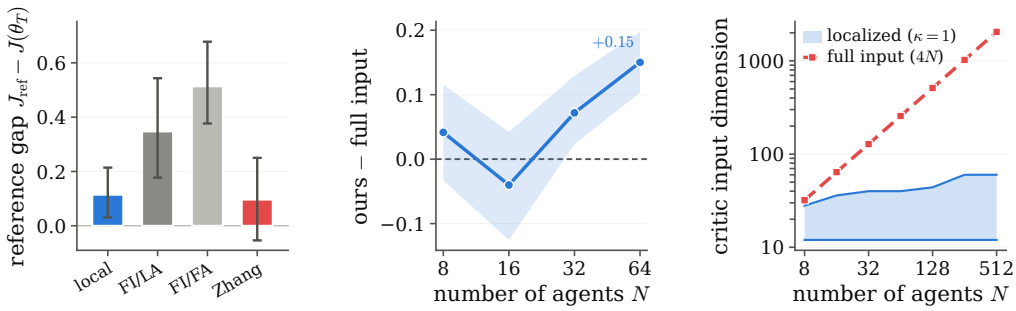
*For any $\epsilon > 0$, choose the optimization and actor-sampling budgets of Appendix E.6. If the resulting run satisfies the hypotheses of Theorem 4.5 on the same prescribed policy domain, then $\mathbb{E}[J_b^\star - J(\theta_\tau)] \leq B_{\mathrm{fixed}}/\sqrt{\mu} + \epsilon$, with budget orders $T = \mathcal{O}(\epsilon^{-2})$, $K = \mathcal{O}(\epsilon^{-4})$, $M = \mathcal{O}(\epsilon^{-2})$, $H_{\mathrm{a}} = \mathcal{O}(\log(1/\epsilon))$, and at most $T(K/(1-\gamma) + M + H_{\mathrm{a}})$ expected transitions. Appendix E.7 gives the radius $\kappa_\star(\epsilon)$ that drives the localization part of $B_{\mathrm{fixed}}$ below $\epsilon$ and the per-agent cost it carries.*

## 5 Numerical Experiments

We evaluate how localizing critic inputs affects solution quality in synthetic ring MDPs and per-agent cost across synthetic and CSMA wireless network environments.

### 5.1 Synthetic rings

**Setup.** The ring MDP has binary local states and actions, one-hop reward coupling ($r_R = 1$), $b = 0$, and $\gamma = 0.9$. At $N = 8$ we evaluate $J(\theta)$ exactly by finite-state enumeration and report the gap to $J_{\mathrm{ref}}$, the highest return from four exact-gradient restarts, shared by all learners on an instance and not assumed globally optimal; a separate paired study uses Monte Carlo returns for $N \in \{8, 16, 32, 64\}$. We compare three learners. *Local* follows Section 3: agent $i$'s critic reads $y_i^\kappa$ and its actor aggregates over $\mathcal{N}_i^\kappa$. *FI/LA* (full input, localized aggregation) uses global state–action




(a) Reference return gap. (b) Local minus FI/LA. (c) Critic input dimension.

Figure 2: **Solution quality and critic input scaling.** (a) Exactly evaluated return gap to the best-found localized reference on $N = 8$ rings, lower is better. (b) Paired final-return difference between Local and FI/LA on sampled rings, positive values favour Local. (c) Maximum critic input dimension across agents at $\kappa = 1$; the band spans the four graph families, the dashed line the full input.

critic inputs, and *FI/FA* (full input, full aggregation) also aggregates over all agents. All three share the learning rule, sampling scheme, and budgets $T$, $K$, $M$ and $H_{\text{a}}$, so Local against FI/LA isolates the critic input scope and FI/LA against FI/FA the actor aggregation scope. All studies share seeds, each fixing an MDP instance and optimization randomness, with bootstrap 95% intervals: marginal for individual learners and paired for matched-seed differences.

**Exact evaluation.** At $N = 8$ the local critic uses 12 of the 32 encoded state–action coordinates, and Local attains a mean reference gap of $0.1142$ $[0.032, 0.214]$ against $0.3472$ $[0.179, 0.544]$ for FI/LA and $0.5134$ $[0.378, 0.680]$ for FI/FA (Figure 2a). Thus, restricted input preserves solution quality at this budget, consistent with the opposing localization and critic-estimation terms in (19), whose contributions these returns cannot separate.

**Across network sizes.** The paired Local-minus-FI/LA difference has a 95% interval containing zero at $N = 8$ and $N = 16$ and lying above zero at $N = 32$ and $N = 64$, at $0.072$ $[0.023, 0.129]$ and $0.150$ $[0.103, 0.197]$ (Figure 2b). Localized input thus gains an advantage as the global configuration grows at a fixed budget, though these returns do not establish dimension as the sole cause.

### 5.2 CSMA Wireless Network Optimization

**Setup.** The carrier-sense multiple access (CSMA) environment lets a transmission succeed only when no neighbour transmits. We place it on rings, two-dimensional tori, and topologies drawn from the EuroRoad and US Power networks in a public repository (Rossi & Ahmed, 2015), at $N$ from 8 to 512 with $\kappa = 1$.

**Input dimension and communication.** At $N = 512$ the largest local critic input is 12 (ring), 20 (torus), 44 (EuroRoad) and 60 (US Power), against $4N = 2048$ for a full input, and an actor batch needs at most $M\Delta_{\mathcal{G}}$ remote critic scalars (Figure 2c). These quantities instantiate the per-agent bounds of Corollary 4.6 and depend only on the graph and the encoding, not on learning performance. The ring and the torus keep this input bounded independently of $N$, whereas the two repository topologies have a maximum degree growing from 8 to 10 and from 9 to 14 between $N = 64$ and $N = 512$, and $d_\kappa$ with it.

## 6 Conclusion

We studied finite-hop neural actor–critic learning in cooperative networked MDPs. A conditional Bellman formulation separates the value learned by local TD from the global value the policy gradient needs, and spatial decay bounds the discrepancy; the result is a finite-sample critic bound and an $\mathcal{O}(1/\sqrt{T})$ rate for the projected gradient mapping and, under gradient domination, for the policy gap above an explicit floor. In the experiments the finite-hop critic matches or improves on the full-input control at matched budgets, at a per-agent dimension set by the neighborhood.

### AI USE STATEMENT

Generative AI tools were used to assist with editing and polishing the manuscript for clarity, concision, grammar, and organization, including improving the presentation of mathematical arguments. They were also used to refine the visual presentation of figures, such as their layout, typography, color scheme, and labeling. All mathematical claims and proofs were independently checked by the authors, and all figure revisions were verified against the underlying data to ensure that they did not alter the scientific content or conclusions. The authors reviewed all AI-assisted revisions and take full responsibility for the final content of this work, including its text, claims, and visual materials.

### REPRODUCIBILITY STATEMENT

We support the reproducibility of both the theoretical and empirical results. Section 2 specifies the networked MDP, localized policy class, and assumptions, while Section 3 provides the critic and actor updates together with complete algorithmic pseudocode. Section 4 states the conditions and finite-time guarantees, and the appendix provides the supporting lemmas and complete proofs, including the finite-width and finite-sample analysis of the neural critic. Section 5 and Appendix F describe the environments, baselines, hyperparameters, random-seed protocol, evaluation metrics, and confidence-interval construction.

## REFERENCES

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

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1691–1692. PMLR, 2018. URL https://proceedings.mlr.press/v75/bhandari18a.html.

Qi Cai, Zhuoran Yang, Jason D. Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11312–11322. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/98baeb82b676b662e12a7af8ad9212f6-Abstract.html.

Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 9(1): 427–438, 2013. doi: 10.1109/TII.2012.2219061.

Tianshu Chu, Jie Wang, Lara Codecà, and Zhaojian Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3): 1086–1095, 2020. doi: 10.1109/TITS.2019.2901791.

Simon S. Du, Xiyu Zhai, Barnabás Póczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eK3i09YQ.

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9827–9869. PMLR, 2023. URL https://proceedings.mlr.press/v202/fatkhullin23a.html.

David Gamarnik, David A. Goldberg, and Theophane Weber. Correlation decay in random decision networks. *Mathematics of Operations Research*, 39(2):229–261, 2014. doi: 10.1287/moor.2013.0609.

Mudit Gaur, Amrit Bedi, Di Wang, and Vaneet Aggarwal. Closing the gap: Achieving global convergence (last iterate) of Actor-Critic under Markovian sampling with neural network parametrization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15153–15179. PMLR, 2024. URL https://proceedings.mlr.press/v235/gaur24a.html.

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.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8571–8580. Curran Associates, Inc., 2018.

Sham M. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1531–1538. MIT Press, 2001.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning*, pp. 267–274. Morgan Kaufmann, 2002.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pp. 795–811, 2016.

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

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, 2nd edition, 2017.

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. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/412604be30f701b1b1e3124c252065e6-Abstract.html.

Xiaolan Liu, Jiadong Yu, Zhiyong Feng, and Yue Gao. Multi-agent reinforcement learning for resource allocation in IoT networks with edge computing. *China Communications*, 17(9):220–236, 2020. doi: 10.23919/JCC.2020.09.017.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6379–6390. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/68a9750337a418a86fe06c1991a1d64c-Abstract.html.

Yasar Sinan Nasir and Dongning Guo. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks. *IEEE Journal on Selected Areas in Communications*, 37(10): 2239–2250, 2019. doi: 10.1109/JSAC.2019.2933973.

Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233, 2007. doi: 10.1109/JPROC.2006.887293.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(79):1–32, 2015. doi: 10.1214/EJP.v20-4039.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046.

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. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/168efc366c449fab9c2843e9b54e2a18-Abstract.html.

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. URL https://proceedings.mlr.press/v120/qu20a.html.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4295–4304. PMLR, 2018. URL https://proceedings.mlr.press/v80/rashid18a.html.

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, 2025. URL https://proceedings.mlr.press/v258/ren25a.html.

Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1):4292–4293, 2015. doi: 10.1609/aaai.v29i1.9277. URL https://ojs.aaai.org/index.php/AAAI/article/view/9277.

R. Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2803–2830. PMLR, 2019. URL https://proceedings.mlr.press/v99/srikant19a.html.

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, Cambridge, MA, 2nd 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. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

John N. Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997. doi: 10.1109/9.580874.

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.

Jianhong Wang, Wangkun Xu, Yunjie Gu, Wenbin Song, and Tim C. Green. Multi-agent reinforcement learning for active voltage control on power distribution networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3271–3284. Curran Associates, Inc., 2021.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24611–24624. Curran Associates, Inc., 2022. doi:

10.52202/068431-1787. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract.html.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5872–5881. PMLR, 2018. URL https://proceedings.mlr.press/v80/zhang18n.html.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In *Handbook of Reinforcement Learning and Control*, pp. 321–384. Springer, 2021. doi: 10.1007/978-3-030-60990-0_12.

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):13:1–13:51, March 2023. doi: 10.1145/3579443. URL https://doi.org/10.1145/3579443.

Zhiyao Zhang, Myeung Suk Oh, Fnu Hairi, Ziyue Luo, Alvaro Velasquez, and Jia Liu. Finite-time global optimality convergence in deep neural actor-critic methods for decentralized Multi-Agent reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 75853–75877. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25bj.html.

Zhaoyi Zhou, Zaiwei Chen, Yiheng Lin, and Adam Wierman. Convergence rates for localized Actor-Critic in networked Markov potential games. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2563–2573. PMLR, 2023. URL https://proceedings.mlr.press/v216/zhou23b.html.

# A Related Work and Limitations

## A.1 Related Work

This work connects three lines of research: localized RL for networked MDPs, actor–critic theory with neural function approximation, and decentralized MARL with graph-based communication.

**1) Localized RL for networked MDPs.** The exponential decay property was introduced for networked MDPs with local interactions, where the influence of a distant state–action component on a local action-value decays geometrically with graph distance. Qu et al. (2020b) used this property to analyze tabular localized actor–critic in the discounted setting, and Qu et al. (2020a) developed the average-reward counterpart. Subsequent work retained the decay mechanism while varying the representation. Lin et al. (2021) handle stochastic interaction networks with time-varying and possibly nonlocal dependencies through state aggregation. Zhang et al. (2023) analyze entropy-regularized localized policy iteration on finite state spaces and obtain a policy-gap bound with a locality term polynomial in $\kappa$. Zhou et al. (2023) give convergence rates for localized actor–critic in networked Markov potential games, with averaged Nash regret as the performance criterion. Ren et al. (2025) construct spectral local representations for continuous networked MDPs under linear critics. Gu et al. (2025) study localized training and decentralized execution for mean-field MARL on a graph-structured state model with exchangeable agents, and analyze a two-layer overparameterized critic. A contemporaneous work by Wang et al. (2026) studies continuous-state–action networked MARL using a localized LSTD critic with spectral random features and a conditional-expectation Bellman target, and establishes an averaged stationarity guarantee. Its critic is linear in fixed random features, whereas our analysis covers a trainable multilayer critic and a policy-gap guarantee. We retain the fixed-agent, local-dependence model of this line but replace its tabular, aggregated, or spectral-linear critics with a trainable multilayer network. This extension is nontrivial because finite-hop observations are generally non-Markovian, whereas standard TD analyses rely on an autonomous Markovian target.

**2) Actor–critic with neural function approximation.** Finite-time analyses of actor–critic methods have progressed from linear critics (Konda & Tsitsiklis, 1999; Bhandari et al., 2018; Srikant &

Ying, 2019) to neural ones. Cai et al. (2019) established finite-sample guarantees for neural TD in the overparameterized regime associated with neural tangent kernel analysis (Jacot et al., 2018; Du et al., 2019). On the actor side, global performance guarantees have been obtained through gradient domination and compatible function approximation (Kakade, 2001; Agarwal et al., 2021), or through Fisher non-degeneracy (Fatkhullin et al., 2023). These analyses provide the foundations for neural actor–critic theory, but they do not address a critic restricted to generally non-Markovian finite-hop observations on a fixed agent network.

**3) Decentralized MARL with graph-based communication.** A parallel literature studies consensus-based decentralized RL, in which agents exchange parameters or value estimates over a communication graph rather than spatially truncating the value function (Zhang et al., 2018; 2021). Zhang et al. (2025) analyze a DNN-based actor–critic in this setting and obtain a finite-time global performance guarantee using a critic that takes the joint state as input. Widely used deep MARL methods, including MADDPG (Lowe et al., 2017), QMIX (Rashid et al., 2018), and MAPPO (Yu et al., 2022), likewise train neural value models with centralized state access, but do not provide finite-time guarantees of the type considered here. In contrast, our critic uses only finite-hop inputs and local scalar-value communication, at the price of a comparator restricted to the localized policy class.

### A.2 Limitations

Like the localized methods it builds on, the algorithm optimizes within a prescribed localized class, and the guarantee is stated relative to that class. Characterizing what the restriction costs, that is, the size of $J^\star - J_b^\star$ under spatial decay, is open for this line of work. Spatial decay constrains the value geometry but not the policy geometry: it implies neither gradient domination, nor Fisher nondegeneracy, nor a small compatible-advantage error. These policy-side quantities therefore enter as separate conditions, imposed only on the compact set $\Theta_b$ that the blockwise projection preserves (Remark E.18). The floor $B_{\mathrm{fixed}}/\sqrt{\mu}$ of (22) is not uniform in the network size. Inside $B_{\mathrm{fixed}}$ the aggregated score bound $L_\psi = (\sum_i L_{\psi,i}^2)^{1/2}$ that enters $C_{\mathrm{loc},\kappa}$ grows like $\sqrt{N}$, the compatible-advantage error $\varepsilon_{\mathrm{bias}}$ and the constant $C_{\mathrm{hc}}$ are instance-dependent, and the width needed for a target $\varepsilon_{\mathrm{app}}(m)$ inside $\chi_m$ may grow as well; outside it, the gradient-domination constant $\mu$ divides, and Appendix E.7 exhibits a family on which $\sqrt{\mu} = \Theta(N^{-1/2})$, so that $1/\sqrt{\mu}$ contributes a further factor $\sqrt{N}$ rather than cancelling one. Appendix E.7 records the resulting radius schedule $\kappa_\star(\epsilon) = \Theta(\log(N/\epsilon)/\log(1/\rho))$ and the sense in which locality nevertheless survives. Corollary 4.6 accordingly bounds per-agent representation and finite-hop messaging uniformly in $N$, and is a statement about per-agent cost rather than total learning complexity.

On the critic side, Theorem E.5 is stated in terms of the linearization errors $\varepsilon_f, \varepsilon_\phi$ and the tangent approximation error $\varepsilon_{\mathrm{app}}(m)$, so that an architecture-specific bound on those quantities can be substituted without revisiting the TD analysis. Bounds of the form (52)–(53) are available in the multilayer neural actor–critic literature at a fixed parameter radius (Appendix E.8), but we do not transcribe a derivation for the parameterization used here, in which the input and output layers are fixed at their random draw; the constants are therefore carried as hypotheses. The tangent-approximation error $\varepsilon_{\mathrm{app}}(m)$ is a condition on the target rather than on the architecture and is assumed throughout, as it is in that literature (Remark E.7). Corollary 4.6 is therefore stated as accuracy above a fixed representation floor.

The analyzed samplers use a simulator that can restart the global system from $\mu_0$, as is standard in this literature. Learning from a single uninterrupted trajectory is a stronger requirement, and extending the critic bound to that setting would carry a mixing-time factor.

Finally, bounded finite-hop input dimension and value-message count bound per-agent cost: acquiring the neighborhood observations is charged separately, and the guarantee controls the expected optimality gap rather than a high-probability one. On the empirical side, the controlled ring has small local state spaces, so its $\kappa$-hop window carries only $64$ configurations and the localization the theory is about discards very little there; the queue sweep of Appendix F.4 raises the window to $262144$ configurations, but both environments remain small and structured, and more demanding local estimation problems are left to future work. Across these instances the smallest admissible radius is also the best one at the budgets tested, so the estimation term of (19) dominates the localization term throughout and the radius schedule $\kappa_\star(\epsilon)$ is not exercised empirically.

# B TECHNICAL RESULTS FOR SPATIAL LOCALIZATION

This appendix provides the technical results supporting Sections 2 and 3. Throughout, $\|f\|_{L_\infty(\nu)}$ denotes the smallest constant $c$ with $|f| \le c$ almost surely under $\nu$, the local state and action spaces are finite, so the conditional laws used below exist, and identities involving conditional expectations hold almost everywhere under the relevant visitation distribution. We keep the pointwise notation of Section 2.1 for $P$, $\pi_\theta$, and $\nu_\theta$, and write $Y_i^\kappa$ for a random local configuration, $y_i^\kappa(s,a)$ for its realization, and $X_i^\kappa = \xi_i(Y_i^\kappa)$ for its neural encoding. Four value objects recur and are kept distinct: $\widetilde{Q}_{i,\kappa}^\theta$ is the conditional truncation of the global value, $Q_{i,\kappa}^\theta$ the conditional finite-hop Bellman fixed point, $\widehat{Q}_{i,\kappa,t}$ the raw round-$t$ neural critic, and $\overline{Q}_{i,\kappa,t}$ its clipped output used by the actor.

## B.1 A FINITE-PROPAGATION SUFFICIENT CONDITION

Definition 2.1 is stated directly because it also covers models whose decay is sharper than that implied by finite-speed propagation. The following result gives a sufficient condition under the dynamics of Section 2.1.

**Proposition B.1** (Finite propagation implies exponential decay)**.** *Suppose the transition kernel has the one-hop factorization* (1)*, the policy is $b$-localized, and $|r_i| \le R_{\max}$ with reward radius $r_R$. For every integer $\ell \ge r_R$, define*

$$\tau(\ell) \coloneqq \left\lfloor \frac{\ell - r_R}{1+b} \right\rfloor.$$

*Then, for any two initial state–action pairs $(s,a)$ and $(\widetilde{s},\widetilde{a})$ that agree on $\mathcal{N}_i^\ell$,*

$$\left|Q_i^\theta(s,a) - Q_i^\theta(\widetilde{s},\widetilde{a})\right| \le \frac{2R_{\max}}{1-\gamma}\gamma^{\tau(\ell)+1}. \tag{23}$$

*Consequently, Definition 2.1 holds with*

$$\rho = \gamma^{1/(1+b)}, \qquad c_{\mathrm{EDP}} = \frac{2R_{\max}}{1-\gamma}\gamma^{-r_R/(1+b)}. \tag{24}$$

*In particular, when $b = r_R = 0$, one may take $\rho = \gamma$ and $c_{\mathrm{EDP}} = 2R_{\max}/(1-\gamma)$.*

*Proof.* Couple the two trajectories by using the same action and transition randomness whenever their local conditioning variables agree. Because the dynamics are one-hop and the policy has radius $b$, a discrepancy can move toward agent $i$ by at most $1+b$ hops per time step. Inductively, the two state–action configurations agree on $\mathcal{N}_i^{\ell-t(1+b)}$ at time $t$. Hence the reward variables of agent $i$ agree for every $t$ such that $t(1+b)+r_R \le \ell$, namely for $t = 0,\dots,\tau(\ell)$. The remaining discounted tail satisfies

$$\left|Q_i^\theta(s,a) - Q_i^\theta(\widetilde{s},\widetilde{a})\right| \le 2R_{\max}\sum_{t=\tau(\ell)+1}^{\infty}\gamma^t = \frac{2R_{\max}}{1-\gamma}\gamma^{\tau(\ell)+1},$$

which proves (23). To obtain (24), write $\ell = r_R + (1+b)q + r$ with $0 \le r < 1+b$, so that $(\ell+1-r_R)/(1+b) = q + (r+1)/(1+b)$ and

$$\gamma^{q+1} \le \gamma^{-r_R/(1+b)}\left(\gamma^{1/(1+b)}\right)^{\ell+1},$$

where the inequality uses $q+1 \ge q + (r+1)/(1+b)$, valid because $r$ is an integer with $r < 1+b$ and hence $r \le b$. □

The value $Q_i^\theta$ in Proposition B.1 is defined under the true dynamics. The restart sampler introduced later is therefore irrelevant to the decay argument.

# C SUPPORTING RESULTS FOR THE LOCALIZED ACTOR–CRITIC

This appendix proves the localization statements used in Section 3.

### C.1 Conditional finite-hop Bellman target

We first relate the global action-value to the finite-hop Bellman fixed point.

**Lemma C.1** (Conditional truncation)**.** *Under Definition 2.1, for every $i$, $\theta \in \Theta_b$, and integer $\kappa \geq r_R$,*

$$\left|Q_i^\theta(s,a) - \widetilde{Q}_{i,\kappa}^\theta\big(y_i^\kappa(s,a)\big)\right| \leq c_{\mathrm{EDP}}\rho^{\kappa+1} \tag{25}$$

*for $\nu_\theta$-almost every $(s,a)$. Moreover, $\|\widetilde{Q}_{i,\kappa}^\theta\|_{L_\infty(\nu_{i,\kappa,\theta})} \leq Q_{\max}$.*

*Proof.* Fix $y$ in the support of $\nu_{i,\kappa,\theta}$ and let $\eta_{i,\theta}(\cdot \mid y)$ be the conditional distribution of $(S,A)$ given $Y_i^\kappa = y$. Every completion in its conditional support agrees with $(s,a)$ on $\mathcal{N}_i^\kappa$. Definition 2.1 therefore gives

$$|Q_i^\theta(s,a) - Q_i^\theta(\bar{s},\bar{a})| \leq c_{\mathrm{EDP}}\rho^{\kappa+1}.$$

Taking expectations under $\eta_{i,\theta}$ proves (25). The norm bound follows from $|Q_i^\theta| \leq Q_{\max}$. □

Let

$$\begin{aligned} \varrho_\theta(s,a) &\coloneqq \mu_0(s)\pi_\theta(a \mid s), \\ \mathcal{P}_\theta\big((s,a),(s',a')\big) &\coloneqq P(s' \mid s,a)\pi_\theta(a' \mid s'). \end{aligned} \tag{26}$$

The discounted occupancy satisfies

$$\nu_\theta = (1-\gamma)\varrho_\theta + \gamma\nu_\theta\mathcal{P}_\theta. \tag{27}$$

Hence $\nu_\theta\mathcal{P}_\theta$ vanishes wherever $\nu_\theta$ does, and the conditional kernel in (11) defines an operator on $L_\infty(\nu_{i,\kappa,\theta})$.

*Proof of Proposition 3.2.* For bounded $f$ and $g$,

$$\begin{aligned} |(T_{i,\kappa}^\theta f)(y) - (T_{i,\kappa}^\theta g)(y)| &\leq \gamma\mathbb{E}[|f(Y_i^{\kappa,+}) - g(Y_i^{\kappa,+})| \mid Y_i^\kappa = y] \\ &\leq \gamma\|f-g\|_{L_\infty(\nu_{i,\kappa,\theta})}. \end{aligned}$$

Thus $T_{i,\kappa}^\theta$ is a contraction and has a unique fixed point $Q_{i,\kappa}^\theta$; bounded rewards imply $\|Q_{i,\kappa}^\theta\|_{L_\infty(\nu_{i,\kappa,\theta})} \leq Q_{\max}$.

The global Bellman identity and the tower property give

$$\begin{aligned} \widetilde{Q}_{i,\kappa}^\theta(y) &= \mathbb{E}[r_i(S,A) + \gamma Q_i^\theta(S^+,A^+) \mid Y_i^\kappa = y], \\ (T_{i,\kappa}^\theta\widetilde{Q}_{i,\kappa}^\theta)(y) &= \mathbb{E}[r_i(S,A) + \gamma\widetilde{Q}_{i,\kappa}^\theta(Y_i^{\kappa,+}) \mid Y_i^\kappa = y]. \end{aligned}$$

Lemma C.1 implies

$$\|T_{i,\kappa}^\theta\widetilde{Q}_{i,\kappa}^\theta - \widetilde{Q}_{i,\kappa}^\theta\|_{L_\infty(\nu_{i,\kappa,\theta})} \leq \gamma c_{\mathrm{EDP}}\rho^{\kappa+1}. \tag{28}$$

Applying the contraction inequality once more proves

$$\|Q_{i,\kappa}^\theta - \widetilde{Q}_{i,\kappa}^\theta\|_{L_\infty(\nu_{i,\kappa,\theta})} \leq \frac{\gamma c_{\mathrm{EDP}}}{1-\gamma}\rho^{\kappa+1}, \tag{29}$$

as claimed. □

Combining Lemma C.1 with (29) yields

$$\big|Q_i^\theta(s,a) - Q_{i,\kappa}^\theta\big(y_i^\kappa(s,a)\big)\big| \leq \frac{c_{\mathrm{EDP}}}{1-\gamma}\rho^{\kappa+1}. \tag{30}$$

The operator is conditional rather than autonomous: variables outside $\mathcal{N}_i^\kappa$ that affect the successor are integrated out by the conditional expectation in (11).

### C.2 Localized policy-gradient approximation

We next prove the bias bound for the localized $Q$-form direction, with $\kappa \geq \max\{r_R, b\}$ as fixed in Section 3.1.

*Proof of Lemma 3.1.* If $j \notin \mathcal{N}_i^\kappa$, then undirectedness implies $i \notin \mathcal{N}_j^\kappa$, so $\widetilde{Q}_{j,\kappa}^\theta(Y_j^\kappa)$ does not depend on $A_i$. Conditioning on $(S, A_{-i})$ and using the product policy gives

$$\mathbb{E}\left[\psi_i^\theta(S, A_i)\widetilde{Q}_{j,\kappa}^\theta(Y_j^\kappa) \mid S, A_{-i}\right] = \widetilde{Q}_{j,\kappa}^\theta(Y_j^\kappa)\,\mathbb{E}\left[\nabla_{\theta_i} \log \pi_{\theta_i}(A_i \mid S_{\mathcal{N}_i^b}) \mid S, A_{-i}\right] = 0.$$

Adding these zero terms to (9) and using (6) yields

$$\nabla_{\theta_i} J(\theta) - g_{i,\kappa}(\theta) = \frac{1}{N(1-\gamma)}\mathbb{E}_{\nu_\theta}\left[\psi_i^\theta(S, A_i)\sum_{j=1}^{N}\big(Q_j^\theta(S, A) - \widetilde{Q}_{j,\kappa}^\theta(Y_j^\kappa)\big)\right]. \tag{31}$$

Lemma C.1 and the score bound imply

$$\|\nabla_{\theta_i} J(\theta) - g_{i,\kappa}(\theta)\|_2 \leq \frac{L_{\psi,i} c_{\mathrm{EDP}}}{1-\gamma}\rho^{\kappa+1}.$$

Squaring and summing over $i$ completes the proof. □

*Remark* C.2 (Use of $Q$-values in the actor update). For an exact local Bellman fixed point, the TD residual has conditional mean zero given $Y_i^\kappa$. Since $\kappa \geq b$, the score $\psi_i^\theta(S, A_i)$ is measurable with respect to $Y_i^\kappa$, so a private score-weighted TD residual also has zero mean and does not estimate (6). The actor therefore uses the localized $Q$-form direction in (15).

## D Sampling and Implementation Details

This appendix provides the sampling, neural-critic, and communication details omitted from the main text.

### D.1 Restart sampling

The restart chain provides discounted-occupancy samples while preserving the true Bellman target. Define

$$\mathcal{K}_\theta((s, a), \cdot) \coloneqq \gamma\mathcal{P}_\theta((s, a), \cdot) + (1-\gamma)\varrho_\theta(\cdot). \tag{32}$$

Throughout, $\|P - Q\|_{\mathrm{TV}} \coloneqq \sup_A |P(A) - Q(A)|$.

**Proposition D.1** (Stationarity and mixing)**.** *The distribution $\nu_\theta$ is stationary for $\mathcal{K}_\theta$, and*

$$\sup_z \|\mathcal{K}_\theta^h(z, \cdot) - \nu_\theta\|_{\mathrm{TV}} \leq \gamma^h, \qquad h \in \mathbb{Z}_{\geq 0}. \tag{33}$$

*Proof.* The occupancy recursion (27) gives $\nu_\theta\mathcal{K}_\theta = \nu_\theta$. Moreover, $\mathcal{K}_\theta(z, \cdot) \geq (1-\gamma)\varrho_\theta(\cdot)$. Coupling two chains at their first common restart (a Doeblin-type argument; cf. Levin & Peres, 2017) proves (33). □

The uniform geometric rate in (33) allows the actor-batch error to be controlled by a burn-in term plus an i.i.d.-like sampling term; sharper deviation inequalities for functions of a mixing chain are available if a variance-adapted rate is preferred (Paulin, 2015).

Whenever a transition of this restart chain is recorded, the true successor $S^+ \sim P(\cdot \mid S, A)$ and $A^+ \sim \pi_\theta(\cdot \mid S^+)$ are recorded first and the restart coin is drawn only afterwards, since using the post-restart pair would evaluate the restart kernel $\mathcal{K}_\theta$ rather than $\mathcal{P}_\theta$. The critic sampler of Appendix E.2 draws no restart coin at all: it obtains $(S, A) \sim \nu_\theta$ exactly by stopping a $\varrho_\theta$-started trajectory at a geometric time and then draws the true successor, as in Algorithm 2.

### D.2 NEURAL CRITIC

This subsection specifies the neural parameterization and projected parameter domain used by the local critics. We first state the encoding assumption that the critic input map is required to satisfy.

**Assumption D.2** (Measurable information-preserving local encoding)**.** For each agent $i$ the encoding map $\xi_i \colon \mathcal{Y}_i^\kappa \to \mathbb{R}^{d_\kappa}$ of the finite-hop configuration space is measurable and satisfies four conditions. *(i) Information preservation:* $\xi_i$ is injective on the support of $\nu_{i,\kappa,\theta}$, so that $\sigma(\xi_i(Y_i^\kappa)) = \sigma(Y_i^\kappa)$ and no local information used by the conditional Bellman target is discarded. *(ii) Canonical neighbor ordering:* the $\kappa$-hop neighbors of $i$ are written into a fixed slot order determined by graph distance, with ties broken by a fixed global node index, so that each occupied slot carries a definite node identity rather than an exchangeable one. *(iii) Zero-padding:* when $|\mathcal{N}_i^\kappa| < n_\kappa$ the unused slots are set to zero, which must leave a padded slot distinguishable from an occupied one; for the one-hot encodings of finite local spaces used throughout, the all-zero block is outside the encoding range and this holds automatically, so the input dimension is $d_\kappa = n_\kappa d_0$ as stated. If an encoding does not have this property, a per-slot occupancy indicator is appended and $d_\kappa = n_\kappa(d_0 + 1)$, which changes none of the bounds below. Heterogeneous per-node dimensions are accommodated by fixing $d_0$ as the maximal per-node block and zero-padding smaller blocks within their slot. *(iv) Normalization:* $\|\xi_i(y)\|_2 = 1$ for every $y$, which fixes the input scale entering the network-geometry constants below. This is a property of the encoding map and not of the raw blocks: concatenating one-hot blocks for the $|\mathcal{N}_i^\kappa|$ occupied slots gives a vector of constant norm $\sqrt{2|\mathcal{N}_i^\kappa|}$, independent of $y$, so $\xi_i$ divides by that constant. The rescaling is by a fixed per-agent factor, so it changes neither the injectivity in (i) nor the slot structure in (ii)–(iii), and $d_\kappa$ is unchanged.

For a local configuration $y$, we write $x = \xi_i(y) \in \mathbb{R}^{d_\kappa}$ for its encoding. As defined in Section 3.3, $d_\kappa = n_\kappa d_0$, which is bounded uniformly in $N$ for fixed $\kappa$ on a bounded-degree graph family. Injectivity in Assumption D.2(i) is what makes the conditional Bellman target of Proposition 3.2 a well-defined function of the encoded input, and the slot identity in (ii)–(iii) is what makes the per-agent critic a function of a fixed-width input rather than of an unordered set. For matrices of the same dimension, we use the Frobenius inner product and its induced norm:

$$\langle A, B\rangle_{\mathrm{F}} \coloneqq \operatorname{tr}(A^\top B), \qquad \|A\|_{\mathrm{F}} \coloneqq \sqrt{\langle A, A\rangle_{\mathrm{F}}}. \tag{34}$$

The subscript F distinguishes the Frobenius inner product from the Fisher matrix $\mathcal{F}(\theta)$ introduced in Appendix E.4 and from the actor-error constant $F_\kappa$ of (69).

For integers $D, m \geq 1$, agent $i$ uses a depth-$D$, width-$m$ fully connected ReLU network. For $x \in \mathbb{R}^{d_\kappa}$, define

$$\begin{aligned} z_i^{(0)} &= H_i x, \\ z_i^{(h)} &= \frac{1}{\sqrt{m}} \operatorname{ReLU}\Big(W_i^{(h)} z_i^{(h-1)}\Big), \qquad h = 1, \dots, D, \\ \widehat{Q}_{i,\kappa}(x; W_i) &= \frac{1}{\sqrt{m}}\, v_i^\top z_i^{(D)}, \end{aligned} \tag{35}$$

where

$$H_i \in \mathbb{R}^{m \times d_\kappa}, \qquad W_i^{(h)} \in \mathbb{R}^{m \times m}, \qquad v_i \in \mathbb{R}^m,$$

and

$$W_i \coloneqq \big(W_i^{(1)}, \dots, W_i^{(D)}\big)$$

collects the trainable hidden-layer weights. The policy being evaluated does not appear as an explicit input to the network; its effect is captured by the data distribution, the Bellman target, and the trained parameter $W_{i,t}^{\mathrm{out}}$.

For a generic fresh initialization draw, write $\omega_i \coloneqq (H_i, v_i, W_{i,\mathrm{init}})$ and suppress the outer-round index in the network and projection definitions below. Following the neural parameterization of Zhang et al. (2025), the entries of $H_i$ and $W_{i,\mathrm{init}}^{(h)}$ are initialized independently from $\mathcal{N}(0, 2)$, while the entries of $v_i$ are initialized from $\mathcal{N}(0, 1)$. The input and output weights $H_i$ and $v_i$ are fixed, and only the hidden-layer weights $W_i$ are trained.

The $1/\sqrt{m}$ output factor in (35) keeps the initialization scale and the tangent-feature norms bounded uniformly in the width, and we record its effect here. With $\|\xi_i(y)\|_2 = 1$ and $\mathcal{N}(0, 2)$ entries, the forward recursion preserves $\mathbb{E}\|z_i^{(h)}\|_2^2 = \Theta(m)$ across layers: each layer multiplies

the squared norm by $\approx 2m$ through $W_i^{(h)}$, halves it through the ReLU, and rescales by $1/m$. Consequently $\mathrm{Var}(v_i^\top z_i^{(D)} \mid z_i^{(D)}) = \|z_i^{(D)}\|_2^2 = \Theta(m)$, and the last-layer gradient satisfies $\|\nabla_{W^{(D)}}(v_i^\top z_i^{(D)})\|_{\mathrm{F}} = \Theta(\sqrt{m})$. Without the $1/\sqrt{m}$ factor both the initial output and its gradient would grow like $\sqrt{m}$, so increasing the width would inflate rather than shrink the linearization error. With the factor, both are $\Theta(1)$, which is what Assumption E.3 below requires. Rescaling the output is equivalent to drawing $v_i \sim \mathcal{N}(0, I_m/m)$ and keeping $\widehat{Q}_{i,\kappa} = v_i^\top z_i^{(D)}$; we use the explicit factor throughout and do not also rescale the initialization law. When dependence on the draw matters, we write $\widehat{Q}_{i,\kappa}^{\omega_i}(x; W)$ for the network with $(H_i, v_i)$ specified by $\omega_i$, and $\mathcal{B}_i(B; \omega_i)$ for the set (36) centered at the corresponding $W_{i,\mathrm{init}}$; otherwise this dependence is suppressed.

The trainable weights are restricted to the layerwise neighborhood

$$\mathcal{B}_i(B) \coloneqq \Big\{ W = \big(W^{(1)}, \dots, W^{(D)}\big) : \|W^{(h)} - W_{i,\mathrm{init}}^{(h)}\|_{\mathrm{F}} \le B, \quad h = 1, \dots, D \Big\}. \tag{36}$$

Projection onto $\mathcal{B}_i(B)$ is performed independently across layers:

$$\big[\Pi_{\mathcal{B}_i(B)}(W)\big]^{(h)} = W_{i,\mathrm{init}}^{(h)} + \min\left\{1, \frac{B}{\|W^{(h)} - W_{i,\mathrm{init}}^{(h)}\|_{\mathrm{F}}}\right\} \cdot \Big(W^{(h)} - W_{i,\mathrm{init}}^{(h)}\Big), \qquad h = 1, \dots, D, \tag{37}$$

where the multiplier is defined as one when $W^{(h)} = W_{i,\mathrm{init}}^{(h)}$.

At outer iteration $t$, after $\theta_t$ is fixed and before critic training, agent $i$ draws a fresh copy

$$\omega_{i,t} \coloneqq \big(H_{i,t}, v_{i,t}, W_{i,t,\mathrm{init}}\big).$$

These draws are independent across agents and rounds and are independent of $\mathcal{H}_t$; conditional on $\mathcal{H}_t$, they are also independent of all round-$t$ critic-chain randomness. Let $\widehat{Q}_{i,\kappa,t}$ and $\mathcal{B}_{i,t}(B)$ denote the network and projection set instantiated by $\omega_{i,t}$. Agent $i$ initializes

$$W_{i,t,0} = \overline{W}_{i,t,0} = W_{i,t,\mathrm{init}}. \tag{38}$$

All components of $\omega_{i,t}$ remain fixed within round $t$.

During critic training, the averaged parameter is updated as (Polyak & Juditsky, 1992)

$$\overline{W}_{i,t,k+1} = \frac{k}{k+1}\overline{W}_{i,t,k} + \frac{1}{k+1}W_{i,t,k}, \qquad k = 0, \dots, K-1. \tag{39}$$

Consequently, the actor-facing parameter of Section 3.3 is realized online as

$$W_{i,t}^{\mathrm{out}} = \overline{W}_{i,t,K} = \frac{1}{K}\sum_{k=0}^{K-1} W_{i,t,k}. \tag{40}$$

Because $\mathcal{B}_{i,t}(B)$ is convex, the averaged iterate remains in $\mathcal{B}_{i,t}(B)$. The average runs over the $K$ iterates that are actually evaluated against a TD tuple, which is the index range controlled by the telescoping argument in Theorem E.5; the terminal iterate $W_{i,t,K}$ is not averaged in.

Algorithm 2 collects the complete critic-side sampling and update order. All local updates are derived from the same underlying global transition, but each agent accesses only its own local configuration and forms its own $\kappa$-hop encoding; in particular, the generally non-Markov local process is never treated as an autonomous Markov chain.

Each critic tuple in Algorithm 2 is drawn independently through a geometric horizon rather than by advancing a single restart chain. Since $\nu_\theta = (1-\gamma)\sum_{h\ge 0}\gamma^h \varrho_\theta \mathcal{P}_\theta^h$ by (27), stopping a $\varrho_\theta$-started trajectory after $G \sim \mathrm{Geom}_0(1-\gamma)$ steps yields a sample distributed exactly as $\nu_\theta$, with no burn-in bias and no dependence across tuples. This licenses the conditional unbiasedness used in Theorem E.5: the $k$-th tuple is independent of the current iterate $W_{i,t,k}$, whereas along a single chain the iterate depends on past states that are correlated with the current one. The price is measured in environment transitions rather than in bias, since

$$\mathbb{E}\left[\sum_{k=0}^{K-1}(G_{t,k}+1)\right] = \frac{K}{1-\gamma}. \tag{41}$$

**Algorithm 2** Fresh Agent-Local Neural–TD Critic at Outer Iteration $t$

**Input:** fixed policy $\theta_t$; $\kappa, K, D, m, \beta, B$.
Conditional on $\mathcal{H}_t$, draw $\omega_{i,t} = (H_{i,t}, v_{i,t}, W_{i,t,\mathrm{init}})$ from their generic fresh laws, mutually independently and independently of all critic-chain randomness.
Instantiate $\widehat{Q}_{i,\kappa,t}$ and $\mathcal{B}_{i,t}(B)$ from $\omega_{i,t}$; keep $H_{i,t}$ and $v_{i,t}$ fixed within round $t$.
For every $i$, set $W_{i,t,0} = \overline{W}_{i,t,0} = W_{i,t,\mathrm{init}}$ as in (38).
**for** $k = 0, \dots, K-1$ **do**
  Draw $G_{t,k} \sim \mathrm{Geom}_0(1-\gamma)$, i.e. $\Pr[G = h] = (1-\gamma)\gamma^h$ for $h \geq 0$, and $Z_{t,k,0} \sim \varrho_{\theta_t}$, independently of $\{\omega_{i,t}\}_{i=1}^N$ and of all other tuples.
  Evolve $Z_{t,k,h+1} \sim \mathcal{P}_{\theta_t}(Z_{t,k,h}, \cdot)$ for $h = 0, \dots, G_{t,k} - 1$ and set $(S_{t,k}, A_{t,k}) \leftarrow Z_{t,k,G_{t,k}}$, so that $(S_{t,k}, A_{t,k}) \sim \nu_{\theta_t}$ exactly.
  Draw the true successor $S^+_{t,k} \sim P(\cdot \mid S_{t,k}, A_{t,k})$ and then $A^+_{t,k} \sim \pi_{\theta_t}(\cdot \mid S^+_{t,k})$.
  **for** $i = 1, \dots, N$ **in parallel do**
    Form $X^{\kappa}_{i,t,k}$ and $X^{\kappa,+}_{i,t,k}$ from the current and true-successor pairs as in Section 3.3.
    Compute the *unclipped* TD residual

$$\delta_{i,t,k} \leftarrow \widehat{Q}_{i,\kappa,t}(X^{\kappa}_{i,t,k}; W_{i,t,k}) - r_i(S_{t,k}, A_{t,k}) - \gamma \widehat{Q}_{i,\kappa,t}(X^{\kappa,+}_{i,t,k}; W_{i,t,k}).$$

    Apply the projected semi-gradient update in (13) to obtain $W_{i,t,k+1}$.
    Update $\overline{W}_{i,t,k+1}$ by (39).
  **end for**
**end for**
Set $W^{\mathrm{out}}_{i,t} \leftarrow \overline{W}_{i,t,K}$ for every $i$ as in (40), and construct the actor-facing clipped critic $\overline{Q}_{i,\kappa,t}$ by (14).
**Output:** $\{W^{\mathrm{out}}_{i,t}, \overline{Q}_{i,\kappa,t}\}_{i=1}^N$.

The actor draws its batch from the restart chain of Appendix D.1 and the critic draws independent geometric-horizon tuples; both are generated by the true coupled system, so the localized observation model and the conditional Bellman operator apply to either sampler.

The residual, projection, and averaging steps in Algorithm 2 use the raw network output. Clipping is applied only after critic training for actor-facing evaluation; it does not modify the weights or any TD target. No critic parameter, gradient, or iterate is communicated between agents. The only critic-derived messages are the clipped scalars $q_{i,t,\ell}$ used later by the actor.

The tangent critic used in Appendix E.2 is the first-order expansion of (35) with respect to the trainable matrices, in the spirit of neural-tangent analyses (Jacot et al., 2018; Du et al., 2019; Cai et al., 2019). In particular,

$$\left\langle \nabla_{W_i^{(h)}} \widehat{Q}_{i,\kappa}(x; W_{i,\mathrm{init}}), W_i^{(h)} - W_{i,\mathrm{init}}^{(h)} \right\rangle_{\mathrm{F}}$$

is the directional derivative contributed by layer $h$.

### D.3 Locality of the finite-hop exchange

The projected actor update (16) is fully decentralized: each block $\theta_i$ is updated from its own local direction $h_{i,t}$ and projected onto $\Theta_{b,i}$, with no network-wide scalar or vector coordination. The only communication is the finite-hop exchange of value messages, whose cost we now bound. Let $\Delta_{\mathcal{G}} := \max_j \deg(j)$ denote the maximum graph degree, a quantity distinct from the per-node encoding dimension $d_0$. Then

$$n_\kappa \leq 1 + \Delta_{\mathcal{G}} \sum_{q=0}^{\kappa-1} (\Delta_{\mathcal{G}} - 1)^q. \tag{42}$$

Recall the critic input dimension $d_\kappa = n_\kappa d_0$ of Section 3.3. Thus, for fixed $\kappa$ over a graph family with uniformly bounded $\Delta_{\mathcal{G}}$, the critic input dimension and the number of scalar critic values consumed per actor sample are uniformly bounded in $N$. Generating the action block under a $b$-localized policy uses state observations in $\mathcal{N}_i^{\kappa+b}$, while value estimates are exchanged within $\kappa$ hops. Total computation scales at least linearly in $N$, with no network-wide coordination.

### D.4 NEURAL-TD CRITIC ERROR

For the averaged critic at outer iteration $t$, define

$$\varepsilon_{\mathrm{TD},i,t} := \left( \mathbb{E}\left[ \left\| \overline{Q}_{i,\kappa,t} - Q_{i,\kappa}^{\theta_t} \right\|_{L_2(\nu_{i,\kappa,\theta_t})}^2 \,\middle|\, \mathcal{H}_t \right] \right)^{1/2}. \tag{43}$$

Because $\nu_{i,\kappa,\theta_t}$ is a probability measure, $\|f\|_{L_2(\nu_{i,\kappa,\theta_t})} \le \|f\|_{L_\infty(\nu_{i,\kappa,\theta_t})}$. Combining this embedding with Proposition 3.2 gives

$$\left( \mathbb{E}\left[ \left\| \overline{Q}_{i,\kappa,t} - \widetilde{Q}_{i,\kappa}^{\theta_t} \right\|_{L_2(\nu_{i,\kappa,\theta_t})}^2 \,\middle|\, \mathcal{H}_t \right] \right)^{1/2} \le \varepsilon_{\mathrm{TD},i,t} + \frac{\gamma c_{\mathrm{EDP}}}{1-\gamma} \rho^{\kappa+1}. \tag{44}$$

Appendix E.2 bounds $\varepsilon_{\mathrm{TD},i,t}$ from a fixed-policy expected risk under independent tuples. Fresh per-round initialization permits conditioning on $\mathcal{H}_t$ before applying that bound. A neural-tangent condition provides one sufficient instantiation for the conditional Bellman target.

## E PROOFS FOR THE FINITE-TIME ANALYSIS

This appendix proves the results stated in Section 4.

### E.1 $L_2$ GEOMETRY OF THE CONDITIONAL BELLMAN OPERATOR

We first establish the contraction property used by the neural-TD analysis. Let $\nu^+_{i,\kappa,\theta}$ denote the distribution of $Y_i^{\kappa,+}$ when $(S,A) \sim \nu_\theta$ and $(S^+, A^+)$ is generated by the true kernel $\mathcal{P}_\theta$ in (26).

**Lemma E.1** ($L_2$ contraction)**.** *For every $i$ and $\theta \in \Theta_b$,*

$$\gamma \nu^+_{i,\kappa,\theta} \le \nu_{i,\kappa,\theta}, \qquad \|T^\theta_{i,\kappa} f - T^\theta_{i,\kappa} g\|_{L_2(\nu_{i,\kappa,\theta})} \le \sqrt{\gamma}\, \|f-g\|_{L_2(\nu_{i,\kappa,\theta})}. \tag{45}$$

*Proof.* The occupancy recursion (27) implies $\gamma \nu_\theta \mathcal{P}_\theta \le \nu_\theta$. Marginalizing to $Y_i^\kappa$ gives the first inequality. Conditional Jensen's inequality then yields

$$\begin{aligned} \|T^\theta_{i,\kappa} f - T^\theta_{i,\kappa} g\|^2_{L_2(\nu_{i,\kappa,\theta})} &\le \gamma^2 \mathbb{E}_{Y \sim \nu_{i,\kappa,\theta}} \mathbb{E}\big[ |f(Y_i^{\kappa,+}) - g(Y_i^{\kappa,+})|^2 \mid Y_i^\kappa = Y \big] \\ &= \gamma^2 \|f-g\|^2_{L_2(\nu^+_{i,\kappa,\theta})} \\ &\le \gamma \|f-g\|^2_{L_2(\nu_{i,\kappa,\theta})}. \end{aligned}$$

Taking square roots completes the proof. □

The contraction modulus is $\gamma$ in $L_\infty$ but $\sqrt{\gamma}$ in $L_2(\nu_{i,\kappa,\theta})$; the latter determines the approximation factor in the projected neural-TD bound.

### E.2 FINITE-WIDTH, FINITE-SAMPLE CRITIC BOUND AND CONDITIONAL TRANSFER

We first state the fixed-policy critic bound used by the transfer. Fix an agent $i$ and a deterministic policy parameter $\theta \in \Theta_b$. Draw a fresh initialization $\omega_i$ and, independently of it, a collection $\mathsf{D}_\theta^{\mathrm{iid}}$ of $K$ independent geometric-horizon tuples generated exactly as in Algorithm 2, so that each current state–action pair has law $\nu_\theta$ and each successor is drawn from the true transition kernel followed by $\pi_\theta$. Let $W_i^{\mathrm{out,iid}}(\theta; \omega_i, \mathsf{D}_\theta^{\mathrm{iid}})$ be the output of the projected critic recursion of Algorithm 2 on this collection, and define

$$\overline{Q}^{\mathrm{out,iid}}_{i,\kappa}(\cdot; \theta, \omega_i, \mathsf{D}_\theta^{\mathrm{iid}}) := \mathrm{clip}_{[-Q_{\max}, Q_{\max}]}\Big( \widehat{Q}^{\omega_i}_{i,\kappa}\Big( \xi_i(\cdot); W_i^{\mathrm{out,iid}}(\theta; \omega_i, \mathsf{D}_\theta^{\mathrm{iid}}) \Big) \Big). \tag{46}$$

**Assumption E.2** (Fixed-policy expected critic accuracy under independent tuples)**.** There is a deterministic $\varepsilon_{\mathrm{iid}} \ge 0$ such that, for every agent $i$ and every deterministic $\theta \in \Theta_b$,

$$\left( \mathbb{E}_{\omega_i, \mathsf{D}_\theta^{\mathrm{iid}}} \left[ \left\| \overline{Q}^{\mathrm{out,iid}}_{i,\kappa}(\cdot; \theta, \omega_i, \mathsf{D}_\theta^{\mathrm{iid}}) - Q^\theta_{i,\kappa} \right\|^2_{L_2(\nu_{i,\kappa,\theta})} \right] \right)^{1/2} \le \varepsilon_{\mathrm{iid}}. \tag{47}$$

The risk map in $\theta$ is measurable. For each fixed policy, the expectation averages over a fresh initialization and an independent collection of critic tuples.

We next give a neural-tangent sufficient condition for Assumption E.2. For a generic fresh initialization $\omega_i$, define the tangent model

$$\widehat{Q}^{\tan,\omega_i}_{i,\kappa}(x;W) \coloneqq \widehat{Q}^{\omega_i}_{i,\kappa}(x;W_{i,\mathrm{init}}) + \sum_{h=1}^{D} \left\langle \nabla_{U^{(h)}} \widehat{Q}^{\omega_i}_{i,\kappa}(x;U)\Big|_{U=W_{i,\mathrm{init}}}, W^{(h)} - W^{(h)}_{i,\mathrm{init}} \right\rangle_{\mathrm{F}}. \quad (48)$$

Here $U = (U^{(1)}, \dots, U^{(D)})$ is a dummy parameter tuple used only to specify the derivative, whereas $W = (W^{(1)}, \dots, W^{(D)})$ is the candidate parameter of the tangent model. For each fixed $\omega_i$ the tangent features are bounded on the unit sphere, so $W \mapsto q_W$ is Lipschitz into $L_2(\nu_{i,\kappa,\theta})$ and the image of the compact convex set $\mathcal{B}_i(B;\omega_i)$ is again compact and convex. Define the nonempty compact convex tangent class

$$\mathcal{Q}^{\tan}_{i,\kappa}(\omega_i) \coloneqq \left\{ y \mapsto \widehat{Q}^{\tan,\omega_i}_{i,\kappa}(\xi_i(y);W) : W \in \mathcal{B}_i(B;\omega_i) \right\}.$$

Let $\Pi^{\tan,\omega_i}_{i,\kappa,\theta}$ be its $L_2(\nu_{i,\kappa,\theta})$ projection. Lemma E.1 implies that $\Pi^{\tan,\omega_i}_{i,\kappa,\theta} T^{\theta}_{i,\kappa}$ is a $\sqrt{\gamma}$-contraction. Denote its fixed point by $Q^{\star,\tan,\theta,\omega_i}_{i,\kappa}$ and define

$$e_{\mathrm{app},i}(\theta;\omega_i) \coloneqq \left\| \Pi^{\tan,\omega_i}_{i,\kappa,\theta} Q^{\theta}_{i,\kappa} - Q^{\theta}_{i,\kappa} \right\|_{L_2(\nu_{i,\kappa,\theta})}. \quad (49)$$

Write $R_0 \coloneqq B\sqrt{D}$ and $R_{\mathcal{B}} \coloneqq 2B\sqrt{D}$, so that $\|W - W_{i,\mathrm{init}}\|_{\mathrm{F}} \le R_0$ and $\|W - W'\|_{\mathrm{F}} \le R_{\mathcal{B}}$ for all $W, W' \in \mathcal{B}_i(B;\omega_i)$.

**Assumption E.3** (Finite-width local neural geometry)**.** There are deterministic constants $Q_0, G_0 < \infty$, deterministic functions $\varepsilon_f(m), \varepsilon_\phi(m) \ge 0$, and an event $\mathcal{E}_{i,\omega}$ with $\mathbb{P}_{\omega_i}(\mathcal{E}_{i,\omega}) \ge 1 - \delta_{\mathrm{net}}$ such that, on $\mathcal{E}_{i,\omega}$, uniformly over $x$ and over $W \in \mathcal{B}_i(B;\omega_i)$,

$$\left|\widehat{Q}^{\omega_i}_{i,\kappa}(x;W_{i,\mathrm{init}})\right| \le Q_0, \quad (50)$$

$$\left\|\nabla_W \widehat{Q}^{\omega_i}_{i,\kappa}(x;W_{i,\mathrm{init}})\right\|_{\mathrm{F}} \le G_0, \quad (51)$$

$$\left|\widehat{Q}^{\omega_i}_{i,\kappa}(x;W) - \widehat{Q}^{\tan,\omega_i}_{i,\kappa}(x;W)\right| \le \varepsilon_f(m), \quad (52)$$

$$\left\|\nabla_W \widehat{Q}^{\omega_i}_{i,\kappa}(x;W) - \nabla_W \widehat{Q}^{\omega_i}_{i,\kappa}(x;W_{i,\mathrm{init}})\right\|_{\mathrm{F}} \le \varepsilon_\phi(m). \quad (53)$$

Moreover the tangent approximation error obeys

$$\left(\mathbb{E}_{\omega_i}\left[\mathbf{1}_{\mathcal{E}_{i,\omega}}\, e_{\mathrm{app},i}(\theta;\omega_i)^2\right]\right)^{1/2} \le \varepsilon_{\mathrm{app}}(m) \quad (54)$$

for every agent $i$ and deterministic $\theta \in \Theta_b$. The approximation condition is required only on the event where the network-geometry bounds hold; on its complement, actor-facing clipping bounds the critic risk directly by $2Q_{\max}$.

Assumption E.3 constrains only the local geometry and representation capacity of the network at initialization. Unlike (47), it does not presuppose that the TD iterates have approached any Bellman fixed point; that is what Theorem E.5 below establishes.

*Remark* E.4 (Sources of the linearization bounds). We state (52) and (53) directly rather than deriving them from a bounded Hessian, because the architecture (35) uses ReLU activations and is therefore not twice differentiable in $W$; a bound on $\sup_W \|\nabla^2_W \widehat{Q}_{i,\kappa}\|_{\mathrm{op}}$ is unavailable. For a smooth activation with $\sup_{W\in\mathcal{B}_i(B)} \|\nabla^2_W \widehat{Q}_{i,\kappa}(x;W)\|_{\mathrm{op}} \le L_D/\sqrt{m}$, Taylor's theorem and the mean-value theorem give (52) and (53) with $\varepsilon_f = L_D R_0^2/(2\sqrt{m})$ and $\varepsilon_\phi = L_D R_0/\sqrt{m}$, so both are $\mathcal{O}(m^{-1/2})$. For the ReLU architecture (35) the two bounds instead require an activation-stability argument that accounts jointly for depth, width and projection radius, together with the initialization probability on which it holds. Such arguments exist for multilayer ReLU critics at a fixed parameter radius, and Appendix E.8 records the form they take and the rate they give; we do not transcribe one for the parameterization used here, and all statements below are written in terms of $\varepsilon_f(m)$ and $\varepsilon_\phi(m)$, so that any architecture-specific rate is substituted with its own constants and failure probability.

Define

$$\begin{aligned} G &\coloneqq G_0 + \varepsilon_\phi(m), \\ Q_B &\coloneqq Q_0 + G_0 R_0 + \varepsilon_f(m), \\ G_{\mathrm{TD}} &\coloneqq G\big[R_{\max} + (1+\gamma)Q_B\big], \\ \varepsilon_g &\coloneqq (1+\gamma)G\,\varepsilon_f(m) + \big[R_{\max} + (1+\gamma)Q_B\big]\varepsilon_\phi(m), \end{aligned} \quad (55)$$

where $R_{\max}$ bounds $|r_i|$. On $\mathcal{E}_{i,\omega}$ the triangle inequality with (50)–

**Theorem E.5** (Finite-width and finite-sample local neural TD)**.** *Suppose Assumption E.3 holds and that the displayed random functions admit jointly measurable versions. Run the projected neural-TD recursion of Algorithm 2 for $K$ steps on independent geometric-horizon tuples with step size*

$$\beta = \frac{R_{\mathcal{B}}}{G_{\mathrm{TD}}\sqrt{K}}, \tag{56}$$

*and output the average $W_i^{\mathrm{out}}$ of* (40)*. Then the risk bound* (47) *of Assumption E.2 holds with*

$$\varepsilon_{\mathrm{iid}} = \left[\frac{R_{\mathcal{B}} G_{\mathrm{TD}}}{(1-\sqrt{\gamma})\sqrt{K}} + \frac{R_{\mathcal{B}}\,\varepsilon_g}{1-\sqrt{\gamma}}\right]^{1/2} + \varepsilon_f(m) + \frac{\varepsilon_{\mathrm{app}}(m)}{1-\sqrt{\gamma}} + 2Q_{\max}\sqrt{\delta_{\mathrm{net}}}, \tag{57}$$

*uniformly over agents $i$ and deterministic policies $\theta \in \Theta_b$. The measurability clause of Assumption E.2 is inherited from the assumed measurable versions and is what Proposition E.8 uses.*

Applying $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ to the bracket in (57) separates the $K$-dependent part from the rest, which gives the form $\varepsilon_{\mathrm{iid}} \le A_{\mathrm{crit}} K^{-1/4} + \chi_m$ quoted in Proposition 4.3, with

$$A_{\mathrm{crit}} \coloneqq \left(\frac{R_{\mathcal{B}} G_{\mathrm{TD}}}{1-\sqrt{\gamma}}\right)^{1/2}, \qquad \chi_m \coloneqq \left(\frac{R_{\mathcal{B}}\varepsilon_g}{1-\sqrt{\gamma}}\right)^{1/2} + \varepsilon_f(m) + \frac{\varepsilon_{\mathrm{app}}(m)}{1-\sqrt{\gamma}} + 2Q_{\max}\sqrt{\delta_{\mathrm{net}}}. \tag{58}$$

By (55) and $R_{\mathcal{B}} = 2B\sqrt{D}$, the constant $A_{\mathrm{crit}}$ depends on $D$, $B$, $Q_0$, $G_0$, $R_{\max}$, $\gamma$, and upper bounds on $\varepsilon_f, \varepsilon_\phi$, but not on $K$.

*Proof.* Fix an agent $i$ and a deterministic $\theta \in \Theta_b$, and suppress these indices. We first work on the event $\mathcal{E}_{i,\omega}$ and condition on the initialization $\omega_i$.

*Step 1: variational inequality for the tangent class.* Write $q_W \coloneqq \widehat{Q}_{i,\kappa}^{\tan,\omega_i}(\xi_i(\cdot); W)$ for $W \in \mathcal{B}_i(B;\omega_i)$, and let $q^\star \coloneqq Q_{i,\kappa}^{\star,\tan,\theta,\omega_i} = q_{W^\star}$ be the fixed point of $\Pi_{i,\kappa,\theta}^{\tan,\omega_i} T_{i,\kappa}^\theta$, which exists and is unique because that map is a $\sqrt{\gamma}$-contraction. Since $q^\star$ is the $L_2(\nu_{i,\kappa,\theta})$ projection of $T_{i,\kappa}^\theta q^\star$ onto the convex set $\mathcal{Q}_{i,\kappa}^{\tan}(\omega_i)$,

$$\big\langle q^\star - T_{i,\kappa}^\theta q^\star,\, q_W - q^\star \big\rangle_{L_2(\nu_{i,\kappa,\theta})} \ge 0 \qquad \text{for all } W \in \mathcal{B}_i(B;\omega_i). \tag{59}$$

Adding and subtracting $T_{i,\kappa}^\theta q^\star$ and applying Lemma E.1 with Cauchy–Schwarz gives

$$\big\langle q_W - T_{i,\kappa}^\theta q_W,\, q_W - q^\star \big\rangle_{L_2(\nu_{i,\kappa,\theta})} \ge (1-\sqrt{\gamma})\|q_W - q^\star\|^2_{L_2(\nu_{i,\kappa,\theta})}. \tag{60}$$

*Step 2: conditional unbiasedness.* For the $k$-th tuple, let $Y_k^\kappa \sim \nu_{i,\kappa,\theta}$ be the local configuration, let $Y_k^{\kappa,+}$ be its true successor, and write $X_k \coloneqq \xi_i(Y_k^\kappa)$ and $X_k^+ \coloneqq \xi_i(Y_k^{\kappa,+})$ for their encodings, so that $\mathsf{Z}_k = (X_k, R_k, X_k^+)$. Let

$$g_k^{\tan} \coloneqq \big[q_{W_k}(X_k) - R_k - \gamma q_{W_k}(X_k^+)\big]\nabla_W \widehat{Q}_{i,\kappa}^{\omega_i}(X_k; W_{i,\mathrm{init}})$$

and let $g_k^{\mathrm{nn}}$ be the semi-gradient actually used by Algorithm 2, namely the same expression with $q_{W_k}$ replaced by $\widehat{Q}_{i,\kappa}^{\omega_i}(\cdot; W_k)$ and the feature replaced by $\nabla_W \widehat{Q}_{i,\kappa}^{\omega_i}(X_k; W_k)$. Because $\widehat{Q}_{i,\kappa}^{\tan,\omega_i}(x;\cdot)$ is affine with gradient frozen at $W_{i,\mathrm{init}}$,

$$\big\langle \nabla_W \widehat{Q}_{i,\kappa}^{\omega_i}(X_k; W_{i,\mathrm{init}}),\, W_k - W^\star \big\rangle_{\mathrm{F}} = q_{W_k}(X_k) - q^\star(X_k).$$

Conditionally on $\omega_i$, the iterate $W_k$ is a function of $\mathsf{Z}_0, \dots, \mathsf{Z}_{k-1}$ and the tuples are i.i.d., so $\mathsf{Z}_k$ is independent of $W_k$; since $\mathcal{E}_{i,\omega} \in \sigma(\omega_i)$, restricting to that event does not disturb this independence, and since $Y_k^\kappa \sim \nu_{i,\kappa,\theta}$ exactly,

$$\mathbb{E}\big[\langle g_k^{\tan}, W_k - W^\star\rangle_{\mathrm{F}} \;\big|\; W_k\big] = \big\langle q_{W_k} - T_{i,\kappa}^\theta q_{W_k},\, q_{W_k} - q^\star \big\rangle_{L_2(\nu_{i,\kappa,\theta})}. \tag{61}$$

This is the step that fails for a single sequential restart chain, where $W_k$ depends on past states correlated with $X_k$.

*Step 3: perturbation and recursion.* By (52), (53), and (55), $\|g_k^{\mathrm{nn}} - g_k^{\tan}\|_{\mathrm{F}} \le \varepsilon_g$ and $\|g_k^{\mathrm{nn}}\|_{\mathrm{F}} \le G_{\mathrm{TD}}$. Nonexpansiveness of the projection onto the convex set $\mathcal{B}_i(B;\omega_i)$, followed by (60), (61), $\|W_k - W^\star\|_{\mathrm{F}} \le R_{\mathcal{B}}$, and taking expectations, yields

$$2\beta(1-\sqrt{\gamma})\mathbb{E}\|q_{W_k} - q^\star\|^2_{L_2(\nu_{i,\kappa,\theta})} \le \mathbb{E}\|W_k - W^\star\|_{\mathrm{F}}^2 - \mathbb{E}\|W_{k+1} - W^\star\|_{\mathrm{F}}^2 + 2\beta R_{\mathcal{B}}\varepsilon_g + \beta^2 G_{\mathrm{TD}}^2. \quad (62)$$

Summing over $k = 0, \dots, K-1$, dividing by $2\beta(1-\sqrt{\gamma})K$, and using $\|W_0 - W^\star\|_{\mathrm{F}} \le R_{\mathcal{B}}$ for the initialization $W_0 = W_{i,\mathrm{init}}$ gives

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\|q_{W_k} - q^\star\|^2_{L_2(\nu_{i,\kappa,\theta})} \le \frac{R_{\mathcal{B}}^2}{2\beta(1-\sqrt{\gamma})K} + \frac{R_{\mathcal{B}}\varepsilon_g}{1-\sqrt{\gamma}} + \frac{\beta G_{\mathrm{TD}}^2}{2(1-\sqrt{\gamma})}. \quad (63)$$

Since $W \mapsto q_W$ is affine and $W_i^{\mathrm{out}} = K^{-1}\sum_{k=0}^{K-1} W_{i,k}$ by (40), Jensen's inequality bounds $\mathbb{E}\|q_{W_i^{\mathrm{out}}} - q^\star\|^2$ by the left-hand side of (63). Substituting (56) gives

$$\mathbb{E}\|q_{W_i^{\mathrm{out}}} - q^\star\|^2_{L_2(\nu_{i,\kappa,\theta})} \le \frac{R_{\mathcal{B}} G_{\mathrm{TD}}}{(1-\sqrt{\gamma})\sqrt{K}} + \frac{R_{\mathcal{B}}\varepsilon_g}{1-\sqrt{\gamma}}. \quad (64)$$

*Step 4: back to the network and to $Q^\theta_{i,\kappa}$.* Taking square roots in (64) and adding (52) gives the same root-mean-square bound for $\widehat{Q}^{\omega_i}_{i,\kappa}(\xi_i(\cdot); W_i^{\mathrm{out}})$ with an extra $\varepsilon_f(m)$. For the remaining distance to the conditional Bellman fixed point, nonexpansiveness of the projection, the two fixed-point identities, and Lemma E.1 give

$$\left\| Q^{\star,\tan,\theta,\omega_i}_{i,\kappa} - Q^\theta_{i,\kappa} \right\|_{L_2(\nu_{i,\kappa,\theta})} \le \sqrt{\gamma} \left\| Q^{\star,\tan,\theta,\omega_i}_{i,\kappa} - Q^\theta_{i,\kappa} \right\|_{L_2(\nu_{i,\kappa,\theta})} + e_{\mathrm{app},i}(\theta;\omega_i).$$

Consequently,

$$\left( \mathbb{E}_{\omega_i}\left[ \mathbf{1}_{\mathcal{E}_{i,\omega}} \left\| Q^{\star,\tan,\theta,\omega_i}_{i,\kappa} - Q^\theta_{i,\kappa} \right\|^2_{L_2(\nu_{i,\kappa,\theta})} \right] \right)^{1/2} \le \frac{\varepsilon_{\mathrm{app}}(m)}{1-\sqrt{\gamma}}. \quad (65)$$

Bounded rewards imply $\|Q^\theta_{i,\kappa}\|_{L_\infty(\nu_{i,\kappa,\theta})} \le Q_{\max}$, so clipping cannot increase the distance to $Q^\theta_{i,\kappa}$ in $L_2(\nu_{i,\kappa,\theta})$. Suppressing the common $(\theta, \omega_i, \mathsf{D}_\theta^{\mathrm{iid}})$ arguments, for every realization of $(\omega_i, \mathsf{D}_\theta^{\mathrm{iid}})$

$$\begin{aligned} &\left\| \overline{Q}^{\mathrm{out,iid}}_{i,\kappa} - Q^\theta_{i,\kappa} \right\|_{L_2(\nu_{i,\kappa,\theta})} \\ &\le \left\| \widehat{Q}^{\omega_i}_{i,\kappa}(\xi_i(\cdot); W_i^{\mathrm{out,iid}}) - Q^{\star,\tan,\theta,\omega_i}_{i,\kappa} \right\|_{L_2(\nu_{i,\kappa,\theta})} + \left\| Q^{\star,\tan,\theta,\omega_i}_{i,\kappa} - Q^\theta_{i,\kappa} \right\|_{L_2(\nu_{i,\kappa,\theta})}. \end{aligned}$$

Minkowski's inequality on the joint initialization–sample probability space, followed by (64), (52), and (65), bounds the expected risk restricted to $\mathcal{E}_{i,\omega}$ by the first three terms of (57).

*Step 5: the exceptional event.* Off $\mathcal{E}_{i,\omega}$ both functions inside the norm take values in $[-Q_{\max}, Q_{\max}]$, so their $L_2(\nu_{i,\kappa,\theta})$ distance is at most $2Q_{\max}$. Splitting the expectation on $\mathcal{E}_{i,\omega}$ and its complement and applying Minkowski's inequality once more adds at most $2Q_{\max}\sqrt{\delta_{\mathrm{net}}}$, giving (57). Every bound used is uniform over agents and over deterministic $\theta \in \Theta_b$. □

**Corollary E.6** (Explicit critic rate)**.** *Suppose $\varepsilon_f(m), \varepsilon_\phi(m) \le C_{\mathrm{lin}} m^{-a}$ for some $a > 0$, and suppose the target admits tangent approximation at rate $\varepsilon_{\mathrm{app}}(m) \le C_{\mathrm{app}} m^{-s}$ for some $s > 0$. Regard $D, B, Q_0, G_0, R_{\max}, C_{\mathrm{lin}}$, and $\gamma$ as fixed. Then there is $C_{\mathrm{crit}} < \infty$, depending only on those quantities, with*

$$\varepsilon_{\mathrm{iid}} \le C_{\mathrm{crit}}\big(K^{-1/4} + m^{-a/2}\big) + \frac{C_{\mathrm{app}}}{1-\sqrt{\gamma}} m^{-s} + 2Q_{\max}\sqrt{\delta_{\mathrm{net}}}. \quad (66)$$

*Consequently, for any $\epsilon > 0$ the finite choices*

$$K \ge \Big(\frac{4C_{\mathrm{crit}}}{\epsilon}\Big)^4, \quad m \ge \max\left\{ \Big(\frac{4C_{\mathrm{crit}}}{\epsilon}\Big)^{2/a}, \Big(\frac{4C_{\mathrm{app}}}{(1-\sqrt{\gamma})\epsilon}\Big)^{1/s} \right\}, \quad \delta_{\mathrm{net}} \le \frac{\epsilon^2}{64 Q_{\max}^2} \quad (67)$$

*give $\varepsilon_{\mathrm{iid}} \le \epsilon$ using an expected $\mathcal{O}\big(\epsilon^{-4}/(1-\gamma)\big)$ environment transitions per outer iteration by* (41).

These rate substitutions are conditional on the stated network estimates. If the constants of those estimates depend on the initialization confidence, that dependence must be retained: obtaining a prescribed $\delta_{\mathrm{net}}$ requires a valid initialization bound for the same architecture and projection radius, so $\delta_{\mathrm{net}}$ is not an independent tuning parameter. The finite-budget statement that does not require a vanishing representation floor is Corollary 4.6.

*Proof.* Under the stated rates, (55) gives $\varepsilon_g = \mathcal{O}(m^{-a})$, so the gradient linearization error $\varepsilon_\phi(m)$ and the Frobenius error $\varepsilon_g(m)$ obey the same rate, with $G_{\mathrm{TD}}$ and $Q_B$ bounded by constants, so the bracket in (57) is $\mathcal{O}(K^{-1/2} + m^{-a})$ and its square root is $\mathcal{O}(K^{-1/4} + m^{-a/2})$. Adding $\varepsilon_f(m) = \mathcal{O}(m^{-a})$ preserves this order, yielding (66). The choices in (67) make each of the four terms at most $\epsilon/4$. □

*Remark* E.7 (The remaining approximation term). Corollary E.6 makes the optimization error explicit in $K$ and the linearization error explicit in $m$, but it retains a hypothesis on $\varepsilon_{\mathrm{app}}(m)$, namely that $Q^\theta_{i,\kappa}$ is approximable by the tangent class at a polynomial rate. This is an assumption on the regularity of the target, and it cannot be removed. Without further regularity, a measurable $Q^\theta_{i,\kappa}$ admits no uniform width–accuracy rate and need not lie in the projected tangent class, so a smoothness or bounded-RKHS-norm condition enters here. The same term appears as an approximation assumption in the neural-TD and neural actor–critic literature (Cai et al., 2019; Gaur et al., 2024; Zhang et al., 2025). What changes relative to (47) is the location of the assumption: it is now a statement about the representability of the target by the initialization-induced tangent class, not about the outcome of the TD recursion. The trivial substitution $\varepsilon_{\mathrm{iid}} = 2Q_{\max}$ remains valid but is no longer the only available reading of the bound.

**Proposition E.8** (Conditional transfer to the per-iteration interface)**.** *Suppose that, conditionally on $\mathcal{H}_t$, the round-$t$ critic initialization has its fresh generic law and is independent of the round-$t$ critic tuples, and that those tuples are the independent geometric-horizon draws of Algorithm 2. Then, for every $t$ with $\theta_t \in \Theta_b$, Assumption 4.2 holds with*

$$\varepsilon_{\mathrm{TD}} = \varepsilon_{\mathrm{iid}}, \tag{68}$$

*where $\varepsilon_{\mathrm{iid}}$ is given explicitly by* (57)*.*

*Proof.* Fix an agent $i$ and outer iteration $t$, condition on $\mathcal{H}_t$, and write $\theta = \theta_t$. Under this conditioning $\theta$ is deterministic, and by hypothesis the round-$t$ initialization and the round-$t$ critic tuples have exactly the product law used in Theorem E.5: a fresh $\omega_{i,t}$ independent of $K$ tuples that are i.i.d. with $(S_{t,k}, A_{t,k}) \sim \nu_\theta$ and true successors. No burn-in correction is needed, because each geometric-horizon draw is already exactly $\nu_\theta$-distributed rather than approaching $\nu_\theta$ along a chain. Theorem E.5 is uniform over agents and over deterministic $\theta \in \Theta_b$, so it applies with $\theta = \theta_t$ on the event $\{\theta_t \in \Theta_b\}$, which is $\mathcal{H}_t$-measurable. This is precisely (18) with $\varepsilon_{\mathrm{TD}} = \varepsilon_{\mathrm{iid}}$. □

*Remark* E.9 (Relation to existing neural-TD results). Theorem E.5 is stated for the conditional Bellman target $T^\theta_{i,\kappa}$ of Lemma E.1, which the finite-hop critic actually regresses on. For a two-layer network and i.i.d. stationary samples, Cai et al. (2019, Theorem 4.6) bounds the mean-square error to a tangent-model stationary point and their Proposition 4.7 supplies the approximation term; our Step 2 uses independent tuples for the same reason their analysis does. Gaur et al. (2024) handle multi-layer networks under Markovian sampling, at the cost of a mixing-time analysis we avoid by sampling the discounted occupancy directly. Zhang et al. (2025) obtain $(K, m, D)$ dependence for a different critic and sampling scheme while still retaining an approximation assumption, as we do in Corollary E.6; see Remark E.7.

Clipping makes the premise integrable without a separate network-output envelope and always permits the trivial choice $\varepsilon_{\mathrm{iid}} = 2Q_{\max}$; the point of (57) is that a nontrivial choice is now available. If a fixed-policy theorem instead gives $\|\overline{Q}^{\mathrm{out,iid}}_{i,\kappa} - Q^\theta_{i,\kappa}\|_{L_2(\nu_{i,\kappa,\theta})} \le r$ on an event of probability at least $1-\delta$, the bounded loss gives $\varepsilon^2_{\mathrm{iid}} \le r^2 + 4Q^2_{\max}\delta$.

*Remark* E.10 (Choice of the actor sampler). Only the critic sampler is replaced. The actor still uses the restart chain of Appendix D.1 with burn-in $H_{\mathrm{a}}$, so the $\gamma^{H_{\mathrm{a}}/2}$ and $M^{-1/2}$ terms in (19) are unchanged. Applying the argument of Theorem E.5 to a single sequential critic chain would require a Markovian stochastic-approximation analysis and would replace $K^{-1/4}$ by $\widetilde{\mathcal{O}}((\tau_{\mathrm{mix}}/K)^{1/4})$; the stationary-to-burn-in coupling of Proposition D.1 compares a nonstationary path with a stationary

one and does not by itself remove the time correlation within the stationary path, so it cannot substitute for that analysis. Sampling independent tuples buys the cleaner statement at the cost of the $1/(1-\gamma)$ transition factor in (41).

*Remark* E.11 (Conditional use of a fixed-policy bound). Conditionally on $\mathcal{H}_t$, the realized $\theta_t$ is fixed before $\omega_{i,t}$ and the critic data are drawn. Therefore the fixed-policy expectation in Assumption E.2 may be applied directly at $\theta_t$. Its deterministic constant is uniform over possible iterates, while the expectation is taken over the fresh round-$t$ initialization and critic data.

### E.3 Policy-gradient estimation error

The constants of Section 4.2 entering the actor-direction error (19) are, in displayed form,

$$\begin{aligned} C_{\mathrm{loc},\kappa} &:= \frac{L_\psi c_{\mathrm{EDP}}}{1-\gamma} + \frac{n_\kappa L_\psi \gamma c_{\mathrm{EDP}}}{N(1-\gamma)^2}, \quad C_{\mathrm{TD},\kappa} := \frac{n_\kappa L_\psi}{N(1-\gamma)}, \\ F_\kappa &:= \frac{n_\kappa L_\psi Q_{\max}}{N(1-\gamma)}, \qquad\qquad C_\gamma := 1 + \frac{8\gamma}{1-\gamma}, \end{aligned} \tag{69}$$

where $n_\kappa$ is the largest $\kappa$-hop neighborhood size defined in Section 3.3. Throughout the remainder of this appendix, $J_{\mathrm{rng}}$ is as defined in Section 4.2. We now combine critic error with restart-chain sampling.

**Lemma E.12** (Vector-valued restart-chain batch)**.** *Let $(Z_\ell)_{\ell\ge 0}$ be a restart chain with stationary distribution $\nu_\theta$, run for $H \in \mathbb{Z}_{\ge 0}$ burn-in steps before recording $M \in \mathbb{Z}_{>0}$ consecutive samples. If $f$ is Euclidean-valued and $\|f(z)\|_2 \le F$, then*

$$\left(\mathbb{E}\left\|\frac{1}{M}\sum_{\ell=1}^{M} f(Z_{H+\ell}) - \mathbb{E}_{Z\sim\nu_\theta}[f(Z)]\right\|_2^2\right)^{1/2} \le F\left(2\gamma^{H/2} + \sqrt{\frac{C_\gamma}{M}}\right). \tag{70}$$

*Proof.* For a stationary copy, let $\bar f = \mathbb{E}_{\nu_\theta} f$. For $r \in \mathbb{Z}_{\ge 1}$, define $m_r(z) := \mathbb{E}[f(Z_r) - \bar f \mid Z_0 = z]$. By (33) and the dual characterization of total variation,

$$\|m_r(z)\|_2 \le 2F\gamma^r.$$

Since $\|f(z) - \bar f\|_2 \le 2F$, it follows that

$$\left|\mathbb{E}\langle f(Z_0) - \bar f, f(Z_r) - \bar f\rangle\right| \le 4F^2\gamma^r.$$

Moreover, the variance identity $\mathbb{E}\|f(Z_0) - \bar f\|_2^2 = \mathbb{E}\|f(Z_0)\|_2^2 - \|\bar f\|_2^2$ together with $\|f\|_2 \le F$ gives $\mathbb{E}\|f(Z_0) - \bar f\|_2^2 \le F^2 - \|\bar f\|_2^2 \le F^2$; note that this is the centered second moment, so the cruder bound $(2F)^2$ is not needed here. Expanding the squared norm of the stationary batch average therefore gives

$$\mathbb{E}\left\|\frac{1}{M}\sum_{\ell=1}^{M}(f(Z_\ell) - \bar f)\right\|_2^2 \le \frac{F^2}{M} + \frac{8F^2}{M^2}\sum_{r=1}^{M-1}(M-r)\gamma^r \le \frac{F^2}{M} + \frac{8F^2}{M}\cdot\frac{\gamma}{1-\gamma} = \frac{C_\gamma F^2}{M},$$

where the last step used $\sum_{r\ge 1}(M-r)\gamma^r \le M\gamma/(1-\gamma)$. For an arbitrary initial state, couple the chain after burn-in to a stationary copy. Couple the law of $Z_H$ to $\nu_\theta$ maximally and drive both chains with common randomness afterwards; Proposition D.1 bounds the coupling failure probability by $\gamma^H$, the two blocks coincide on the success event, and the two averages differ by at most $2F$ on failure. Thus the mean square of their difference is at most $4F^2\gamma^H$, so its root mean square is at most $2F\gamma^{H/2}$; Minkowski's inequality in $L_2$ completes the proof. □

*Proof of Proposition 4.4.* For compactness, write

$$\mathbb{E}_t[\,\cdot\,] := \mathbb{E}[\,\cdot \mid \mathcal{H}_t],$$

and let

$$\mathcal{G}_t := \sigma\big(\mathcal{H}_t, \{\omega_{j,t}, W_{j,t}^{\mathrm{out}}\}_{j=1}^N\big).$$

Thus $\mathcal{H}_t \subseteq \mathcal{G}_t$, and the tower property may be used between the two conditional expectations. Conditionally on $\mathcal{G}_t$, the complete instantiated critics are fixed. By the independence hypothesis in the proposition, the fresh actor restart chain retains its stated law after this conditioning.

*Step 1 (actor-sampling error given fixed critics).* Condition on $\mathcal{G}_t$ and recall the clipped critic output $\overline{Q}_{j,\kappa,t}$ from (14). Define its population actor direction

$$\overline{g}_{i,\kappa,t} \coloneqq \frac{1}{N(1-\gamma)} \mathbb{E}_{\nu_{\theta_t}} \left[ \psi_i^{\theta_t}(S, A_i) \sum_{j \in \mathcal{N}_i^\kappa} \overline{Q}_{j,\kappa,t}(Y_j^\kappa) \right].$$

Let $\overline{g}_{\kappa,t}$ be the concatenation of these blocks. Conditionally on $\mathcal{G}_t$ the critics $\{W_{j,t}^{\mathrm{out}}\}_j$ are deterministic, so $\overline{g}_{\kappa,t}$ is a fixed vector and the only randomness is the actor restart chain, which is independent of critic training given $\mathcal{H}_t$ by hypothesis. Clipping in (14) gives $|q_{j,t,\ell}| \le Q_{\max}$, so the single-sample statistic $f$ obeys

$$\|f(z)\|_2^2 \le \frac{n_\kappa^2 Q_{\max}^2}{N^2(1-\gamma)^2} \sum_{i=1}^N L_{\psi,i}^2 = F_\kappa^2,$$

and Lemma E.12 gives

$$\left(\mathbb{E}\left[\|h_t - \overline{g}_{\kappa,t}\|_2^2 \,\middle|\, \mathcal{G}_t\right]\right)^{1/2} \le F_\kappa \left( 2\gamma^{H_{\mathrm{a}}/2} + \sqrt{\frac{C_\gamma}{M}} \right). \tag{71}$$

*Step 2 (critic error, folded in blockwise).* Set

$$\zeta_{j,t} \coloneqq \left\| \overline{Q}_{j,\kappa,t} - Q_{j,\kappa}^{\theta_t} \right\|_{L_2(\nu_{j,\kappa,\theta_t})}, \qquad \eta_\kappa \coloneqq \frac{\gamma c_{\mathrm{EDP}}}{1-\gamma} \rho^{\kappa+1}.$$

Assumption 4.2 directly gives, for every $j$,

$$\left(\mathbb{E}_t[\zeta_{j,t}^2]\right)^{1/2} \le \varepsilon_{\mathrm{TD}}. \tag{72}$$

For every realization of the critics, the score bound, Cauchy–Schwarz in $L_2(\nu_{j,\kappa,\theta_t})$, the probability-measure embedding $\|f\|_{L_2(\nu_{j,\kappa,\theta_t})} \le \|f\|_{L_\infty(\nu_{j,\kappa,\theta_t})}$, and (29) imply, block by block,

$$\|\overline{g}_{i,\kappa,t} - g_{i,\kappa}(\theta_t)\|_2 \le \frac{L_{\psi,i}}{N(1-\gamma)} \sum_{j \in \mathcal{N}_i^\kappa} (\zeta_{j,t} + \eta_\kappa). \tag{73}$$

Conditional Minkowski's inequality and (72) yield

$$\begin{aligned} \left(\mathbb{E}_t\left[\|\overline{g}_{i,\kappa,t} - g_{i,\kappa}(\theta_t)\|_2^2\right]\right)^{1/2} &\le \frac{L_{\psi,i}}{N(1-\gamma)} \sum_{j \in \mathcal{N}_i^\kappa} \left( \left(\mathbb{E}_t[\zeta_{j,t}^2]\right)^{1/2} + \eta_\kappa \right) \\ &\le \frac{n_\kappa L_{\psi,i}}{N(1-\gamma)} (\varepsilon_{\mathrm{TD}} + \eta_\kappa). \end{aligned}$$

Finally, summing the blocks and using $L_\psi^2 = \sum_i L_{\psi,i}^2$ give

$$\begin{aligned} \left(\mathbb{E}_t\left[\|\overline{g}_{\kappa,t} - g_\kappa(\theta_t)\|_2^2\right]\right)^{1/2} &= \left( \sum_{i=1}^N \mathbb{E}_t\left[\|\overline{g}_{i,\kappa,t} - g_{i,\kappa}(\theta_t)\|_2^2\right] \right)^{1/2} \\ &\le C_{\mathrm{TD},\kappa} \varepsilon_{\mathrm{TD}} + \frac{n_\kappa L_\psi \gamma c_{\mathrm{EDP}}}{N(1-\gamma)^2} \rho^{\kappa+1}. \end{aligned} \tag{74}$$

*Step 3 (assembly).* Applying the tower property to (71) bounds $\left(\mathbb{E}_t[\|h_t - \overline{g}_{\kappa,t}\|_2^2]\right)^{1/2}$ by the same sampling term. Lemma 3.1 gives the deterministic localization bound

$$\|g_\kappa(\theta_t) - \nabla J(\theta_t)\|_2 \le \frac{L_\psi c_{\mathrm{EDP}}}{1-\gamma} \rho^{\kappa+1}. \tag{75}$$

Combining (71)–(75) by Minkowski's inequality in $L_2$ proves (20). □

### E.4 A SUFFICIENT POLICY-GEOMETRY CONDITION

We derive the gradient-domination condition used in the main theorem from Fisher nondegeneracy and compatible function approximation (Sutton et al., 1999; Kakade, 2001; Agarwal et al., 2021). Let $\psi^\theta(s,a) = (\psi_1^\theta(s,a_1), \dots, \psi_N^\theta(s,a_N))$ and define

$$\mathcal{F}(\theta) \coloneqq \mathbb{E}_{\nu_\theta}[\psi^\theta(S,A)\psi^\theta(S,A)^\top].$$

Assume that $J_b^\star$ is attained by some $\theta_b^\star \in \Theta_b$, that the comparator occupancy $\nu_b^\star$ is absolutely continuous with respect to $\nu_\theta$ for every $\theta \in \Theta_b$ (common support), so that $\psi^\theta$ is $\nu_b^\star$-integrable and the expectations below are well defined, and let $\nu_b^\star(s,a) = d_{\theta_b^\star}(s)\pi_{\theta_b^\star}(a \mid s)$. Define the team advantage

$$A^\theta(s,a) \coloneqq Q^\theta(s,a) - \mathbb{E}_{A' \sim \pi_\theta(\cdot|s)}[Q^\theta(s,A')].$$

**Proposition E.13** (Fisher-compatible gradient domination)**.** *Let $\mu_F > 0$, $G_\star > 0$, and $\varepsilon_{\mathrm{bias}} \geq 0$. Suppose that, for every $\theta \in \Theta_b$, the smallest positive eigenvalue of $\mathcal{F}(\theta)$ is at least $\mu_F$, $\nabla J(\theta)$ lies in the range of $\mathcal{F}(\theta)$, and*

$$w_\theta \coloneqq \mathcal{F}(\theta)^\dagger \nabla J(\theta)$$

*satisfies*

$$\mathbb{E}_{\nu_b^\star}\Big[\big(A^\theta(S,A) - (1-\gamma)w_\theta^\top \psi^\theta(S,A)\big)^2\Big] \leq \varepsilon_{\mathrm{bias}}, \tag{76}$$

$$\mathbb{E}_{\nu_b^\star}[\|\psi^\theta(S,A)\|_2^2] \leq G_\star^2. \tag{77}$$

*Then, for every $\theta \in \Theta_b$,*

$$\frac{\mu_F}{G_\star}\big(J_b^\star - J(\theta)\big) \leq \|\nabla J(\theta)\|_2 + \frac{\mu_F}{G_\star(1-\gamma)}\sqrt{\varepsilon_{\mathrm{bias}}}. \tag{78}$$

*Proof.* The standard performance-difference argument (Kakade & Langford, 2002; Agarwal et al., 2021, Lemma 2) gives

$$J_b^\star - J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{\nu_b^\star}[A^\theta(S,A)].$$

The compatible-advantage bound (76), together with (77) and Cauchy–Schwarz, implies

$$J_b^\star - J(\theta) \leq G_\star \|w_\theta\|_2 + \frac{\sqrt{\varepsilon_{\mathrm{bias}}}}{1-\gamma}.$$

Fisher nondegeneracy gives $\|w_\theta\|_2 \leq \|\nabla J(\theta)\|_2/\mu_F$. Rearranging proves (78). □

Two remarks on Proposition E.13 are in order. First, (78) is the unconstrained form, stated through $\nabla J$ rather than through the gradient mapping of (17). The two are related by the exact identity

$$\big\|\nabla J(\theta) - \mathcal{G}_\alpha(\theta)\big\|_2 = \frac{1}{\alpha}\operatorname{dist}\big(\theta + \alpha\nabla J(\theta),\, \Theta_b\big), \tag{79}$$

since $\alpha\big(\nabla J(\theta) - \mathcal{G}_\alpha(\theta)\big) = \big(\theta + \alpha\nabla J(\theta)\big) - \Pi_{\Theta_b}\big(\theta + \alpha\nabla J(\theta)\big)$, the normal-cone component discarded by the projection. Accordingly (78) supplies clause (iii) with $\sqrt{\mu} = \mu_F/G_\star$ and $C_{\mathrm{hc}} = \mu_F/[G_\star(1-\gamma)]$ at every $\theta$ whose ascent step is feasible, and in general up to the deficit (79); clause (iii) as stated asks in addition that this deficit not create a nonoptimal constrained stationary point, and Proposition E.14 verifies that requirement in closed form on a network family. Second, the bounded-score and smoothness clauses of Assumption 4.1 are separate conditions and must be verified on their own.

The next result exhibits a family on which every constant in Assumption 4.1 and Definition 2.1 is available in closed form. The construction couples states across neighbours and puts the action interaction in the reward, while keeping the state transition free of the actions.

**Proposition E.14** (A network family with explicit policy constants)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be undirected with maximum degree $\Delta_{\mathcal{G}}$, and let $\mathcal{S}_i = \mathcal{A}_i = \{0,1\}$ with*

$$P_i(S_i' = 1 \mid s,a) = \frac{1}{4} + \frac{1}{2|\mathcal{N}_i|}\sum_{j \in \mathcal{N}_i} s_j, \tag{80}$$

*and* $\mu_0$ *the product of independent* $\mathrm{Bernoulli}(1/2)$ *laws. For* $\eta > 0$ *and* $0 < \zeta\Delta_{\mathcal{G}} < 1$ *let*

$$r_i(s,a) = a_i - \frac{\zeta}{2} \sum_{j:\{i,j\}\in\mathcal{E}} a_i a_j + \eta s_i, \tag{81}$$

*so that* $r_R = 1$ *and* $R_{\max} = 1 + \zeta\Delta_{\mathcal{G}}/2 + \eta$*. Take the localized class to be the state-independent product policies*

$$\Theta_b := [-R, R]^N, \qquad \pi_{\theta_i}(1) = p_i = \sigma(\theta_i),$$

*which lies in the* $b = 0$ *class of* (2) *with* $R_\Theta = R$*, and write* $p_- = \sigma(-R)$*,* $p_+ = \sigma(R) = 1 - p_-$*. Then clauses* (i)–(ii) *of Assumption 4.1 hold with*

$$L_{\psi,i} = 1, \qquad L_J = \frac{1/4 + \zeta\Delta_{\mathcal{G}}/16}{N(1-\gamma)},$$

*the hypotheses of Proposition E.13 hold with*

$$\mu_F = p_- p_+, \qquad G_\star = \sqrt{N}, \qquad \varepsilon_{\mathrm{bias}} = \left(\frac{\zeta|\mathcal{E}|}{N}\right)^2, \tag{82}$$

*so that* (78) *holds with*

$$\sqrt{\mu} = \frac{p_- p_+}{\sqrt{N}}, \qquad C_{\mathrm{hc}} = \frac{p_- p_+}{\sqrt{N}(1-\gamma)}.$$

*Clause* (iii) *itself holds on this family directly and with no residual: for every* $R \geq \frac{1}{2}$*, every actor step* $\alpha \leq 1/L_J$*, and every* $\theta \in \Theta_b$*,*

$$\sqrt{\mu_{\mathcal{G}}}\big(J_b^\star - J(\theta)\big) \leq \big\|\mathcal{G}_\alpha(\theta)\big\|_2, \qquad \sqrt{\mu_{\mathcal{G}}} := \frac{2(1-\zeta\Delta_{\mathcal{G}})\, p_- p_+}{R\sqrt{N}}, \tag{83}$$

*so clause* (iii) *may be taken with* $\sqrt{\mu} = \sqrt{\mu_{\mathcal{G}}}$ *and* $C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}} = 0$*. Finally, Definition 2.1 holds with* $c_{\mathrm{EDP}} = \eta/(1-\gamma)$ *and* $\rho = \gamma$*. All of these constants are independent of the optimization horizon.*

*Proof.* The transition (80) takes values in $[1/4, 3/4]$, depends on $s_{\mathcal{N}_i}$ only, and does not depend on $a$; the next local states are conditionally independent given $(s,a)$ because the driving randomness is independent across agents. The reward (81) depends on $(s_i, a_{\mathcal{N}_i})$, so $r_R = 1$.

*Objective.* Because the state process does not depend on the actions and the policy does not depend on the states, the actions at each time are independent across agents with $\mathbb{E}[A_i] = p_i$, and $\sum_i \sum_{j:\{i,j\}\in\mathcal{E}} p_i p_j = 2\sum_{\{i,j\}\in\mathcal{E}} p_i p_j$. Hence

$$J(\theta) = C_{\mathrm{s}} + \frac{1}{N(1-\gamma)} \left[\sum_i p_i - \zeta \sum_{\{i,j\}\in\mathcal{E}} p_i p_j\right], \tag{84}$$

with $C_{\mathrm{s}}$ independent of $\theta$. Since $1 - \zeta\sum_{j:\{i,j\}\in\mathcal{E}} p_j \geq 1 - \zeta\Delta_{\mathcal{G}} > 0$, the objective is strictly increasing in each $p_i$, so $J_b^\star$ is attained at $\theta^\star = (R, \ldots, R)$.

*Scores and smoothness.* The score is $\psi_i = a_i - p_i$, so $|\psi_i| \leq 1$ and $L_{\psi,i} = 1$. Writing $p_i' = p_i(1-p_i)$ and using $|p_i'| \leq 1/4$ and $|p_i''| \leq 1/4$, differentiating (84) gives

$$\left|\frac{\partial^2 J}{\partial\theta_i^2}\right| \leq \frac{1/4}{N(1-\gamma)}, \qquad \left|\frac{\partial^2 J}{\partial\theta_i \partial\theta_j}\right| \leq \frac{\zeta/16}{N(1-\gamma)} \ \text{ for } \{i,j\} \in \mathcal{E},$$

and zero otherwise, so the largest absolute row sum, and hence the operator norm of the Hessian, is at most $L_J$ as stated.

*Fisher matrix and comparator score bound.* The actions are conditionally independent with $\mathbb{E}[\psi_i] = 0$ and $\mathbb{E}[\psi_i^2] = p_i(1-p_i)$, so

$$\mathcal{F}(\theta) = \mathrm{diag}\big(p_i(1-p_i)\big)_{i=1}^N \succeq p_- p_+ I$$

for every $\theta \in \Theta_b$, which gives $\mu_F = p_- p_+$ and makes $\nabla J(\theta)$ lie in the range of $\mathcal{F}(\theta)$. Also $\|\psi^\theta\|_2^2 = \sum_i \psi_i^2 \leq N$ pointwise, so (77) holds with $G_\star = \sqrt{N}$.

*Compatible-advantage error.* The stage reward is the only action-dependent term and the state law is action-free, so $A^\theta(s,a) = \bar{r}(s,a) - \mathbb{E}_{A\sim\pi_\theta}[\bar{r}(s,A)]$. Substituting $u_i := a_i - p_i$ and $a_i a_j - p_i p_j = u_i u_j + p_j u_i + p_i u_j$ gives

$$A^\theta(s,a) = \frac{1}{N}\sum_i \left(1 - \zeta \sum_{j:\{i,j\}\in\mathcal{E}} p_j\right) u_i - \frac{\zeta}{N}\sum_{\{i,j\}\in\mathcal{E}} u_i u_j.$$

From (84) and $\mathcal{F}(\theta) = \mathrm{diag}(p_i')$,

$$w_{\theta,i} = \big(\mathcal{F}(\theta)^\dagger \nabla J(\theta)\big)_i = \frac{1 - \zeta\sum_{j:\{i,j\}\in\mathcal{E}} p_j}{N(1-\gamma)},$$

so $(1-\gamma) w_\theta^\top \psi^\theta$ reproduces the first sum exactly and

$$A^\theta - (1-\gamma) w_\theta^\top \psi^\theta = -\frac{\zeta}{N}\sum_{\{i,j\}\in\mathcal{E}} u_i u_j.$$

Since $|u_i| \le 1$, this residual is bounded by $\zeta|\mathcal{E}|/N$ pointwise, hence in mean square under $\nu_b^\star$, which gives (82). Proposition E.13 then supplies the stated $\sqrt{\mu}$ and $C_{\mathrm{hc}}$.

*Gradient-mapping domination.* Write $u_i := R - \theta_i \in [0, 2R]$ and $c_i := 1 - \zeta\sum_{j:\{i,j\}\in\mathcal{E}} p_j \ge 1 - \zeta\Delta_{\mathcal{G}} > 0$. Differentiating (84) gives $\partial_{\theta_i} J = c_i p_i'/[N(1-\gamma)] \ge a := (1-\zeta\Delta_{\mathcal{G}})p_- p_+/[N(1-\gamma)] > 0$, so every partial derivative is strictly positive and $\theta^\star = (R,\dots,R)$ is the unique maximizer. Because $\Theta_b$ is the box $[-R,R]^N$ and each partial derivative is positive, the projection in $\mathcal{G}_\alpha$ can clip only at the upper endpoint, so

$$\mathcal{G}_\alpha(\theta)_i = \min\left\{\partial_{\theta_i} J,\ \frac{u_i}{\alpha}\right\} \ge \min\left\{a,\ \frac{u_i}{\alpha}\right\} \ge 0.$$

If $u_i/\alpha \ge a$ the middle term equals $a \ge a u_i/(2R)$ because $u_i \le 2R$; otherwise it equals $u_i/\alpha \ge a u_i/(2R)$ because $\alpha a \le 2R$, which holds since $\alpha \le 1/L_J \le 4N(1-\gamma)$, $a \le 1/[4N(1-\gamma)]$ and $R \ge \frac{1}{2}$. Hence $\mathcal{G}_\alpha(\theta)_i \ge a u_i/(2R)$ for every $i$, and since all entries are nonnegative,

$$\big\|\mathcal{G}_\alpha(\theta)\big\|_2 \ge \frac{1}{\sqrt{N}}\sum_{i=1}^N \mathcal{G}_\alpha(\theta)_i \ge \frac{a}{2R\sqrt{N}}\sum_{i=1}^N u_i.$$

On the other side, $p_+^2 - p_i p_j = p_+(p_+ - p_j) + p_j(p_+ - p_i) \ge 0$, so dropping that term in (84) and using that $\sigma$ is $\frac{1}{4}$-Lipschitz gives

$$J_b^\star - J(\theta) \le \frac{1}{N(1-\gamma)}\sum_{i=1}^N (p_+ - p_i) \le \frac{1}{4N(1-\gamma)}\sum_{i=1}^N u_i.$$

Combining the two displays and substituting $a$ yields (83).

*Exponential decay.* Fix $\ell \ge r_R = 1$ and two pairs with $y_i^\ell(s,a) = y_i^\ell(\widetilde{s},\widetilde{a})$. The time-zero reward $r_i$ is a function of $(s_i, a_{\mathcal{N}_i})$ and therefore agrees. For $t \ge 1$ the actions are drawn from the state-independent policy, so the action-dependent part of $\mathbb{E}[r_i(S_t, A_t)]$ is a constant independent of the initial pair, and the only remaining dependence is $\eta\,\mathbb{E}[S_{i,t} \mid s]$. Iterating (80) shows that $\mathbb{E}[S_{i,t} \mid s]$ is a function of $s_{\mathcal{N}_i^t}$ alone, so the two values agree for every $t \le \ell$. For $t > \ell$ they differ by at most one, whence

$$\big|Q_i^\theta(s,a) - Q_i^\theta(\widetilde{s},\widetilde{a})\big| \le \eta \sum_{t=\ell+1}^\infty \gamma^t = \frac{\eta}{1-\gamma}\gamma^{\ell+1},$$

uniformly over $\theta \in \Theta_b$. □

Through (83) the same family satisfies clause (iii) with no policy-side residual at all, at the cost of a route that uses the explicit form of $J$ rather than the generic compatible-advantage argument; the $\varepsilon_{\mathrm{bias}}$ of (82) is therefore a property of the sufficient condition of Proposition E.13 and not of the family. The family verifies the policy-side and decay conditions only: the constants of the environments in Appendix F, containment of the unprojected update in $[-R,R]^N$ for all horizons, and the critic representation error are all outside its scope.

### E.5 PROJECTED ASCENT AND PROOF OF THE MAIN THEOREM

We complete the proof by analyzing the projected ascent step (16). Write

$$g_t \coloneqq \nabla J(\theta_t), \qquad e_t \coloneqq h_t - g_t, \qquad \Delta_t \coloneqq J_b^\star - J(\theta_t) \geq 0,$$

the last because $\theta_t \in \Theta_b$ and $J_b^\star = \sup_{\theta\in\Theta_b} J(\theta)$. Since $\theta_{t+1} = \Pi_{\Theta_b}(\theta_t + \alpha h_t)$, write the realized displacement and the empirical (sample) gradient mapping as

$$S_t \coloneqq \theta_{t+1} - \theta_t, \qquad \widehat{\mathcal{G}}_t \coloneqq \frac{1}{\alpha} S_t = \frac{1}{\alpha}\big(\Pi_{\Theta_b}(\theta_t + \alpha h_t) - \theta_t\big),$$

so that $\mathcal{G}_\alpha(\theta_t) = \frac{1}{\alpha}(\Pi_{\Theta_b}(\theta_t + \alpha g_t) - \theta_t)$ is its population counterpart from (17). Because the Euclidean projection onto the convex set $\Theta_b$ is nonexpansive,

$$\big\|\widehat{\mathcal{G}}_t - \mathcal{G}_\alpha(\theta_t)\big\|_2 = \frac{1}{\alpha}\big\|\Pi_{\Theta_b}(\theta_t + \alpha h_t) - \Pi_{\Theta_b}(\theta_t + \alpha g_t)\big\|_2 \leq \|h_t - g_t\|_2 = \|e_t\|_2. \tag{85}$$

**Lemma E.15** (Actor feasibility)**.** *For any step sizes and any realized directions, if $\theta_0 \in \Theta_b$ then the projected update* (16) *satisfies $\theta_t \in \Theta_b$ for every $t \geq 0$.*

*Proof.* Each block satisfies $\theta_{i,t+1} = \Pi_{\Theta_{b,i}}(\cdot) \in \Theta_{b,i}$ by definition of the Euclidean projection, and $\theta_0 \in \Theta_b = \prod_i \Theta_{b,i}$; hence $\theta_t = (\theta_{1,t}, \ldots, \theta_{N,t}) \in \Theta_b$ for all $t$. □

**Lemma E.16** (One-step sufficient increase)**.** *Suppose clause (ii) of Assumption 4.1 holds and the actor step satisfies $\alpha \leq 1/L_J$. Then for every outer iteration $t$,*

$$J(\theta_{t+1}) \geq J(\theta_t) + \frac{\alpha}{4}\big\|\widehat{\mathcal{G}}_t\big\|_2^2 - \alpha\|e_t\|_2^2. \tag{86}$$

*Proof.* Both $\theta_t, \theta_{t+1} \in \Theta_b$ by Lemma E.15, and $\Theta_b$ is convex, so the segment $[\theta_t, \theta_{t+1}] \subset \Theta_b$ and $L_J$-smoothness gives

$$J(\theta_{t+1}) \geq J(\theta_t) + \langle g_t, S_t\rangle - \frac{L_J}{2}\|S_t\|_2^2, \tag{87}$$

where $S_t = \theta_{t+1} - \theta_t = \alpha\widehat{\mathcal{G}}_t$. Because $\theta_{t+1} = \Pi_{\Theta_b}(\theta_t + \alpha h_t)$ is the Euclidean projection of $\theta_t + \alpha h_t$ onto the convex set $\Theta_b$, the obtuse-angle (variational) inequality gives, for the feasible point $\theta_t \in \Theta_b$,

$$\langle(\theta_t + \alpha h_t) - \theta_{t+1},\, \theta_t - \theta_{t+1}\rangle \leq 0,$$

that is, $\alpha\langle h_t, S_t\rangle \geq \|S_t\|_2^2$. Splitting $h_t = g_t + e_t$,

$$\langle g_t, S_t\rangle \geq \frac{1}{\alpha}\|S_t\|_2^2 - \langle e_t, S_t\rangle \geq \frac{1}{\alpha}\|S_t\|_2^2 - \|e_t\|_2\|S_t\|_2. \tag{88}$$

Substituting (88) into (87) and using $\alpha \leq 1/L_J$, i.e. $L_J/2 \leq 1/(2\alpha)$,

$$J(\theta_{t+1}) \geq J(\theta_t) + \frac{1}{\alpha}\|S_t\|_2^2 - \|e_t\|_2\|S_t\|_2 - \frac{1}{2\alpha}\|S_t\|_2^2 = J(\theta_t) + \frac{1}{2\alpha}\|S_t\|_2^2 - \|e_t\|_2\|S_t\|_2.$$

By Young's inequality $\|e_t\|_2\|S_t\|_2 \leq \frac{1}{4\alpha}\|S_t\|_2^2 + \alpha\|e_t\|_2^2$, so

$$J(\theta_{t+1}) \geq J(\theta_t) + \frac{1}{4\alpha}\|S_t\|_2^2 - \alpha\|e_t\|_2^2.$$

Finally $\|S_t\|_2^2 = \alpha^2\|\widehat{\mathcal{G}}_t\|_2^2$ gives (86). □

We first record the mapping-based stationarity rate, then convert it to a policy gap through clause (iii).

**Proposition E.17** (Averaged gradient-mapping stationarity)**.** *Suppose clauses (i)–(ii) of Assumption 4.1, Definition 2.1, Assumption 4.2, and the actor-sampling condition of Proposition 4.4 hold, that $\theta_0 \in \Theta_b$, and that $\alpha \leq 1/L_J$. Let $\tau$ be uniform on $\{0, \ldots, T-1\}$, independent of the run. Then*

$$\mathbb{E}\big\|\mathcal{G}_\alpha(\theta_\tau)\big\|_2^2 \leq \frac{8J_{\mathrm{rng}}}{\alpha T} + 10\,\varepsilon_{\mathrm{AC}}^2. \tag{89}$$

*Proof.* Rearranging (86), taking total expectations, and using $\|\widehat{\mathcal{G}}_t\|_2^2 \geq \frac{1}{2}\|\mathcal{G}_\alpha(\theta_t)\|_2^2 - \|e_t\|_2^2$ from (85) and $\|u+v\|_2^2 \leq 2\|u\|_2^2 + 2\|v\|_2^2$,

$$\frac{\alpha}{8}\mathbb{E}\|\mathcal{G}_\alpha(\theta_t)\|_2^2 \leq \mathbb{E}[J(\theta_{t+1}) - J(\theta_t)] + \frac{5\alpha}{4}\mathbb{E}\|e_t\|_2^2.$$

Proposition 4.4 and the tower property give $\mathbb{E}\|e_t\|_2^2 \leq \varepsilon_{\mathrm{AC}}^2$. Summing over $t = 0, \ldots, T-1$, telescoping $J$, and using $J(\theta_T) - J(\theta_0) \leq J_{\mathrm{rng}}$,

$$\frac{\alpha}{8}\sum_{t=0}^{T-1}\mathbb{E}\|\mathcal{G}_\alpha(\theta_t)\|_2^2 \leq J_{\mathrm{rng}} + \frac{5\alpha}{4}T\,\varepsilon_{\mathrm{AC}}^2.$$

Dividing by $\alpha T/8$ and using $\mathbb{E}\|\mathcal{G}_\alpha(\theta_\tau)\|_2^2 = T^{-1}\sum_{t<T}\mathbb{E}\|\mathcal{G}_\alpha(\theta_t)\|_2^2$ gives (89). □

*Proof of Theorem 4.5.* By clause (iii), $\sqrt{\mu}\,\Delta_t \leq \|\mathcal{G}_\alpha(\theta_t)\|_2 + C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}}$ for every $t$, so averaging over $\tau$ and using Jensen's inequality $\mathbb{E}\|\mathcal{G}_\alpha(\theta_\tau)\|_2 \leq (\mathbb{E}\|\mathcal{G}_\alpha(\theta_\tau)\|_2^2)^{1/2}$,

$$\sqrt{\mu}\,\mathbb{E}[\Delta_\tau] \leq \left(\mathbb{E}\|\mathcal{G}_\alpha(\theta_\tau)\|_2^2\right)^{1/2} + C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}}.$$

Substituting Proposition E.17 and dividing by $\sqrt{\mu}$ gives

$$\mathbb{E}[\Delta_\tau] \leq \frac{1}{\sqrt{\mu}}\left(\sqrt{\frac{8J_{\mathrm{rng}}}{\alpha T} + 10\,\varepsilon_{\mathrm{AC}}^2} + C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}}\right), \tag{90}$$

which is (21). □

Proposition E.17 uses only bounded scores, smoothness, the critic and actor-sampling conditions, and the projection that supplies domain containment. The constants of (19) displayed at the start of Appendix E.3 depend on $L_\psi$, $c_{\mathrm{EDP}}$, $n_\kappa$, $N$, $\gamma$, and $Q_{\max}$ only, so Proposition 4.4 is available without clause (iii) of Assumption 4.1. The price is that (89) bounds a projected-gradient-mapping norm at a randomly selected iterate rather than the policy gap.

*Remark* E.18 (Feasibility and the role of the compact domain). Feasibility is unconditional and requires no step-size or horizon condition. The blockwise projection preserves the product-policy factorization and the prescribed observation radius: by (2) each $\pi_{\theta_i}$ depends on $s_{\mathcal{N}_i^b}$ only, and (16) projects the $\theta_i$-block back onto $\Theta_{b,i}$, which lies in the $\theta_i$-block. By Lemma E.15 it keeps every iterate in the compact convex set $\Theta_b$, so the constants of Assumption 4.1 are required only on $\Theta_b$ and $J_b^\star$ is a bona fide constrained optimum. This is what the bounded domain buys: the constants can otherwise degrade along a trajectory, since for a softmax parameterization $\|\theta_i\|_2 \to \infty$ saturates $\pi_{\theta_i}$, the localized Fisher information degenerates, and the $\mu$ of Appendix E.4 tends to 0, inflating the right-hand side of (21); the domain $\Theta_{b,i}$ excludes this without disturbing the product, $b$-localized structure, which the blockwise projection preserves exactly. The projected unnormalized step needs no separate admissibility or projection-inactivity condition: the projection itself supplies containment, and the gradient-mapping analysis of Lemma E.16 and Proposition E.17 handles boundary iterates directly.

### E.6 Accuracy above the fixed error floor

Fix the policy domain, the localization radius, the critic architecture, the projection radius, and valid neural and policy constants, and let $B_{\mathrm{fixed}}$ be as in (22). By Proposition 4.3, $\varepsilon_{\mathrm{TD}} \leq A_{\mathrm{crit}}K^{-1/4} + \chi_m$, so substituting (19) into the bound (21) of Theorem 4.5 and applying $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ gives

$$\begin{aligned}\mathbb{E}\left[J_b^\star - J(\theta_\tau)\right] \leq \frac{B_{\mathrm{fixed}}}{\sqrt{\mu}} + \frac{1}{\sqrt{\mu}}\Bigg[\sqrt{\frac{8J_{\mathrm{rng}}}{\alpha T}} \\ + \sqrt{10}\left(C_{\mathrm{TD},\kappa}A_{\mathrm{crit}}K^{-1/4} + 2F_\kappa\gamma^{H_{\mathrm{a}}/2} + F_\kappa\sqrt{\tfrac{C_\gamma}{M}}\right)\Bigg],\end{aligned} \tag{91}$$

which separates the fixed floor $B_{\mathrm{fixed}}$ of (22), collecting the localization term $C_{\mathrm{loc},\kappa}\rho^{\kappa+1}$ and the representation floor $C_{\mathrm{TD},\kappa}\chi_m$, from the four controllable contributions: the optimization horizon

$T$, the TD budget $K$, the actor burn-in $H_{\mathrm{a}}$, and the actor batch $M$. For a target excess error $\epsilon > 0$ it therefore suffices to make the bracket at most $\sqrt{\mu}\,\epsilon$. Splitting $\sqrt{\mu}\,\epsilon$ equally across the four terms, choose integer budgets satisfying

$$\begin{aligned} T &\geq \frac{128\, J_{\mathrm{rng}}}{\alpha\mu\epsilon^2}, & K &\geq \left(\frac{4\sqrt{10}\, C_{\mathrm{TD},\kappa} A_{\mathrm{crit}}}{\sqrt{\mu}\,\epsilon}\right)^4, \\ M &\geq \frac{160\, C_\gamma F_\kappa^2}{\mu\epsilon^2}, & H_{\mathrm{a}} &\geq 2\left\lceil \frac{\log\bigl(\max\{1, 8\sqrt{10}\, F_\kappa/(\sqrt{\mu}\,\epsilon)\}\bigr)}{\log(1/\gamma)} \right\rceil. \end{aligned} \tag{92}$$

Each of the four terms in the bracket is then at most $\sqrt{\mu}\,\epsilon/4$, so the second term of (91) is at most $\epsilon$. Hence, provided the realized trajectory remains in the prescribed policy domain and the remaining hypotheses of Theorem 4.5 hold,

$$\mathbb{E}\left[J_b^\star - J(\theta_\tau)\right] \leq \frac{B_{\mathrm{fixed}}}{\sqrt{\mu}} + \epsilon.$$

The optimization term enters under the square root while the critic and actor-sampling terms enter linearly, so $T$ and $M$ scale with $(\mu\epsilon^2)^{-1}$, $K$ with $(\mu\epsilon^2)^{-2}$, and $H_{\mathrm{a}}$ logarithmically. The budget orders are $T = \mathcal{O}(\epsilon^{-2})$, $K = \mathcal{O}(\epsilon^{-4})$, $M = \mathcal{O}(\epsilon^{-2})$, and $H_{\mathrm{a}} = \mathcal{O}(\log(1/\epsilon))$. The expected number of environment transitions is at most

$$T\left(\frac{K}{1-\gamma} + M + H_{\mathrm{a}}\right), \tag{93}$$

and with the other constants fixed the critic contribution has order $\mathcal{O}(\epsilon^{-6}/(1-\gamma))$.

These choices vary neither the localization radius, nor the network width, nor the projection radius, nor the initialization confidence. Reducing $B_{\mathrm{fixed}}$ itself is a different question. Its critic part $\sqrt{10}\, C_{\mathrm{TD},\kappa}\chi_m/\sqrt{\mu}$ requires width and initialization control: it is at most $\epsilon$ once $\chi_m \leq \sqrt{\mu}\,\epsilon/(\sqrt{10}\, C_{\mathrm{TD},\kappa})$, which under the width-decay rates assumed in Corollary E.6 holds once

$$\begin{aligned} m &\geq \max\left\{ \left(\frac{3\sqrt{10}\, C_{\mathrm{crit}} C_{\mathrm{TD},\kappa}}{\sqrt{\mu}\,\epsilon}\right)^{2/a}, \left(\frac{3\sqrt{10}\, C_{\mathrm{app}} C_{\mathrm{TD},\kappa}}{(1-\sqrt{\gamma})\sqrt{\mu}\,\epsilon}\right)^{1/s} \right\}, \\ \delta_{\mathrm{net}} &\leq \frac{\mu\epsilon^2}{360\, C_{\mathrm{TD},\kappa}^2 Q_{\max}^2}, \end{aligned} \tag{94}$$

splitting $\chi_m$ across its linearization, tangent-approximation, and initialization terms. Its localization part $\sqrt{10}\, C_{\mathrm{loc},\kappa}\rho^{\kappa+1}/\sqrt{\mu}$ decreases geometrically in $\kappa$, and Appendix E.7 gives the radius $\kappa_\star(\epsilon)$ at which it falls below a target together with the per-agent cost that radius carries. Raising $\kappa$ is not free at a fixed architecture and a fixed sample budget: it enlarges the critic input dimension $d_\kappa = n_\kappa d_0$ and the number of distinct local configurations the critic must fit, which the bound sees only through $\varepsilon_{\mathrm{app}}(m)$ inside $\chi_m$; Appendices F.2 and F.3 measure the resulting trade-off directly. The compatible-advantage contribution $C_{\mathrm{hc}}\sqrt{\varepsilon_{\mathrm{bias}}}/\sqrt{\mu}$ is not reducible by any budget. Appendix E.8 records what is and is not available for these quantities.

### E.7 The locality radius required for a nontrivial floor

Corollary 4.6 holds $\kappa$ fixed, so its per-agent accounting is uniform in $N$ by construction. The question that accounting leaves open is how large $\kappa$ must be for the floor itself to be small, and whether the resulting window is still small relative to the network. We answer both in closed form. Since $n_\kappa \leq N$,

$$C_{\mathrm{loc},\kappa} \leq \overline{C}_{\mathrm{loc}} \coloneqq \frac{L_\psi c_{\mathrm{EDP}}}{1-\gamma} + \frac{L_\psi \gamma c_{\mathrm{EDP}}}{(1-\gamma)^2} = \frac{L_\psi c_{\mathrm{EDP}}}{(1-\gamma)^2}, \tag{95}$$

so writing

$$\Gamma_{\mathrm{loc}} \coloneqq \frac{\sqrt{10}\, L_\psi c_{\mathrm{EDP}}}{(1-\gamma)^2\sqrt{\mu}}, \qquad \kappa_\star(\epsilon) \coloneqq \max\left\{ r_R,\ b,\ \left\lceil \frac{\log\bigl(\max\{1, \Gamma_{\mathrm{loc}}/\epsilon\}\bigr)}{\log(1/\rho)} \right\rceil - 1 \right\}, \tag{96}$$

the localization part of $B_{\mathrm{fixed}}/\sqrt{\mu}$ is at most $\epsilon$ for every $\kappa \geq \kappa_\star(\epsilon)$. Combining (96) with the ball bound (42) gives the per-agent cost at that radius.

**Corollary E.19** (Locality at a target localization floor)**.** *Let $\epsilon > 0$ and $\kappa = \kappa_\star(\epsilon)$. On a graph family of maximum degree $\Delta_{\mathcal{G}}$ the critic input dimension satisfies $d_{\kappa_\star} = n_{\kappa_\star} d_0$ with*

$$n_{\kappa_\star} \le \begin{cases} 2\kappa_\star + 1 = \mathcal{O}\left(\dfrac{\log(\Gamma_{\mathrm{loc}}/\epsilon)}{\log(1/\rho)}\right), & \Delta_{\mathcal{G}} = 2, \\ 1 + \dfrac{\Delta_{\mathcal{G}}(\Delta_{\mathcal{G}}-1)^{\kappa_\star}}{\Delta_{\mathcal{G}}-2} = \mathcal{O}\big(\big(\Gamma_{\mathrm{loc}}/\epsilon\big)^{s}\big), \quad s := \dfrac{\log(\Delta_{\mathcal{G}}-1)}{\log(1/\rho)}, & \Delta_{\mathcal{G}} \ge 3. \end{cases} \tag{97}$$

*More generally, if the family has polynomial ball growth $n_\kappa \le c_{\mathcal{G}} \kappa^{D_{\mathcal{G}}}$, then $d_{\kappa_\star} = \mathcal{O}\big(d_0(\log(\Gamma_{\mathrm{loc}}/\epsilon))^{D_{\mathcal{G}}}\big)$.*

Three consequences of this schedule are worth stating explicitly.

First, $\Gamma_{\mathrm{loc}}$ is not uniform in $N$. In the family of Proposition E.14 one has $L_\psi = \sqrt{N}$, $c_{\mathrm{EDP}} = \eta/(1-\gamma)$, $\rho = \gamma$ and $\sqrt{\mu} = 2(1-\zeta\Delta_{\mathcal{G}})p_- p_+/(R\sqrt{N})$, hence

$$\Gamma_{\mathrm{loc}} = \frac{\sqrt{10}\, R\, \eta\, N}{2(1-\zeta\Delta_{\mathcal{G}})p_- p_+ (1-\gamma)^3} = \Theta(N), \qquad \kappa_\star(\epsilon) = \Theta\left(\frac{\log(N/\epsilon)}{\log(1/\gamma)}\right). \tag{98}$$

The return range of that class is $\mathcal{O}(1)$ by (84), so on it a radius growing logarithmically in $N$ already suffices to make Theorem 4.5 informative.

Second, locality nevertheless survives on the graphs this line of work targets. On the ring of (98), $\Delta_{\mathcal{G}} = 2$ and (97) gives $d_{\kappa_\star} = \Theta\big(d_0 \log(N/\epsilon)\big)$, against the $4N$ of a full-input critic; the same holds for grids and, more generally, for any family of polynomial ball growth. On families whose balls grow exponentially the schedule (96) yields $d_{\kappa_\star} = \mathcal{O}\big(d_0(\Gamma_{\mathrm{loc}}/\epsilon)^s\big)$, which is a strict saving over the global input only when

$$\rho\,(\Delta_{\mathcal{G}} - 1) < 1, \tag{99}$$

that is, when the spatial-decay rate beats the branching factor. Condition (99) is checkable from $\gamma$, $b$ and the graph alone through $\rho = \gamma^{1/(1+b)}$, and it is the precise sense in which the finite-hop design is a saving rather than a reparameterization.

Third, the $\sqrt{N}$ in $L_\psi$ is not an artifact of the truncation argument and does not admit a cancellation across agents. The blockwise identity (31) already carries the factor $1/N$ in front of the sum over $j$, so the per-block bias $L_{\psi,i} c_{\mathrm{EDP}} \rho^{\kappa+1}/(1-\gamma)$ is free of $N$; the $\sqrt{N}$ enters only when the $N$ blocks are aggregated in the Euclidean norm through $L_\psi = (\sum_i L_{\psi,i}^2)^{1/2}$, and the exact gradient $\|\nabla J\|_2$ carries the same factor. What makes the floor $N$-dependent is therefore not the localization step but the comparison of an $\ell_2$ stationarity measure against an $\mathcal{O}(1)$ objective range, which enters through $1/\sqrt{\mu} = G_\star/\sqrt{\mu_F}$ with $G_\star = \Theta(\sqrt{N})$ in Proposition E.13. Schedule (96) is the accounting that this scaling forces, not a deficiency of Lemma 3.1.

For fixed $\kappa$ on a bounded-degree graph with uniformly bounded per-node encoding dimension, $n_\kappa$ and $d_\kappa$ are uniformly bounded in $N$. Total computation scales at least linearly in $N$, and no network-wide coordination is used: the projected update (16) is fully decentralized.

### E.8 Representation conditions and their scope

The assumptions of Section 4 are stated at the level of the constants $c_{\mathrm{EDP}}, \rho$ (Definition 2.1), $\mu, C_{\mathrm{hc}}, \varepsilon_{\mathrm{bias}}$ (Assumption 4.1), and $\varepsilon_{\mathrm{TD}}$ (Assumption 4.2). We record here what can and cannot be concluded about their realizability, separating the conditions that are properties of the *architecture* from the one that is a property of the *target*.

The architecture-side conditions are the network-geometry bounds of Assumption E.3: the envelopes $Q_0, G_0$ and the linearization errors $\varepsilon_f(m), \varepsilon_\phi(m)$, holding on an initialization event of probability at least $1 - \delta_{\mathrm{net}}$. These do not refer to the Bellman target, to the policy, or to the graph; they are statements about a depth-$D$, width-$m$ network on a bounded parameter ball, and bounds of exactly this form are available in the neural actor–critic literature. For a depth-$D$ ReLU network constrained to a layerwise Frobenius ball $\|W^{(h)} - W^{(h)}_{\mathrm{init}}\|_{\mathrm{F}} \le B$ of *constant* radius, Gaur et al. (2024) use, from **?**, Lemma F.3, a bound of the form

$$\big|\widehat{Q}_{i,\kappa}(x;W) - \widehat{Q}^{\tan}_{i,\kappa}(x;W)\big| = \widetilde{\mathcal{O}}\Big(m^{-1/6} D^{5/2}\Big) \tag{100}$$

uniformly over inputs and over the ball, on an initialization event whose failure probability is exponentially small in a positive power of the width, and carry it through a multilayer neural actor–critic analysis. This is the quantity (52) asks for, and in the notation of Corollary E.6 it is the case $a = 1/6$ with the depth absorbed into $C_{\mathrm{lin}}$. The companion gradient bound (53) is of the same type, and we state both as hypotheses rather than transcribing a derivation for the parameterization used here, in which the input and output layers are fixed at their random draw and only the hidden matrices are trained. Two features of (100) are worth recording because they shape what the critic floor can be. It is a width rate at a radius that does not grow with the problem, so $B$ must be treated as a fixed architectural choice and not as a quantity enlarged to buy representation power; and it degrades in depth, so $D$ enters the floor adversely, as it already does through $R_0 = B\sqrt{D}$ in (58).

The target-side condition is $\varepsilon_{\mathrm{app}}(m)$ of (54), the error of approximating $Q^{\theta}_{i,\kappa}$ by the tangent class at the chosen radius. This one is not an architectural fact and is not derived here. Remark E.7 explains why: without a regularity condition on the target, a measurable $Q^{\theta}_{i,\kappa}$ admits no uniform width–accuracy rate. The same quantity appears as an assumption in the works this analysis is closest to, in the form $\min_{\theta'} \mathbb{E}[(Q_{\theta'} - T^{\pi}Q_{\theta})^2] \le \varepsilon_{\mathrm{approx}}$ (Gaur et al., 2024, Assumption 4), and likewise in Cai et al. (2019) and **?**. What is specific here is only the target: $\varepsilon_{\mathrm{app}}$ is measured against the *conditional* finite-hop fixed point of Proposition 3.2 rather than against a global Bellman image, so the finite-hop restriction enters the critic problem through which function is being approximated and not through an extra assumption.

Two consequences follow for the floor $\chi_m$ of (58). First, $\chi_m$ does not vanish at fixed width: it is a representation floor, and Corollary 4.6 is stated as accuracy above it rather than as convergence to zero. Second, its dependence on the localization radius is indirect. The constants $A_{\mathrm{crit}}$ and $\chi_m$ are built from $B$, $D$, $Q_0$, $G_0$ and the linearization errors, none of which mentions $\kappa$; the radius enters the critic input dimension $d_\kappa = n_\kappa d_0$, and therefore the difficulty of the regression at a fixed sample budget, only through $\varepsilon_{\mathrm{app}}(m)$ and the constants $Q_0, G_0$ that the encoding scale fixes. The behaviour in $\kappa$ that Appendix F measures is therefore a finite-budget effect that this floor does not resolve on its own.

Because spatial decay does not by itself force the compatible-advantage error to vanish, Theorem 4.5 retains $\varepsilon_{\mathrm{bias}}$ explicitly, and Corollary 4.6 does not assert that the error floor can be driven to zero on every instance.

# F Experimental Details

## F.1 Implementation and protocol

The first three paragraphs cover the synthetic rings of Section 5.1; the last covers the CSMA environment and the topologies of Section 5.2.

**Learners.** *Local* trains a private ReLU multilayer perceptron (MLP) critic per agent on its own $\kappa$-hop encoded inputs and takes the localized actor step of Section 3.3. ***FI/LA*** and ***FI/FA*** differ from it in the critic input alone, or in the critic input and the actor aggregation radius, respectively. Every other component is shared, namely the same environment instances and seeds, the same sampling scheme, the same critic architecture family, and the same $T$, $K$, $M$ and $H_{\mathrm{a}}$.

**Controlled ring MDP.** The environment has binary local states and actions, discount $\gamma = 0.9$, one-hop transition coupling of strength $0.8$, and reward radius $r_R = 1$. Critics have width $64$ and depth $3$ and use $K = 200$ TD updates per outer iteration at step size $0.05$. Restart-chain burn-in is $20$, the actor batch is $M = 64$, and the actor runs for $400$ iterations with constant step $0.4$. Each of the $20$ seeds determines both a generated MDP instance and the optimization randomness, so the intervals reflect both instance and training variability. Means and paired differences use bootstrap $95\%$ confidence intervals with $10{,}000$ resamples.

For each instance, finite-state enumeration computes $J(\theta)$ exactly for any specified policy. The reference policy is the best result of four exact-gradient ascent restarts, each run for $3000$ iterations within $\Theta_b$; denote its return by $J_{\mathrm{ref}}$. We report the exactly evaluated reference gap $J_{\mathrm{ref}} - J(\theta_T)$. Because $J_{\mathrm{ref}}$ is the best value found by this procedure, the reference gap may be negative.

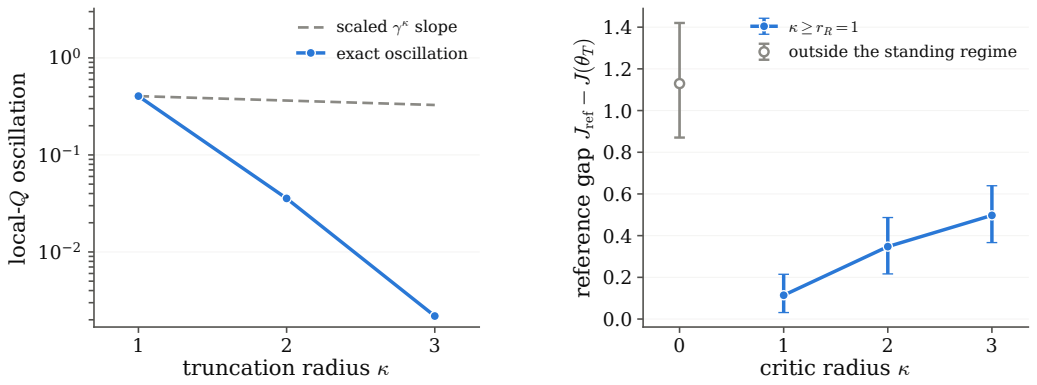
**Large-$N$ comparison protocol.** The sampled-ring study uses $N \in \{8, 16, 32, 64\}$, width 64, depth 3, $K = 200$, $M = 64$, and $T = 400$. Each final return averages 20 episodes of horizon 200, and reported intervals use the paired seeds 0–19; the seeds 20–24 are held out and used only for step-size selection. All runs of this study were executed on NVIDIA H100 NVL GPUs.

**Wireless random access and structural accounting.** The large-$N$ environment is a CSMA-type process: queued neighboring transmitters collide, while an isolated transmission succeeds and receives reward one. Write $\Delta_{\mathcal{G}} := \max_j \deg(j)$ for the largest node degree, as in Corollary 4.6 and (42); this is distinct from the per-node encoding dimension $d_0$ of Section 3.3. For $\kappa = 1$ and binary one-hot state/action encodings, the maximum zero-padded critic input is $d_\kappa = 4(\Delta_{\mathcal{G}} + 1)$, and an actor batch consumes at most $M\Delta_{\mathcal{G}}$ remote critic scalars. These graph-derived worst-agent quantities count critic scalars received from other agents. Monte Carlo returns use horizon 200 and 20 episodes. In the $N = 20$ radius sweep, the 20 paired seeds vary training and evaluation randomness on the same CSMA ring.

### F.2 Solution-quality and learning diagnostics

**Spatial-decay and radius diagnostics.** For each seed, Figure 3a evaluates a separately initialized fixed policy. Exhaustive enumeration computes, for each agent, the largest oscillation of $Q_i^\theta$ over global state–action pairs that agree on the $\kappa$-hop ball; the plot averages these exact per-agent quantities over agents and seeds. This estimates mean fixed-policy oscillation over the sampled policies; Definition 2.1 instead takes a uniform bound over policies. A geometric fit over $\kappa = 1, 2, 3$ gives $\widehat{\rho} \approx 0.07$. The dashed line shows the finite-propagation reference slope $\gamma^\kappa$ scaled at $\kappa = 1$; these instances use $b = 0$, for which Proposition B.1 gives $\rho = \gamma$. The logarithmic panel omits the exactly zero value at $\kappa = 4$.

At a fixed critic budget, the reference gap in Figure 3b falls from 1.13 at the out-of-regime point $\kappa = 0$ to 0.11 at $\kappa = 1$, then rises to 0.35 and 0.50 at $\kappa = 2, 3$. The pattern is consistent with reduced truncation bias followed by increased finite-budget estimation difficulty.




(a) Fixed-policy local-$Q$ oscillation.

(b) Finite-budget radius tradeoff.

Figure 3: Spatial-decay diagnostics on the controlled $N = 8$ ring. Each plotted quantity is evaluated exactly for its specified policy; error bars in (b) are bootstrap 95% confidence intervals over 20 seeds, each defining an MDP instance and optimization randomness.

**Localization radius on the $N = 20$ CSMA ring.** Figure 4 gives a second radius sweep. The mean final return at $\kappa = 1$ is 2.348 $[2.262, 2.436]$; its paired differences from $\kappa = 2$ and $\kappa = 3$ are 0.288 $[0.205, 0.385]$ and 0.476 $[0.399, 0.558]$, respectively. The $\kappa = 0$ point lies outside the standing requirement $\kappa \geq \max\{r_R, b\}$, which equals 1 here since $b = 0$ and $r_R = 1$.

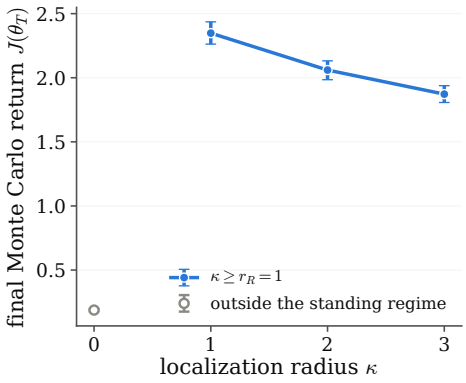



Figure 4: Localization-radius sweep on the fixed $N = 20$ CSMA ring (20 paired seeds; pointwise bootstrap 95% confidence intervals). Here $\kappa$ controls both the critic window and the actor aggregation, and $\kappa = 0$ is outside the $r_R = 1$ standing regime.

### F.3 Direct measurement of the critic risk

We measure the critic risk $\|\widehat{Q}_{i,\kappa} - Q^{\theta}_{i,\kappa}\|_{L_2(\nu_\theta)}$ against the *exact* conditional Bellman fixed point. On the $N = 5$ ring the joint chain is small enough to enumerate, so for a fixed $\theta$ we form the discounted occupancy $\nu_\theta$ in closed form, build the stationary conditional law from it, and solve the resulting conditional Bellman system as a linear fixed point. Two checks confirm the solver: at $\kappa$ equal to the full radius the computed $Q^{\theta}_{i,\kappa}$ reproduces the global $Q^{\theta}_i$ to machine precision ($\sim 10^{-15}$), and the induced localization bias decays geometrically in $\kappa$ at a ratio close to the $\hat{\rho}$ measured independently in Figure 3a. Results average 20 seeds, the same seed budget as the rest of the appendix. All critic-risk numbers below are computed from this exact construction; we do not report the corresponding quantities at sizes where the conditional Bellman system cannot be solved exactly.

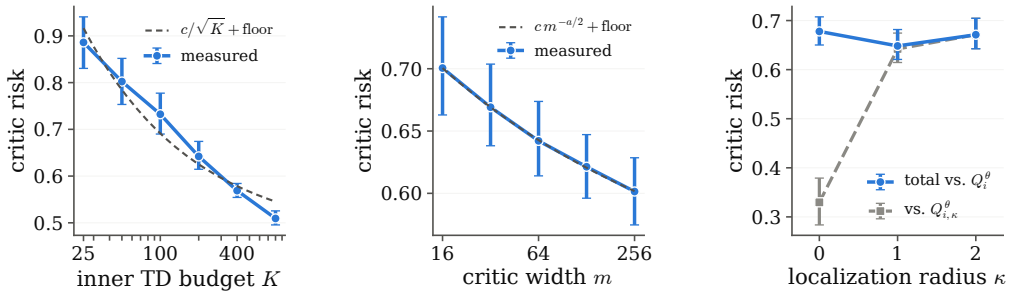
Figure 5 reports the $K$, $m$, and $\kappa$ slices. The risk decreases monotonically in the inner TD budget $K$ ($0.886 \to 0.509$ from $K = 25$ to $K = 800$) and in width $m$ ($0.701 \to 0.601$ from $m = 16$ to $m = 256$); it also decreases in depth $D$ ($0.676 \to 0.612$ from $D = 2$ to $D = 4$), which we measure but do not plot. The dashed $c/\sqrt{K} + \text{floor}$ and $cm^{-a/2} + \text{floor}$ curves are descriptive least-squares references for these finite sweeps.

The $\kappa$ slice illustrates the estimation–localization trade-off in the error decomposition. The estimation error against $Q^{\theta}_{i,\kappa}$ *grows* with $\kappa$ ($0.330 \to 0.671$), because a larger window is a higher-dimensional regression at a fixed sample budget, while the localization bias shrinks. The total error against the global $Q^{\theta}_i$ is smallest at $\kappa = 1$ (0.648, against 0.678 at $\kappa = 0$ and 0.671 at $\kappa = 2$); the differences are small, with $15/20$ seeds favouring $\kappa = 1$ in each comparison. At $\kappa = 2$, the full window for $N = 5$, the two curves coincide and recover the full-window limit. These are fixed-policy measurements of the critic alone; they are not an evaluation of the actor iterate that Theorem 4.5 bounds.

### F.4 A window-size sweep: the structural accounting

The benchmarks above keep the finite-hop window small: at $\kappa = 1$ on a binary ring it has $(|\mathcal{S}_i||\mathcal{A}_i|)^{n_\kappa} = 64$ configurations at every $N$, so increasing the number of agents does not enlarge what a single critic has to represent. This subsection varies the per-agent state cardinality instead, which grows the window geometrically while leaving the graph untouched, and asks whether the finite-hop input is still the better one once the window is large.

**Environment.** The multi-level-queue CSMA process places a queue $q_i \in \{0, \dots, Q_{\text{cap}}\}$ at each node of a ring, with $a_i \in \{0, 1\}$ and $\mu_0$ the product of uniform laws on $\{0, \dots, Q_{\text{cap}}\}$. Writing $\text{att}_i = a_i \mathbf{1}\{q_i \geq 1\}$, a node succeeds when it attempts alone among its neighbours, serves $\min\{q_i, B_{\text{serve}}\}$ packets on success, then receives one arrival with probability $\lambda_{\text{a}}$ and drops it if the




(a) Risk versus inner TD budget $K$. (b) Risk versus critic width $m$. (c) The $\kappa$ trade-off.

Figure 5: Measured critic risk against the exact conditional Bellman fixed point on the $N = 5$ ring, 20 seeds. Dashed curves in (a)–(b) are descriptive least-squares references. In (c), estimation error grows with the input window while total error is minimized at an interior $\kappa$. Bars are bootstrap 95% confidence intervals of the mean.

Table 2: Window cardinality against encoded critic input dimension on the multi-level-queue ring at $N = 64$ and $\kappa = 1$. "Window" counts the distinct $\kappa$-hop state–action configurations, $(2(Q_{\text{cap}} + 1))^{n_\kappa}$ with $n_\kappa = 3$; $d_\kappa = n_\kappa(Q_{\text{cap}} + 3)$ is the zero-padded encoded input the critic actually reads, against $N(Q_{\text{cap}} + 3)$ for a critic given the whole graph.

| $Q_{\text{cap}}$ | $\lvert\mathcal{S}_i\rvert$ | Window configurations | $d_\kappa$ | full-input dim. |
|---|---|---|---|---|
| 1 | 2 | 64 | 12 | 256 |
| 7 | 8 | 4096 | 30 | 640 |
| 31 | 32 | 262144 | 102 | 2176 |

queue is full:

$$r_i = \frac{\min\{q_i, B_{\text{serve}}\}}{B_{\text{serve}}}\,\text{succ}_i - c_{\text{h}}\frac{q_i}{Q_{\text{cap}}} - c_{\text{d}}\,\text{drop}_i,$$

where $c_{\text{h}}$ is the per-unit holding cost and $c_{\text{d}}$ the cost of a dropped arrival. The simulator returns the realized value of this expression, whose only randomness given $(q, a)$ is the arrival indicator inside $\text{drop}_i$; the stage reward of Section 2.1 is its conditional expectation

$$r_i(q_{\mathcal{N}_i}, a_{\mathcal{N}_i}) \coloneqq \mathbb{E}\big[\,\cdot \;\big|\; q, a\,\big] = \frac{\min\{q_i, B_{\text{serve}}\}}{B_{\text{serve}}}\,\text{succ}_i - c_{\text{h}}\frac{q_i}{Q_{\text{cap}}} - c_{\text{d}}\lambda_{\text{a}}\mathbf{1}\{q_i^{\text{post}} = Q_{\text{cap}}\},$$

where $\text{succ}_i$ and the post-service level $q_i^{\text{post}}$ are deterministic functions of $(q_{\mathcal{N}_i}, a_{\mathcal{N}_i})$. Every term is therefore a function of $(q_{\mathcal{N}_i}, a_{\mathcal{N}_i})$, so $r_R = 1$, the transition factorizes as in (1), and $R_{\max} = 1 + c_{\text{h}} + c_{\text{d}}$. The realized value is a bounded unbiased sample of $r_i$, so the conditional Bellman operator (11) and the TD analysis are unaffected by which of the two the simulator reports. At $Q_{\text{cap}} = 1$, $c_{\text{h}} = c_{\text{d}} = 0$, and $B_{\text{serve}} = 1$ the process reduces to the binary random-access model of Appendix F. The runs below use $\lambda_{\text{a}} = 0.6$, $B_{\text{serve}} = 2$, $c_{\text{h}} = 0.3$, and $c_{\text{d}} = 1.0$, chosen so that the queue level enters the optimal localized policy.

**Protocol.** Only $Q_{\text{cap}}$ varies across the sweep and only the critic input scope varies between Local and FI/LA, with the actor loop, the sampling scheme, and every budget held fixed at $\kappa = 1$, $\gamma = 0.9$, $T = 300$, $K = 200$, $M = 64$, evaluation horizon 200 over 20 episodes, and 20 paired seeds. The metric is the final Monte Carlo return; exact enumeration is infeasible at these sizes.

**The structural measurement.** Table 2 records how the two quantities separate at $N = 64$. The number of distinct $\kappa$-hop configurations grows as $(2(Q_{\text{cap}} + 1))^{n_\kappa}$, geometrically in the per-agent state cardinality, while the encoded critic input grows only as $d_\kappa = n_\kappa(Q_{\text{cap}} + 3)$. At $Q_{\text{cap}} = 31$ these are 262144 against 102. This is a property of the window and the encoding, and does not depend on any learning outcome.

**Locality is not a constraint at the largest window.** At $Q_{\mathrm{cap}} = 31$ and $N = 64$ Local with $d_\kappa = 102$ attains $-1.0253 \pm 0.0038$ against $-1.1158 \pm 0.0034$ for FI/LA, whose input is the whole graph at $d_\kappa = 2176$, a paired advantage of $0.0905$ that is positive on all $20$ seeds with minimum $0.0545$. Giving the critic the entire configuration therefore does not pay for its input dimension at this budget.

**What the sweep establishes.** The comparison of Section 5.1 was made on a window of $64$ state–action configurations, where the finite-hop restriction discards little. Here the window carries $262144$ configurations and the conclusion is unchanged: the $\kappa$-hop input remains the better one, at an encoded dimension of $102$ against $2176$. This is the per-agent reading of Corollary 4.6 at a window that is far too large to be traversed at the sampling budget used, so the critic has to generalize across local configurations rather than record them.

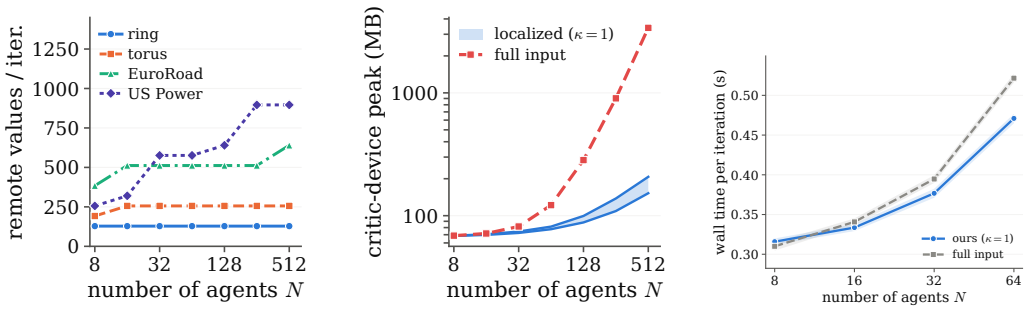
### F.5 Topology construction and worst-agent accounting

The torus uses the nearest factorization of $N$ into a periodic rectangular grid; at $N = 8$ its $2 \times 4$ geometry has three distinct neighbors because the two vertical wrap-around neighbors coincide, while $N \geq 16$ has degree four. EuroRoad and US Power are taken from a public network repository (Rossi & Ahmed, 2015); for each we start in the largest connected component and take a connected induced breadth-first search (BFS) subgraph of the requested size. Irregular local encodings are zero-padded to the largest $\kappa$-hop input used by the batched critic.

Table 3 and Figure 6 summarize the resulting locality and implementation costs. The table also shows the limit of the uniform-in-$N$ reading: the ring and the torus keep $\Delta_{\mathcal{G}}$, and hence $d_\kappa$, fixed as $N$ grows, whereas the two repository graphs do not, their largest degree rising from $8$ and $9$ at $N = 64$ to $10$ and $14$ at $N = 512$. End to end on the ring, the measured critic-side peak device memory grows from $68$ to $152$ MB between $N = 8$ and $N = 512$ for the localized implementation, against $69$ to $3383$ MB for the full-input one (Figure 6b), and the wall time per actor iteration at $N = 64$ is $1.108$ times the localized one for the full input (Figure 6c).

Table 3: Representative worst-agent locality accounting at $\kappa = 1$ and $M = 64$. "Remote values" counts critic values received from other agents per iteration.

| Topology | $N$ | $\Delta_{\mathcal{G}}$ | $d_\kappa$ | remote values / iter. |
|---|---|---|---|---|
| Ring | 64 | 2 | 12 | 128 |
| Ring | 512 | 2 | 12 | 128 |
| 2D torus | 64 | 4 | 20 | 256 |
| 2D torus | 512 | 4 | 20 | 256 |
| EuroRoad | 64 | 8 | 36 | 512 |
| EuroRoad | 512 | 10 | 44 | 640 |
| US Power | 64 | 9 | 40 | 576 |
| US Power | 512 | 14 | 60 | 896 |
| Full input | 64 | – | 256 | – |
| Full input | 512 | – | 2048 | – |




(a) Finite-hop messages. (b) Critic-device peak memory. (c) Wall time per actor iteration.

Figure 6: Locality and implementation measurements. (a) Worst-agent remote finite-hop critic values per actor iteration at $\kappa = 1$ and $M = 64$; the $N = 512$ counts are 128, 256, 640, and 896 on the ring, torus, EuroRoad, and US Power, respectively. Neighborhood observations used to construct critic inputs are counted separately in Appendix D. (b) Measured critic-device peak memory for the localized implementation on four topologies and the full-input critic side on the ring (mean over seeds). From $N = 8$ to $N = 512$, the localized ring critic side grows from 68 to 152 MB, while the full-input critic side grows from 69 to 3383 MB. (c) CUDA-synchronized wall time for the matched local/full-input pair on sampled rings; bars are pointwise bootstrap 95% confidence intervals over 20 seeds. At $N = 64$, the full-input mean is 1.108 times the localized mean (95% CI $[1.089, 1.128]$). Times cover the co-located simulator and training loop; distributed network latency is not included.